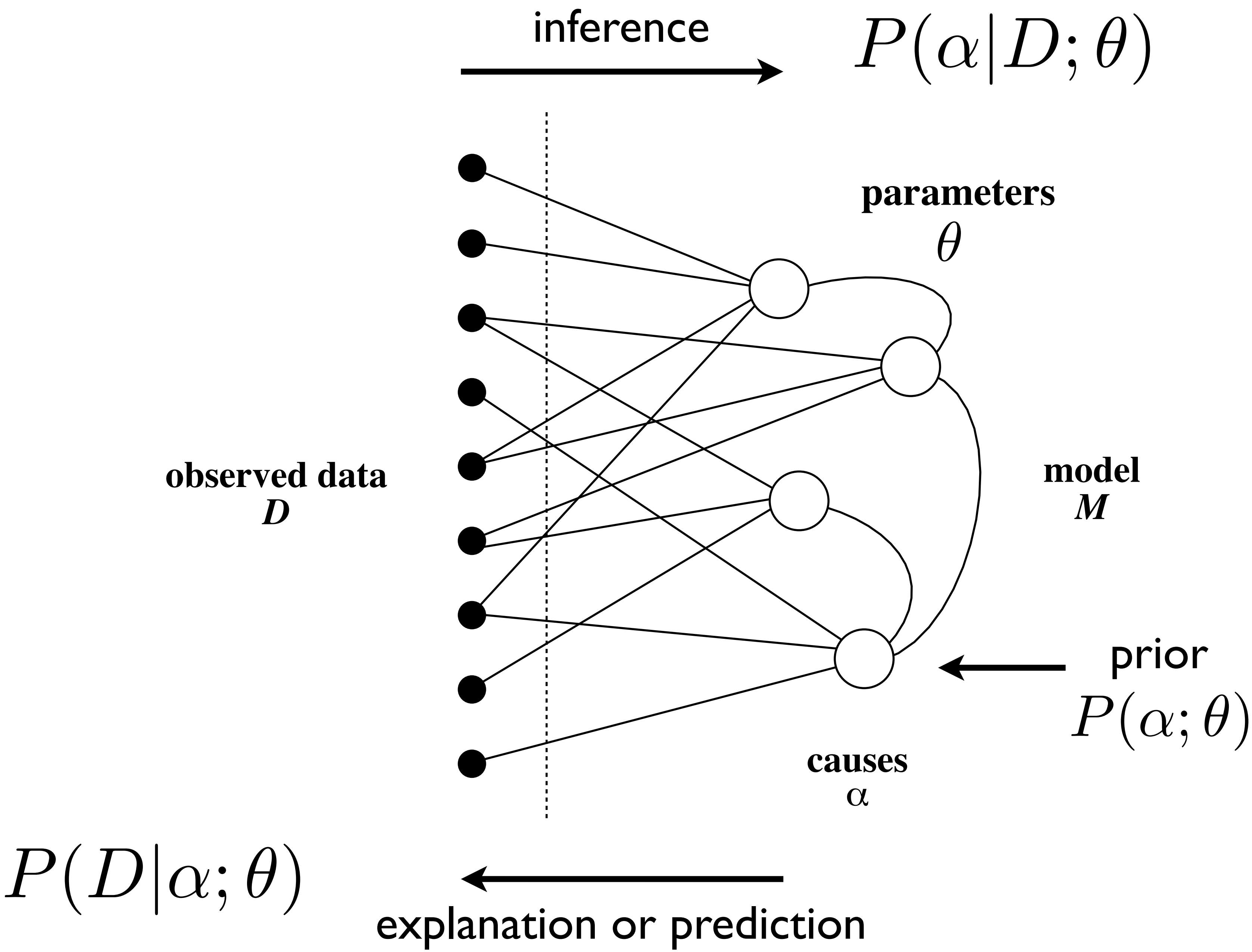
Generative models



Inference:

$$P(\alpha|D;\theta) \propto P(D|\alpha;\theta)\,P(\alpha;\theta)$$

Explanation or prediction:

$$P(D|\hat{\alpha};\theta) \quad \textbf{with} \quad \hat{\alpha} = \arg\max_{\alpha} P(\alpha|D;\theta)$$

Objective for learning:

$$\hat{\theta} = \arg\max_{\theta} \langle \log P(D|\theta) \rangle$$

$$P(D|\theta) = \sum_{\alpha} P(D|\alpha;\theta)\,P(\alpha;\theta)$$

We can keep on going…

$$P(\theta|D) = \frac{\overbrace{P(D|\theta)}^{\text{likelihood}}\,\overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(D)}_{\text{evidence}}}$$

$$P(D) = \int P(D|\theta)\,P(\theta)\,d\theta$$

David MacKay Ph.D. thesis (1991)

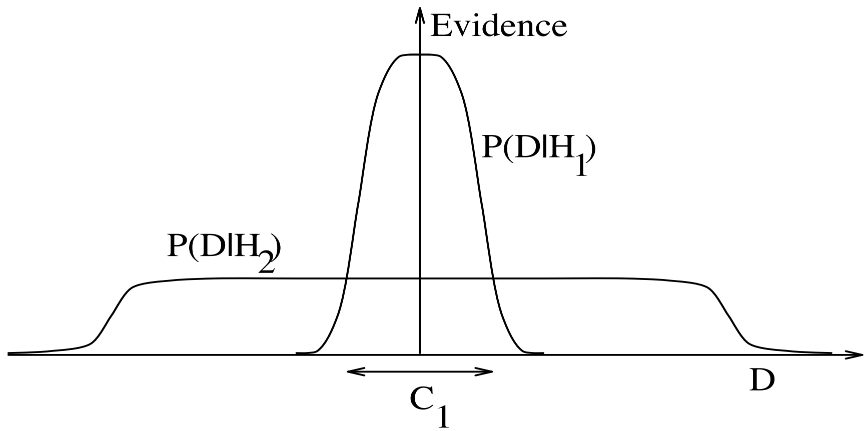


Figure 2.2: **Why Bayes embodies Occam's razor**

This figure gives the basic intuition for why complex models are penalised. The horizontal axis represents the space of possible data sets D. Bayes' rule rewards models in proportion to how much they *predicted* the data that occurred. These predictions are quantified by a normalised probability distribution on D. In this paper, this probability of the data given model $\mathcal{H}_i$, $P(D|\mathcal{H}_i)$, is called the evidence for $\mathcal{H}_i$.

