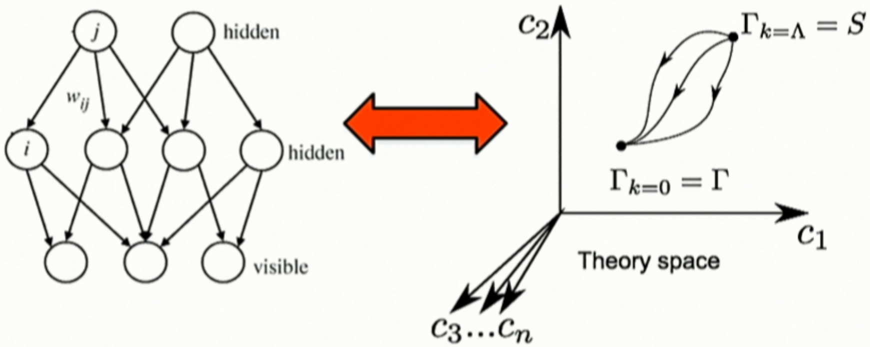
Title: Physical approaches to the extraction of relevant information

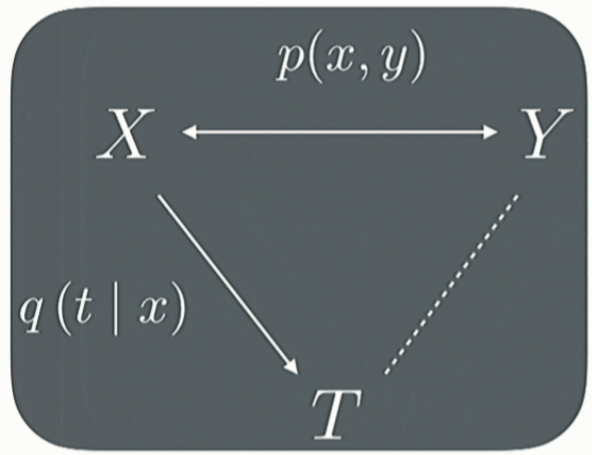
Date: Aug 09, 2016 11:00 AM

URL: http://pirsa.org/16080006

Abstract: In the first part of this talk, I will focus on the physics of deep learning, a popular subfield of machine learning where recent performance on tasks such as visual object recognition rivals human performance. I present work relating greedy training of deep belief networks to a form of variational real-space renormalization. This connection may help explain how deep networks automatically learn relevant features from data and extract independent factors of variation. Next, I turn to the information bottleneck (IB), an information theoretic approach to clustering and compression of relevant information that has been suggested as a framework for deep learning. I present a new variant of IB called the Deterministic Information Bottleneck, arguing that it better captures the notion of compression while retaining relevant information.

Physical approaches to the extraction of relevant information





David J. Schwab
Department of Physics and Astronomy
Northwestern University

What I get paid to do:

- From intracellular signaling to population oscillations: bridging scales in collective behavior

  Molecular Systems Biology (2015)

- Constant exponential growth through very different metabolic strategies

  Cell Reports (2014)

- Quantifying the role of population subdivision in evolution on rugged fitness landscapes

  PLoS Computational Biology (2014)

- Lag normalization in an electrically coupled neural network

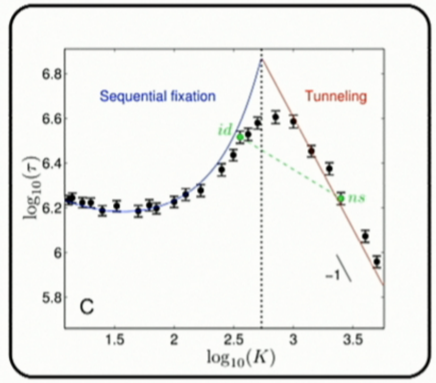
  Nature Neuroscience (2013)

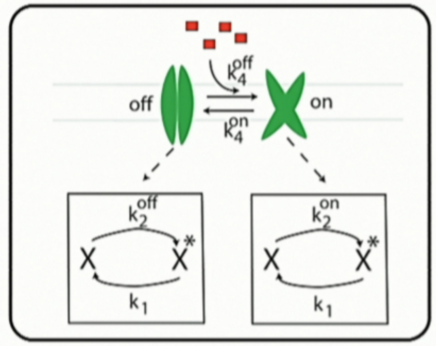
- The energetic costs of cellular computation

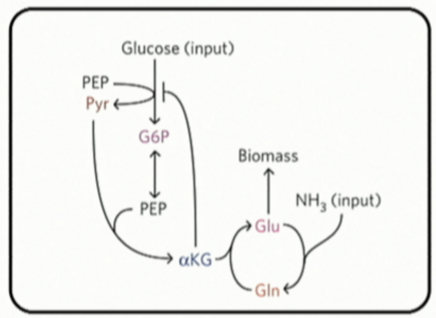
  PNAS (2012)

- **Coordination of carbon and nitrogen metabolism in *e. coli***

  Nature Chem. Bio. (2011)







