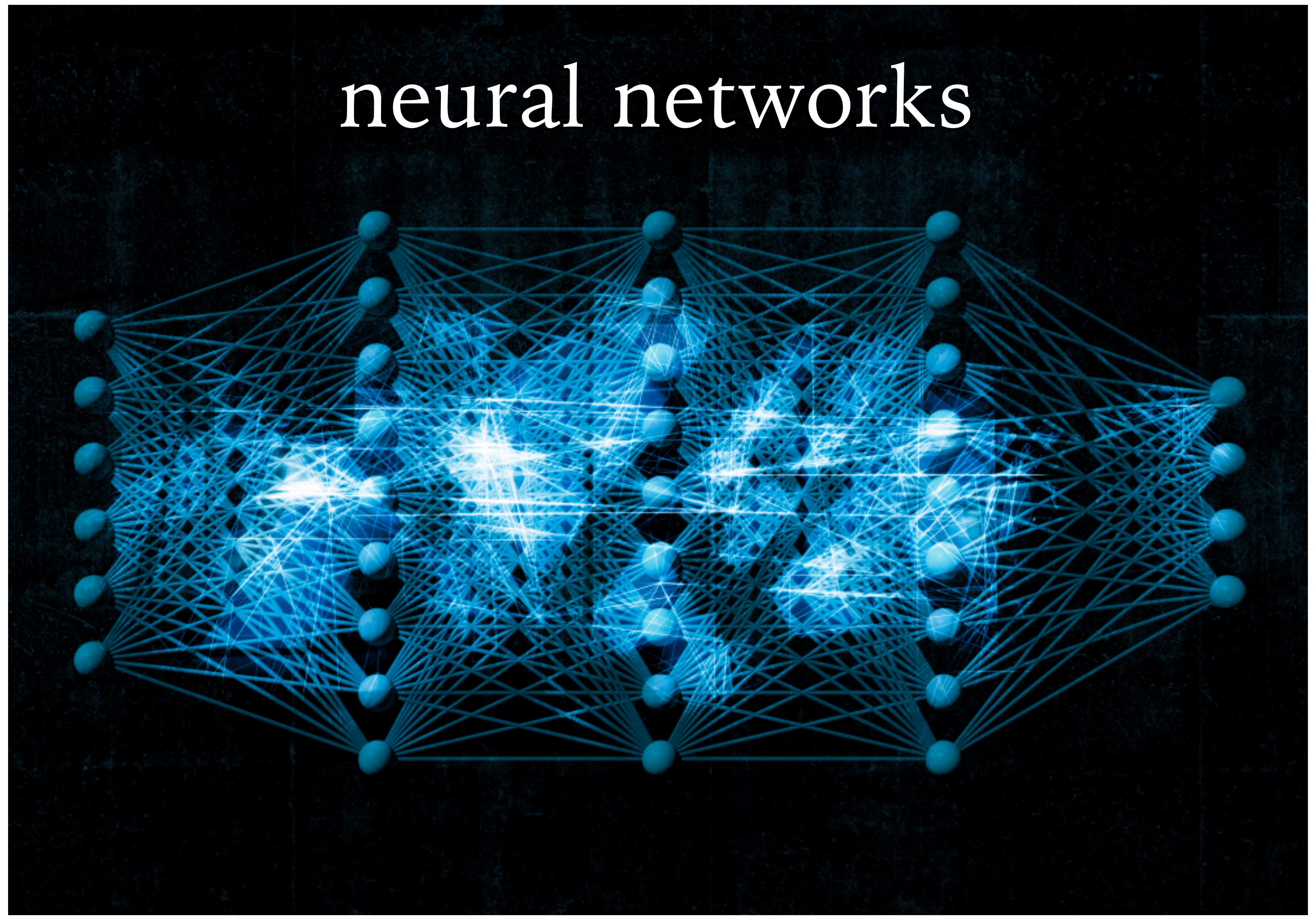
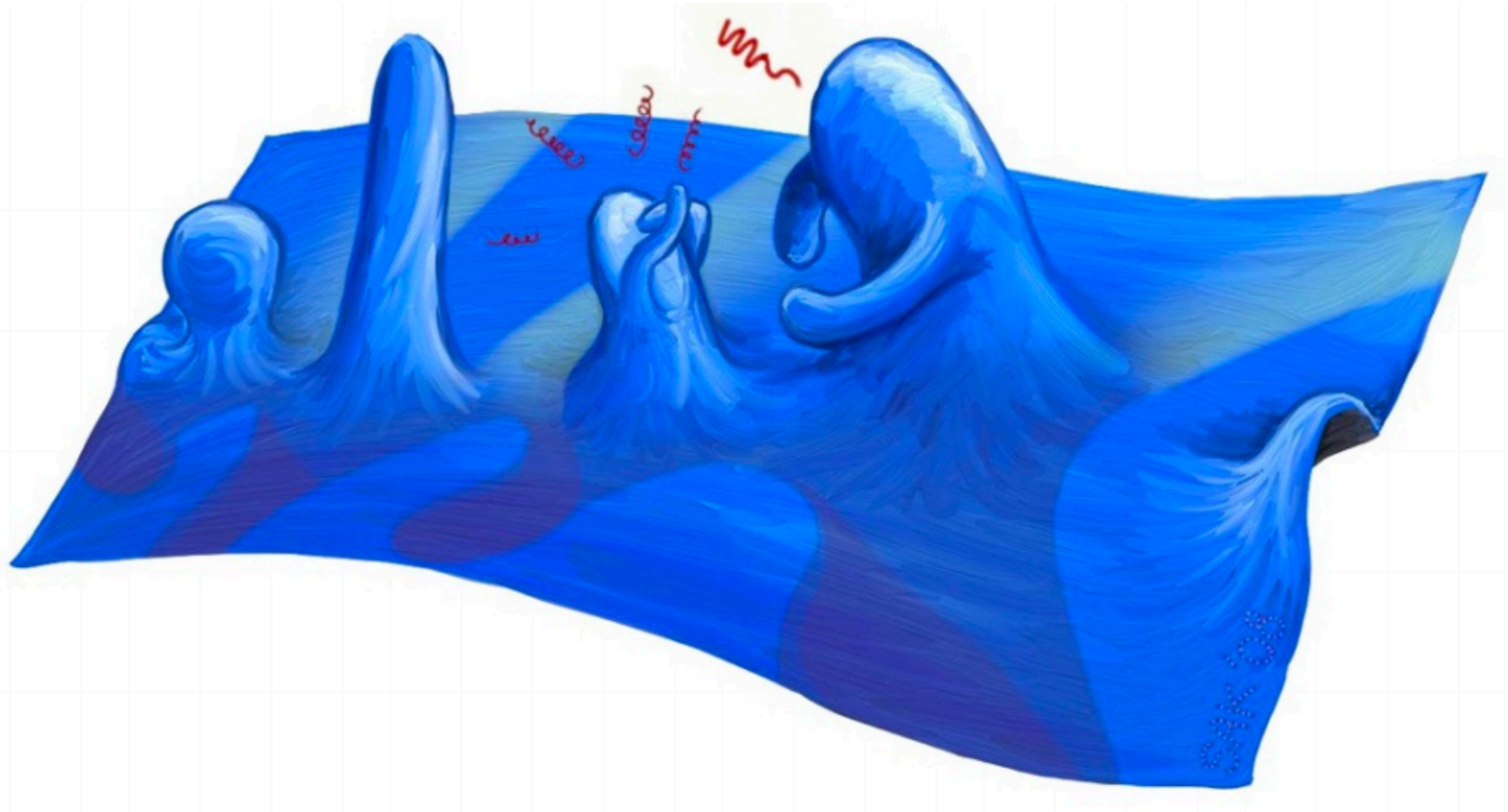