A simple model $\mathcal{H}_1$ makes only a limited range of predictions, shown by $P(D|\mathcal{H}_1)$; a more powerful model $\mathcal{H}_2$, that has, for example, more free parameters than $\mathcal{H}_1$, is able to predict a greater variety of data sets. This means however that $\mathcal{H}_2$ does not predict the data sets in region $\mathcal{C}_1$ as strongly as $\mathcal{H}_1$. Assume that equal prior probabilities have been assigned to the two models. Then if the data set falls in region $\mathcal{C}_1$, the *less powerful* model $\mathcal{H}_1$ will be the *more probable* model.

David MacKay Ph.D. thesis (1991)

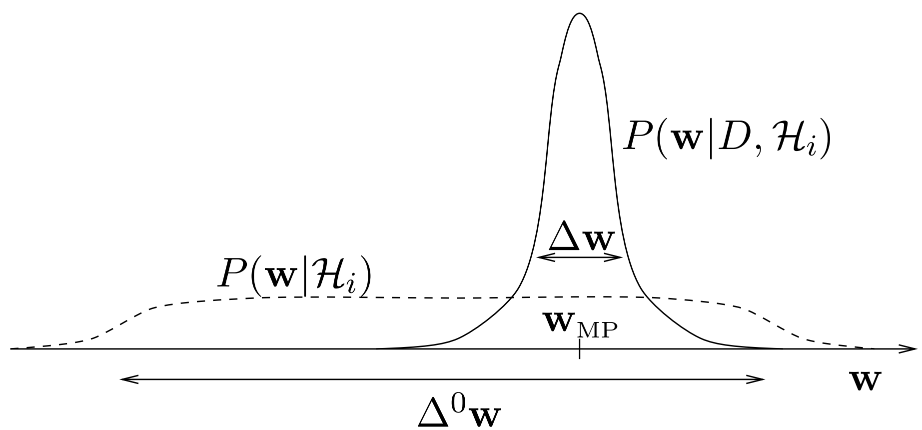


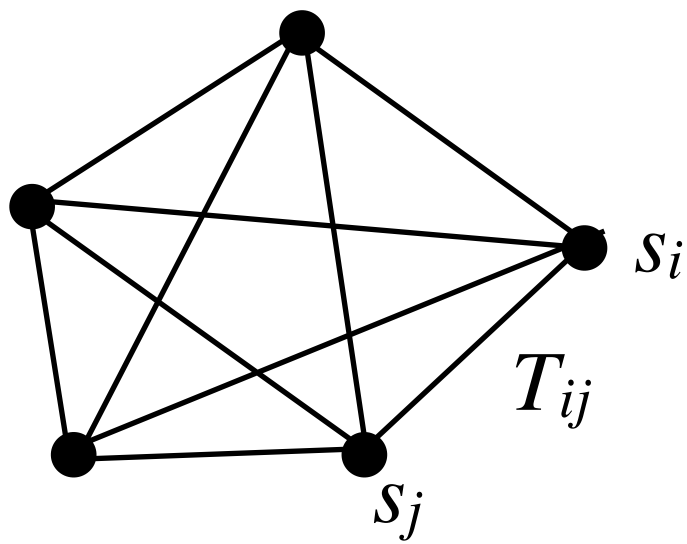
Figure 2.3: **The Occam factor**

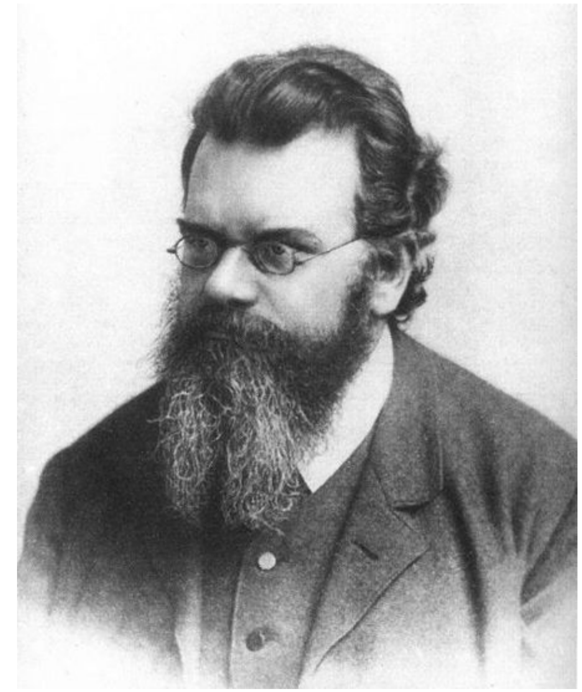
This figure shows the quantities that determine the Occam factor for a hypothesis $\mathcal{H}_i$ having a single parameter $\mathbf{w}$. The prior distribution (dotted line) for the parameter has width $\Delta^0\mathbf{w}$. The posterior distribution (solid line) has a single peak at $\mathbf{w}_{\mathrm{MP}}$ with characteristic width $\Delta\mathbf{w}$. The Occam factor is $\frac{\Delta\mathbf{w}}{\Delta^0\mathbf{w}}$.

$$\underset{\text{Evidence}}{P(D\,|\mathcal{H}_i)} \simeq \underbrace{P(D\,|\mathbf{w}_{\mathrm{MP}},\mathcal{H}_i)}_{\text{Best fit likelihood}} \; \underbrace{P(\mathbf{w}_{\mathrm{MP}}|\mathcal{H}_i)\,\Delta\mathbf{w}}_{\text{Occam factor}}. \tag{2.5}$$

$$\boxed{\text{Occam factor} = \frac{\Delta\mathbf{w}}{\Delta^0\mathbf{w}}}$$

The "Boltzmann machine"
(Hinton & Sejnowski, 1983)





Ludwig Boltzmann

$$E(\mathbf{s}) \quad = \quad -\frac{1}{2}\sum_{ij} T_{ij}\, s_i\, s_j$$

$$P(\mathbf{s}) \quad = \quad \frac{1}{Z} e^{-\beta\, E(\mathbf{s})}$$

Boltzmann-Gibbs distribution

$$
\begin{aligned}
P(\mathbf{x}) &= \frac{1}{Z} e^{\lambda\,\phi(\mathbf{x})} \\
Z &= \int e^{\lambda\,\phi(\mathbf{x})}\, d\mathbf{x}
\end{aligned}
$$