Outline of the talk:

- Deep learning and the renormalization group

- The deterministic information bottleneck

Pankaj Mehta
Boston University

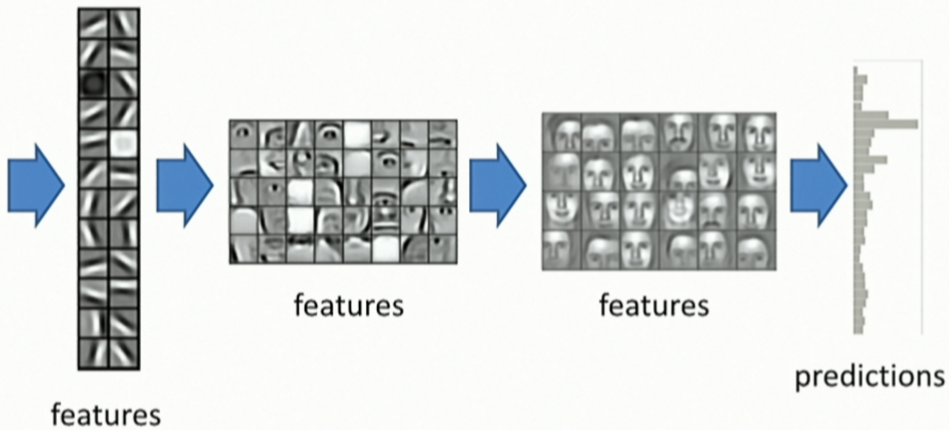
What is "deep learning"?

- Learning multiple levels of representation/abstraction

- Has revolutionized object recognition, speech recognition, many other emerging applications e.g. translation, natural language processing, reinforcement learning

- Many industrial applications - Google, Facebook, Baidu, Microsoft, etc.

- Feature learning with the prior that there are a hierarchy of underlying factors



Motivation

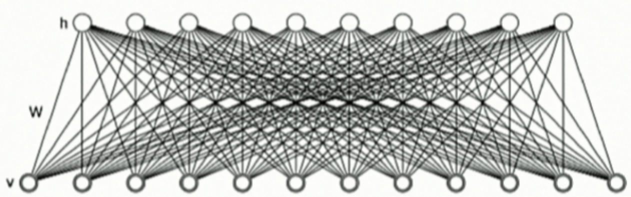
Understanding the success of unsupervised, pre-training in Deep Belief Networks (DBNs) for dimensional reduction – (iterative coarse graining)

Reducing the Dimensionality of Data with Neural Networks

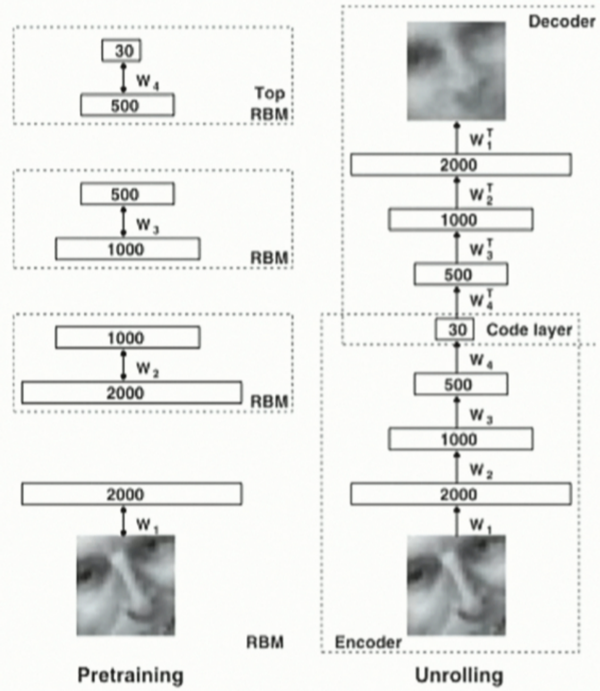
G. E. Hinton* and R. R. Salakhutdinov

28 JULY 2006 VOL 313 SCIENCE www.sciencemag.org

RBM



$$P(v,h;\theta) = \frac{1}{Z(\theta)} \exp(-E(v,h;\theta)),$$
$$Z(\theta) = \sum_{v}\sum_{h} \exp(-E(v,h;\theta)).$$



Belief nets:

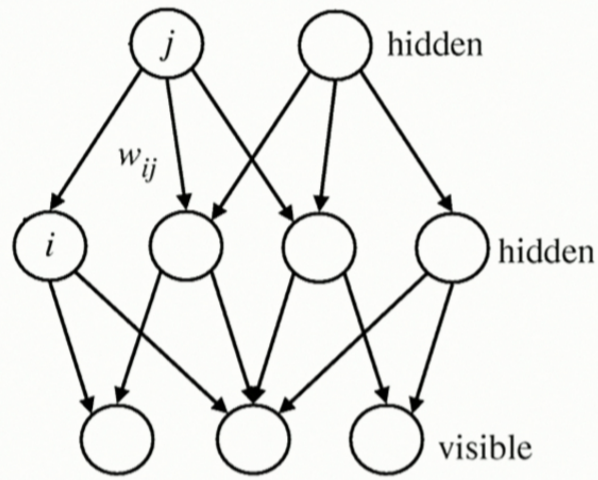
• Directed, acyclic graph of stochastic variables

