Interpreting Artificial Neural Networks in the Context of Theoretical Physics

Sebastian Johann Wetzel

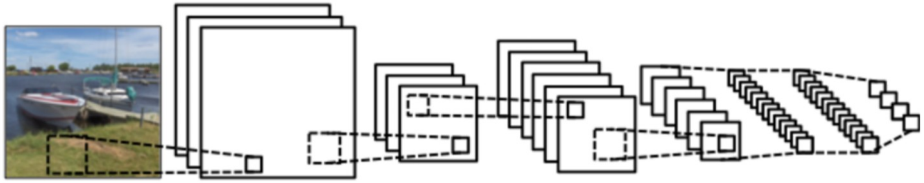
PERIMETER INSTITUTE

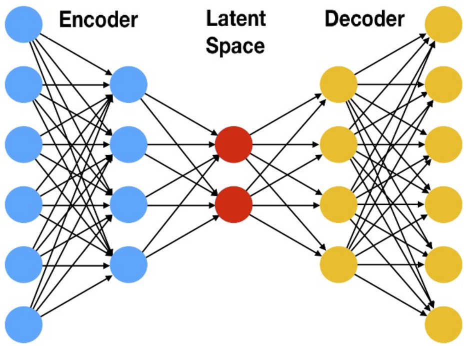
Canada NRC·CNRC

PIQUIL

Success of Artifical Neural Networks

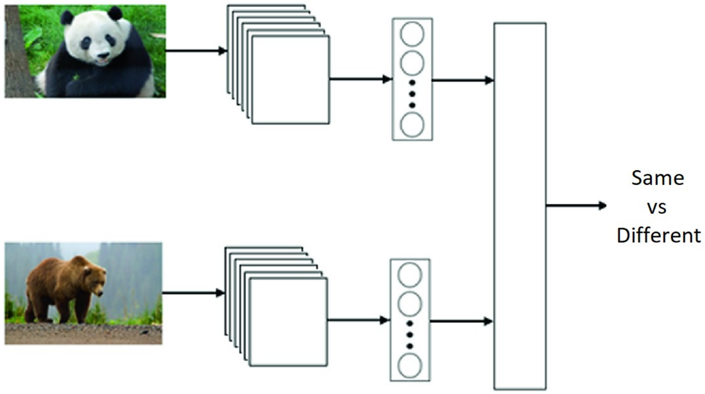
Image Classification
(Convolutional Network)





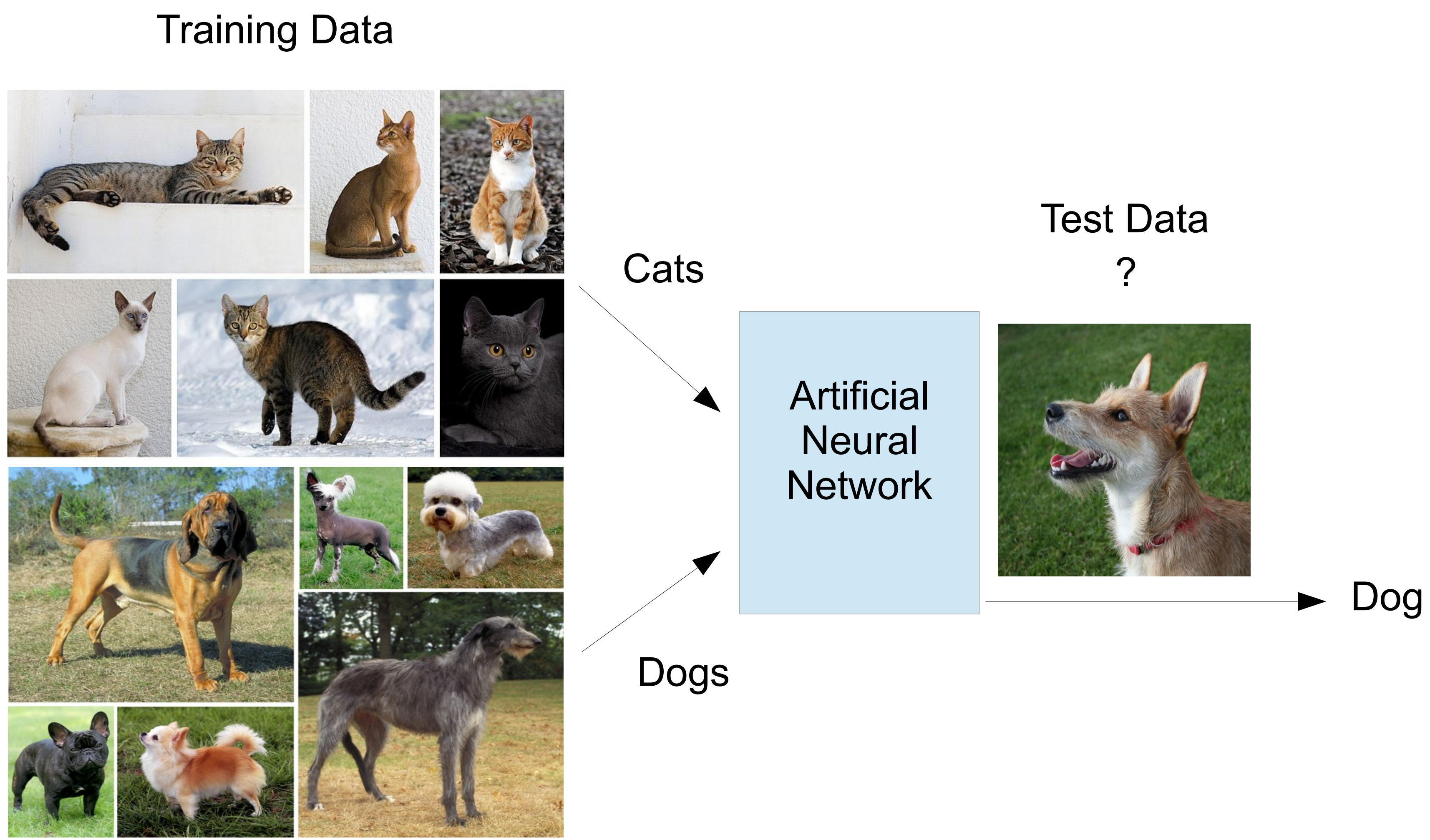
Generative Modelling /
Anomaly Detection
(Autoencoders)

Similarity Detection
(Siamese Network)



(Supervised) Machine Learning with Neural Nets

„Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.“ - Wikipedia



(Supervised) Machine Learning with Neural Nets

„Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.“ - Wikipedia

1) What does the Neural Network actually learn?
2) Can this Knowledge help in Scientific Discovery?

Training Data

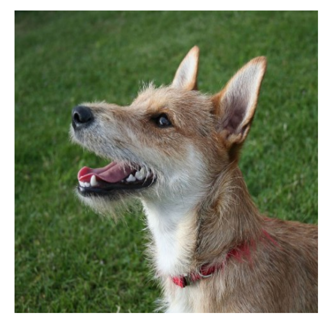
Cats

Dogs

Artificial Neural Network

Test Data

?



