

Physics inspired Machine Learning

Mykola Maksymenko, Research Lead at SoftServe R&D



Research Lead at SoftServe R&D

- PhD in Theoretical Physics
- Max-Planck Institute for the Physics of Complex Systems (Germany)
- Weizmann Institute of Science (Israel)
- Institute for Condensed Matter Physics of NASU

Physics of complex systems

**Quantum phases
of matter**

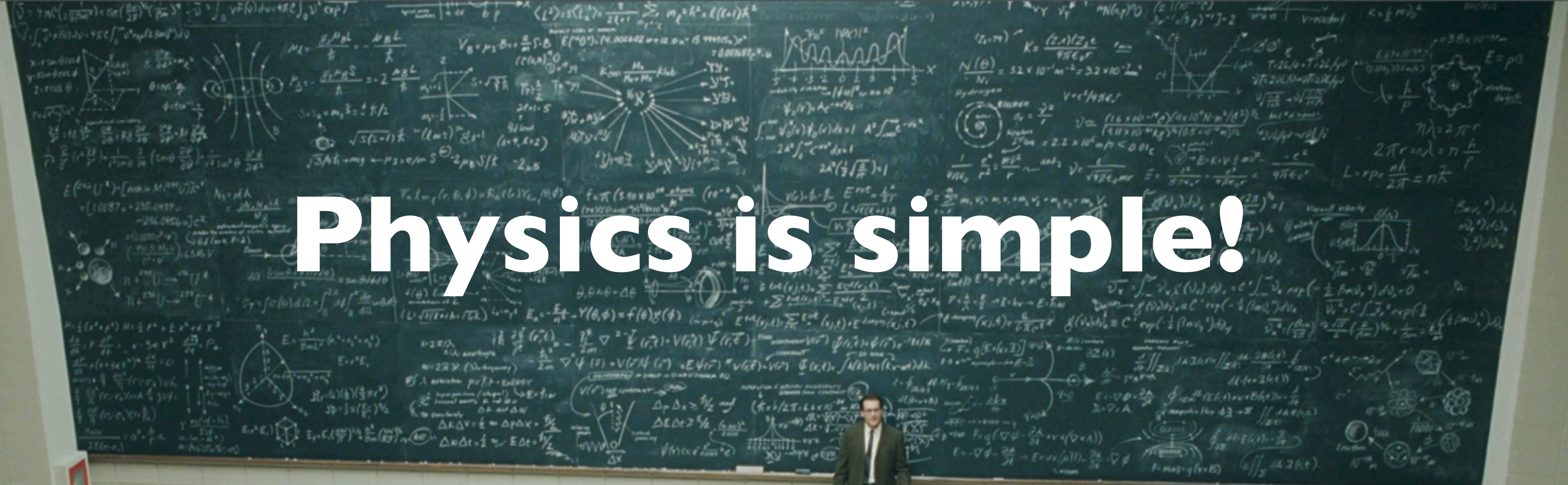
**Complex
Networks**

**Non-equilibrium
systems**

Exotic magnetism



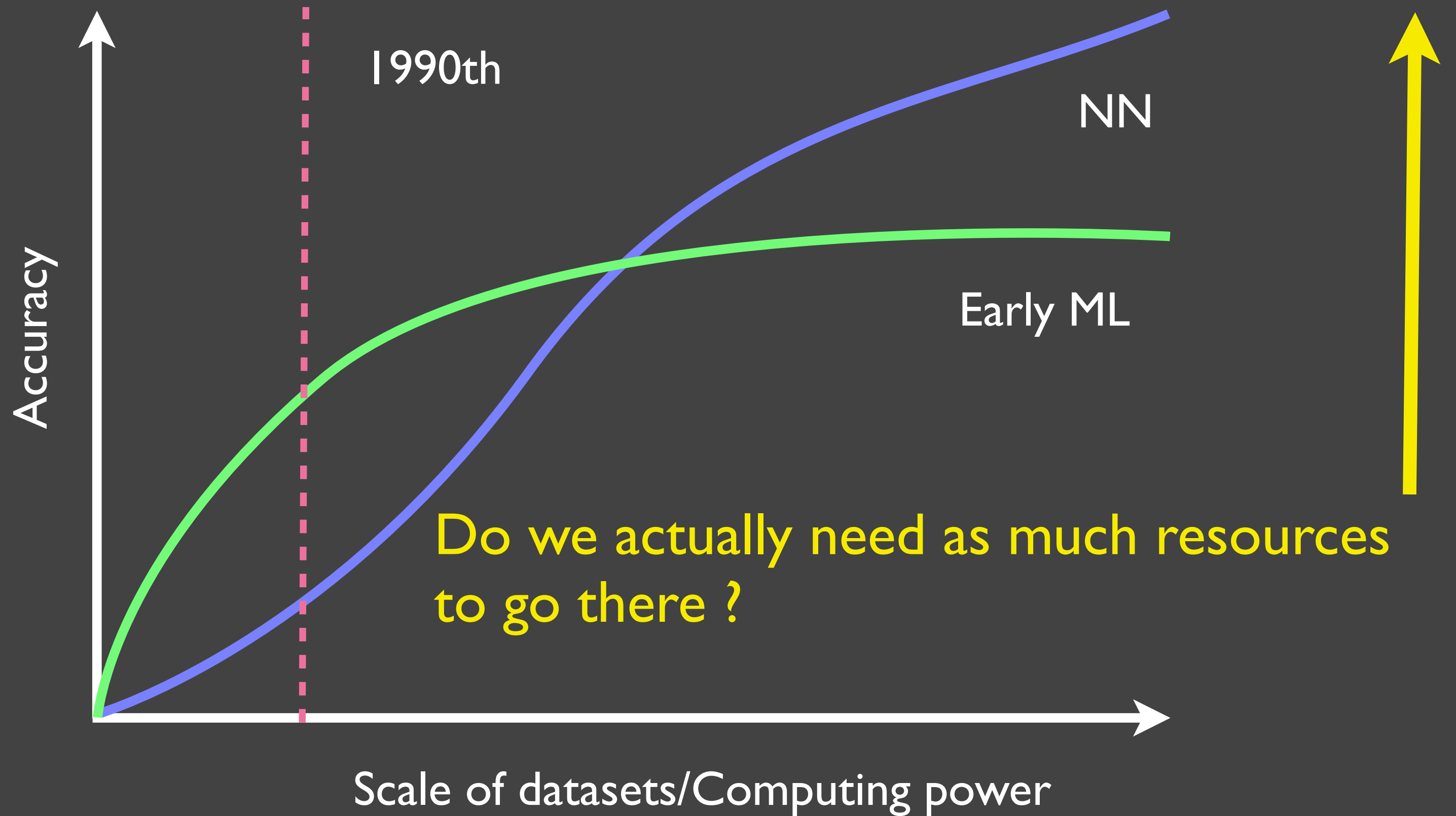
Physics is simple!



Building the Deep Learning architectures



A naive picture of a recent progress



Resources are expensive



Andrej Karpathy ✓

@karpathy

Читати



TitanX runs at $\sim 200\text{W}$ (0.72MJ/h). In $\sim 12\text{h}$ that's $\sim 8\text{MJ}$. Energy content of wood is 20MJ/kg , so running 1 TitanX overnight burns 1 pound of wood

11:00 - 13 жовт. 2015

The goals:

Optimal models

Easier training

Universality

Outline

Physics of Learning

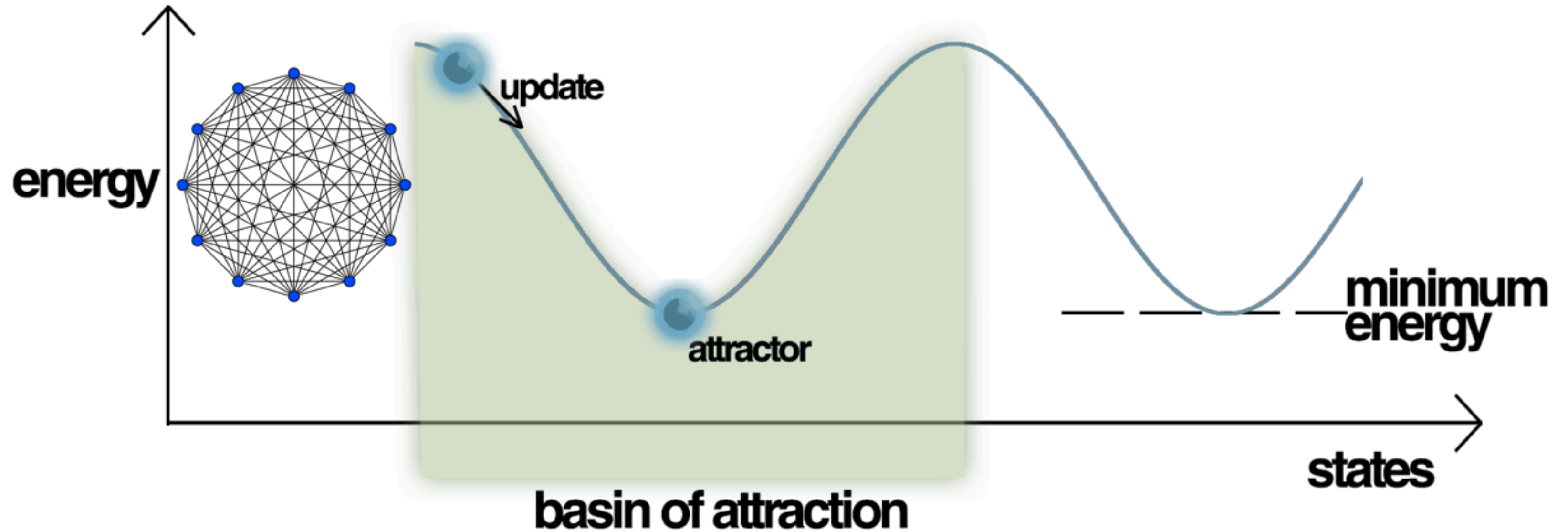
Redundancy of Neural Nets are redundant ?

How physics helps for better learning ?

Neural nets and quantum wave functions

Learning the Tensor networks

Hopfield network J. Hopfield PNAS 1982



Fully recurrent

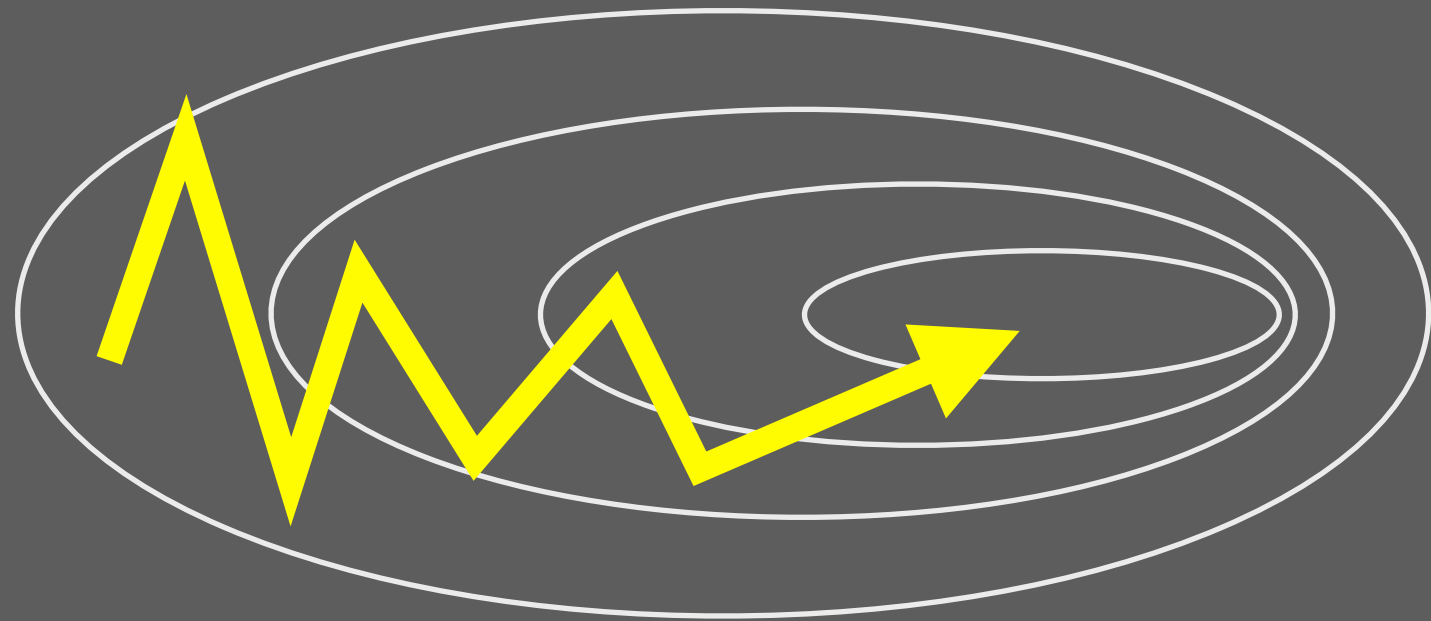
Associative memory as valleys in energy space

Statistical Physics of learning from examples

Seung et. al., PRA 1992

Gradient descent as a Langevin equation

$$\frac{\partial \mathbf{W}}{\partial t} = -\nabla_{\mathbf{W}} E(\mathbf{W}) + \eta(t)$$