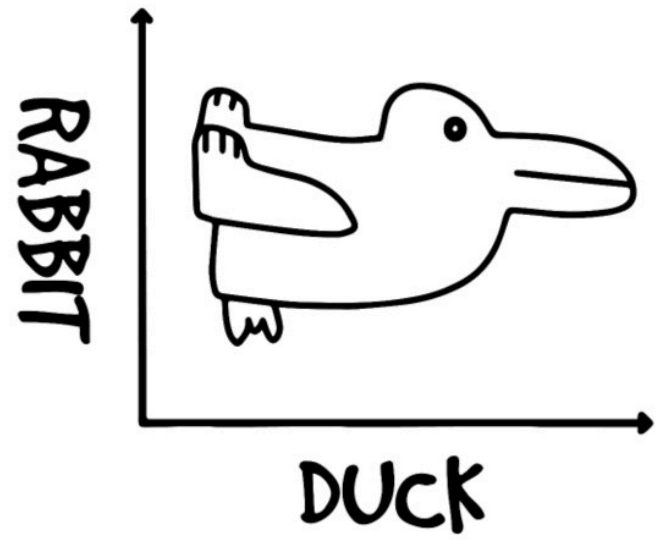
field theories



Machine Learning and (Large-N) Field Theory

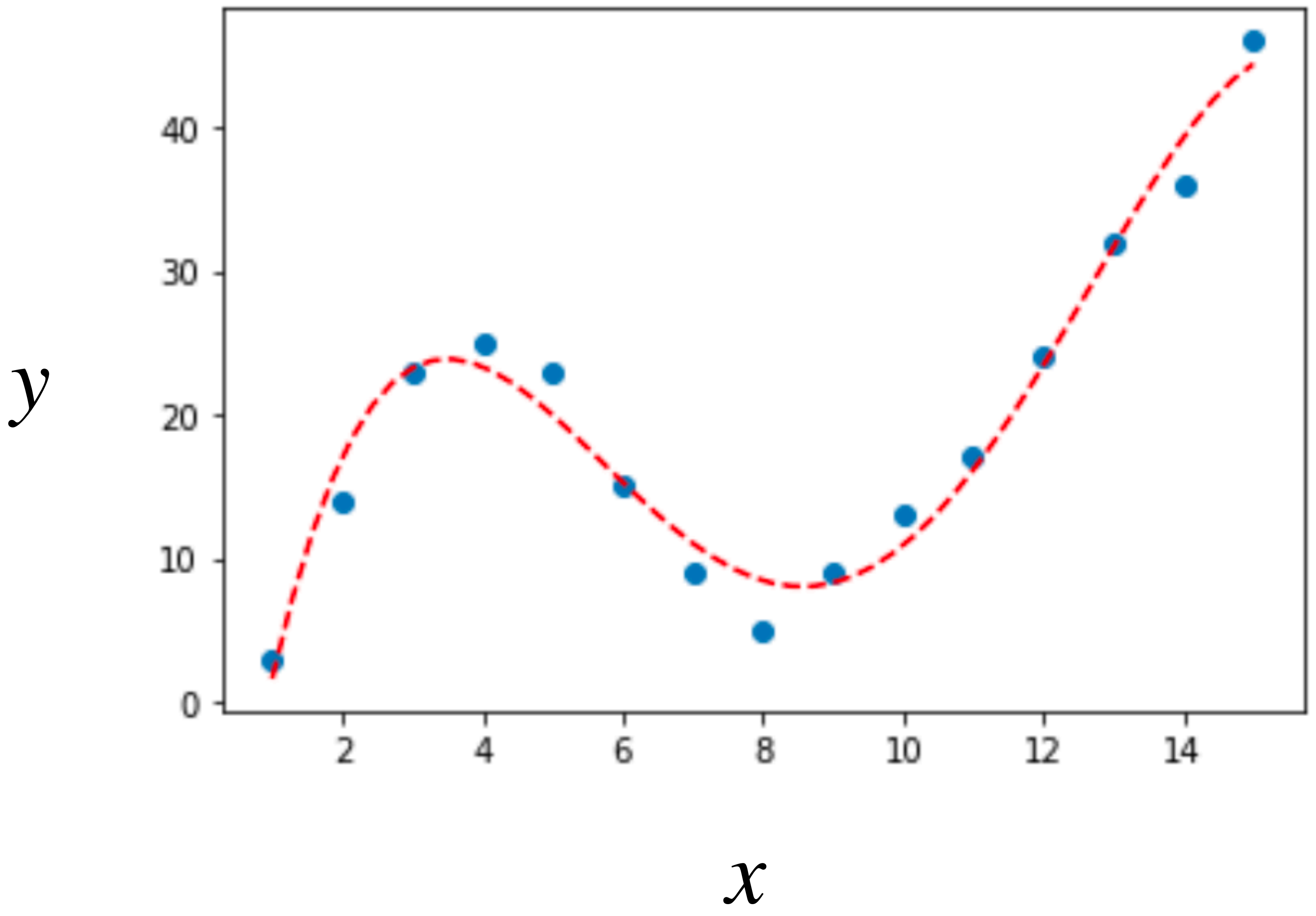
Zhengkang "Kevin" Zhang (University of Utah)

I. Banta, T. Cai, N. Craig, ZZ, 2305.02334 + ZZ, to appear.



What is a (deep) neural network?

Goal (supervised learning): learn a function $y = f(\vec{x})$ from training dataset $(\vec{x}_\alpha, y_\alpha)$.