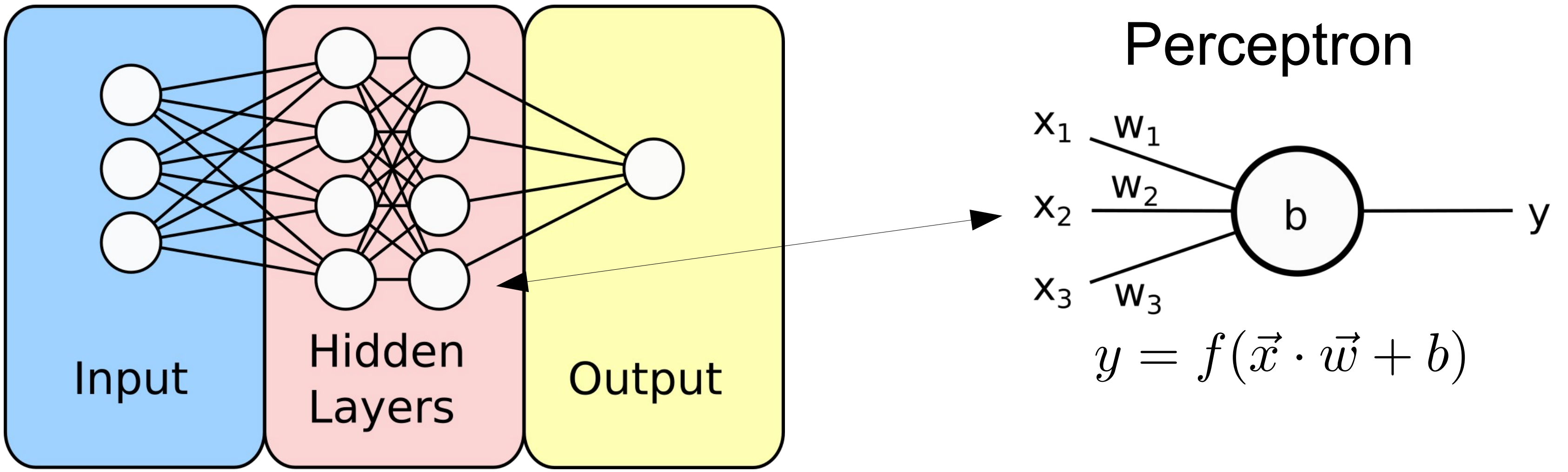
Dog

Overview

- ✗ Artificial Neural Networks
- ✗ Interpretation Techniques
- ✗ Interpretation of Convolutional Neural Networks
- ✗ Interpretation of Autoencoders
- ✗ Interpretation of Siamese Networks

Artificial Neural Networks

Feed forward neural network



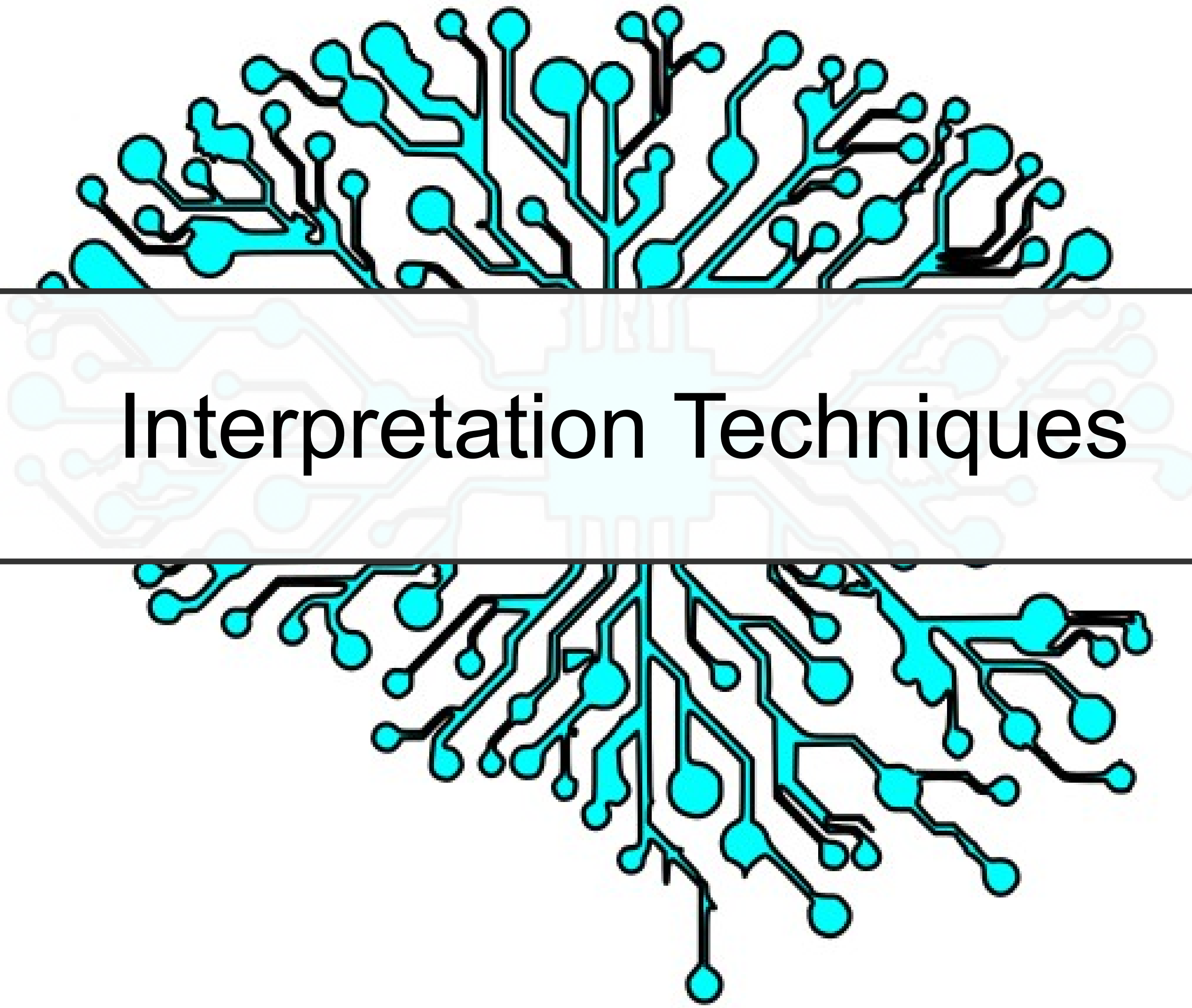
$$y = f(\vec{x} \cdot \vec{w} + b)$$

Input: Data $X = (\vec{x}_1, ..., \vec{x}_n)$, Label $Y = (y_1, ..., y_n)$

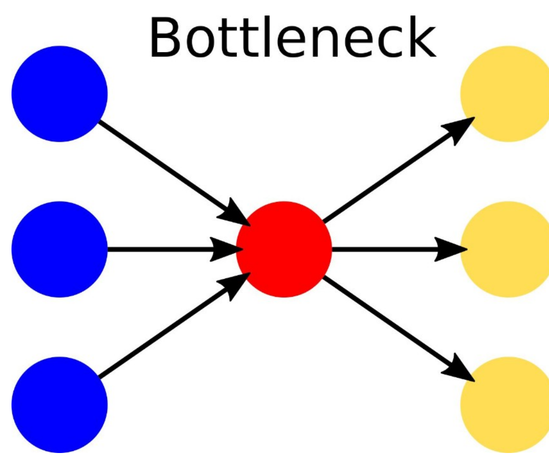
Output: $Y_{pred} = F(X, w_{ij}^L, b_i^L)$

