What Makes Data Suitable for Deep Learning?

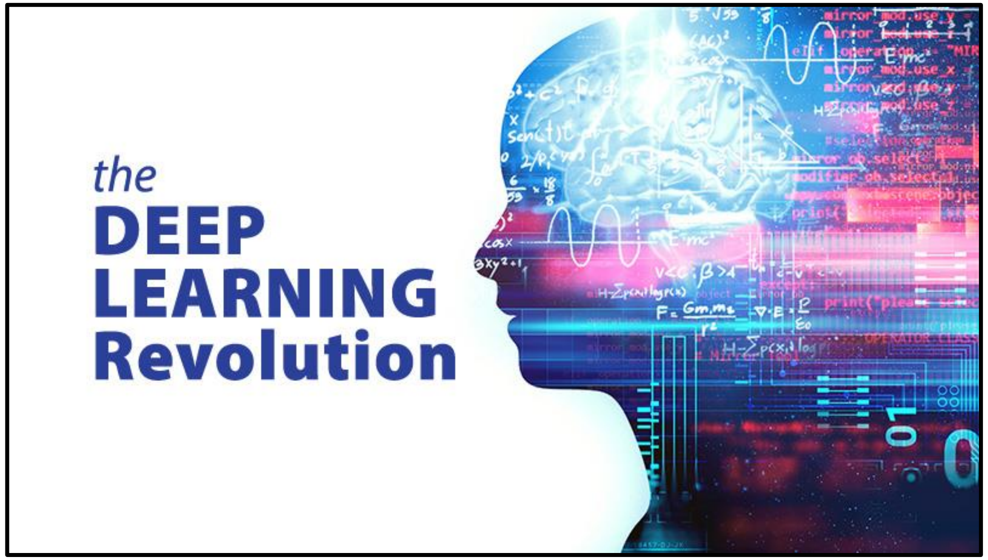
Nadav Cohen

Tel Aviv University

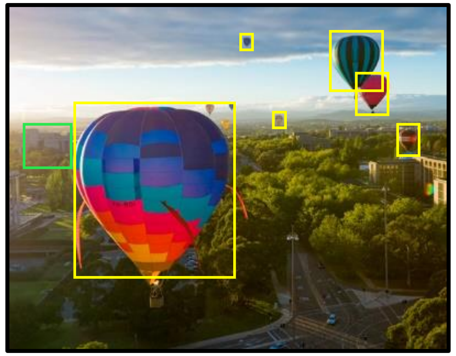
Conference on Foundations of Computational Mathematics

Workshop on Computational Harmonic Analysis and Data Science

15 June 2023

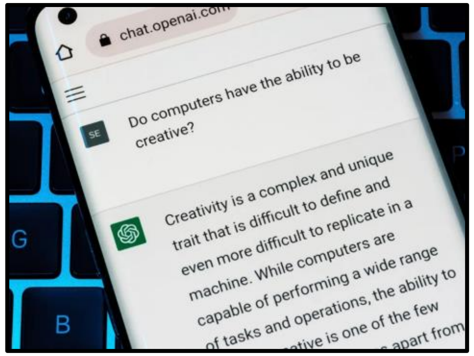


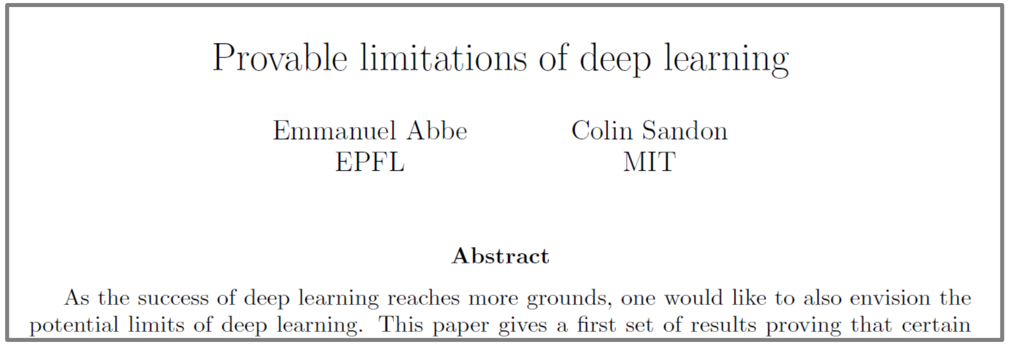
Images



Audio

Text





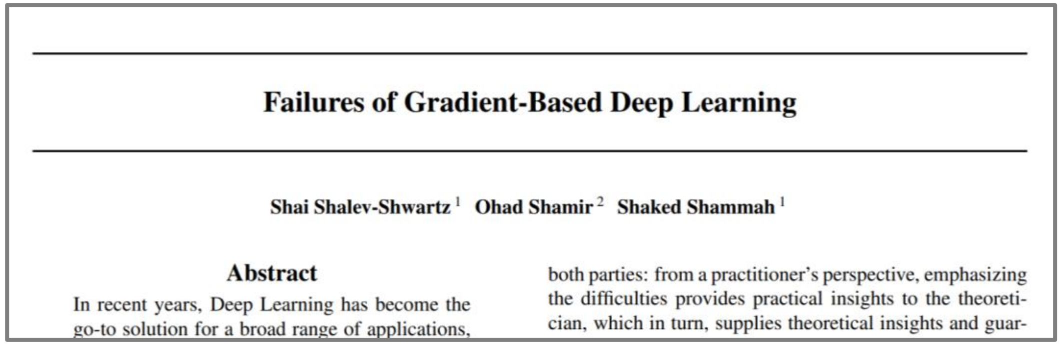
Provable limitations of deep learning

Emmanuel Abbe
EPFL

Colin Sandon
MIT

Abstract

As the success of deep learning reaches more grounds, one would like to also envision the potential limits of deep learning. This paper gives a first set of results proving that certain



Failures of Gradient-Based Deep Learning

Shai Shalev-Shwartz [1] Ohad Shamir [2] Shaked Shammah [1]

Abstract

In recent years, Deep Learning has become the go-to solution for a broad range of applications,

both parties: from a practitioner's perspective, emphasizing the difficulties provides practical insights to the theoretician, which in turn, supplies theoretical insights and guar-

What makes data suitable for deep learning?