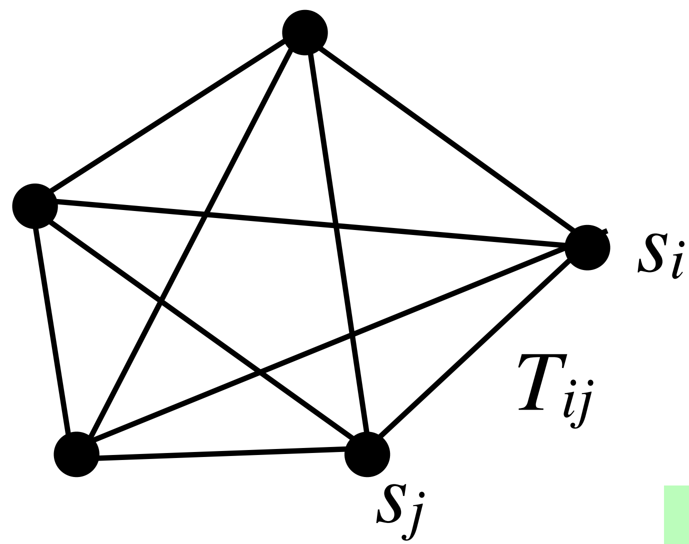
Learning rule:

$$
\begin{aligned}
\Delta\lambda &\propto \frac{\partial}{\partial\lambda}\langle \log P(\mathbf{x})\rangle \\
&= \langle \phi(\mathbf{x})\rangle - \langle \phi(\mathbf{x})\rangle_{P(\mathbf{x})}
\end{aligned}
$$

$$\log P(\mathbf{x}) = \lambda\phi(\mathbf{x}) - \log Z$$

$$\begin{aligned}
\frac{\partial}{\partial\lambda}\langle\log P(\mathbf{x})\rangle &= \frac{\partial}{\partial\lambda}\langle\lambda\,\phi(\mathbf{x}) - \log Z\rangle \\
&= \langle\phi(\mathbf{x}) - \frac{\partial}{\partial\lambda}\log Z\rangle \\
&= \langle\phi(\mathbf{x}) - \frac{1}{Z}\frac{\partial Z}{\partial\lambda}\rangle \\
&= \langle\phi(\mathbf{x}) - \frac{1}{Z}\int\phi(\mathbf{x})\,e^{\lambda\,\phi(\mathbf{x})}d\mathbf{x}\rangle \\
&= \langle\phi(\mathbf{x}) - \int\phi(\mathbf{x})\,P(\mathbf{x})d\mathbf{x}\rangle \\
&= \langle\phi(\mathbf{x})\rangle - \langle\phi(\mathbf{x})\rangle_{P(\mathbf{x})}
\end{aligned}$$

The "Boltzmann machine"
(Hinton & Sejnowski, 1983)



λ_{ij} $\phi_{ij}(\mathbf{s})$

$$E(\mathbf{s}) = -\frac{1}{2}\sum_{ij} T_{ij}\, s_i\, s_j$$

$$P(\mathbf{s}) = \frac{1}{Z} e^{-\beta\, E(\mathbf{s})}$$

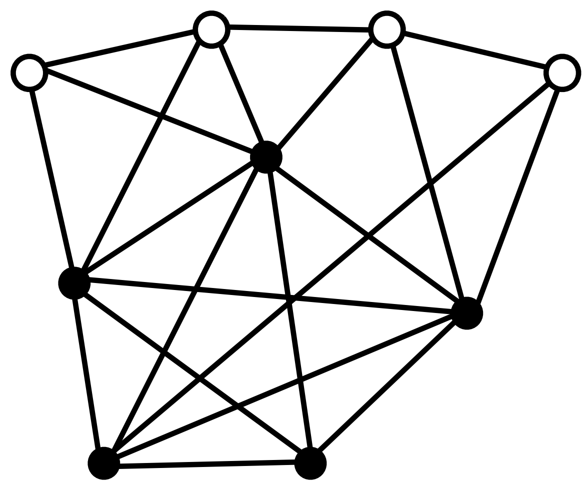
The Boltzmann machine learning rule

$$\begin{aligned} \Delta T_{ij} &\propto \frac{\partial \langle \log P(\mathbf{s}) \rangle}{\partial T_{ij}} \\ &= \beta \left[\langle s_i \, s_j \rangle_{\text{clamped}} - \langle s_i \, s_j \rangle_{\text{free}} \right] \end{aligned}$$

data

$P(\mathbf{s})$

“Boltzmann machine” with hidden units
(Hinton & Sejnowski)



‘hidden’ units, s^h

‘visible’ units, s^v

$$E(\mathbf{s}^v, \mathbf{s}^h) = -\sum_{i,j} T^{vv}_{ij}\, s^v_i s^v_j - \sum_{i,j} T^{vh}_{ij}\, s^v_i s^h_j - \sum_{i,j} T^{hh}\, s^h_i s^h_j$$

$$P(\mathbf{s}^v, \mathbf{s}^h) = \frac{1}{Z} e^{-E(\mathbf{s}^v, \mathbf{s}^h)}$$

$$P(\mathbf{s}^v) = \sum_{\mathbf{s}^h} P(\mathbf{s}^v, \mathbf{s}^h)$$

The Boltzmann machine learning rule

$$\begin{aligned}\Delta T_{ij} &\propto \frac{\partial \log P(\mathbf{s})}{\partial T_{ij}} \\ &= \beta \left[\langle s_i \, s_j \rangle_{\text{clamped}} - \langle s_i \, s_j \rangle_{\text{free}} \right]\end{aligned}$$

Clamped: $\left\{ \begin{array}{lcc} \mathbf{s}^v & = & \mathbf{x} \\ \mathbf{s}^h & \sim & P(\mathbf{s}^h | \mathbf{s}^v) \end{array} \right.$

Free: $[\mathbf{s}^v, \mathbf{s}^h] \sim P(\mathbf{s}^v, \mathbf{s}^h) \equiv P(\mathbf{s})$

Gibbs sampling

To sample from $P(\mathbf{x})$:

$$x_1 \sim P(x_1|x_2, ..., x_n)$$

$$x_2 \sim P(x_2|x_1, x_3, ..., x_n)$$

$$x_3 \sim P(x_3|x_1, x_2, x_4, ..., x_n)$$

$$\vdots$$

$$x_n \sim P(x_n|x_1, ..., x_{n-1})$$

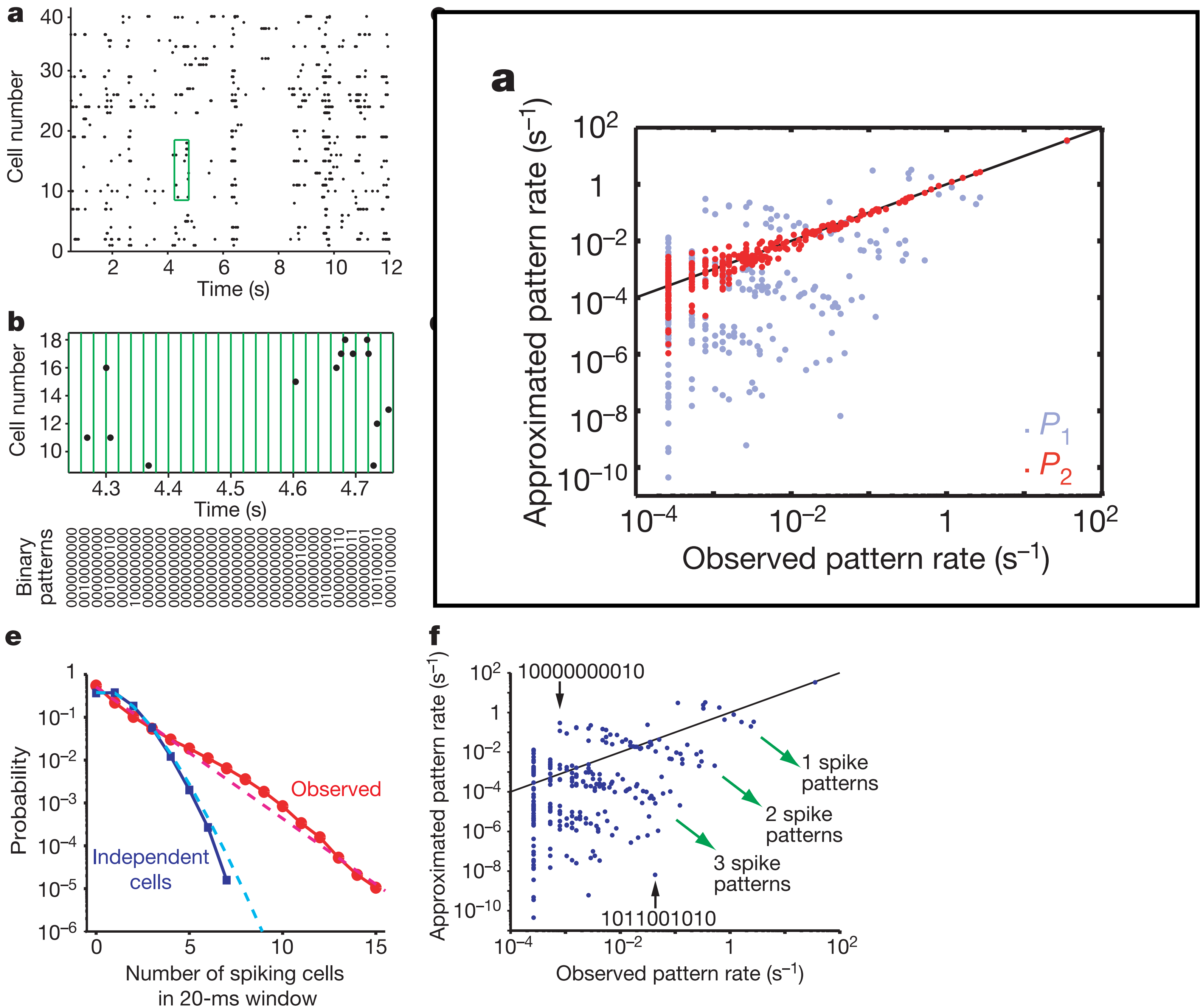
Dynamics

$$
\begin{aligned}
P(s_i = 1 | \{s_{\bar{i}}\}) &= \frac{P(s_i = 1, \{s_{\bar{i}}\})}{P(s_i = 1, \{s_{\bar{i}}\}) + P(s_i = -1, \{s_{\bar{i}}\})} \\
&= \frac{e^{\beta \sum_{j \neq i} T_{ij}\, s_j}}{e^{\beta \sum_{j \neq i} T_{ij}\, s_j} + e^{-\beta \sum_{j \neq i} T_{ij}\, s_j}} \\
&= \frac{1}{1 + e^{-2\beta \sum_{j \neq i} T_{ij}\, s_j}}
\end{aligned}
$$

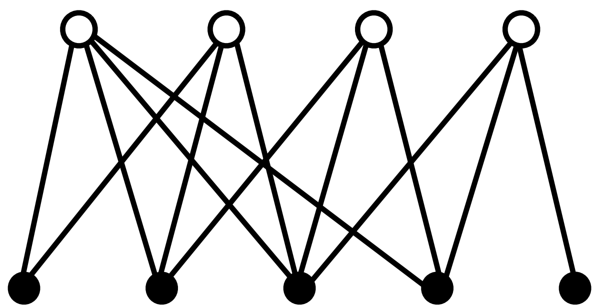
Thus:

$$
\begin{aligned}
P(s_i = 1 | \{s_{\bar{i}}\}) &= \sigma(2\beta\, h_i) \\
h_i &= \sum_{j \neq i} T_{ij}\, s_j
\end{aligned}
$$

Application: modeling activity of neural populations (Schneidman et al.)



Restricted Boltzmann machine (RBM)



‘hidden’ units, s^h

‘visible’ units, s^v

$$E(\mathbf{s}^v, \mathbf{s}^h) = -\sum_{i,j} T^{vh}_{ij}\; s^v_i\; s^h_j$$

$$P(\mathbf{s}^v, \mathbf{s}^h) = \frac{1}{Z} e^{-E(\mathbf{s}^v, \mathbf{s}^h)} \quad \Longrightarrow$$

$$P(s^h_i | s^h_{\bar{i}}, \mathbf{s}^v) = \sigma(\sum_j T^{hv}_{ij}\; s^v_j)$$

$$P(s^v_i | s^v_{\bar{i}}, \mathbf{s}^h) = \sigma(\sum_j T^{vh}_{ij}\; s^h_j)$$

$$P(\mathbf{s}^v) = \sum_{\mathbf{s}^h} P(\mathbf{s}^v, \mathbf{s}^h) \; = \; \frac{1}{Z} e^{\sum_i \log(1+e^{T^{vh}_i \cdot \mathbf{s}^v})}$$

OPEN ACCESS Freely available online

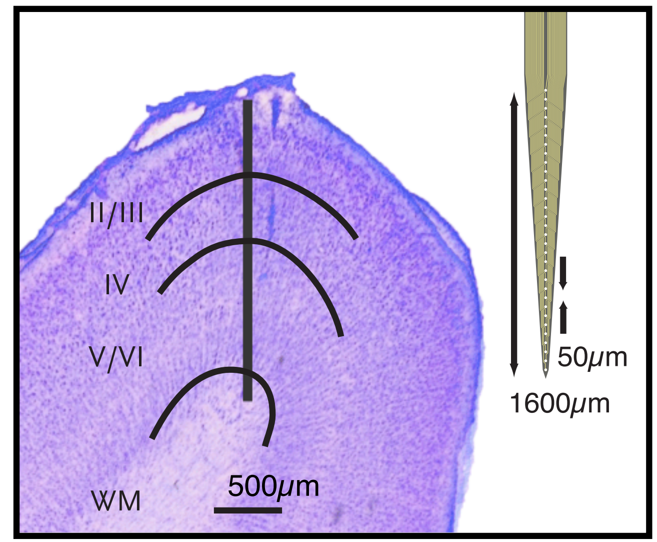
PLOS COMPUTATIONAL BIOLOGY

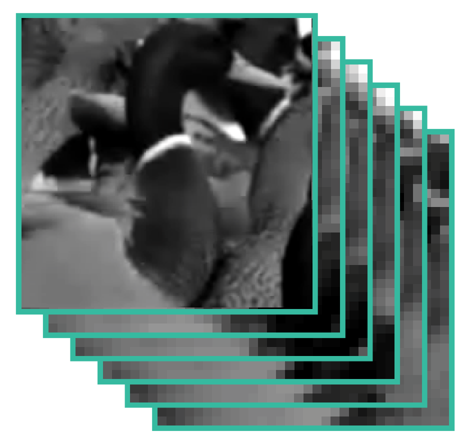
Modeling Higher-Order Correlations within Cortical Microcolumns

