Disentangling Representations in Restricted Boltzmann Machines without Adversaries

J. Fernandez-de-Cossio-Diaz

PSL Junior Fellow AI
Ecole Normale Supérieure, Paris

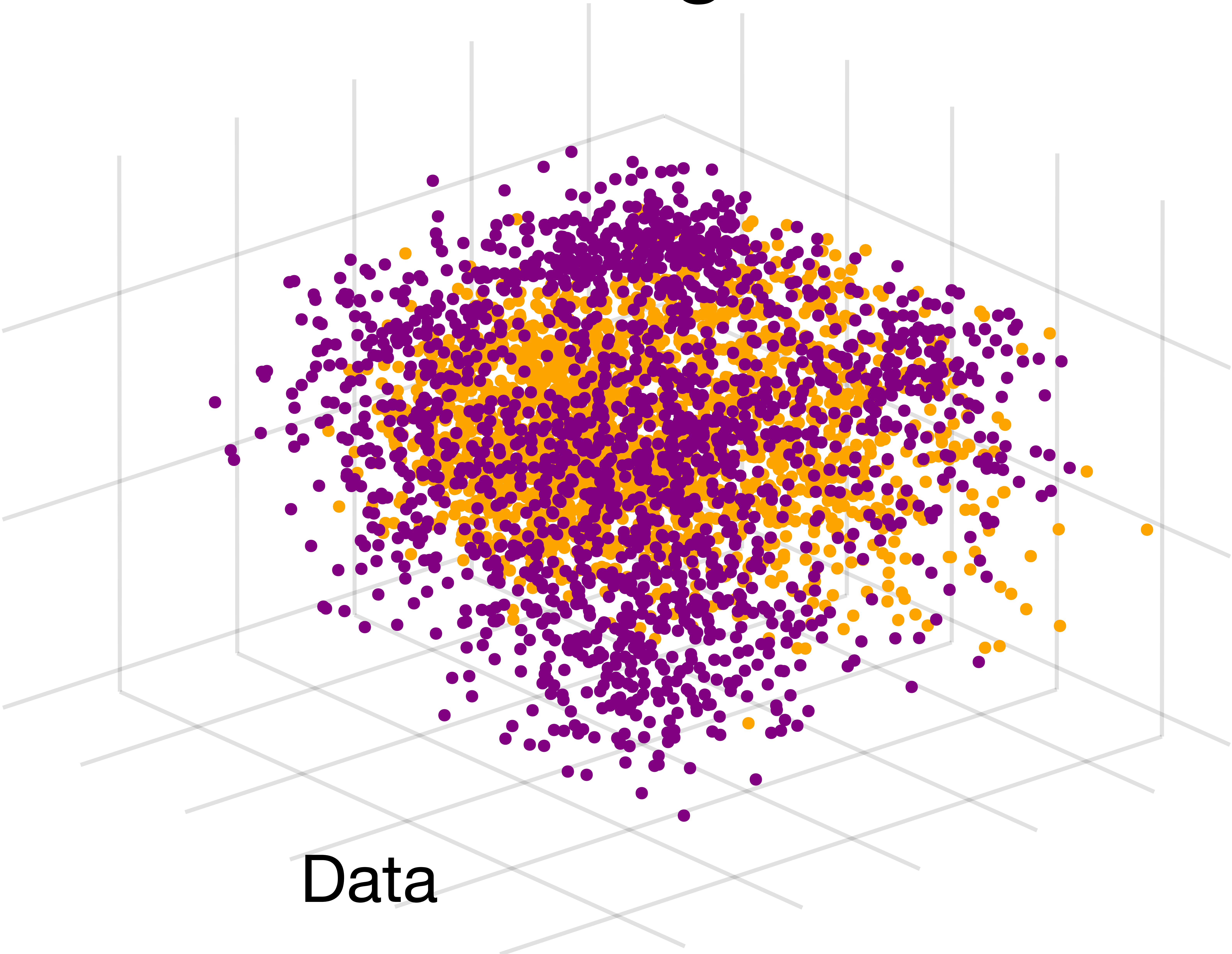
In collaboration with:

S. Cocco & R. Monasson

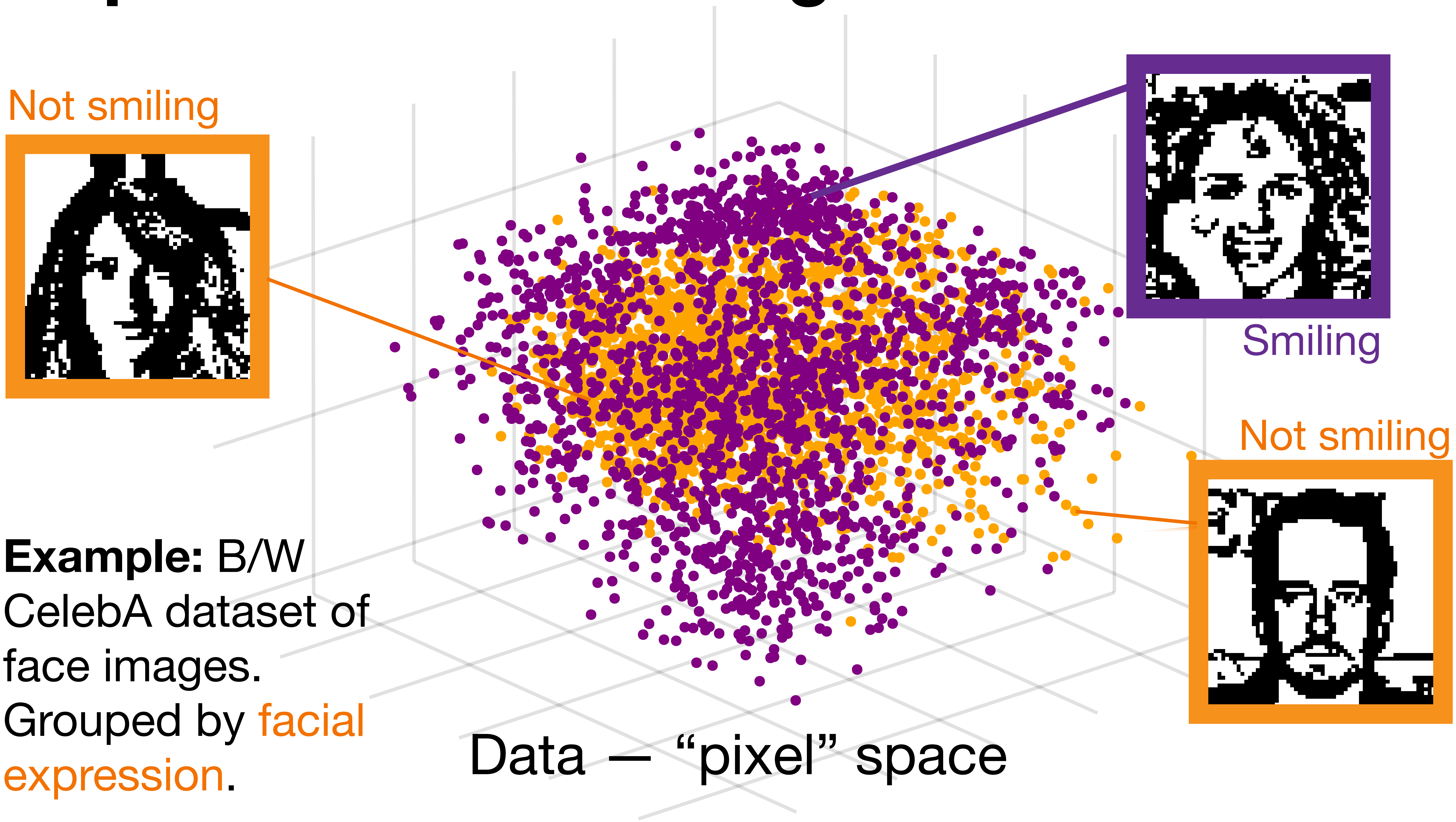
JFdCD, S.Cocco, R.Monasson, PRX 13, 021003 (2023)

Trento 29/06/2023, ECT* Machine Learning for Lattice Field Theory and Beyond

Representation learning



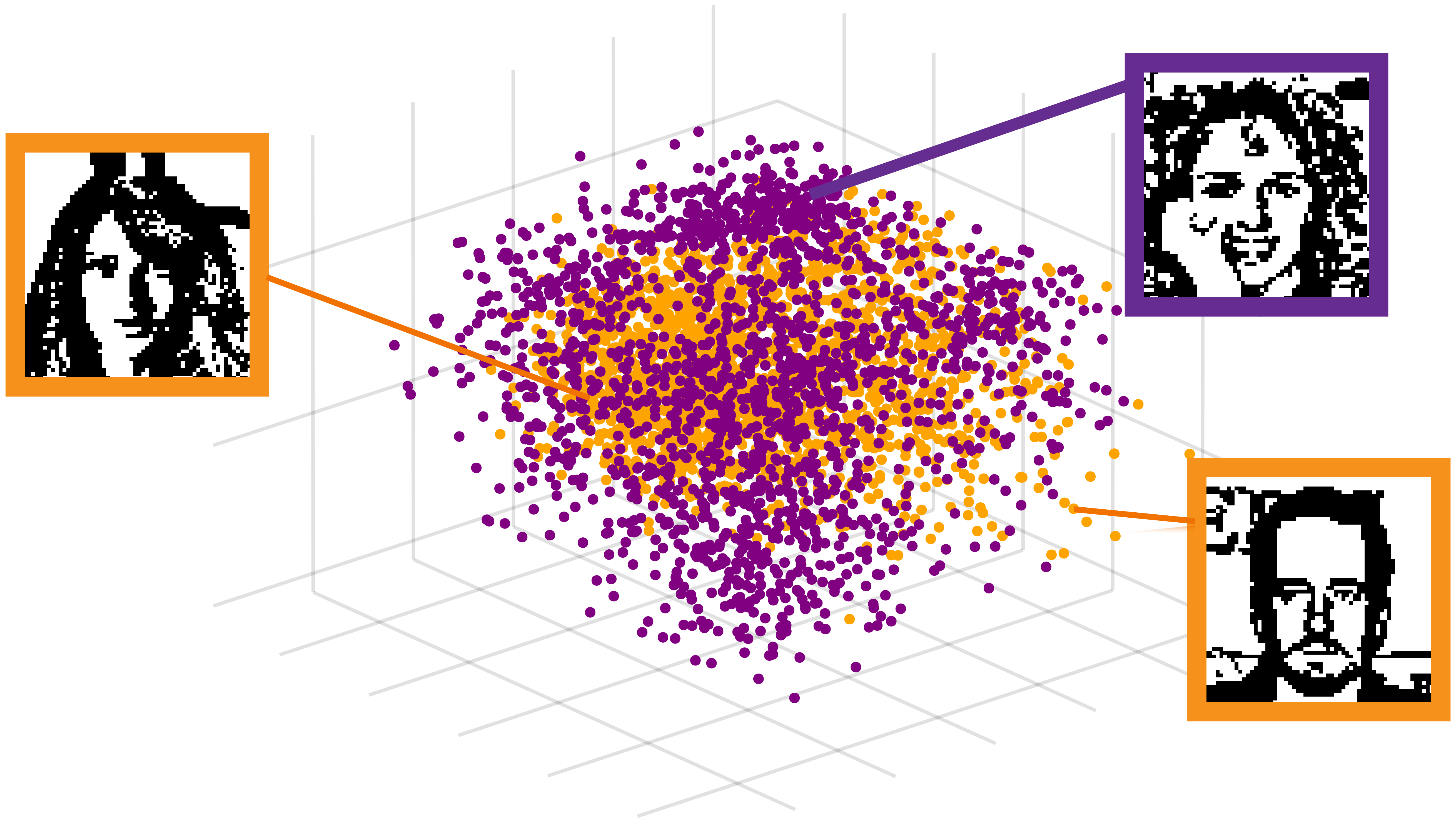
Representation learning



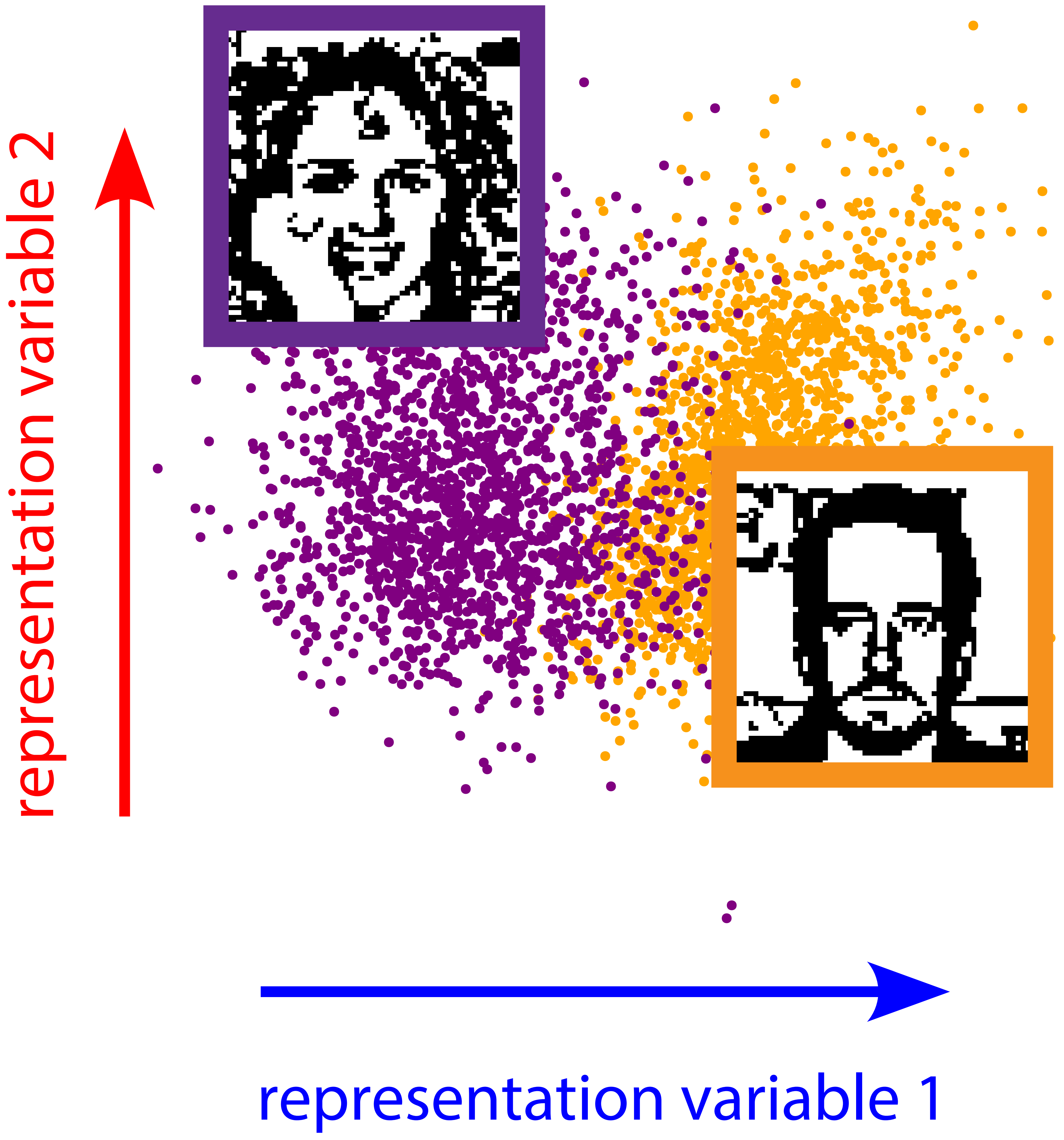
Example: B/W CelebA dataset of face images. Grouped by facial expression.