Sources

What Makes Data Suitable for a Locally Connected Neural Network? A Necessary and Sufficient Condition Based on Quantum Entanglement

Yotam Alexander + Nimrod De La Vega + Noam Razin + **C**

arXiv

On the Ability of Graph Neural Networks to Model Interactions Between Vertices

Noam Razin + Tom Verbin + **C**

arXiv

Outline

- The Role of Data Distributions in Deep Learning
- An Appeal to Quantum Physics
- Characterization of Data Suitable for Neural Networks
- Conclusion

Statistical Learning Setup

$\mathcal{X}$ – instance space

$\mathcal{Y}$ – label space

$\mathcal{D}$ – distribution over $\mathcal{X} \times \mathcal{Y}$ (unknown)

$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ – loss function

Task

Given training set $S = \{(x_m, y_m)\}_{m=1}^M$ drawn i.i.d. from $\mathcal{D}$, return hypothesis $h : \mathcal{X} \to \mathcal{Y}$ that minimizes population loss:

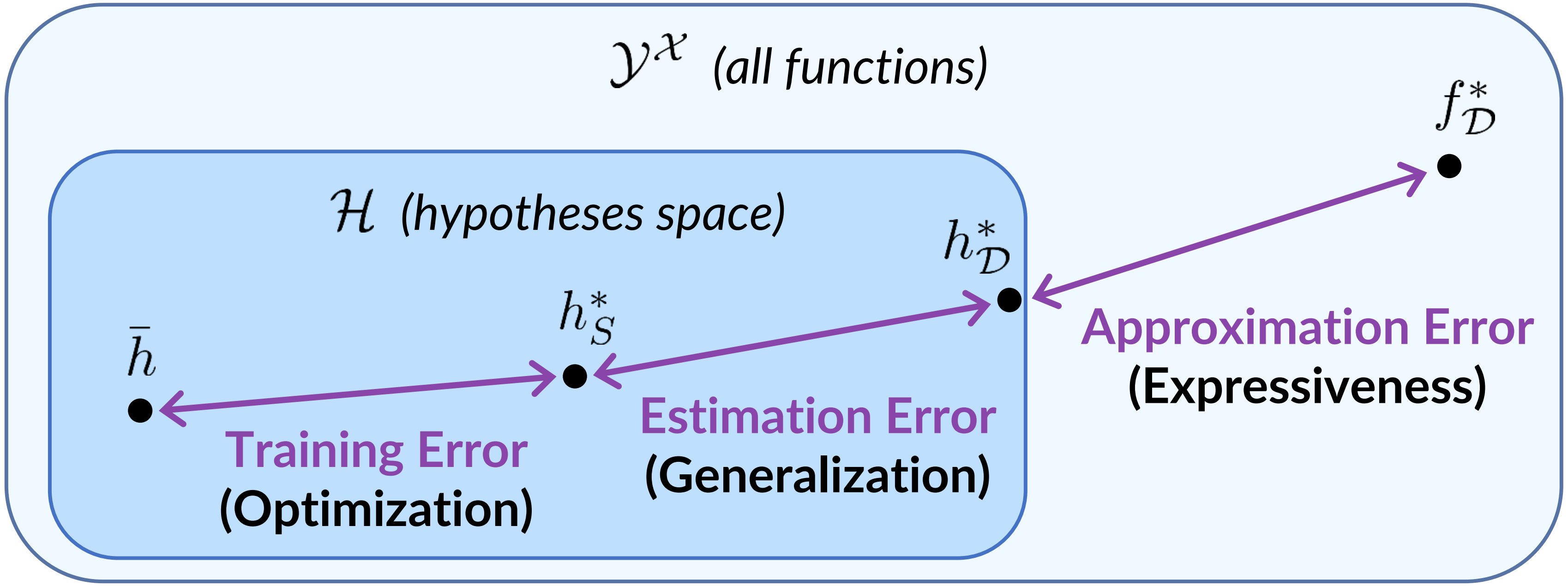
$$L_{\mathcal{D}}(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(y, h(x))]$$

Approach

Predetermine hypotheses space $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and return $h \in \mathcal{H}$ that minimizes empirical loss:

$$L_S(h) := \frac{1}{M} \sum_{m=1}^M \ell(y_m, h(x_m))$$

Three Pillars of Statistical Learning Theory: Expressiveness, Generalization and Optimization



$f^*_{\mathcal{D}}$ – ground truth (minimizer of population loss over $\mathcal{Y}^{\mathcal{X}}$)

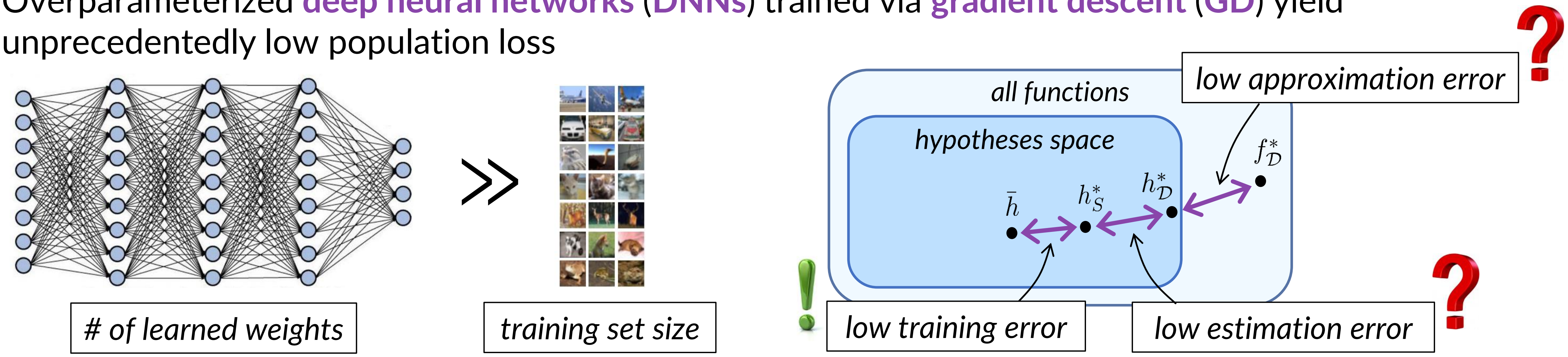
$h^*_{\mathcal{D}}$ – optimal hypothesis (minimizer of population loss over $\mathcal{H}$)

h^*_S – empirically optimal hypothesis (minimizer of empirical loss over $\mathcal{H}$)

