# The deterministic information bottleneck

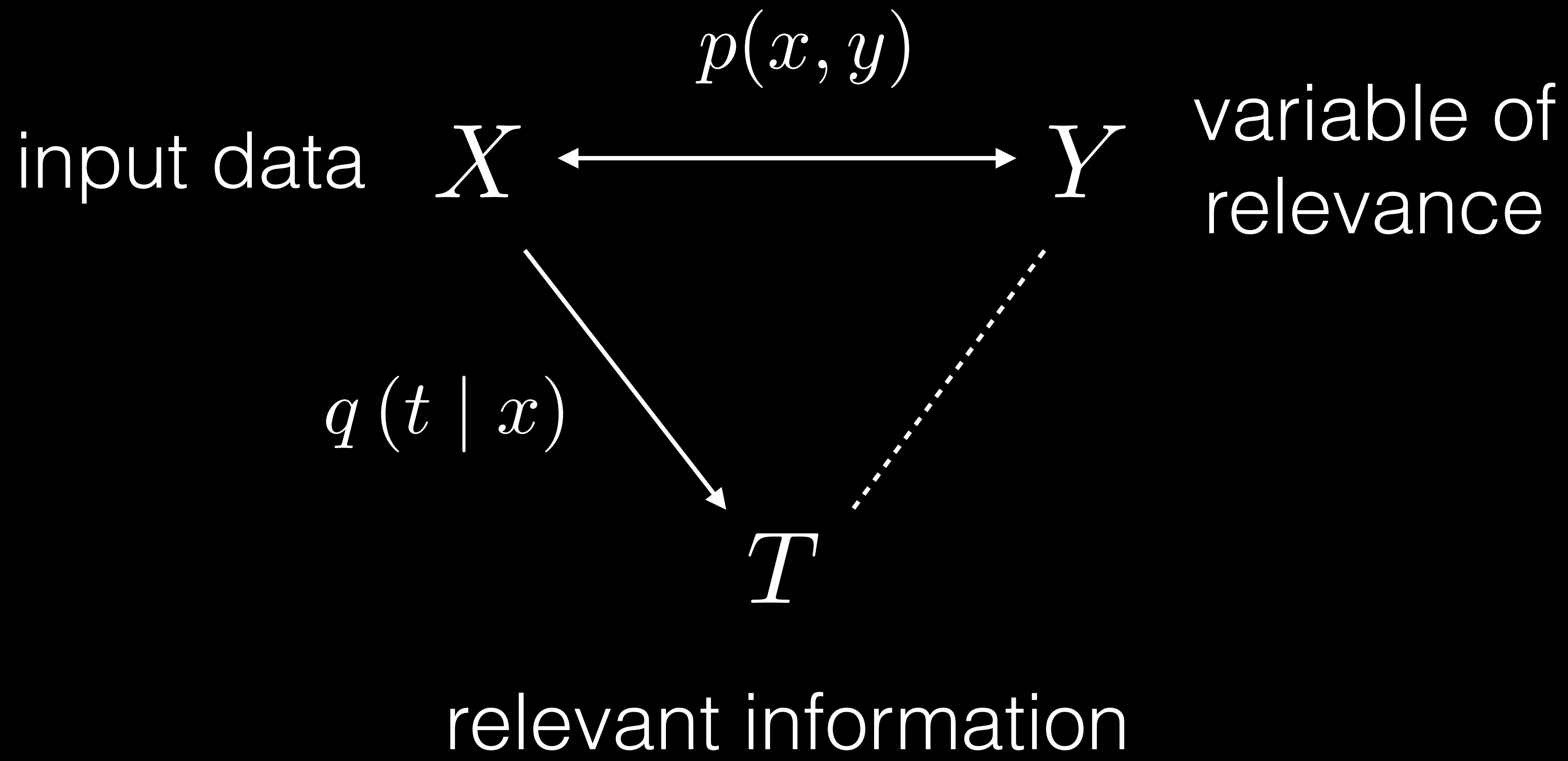

DJ Strouse
Princeton University