• Data are "visible" variables

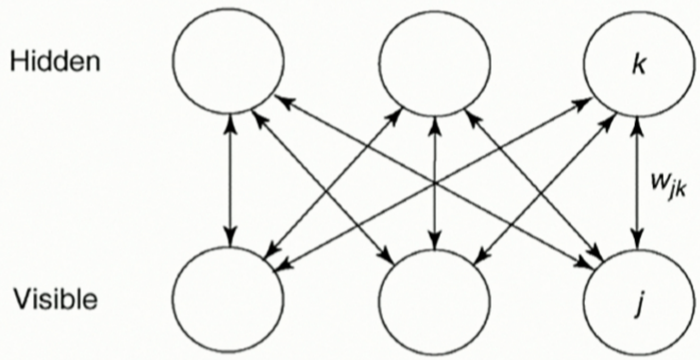
• Would like to perform:

- *Inference*: infer hiddens given data

- *Learning*: adjust interactions to make network more likely to generate observed data



Hinton's roundabout breakthrough: Restricted Boltzmann Machines



Can be learned efficiently, e.g. contrastive divergence

$$-E(\mathbf{v},\mathbf{h}) = \sum_i c_i v_i + \sum_j b_j h_j + \sum_{i,j} w_{ij} v_i h_j$$

Energy of joint configuration, bipartite graph

Interactions of all orders:

$$p(\mathbf{v}) = \int d\mathbf{h} p(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\exp[-E_R(\mathbf{v})]$$

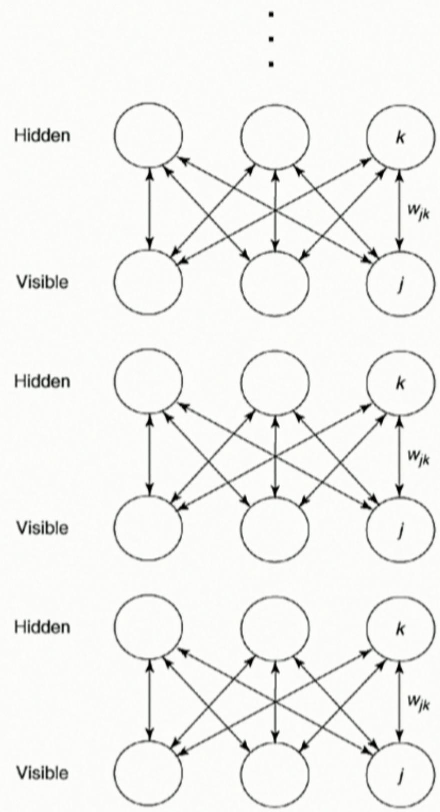
$$E_R(\mathbf{v}) = -\sum_j \log(1 + e^{\sum_i w_{ij} v_i + b_j}) - \sum_j c_i v_i$$

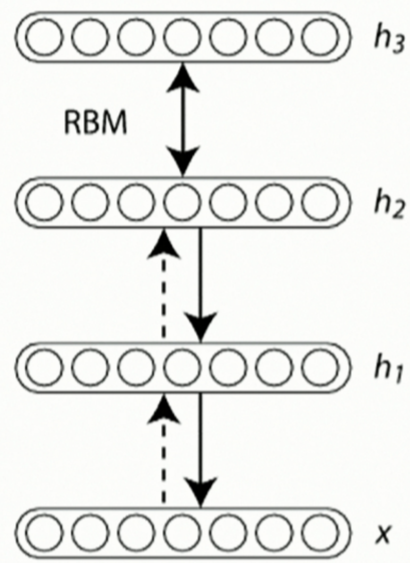
Greedy layer-by-layer training of RBMs form a Deep Belief Network

Fit lowest layer RBM

Use hidden activities as data for next RBM

Repeat as necessary...





Kadanoff's Variational RG

- Couple coarse grained and microscopic degrees of freedom and integrate out (marginalize) microscopic variables

- Introduce new operator that defines coupling:

$$e^{-\mathbf{H}_{\lambda}^{RG}[\{h_j\}]} \equiv \mathrm{Tr}_{v_i} e^{\mathbf{T}_{\lambda}(\{v_i\},\{h_j\})-\mathbf{H}(\{v_i\})}$$

- Free Energy is invariant under transform if:

$$\mathrm{Tr}_{h_j} e^{\mathbf{T}_{\lambda}(\{v_i\},\{h_j\})} = 1$$

(Simultaneously want to choose T to make the trace over v tractable.)