Representation learning

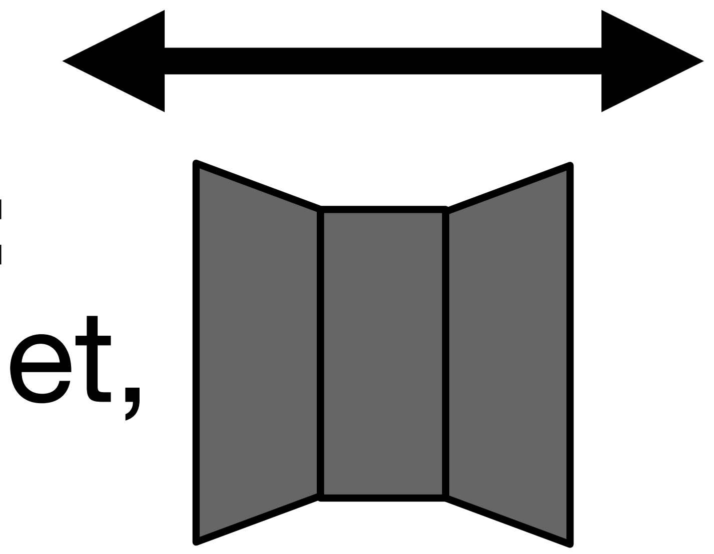
Data — “pixel” space



Representation space

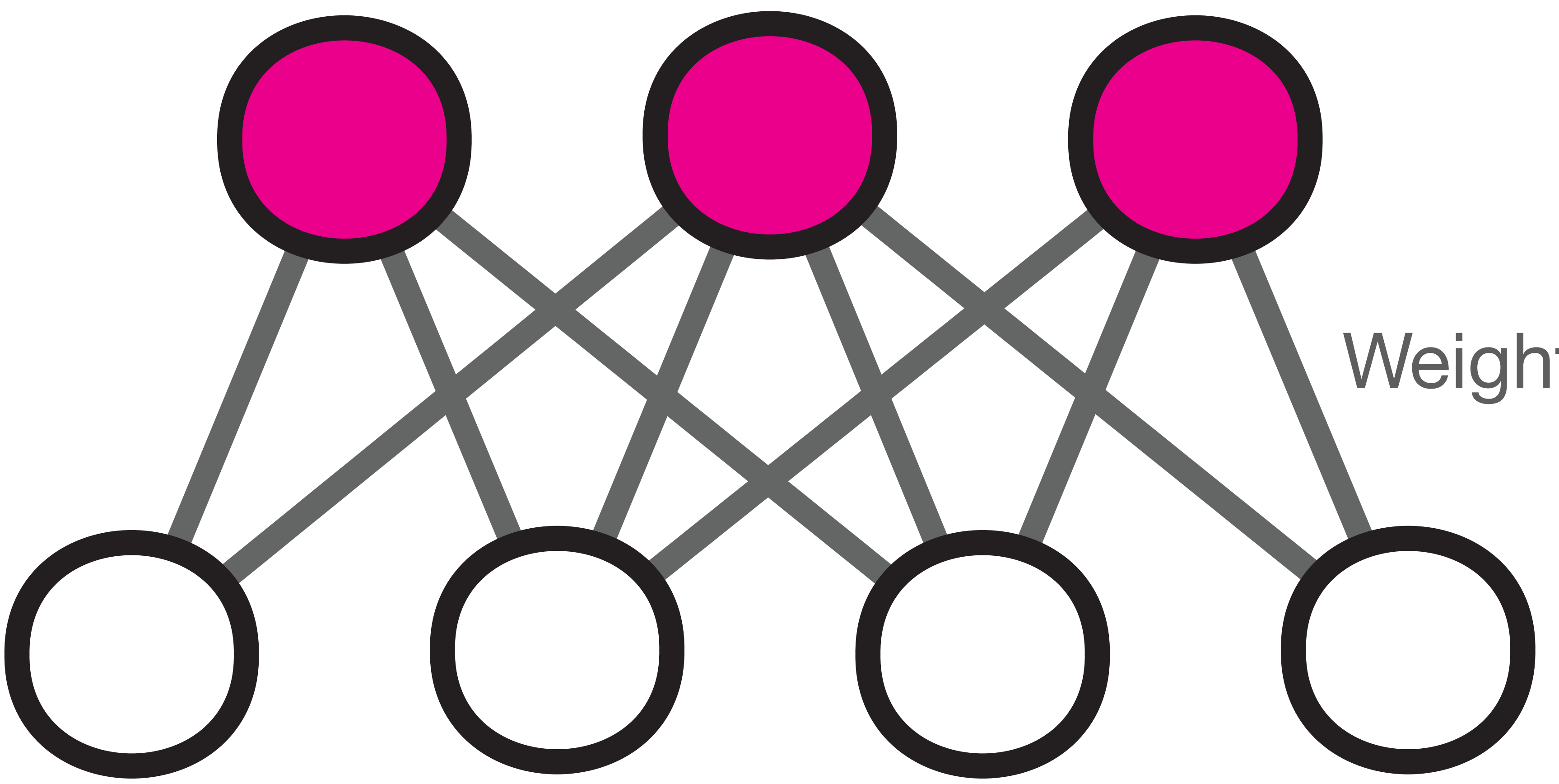


Mapping:
e.g. neural net,
...



Restricted Boltzmann machines

Simple generative model, implementing data / representation duality

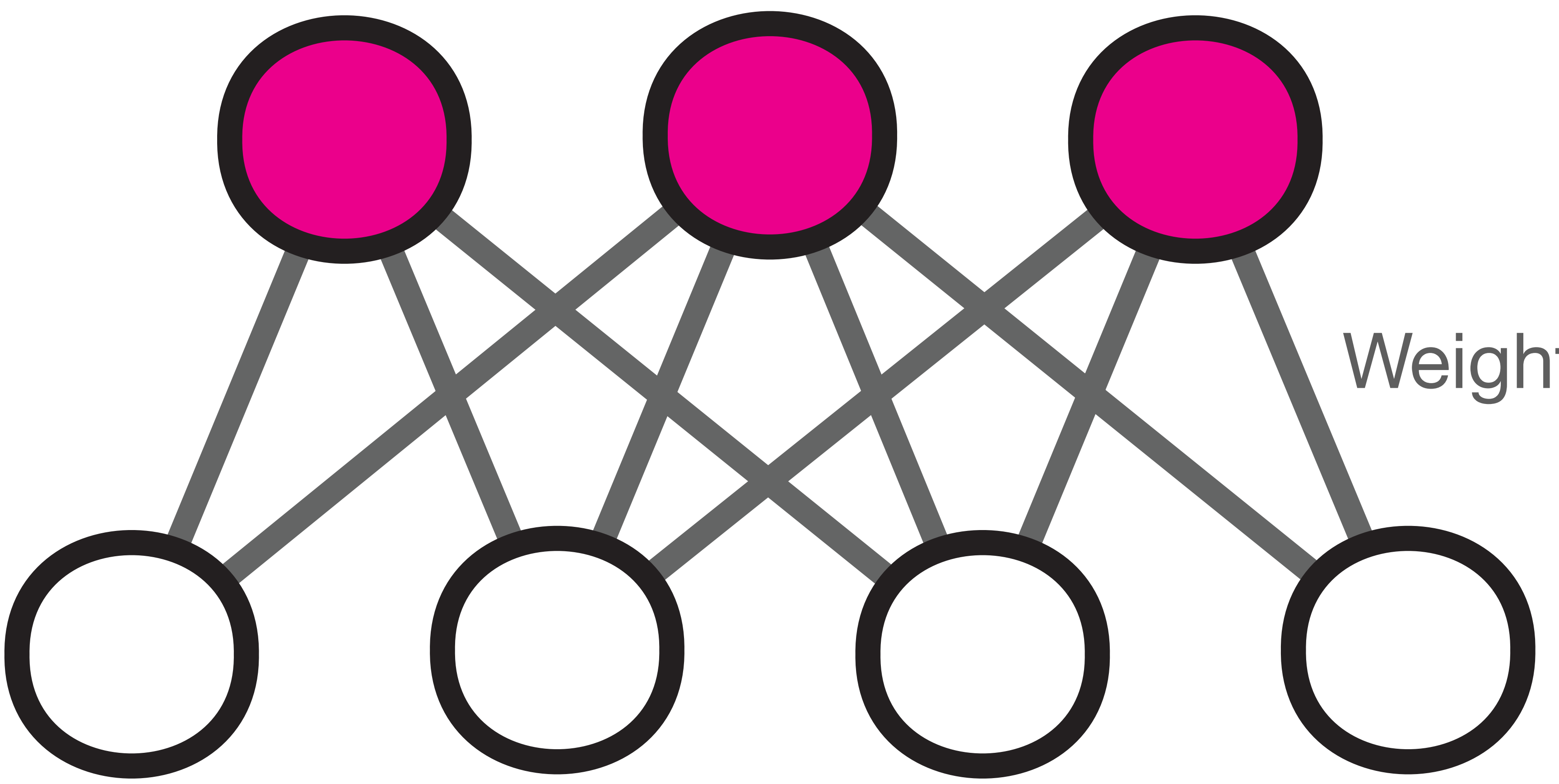


Energy function:

$$E(\mathbf{v}, \mathbf{h}) = \sum_{\mu} \mathscr{V}_i(v_i) + \sum_{\mu} \mathscr{U}_\mu(h_\mu) - \sum_{i\mu} w_{i\mu} v_i h_\mu$$

Restricted Boltzmann machines

Simple generative model, implementing data / representation duality



Energy function:

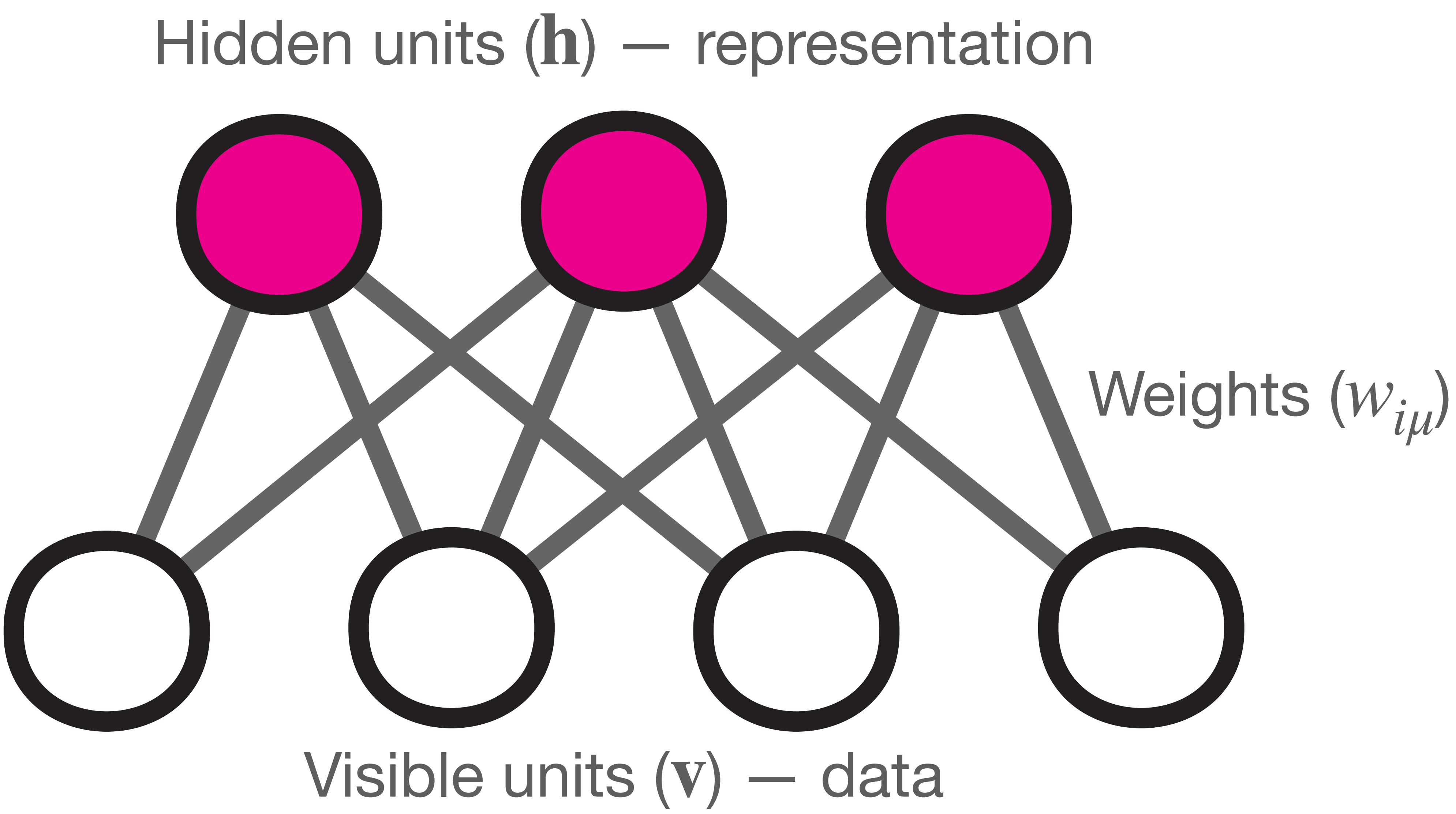
$$E(\mathbf{v}, \mathbf{h}) = \sum_{\mu} \mathscr{V}_i(v_i) + \sum_{\mu} \mathscr{U}_\mu(h_\mu) - \sum_{i\mu} w_{i\mu} v_i h_\mu$$