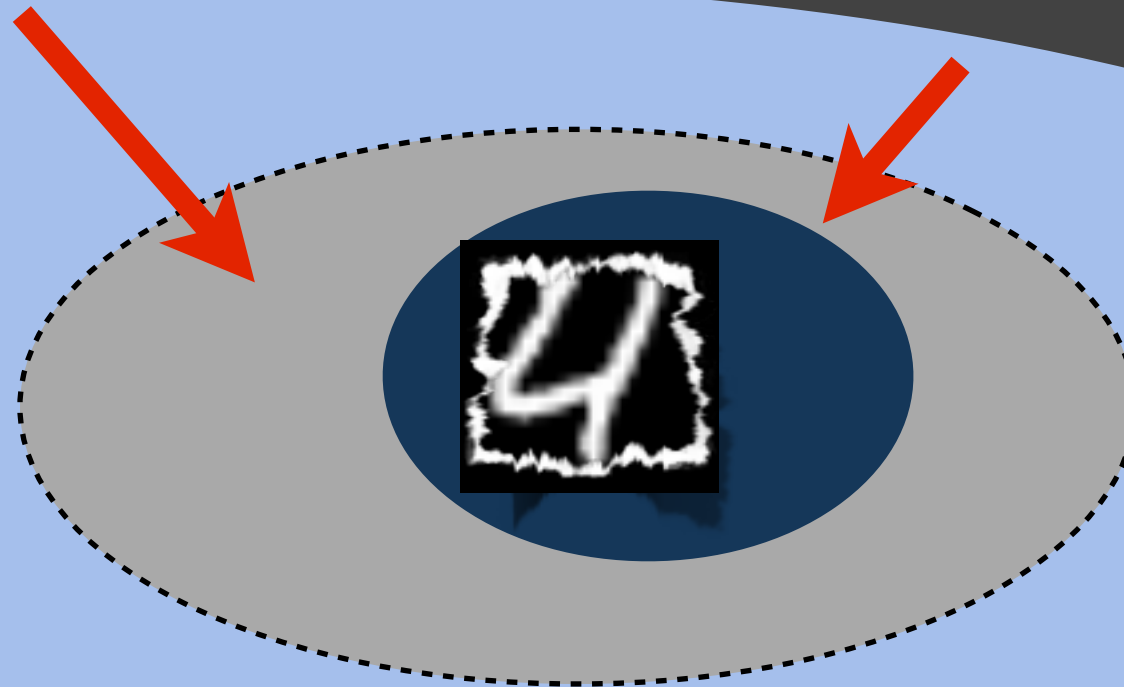
$$P(W) = \frac{e^{-\beta E(W)}}{\mathcal{Z}}$$

in the long time limit

Curse of dimensionality in ML

Tunable NN capacity

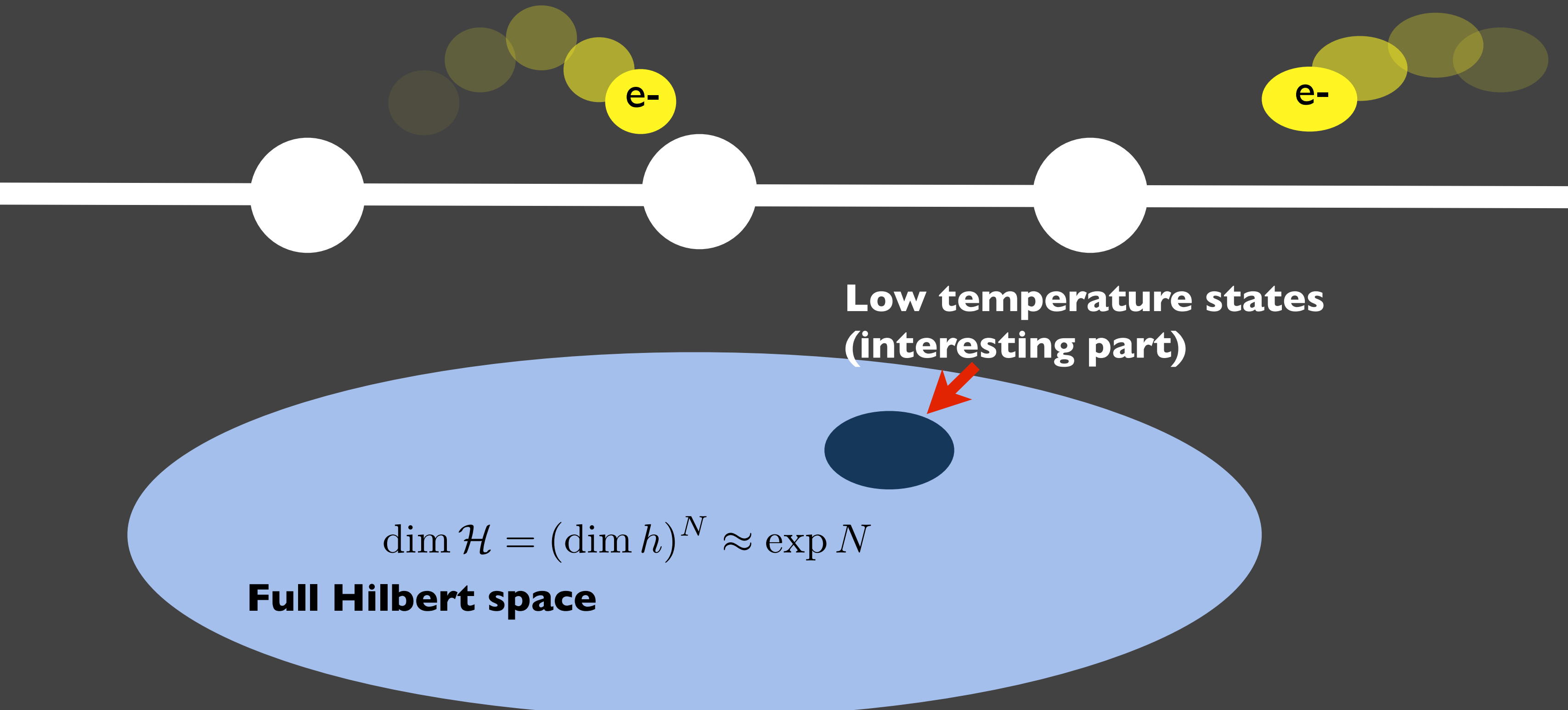
**MNIST pictures
space**



Full space of pixel states

$$2^{(28 \times 28)}$$

Grows of Quantum State space





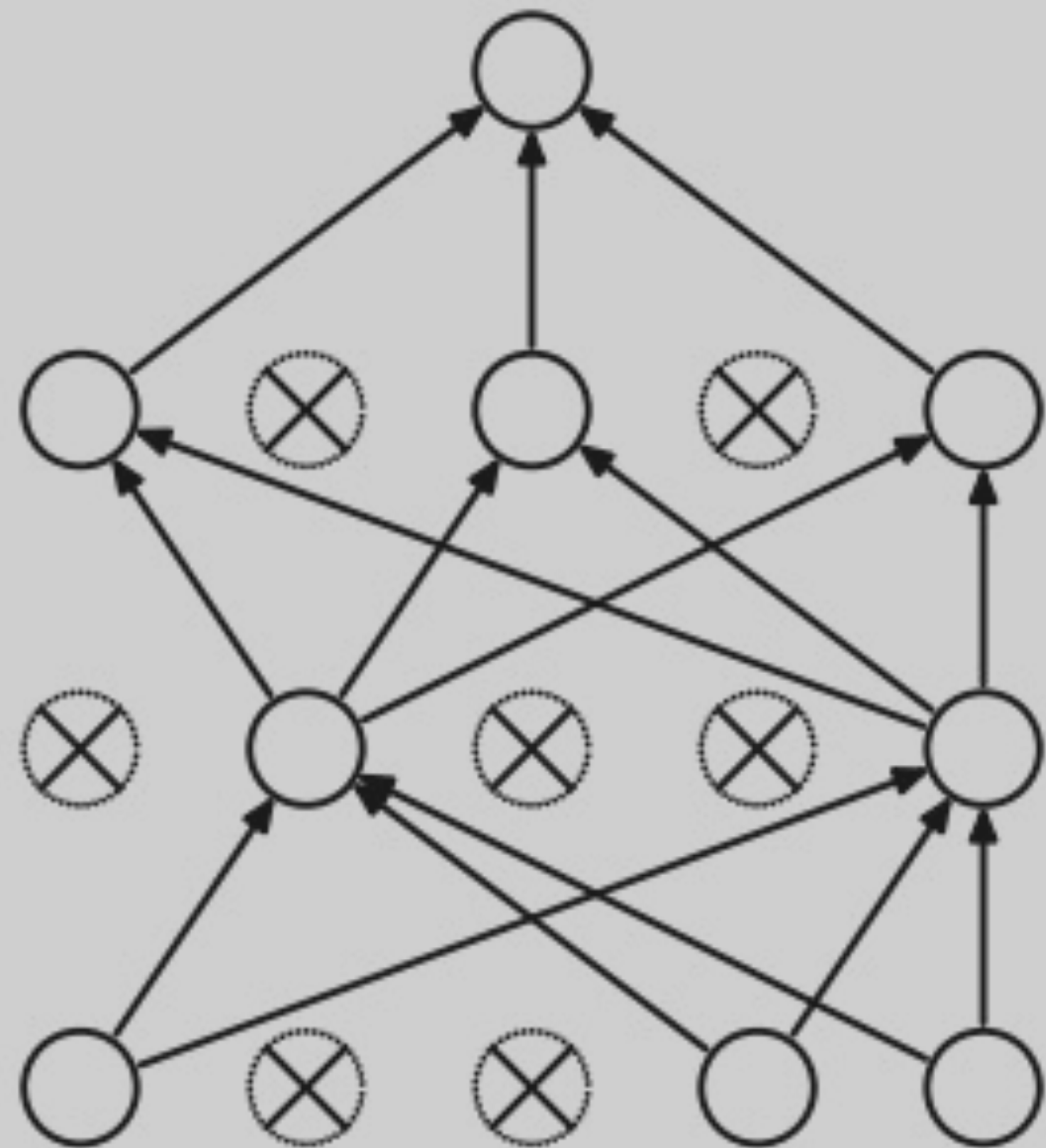
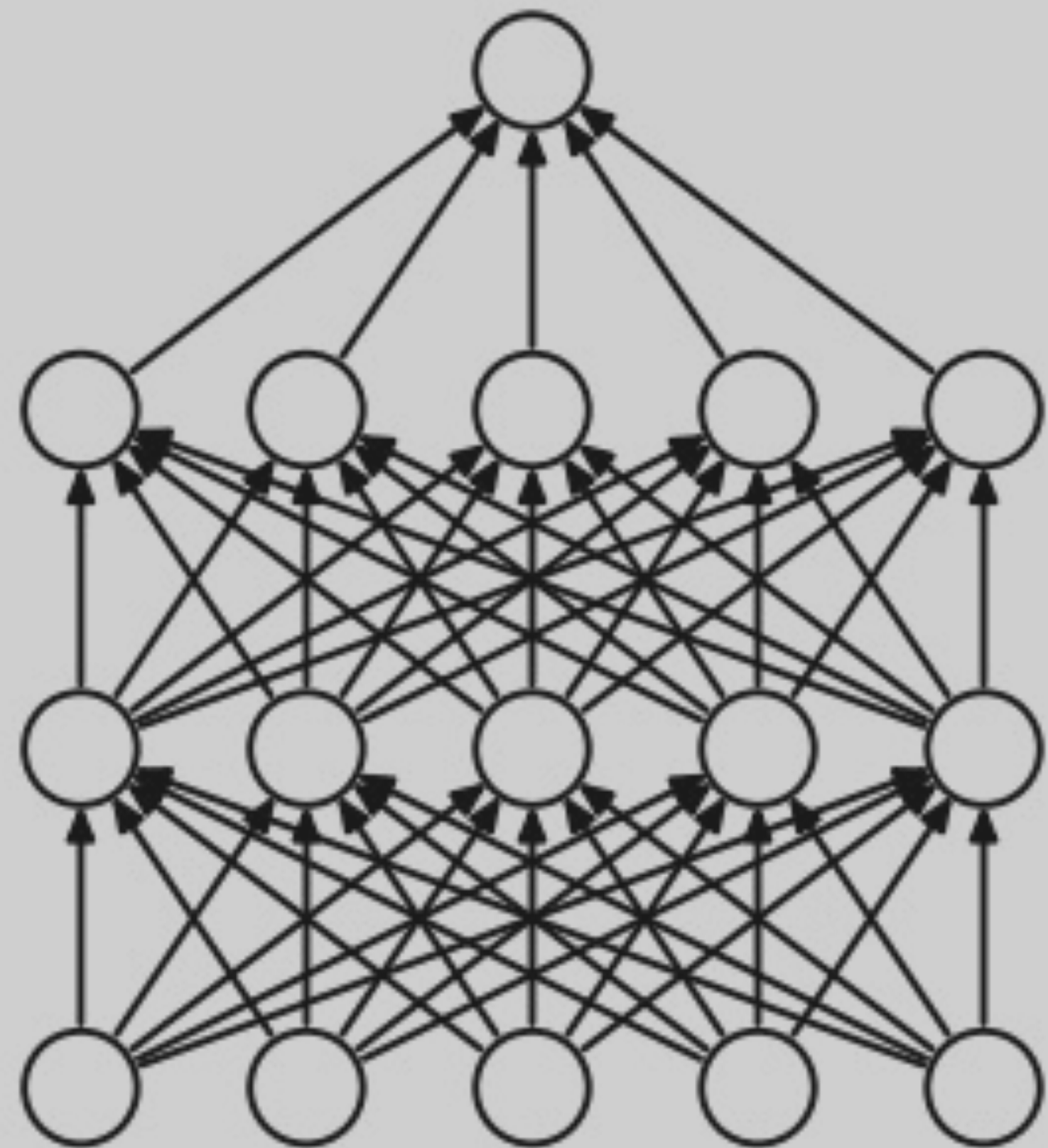
**Fundamental
Research**