- Variational parameters chosen to minimize free energy difference (or bound it in some way)

$$\Delta F = F_{\theta}[\mathbf{h}] - F[\mathbf{v}]$$

Mapping between DBNs and Variational RG

- Can map two schemes to each other through following relation:

$$T_\theta(\mathbf{v}, \mathbf{h}) = -E_\theta(\mathbf{v}, \mathbf{h}) + H(\mathbf{v})$$

- Can show under this identification that preserving Free Energy is same as exactly modeling true distribution with variational distribution

$$Tr_{\mathbf{h}} e^{T_\theta(\mathbf{v}, \mathbf{h})} = 1 \leftrightarrow D_{KL}(P(\mathbf{v}) || P_\theta(\mathbf{v})) = 0$$

- RG Hamiltonian is exactly the "Hamiltonian" describing the hidden, coarse-grained degrees of freedom

$$H_\theta^{RG}(\mathbf{h}) = H_\theta^{RBM}(\mathbf{h}) \equiv -\log Tr_{\mathbf{v}} P(\mathbf{v}, \mathbf{h}; \theta) - \log \mathcal{Z}(\theta)$$

Mapping between DBNs and Variational RG

- Can map two schemes to each other through following relation:

$$T_\theta(\mathbf{v},\mathbf{h}) = -E_\theta(\mathbf{v},\mathbf{h}) + H(\mathbf{v})$$

- Can show under this identification that preserving Free Energy is same as exactly modeling true distribution with variational distribution

$$Tr_{\mathbf{h}} e^{T_\theta(\mathbf{v},\mathbf{h})} = 1 \leftrightarrow D_{KL}(P(\mathbf{v})||P_\theta(\mathbf{v})) = 0$$

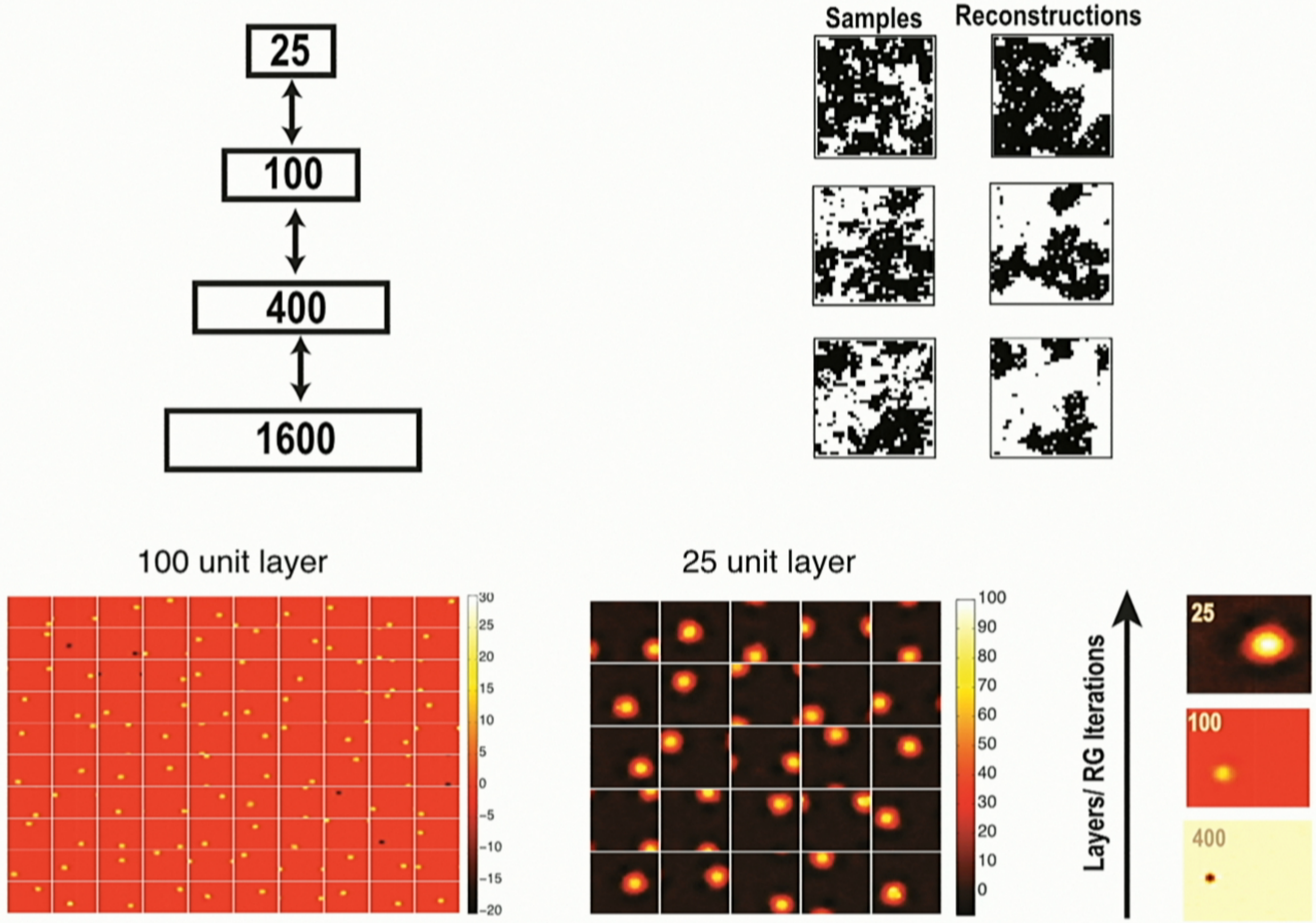
- RG Hamiltonian is exactly the “Hamiltonian” describing the hidden, coarse-grained degrees of freedom