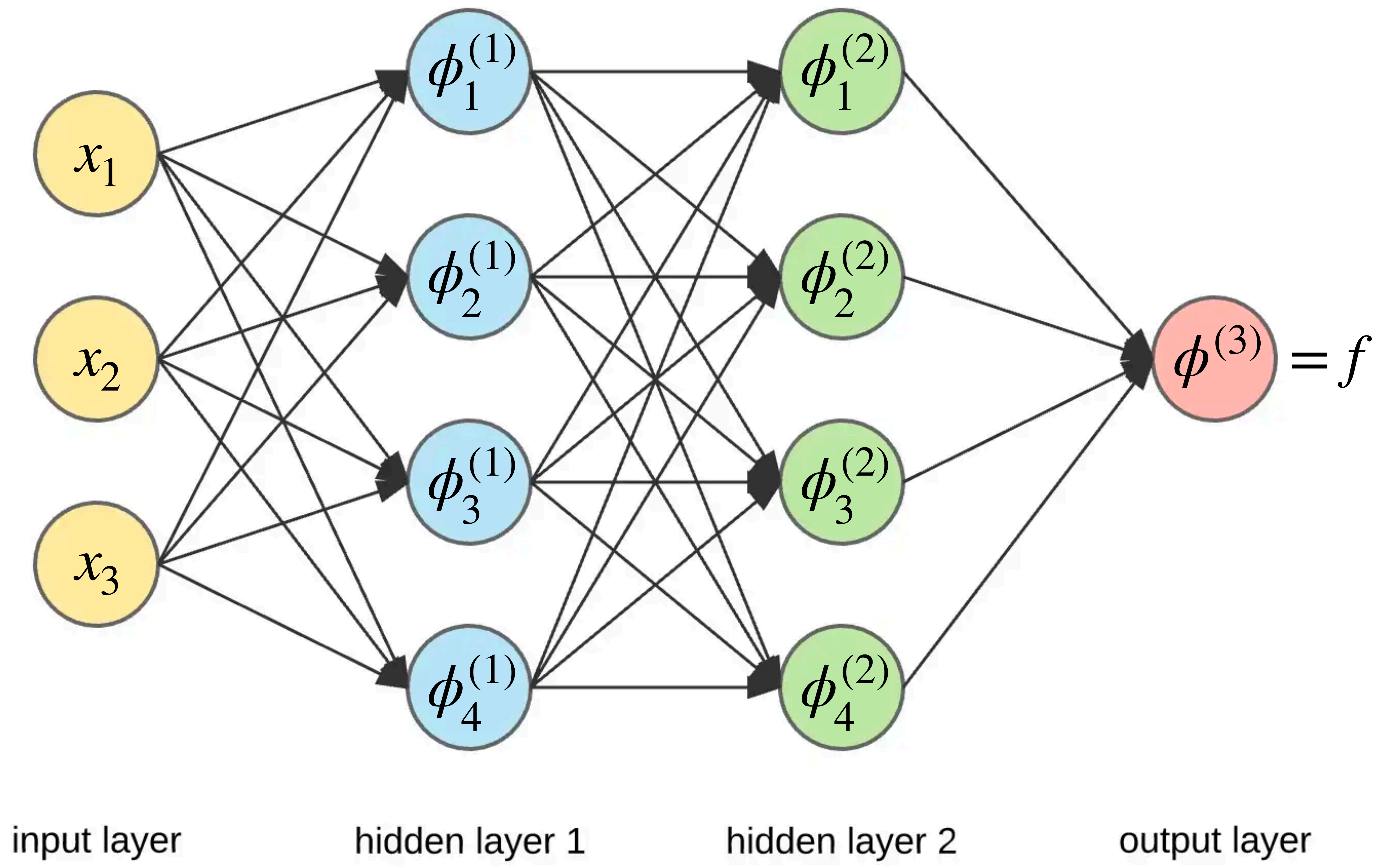


x Image	y Label
	Cat
	Cat
	Dog
	Dog

What is a (deep) neural network?

Goal (supervised learning): learn a function $y = f(\vec{x})$ from training dataset $(\vec{x}_\alpha, y_\alpha)$.

Neural network = parameterized function with a huge number of parameters ("expressive" enough to represent complicated functions).



$$\phi_i^{(1)}(\vec{x}) = \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j + b_i^{(1)},$$

$$\phi_i^{(\ell)}(\vec{x}) = \sum_{j=1}^{n_{\ell-1}} W_{ij}^{(\ell)} \sigma\left(\phi_j^{(\ell-1)}(\vec{x})\right) + b_i^{(\ell)} \quad (\ell \geq 2).$$

nonlinear activation function (e.g. tanh)

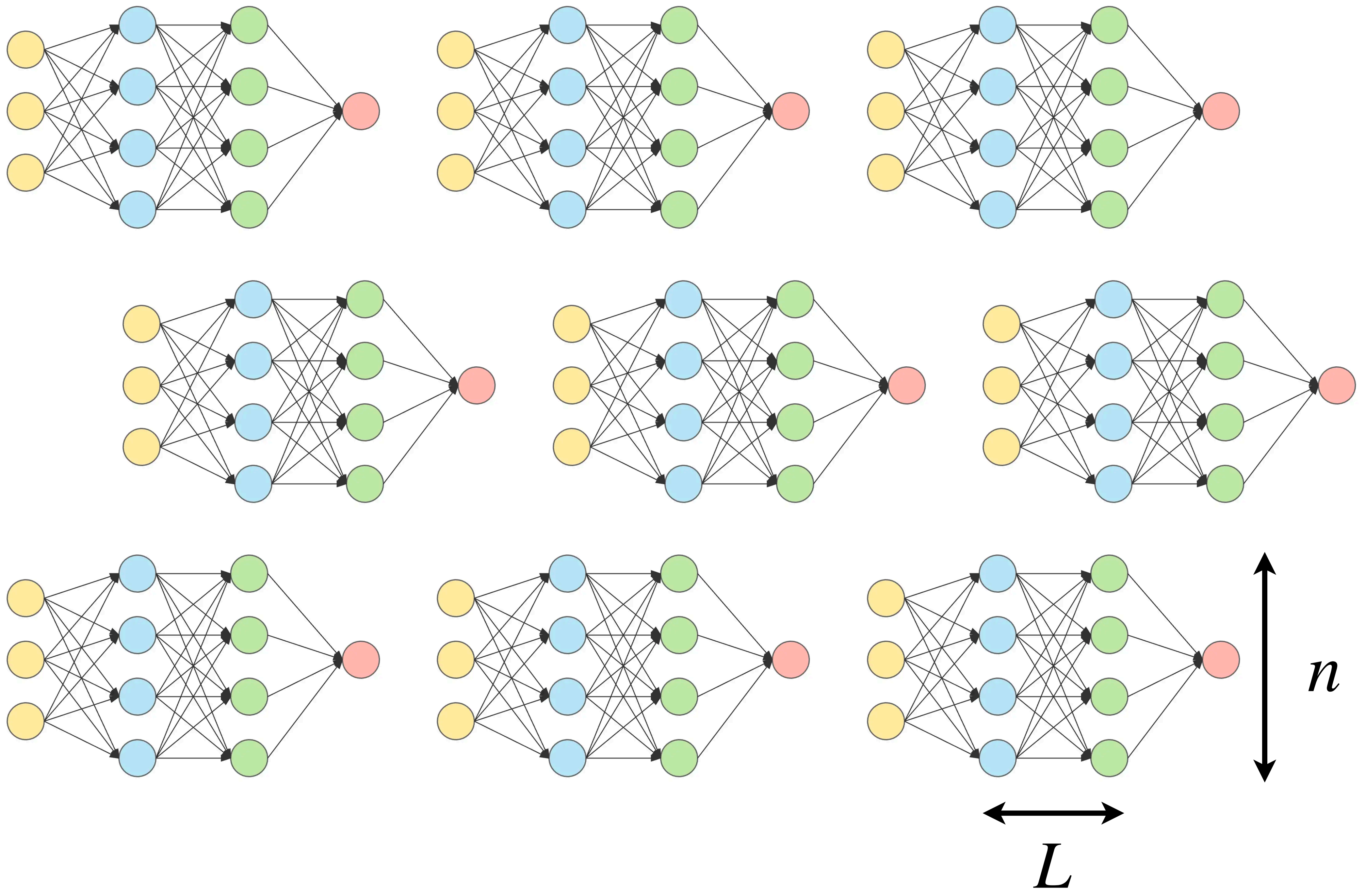
weights biases

trainable parameters:

- randomly initialized
- then updated to fit training data

Neural networks $\leftrightarrow$ field theories

Ensemble of networks, randomly initialized.



Neurons $\leftrightarrow$ scalar fields $\phi(\vec{x})$.

Ensemble statistics $\leftrightarrow$ action: $P(\phi) = e^{-S[\phi]}$.

$$\left\langle \phi^{(\ell)}_{i_1}(\vec{x}_1)\ \dots\ \phi^{(\ell)}_{i_{2k}}(\vec{x}_{2k}) \right\rangle = \int \mathcal{D}\phi\, \phi^{(\ell)}_{i_1}(\vec{x}_1)\ \dots\ \phi^{(\ell)}_{i_{2k}}(\vec{x}_{2k})\, e^{-S[\phi]}$$

Evolution with layer ℓ $\leftrightarrow$ RG flow.

Infinitely-wide networks* $(n \to \infty)$ $\leftrightarrow$ free theories.

* Neal '96. Williams '96.