Goal: choose w_{ij}^L and b_i^L such that $Y_{pred} \approx Y$

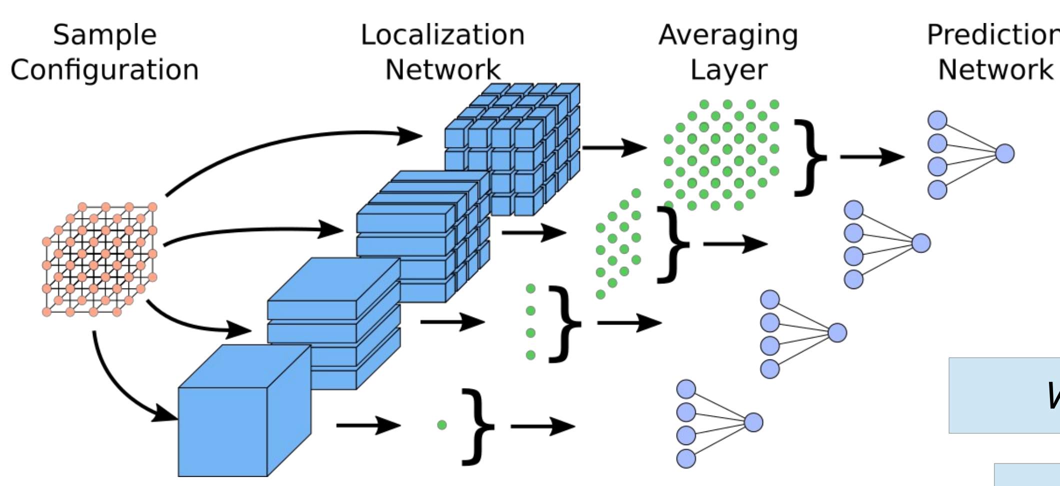
Interpretation Techniques



Bottleneck Interpretation
+Correlation Probing Neural Network



+



Wetzel, PRE 2017

Wetzel, Scherzer, PRB 2017

Wetzel, et al, PRR 2020

Looking at the weights

- ✗ No, works but only for the most simple problems.

Influence Functions

Phase Detection with Neural Networks: Interpreting the Black Box

Anna Dawid,[1,2] Patrick Huembeli,[2] Michał Tomza,[1] Maciej Lewenstein,[2,3] and Alexandre Dauphin[2]

- ✗ Remove specific datapoints or features and measure the effect on the performance
- ✗ Largest change in performance indicates the most influential data point or feature

Dawid et al, New Journal of Pyhsics 2020

Dark Matter

Discovering Symbolic Models from Deep Learning with Inductive Biases

Miles Cranmer[1] Alvaro Sanchez-Gonzalez[2] Peter Battaglia[2] Rui Xu[1]

Kyle Cranmer[3] David Spergel[4,1] Shirley Ho[4,3,1,5]

- ✗ Simulate Dark Matter
- ✗ Apply symbolic regression at the output of a graph neural network to recover force equation

Cranmer et al., Neurips 2020

Condensed Matter+Correlator Network

Correlator Convolutional Neural Networks: An Interpretable Architecture for Image-like Quantum Matter Data

Cole Miles,[1] Annabelle Bohrdt,[2,3,4] Ruihan Wu,[5] Christie Chiu,[2,6,7] Muqing Xu,[2] Geoffrey Ji,[2] Markus Greiner,[2] Kilian Q. Weinberger,[5] Eugene Demler,[2] and Eun-Ah Kim[1]

- ✗ Explicit feature engineering layer that probes for correlations
- ✗ Dominant features correspond to dominant correlations in condensed matter system

Miles et al., Arxiv 2020

Physical Concepts

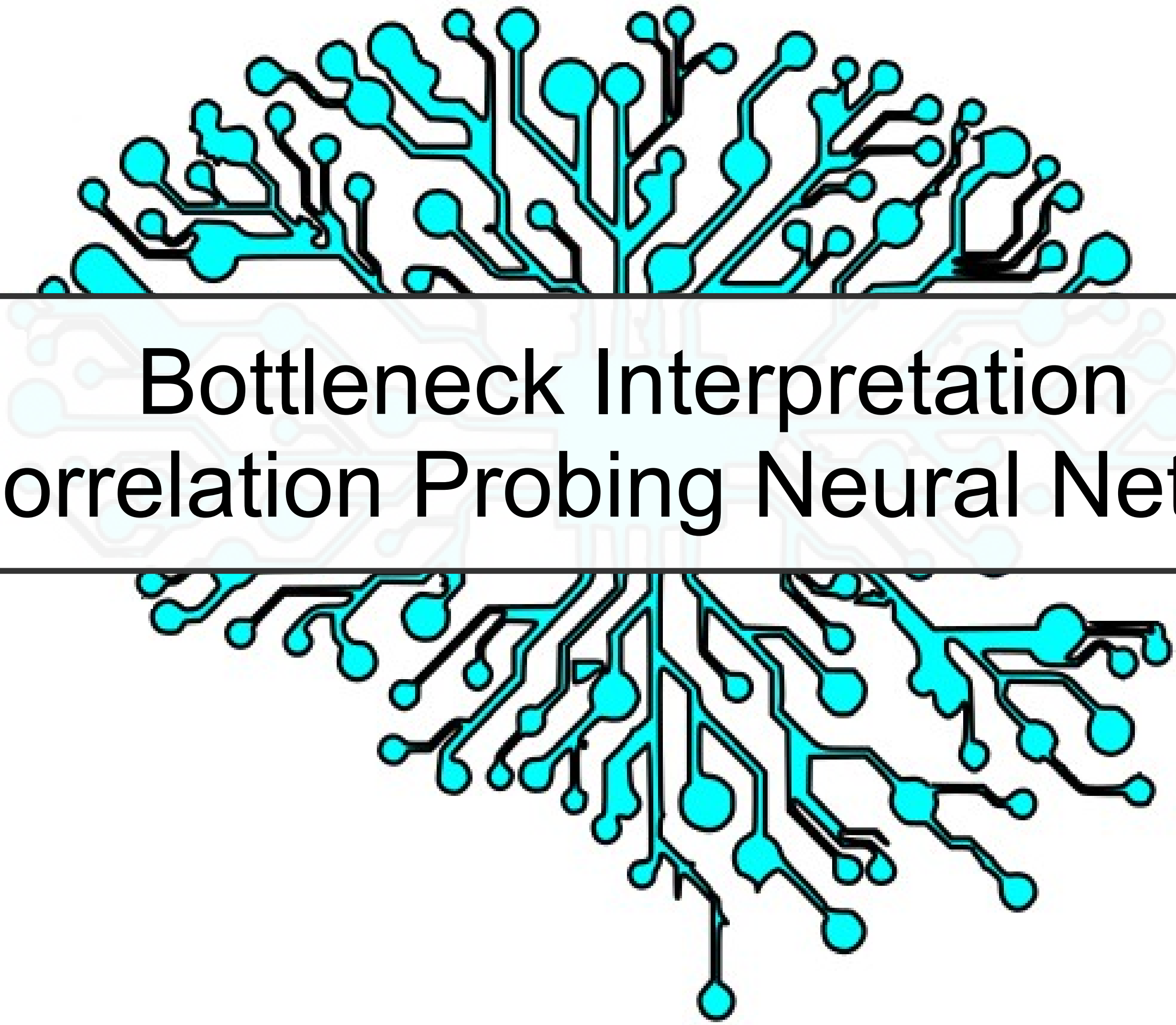
Discovering physical concepts with neural networks

Raban Iten,* Tony Metger,* Henrik Wilming, Lídia del Rio, and Renato Renner
ETH Zürich, Wolfgang-Pauli-Str. 27, 8093 Zürich, Switzerland.
(Dated: January 24, 2020)