David Schwab
Northwestern University



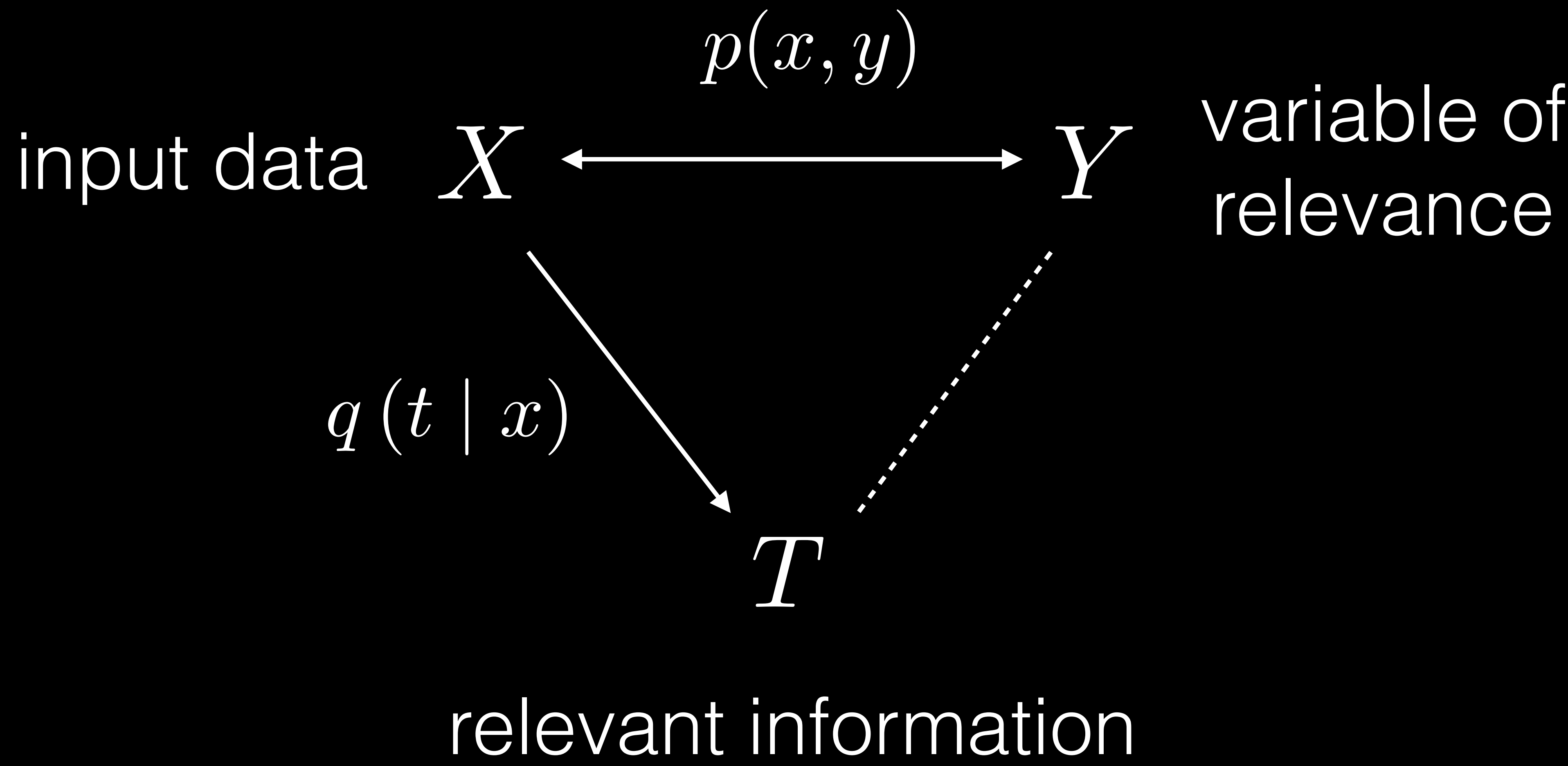
DOE CSGF Program Review 2016
Washington, DC