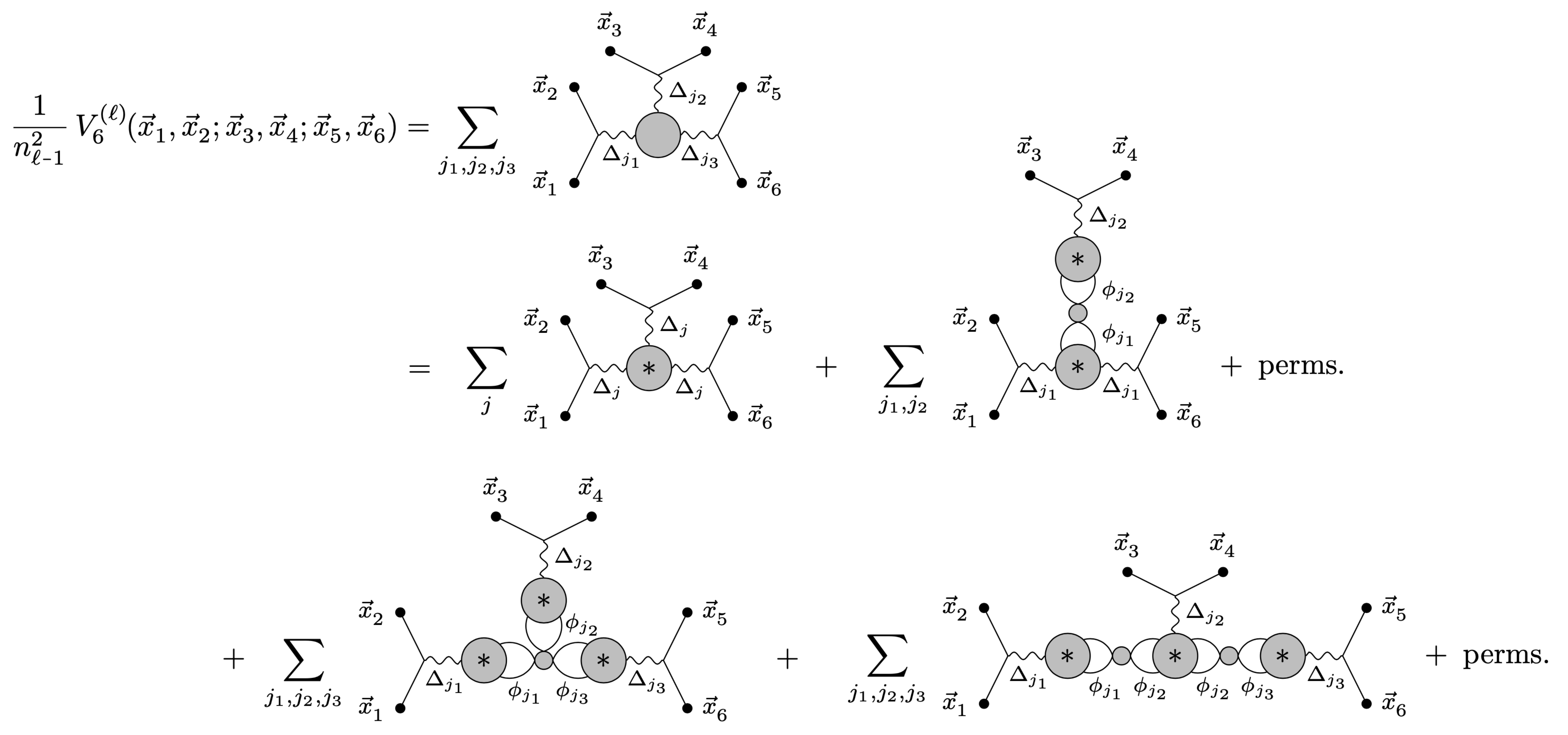
Wide networks $(n \gg L)$ $\leftrightarrow$ weakly-interacting theories (perturbative).

Diagrammatic framework

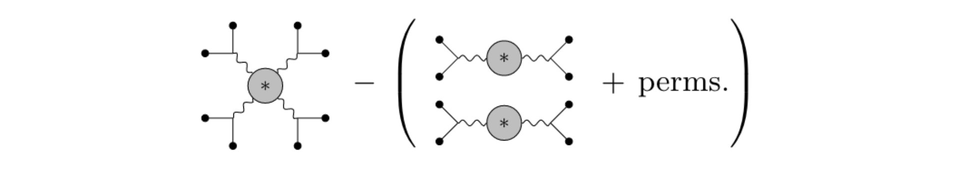
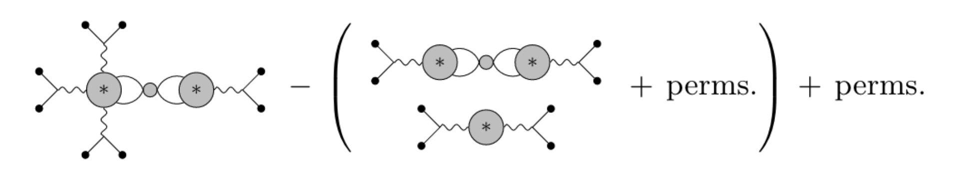
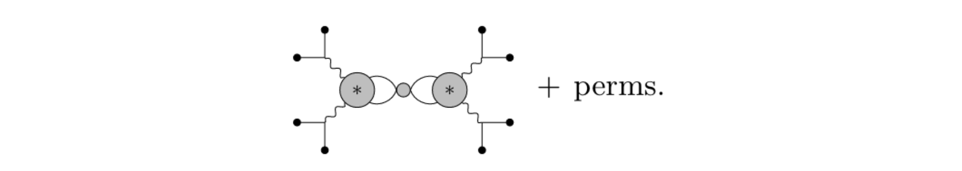
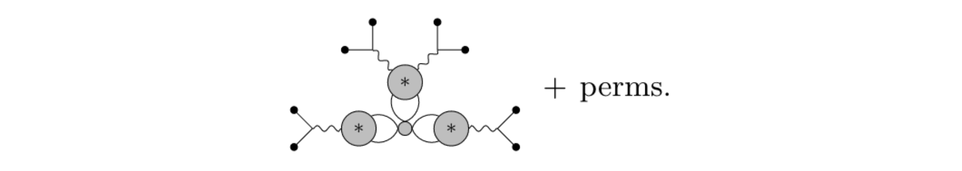
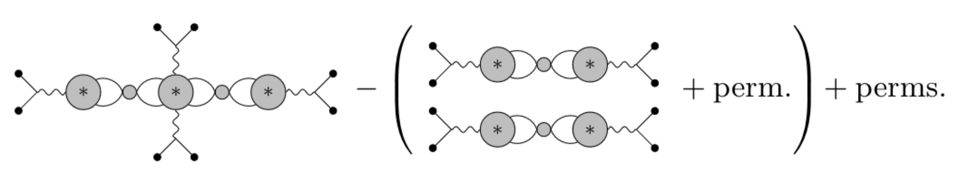
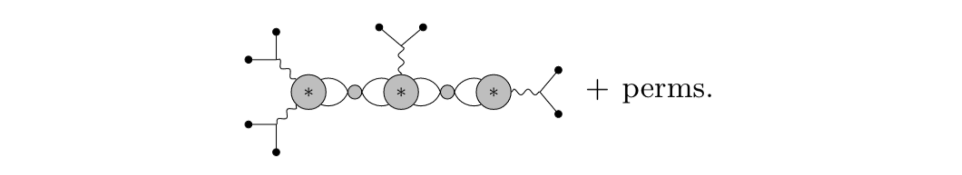
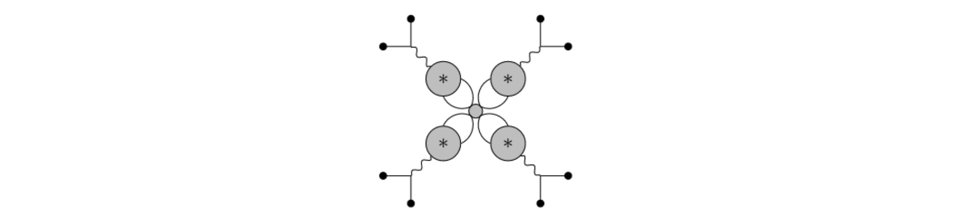
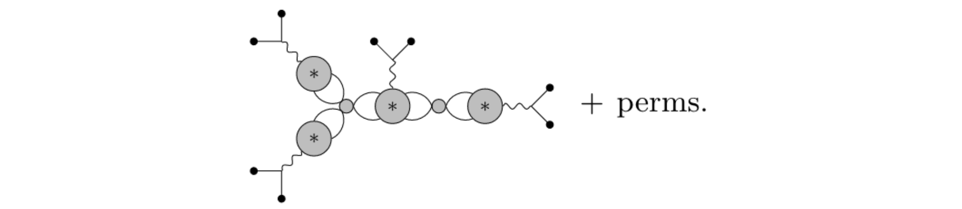
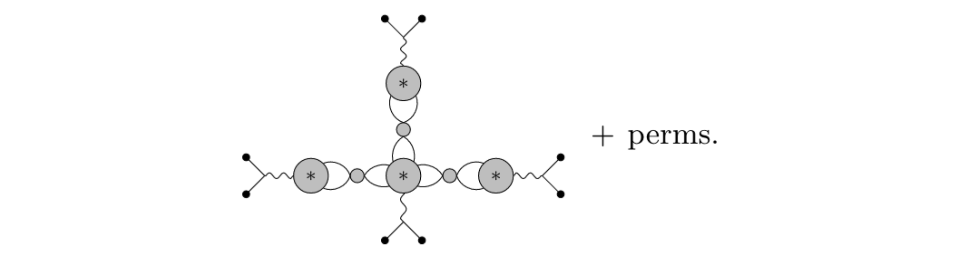
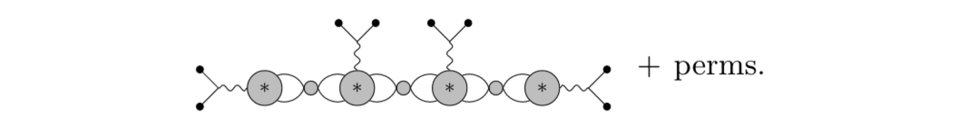
[I. Banta, T. Cai, N. Craig, ZZ, 2305.02334]