Likelihood: $\mathscr{L} = \ln P(\mathbf{v}) = \underbrace{\ln \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}}_{-E_{\text{eff}}(\mathbf{v})} - \ln Z$

Encodes for (higher-order) interactions

Restricted Boltzmann machines

Simple generative model, implementing data / representation duality



Energy function:

$$E(\mathbf{v},\mathbf{h}) = \sum_{\mu} \mathscr{V}_i(v_i) + \sum_{\mu} \mathscr{U}_\mu(h_\mu) - \sum_{i\mu} w_{i\mu} v_i h_\mu$$

Likelihood: $\mathscr{L} = \ln P(\mathbf{v}) = \ln \underbrace{\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}}_{-E_{\text{eff}}(\mathbf{v})} - \ln Z$

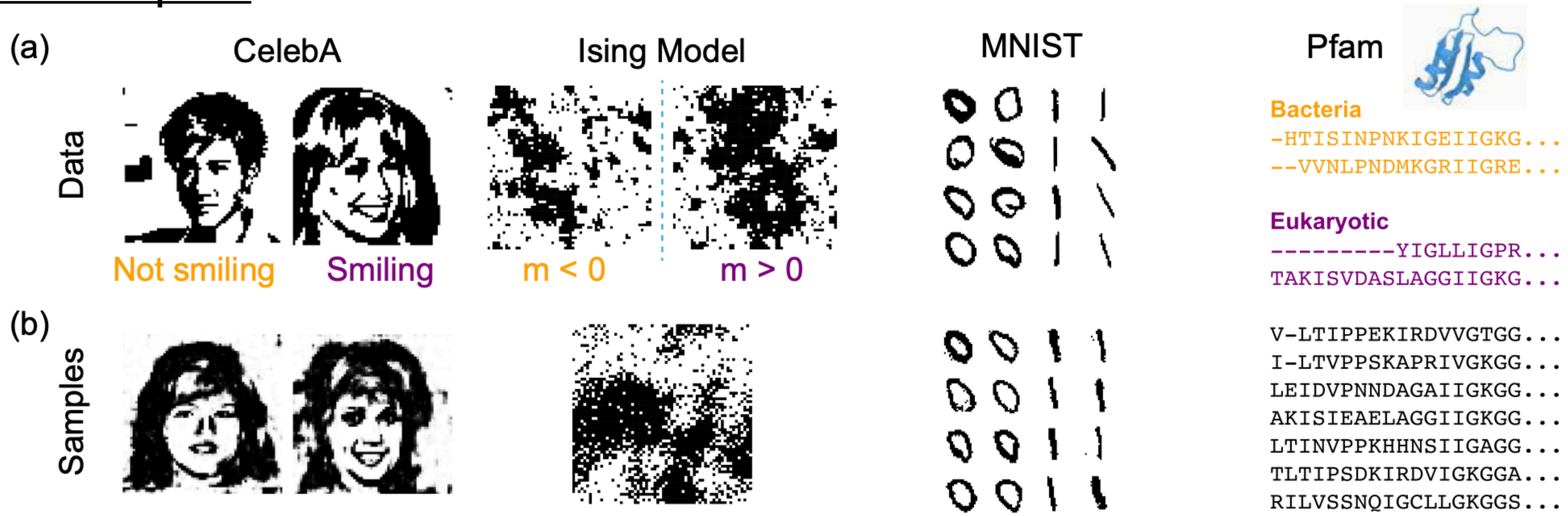
Model trained by max. likelihood of the data

$$\frac{\partial \ln P(\text{data})}{\partial \omega} = \left\langle \frac{\partial(-E_{\text{eff}}(\mathbf{v}))}{\partial \omega} \right\rangle_{\text{data}} - \left\langle \frac{\partial(-E_{\text{eff}}(\mathbf{v}))}{\partial \omega} \right\rangle_{\text{model}}$$

Encodes for (higher-order) interactions

RBM as generative models across diverse datasets

Some examples

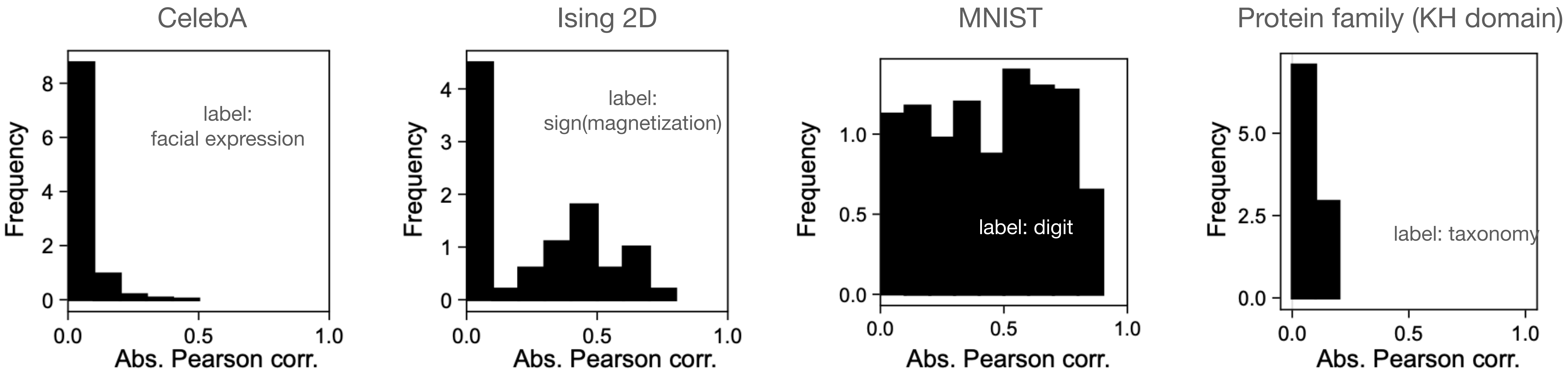


Other recent applications (in biology) …

Immunology: Bravi, et al. bioRxiv (2022): 2022-12; Cell systems 12.2 (2021): 195-202.
RNA: Di Gioacchino, et al. PLoS CB 18.9 (2022): e1010561, JFdCD, et al. bioRxiv (2023): 2023-05
Proteins: Tubiana Elife 8 (2019): e39397.

RBM as generative models across diverse datasets

Nature of the representations

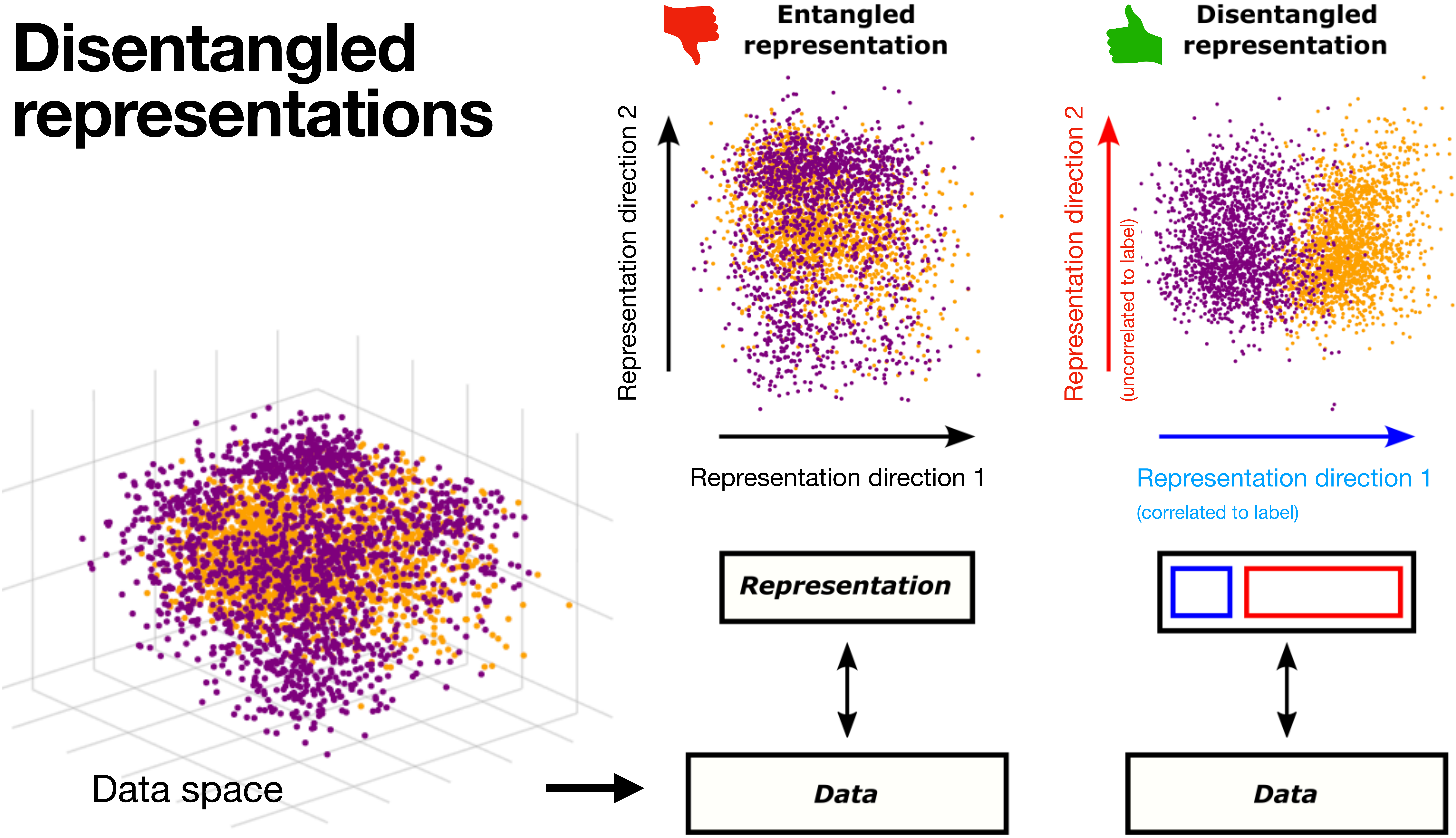


Correlation between hidden units and label

Information about features is distributed across a large number of hidden units

Representations are generally *entangled*.

Disentangled representations



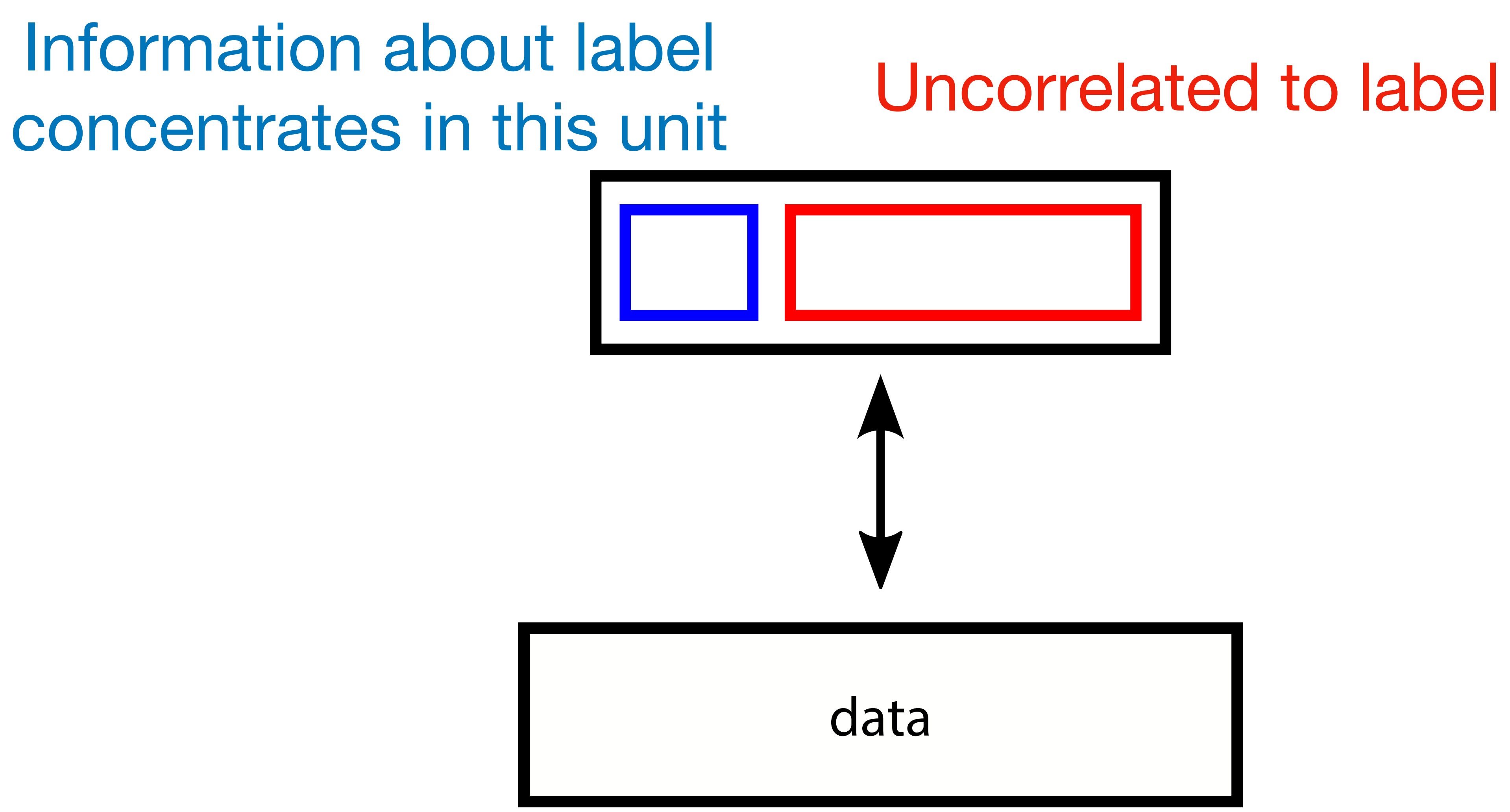
Adversarial formulation

Setup: Dataset $\mathbf{v}^1, \ldots, \mathbf{v}^B$ with binary labels $u^1, \ldots, u^B$ *e.g.* faces smiling vs. or not