Industry

Redundancy of Neural Nets

NN overfit - need Regularized training

Srivastava et.al. JMLR 2014



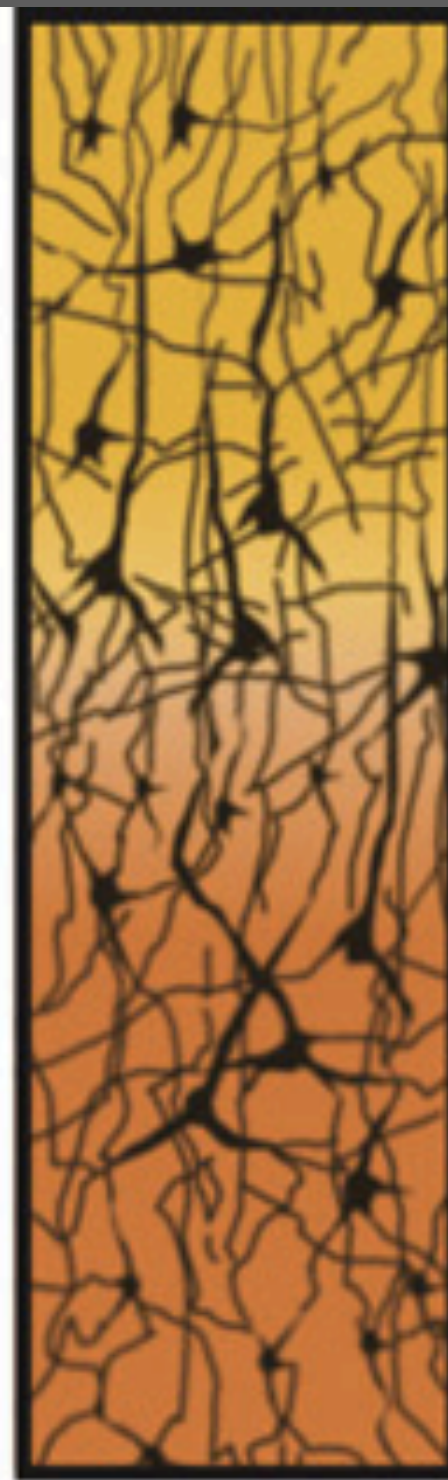
Compression with pruning



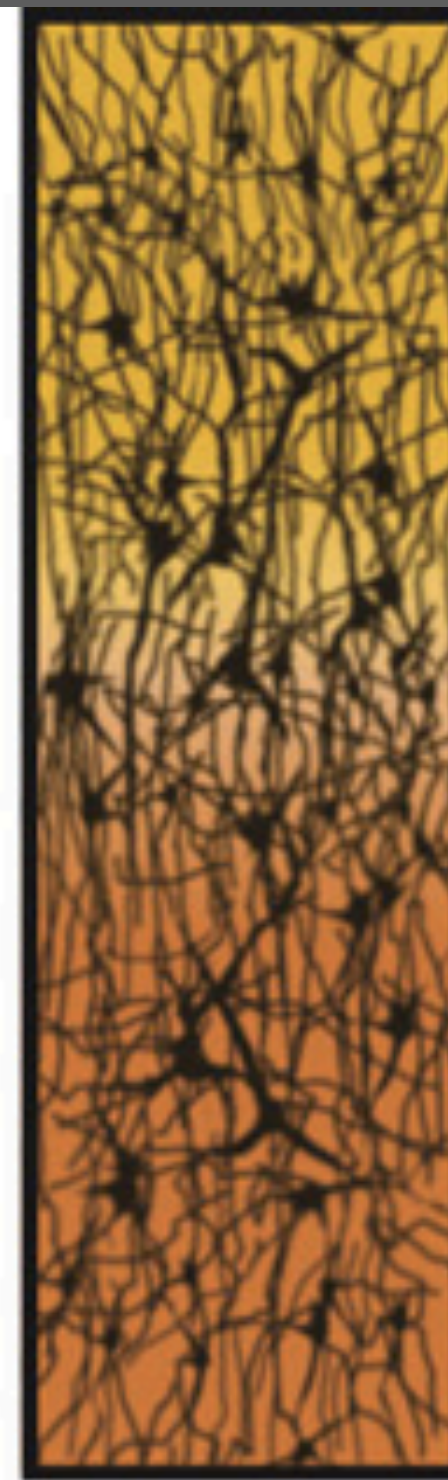
Newborn



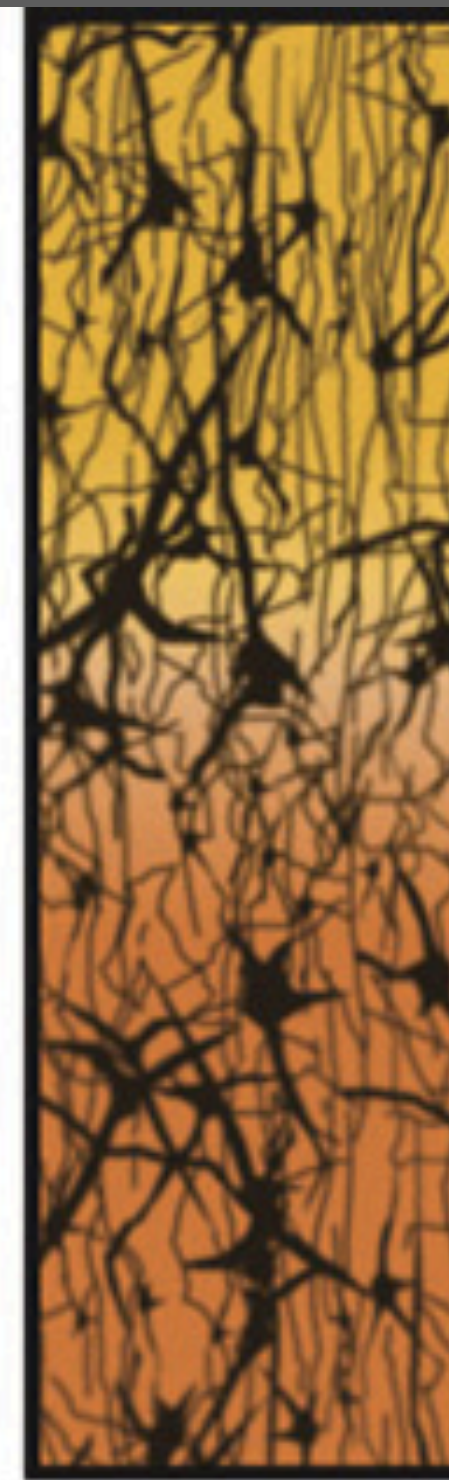
1 Month



9 Months



2 Years

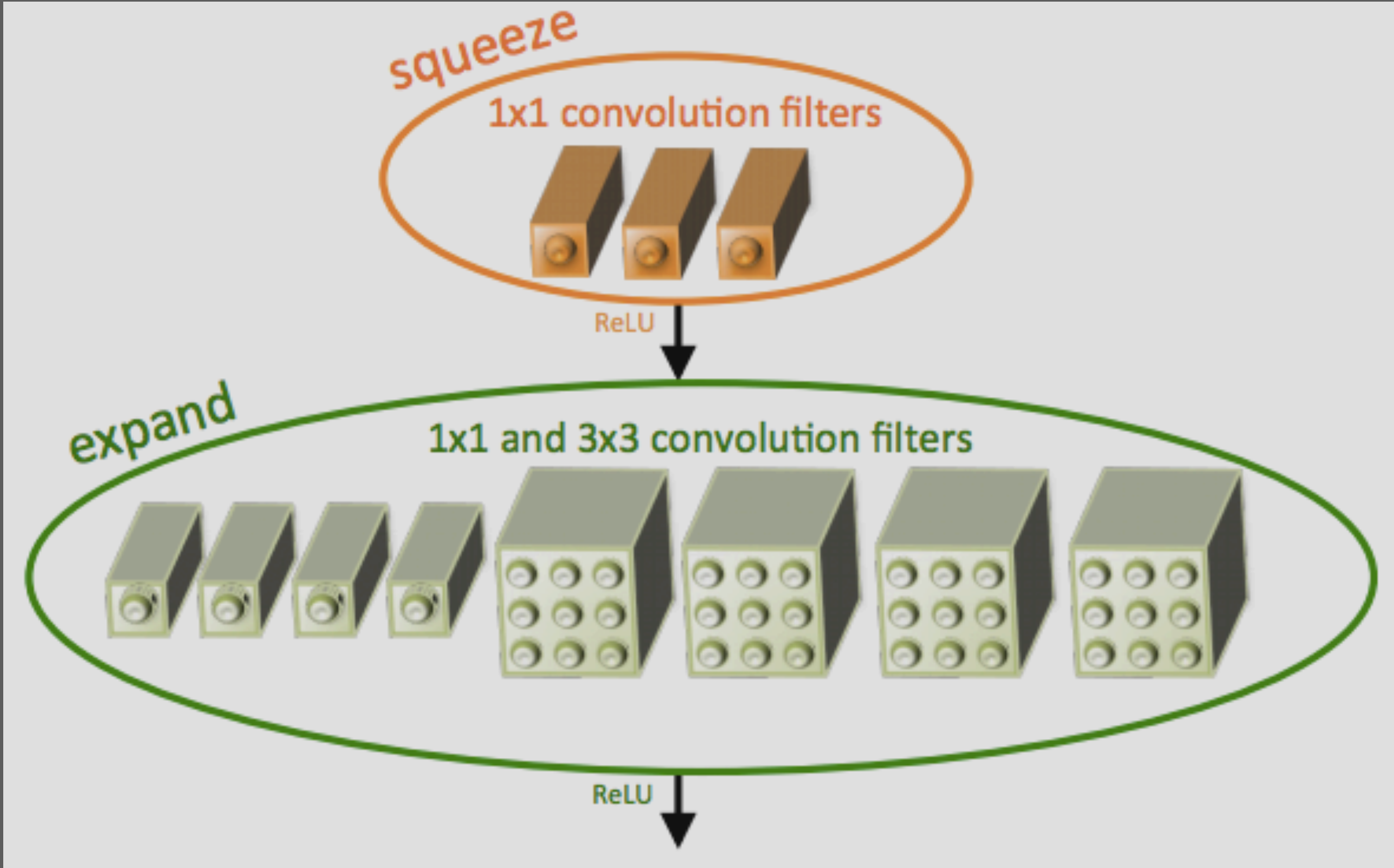


Adult

Compression with pruning

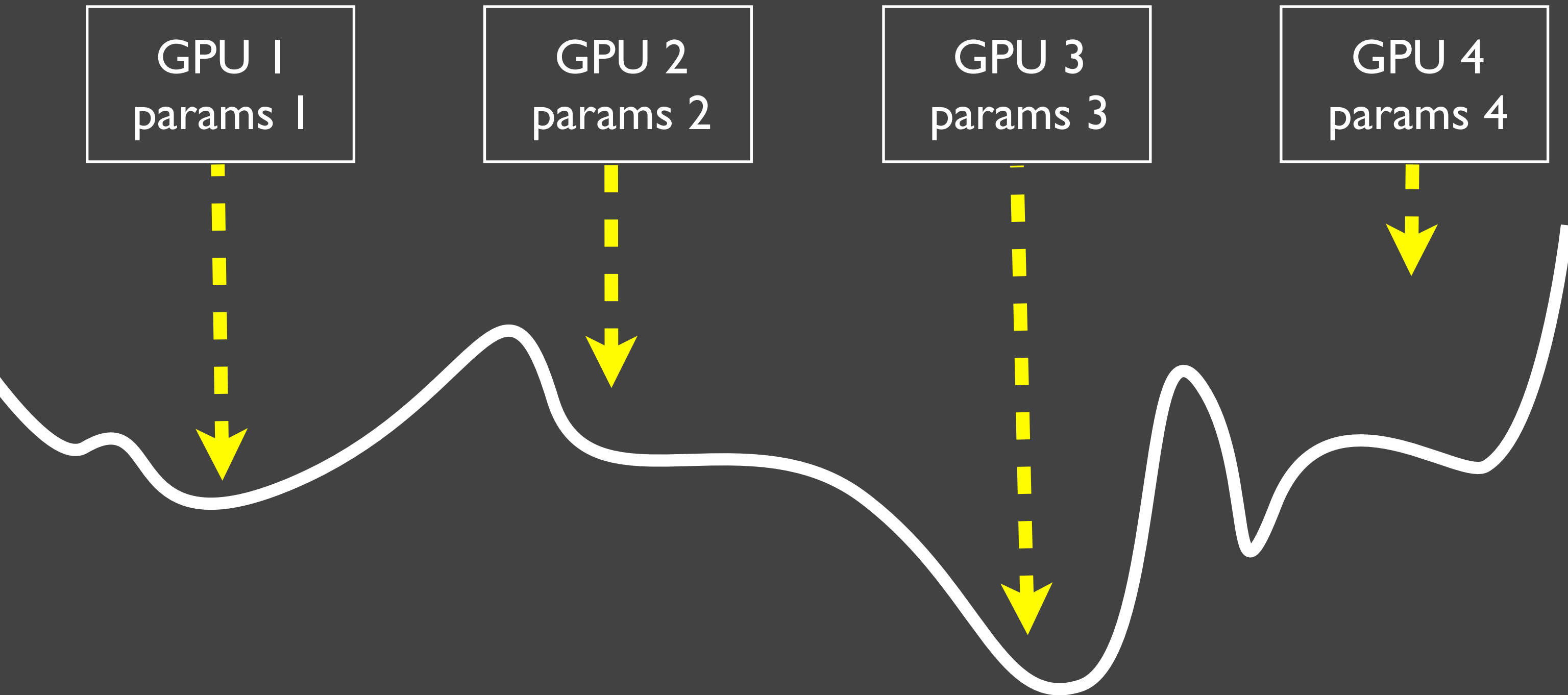
Network	Top-1 Error	Top-5 Error	Parameters	Compression Rate
LeNet-300-100 Ref	1.64%	-	267K	
LeNet-300-100 Pruned	1.59%	-	22K	12×
LeNet-5 Ref	0.80%	-	431K	
LeNet-5 Pruned	0.77%	-	36K	12×
AlexNet Ref	42.78%	19.73%	61M	
AlexNet Pruned	42.77%	19.67%	6.7M	9×
VGG16 Ref	31.50%	11.32%	138M	
VGG16 Pruned	31.34%	10.88%	10.3M	13×