In addition to reproducing known results for lower-point correlators, we were able to push $1/n$ calculations to higher orders using diagrams.

6-point



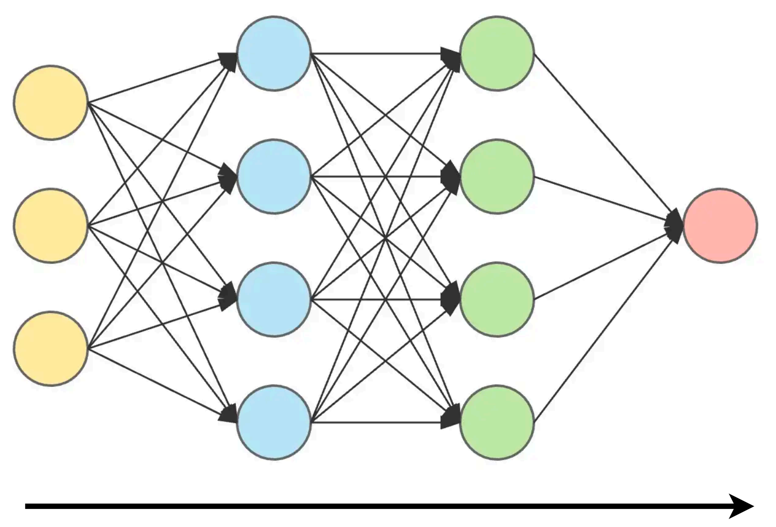
8-point

# of j sums	diagrams	degenerate limit result $/ (C_W^{(\ell)})^4$
1		$\frac{1}{n_{\ell-1}^3}\left(\langle\Delta^4\rangle - 3\langle\Delta^2\rangle^2\right)$
2		$\frac{1}{4}\cdot\frac{V_4^{(\ell-1)}}{n_{\ell-1}^2 n_{\ell-2}}\cdot 4\left(\langle\partial^2\Delta^3\rangle\langle\partial^2\Delta\rangle - 3\langle\partial^2\Delta\rangle^2\langle\Delta^2\rangle\right)$
		$\frac{1}{4}\cdot\frac{V_4^{(\ell-1)}}{n_{\ell-1}^2 n_{\ell-2}}\cdot 3\langle\partial^2\Delta^2\rangle^2$
3		$\frac{1}{8}\cdot\frac{V_6^{(\ell-1)}}{n_{\ell-1} n_{\ell-2}^2}\cdot 6\langle\partial^2\Delta^2\rangle\langle\partial^2\Delta\rangle^2$
		$\frac{1}{16}\cdot\frac{\left(V_4^{(\ell-1)}\right)^2}{n_{\ell-1} n_{\ell-2}^2}\cdot 6\left(\langle\partial^4\Delta^2\rangle\langle\partial^2\Delta\rangle^2 - 2\langle\partial^2\Delta\rangle^4\right)$
		$\frac{1}{16}\cdot\frac{\left(V_4^{(\ell-1)}\right)^2}{n_{\ell-1} n_{\ell-2}^2}\cdot 12\langle\partial^4\Delta\rangle\langle\partial^2\Delta^2\rangle\langle\partial^2\Delta\rangle$
4		$\frac{1}{16}\cdot\frac{V_8^{(\ell-1)}}{n_{\ell-2}^3}\langle\partial^2\Delta\rangle^4$
		$\frac{1}{32}\cdot\frac{V_6^{(\ell-1)}V_4^{(\ell-1)}}{n_{\ell-2}^3}\cdot 12\langle\partial^4\Delta\rangle\langle\partial^2\Delta\rangle^3$
		$\frac{1}{64}\cdot\frac{\left(V_4^{(\ell-1)}\right)^3}{n_{\ell-2}^3}\cdot 4\langle\partial^6\Delta\rangle\langle\partial^2\Delta\rangle^3$
		$\frac{1}{64}\cdot\frac{\left(V_4^{(\ell-1)}\right)^3}{n_{\ell-2}^3}\cdot 12\langle\partial^4\Delta\rangle^2\langle\partial^2\Delta\rangle^2$

Criticality

[I. Banta, T. Cai, N. Craig, ZZ, 2305.02334]

input (UV)



output (IR)

Exponential scaling (generic) $\leftrightarrow$ flow to trivial fixed point.

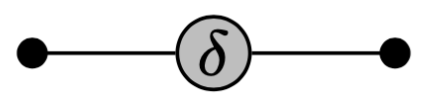
Tune to criticality $\Rightarrow$ power-law scaling $\leftrightarrow$ nontrivial fixed point.

Raghu et al '16. Poole et al '16. Schoenholz et al '16.