$$H_\theta^{RG}(\mathbf{h}) = H_\theta^{RBM}(\mathbf{h}) \equiv -\log Tr_{\mathbf{v}} P(\mathbf{v},\mathbf{h};\theta) - \log \mathcal{Z}(\theta)$$

Comparing DBNs and Variational RG

Property	Variational RG	Deep Belief Networks
How input distribution is defined	Hamiltonian defining P(v)	Data samples drawn from P(v)
How interactions are defined	T(v,h)	E(v,h)
Exact transformation	$Tr_{\mathbf{h}} e^{T(\mathbf{v},\mathbf{h})} = 1$	KL divergence between P(v) and variational distribution is zero
Approximations	Minimize or bound free energy differences	Minimize the KL divergence
Method	Analytic (mostly)	Numerical
What happens under coarse-graining	Relevant operators grow/irrelevant shrink	New features emerge

Deep belief networks implement a form of RG



Ongoing work:

- Flow away from criticality when ascending layers
- Do deep networks work best for near-critical data?

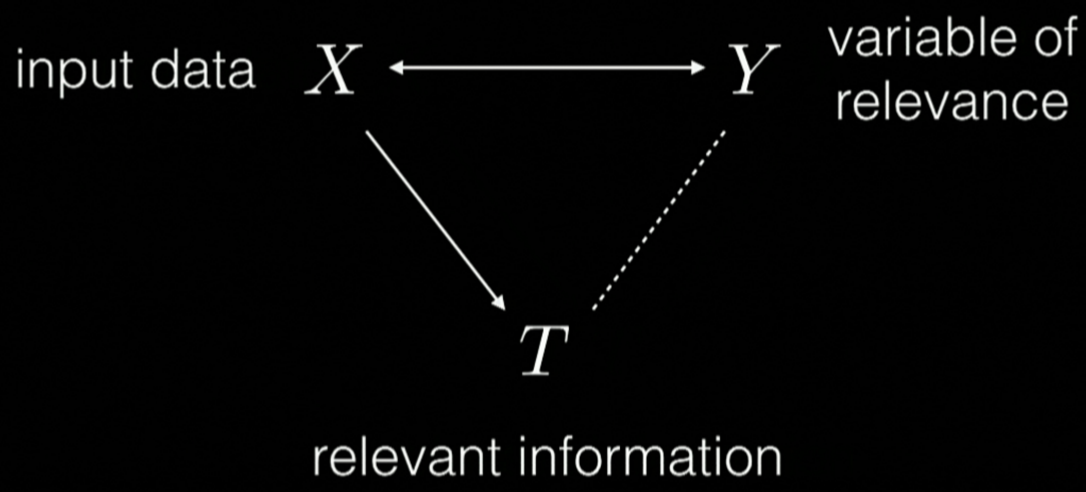
Outline of the talk:

- Deep learning and the renormalization group

- The deterministic information bottleneck

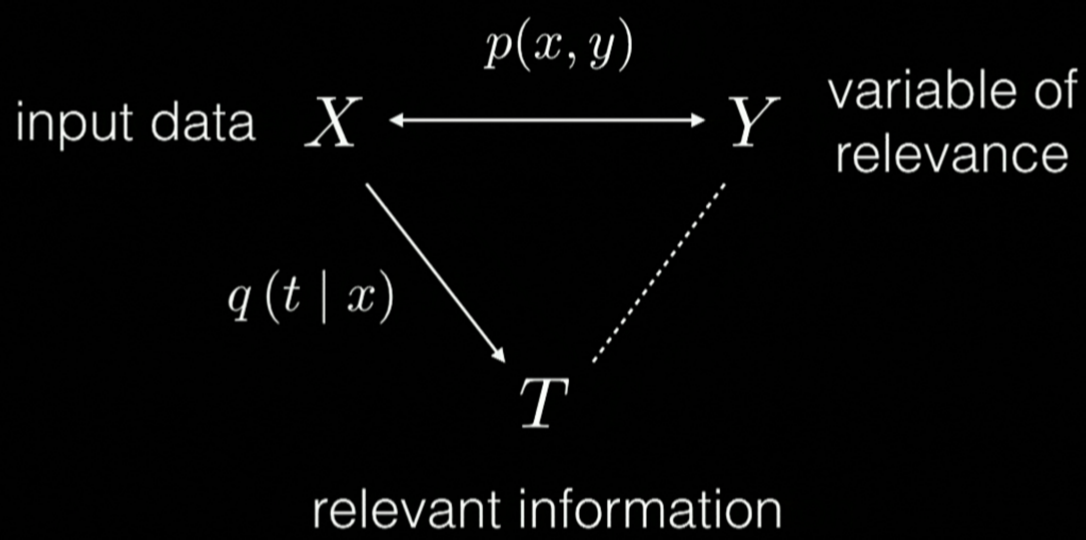
DJ Strouse
Princeton University

Information bottleneck (IB)