Better architectures pop up regularly

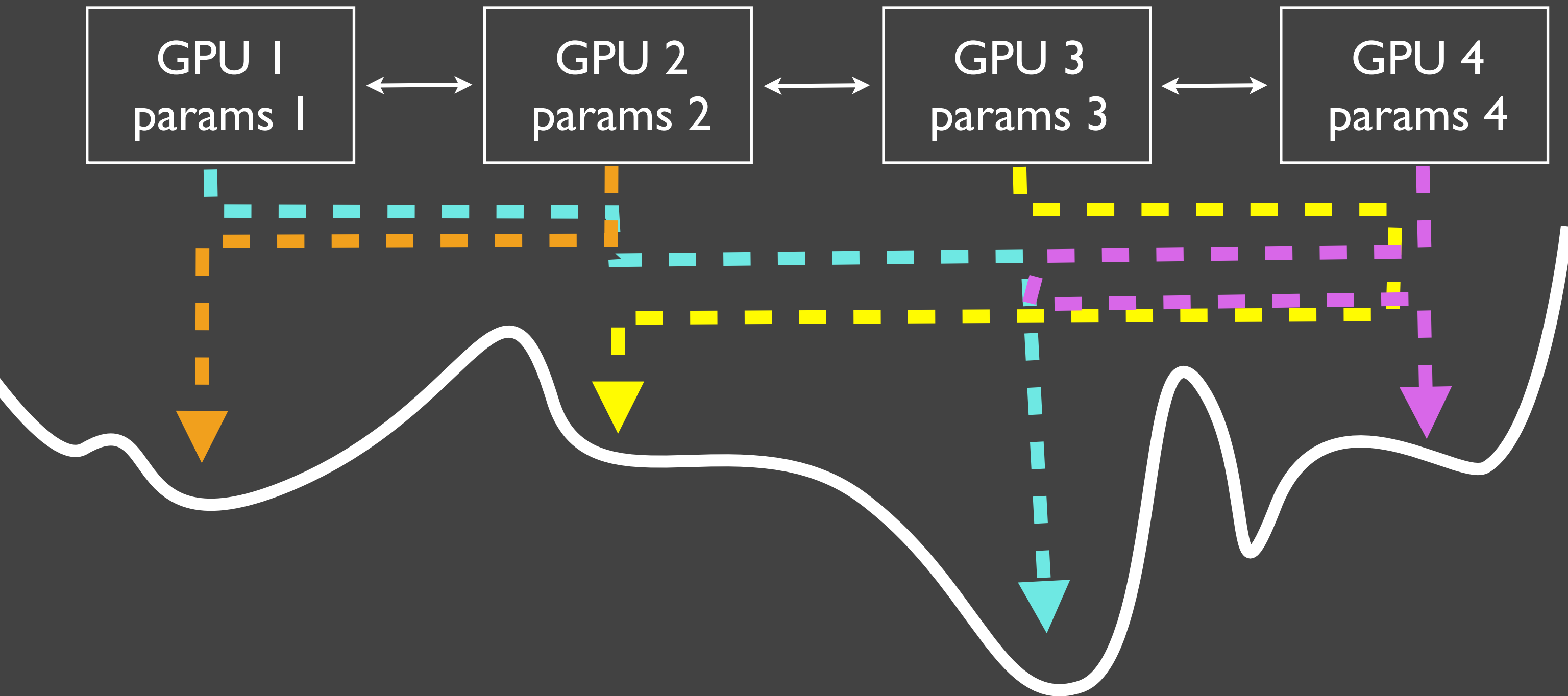


Physics for better learning

Typical search for good parameters

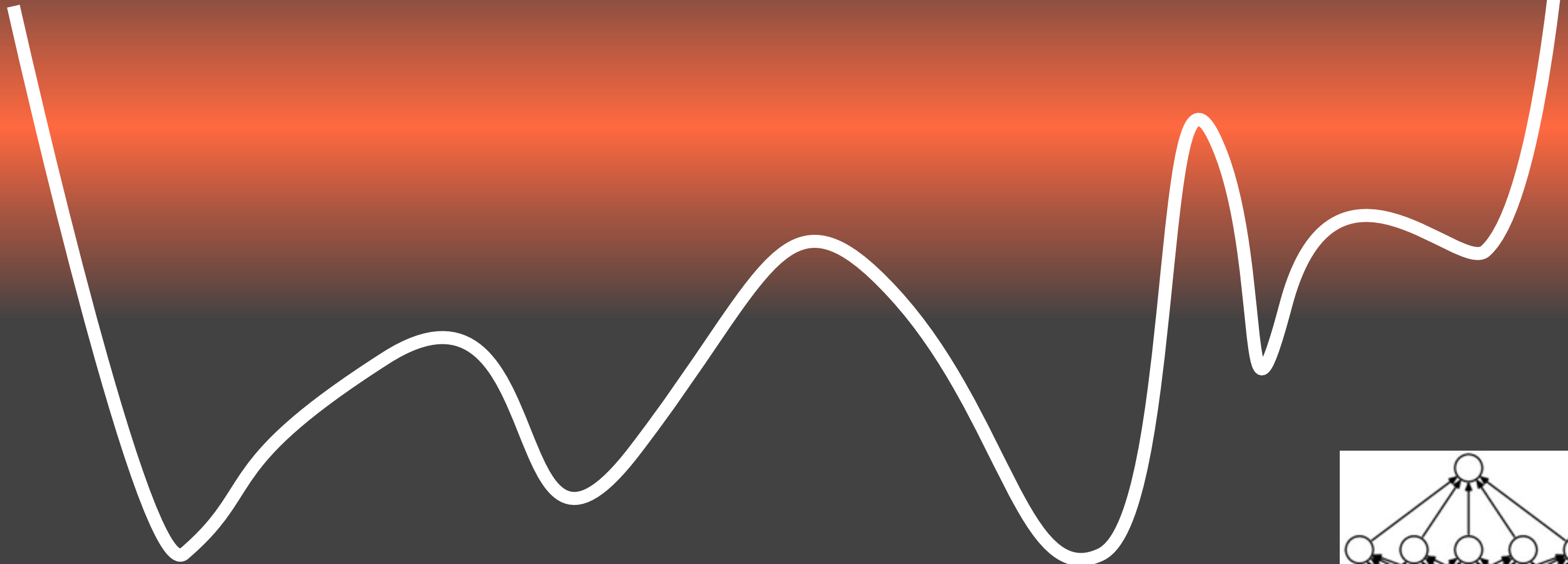


Physicist search for good parameters



Dropout as a temperature

Cost function

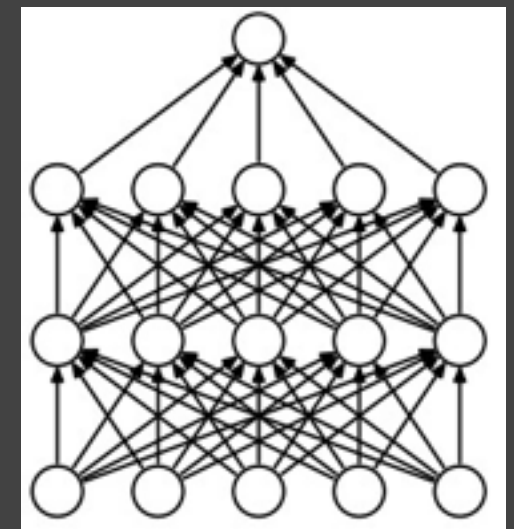


Disconnected
neurons

$p=1$



Also typical picture in spin-glass.



$p=0$

Dropout as a temperature

Statistical Physics of learning from examples

Seung et. al., PRA 1992

$$\langle\langle E(\mathbf{W}) \rangle\rangle = P\epsilon(\mathbf{W})$$

$P = \alpha N$ number of examples scales with respect to network size

$$P_0(\mathbf{W}) = \frac{\exp[-N\beta\alpha\epsilon(\mathbf{W})]}{Z} \quad \text{in high-temperature limit}$$

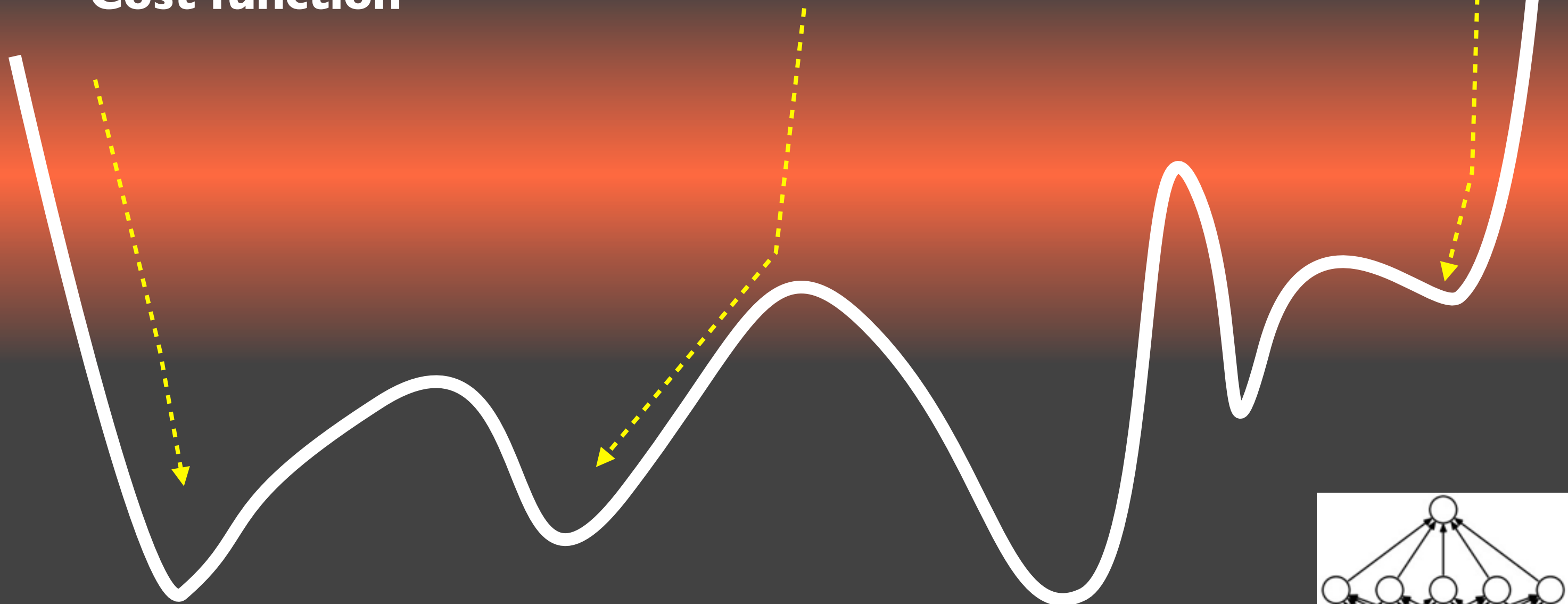
with effective temperature T/α

Annealed Dropout

Rennie et.al. IEEE 2014 (IBM group)

Dropout as a temperature

Cost function

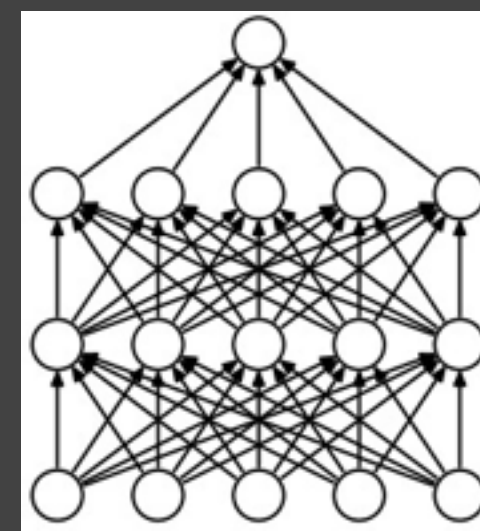


Disconnected
neurons

$p=1$

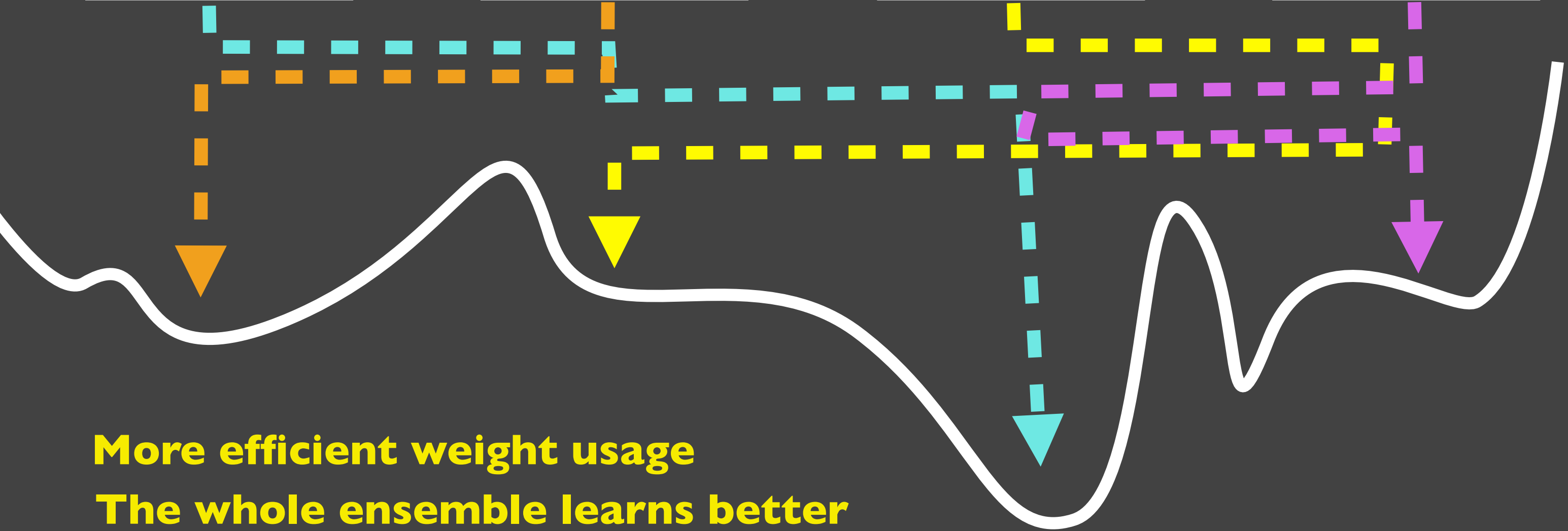
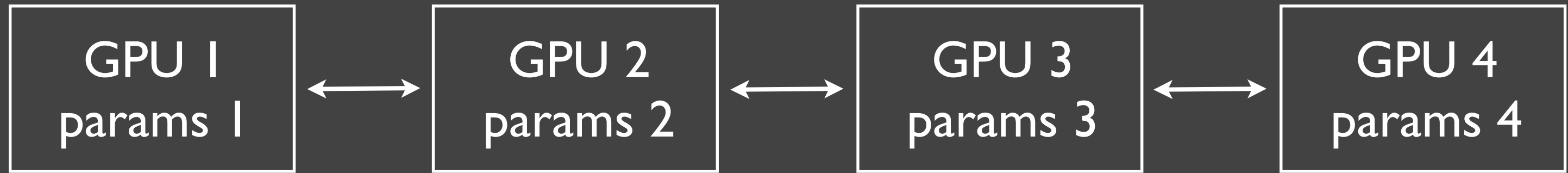
Annealed Dropout

Rennie et.al. IEEE 2014 (IBM group)