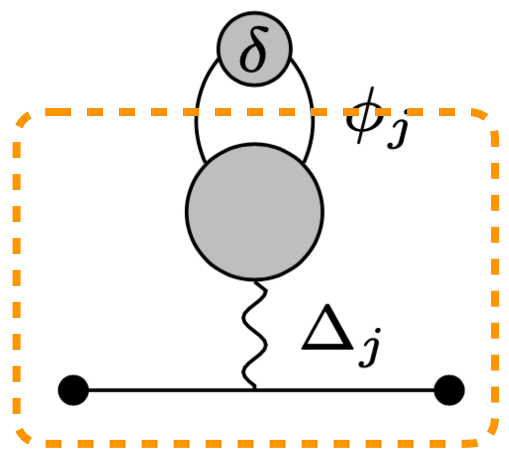
2-point correlator analysis:

$$\langle \mathcal{G}^{(\ell\text{-}1)}(\vec{x}_1,\vec{x}_2)\rangle \to \langle \mathcal{G}^{(\ell\text{-}1)}(\vec{x}_1,\vec{x}_2)\rangle + \delta\langle \mathcal{G}^{(\ell\text{-}1)}(\vec{x}_1,\vec{x}_2)\rangle$$

$$\Rightarrow \quad \delta\langle \mathcal{G}^{(\ell)}(\vec{x}_1,\vec{x}_2)\rangle =$$



$$= \sum_j$$



$$= \int d\vec{y}_1 d\vec{y}_2\, \chi^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{y}_1,\vec{y}_2)\, \delta\langle \mathcal{G}^{(\ell\text{-}1)}(\vec{y}_1,\vec{y}_2)\rangle$$

$$\chi^{(\ell)} = \frac{C_W^{(\ell)}}{2}\left\langle \frac{\delta^2\Delta(\vec{x}_1,\vec{x}_2)}{\delta\phi(\vec{y}_1)\delta\phi(\vec{y}_2)}\right\rangle_{\mathcal{K}_0^{(\ell\text{-}1)}} + \mathcal{O}\left(\frac{1}{n}\right)$$

susceptibility

Tune to criticality:
$$\chi^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{y}_1,\vec{y}_2)\Big|_{\mathcal{K}_0^{(\ell\text{-}1)}=\mathcal{K}^\star} = \frac{1}{2}\Big[\delta(\vec{x}_1-\vec{y}_1)\,\delta(\vec{x}_2-\vec{y}_2)+\delta(\vec{x}_1-\vec{y}_2)\,\delta(\vec{x}_2-\vec{y}_1)\Big]$$

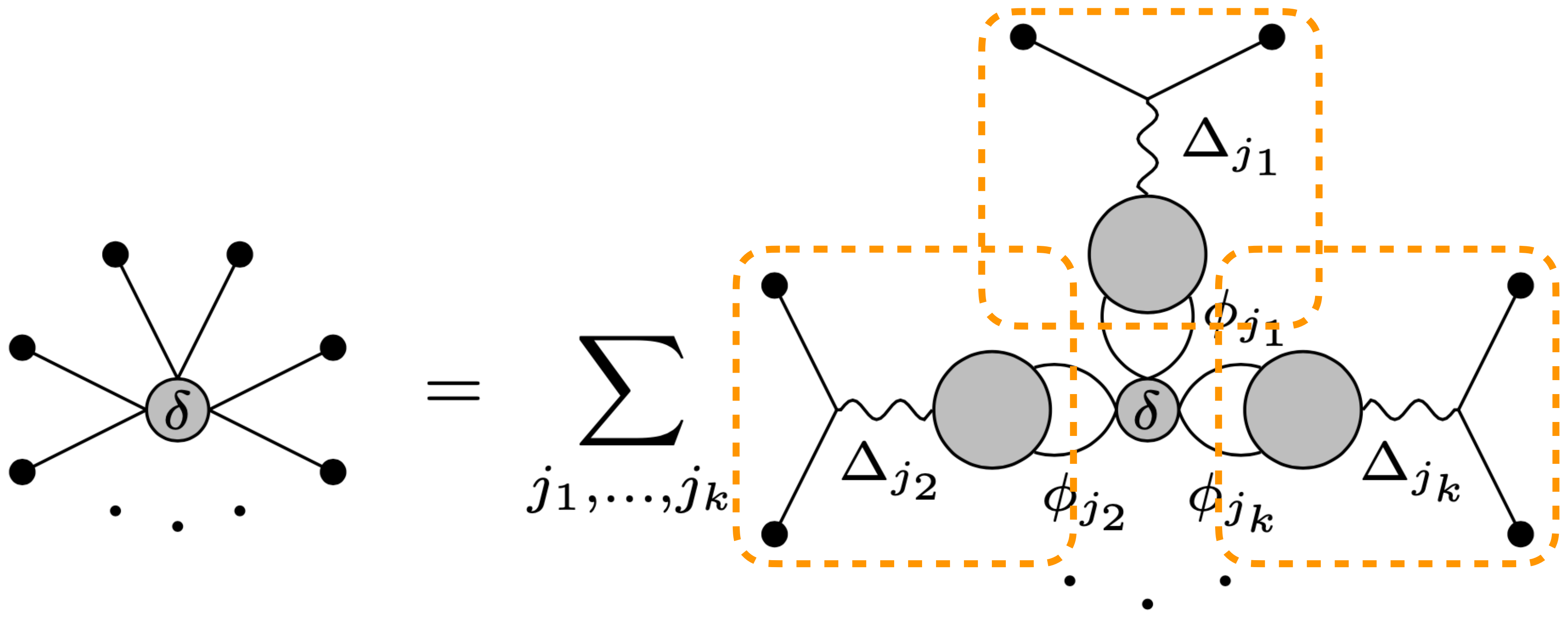
RG fixed point

Structures of RG flow

[I. Banta, T. Cai, N. Craig, ZZ, 2305.02334]

Higher-point correlators?

Common structure:



$$\Rightarrow \quad \left(\frac{n_{\ell\text{-}2}}{n_{\ell\text{-}1}}\right)^{k\text{-}1} \frac{\delta V_{2k}^{(\ell)}(\vec{x}_1,\vec{x}_2;\ldots;\vec{x}_{2k\text{-}1},\vec{x}_{2k})}{\delta V_{2k}^{(\ell-1)}(\vec{y}_1,\vec{y}_2;\ldots;\vec{y}_{2k\text{-}1},\vec{y}_{2k})} = \text{sym.}\left[\prod_{k'=1}^{k} \chi^{(\ell)}(\vec{x}_{2k'\text{-}1},\vec{x}_{2k'};\vec{y}_{2k'\text{-}1},\vec{y}_{2k'})\right]$$

same susceptibility that appeared in the 2-point correlator analysis!

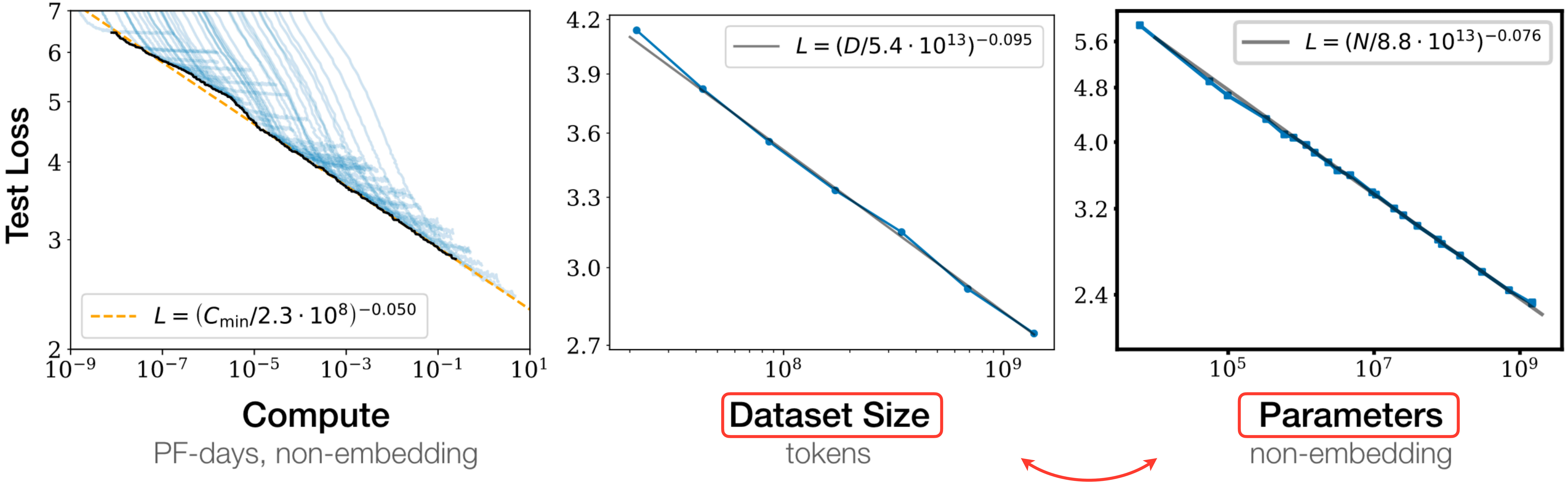
Single criticality condition: $\chi^{(\ell)}(\vec{x}_1,\vec{x}_2;\vec{y}_1,\vec{y}_2)\Big|_{\mathcal{K}_0^{(\ell\text{-}1)}=\mathcal{K}^\star} = \frac{1}{2}\Big[\delta(\vec{x}_1-\vec{y}_1)\,\delta(\vec{x}_2-\vec{y}_2)+\delta(\vec{x}_1-\vec{y}_2)\,\delta(\vec{x}_2-\vec{y}_1)\Big]$