# 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

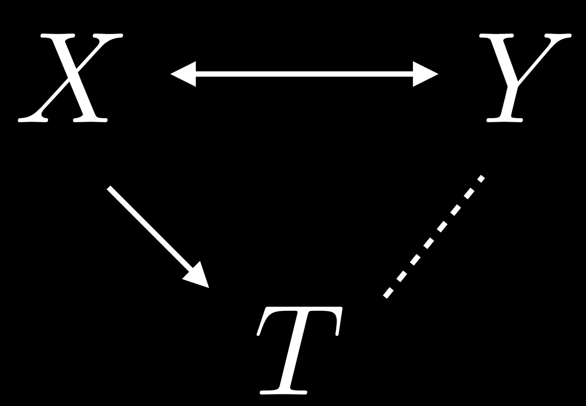
| | $X$ | $T$ | $Y$ |
|---|---|---|---|
| **user segmentation** | demographics & past behavior | cluster ID | future purchase/ click behavior |

# 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? |

*Palmer et al, PNAS, 2015*

# Information bottleneck (IB)




data: $p(x, y)$

free parameter: $\beta > 0$

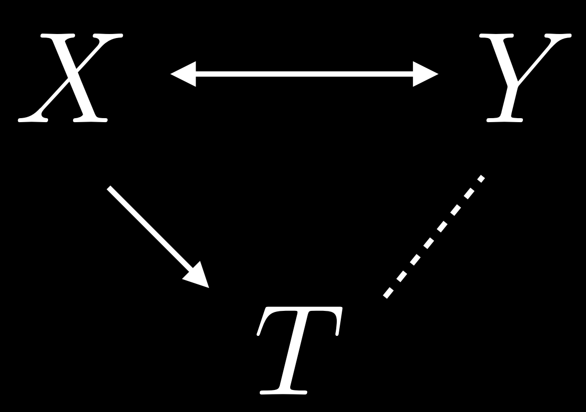
Markov constraint: $T \leftarrow X \longleftrightarrow Y$

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

compression relevance

*Tishby, Pereira, Bialek 1999*

# Information bottleneck (IB)




data: $p(x, y)$

free parameter: $\beta > 0$

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

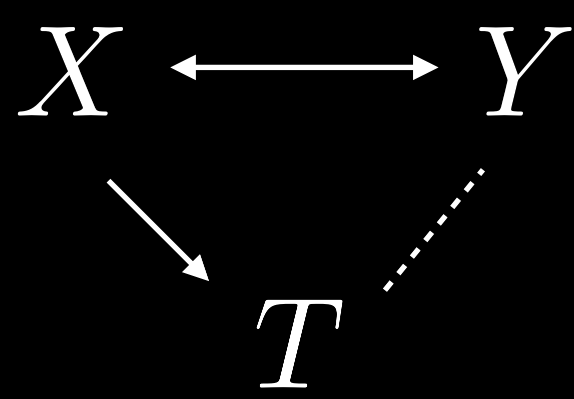
$$\min_{q(t|x)} L\left[q\left(t \mid x\right)\right] = I\left(T;X\right) - \beta I\left(T;Y\right)$$