- ✗ Interpretation of autoencoder latent representation
- ✗ Ask physical questions to be extractable from latent space

Iten et al., PRL 2020

Bottleneck Interpretation
+Correlation Probing Neural Network



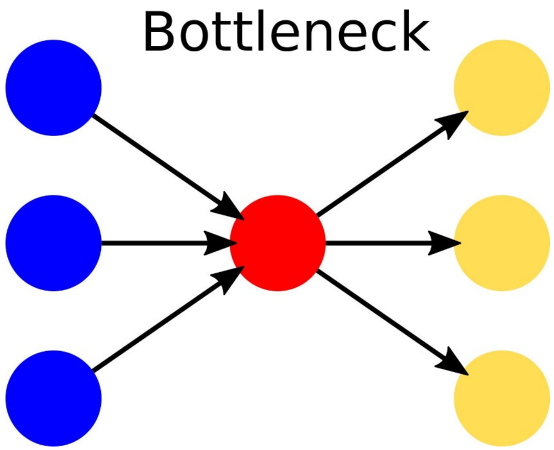
Bottleneck Interpretation

Interpretation is often difficult since information is spread over several neurons and layers

If the neuron contains the information of <u>one</u> single quantity/obervable $Q(S)$ → The output of the neuron can be mapped via a bijective function to the observable

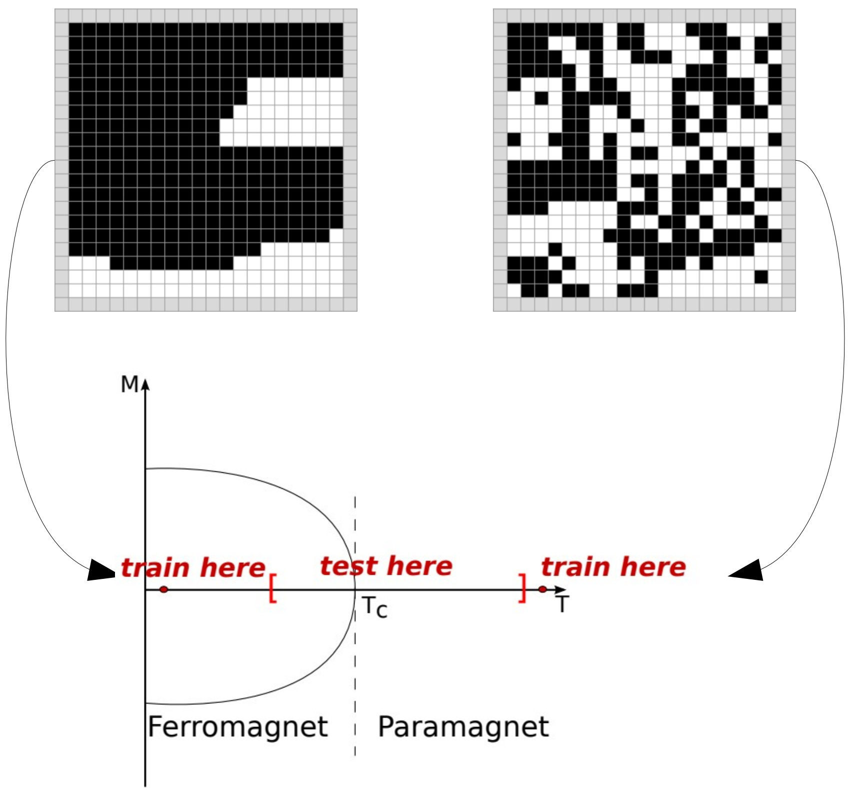
$$F(S) = f(Q(S))$$

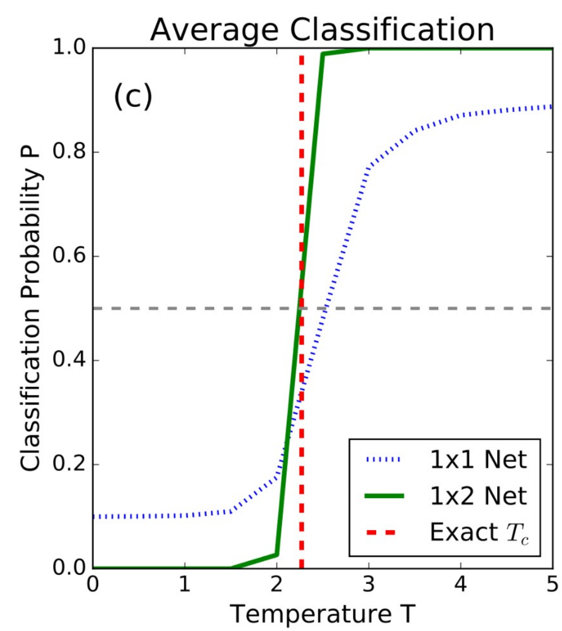
- Idea: identify or enforce bottlenecks in the network
- Perform regression on the output of the bottleneck neuron



Supervised Learning
2d Ising Model

- Data: Monte Carlo samples
- Training at well known points in phase diagram
- Labels: Phase
- Testing in interval containing phase transition
- Estimate within 1% of exact value $T_c = \dfrac{2}{\ln(1+\sqrt{2})}$