$\Rightarrow$ Power-law scaling for all connected correlators!

Neural scaling laws & duality

Many ML models exhibit power law scaling of performance.

[Kaplan et al, 2001.08361] [Sharma, Kaplan, 2004.10802] [Bahri et al, 2102.06701]



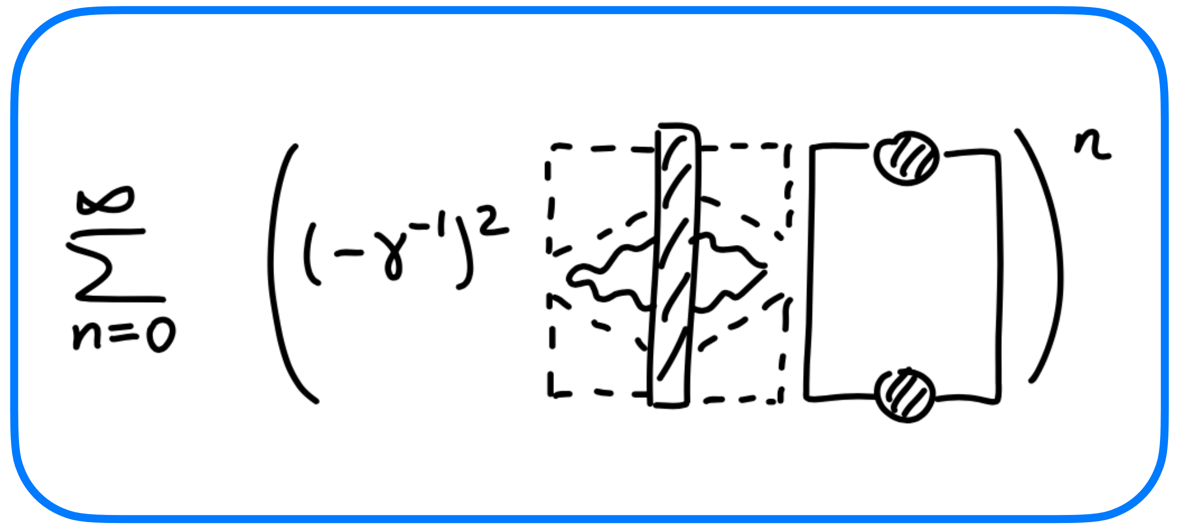
approximately symmetric

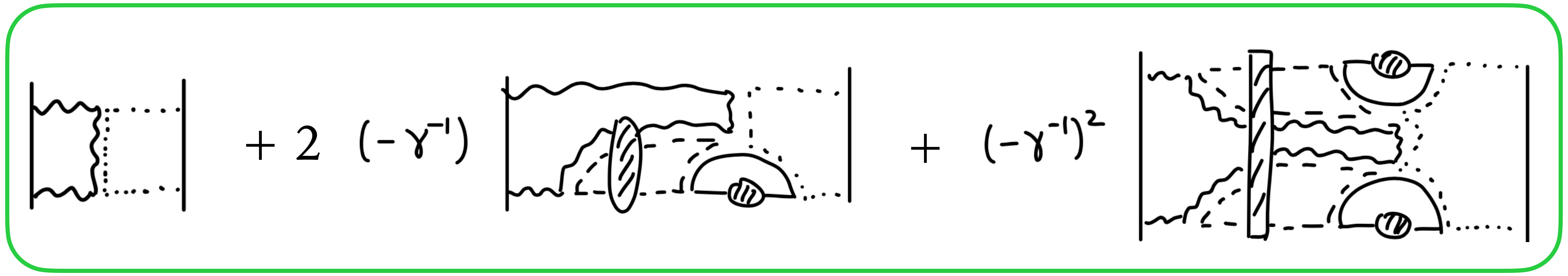
Ising model of neural scaling laws? [Maloney, Roberts, Sully, 2210.16859]

Large-N diagrammatics for neural scaling laws

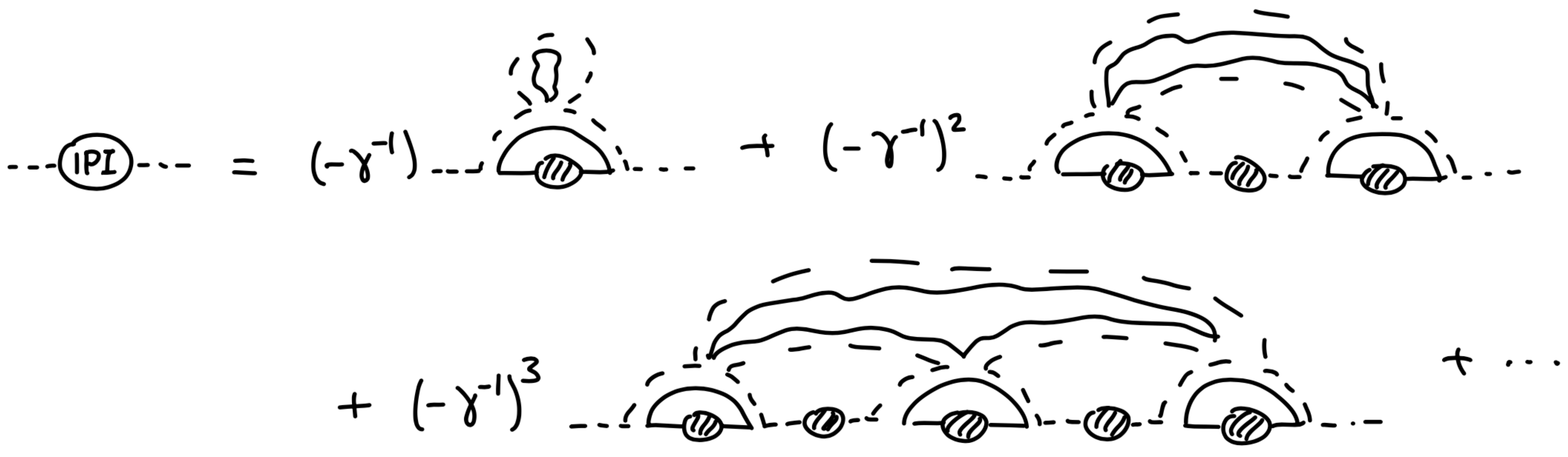
[ZZ, to appear]

$$\langle \text{Test loss} \rangle = \mathcal{R} \cdot (\mathcal{L}_1 + 2\mathcal{L}_2 + \mathcal{L}_3')$$