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

$$q\left(t\right) = \sum_x p\left(x\right) q\left(t \mid x\right)$$

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

# Information bottleneck (IB)

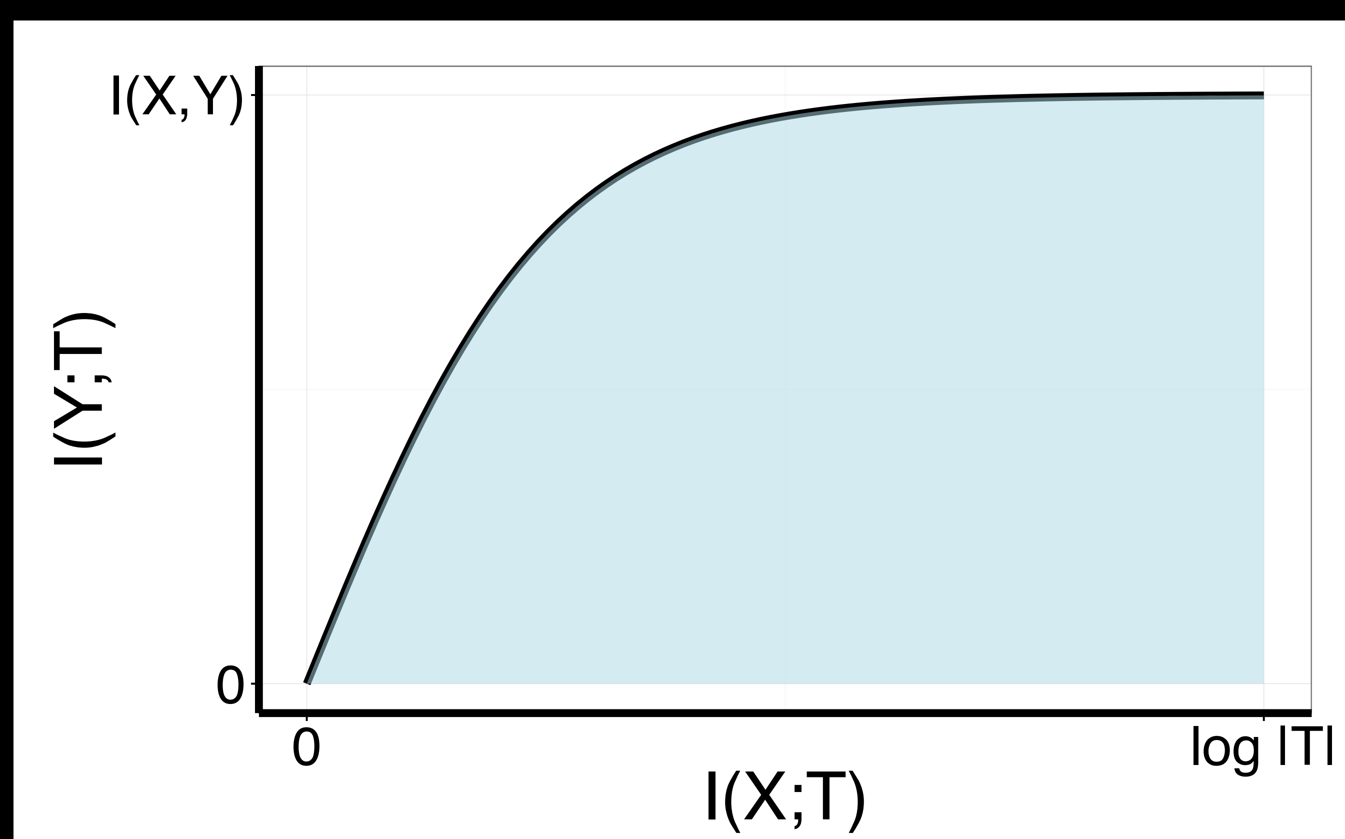



data: $p(x, y)$

free parameter: $\beta > 0$

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

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



# 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

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

$$L_{\text{IB}} = L_{\alpha=1}$$

$$L_{\text{DIB}} = L_{\alpha=0}?$$

# A generalized IB

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