$\bar{h}$ – returned hypotheiss

Deep Learning

Overparameterized **deep neural networks** (**DNNs**) trained via **gradient descent** (**GD**) yield unprecedentedly low population loss



Training error: GD on overparameterized DNN converges to global min with **arbitrary** data distribution (Jacot et al. 2018, Du et al. 2019, Allen-Zhu et al. 2019, Zou et al. 2020)

Approximation error: poly-sized DNN can express hypothesis with low population loss only for **some** data distributions (Telgarsky 2016, **C** et al. 2016)

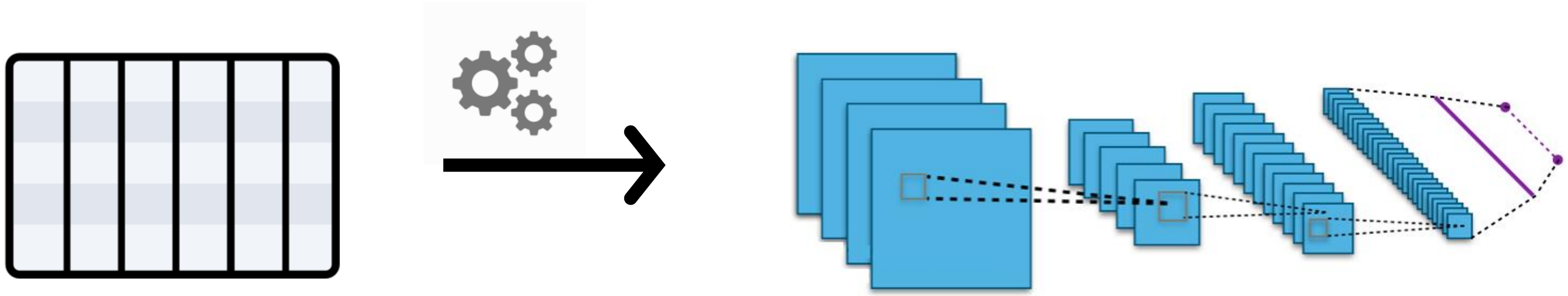
Estimation error: GD on DNN leads to optimal population loss only for **some** data distributions (Shalev-Shwartz et al. 2017, Abbe & Sandon 2018)

What makes a data distribution lead to low approximation/estimation error?

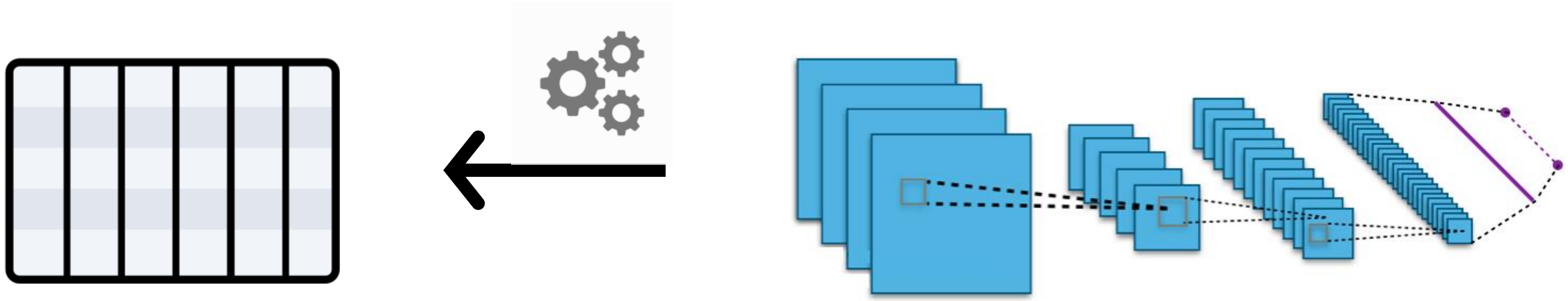
Why Study Suitability of Data for Deep Learning?

Aside from scientific curiosity, can lead to practical methods for:

- Adapting data to neural networks (NNs)