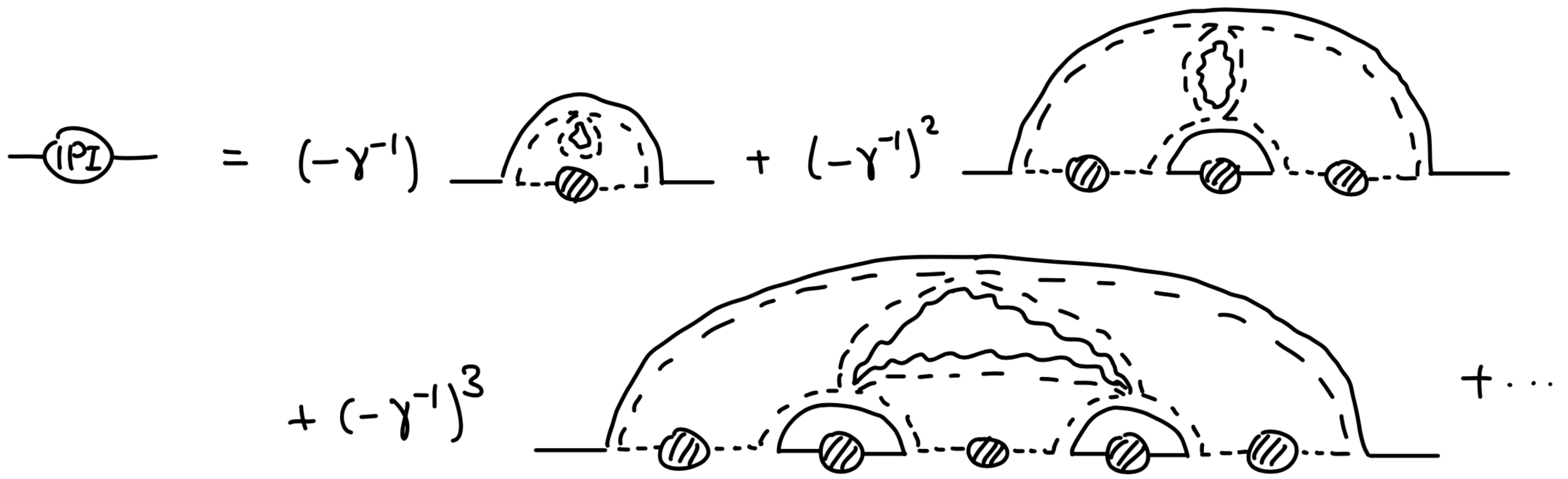




Sets of diagrams are related by a duality transformation, e.g.



dual



Dreams

A theory of ~~everything~~ deep learning (opening the black box)?

Lee et al '17-19. Matthews et al '18. Yang '19-23.

Jacot, Gabriel, Hongler '18.

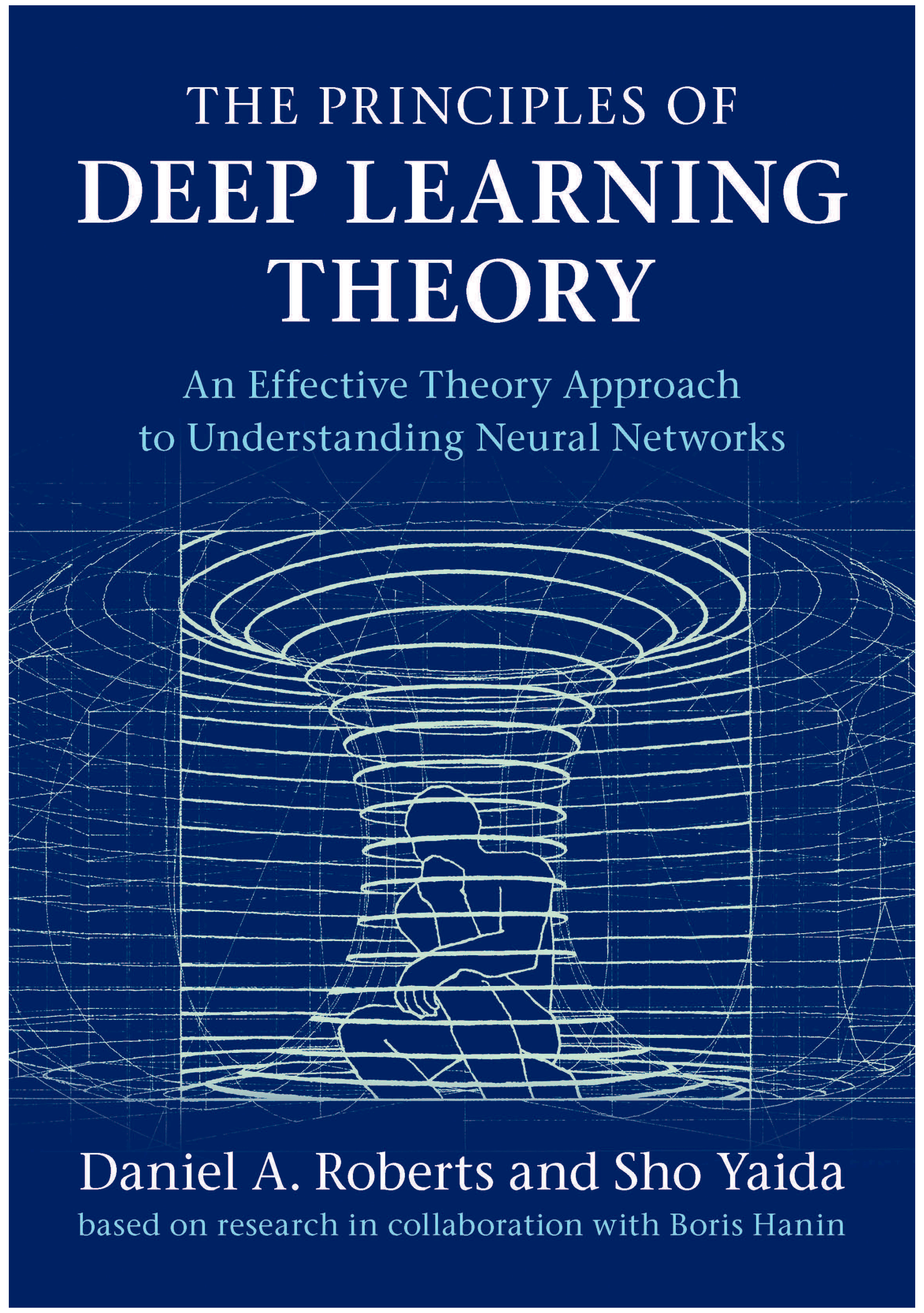
Antognini '19. Huang, Yau '19.

Yaida '19, '22. Hanin, Nica '19. Hanin '21, '22.

Dyer, Gur-Ari '19. Aitken, Gur-Ari '20. Andreassen, Dyer '20.

Naveh, Ringel et al '20, '21. Zavatone-Veth et al '21.

Roberts, Yaida, Hanin '21. (Our work is largely inspired by this book.)



Dreams

A theory of ~~everything~~ deep learning (opening the black box)?

A new angle to learn about field theories?

$$\langle \phi(x_1) \dots \phi(x_k) \rangle = \int d\theta \, P(\theta) \, \phi_\theta(x_1) \dots \phi_\theta(x_k) = \frac{1}{Z} \int \mathcal{D}\phi \, e^{-S[\phi]} \, \phi(x_1) \dots \phi(x_k)$$

parameter/feature space description ←dual→ functional/sample space description

Erbin, Lahoche, Samary '21, '22.

Bachtis, Aarts, Lucini '21.

Grosvenor, Jefferson (+ Erdmenger) '21.

Halverson '21; + Maiti, Stoner '20, '21; + Demirtas, Schwartz '23.