$\max \mathscr{L}(\text{data})$

subject to

$\text{MI}(h, \text{label}) = 0$



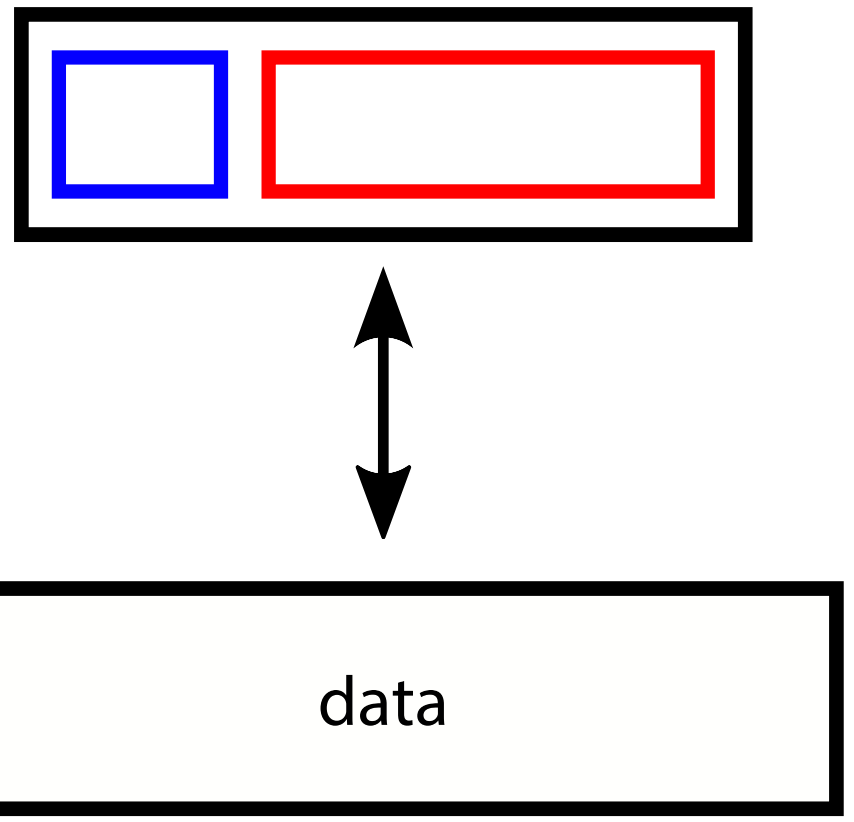
Mutual information (MI)
is difficult to control …

Lample, G, et al. "Fader networks:." NIPS 30 (2017).

Adversarial formulation

Setup: Dataset $\mathbf{v}^1, \ldots, \mathbf{v}^B$ with labels $u^1, \ldots, u^B$ *e.g.* faces smiling vs. or not

Adversarial classifier $\mathscr{L}_{\mathscr{A}}$



$$\max \mathscr{L}(\text{data})$$

subject to

$$\text{MI}(h, \text{label}) = 0$$

“Adversarial” training.
Similar to GAN

$$\max \left(\mathscr{L}(\text{data}) - \alpha \max_{\mathscr{A}} \mathscr{L}_{\mathscr{A}}(\text{labels} \,|\, \text{data}) \right)$$

If the best possible classifier $\mathscr{L}_{\mathscr{A}}$ is bad at predicting the label, then label information has been erased

Lample, G, et al. "Fader networks:." NIPS 30 (2017).

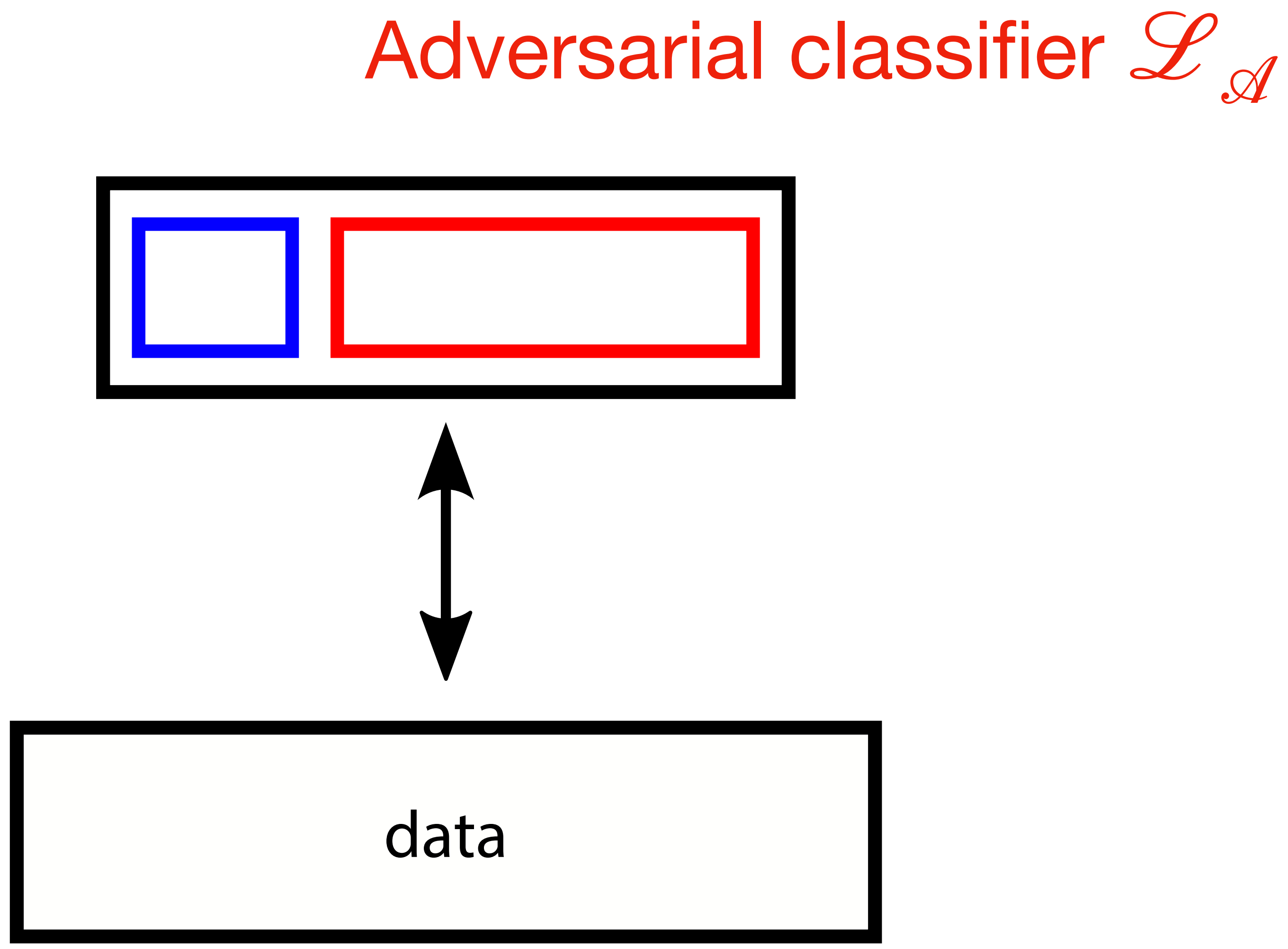
Adversarial formulation

Setup: Dataset $\mathbf{v}^1, \ldots, \mathbf{v}^B$ with labels $u^1, \ldots, u^B$ *e.g.* faces smiling vs. or not

$\max \mathscr{L}(\text{data})$

subject to

$\text{MI}(h, \text{label}) = 0$



Issues:

- Unstable training
- What constraints are imposed by the classifier?
- Likelihood cost?

"Adversarial" training.
Similar to GAN

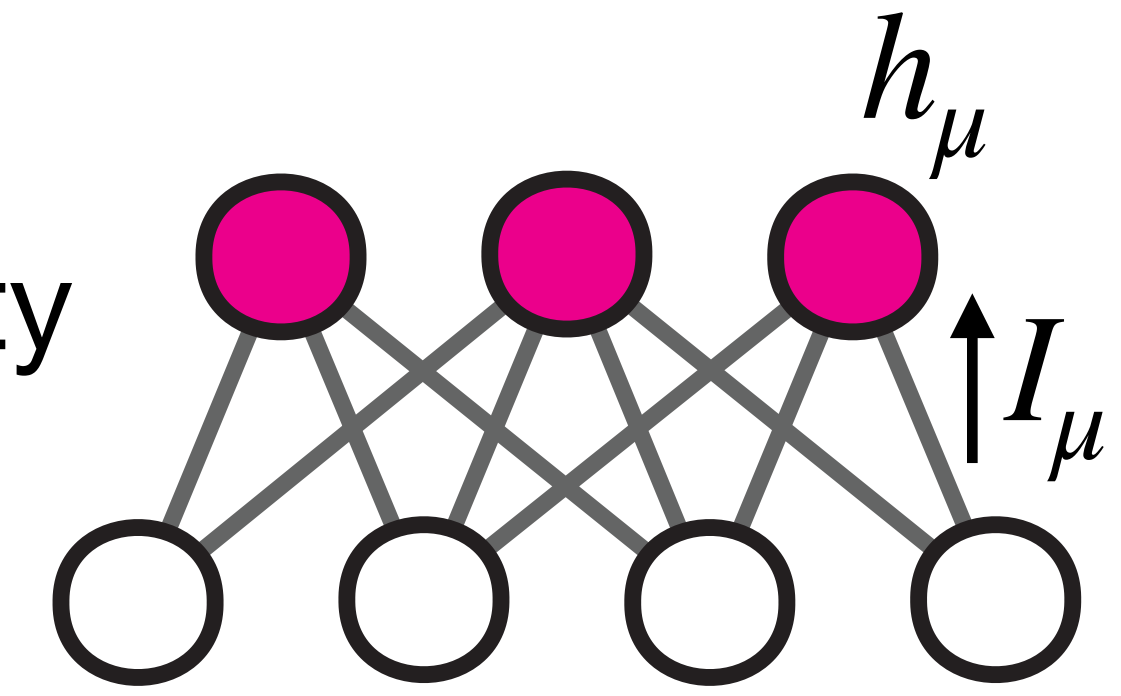
$$\max \left(\mathscr{L}(\text{data}) - \alpha \max_{\mathscr{A}} \mathscr{L}_{\mathscr{A}}(\text{labels} \mid \text{data}) \right)$$

If the best possible classifier $\mathscr{L}_{\mathscr{A}}$ is bad at predicting the label, then label information has been erased

Lample, G, et al. "Fader networks:." NIPS 30 (2017).

A hierarchy of explicit constraints

Setup: Dataset $\mathbf{v}^1, \ldots, \mathbf{v}^B$ with labels $u^1, \ldots, u^B$ *e.g.* faces smiling vs. or not



1) MI$(h$, label$)$ $\leq$ MI$(I = \mathbb{W}^\top \mathbf{v}$, label$)$ — data processing inequality

2) MI$(I = \mathbb{W}^\top \mathbf{v}$, label$) = 0$ implies that:

$$\langle u I_\mu \rangle = 0, \langle u I_\mu I_\nu \rangle = 0, \langle u I_\mu I_\nu I_\gamma \rangle = 0, \ldots$$

(correlations between I_μ and label vanish at all orders)

Explicit conditions on the weights — No "adversary"

$$\sum_i q_i^{(1)} w_{i\mu} = 0, \sum_i q_{ij}^{(2)} w_{i\mu} w_{j\nu} = 0, \text{etc.}, \ldots$$

A hierarchy of explicit constraints

1st-order: $\langle u I_\mu \rangle = 0$

2nd-order: $\langle u I_\mu I_\nu \rangle = 0$

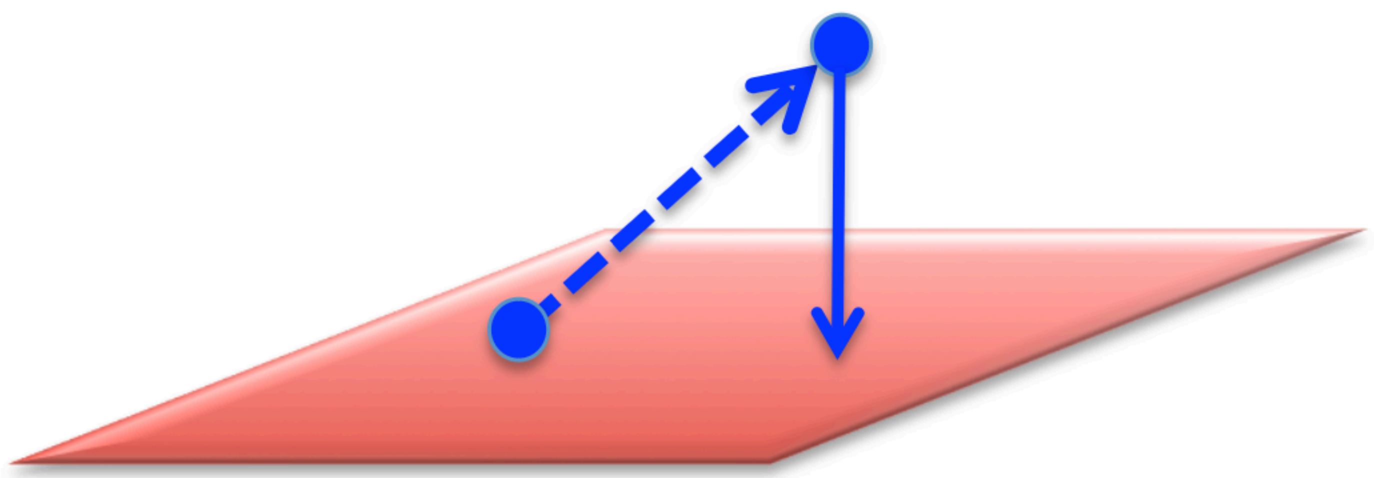
Explicit conditions on the weights — No “adversary”

$$\sum_i q_i^{(1)} w_{i\mu} = 0$$

$$\sum_i q_{ij}^{(2)} w_{i\mu} w_{j\nu} = 0$$

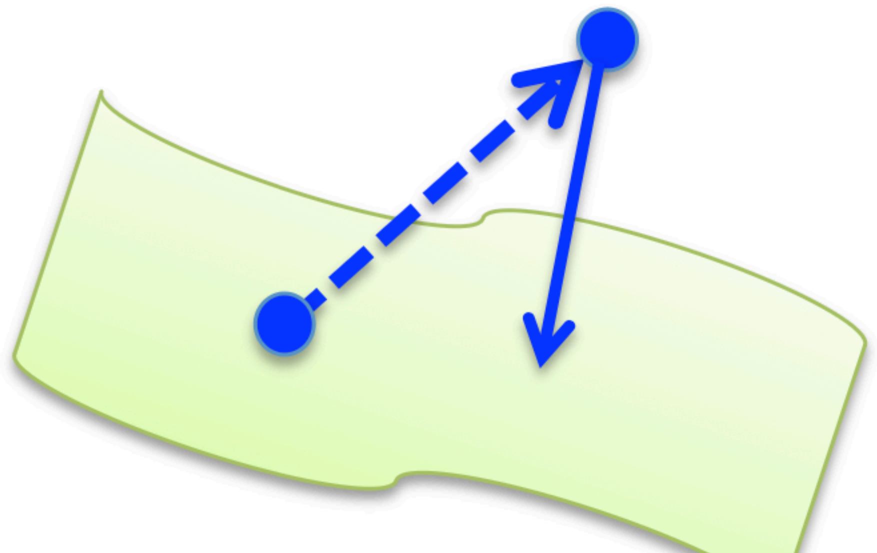
$$q_i^{(1)} = \langle u v_i \rangle - \langle u \rangle \langle v_i \rangle$$

$$q_i^{(2)} = \langle u v_i v_j \rangle - \langle u \rangle \langle v_i v_j \rangle$$