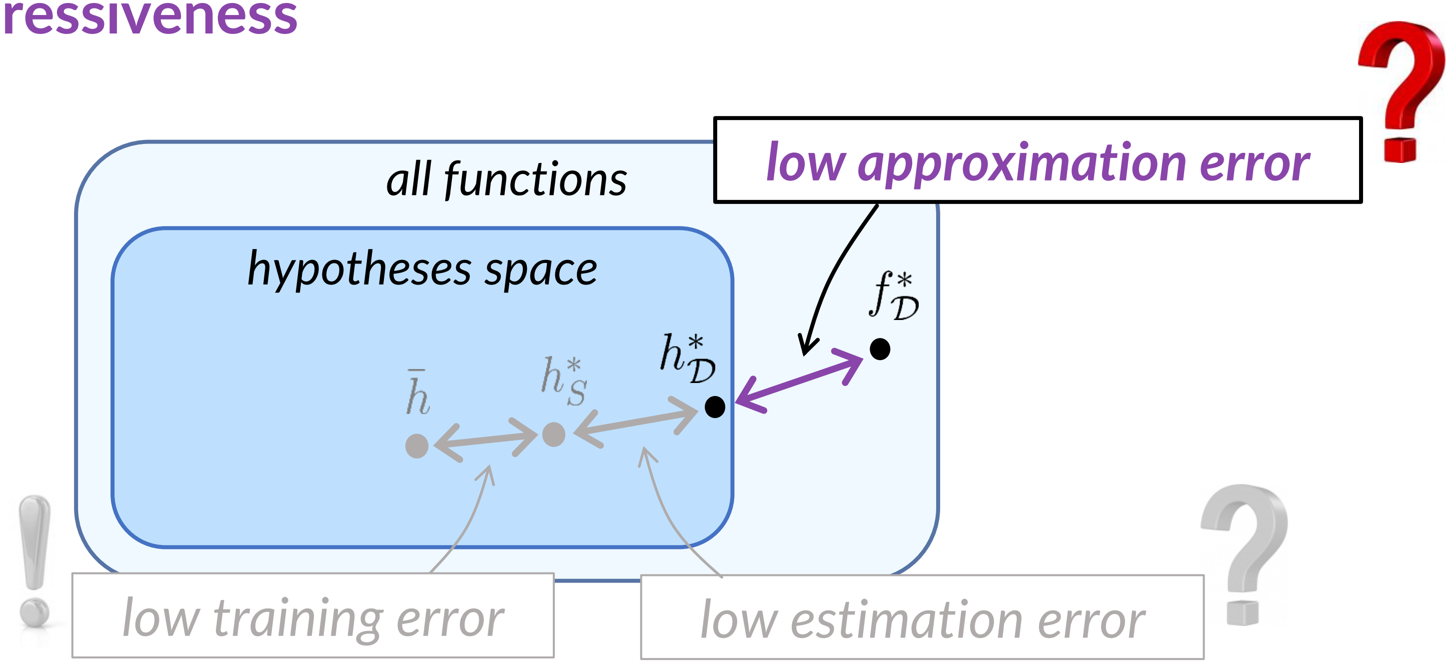


- Adapting NNs to data



Our Focus: Suitability of Data in Terms of Expressiveness

We focus on **expressiveness**



Existing literature:

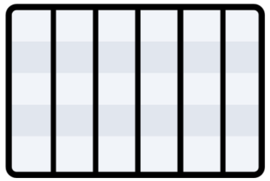
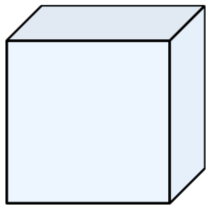
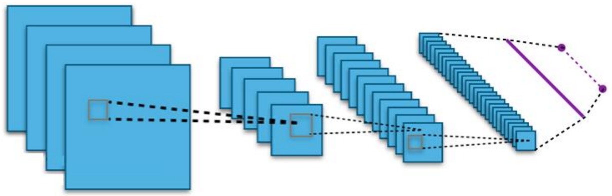
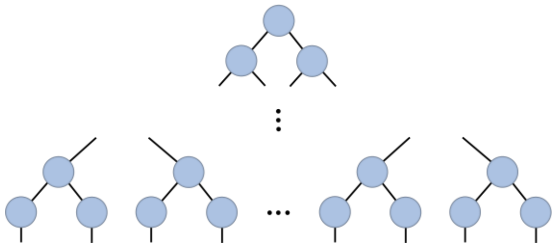
- Restrictive sufficient conditions on data distribution (folklore results, Telgarsky 2016, Zhang et al. 2017)
- Missing characterizations with **necessary and sufficient conditions**

Outline

- The Role of Data Distributions in Deep Learning
- An Appeal to Quantum Physics
- Characterization of Data Suitable for Neural Networks
- Conclusion

Quantum Physics

Discipline that also ties **distributions** with **computational models**

	Deep learning	**Quantum physics**
Distribution	 Data	 Tensor
Computational model	 NN	 Tensor network

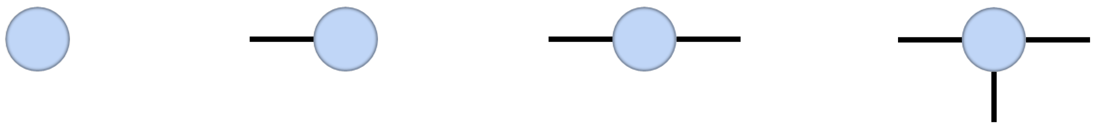
Tensors and Tensor Networks

Tensor $\mathcal{T} \in \mathbb{R}^{D_1 \times \cdots \times D_N}$ – multi-dimensional array

Tensor network (**TN**) – graph in which:

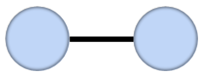
vertices $\longleftrightarrow$ tensors edges $\longleftrightarrow$ axes

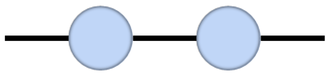
scalar *vector* *matrix* *3-dim tensor*

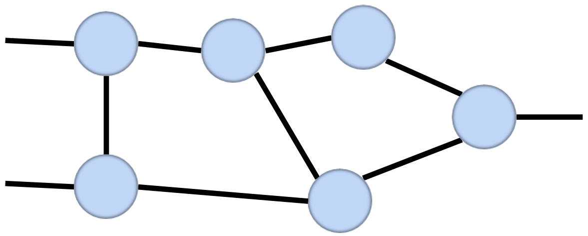


edge connecting two vertices (tensors) represents **contraction**

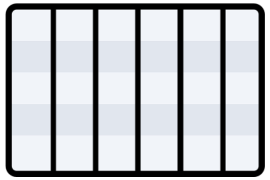
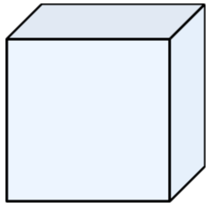
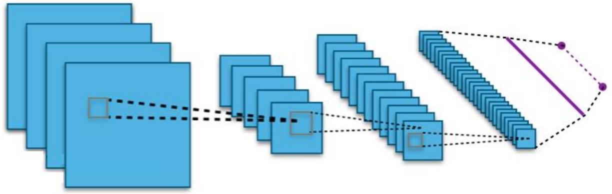
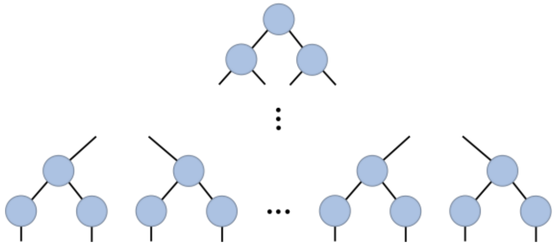
inner product *matrix multiplication*