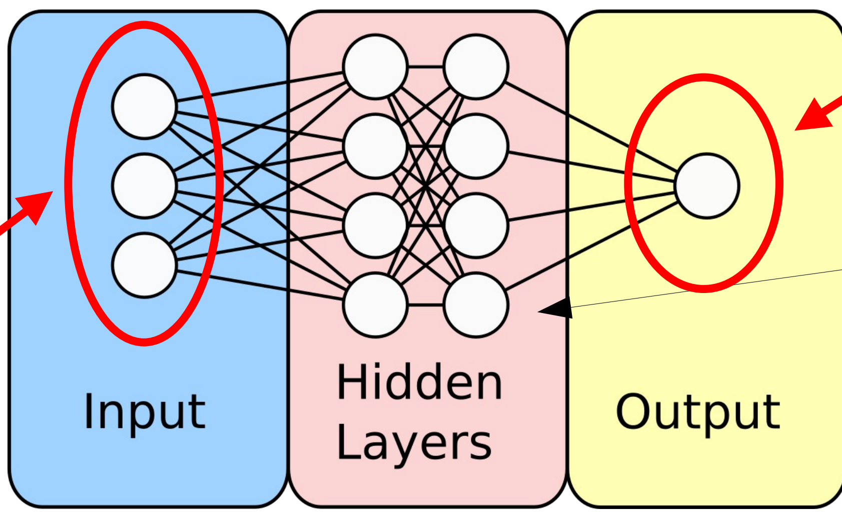
Carrasquila, Melko, Nature 2017

Artificial Neural Networks

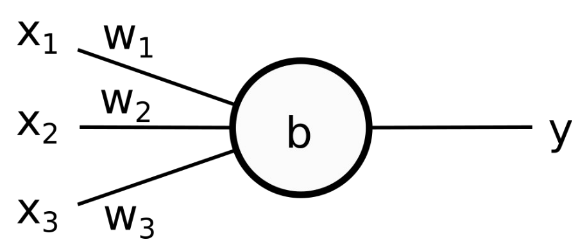
Too Many Features for Regression

Feed forward neural network

Natural Bottleneck



Perceptron



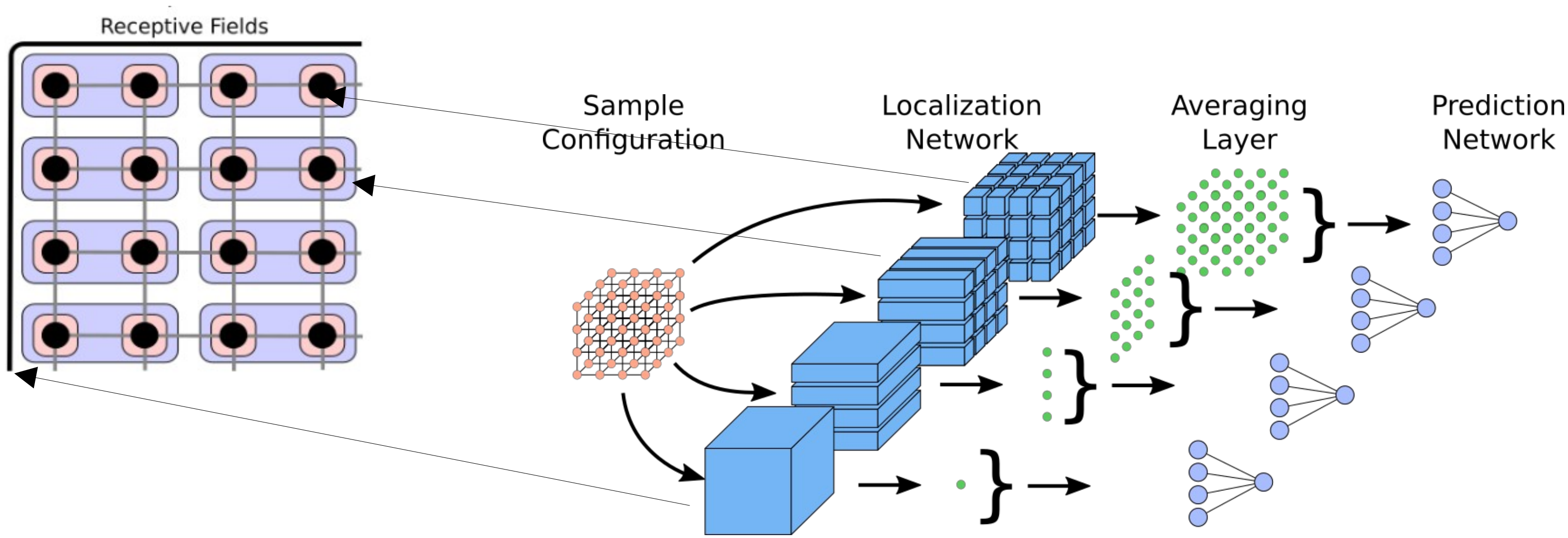
$y = f(\vec{x} \cdot \vec{w} + b)$

Input: Data $X = (\vec{x}_1, \ldots, \vec{x}_n)$, Label $Y = (y_1, \ldots, y_n)$

Output: $Y_{pred} = F(X, w_{ij}^L, b_i^L)$

Goal: choose w_{ij}^L and b_i^L such that $Y_{pred} \approx Y$

Interpretation of Neural Network
2d Ising Model



Correlation Probing Network

Wetzel, Scherzer, PRB 2017

- Correlation Probing Net interpolates between a general NN and a minimal optimal NN which has the same performance
- Interpretation by reducing the NN capacity in an ordered manner until one observes a performance drop
- Inspired by intensive/extensive quantities (averaging layer probes for translational invariance of the quantity $Q(S)$)

Interpretation of Neural Network
2d Ising Model

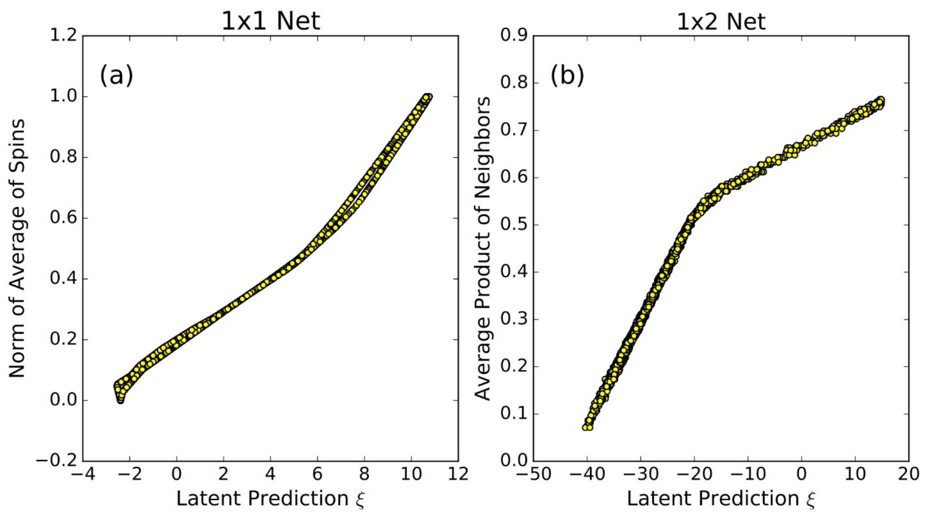
Receptive Field Size	Train Loss	Validation Loss
28×28	$6.1588e-04$	0.0232
1×2	**1.2559e-04**	**1.2105e-07**
1×1	**0.2015**	**0.1886**
baseline	0.6931	0.6931

Decision functions

$$F(S) = \mathrm{sigmoid}(w\,Q(S) + b)$$

- $Q(S) = |1/N \sum_i s_i|$ Magnetization
- $Q(S) = \frac{1}{N} \sum_{<i,j>nn} s_i s_j$ Expected Energy per site

Deduction visually confirmed:

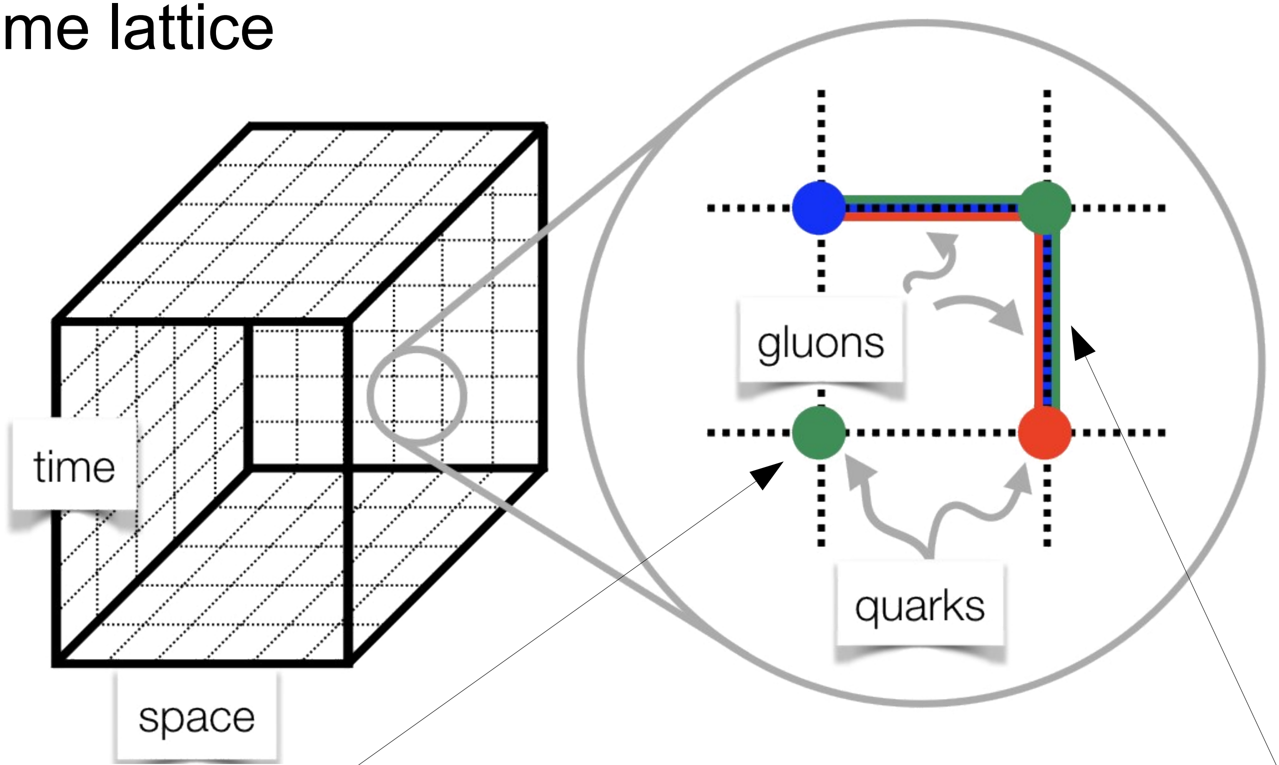


Note:

1x2 Network also has
the Magnetization minimum
which is easier to find!

SU(2) Lattice Gauge Theory

Space time lattice



Quarks on heavy static lattice sites.

Gluons on the connections between lattice sites are described by Matrices

$$\cancel{U_\mu^x \in SU(3)}$$

$$U_\mu^x \in SU(2)$$

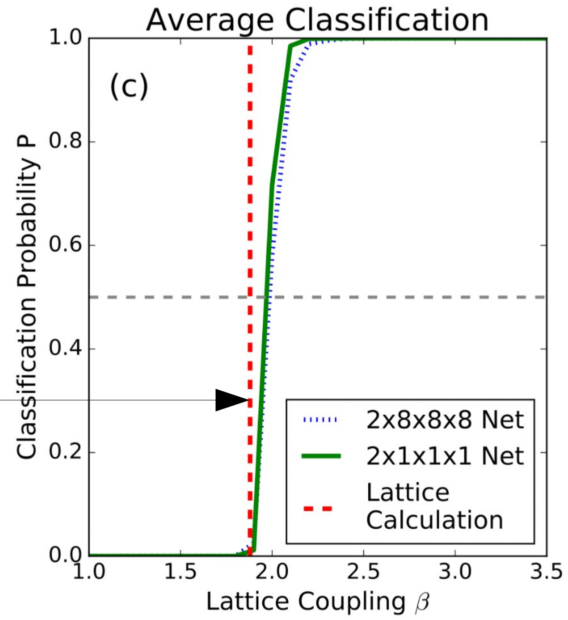
Supervised Learning
SU(2) Lattice Gauge Theory

Data: Monte Carlo samples

$$S_{\text{Wilson}}[U] = \beta_{\text{latt}} \sum_x \sum_{\mu<\nu} \text{Re tr}\left(1 - U^x_{\mu\nu}\right)$$

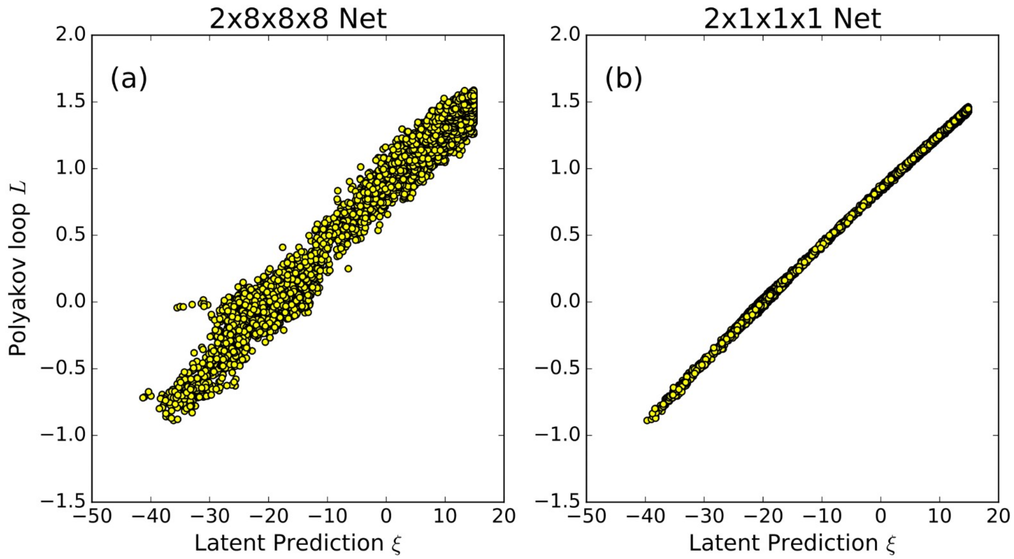
- Training at well known points in phase diagram
- Labels: Phase

Find phase transition close to lattice calculation



Interpretation of Neural Network
SU(2) Gauge Theory

Deduction confirmed by perfect correlation between NN output and Polyakov Loop order parameter

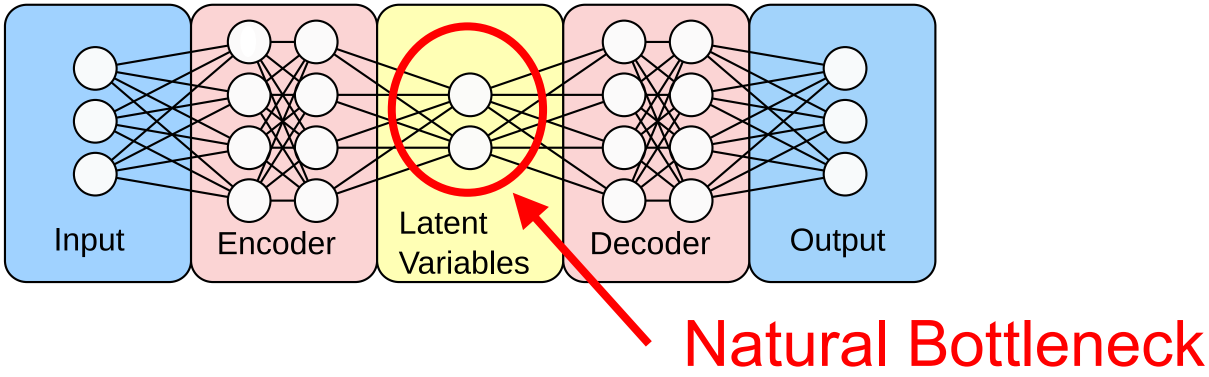


$$F(S) \approx \mathrm{sigmoid}\left(w \left(\frac{2}{N} \sum_{\vec{x}} f(\{U_\mu^{x_0,\vec{x}}\}) \right) + b \right)$$

$$f(\{U_\mu^{x_0}\}) = a_\tau^0 a_\tau^1 - b_\tau^0 b_\tau^1 - c_\tau^0 c_\tau^1 - d_\tau^0 d_\tau^1 = \mathrm{tr}\left(U_\tau^0 U_\tau^1\right)$$