Learning algorithm

- Gradient ascent of likelihood
- Projection to satisfy constraints



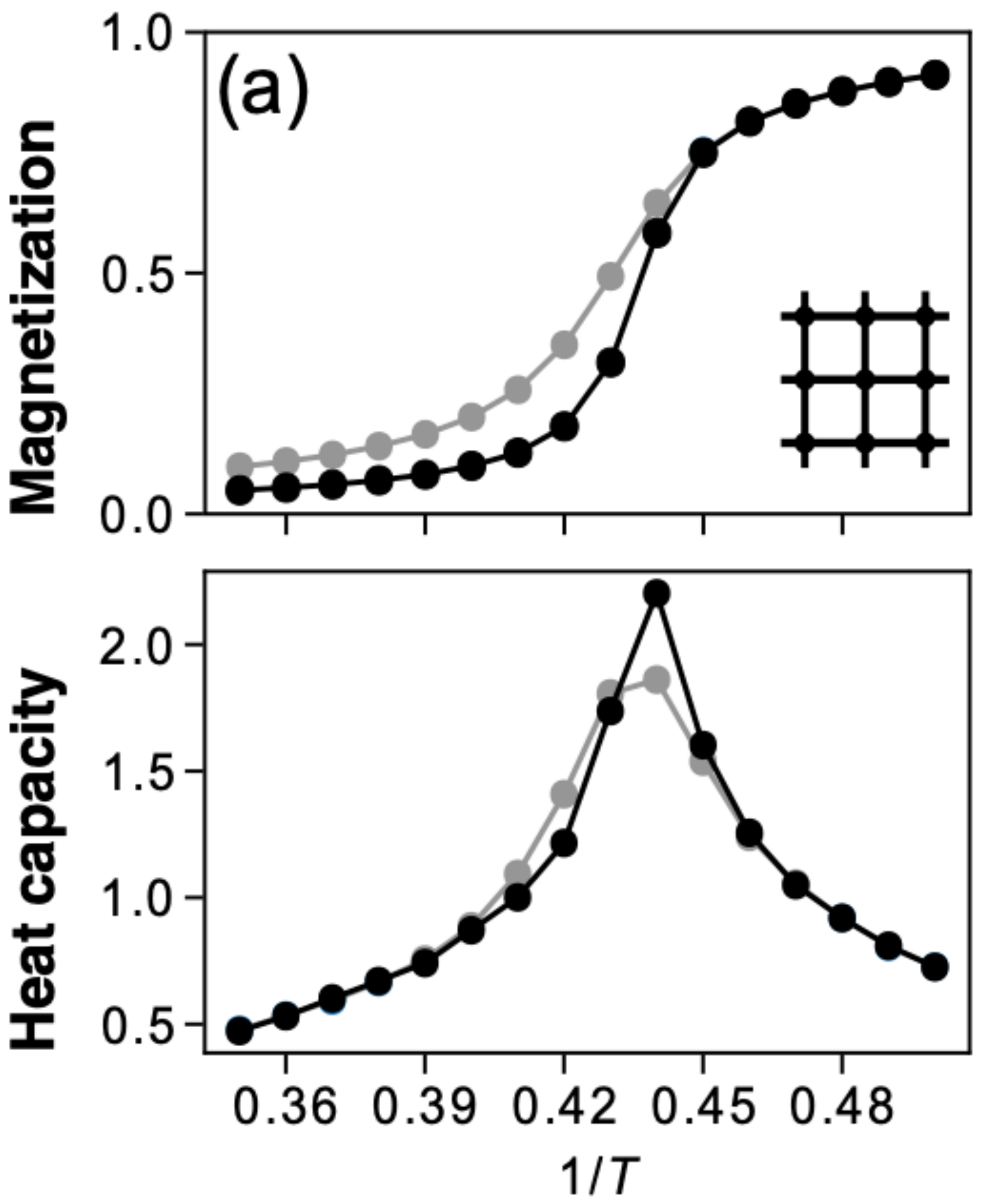
Sylvester

Trained RBM concentrates label information

Constraint imposed on ALL hidden units:
Partial Erasure of label information

Constraint imposed on a subset of hidden units:
Concentrates label information

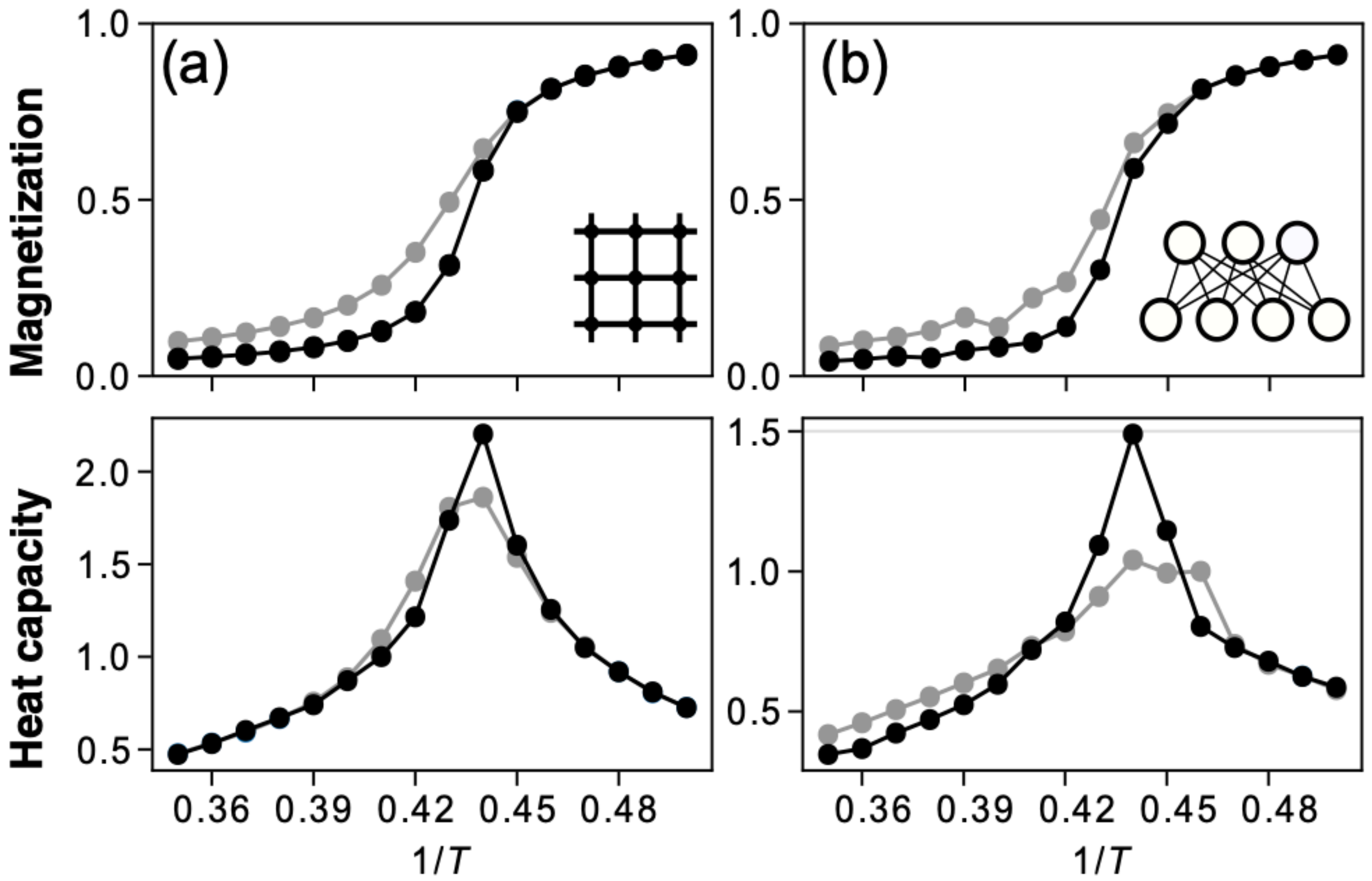
Application to 2D-Ising model



2-Dimensional
Ising model

N = 32x32 and
64x64

Application to 2D-Ising model



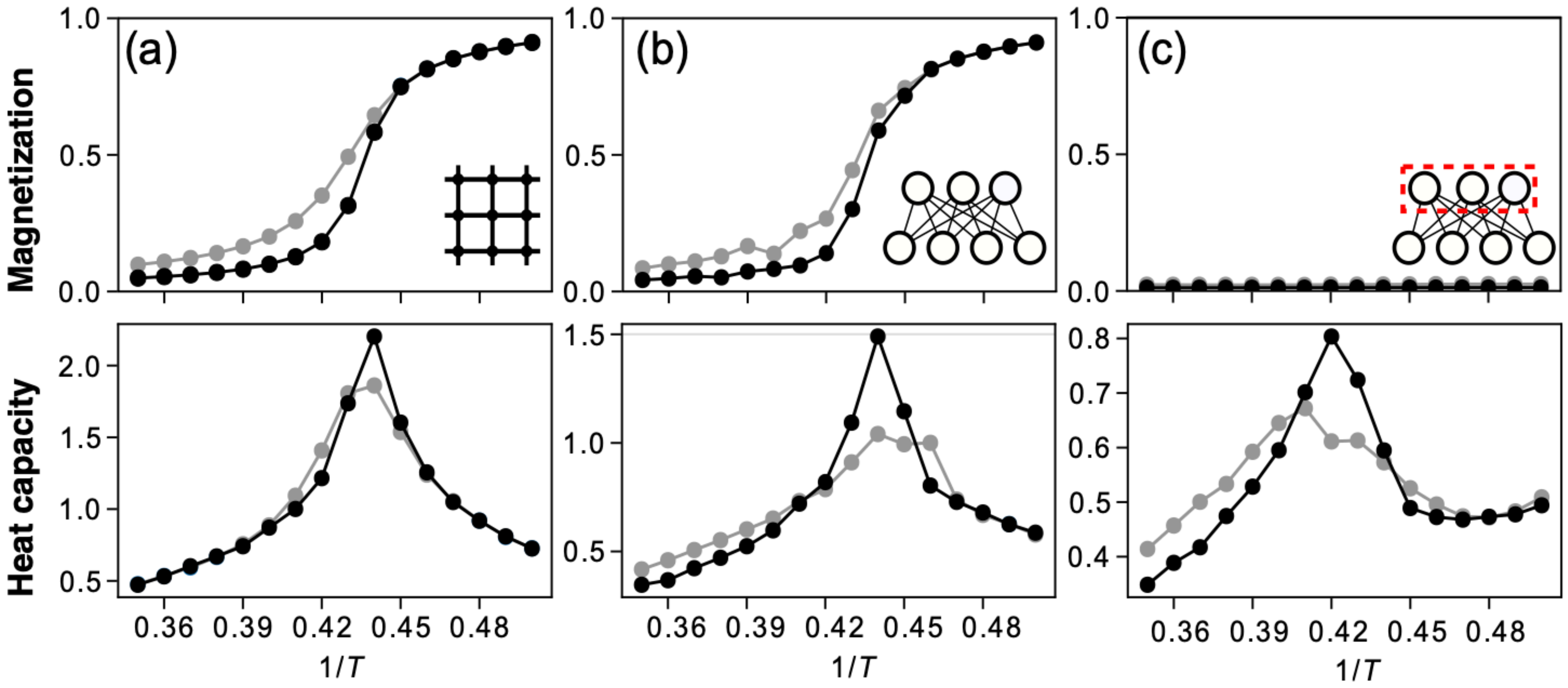
2-Dimensional Ising model

N = 32x32 and 64x64

RBM captures behaviour of observables

Application to 2D-Ising model

Label = sign(m)



1st-order constrained RBM has $m = 0$ but heat capacity has correct behavior.