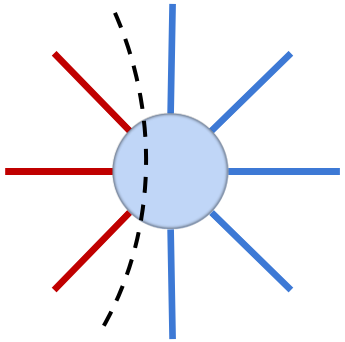
Assessing Suitability of Distribution to Computational Model

	Deep learning	Quantum physics
Distribution	 Data	 Tensor
Computational model	 NN	 TN
Theory for suitability of distribution to model	✗	✓

based on **quantum entanglement**

Quantum Entanglement

Quantifies **dependencies that tensor admits under partitions of its axes**



Let $\mathcal{T} \in \mathbb{R}^{D_1 \times \cdots \times D_N}$ and $\mathcal{I} \subseteq \{1, \ldots, N\}$

$\mathrm{mat}_{\mathcal{I}}(\mathcal{T})$ – arrangement of $\mathcal{T}$ as matrix with axes in $\mathcal{I}$ unrolled as rows

$\{\rho_d := \sigma_d^2 / \sum_{d'} \sigma_{d'}^2\}_d$ – distribution induced by singular values of $\mathrm{mat}_{\mathcal{I}}(\mathcal{T})$

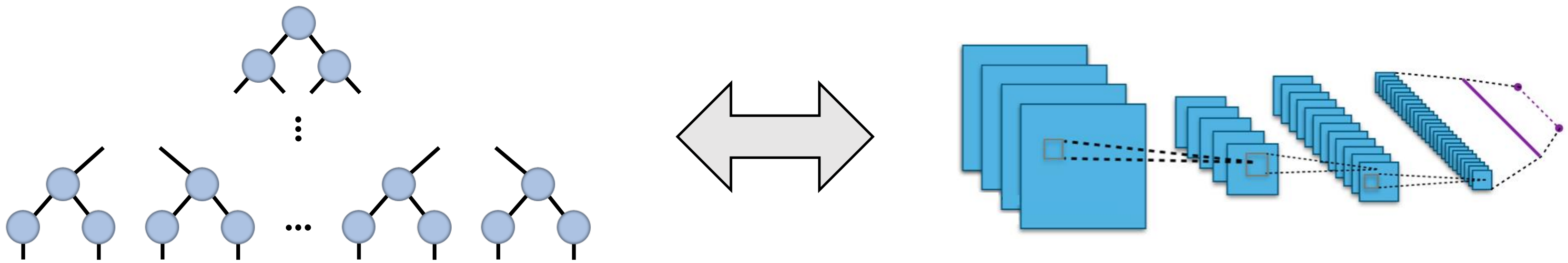
$\mathrm{QE}(\mathcal{T}; \mathcal{I}) := -\sum_d \rho_d \ln \rho_d$ (entropy of $\{\rho_d\}_d$)

Outline

- The Role of Data Distributions in Deep Learning
- An Appeal to Quantum Physics
- **Characterization of Data Suitable for Neural Networks**
- Conclusion

Neural Tensor Networks

We study TNs equivalent to NNs with multiplicative non-linearity



Why?

- The equivalent NNs are competitive empirically
  (**C** et al. 2016a, Stoudenmire 2018)
- Neural TNs enabled analyses of expressiveness and implicit regularization in deep learning
  (**C** et al. 2016b, **C** & Shashua 2017, Levine et al. 2018, Khrulkov et al. 2018, Razin et al. 2021;2022)
- The analyses led to insights and practical tools for widespread NNs
  (**C** et al. 2016b, **C** & Shashua 2017, Levine et al. 2018, Khrulkov et al. 2018, Razin et al. 2021;2022)

Data Tensor

input elements (e.g. audio samples, text tokens)

Classification Setting:

- Instance space $\mathcal{X} = \left\{ \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} \right) : \mathbf{x}^{(n)} \in \mathbb{R}^D \right\}_{m=1}^{M}$
- Label space $\mathcal{Y} = \{+1, -1\}$

Given training set $S = \left\{ \left(\left(\mathbf{x}_m^{(1)}, \ldots, \mathbf{x}_m^{(N)} \right), y_m \right) \right\}_{m=1}^{M}$, define:

Data tensor: $\mathcal{D} := \frac{1}{M} \sum_{m=1}^{M} y_m \cdot \mathbf{x}_m^{(1)} \otimes \cdots \otimes \mathbf{x}_m^{(N)}$

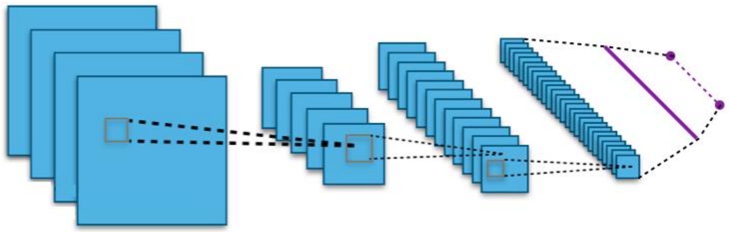
axes correspond to features

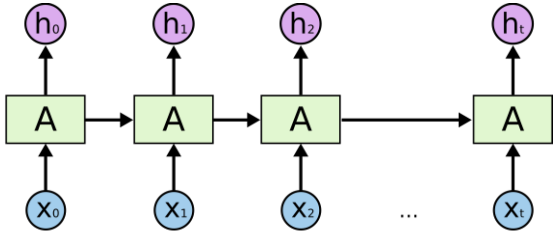
outer product

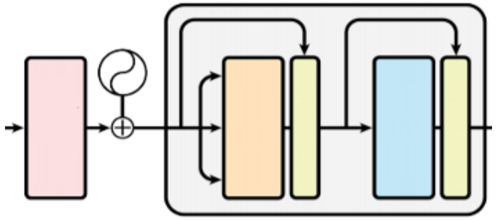
Each subset of features $\mathcal{I}$ induces a quantum entanglement $\mathrm{QE}(\mathcal{D}; \mathcal{I}) \in \mathbb{R}_{\geq 0}$

can be computed efficiently

Locally Connected Neural Networks (Theorem)