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

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

# Solving the DIB

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

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

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

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

# IB vs DIB: summary

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

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

*channel coding with relevance*

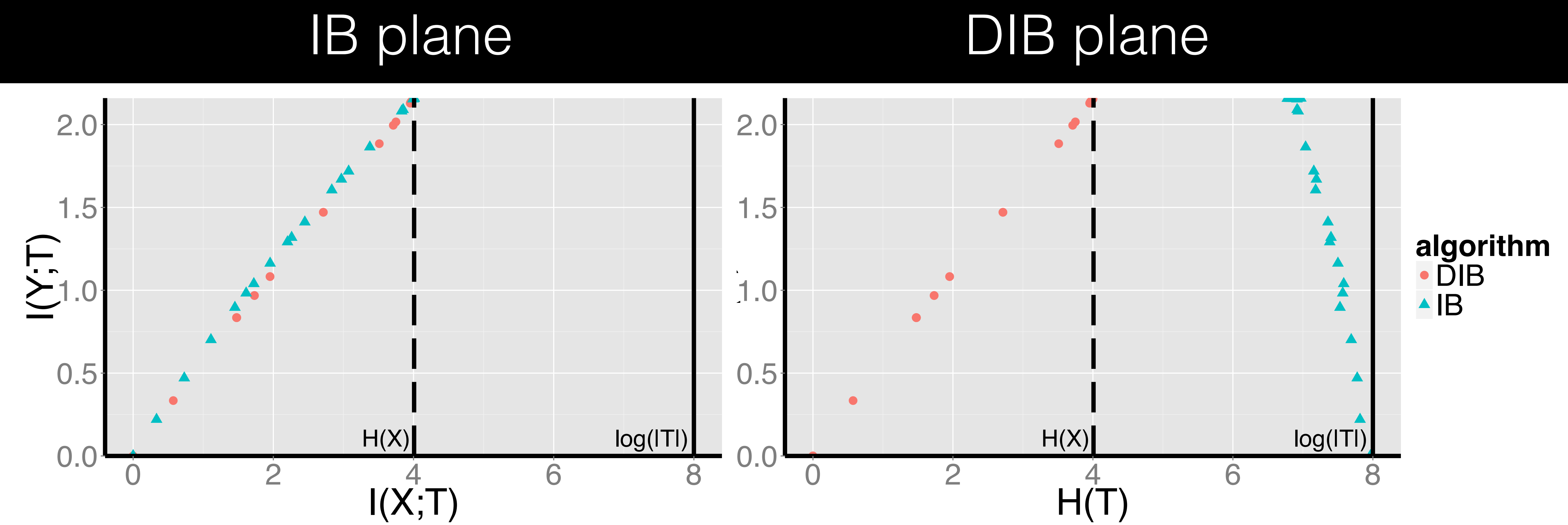
$$L_{\mathrm{DIB}} = H(T) - \beta I(Y;T)$$