- Polyakov Loop

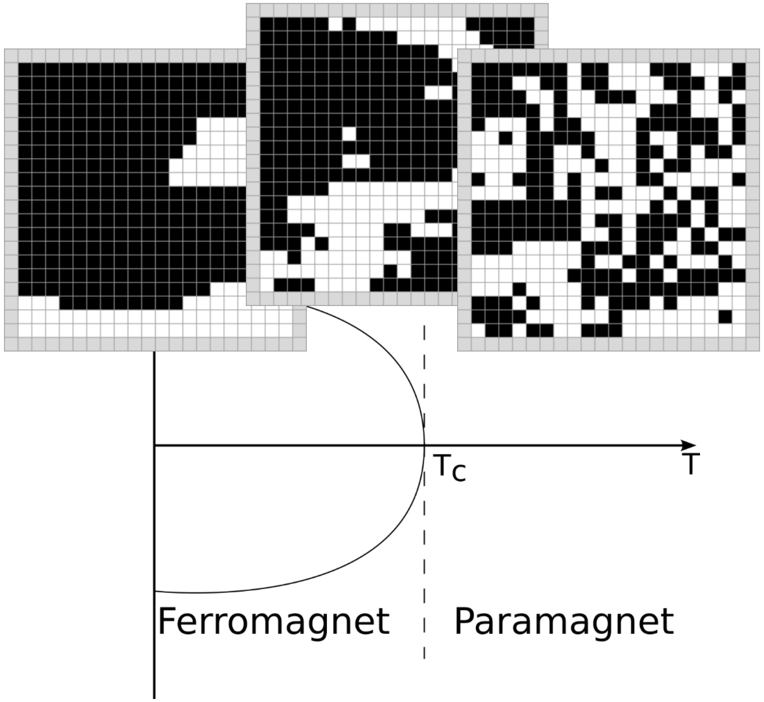
(Variational) Autoencoder
2d Ising Model



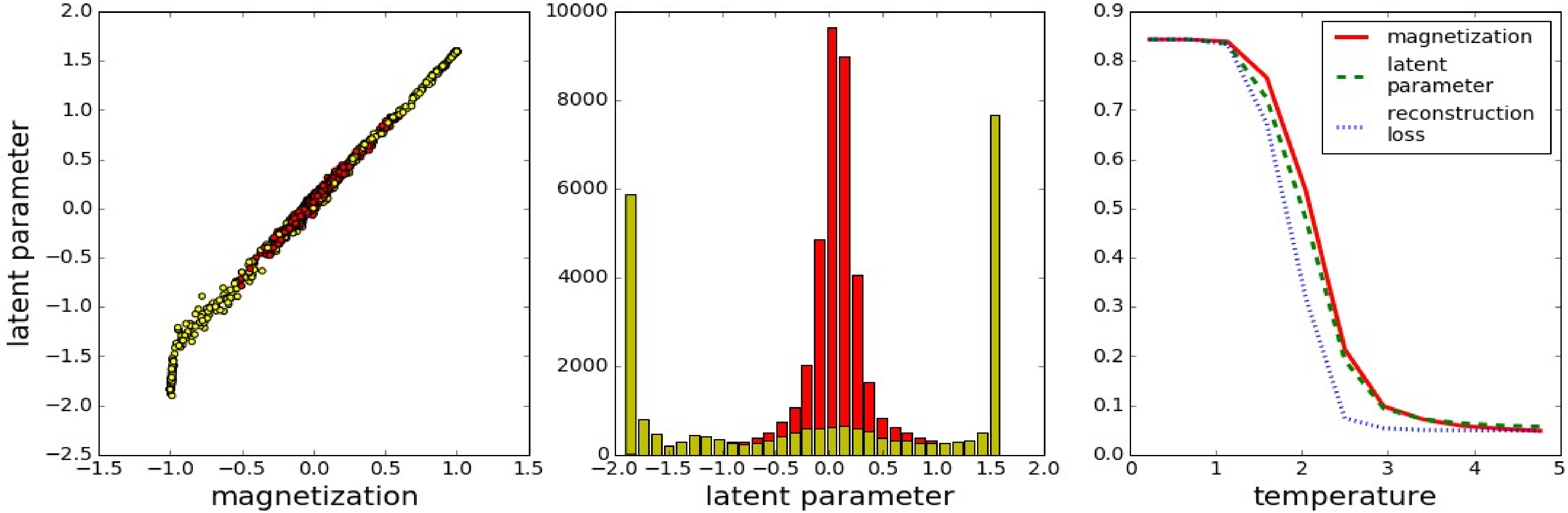
Objective: Minimize Reconstruction error

$$MSE = \frac{1}{N}\sum_k \left\| x_k - F(x_k) \right\|^2$$

- Data: Monte Carlo samples
- Train everywhere in phase diagram
- Labels: None



(Variational) Autoencoder
2d Ising Model

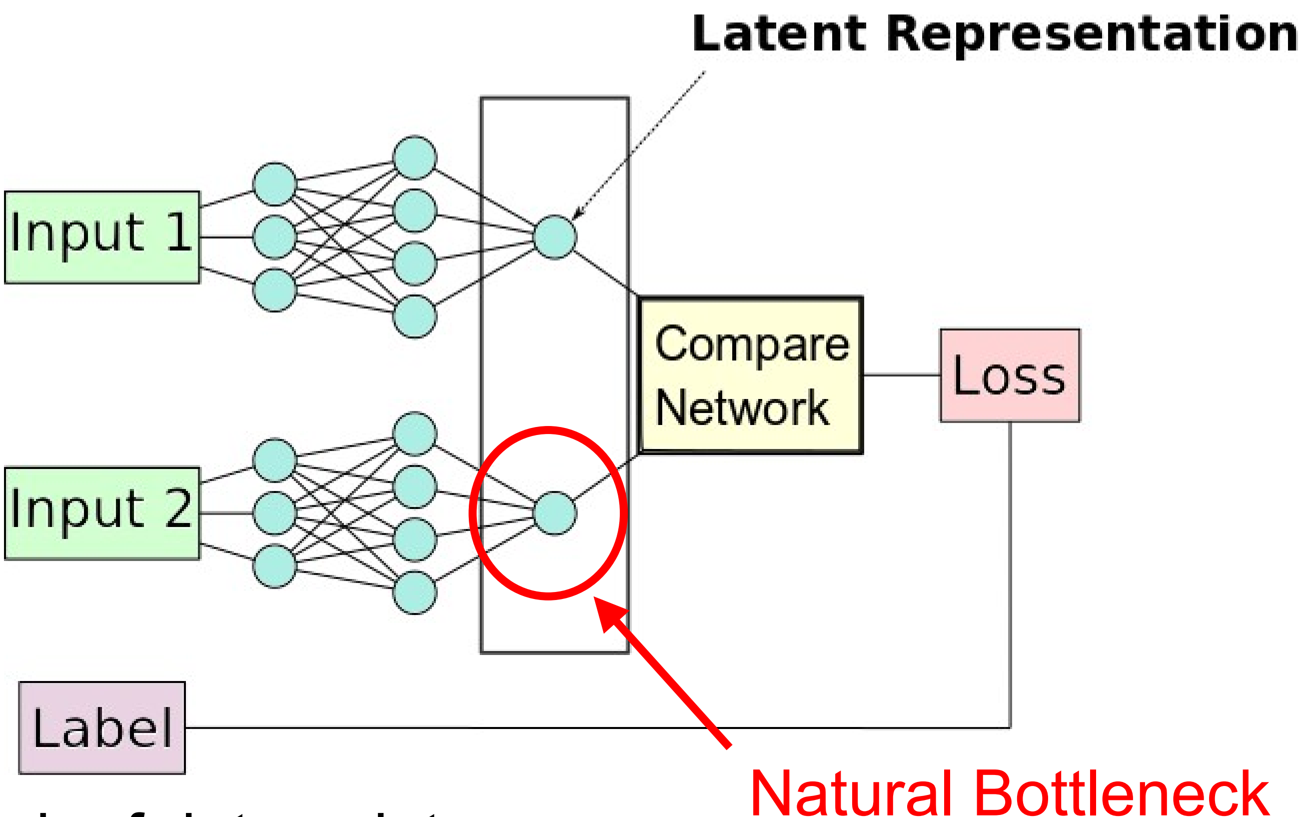


Ferromagnetic Ising model on the square lattice

Wetzel, PRE 2017

- Latent parameter corresponds to magnetization
- Identification of phases: Latent representations are clustered
- Location of phases: Magnetization, latent parameter and reconstruction loss show a steep change at the phase transition.