$p=0$

Learning with parallel tempering

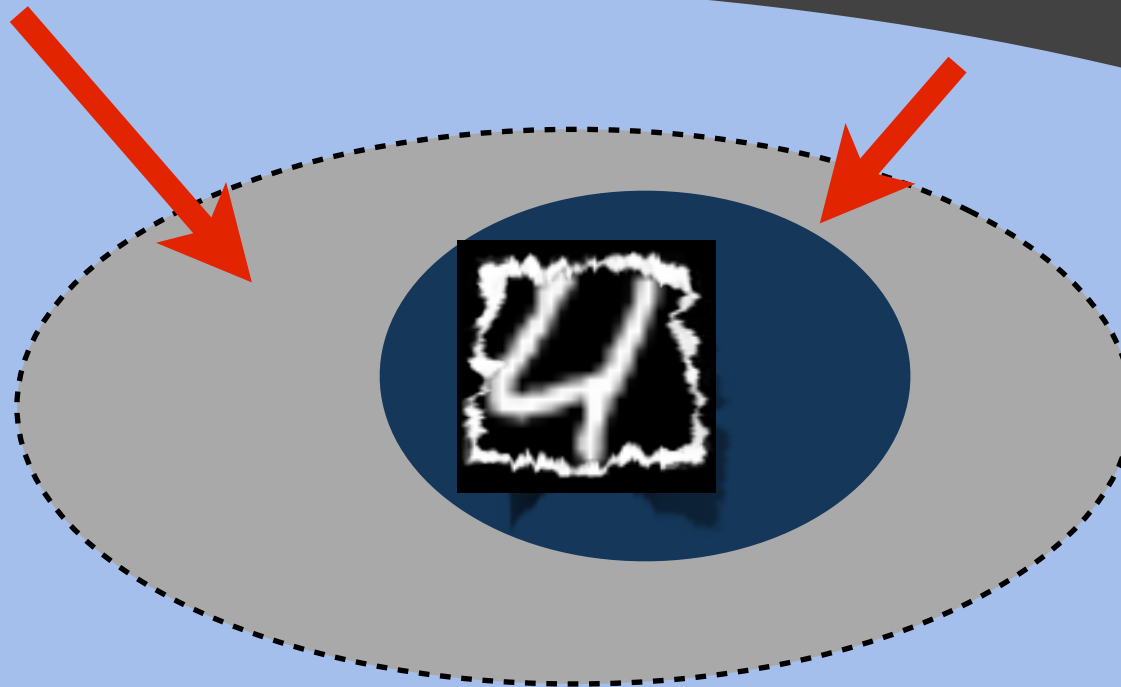


**Neural Nets
and
Quantum wave functions**

Curse of dimensionality in ML

Tunable NN capacity

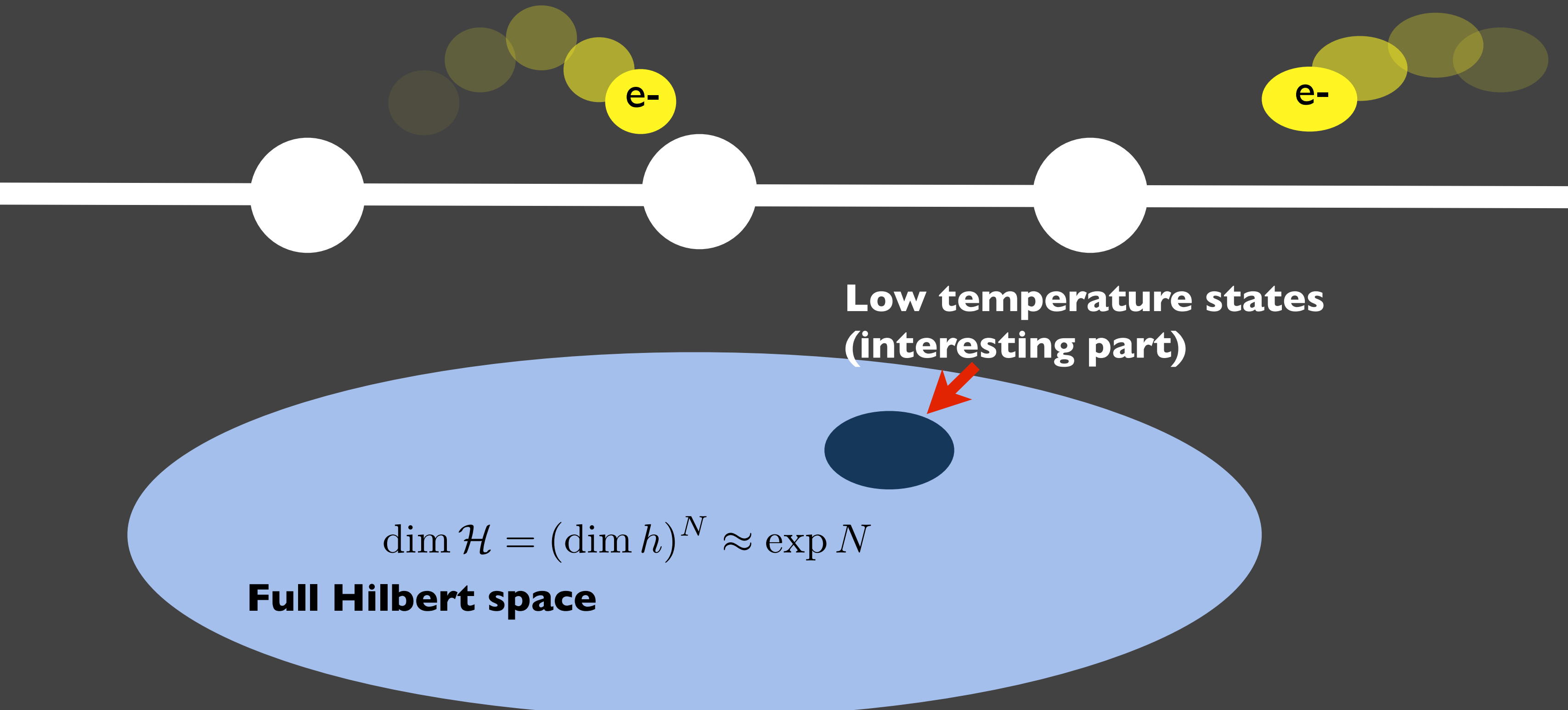
MNIST pictures space



Full space of pixel states

$$2^{(28 \times 28)}$$

Grows of Quantum State space



What is the quantum wave function ?



What is the quantum wave function ?

$$|\phi\rangle = \sum_{j_1=\uparrow,\downarrow\dots j_N=\uparrow,\downarrow} c_{j_1,\dots,j_N} |j_1\rangle \dots |j_N\rangle$$

Large tensor of probabilities for each configuration

$$|\phi(\uparrow\downarrow\uparrow)\rangle = 0.1$$

Graphical notations for tensors



constant



vector v_i



matrix M_{ij}



matrix product $\sum_j M_{ij} M_{jk}$

Representing Quantum states with tensor networks

$$c_{j_1 \dots j_N}$$

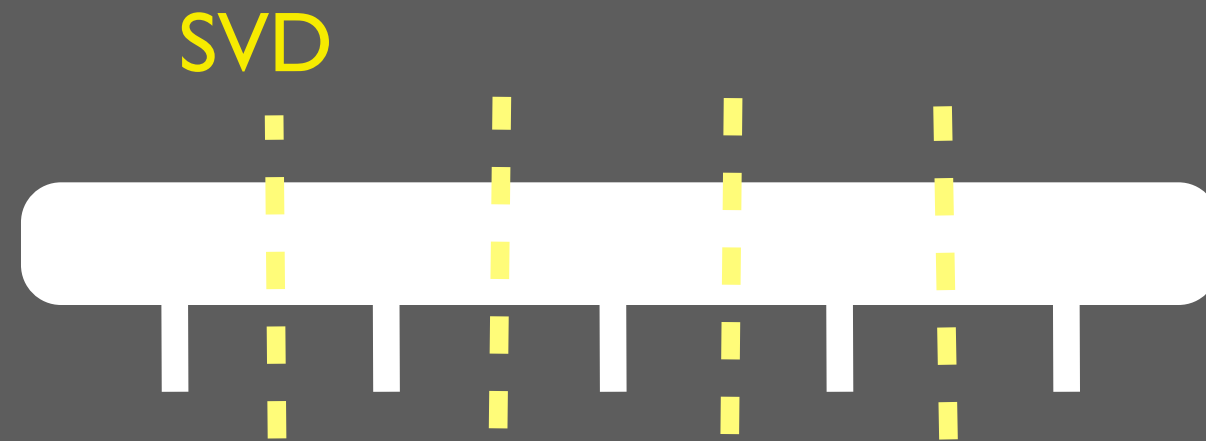


Large tensor of probabilities for each configuration

$$|\phi(\uparrow\downarrow\uparrow)\rangle = 0.1$$

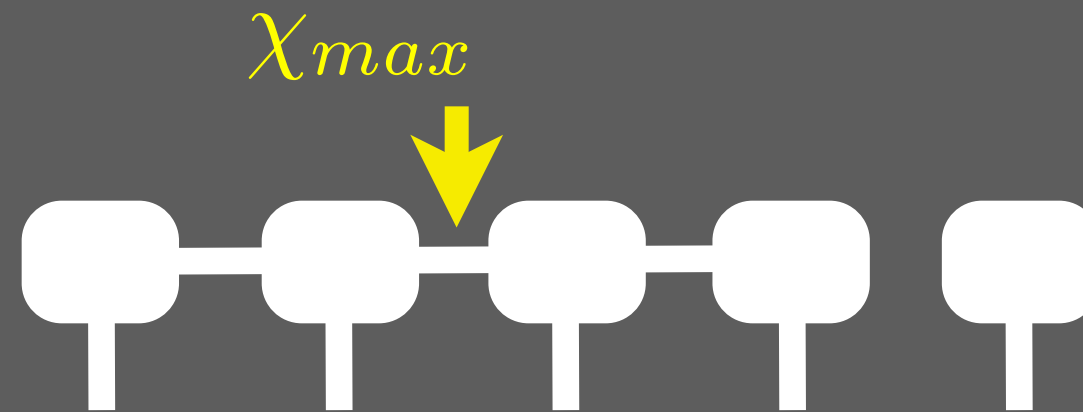
Representing Quantum states with tensor networks

$$C_{j_1 \dots j_N}$$



\approx

$$A_{j_1, \alpha, \beta}^{[1]} A_{j_2, \alpha, \beta}^{[2]} \dots A_{j_N, \alpha, \beta}^{[N]}$$

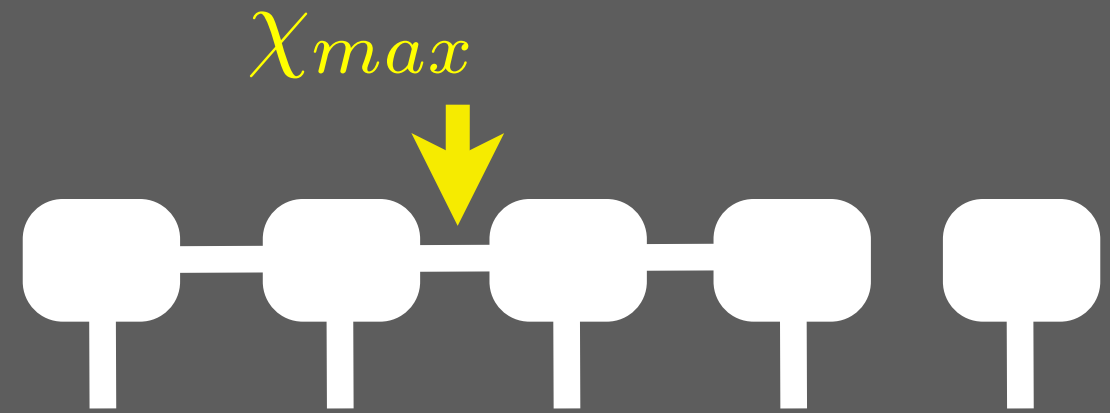


Representing Quantum states with tensor networks



$$\exp(N)$$

vs

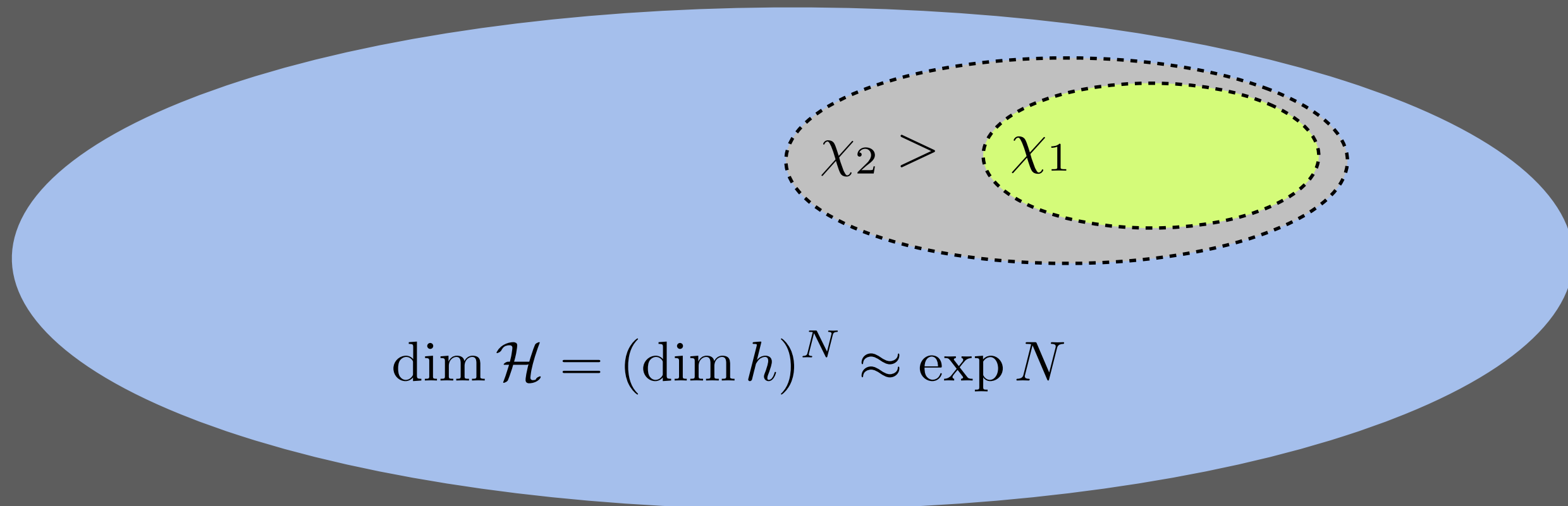


$$N d \chi_{max}^2$$

Strong reduction of complexity

Representing Quantum states with tensor networks

$$\exp(N) \quad \text{vs} \quad Nd\chi_{max}^2$$



~~Quantum~~ Machine Learning with Tensor Networks

Stoudenmire et. al., NIPS 29, 4799 (2016)

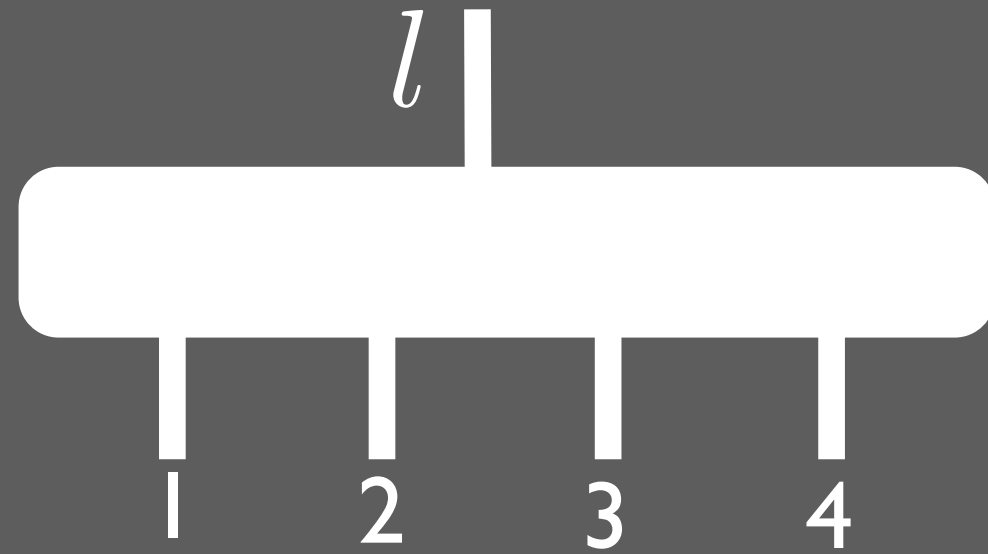
Han et.al., ArXiv:1709.01662 (2017)

ML with Tensor Networks

$$f^l(\mathbf{x}) = W^l \cdot \Phi(\mathbf{x})$$

ML with Tensor Networks

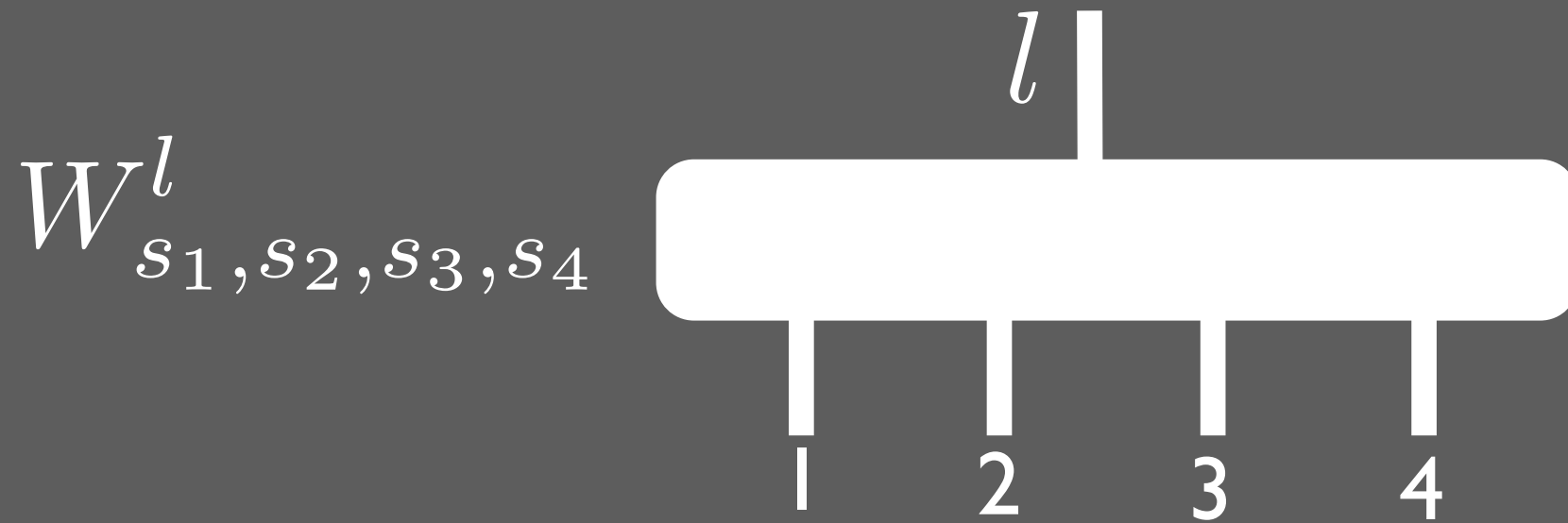
$$f^l(\mathbf{x}) = W^l \cdot \Phi(\mathbf{x})$$