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

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

*source coding with relevance*

# IB vs DIB: experiments



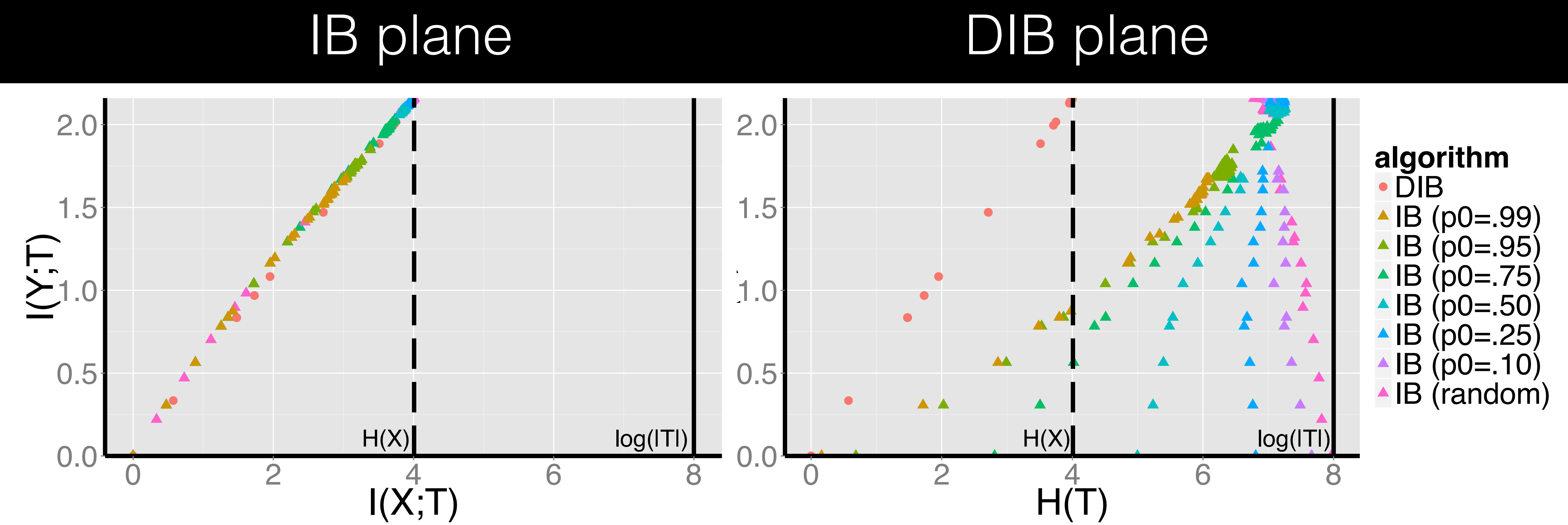

$$I(X,T) = H(T) - H(T \mid X)$$

2 ways to get $I(X,T) = 0$

$$H(T) = H(T \mid X) = 0$$

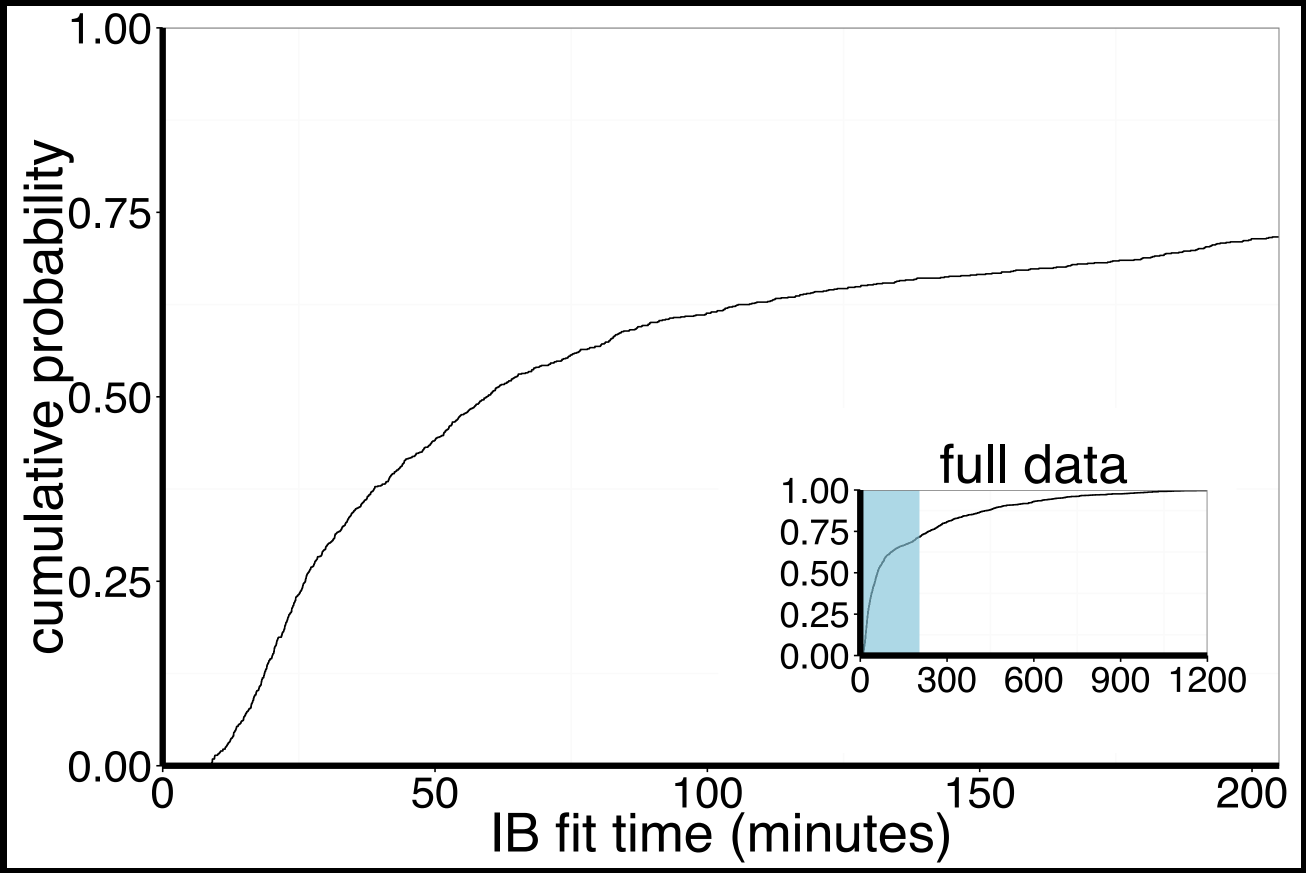
$$H(T) = H(T \mid X) = \text{const} > 0$$