Urs Köster[1]*, Jascha Sohl-Dickstein[2], Charles M. Gray[3], Bruno A. Olshausen[1]

1 Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Berkeley, California, United States of America, **2** Department of Applied Physics, Stanford University and Khan Academy, Palo Alto, California, United States of America, **3** Department of Cell Biology and Neuroscience, Montana State University, Bozeman, Montana, United States of America

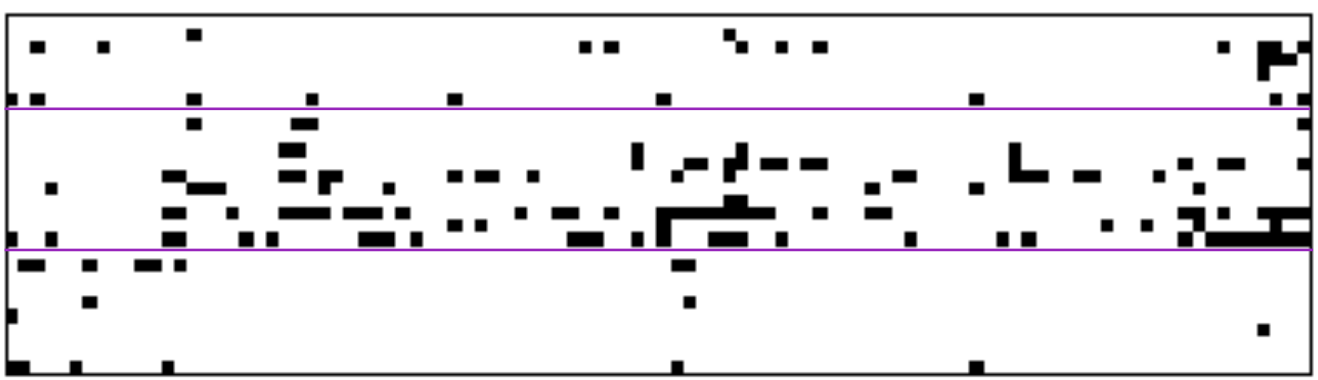
- Silicon polytrode, 32 channels span all laminae
- Anesthetized cat area V1
- Stimuli consist of natural scene movies at 150 fps



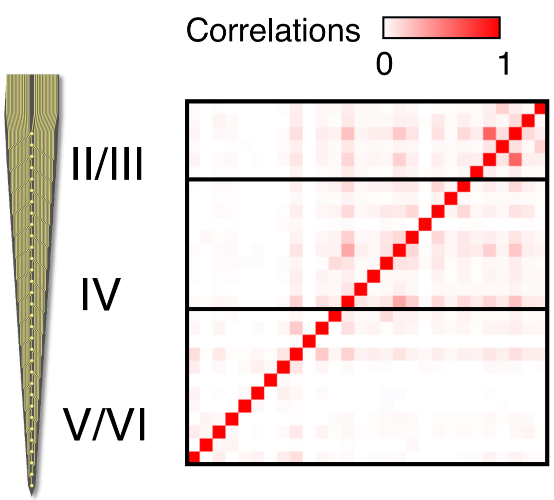


Experiment

- Simultaneous recordings of 20-40 well-isolated cells across all cortical layers of V1
- How can we discover patterns in the activity?



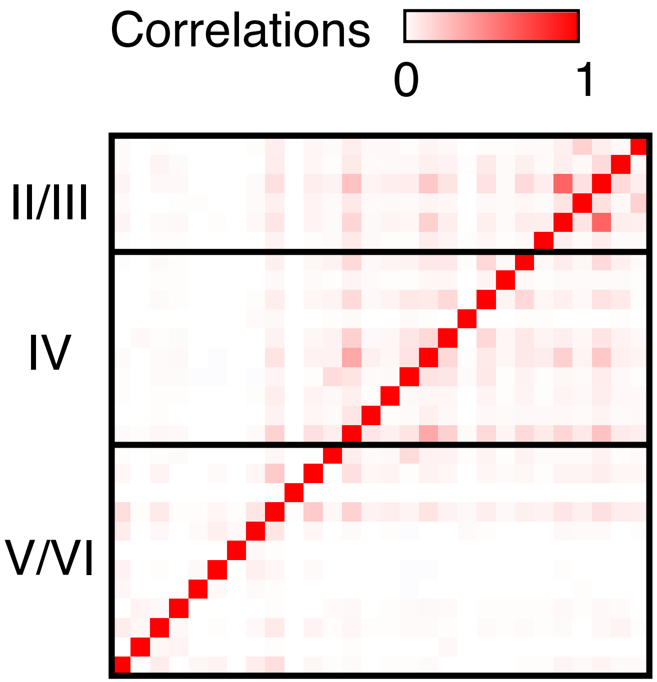
2s of raw data: 28 units discretized in 20ms time bins



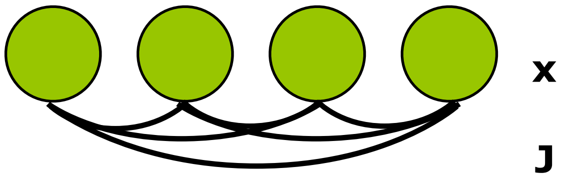
Layer Pairwise correlations

From Correlations to Models

- Ising model: Models pairwise correlations with pairwise coupling
- Limitation: Cannot capture higher order structure



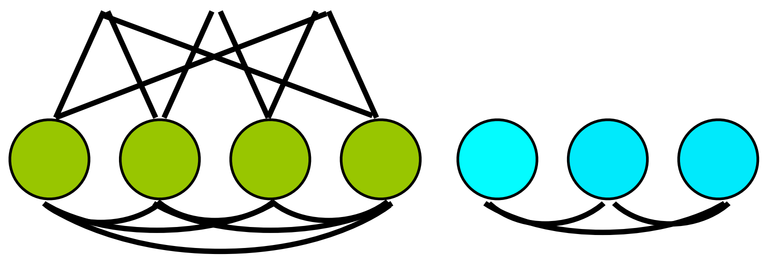
Matrix of correlations between pairs of cells