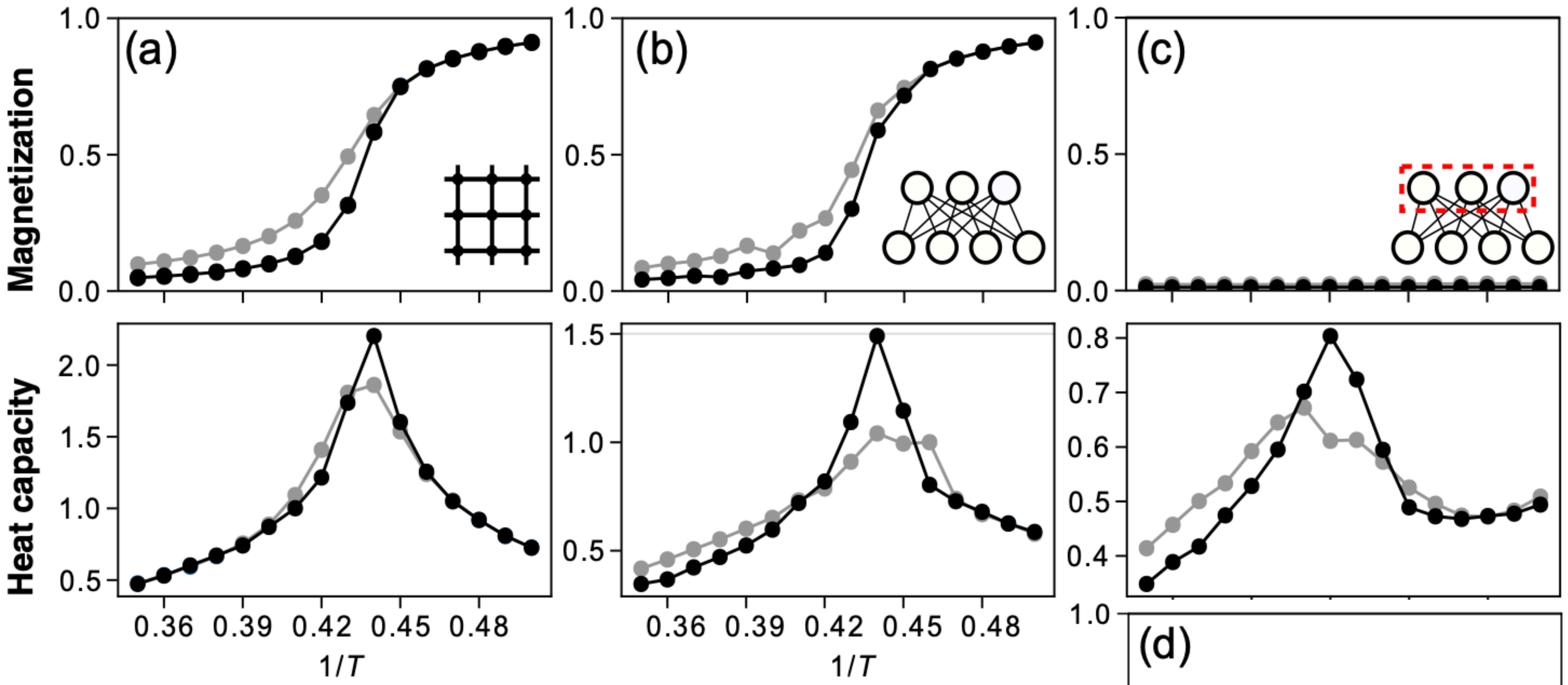
2-Dimensional Ising model

N = 32x32 and 64x64

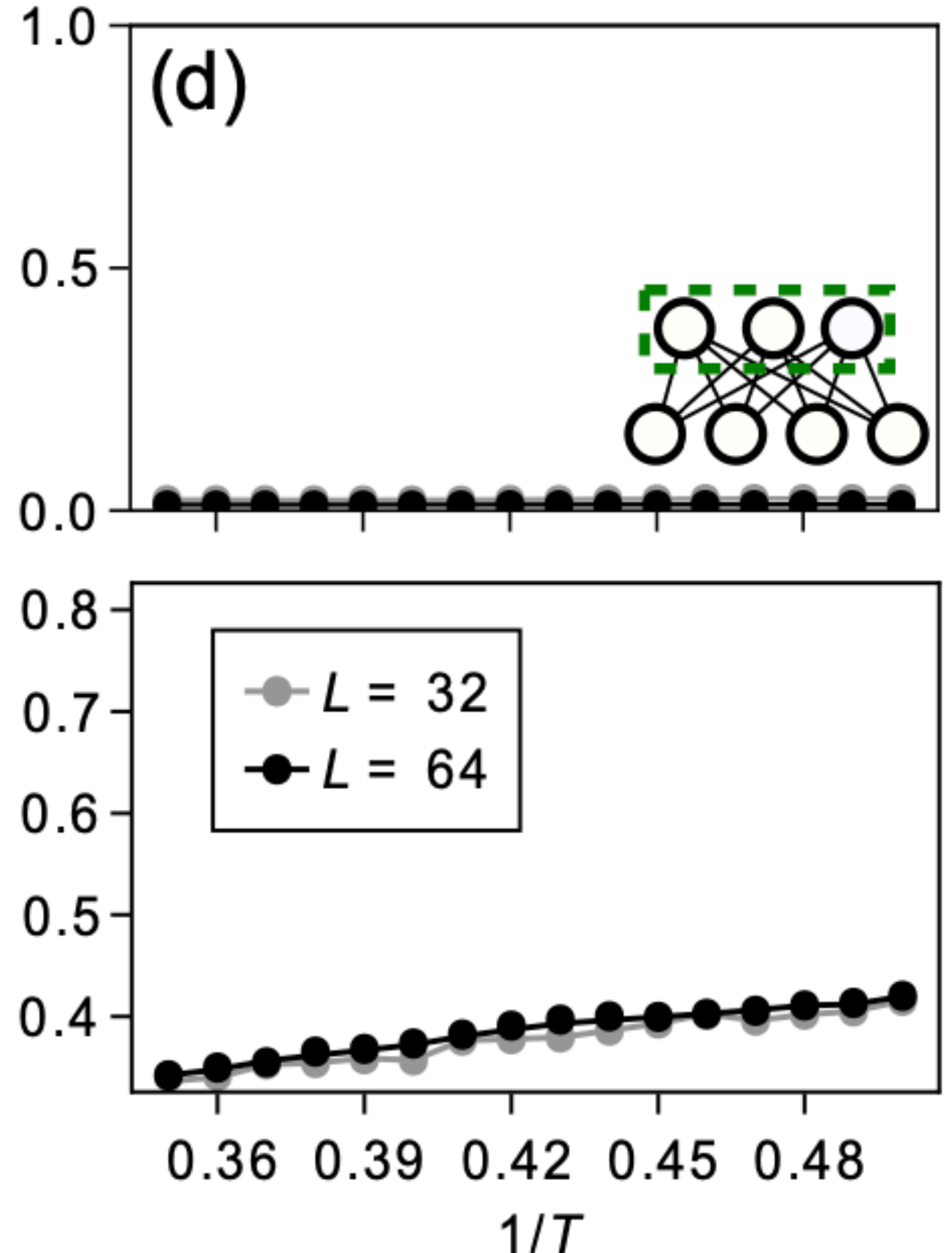
RBM captures behaviour of observables

Application to 2D-Ising model

Label = sign(m)



1st-order constrained RBM has $m = 0$ but heat capacity has correct behavior.



2nd-order constraint -> no correlations

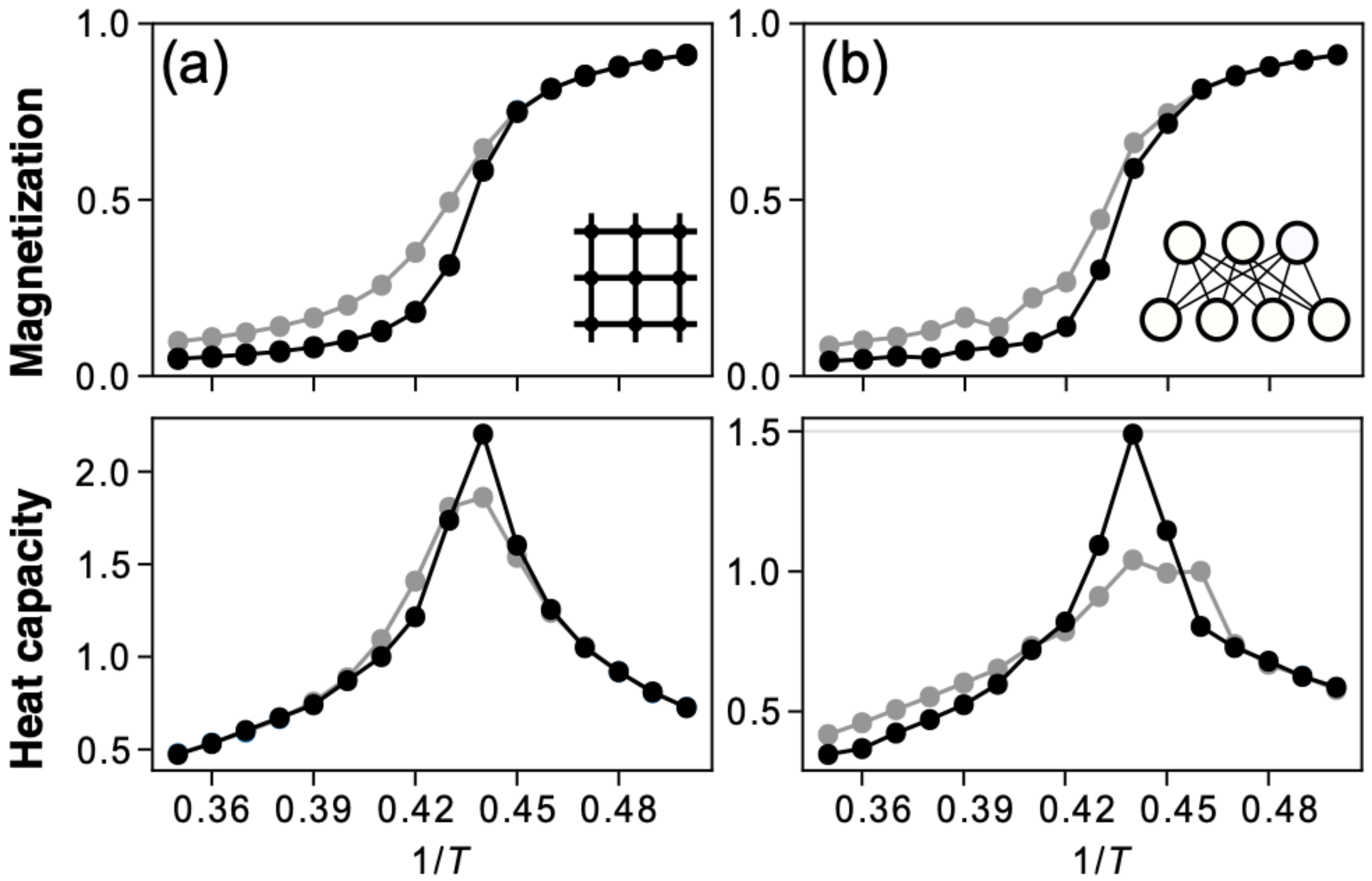
2-Dimensional Ising model

N = 32x32 and 64x64

RBM captures behaviour of observables

Application to 2D-Ising model

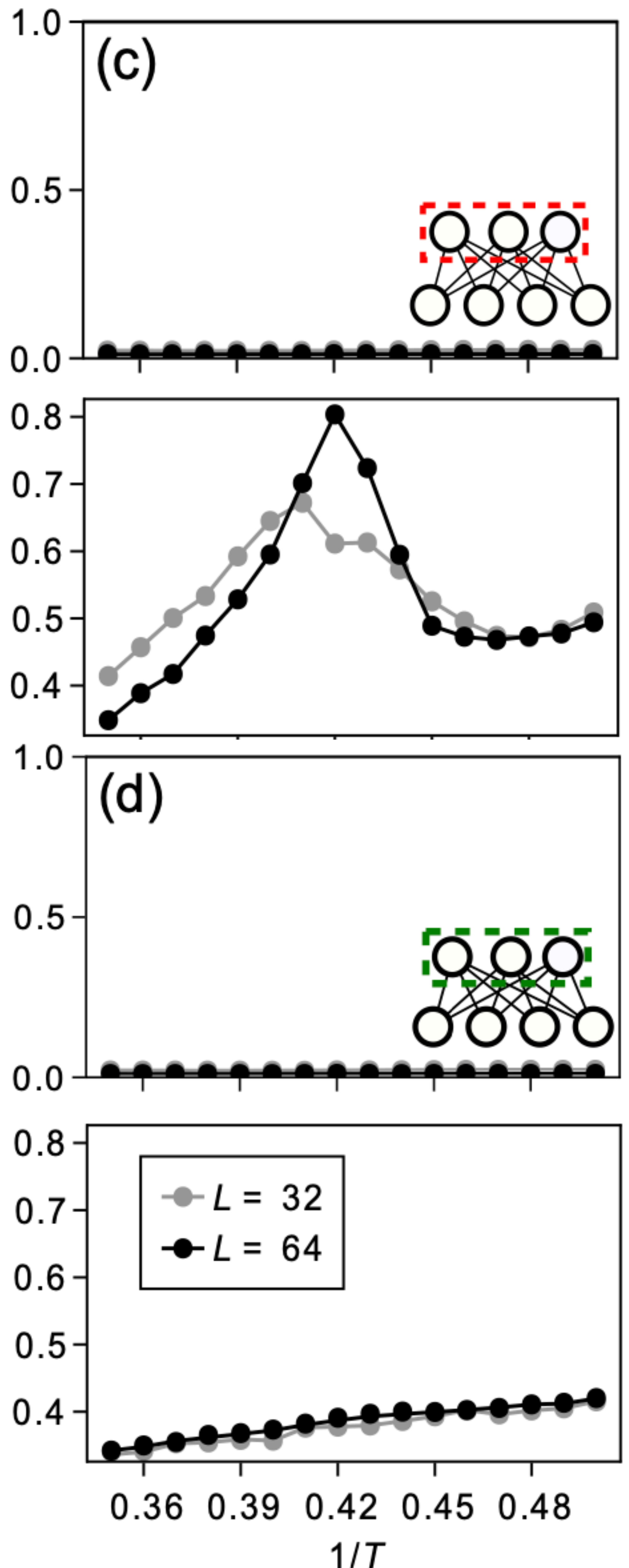
Label = sign(m)