1	2
3	4

Tensorizing weights and data

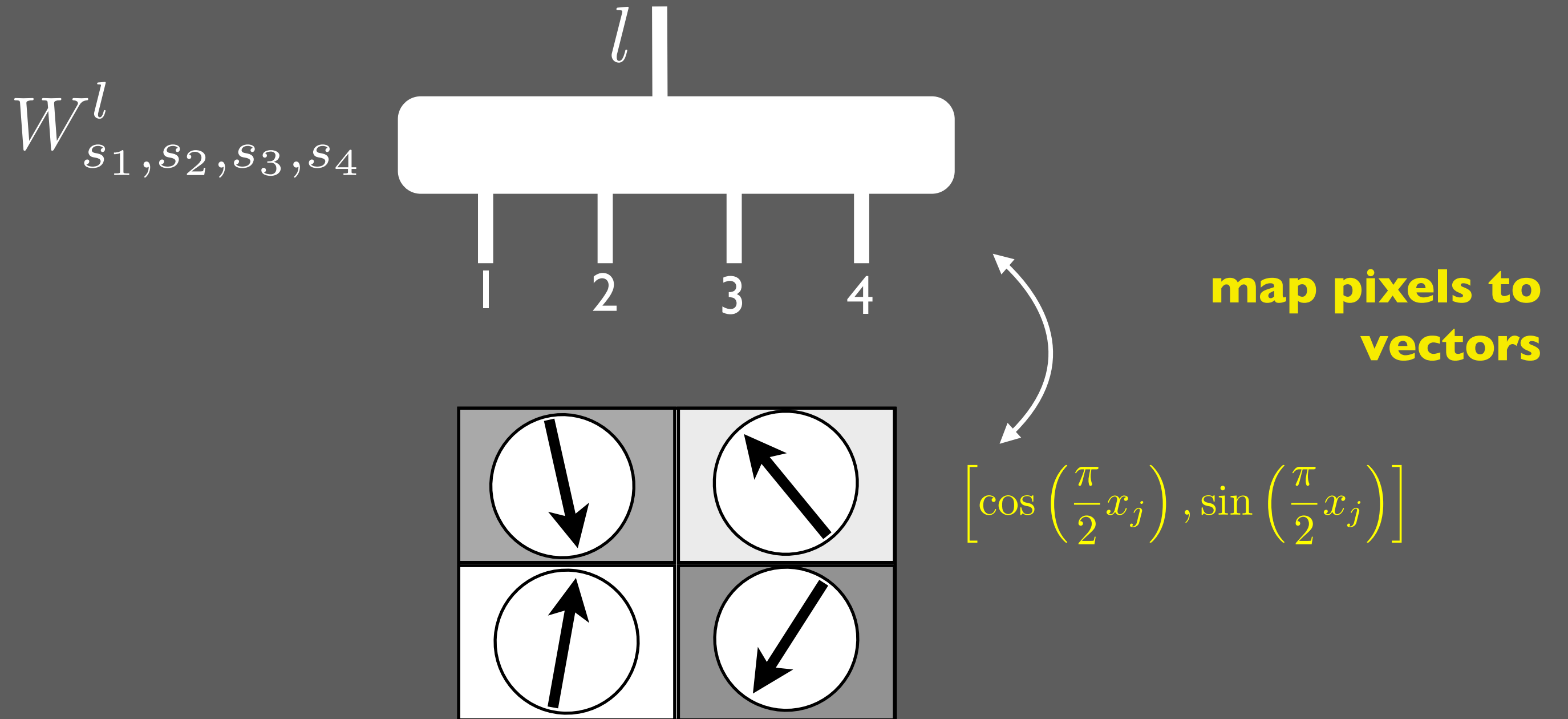
$$f^l(\mathbf{x}) = W^l \cdot \Phi(\mathbf{x})$$



1	2
3	4

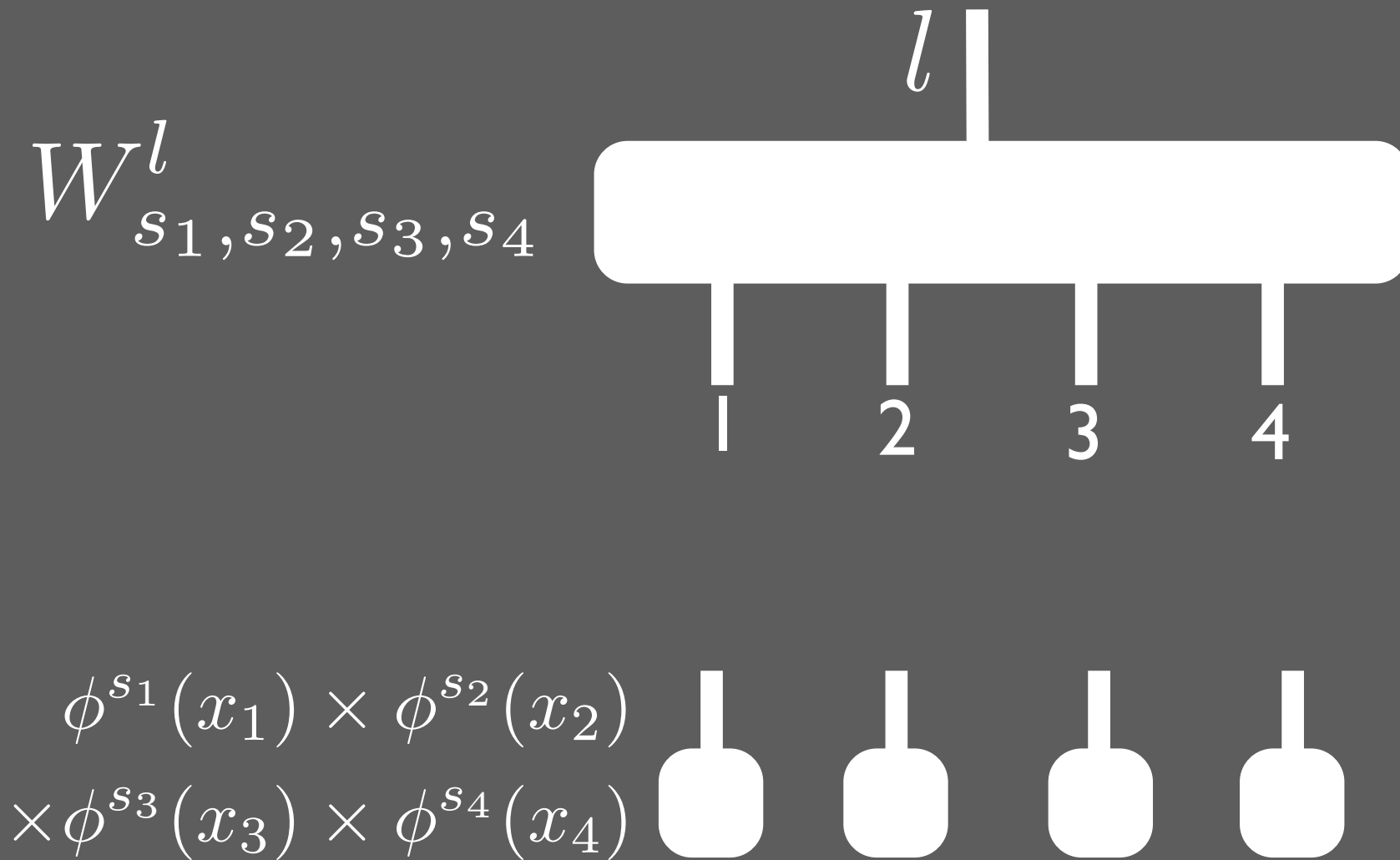
Tensorizing weights and data

$$f^l(\mathbf{x}) = W^l \cdot \Phi(\mathbf{x})$$



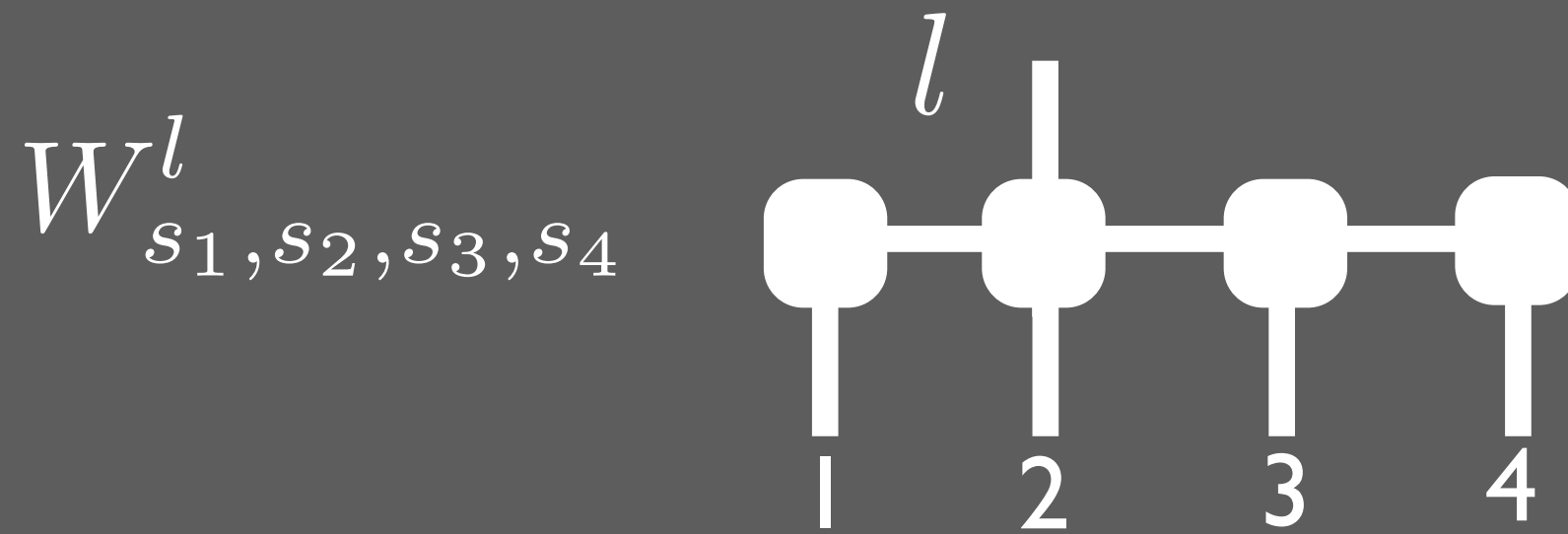
Tensorizing weights and data

$$f^l(\mathbf{x}) = W^l \cdot \Phi(\mathbf{x})$$

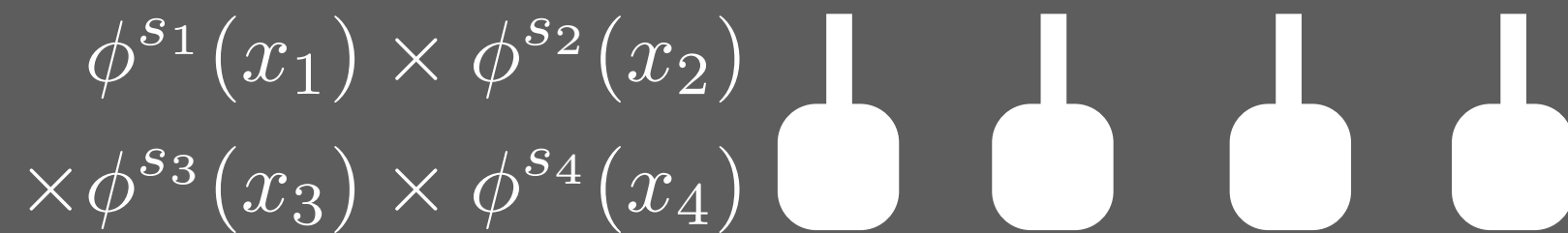


Tensorizing weights and data

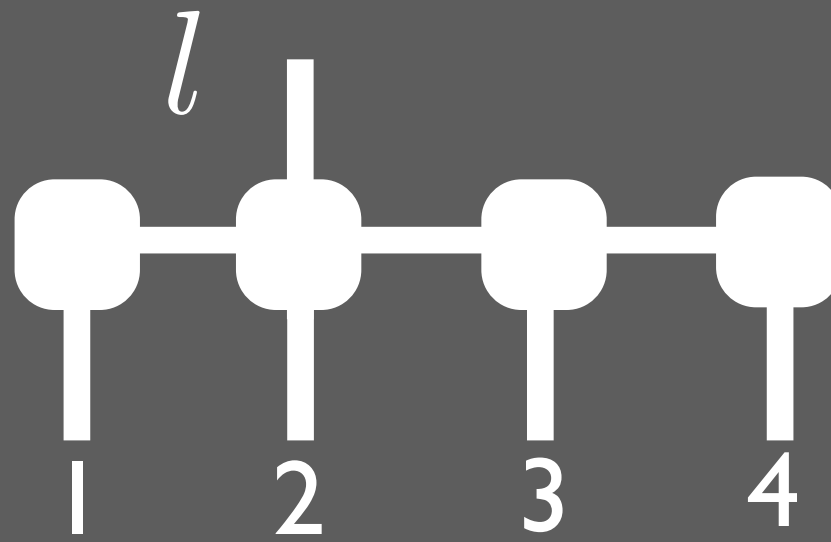
$$f^l(\mathbf{x}) = W^l \cdot \Phi(\mathbf{x})$$