Ising Model:
Pairwise couplings **J** explain correlations between states **x**

Proposed Model

- Ising model, cells connect with pairwise coupling
- Additional hidden units
- Boltzmann Machine
- No connections between hidden units: Restricted Boltzmann Machine (RBM)
- Estimation of parameters is made efficient with Minimum Probability Flow (MPF)



MPF objective

$$K = \sum_{\mathbf{x} \in \mathcal{D}} \sum_{\mathbf{x}' \notin \mathcal{D}} g(\mathbf{x}, \mathbf{x}') \, \exp\left(\frac{1}{2}[E(\mathbf{x}) - E(\mathbf{x}')]\right)$$

$$g(\mathbf{x}, \mathbf{x}') = \begin{cases} 1 & \mathbf{x}' = 1 - \mathbf{x} \\ 1 & H(\mathbf{x}, \mathbf{x}') = 1 \\ 0 & \text{otherwise} \end{cases}$$

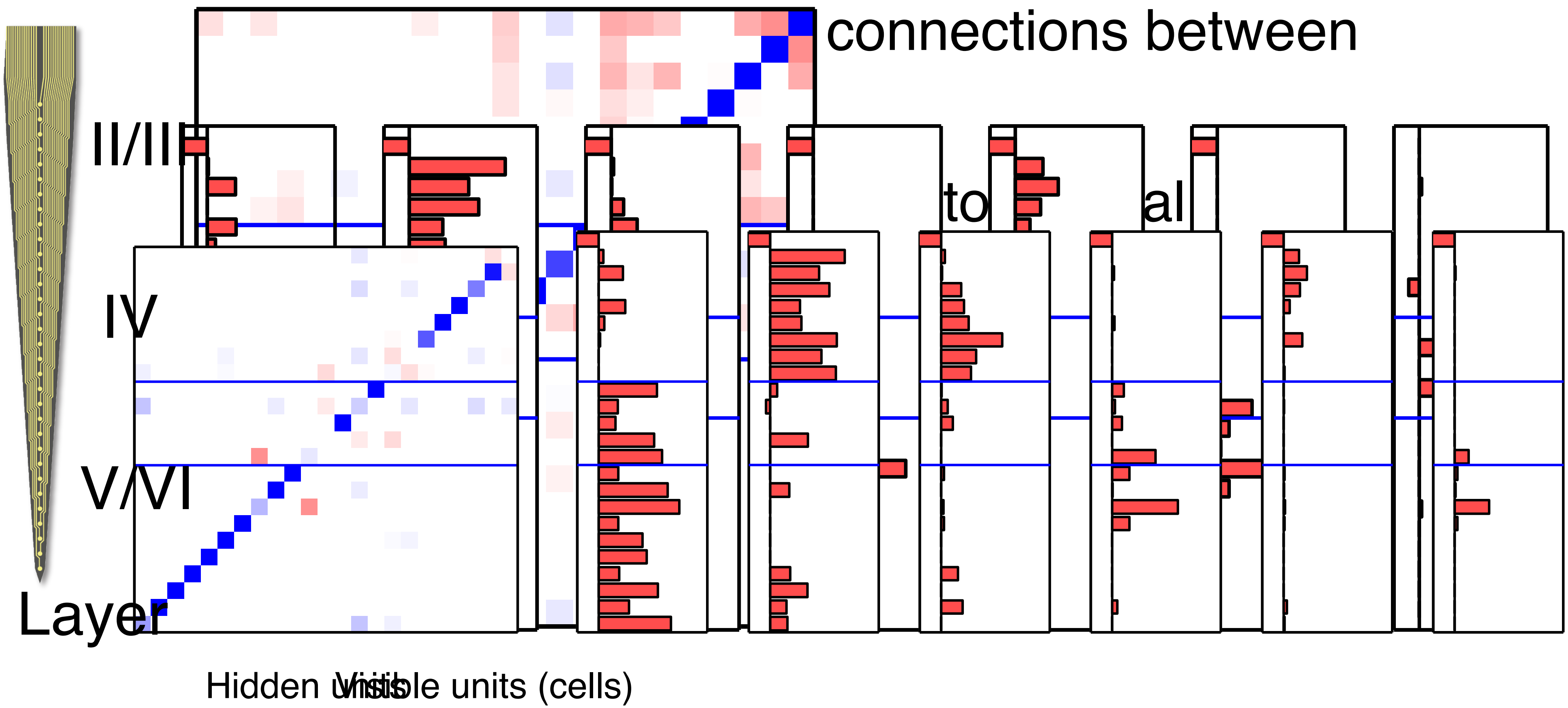
Ising: $$E_I(\mathbf{x}) = -\mathbf{x}^T \mathbf{J} \, \mathbf{x} - \mathbf{b}^T \mathbf{x}$$

RBM: $$E_R(\mathbf{x}) = -\sum_i \log(1 + e^{\mathbf{w}_i^T \mathbf{x} + b_{h,i}}) - \mathbf{b}_v^T \mathbf{x}$$

sRBM: $$E_S(\mathbf{x}) = -\mathbf{x}^T \mathbf{J} \, \mathbf{x} - \sum_i \log(1 + e^{\mathbf{w}_i^T \mathbf{x} + b_{h,i}}) - \mathbf{b}_v^T \mathbf{x}$$