2-Dimensional Ising model

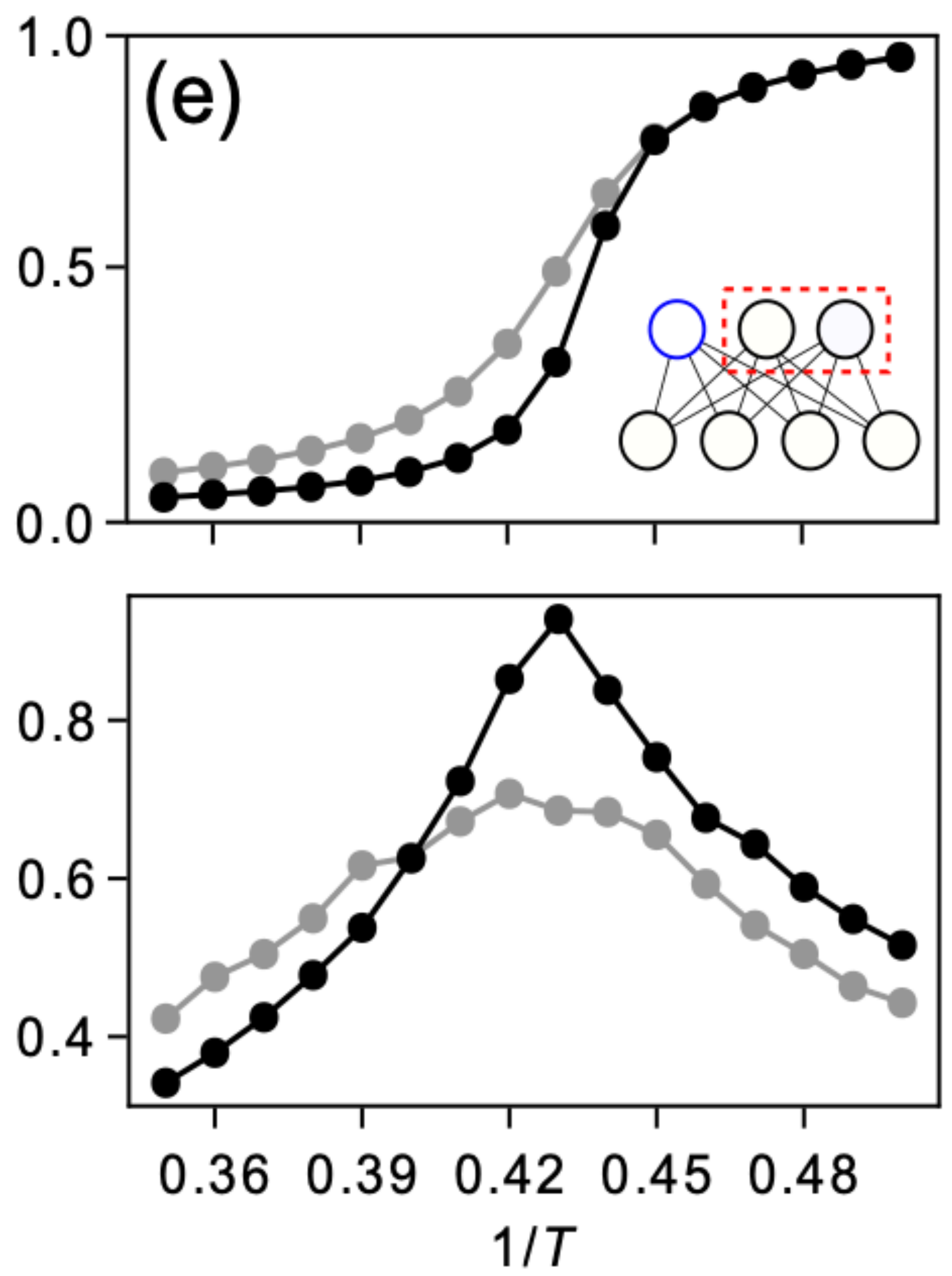
N = 32x32 and 64x64

RBM captures behaviour of observables



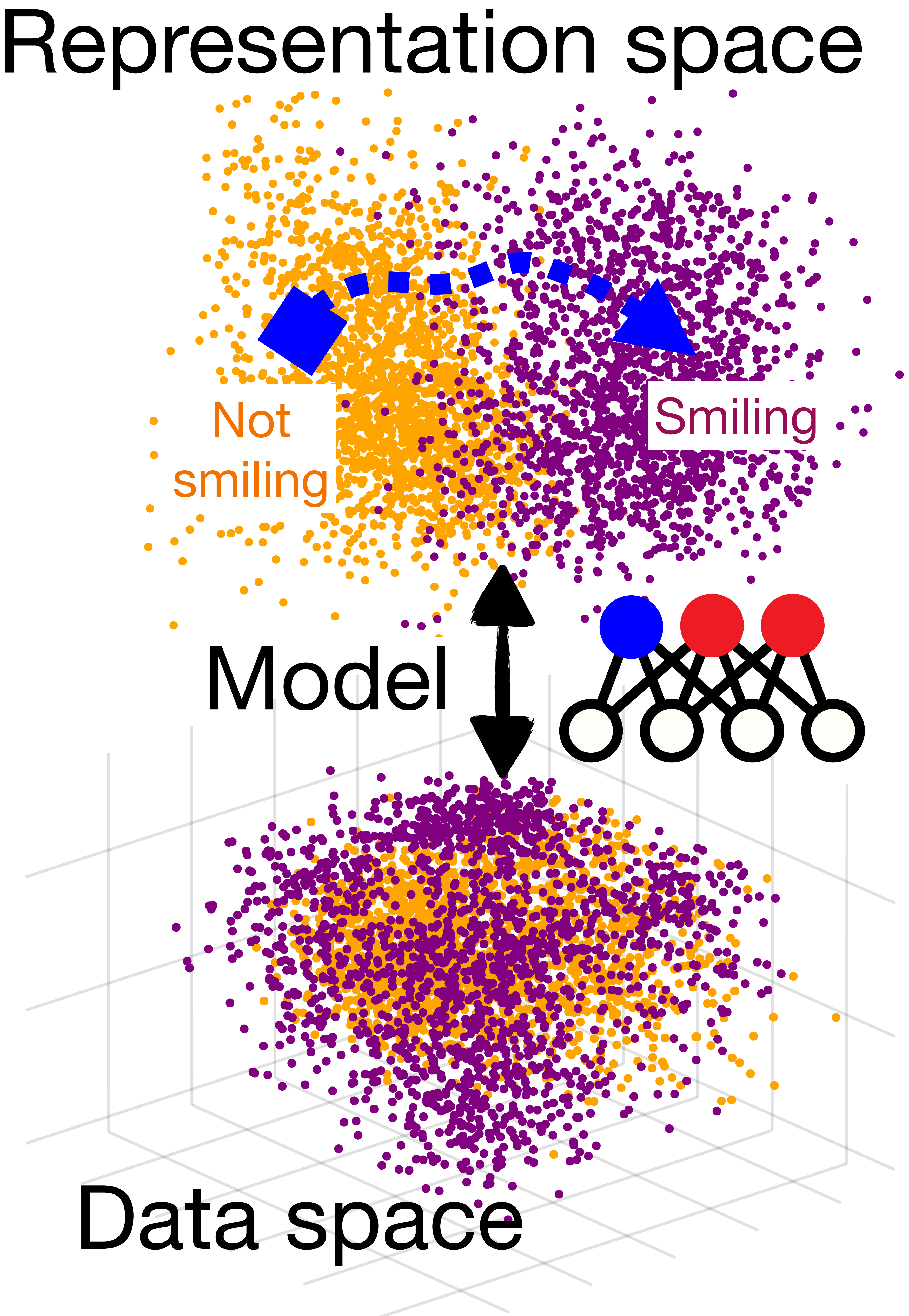
1st-order constrained RBM has $m = 0$ but heat capacity has correct behavior.

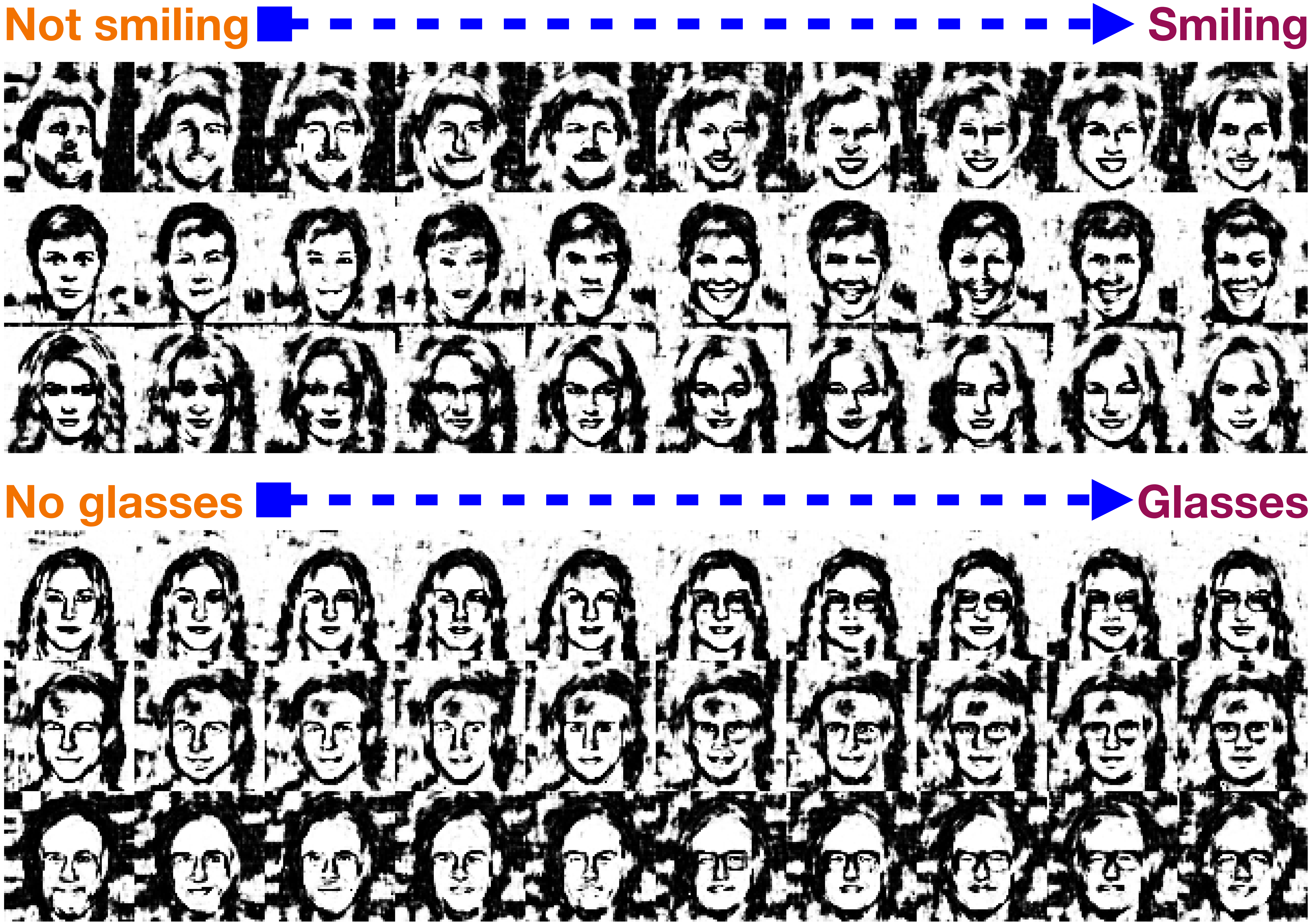
2nd-order constraint -> no correlations



Releasing one hidden unit recovers behavior of all observables

Manipulating data through representation learning





Likelihood cost in a Gaussian setting

$$LL_{\text{Gauss}} = \frac{1}{2} \sum_{\mu} \left(\lambda_\mu - 1 - \ln \lambda_\mu \right)$$

λ_μ eigenvalues of correlation matrix of data

Likelihood cost in a Gaussian setting

$$LL_{\text{Gauss}} = \frac{1}{2} \sum_{\mu} (\lambda_\mu - 1 - \ln \lambda_\mu)$$

λ_μ eigenvalues of correlation matrix of data

Constraint: $\tilde{\mathbf{C}}^{\perp} = \mathbf{P}\tilde{\mathbf{C}}\mathbf{P}$

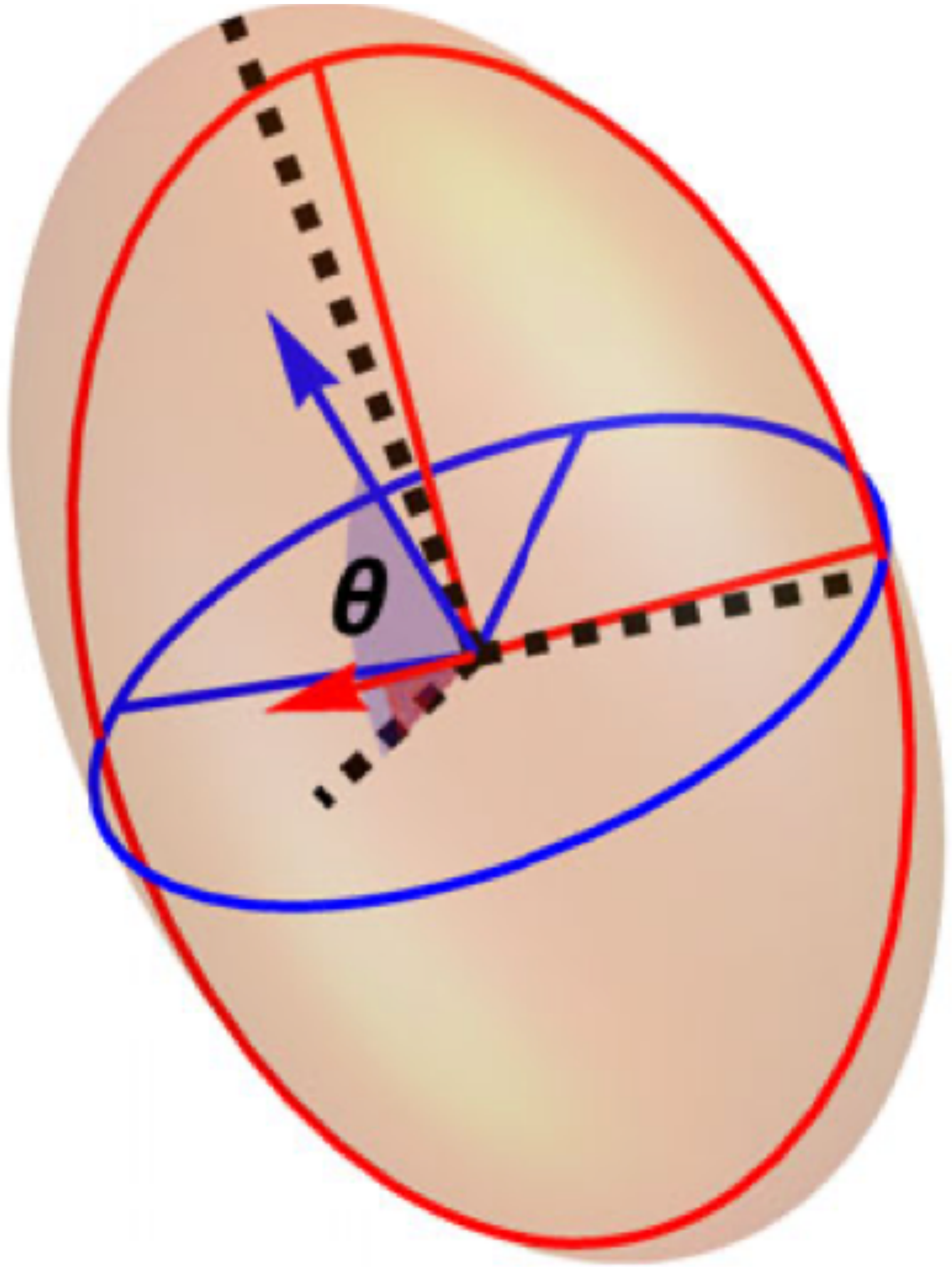
Likelihood cost in a Gaussian setting

$$LL_{\text{Gauss}} = \frac{1}{2}\sum_{\mu} (\lambda_\mu - 1 - \ln \lambda_\mu)$$