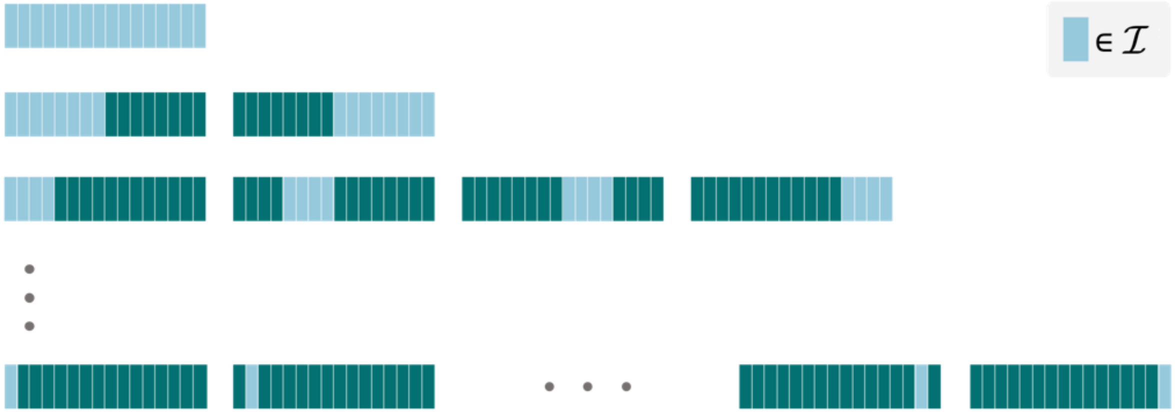




Theorem (informally stated)

A locally connected NN has low approximation error **if and only if** $\mathrm{QE}(\mathcal{D};\mathcal{I})$ is low for every canonical subset of features $\mathcal{I}$

Canonical subsets of $\{1,\ldots,N\}$:



Locally Connected Neural Networks (Proof Sketch)

Theorem (informally stated)
A locally connected NN has low approximation error **if and only if** $\mathrm{QE}(\mathcal{D};\mathcal{I})$ is low for every canonical subset of features $\mathcal{I}$

Proof Sketch

NN has low approximation error $\iff$ equivalent TN can fit expected data tensor $\mathbb{E}[\mathcal{D}]$

Quantum physics theory:

TN equivalent to locally connected NN can fit $\mathbb{E}[\mathcal{D}]$

$\iff$ $\mathrm{QE}(\mathbb{E}[\mathcal{D}];\mathcal{I})$ is low for every canonical subset $\mathcal{I}$

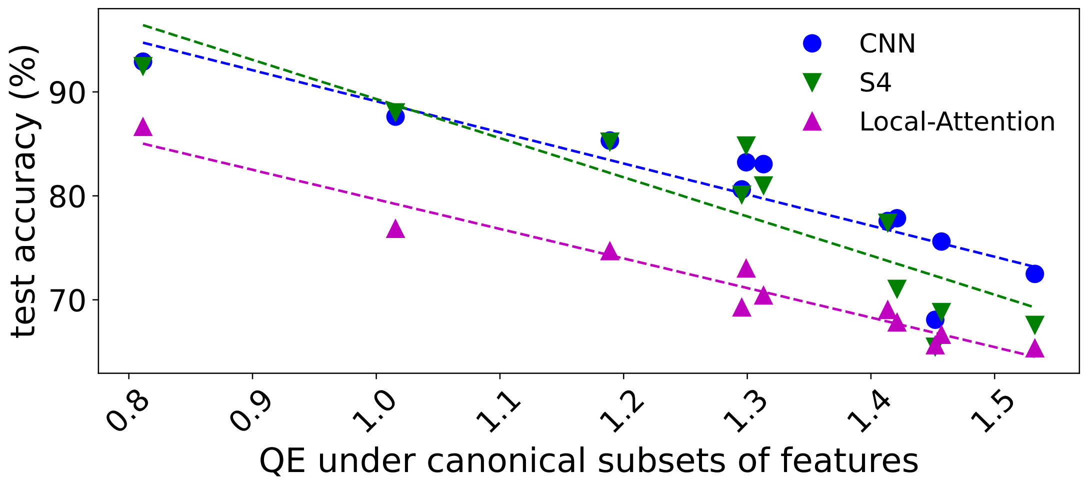
W.h.p. $\mathcal{D} \approx \mathbb{E}[\mathcal{D}] \implies \mathrm{QE}(\mathcal{D};\mathcal{I}) \approx \mathrm{QE}(\mathbb{E}[\mathcal{D}];\mathcal{I})$ for every subset $\mathcal{I}$

Locally Connected Neural Networks (Experiments)

Theorem (informally stated)

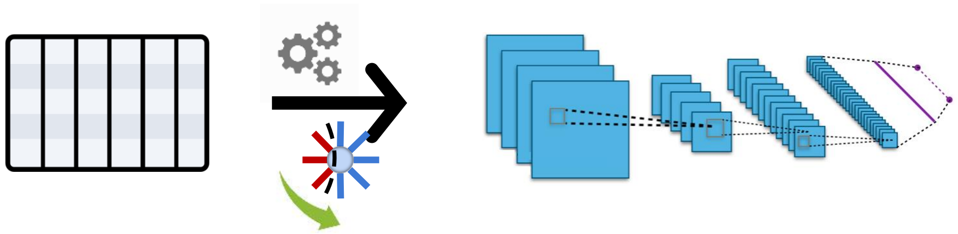
A locally connected NN has low approximation error **if and only if** $\mathrm{QE}(\mathcal{D}; \mathcal{I})$ is low for every canonical subset of features $\mathcal{I}$