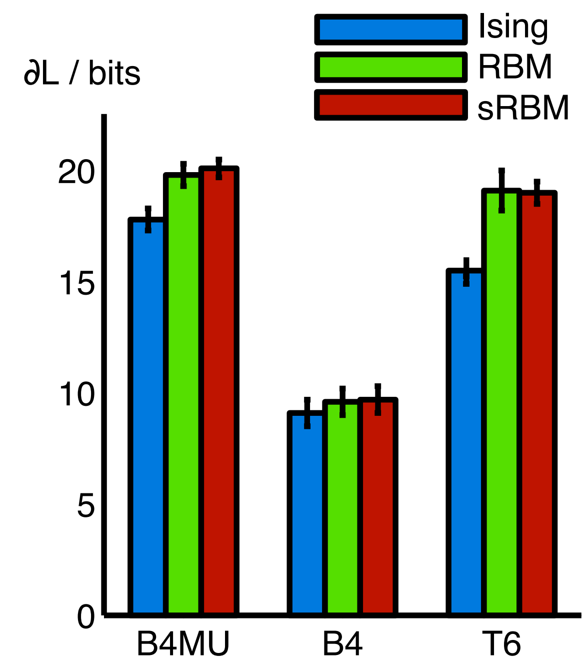
Results: Model structure

- Ising model: Pairwise coupling parameters
- RBM with vertical connections only

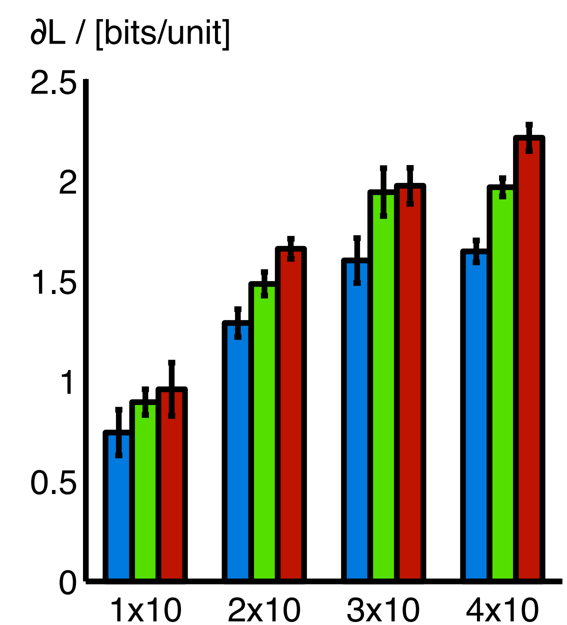


Model comparison

- Normalized probabilities with Annealed Importance Sampling for model comparison
- Model quality measured as likelihood gain over independent model
- Boltzmann machines with hidden units significantly outperform Ising models
- Spatiotemporal data sets to compare model dimensionality