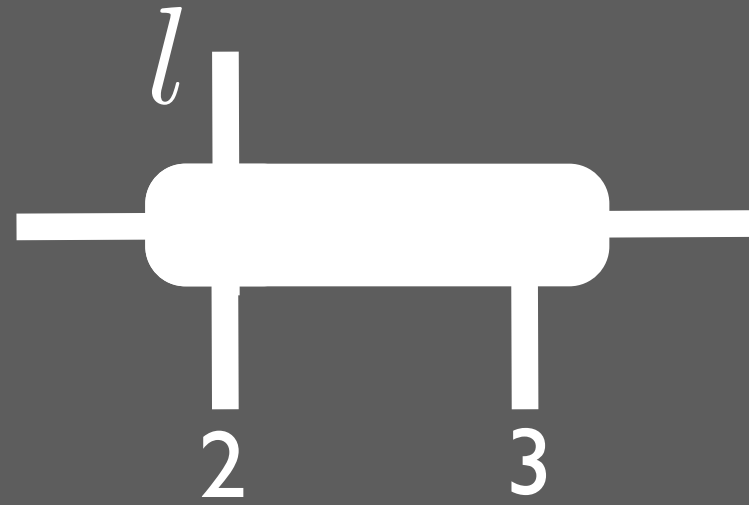
**tensorize weights
matrix**



Updating the weights

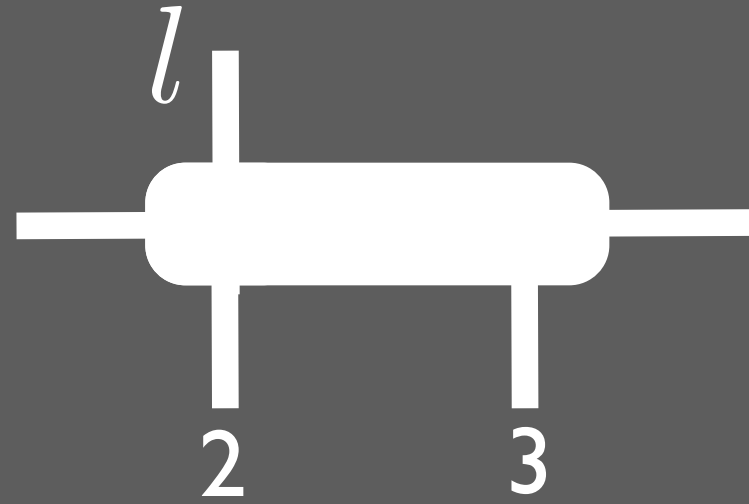


Updating the weights



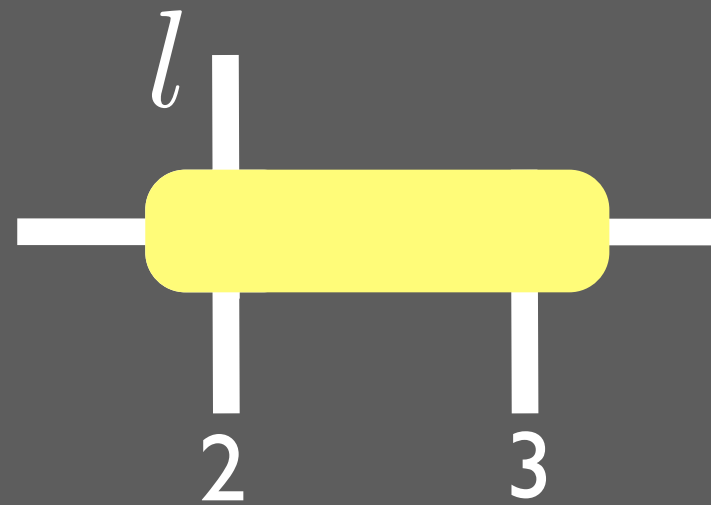
Updating the weights

B^l



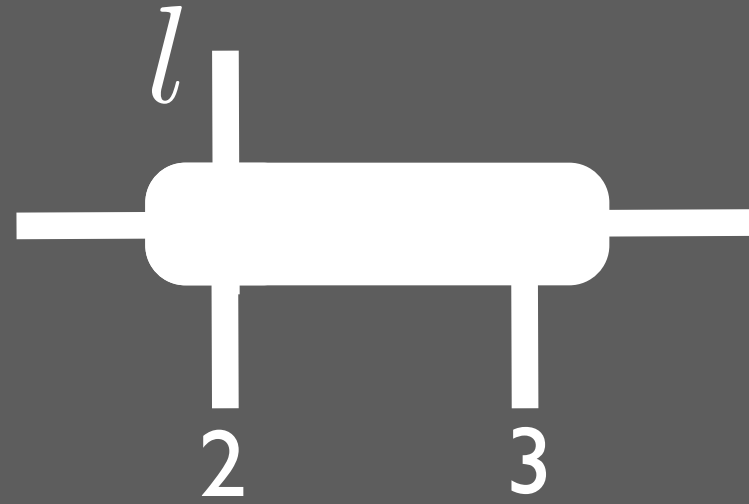
+

ΔB^l



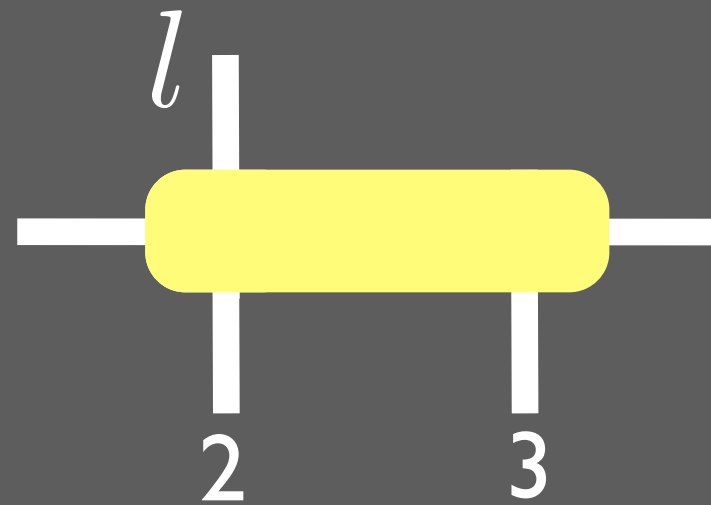
Updating the weights

B^l

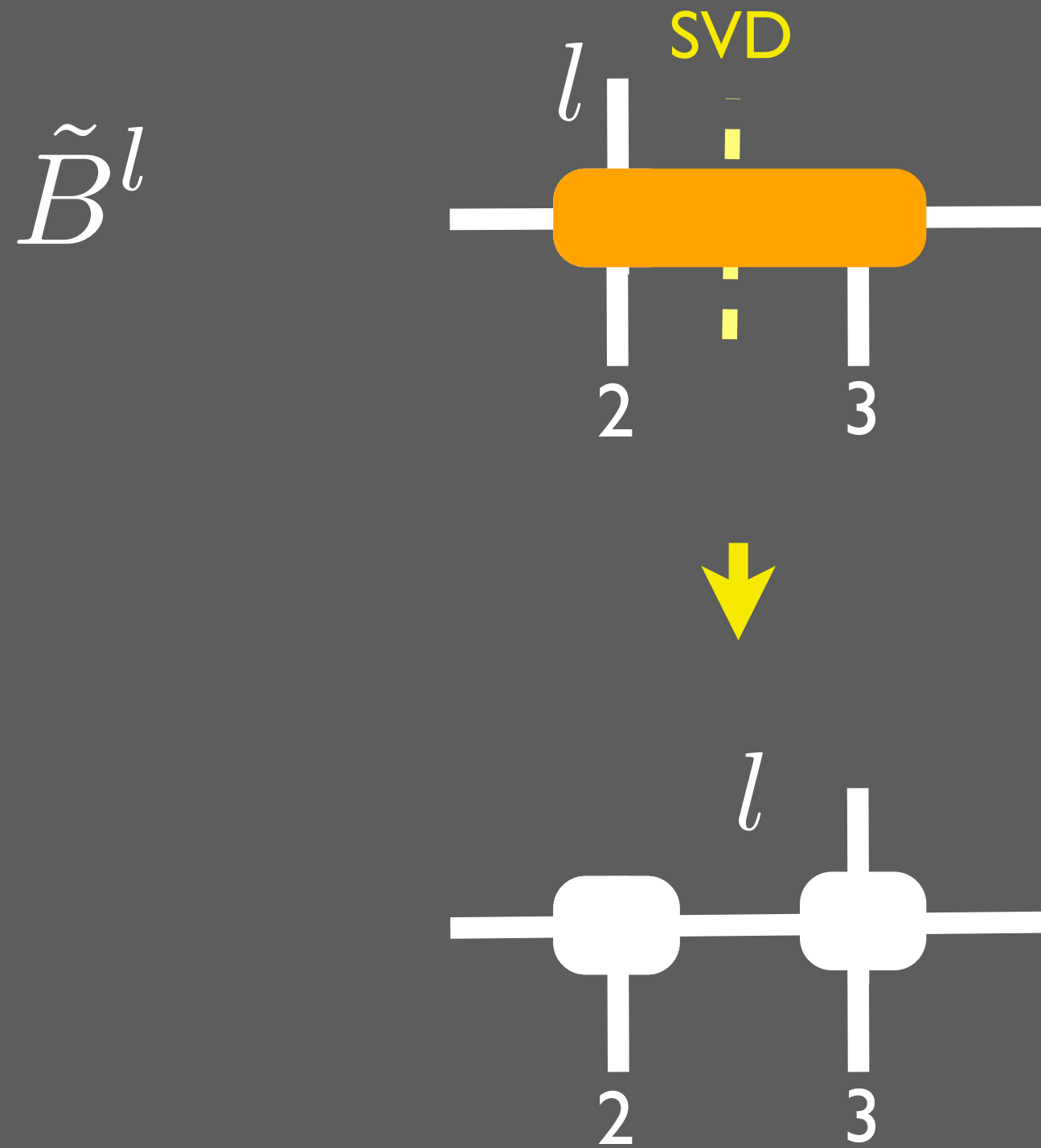


+

ΔB^l



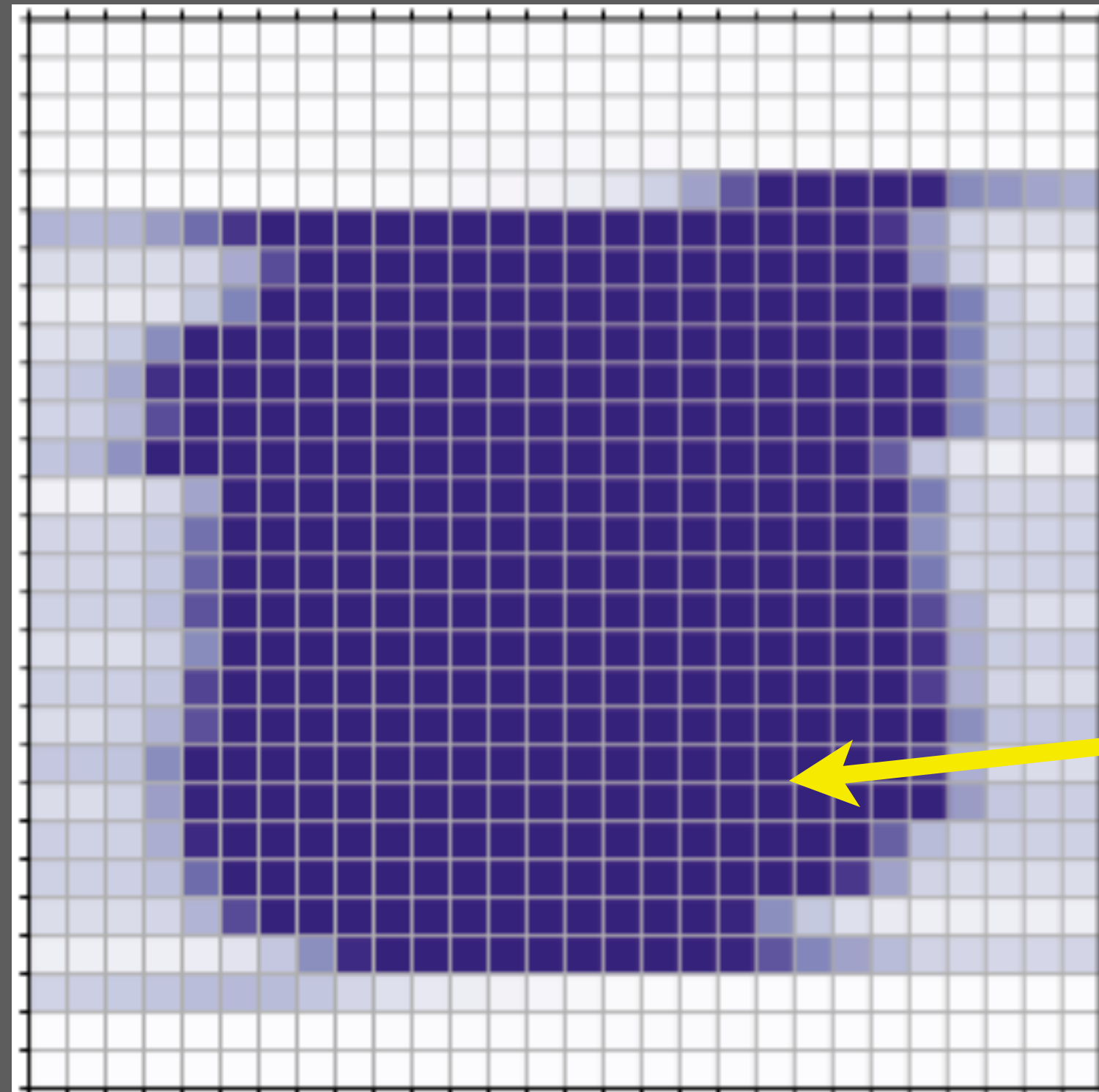
Updating the weights



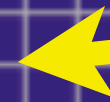
99.03% on ID MNIST

Network adjusts itself according to data complexity

**decorrelated
pixels**



**large bond
dimension**



Han et.al., ArXiv:1709.01662 (2017)

Optimal distribution modelling

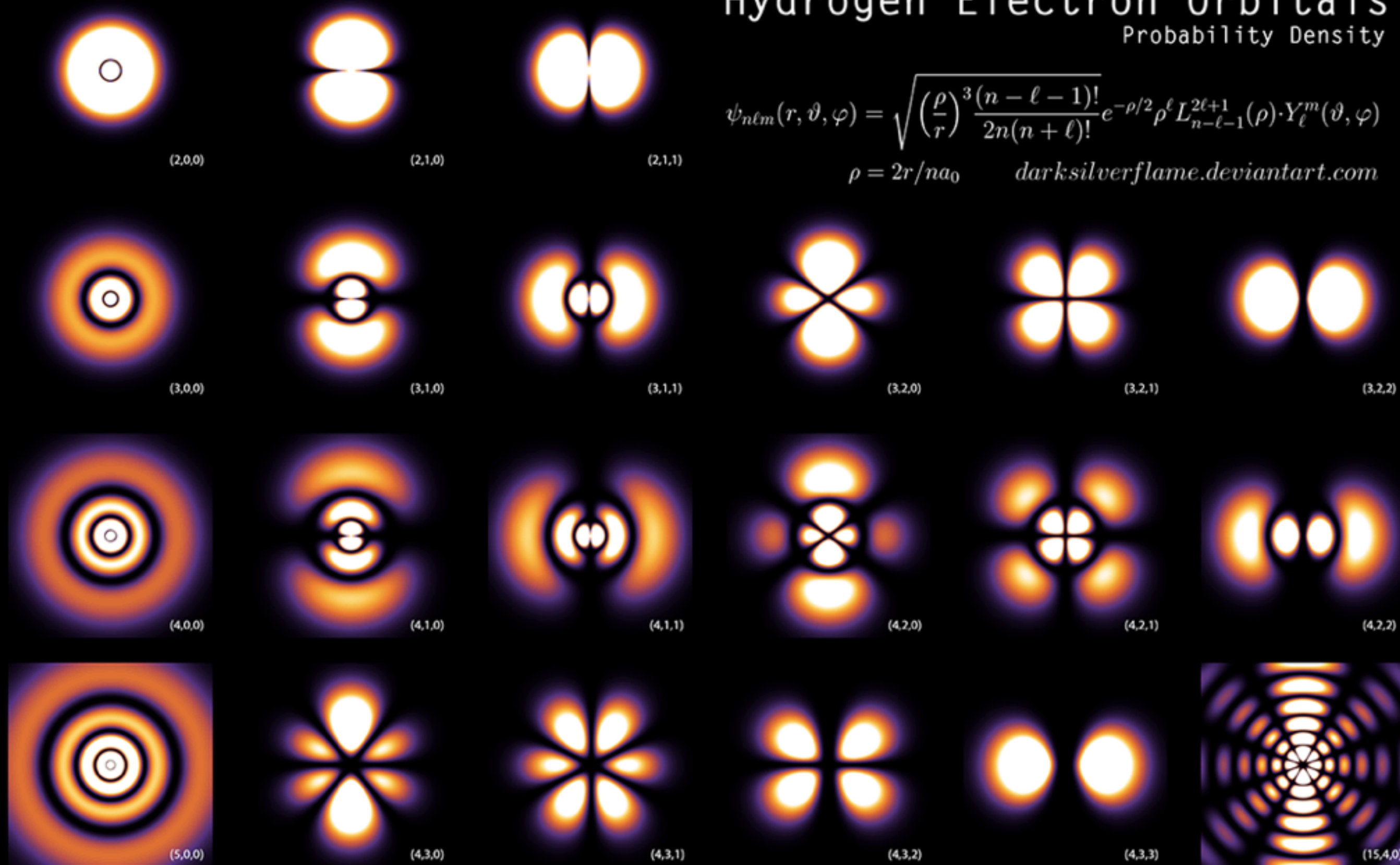
Hydrogen Electron Orbitals

Probability Density

$$\psi_{nlm}(r, \vartheta, \varphi) = \sqrt{\left(\frac{\rho}{r}\right)^3 \frac{(n-l-1)!}{2n(n+l)!}} e^{-\rho/2} \rho^l L_{n-l-1}^{2l+1}(\rho) \cdot Y_l^m(\vartheta, \varphi)$$