λ_μ eigenvalues of correlation matrix of data

Constraint: $\tilde{\mathbf{C}}^{\perp} = \mathbf{P}\tilde{\mathbf{C}}\mathbf{P}$

$$\underbrace{\lambda_1 \geq \lambda_1^{\perp}} \geq \underbrace{\lambda_2 \geq \lambda_2^{\perp}} \geq \ldots \geq \underbrace{\lambda_N \geq \lambda_N^{\perp}} = 0 \text{ (*)}$$



(*) Poincaré separation theorem

Likelihood cost in a Gaussian setting

$$LL_{\text{Gauss}} = \frac{1}{2}\sum_{\mu}(\lambda_\mu - 1 - \ln \lambda_\mu)$$

λ_μ eigenvalues of correlation matrix of data

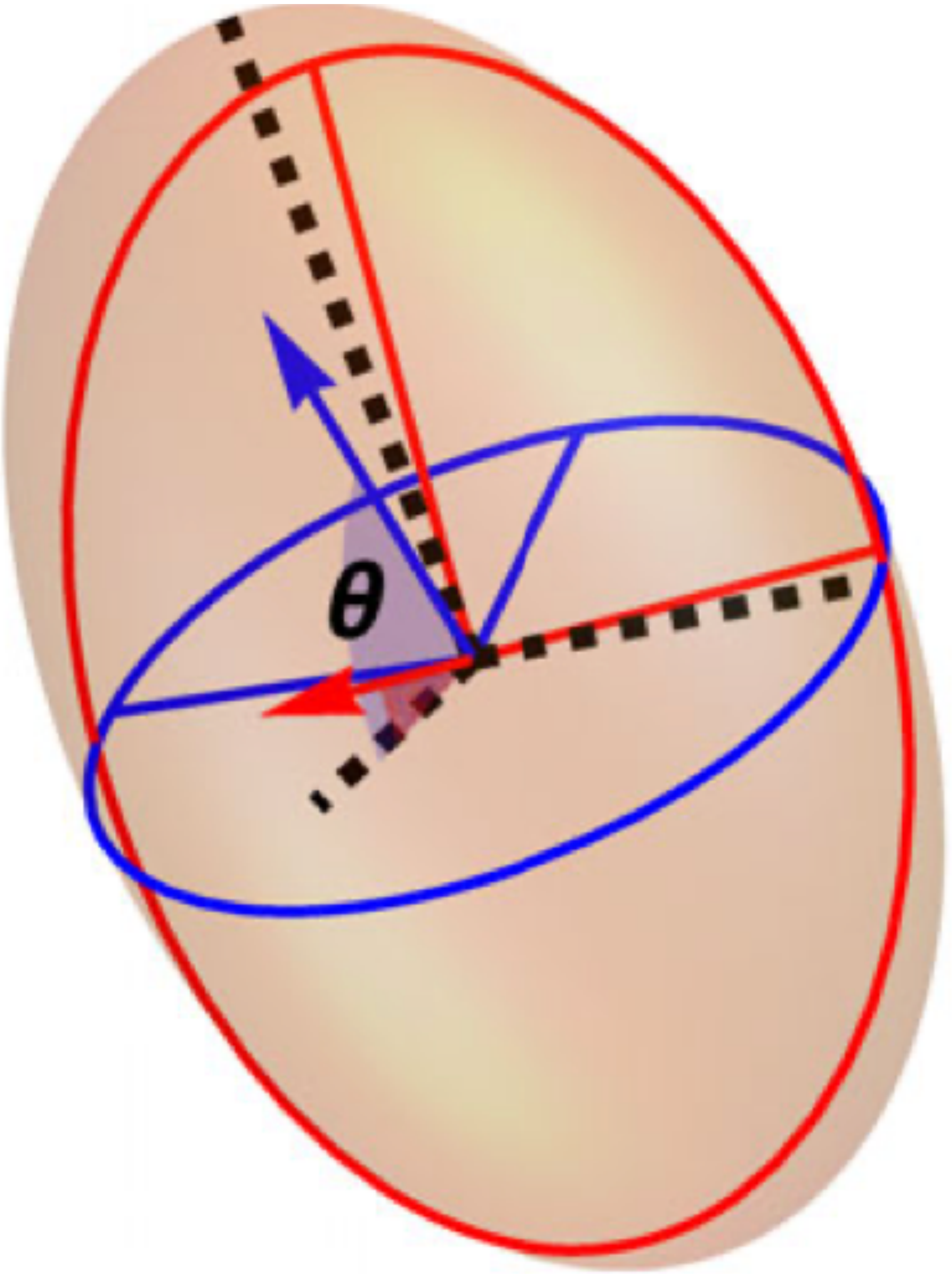
Constraint: $\tilde{\mathbf{C}}^{\perp} = \mathbf{P}\tilde{\mathbf{C}}\mathbf{P}$

$$\lambda_1 \geq \lambda_1^{\perp} \geq \lambda_2 \geq \lambda_2^{\perp} \geq \ldots \geq \lambda_N \geq \lambda_N^{\perp} = 0 \text{ (*)}$$

$\ldots$

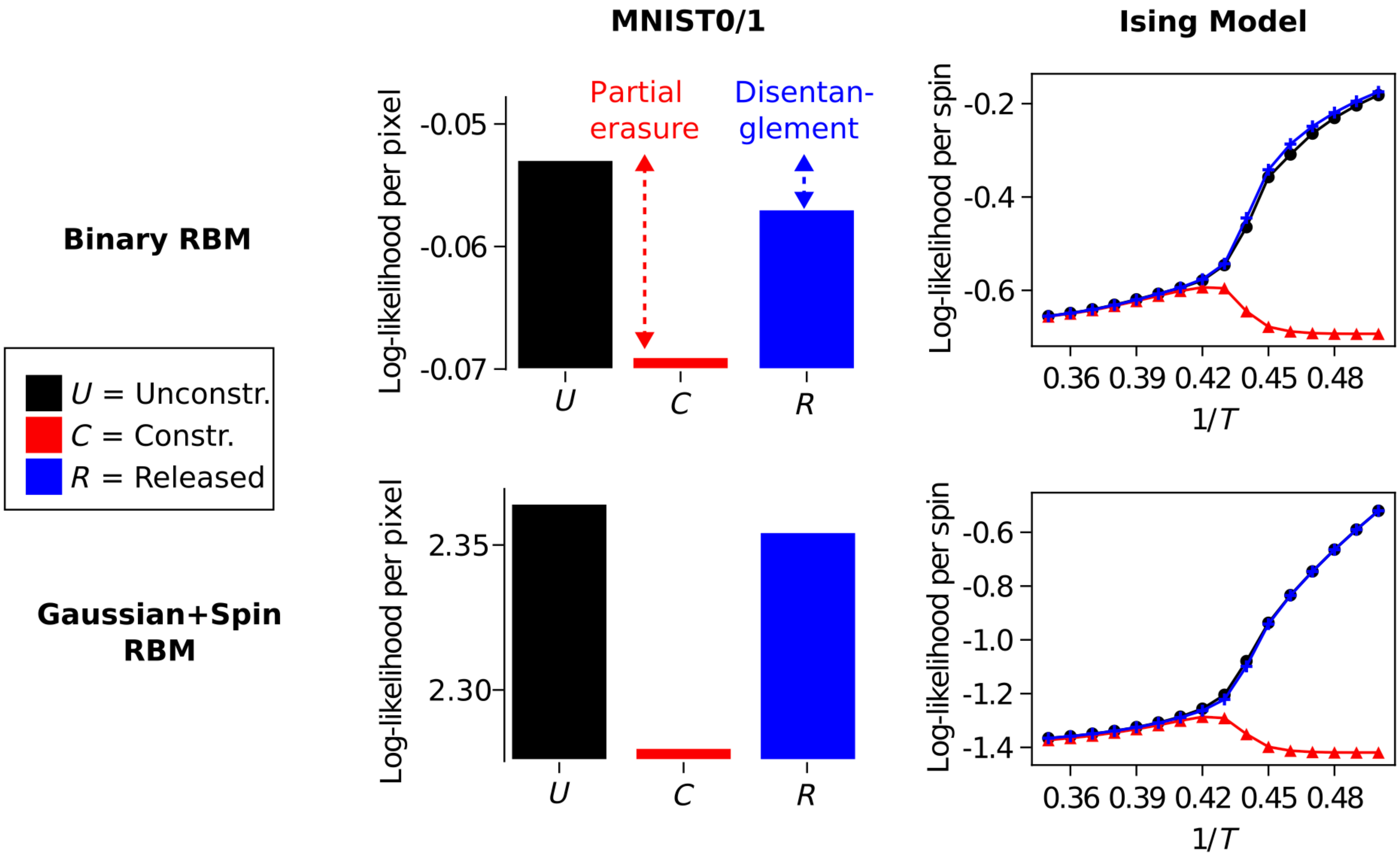
Spectral gaps account for log-likelihood cost of erasure / disentanglement

$$LL_{\text{constr.}} = LL_{\text{Gauss}} - \text{cost}$$



(*) Poincaré separation theorem

Likelihood cost in a Gaussian setting

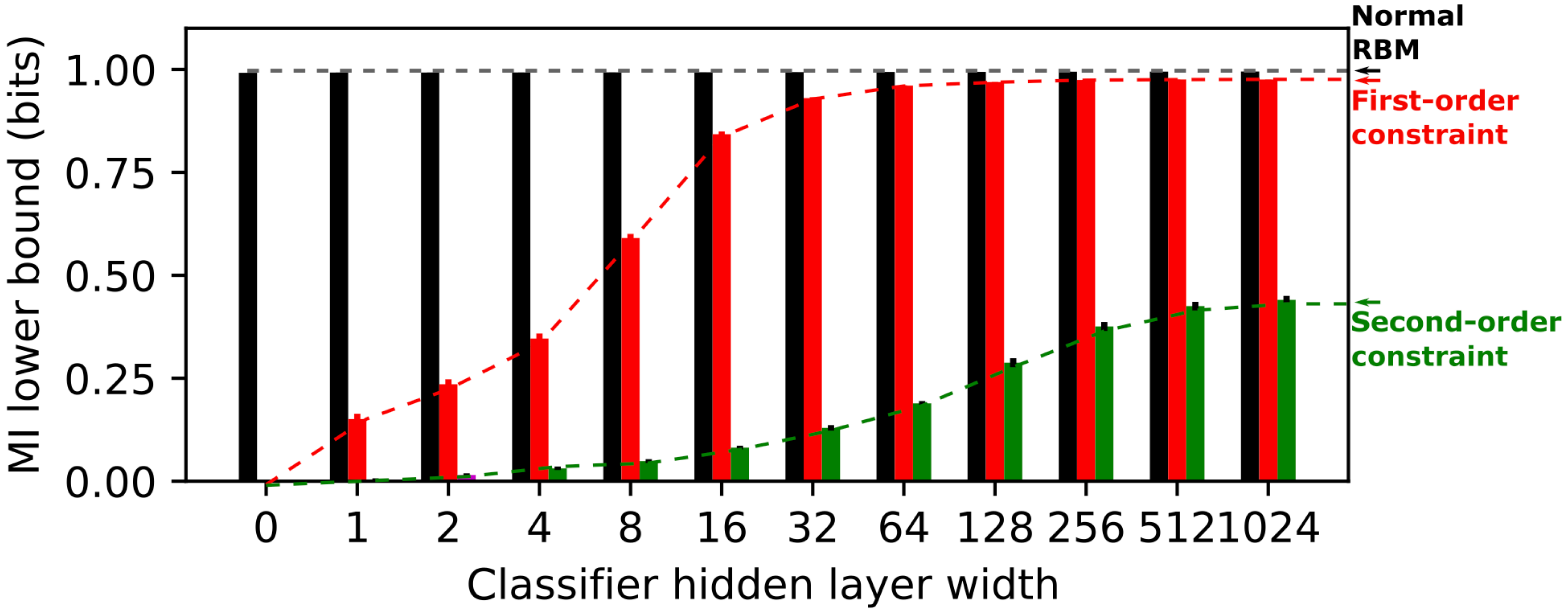


Approximate "erasure" of information

$$\text{MI}(I = \mathbb{W}^\top \mathbf{v}, \text{label}) = 0 \xleftrightarrow[?]{} \langle uI_\mu \rangle = 0, \langle uI_\mu I_\nu \rangle = 0, \langle uI_\mu I_\nu I_\gamma \rangle = 0, \ldots$$

Approximate "erasure" of information

$$\mathrm{MI}(I = \mathbb{W}^{\top}\mathbf{v}, \text{label}) = 0 \quad \overset{?}{\longleftrightarrow} \quad \langle u I_\mu \rangle = 0, \langle u I_\mu I_\nu \rangle = 0, \langle u I_\mu I_\nu I_\gamma \rangle = 0, \ldots$$



Summary:

- Semi-supervised approach
- Concentrate information about attribute on small subset of latent variables
- Transfer attributes from one data-point to another

Perspectives:

- Application to biological sequences. Transfer useful properties between natural sequences (specificity, stability, …). Ex.: WW, HSP.
- **… other?**

https://github.com/cossio/RestrictedBoltzmannMachines.jl

JFdCD, S.Cocco, R.Monasson, PRX 13, 021003 (2023)

Acknowledgements

Simona Cocco

Rémi Monasson

ENS, Paris

Thanks