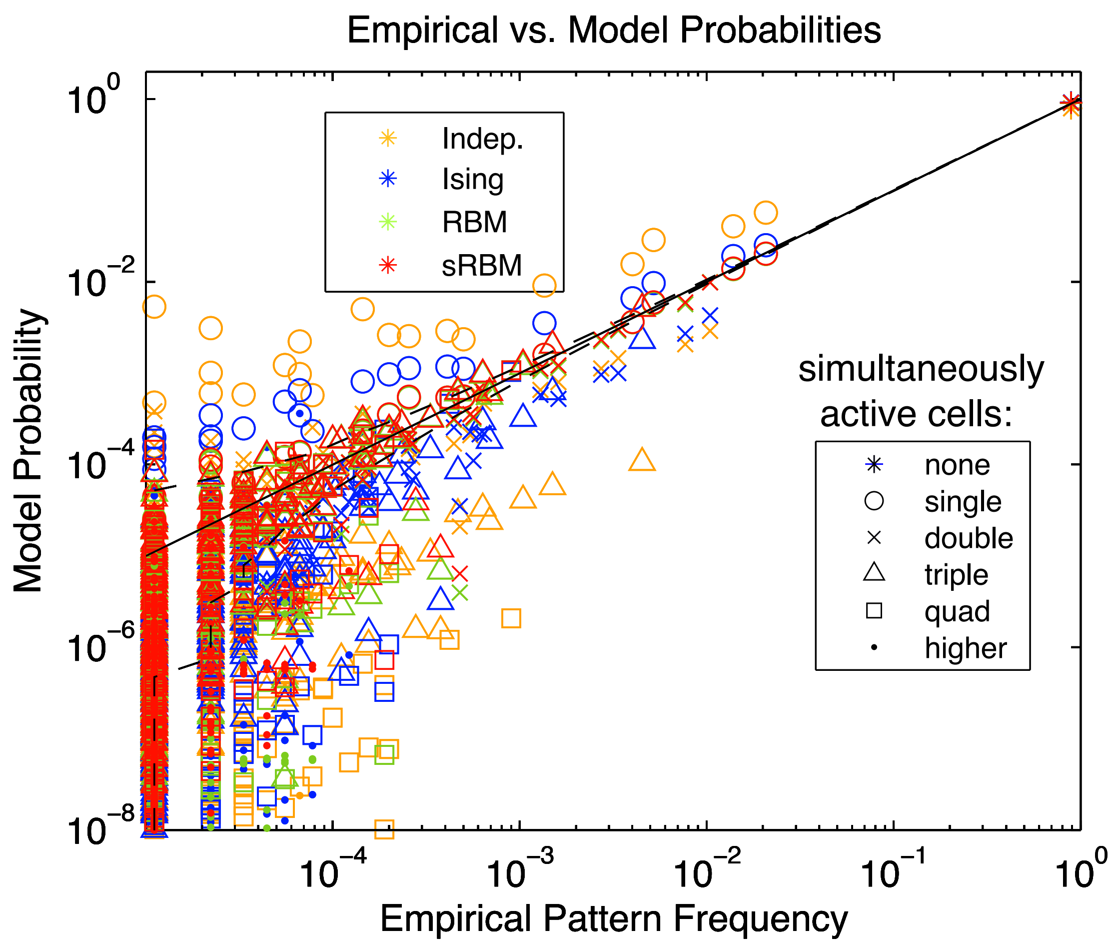
log-likelihood gain



Spatiotemporal models

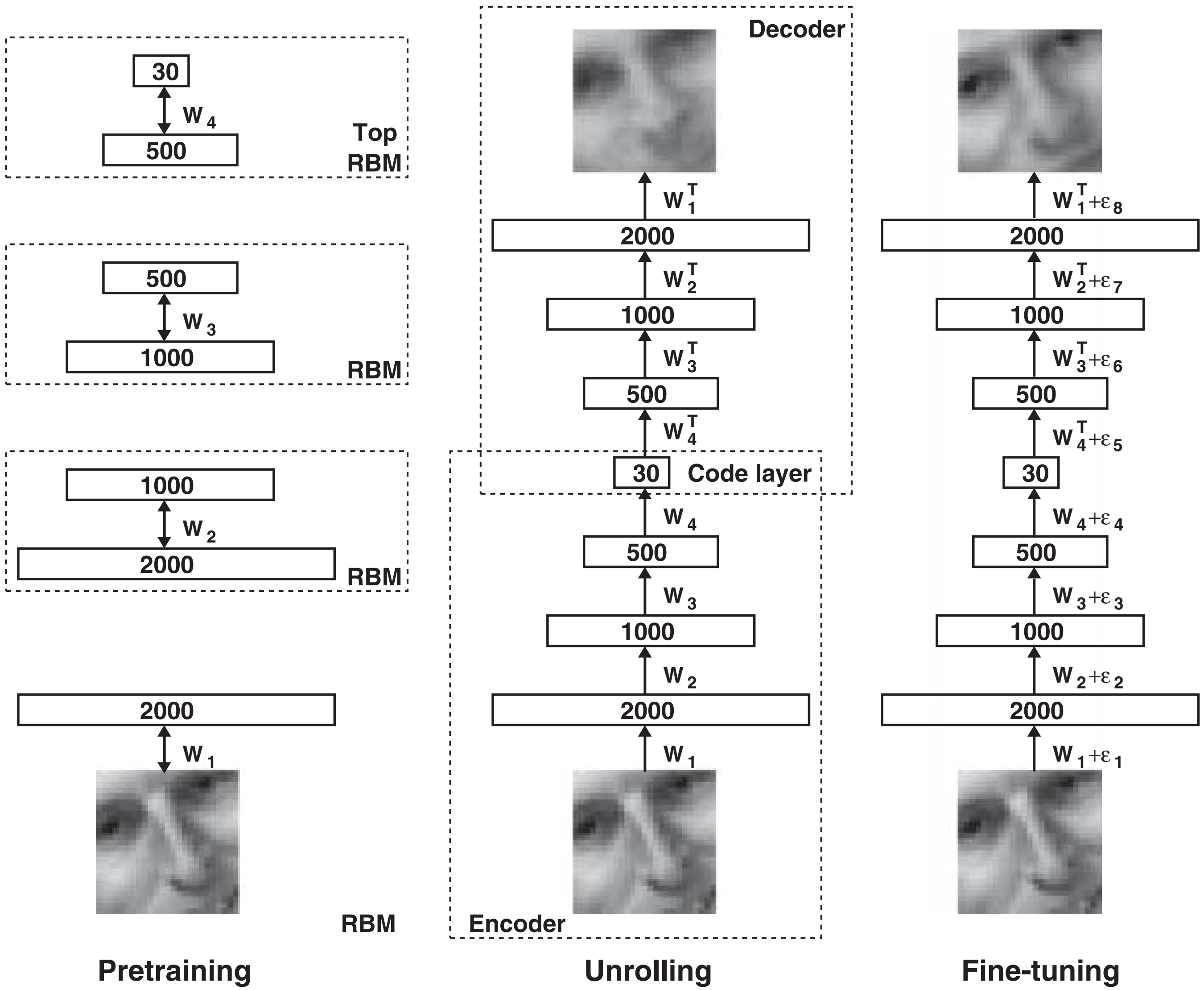
Results: Pattern frequency

- Insight into where and how models fail
- Independent model fails on pairs of cells
- Ising model underestimates triplet activity
- RBMs capture all patterns well



Reducing the Dimensionality of Data with Neural Networks

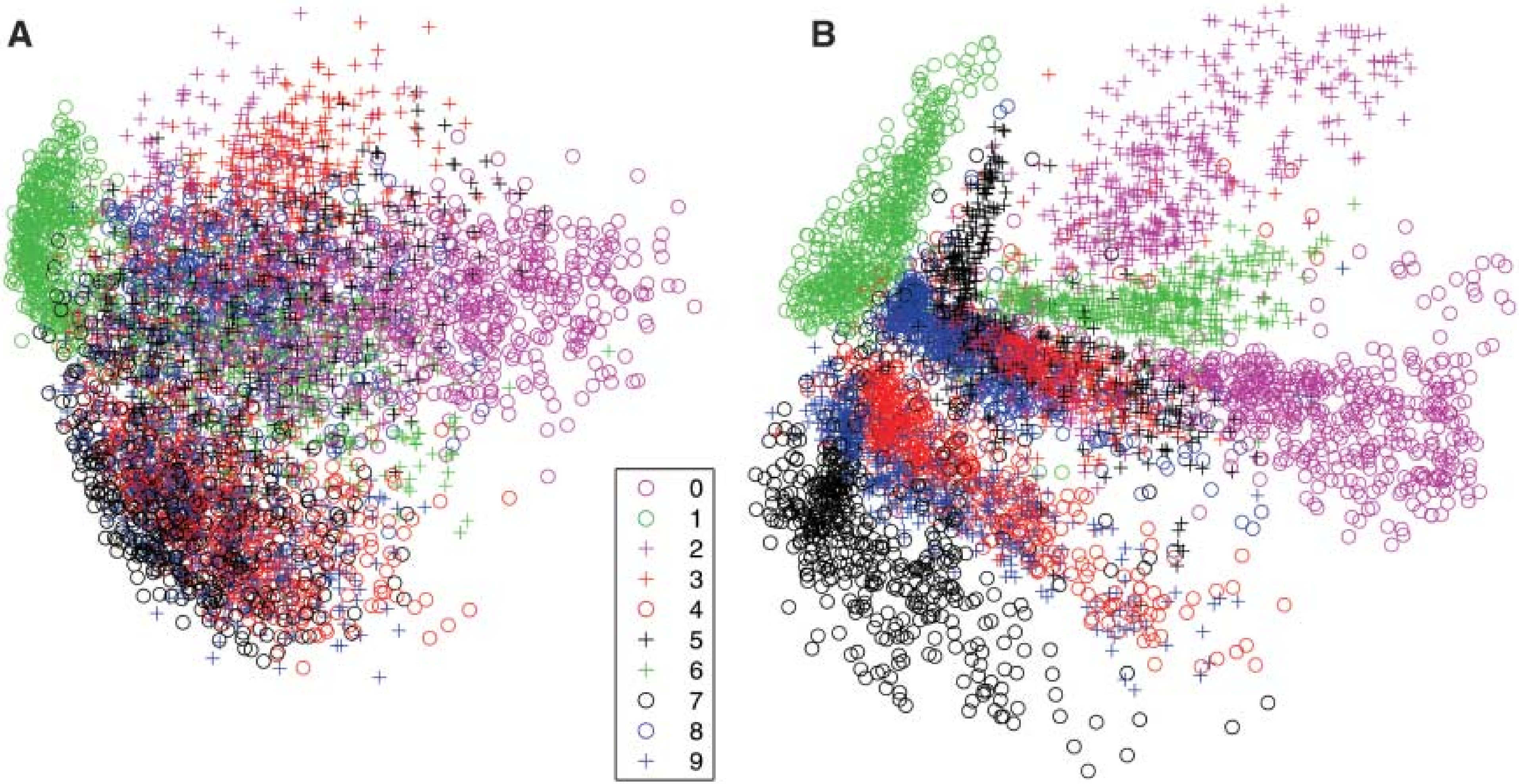
G. E. Hinton* and R. R. Salakhutdinov



$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i \in \text{pixels}} b_i v_i - \sum_{j \in \text{features}} b_j h_j - \sum_{i,j} v_i h_j w_{ij}$$

$$\Delta w_{ij} = \varepsilon\left(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{recon}}\right)$$

Application to hand-written digits



2D PCA

784-1000-500-250-2 autoencoder