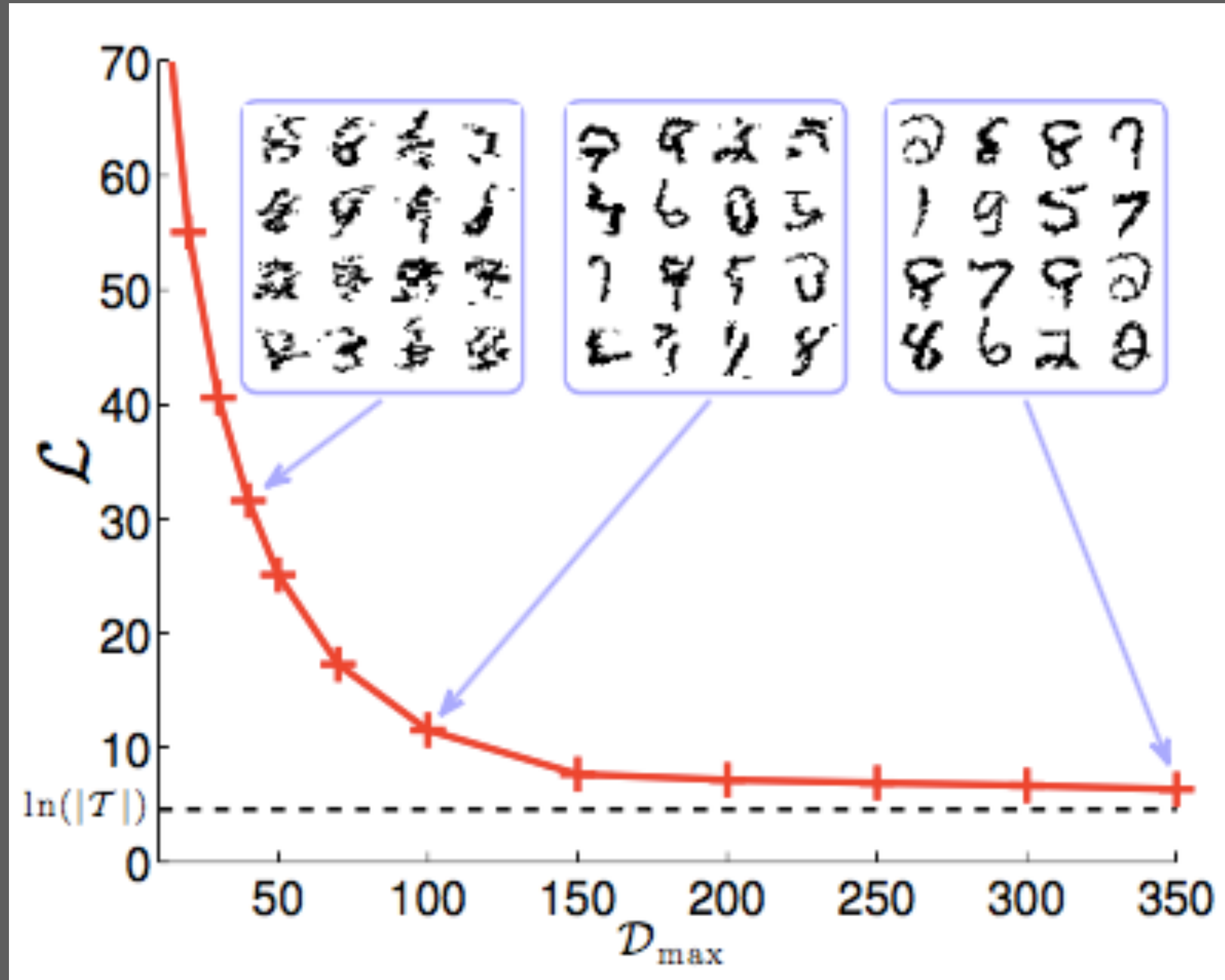
$$\rho = 2r/na_0$$

darksilverflame.deviantart.com

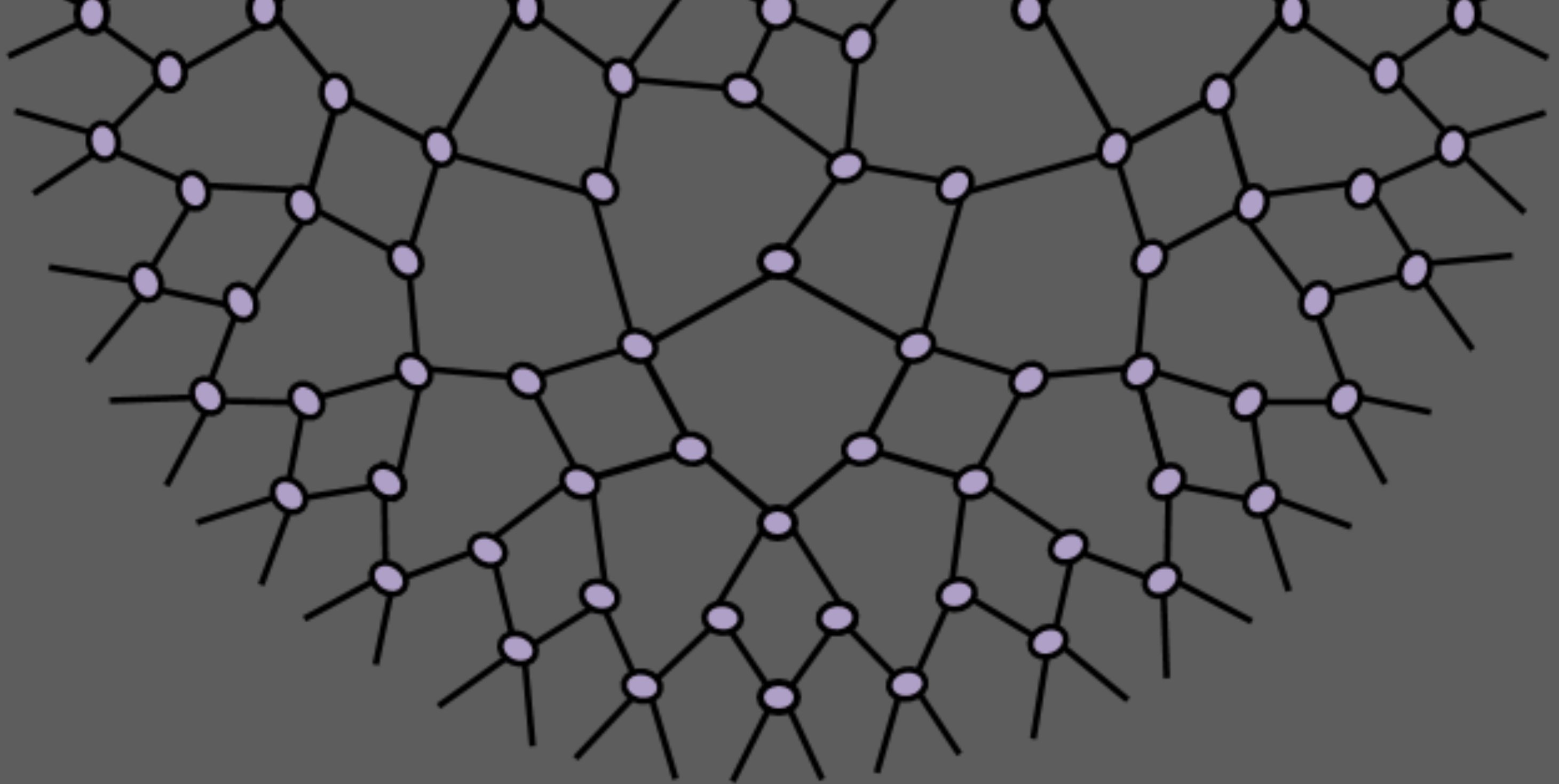


darksilverflame.deviantart.com

Optimal distribution modelling

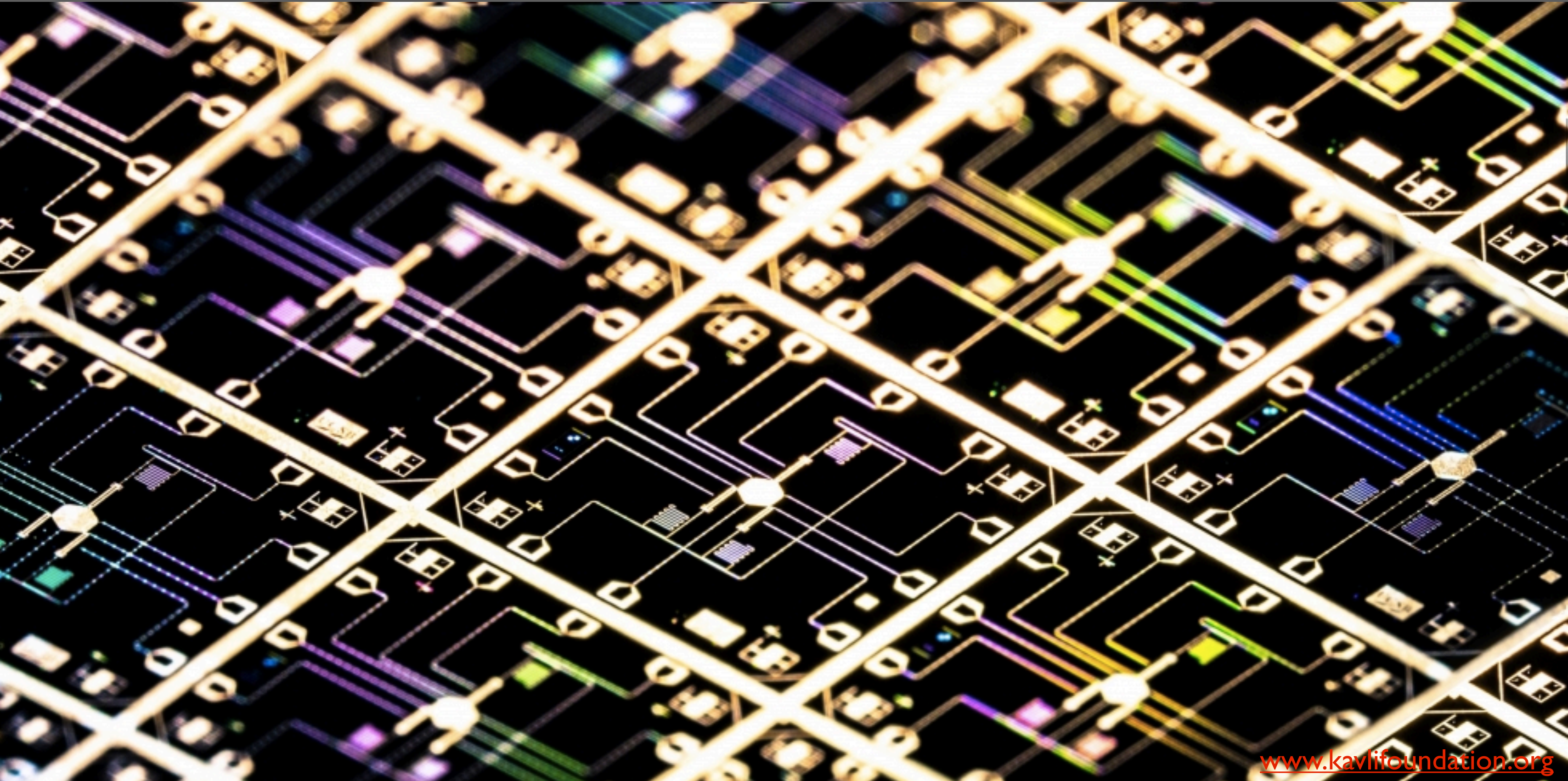


Han et al., ArXiv:1709.01662 (2017)



A Zoo of open architectures and tricks!

Extrapolation to Quantum Qubit arrays ?



Thank you!

Handwritten physics notes covering various topics:

- Top Left:** A diagram showing a central point with several lines radiating outwards, labeled with Y_1, Y_2, Y_3 and Y_4, Y_5, Y_6 . Below it are some calculations involving k_{lab} and M_0 .
- Top Center:** A graph of $\psi_0^2(x) = |\psi_0(x)|^2$ versus x . The x-axis ranges from -5 to 5. The curve shows a central peak at $x=0$ and smaller side peaks. Below the graph is the equation $\psi_0(x) = A e^{-|x|/2}$.
- Top Right:** A diagram of a hydrogen atom with a central nucleus and an orbiting electron. It includes the formula $K = \frac{(Z_1 e)(Z_2 e)}{4\pi\epsilon_0 r}$ and $N(\theta) = \frac{N_i}{N_e} = 32 \times 10^{-1} m^{-2} = 3.2 \times 10^{-7} mm^{-2}$. Other notes include "Hydrogen", "Electron", and "Nucleus".
- Middle Left:** A graph of potential energy $V(r)$ versus distance r . The curve shows a deep well at small r and levels off at a constant value for large r . Equations include $V(r) = \frac{1}{2} \frac{e^2}{r}$ and $E_{rot} = \frac{L^2}{2I}$.
- Middle Right:** A diagram of a right-angled triangle with a hypotenuse. The angle at the bottom left is labeled θ . The hypotenuse is labeled mgc^2 . Below it is the equation $(ER)^2 - p^2 = (mc^2)^2$.
- Bottom Left:** A diagram of a circular ring or disk with a central point and a radius r . It includes the equation $E^{tot}(x, t) = \sum_{n,j} \frac{E_{rot}(x, t) + E_{adv}(x, t)}{2}$.
- Bottom Center:** A diagram of a particle with a central point and a radius r . It includes the equation $E = mc^2$ and $E^2 = p^2 c^2 + m^2 c^4$.
- Bottom Right:** A diagram of a particle with a central point and a radius r . It includes the equation $E = h\nu \rightarrow E = h\omega$ and $P = \frac{h}{\lambda} = \frac{E}{c} \rightarrow E = hc/\lambda$.