Tishby, Pereira, Bialek 1999

Information bottleneck (IB)



statistics: soft sufficient statistic
info theory: lossy compression, distortion ~ relevance
machine learning: maximally informative clustering

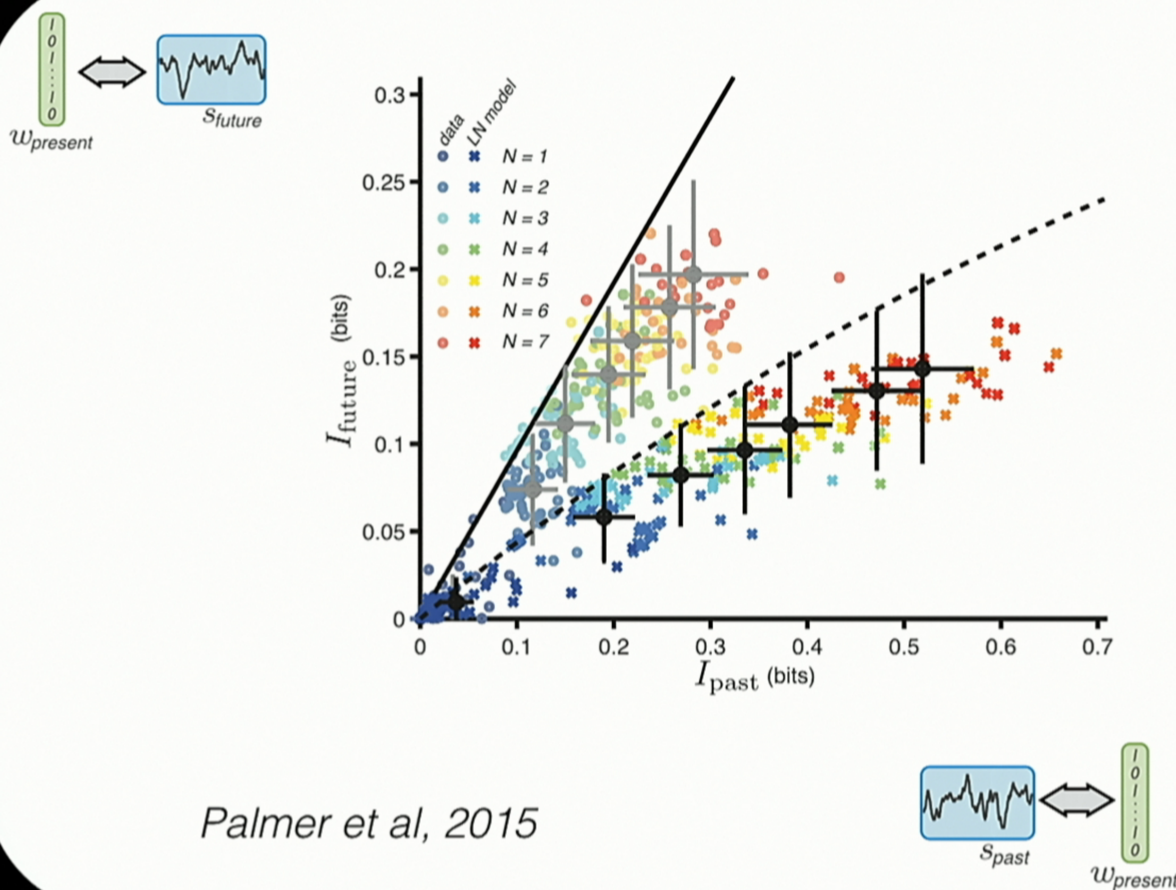
IB examples

	X	T	Y
user segmentation	demographics & past behavior	cluster ID	future purchase/ click behavior

IB examples

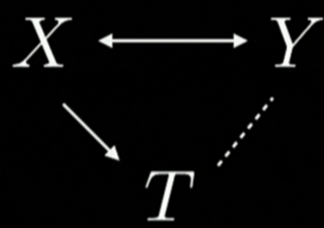
	X	T	Y
user segmentation	demographics & past behavior	cluster ID	future purchase/ click behavior
human attention & memory	sensory input	neural activity/ synaptic changes	future sensory input?