Empirical Demonstration

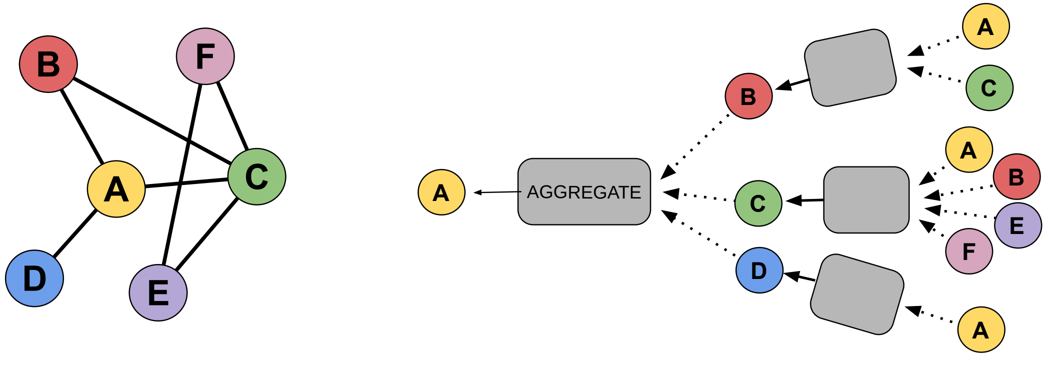


Practical Application

Improving accuracy of locally connected NNs by arranging features such that $\mathrm{QE}(\mathcal{D}; \mathcal{I})$ is low for all canonical subsets $\mathcal{I}$



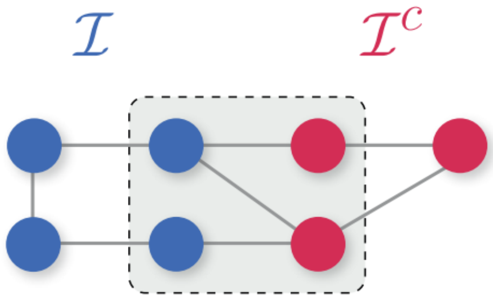
Graph Neural Networks (Theorem)

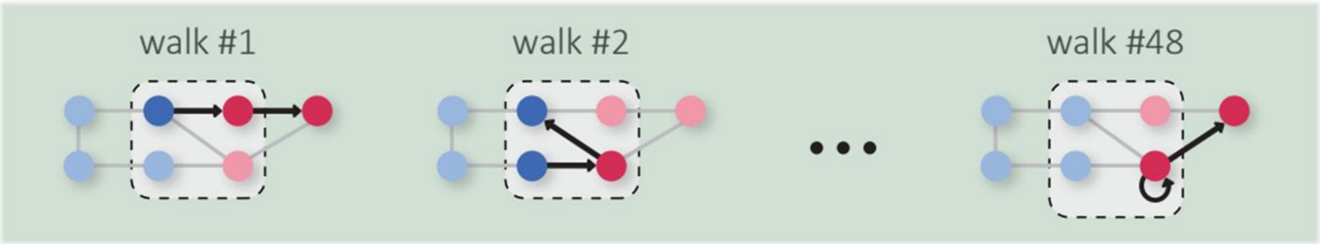


Theorem (informally stated)

If a graph NN has low approximation error, then for each subset of features $\mathcal{I}$: $\mathrm{QE}(\mathcal{D}; \mathcal{I}) = \mathcal{O}(\text{walk-index}(\mathcal{I}))$

# of walks in input graph emanating from boundary of $\mathcal{I}$





Graph Neural Networks (Proof Sketch)

Theorem (informally stated)

If a graph NN has low approximation error, then for each subset of features $\mathcal{I}$: $\mathrm{QE}(\mathcal{D};\mathcal{I}) = \mathcal{O}(\text{walk-index}(\mathcal{I}))$

Proof Sketch

NN has low approximation error $\iff$ equivalent TN can fit expected data tensor $\mathbb{E}[\mathcal{D}]$

Quantum physics theory:

TN equivalent to graph NN can fit $\mathbb{E}[\mathcal{D}]$

$\implies$ $\mathrm{QE}(\mathbb{E}[\mathcal{D}];\mathcal{I}) = \mathcal{O}(\text{walk-index}(\mathcal{I}))$ for every subset $\mathcal{I}$

W.h.p. $\mathcal{D} \approx \mathbb{E}[\mathcal{D}]$ $\implies$ $\mathrm{QE}(\mathcal{D};\mathcal{I}) \approx \mathrm{QE}(\mathbb{E}[\mathcal{D}];\mathcal{I})$ for every subset $\mathcal{I}$

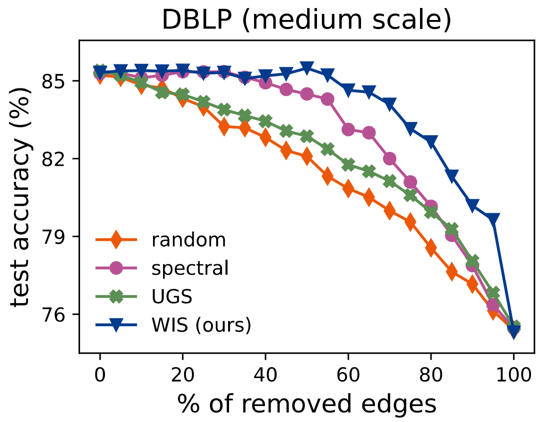
Graph Neural Networks (Experiments)

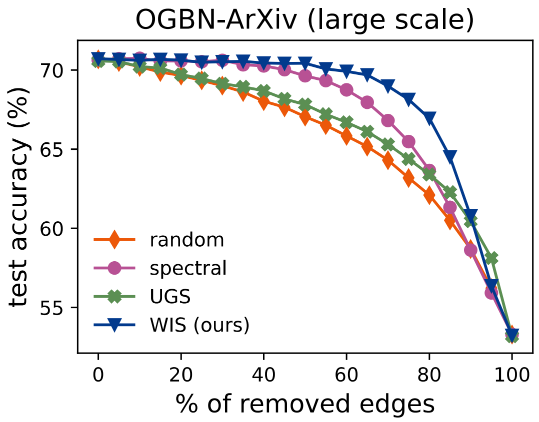
Theorem (informally stated)
If a graph NN has low approximation error, then for each subset of features $\mathcal{I}$: $\mathrm{QE}(\mathcal{D};\mathcal{I}) = \mathcal{O}(\text{walk-index}(\mathcal{I}))$