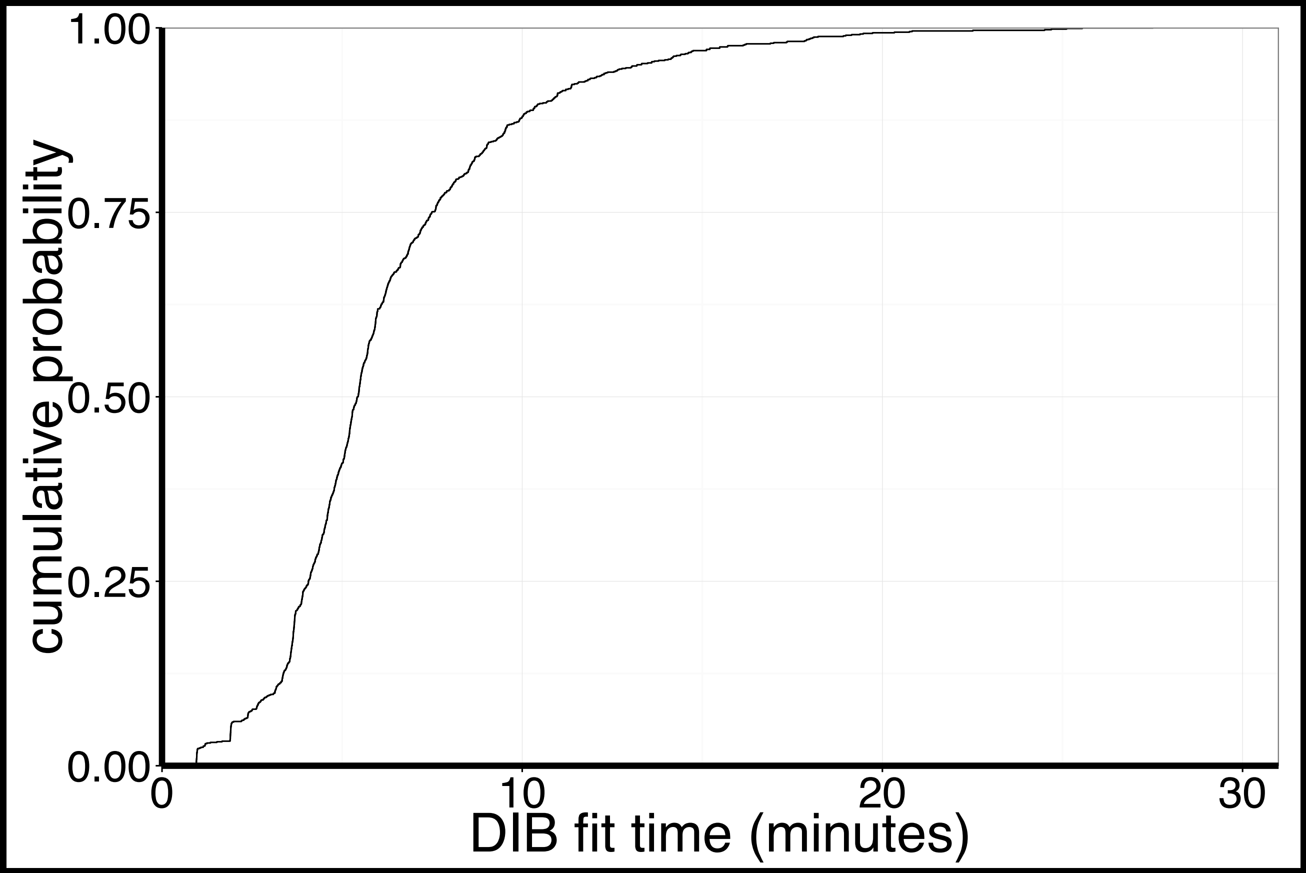
# IB vs DIB: experiments



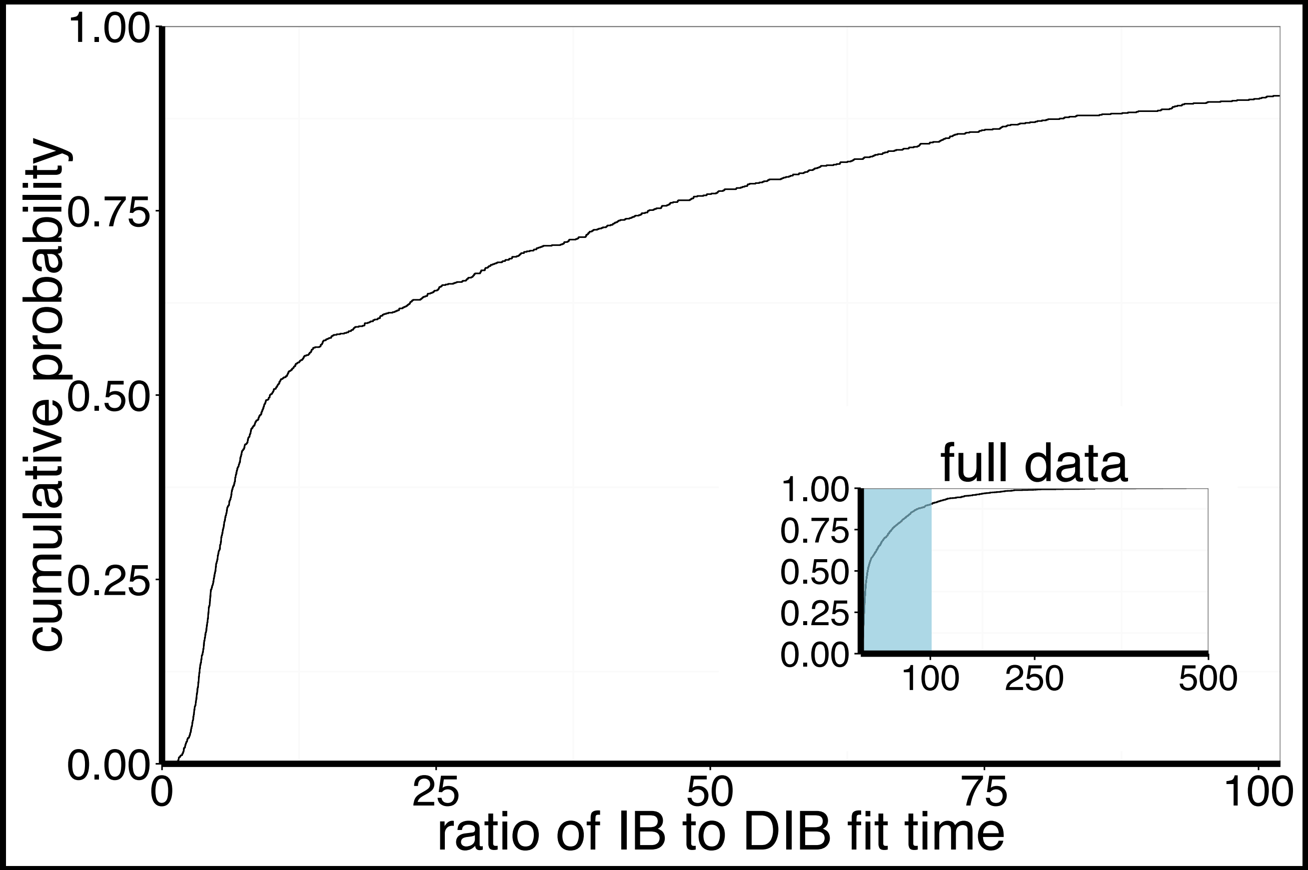

IB and DIB behave similarly in terms of IB cost function…
…but DIB performs far better in terms of its own

# IB vs DIB: efficiency



# IB vs DIB: efficiency



# 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

# Thanks

- David Schwab & Bill Bialek
- The Krell Institute
- The Hertz Foundation