IB examples



Palmer et al, 2015

Information bottleneck (IB)

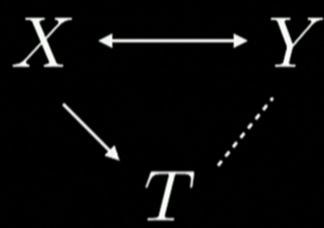


data: $p(x, y)$

$$\min_{q(t|x)} L\left[q\left(t \mid x\right)\right] = I\left(T; X\right) - \beta I\left(T; Y\right)$$

Tishby, Pereira, Bialek 1999

Information bottleneck (IB)



data: $p(x, y)$

free parameter: $\beta > 0$

Markov constraint: $T \leftarrow X \longleftrightarrow Y$

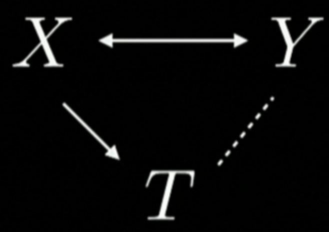
$$\min_{q(t|x)} L\left[q\left(t \mid x\right)\right] = I\left(T; X\right) - \beta I\left(T; Y\right)$$

$$q\left(t \mid x\right) = \frac{q\left(t\right)}{Z\left(x, \beta\right)} \exp\left[-\beta D_{KL}\left[p\left(y \mid x\right) \mid q\left(y \mid t\right)\right]\right]$$

$$q\left(t\right) = \sum_{x} p\left(x\right) q\left(t \mid x\right)$$

$$q\left(y \mid t\right) = \frac{1}{q\left(t\right)} \sum_{x} p\left(y \mid x\right) q\left(t \mid x\right) p\left(x\right)$$

Information bottleneck (IB)



data: $p(x, y)$

free parameter: $\beta > 0$

Markov constraint: $T \leftarrow X \longleftrightarrow Y$

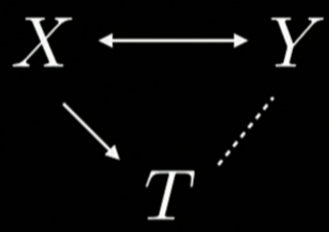
$$\min_{q(t|x)} L\left[q\left(t \mid x\right)\right] = I\left(T; X\right) - \beta I\left(T; Y\right)$$

$$q\left(t \mid x\right) = \frac{q\left(t\right)}{Z\left(x, \beta\right)} \exp\left[-\beta D_{KL}\left[p\left(y \mid x\right) \mid q\left(y \mid t\right)\right]\right]$$

$$q\left(t\right) = \sum_{x} p\left(x\right) q\left(t \mid x\right)$$

$$q\left(y \mid t\right) = \frac{1}{q\left(t\right)} \sum_{x} p\left(y \mid x\right) q\left(t \mid x\right) p\left(x\right)$$

Information bottleneck (IB)

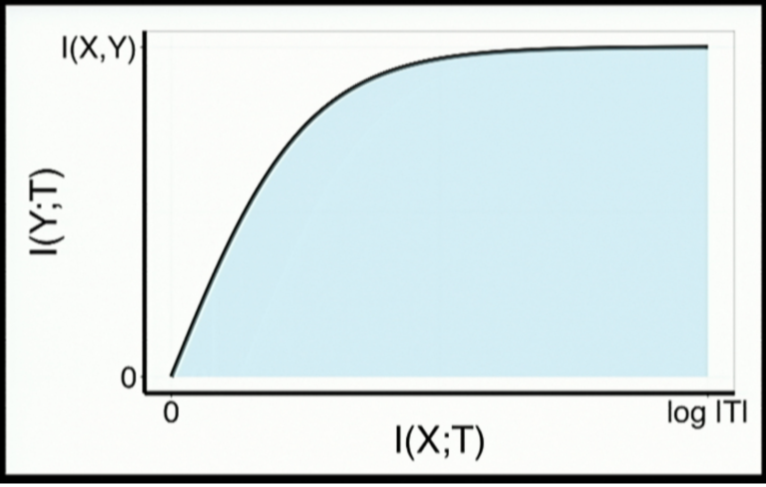


data: $p(x, y)$

free parameter: $\beta > 0$

Markov constraint: $T \leftarrow X \longleftrightarrow Y$

$$\min_{q(t|x)} L\left[q\left(t \mid x\right)\right] = I\left(T; X\right) - \beta I\left(T; Y\right)$$



Measuring compression

$$\min_{q(t|x)} L\left[q\left(t \mid x\right)\right] = I\left(T;X\right) - \beta I\left(T;Y\right)$$

channel coding/
rate distortion theory

$$\min_{q(t|x)} L\left[q\left(t \mid x\right)\right] = H\left(T\right) - \beta I\left(T;Y\right)$$

source coding

Measuring compression

$$\min_{q(t|x)} L\left[q\left(t \mid x\right)\right] = I\left(T;X\right) - \beta I\left(T;Y\right)$$

channel coding/
rate distortion theory

$$\min_{q(t|x)} L\left[q\left(t \mid x\right)\right] = H\left(T\right) - \beta I\left(T;Y\right)$$

source coding

$$\begin{aligned} L_{\mathrm{IB}} - L_{\mathrm{DIB}} &= I(X;T) - H(T) \\ &= -H(T \mid X) \end{aligned}$$

implicit encouragement of stochasticity

A generalized IB

$$L_\alpha \equiv H(T) - \alpha H(T \mid X) - \beta I(Y;T)$$

A generalized IB

$$L_\alpha \equiv H(T) - \alpha H(T \mid X) - \beta I(Y;T)$$

$$q_\alpha(t \mid x) \propto \exp\left[\frac{1}{\alpha}\left(\log q_\alpha(t) - \beta D_{\mathrm{KL}}[p(y \mid x) \mid q_\alpha(y \mid t)]\right)\right]$$