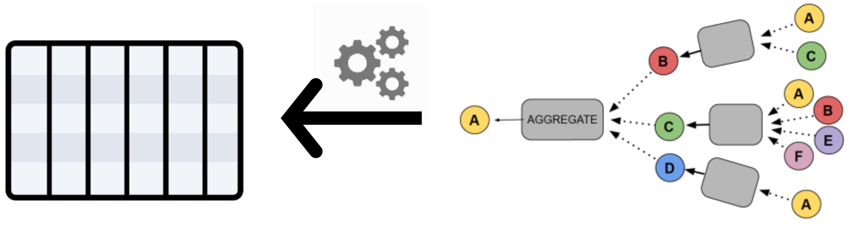
Practical Application

Algorithm for edge sparsification that preserves accuracy of graph NN:

- Select subsets $\mathcal{I}$ for which $\mathrm{QE}(\mathcal{D};\mathcal{I})$ is high compared to $\text{walk-index}(\mathcal{I})$
- Prune edge whose removal reduces $\text{walk-index}(\mathcal{I})$ for selected $\mathcal{I}$ the least







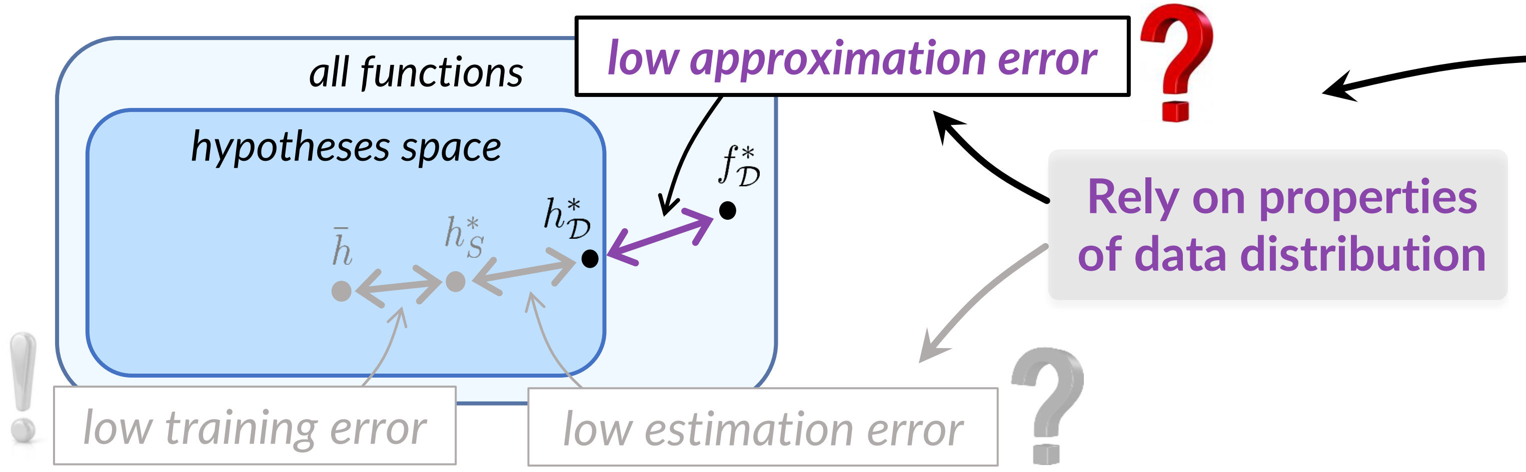
Code

Outline

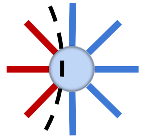
- The Role of Data Distributions in Deep Learning
- An Appeal to Quantum Physics
- Characterization of Data Suitable for Neural Networks
- **Conclusion**

Recap

Overparameterized DNNs trained via GD yield unprecedently low population loss



We study this via tools from **quantum physics**



Locally Connected NNs

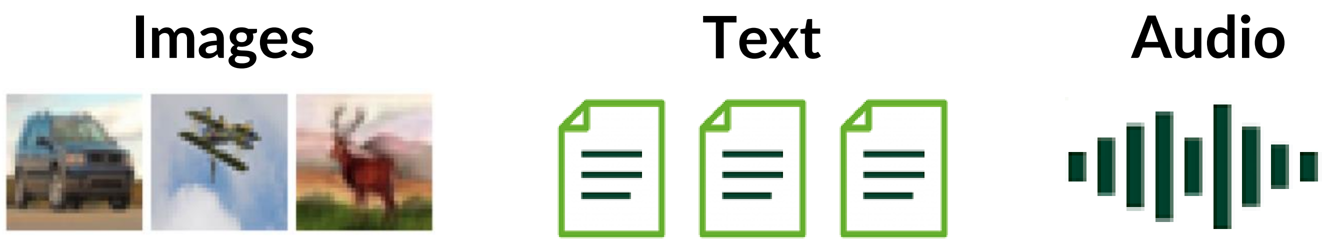
- **Theory:** accurate prediction is possible **if and only if** data admits low entanglement under canonical subsets
- **Practical application:** enhancing suitability of data via feature arrangement

GNNs

- **Theory:** accurate prediction is possible only if walk indices surpass entanglements
- **Practical application:** sparsifying architectures (input graphs) according to data

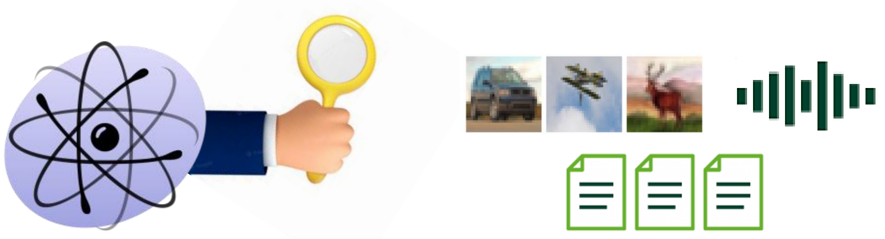
Reasoning About Natural Data via Physics

Deep learning is most commonly applied to data modalities regarded as **natural**



Difficult to formalize since we **lack tools for reasoning about natural data**

Hypothesis: physics will be key to overcoming this difficulty



Thank You!

Work supported by:
Google Research Scholar Award, Google Research Gift, the Yandex Initiative in Machine Learning, the Israel Science Foundation (grant 1780/21), Len Blavatnik and the Blavatnik Family Foundation, Tel Aviv University Center for AI and Data Science, Amnon and Anat Shashua, and Apple scholars in AI/ML PhD fellowship