Solving the DIB

$$L_\alpha \equiv H(T) - \alpha H(T \mid X) - \beta I(Y;T)$$

$$q_\alpha(t \mid x) \propto \exp\left[\frac{1}{\alpha}\left(\log q_\alpha(t) - \beta D_{\mathrm{KL}}[p(y \mid x) \mid q_\alpha(y \mid t)]\right)\right]$$

$$\lim_{\alpha \to 0} q_\alpha(t \mid x) = \delta(t - f(x))$$

$$f(x) = \underset{t}{\operatorname{argmax}}(\log q(t) - \beta D_{\mathrm{KL}}[p(y \mid x) \mid q(y \mid t)])$$

IB vs DIB: summary

$$L_{\text{IB}} = I(X;T) - \beta I(Y;T)$$

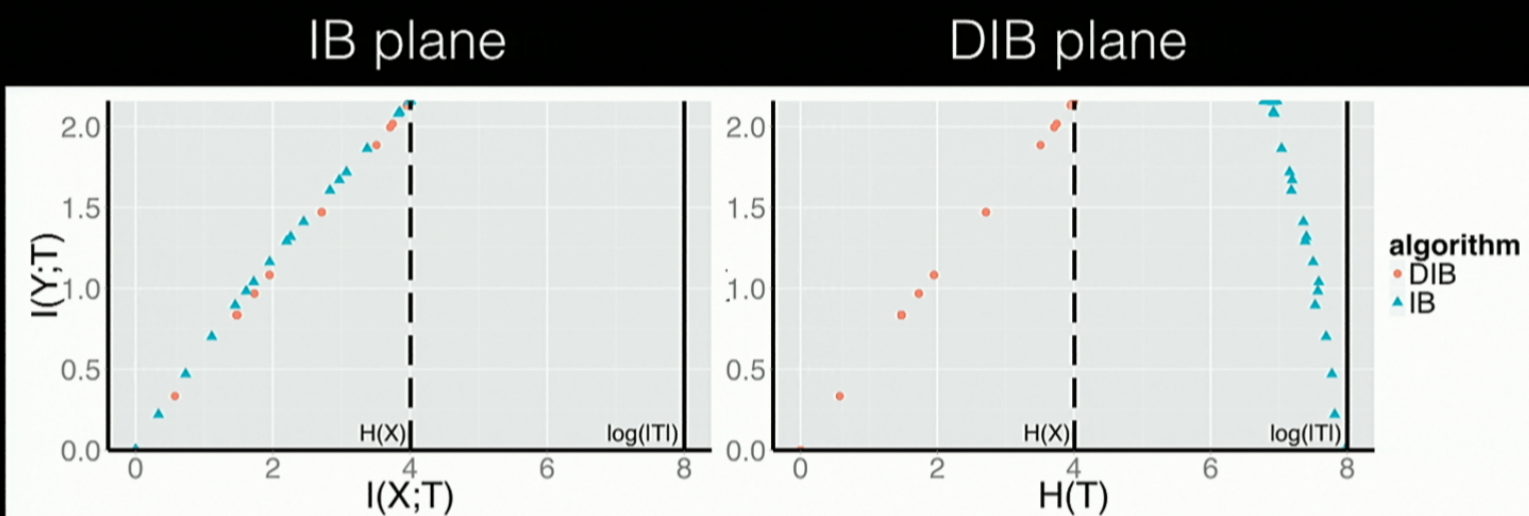
$$q_{\text{IB}}(t \mid x) = \frac{q(t)}{Z(x,\beta)} \exp[-\beta D_{\text{KL}}[p(y \mid x) \mid q(y \mid t)]]$$

$$L_{\text{DIB}} = H(T) - \beta I(Y;T)$$

$$q_{\text{DIB}}(t \mid x) = \delta(t - f(x))$$

$$f(x) = \underset{t}{\operatorname{argmax}}(\log q(t) - \beta D_{\text{KL}}[p(y \mid x) \mid q(y \mid t)])$$

IB vs DIB: experiments



Summary

- proposed new cost functional for extraction of relevant information based on source coding (rather than channel coding)

Summary

- proposed new cost functional for extraction of relevant information based on source coding (rather than channel coding)
- consequence -> deterministic encoder/hard clustering (rather than stochastic/soft)
- IB and DIB exhibit non-trivial differences when fit to data
- DIB fits run 1-2 orders of magnitude faster than IB
- bonus: method to interpolate between IB and DIB