Siamese Neural Networks

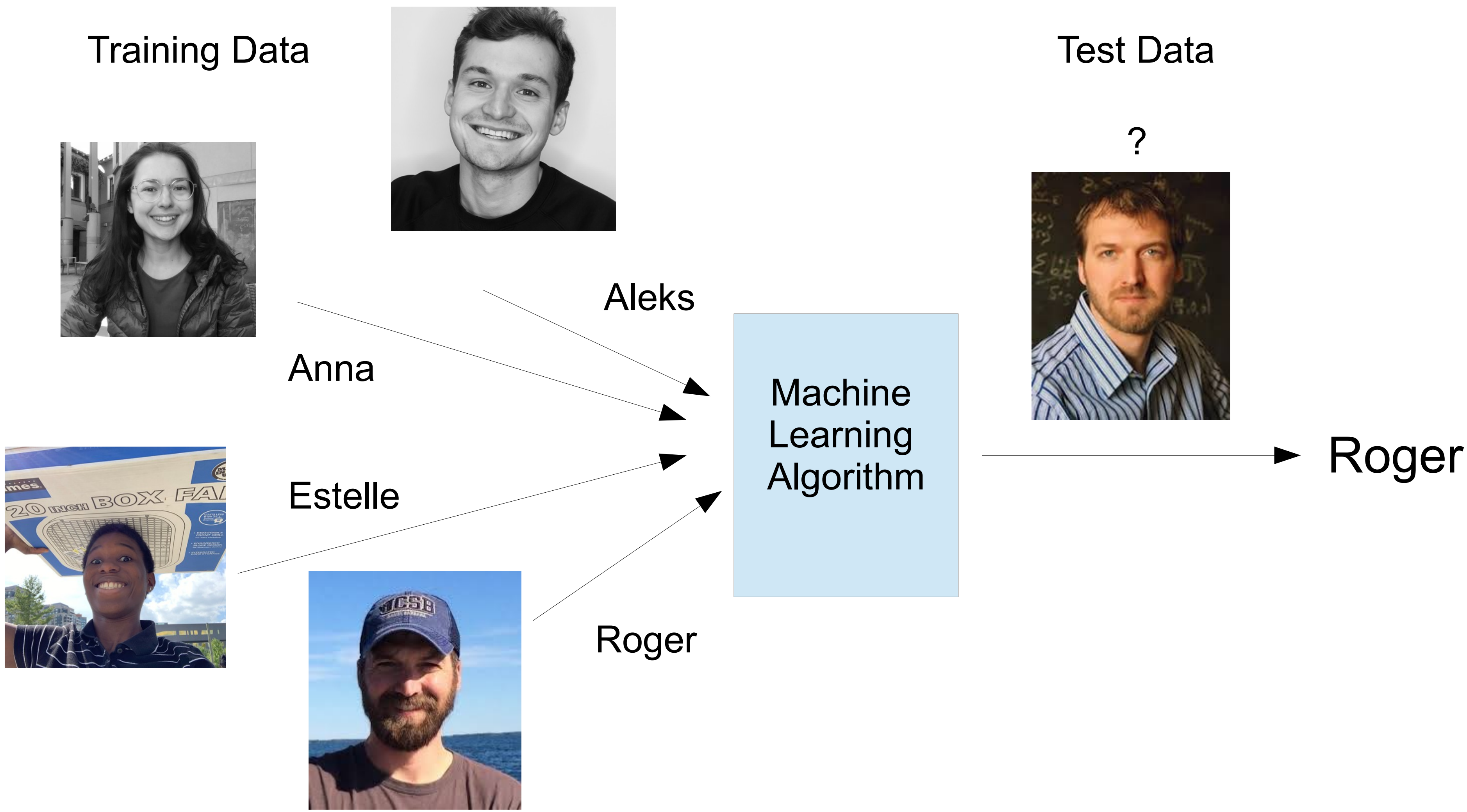


- Input : Pair of data points
- Label : same / different
- Network pair contains identical neural networks with shared weights

Wetzel, et al, PRR 2020

Machine Learning
Multi Class Classification

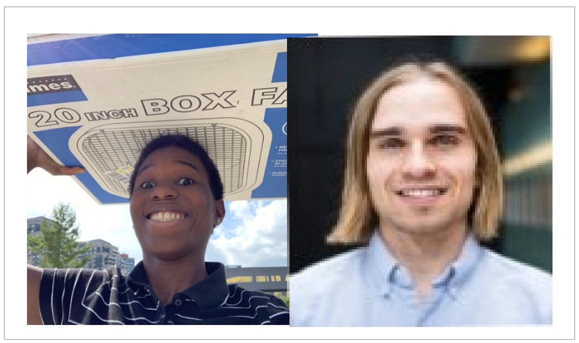
„Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed." - Wikipedia



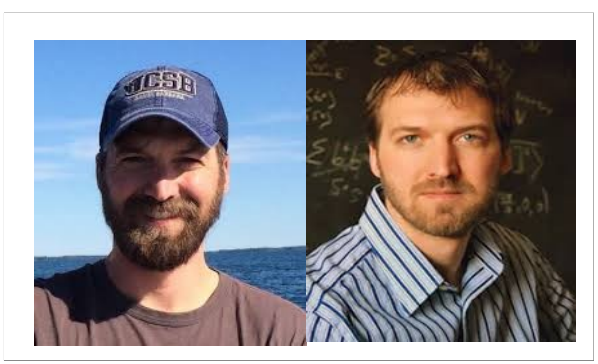
Machine Learning
Infinite Class Classification

Reformulation of the Problem:

- Teach a maching learning algorithm if two pictures show the same class.



different

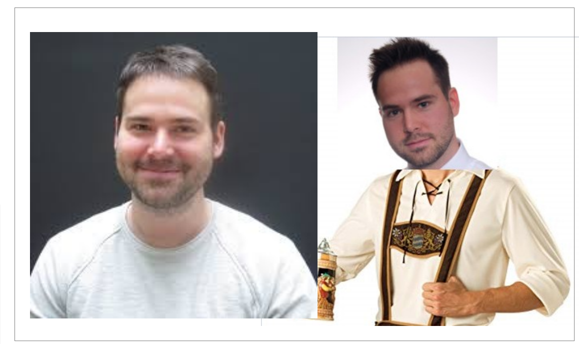


same

Machine Learning Algorithm

Test Data

?



→ same

Reference Picture

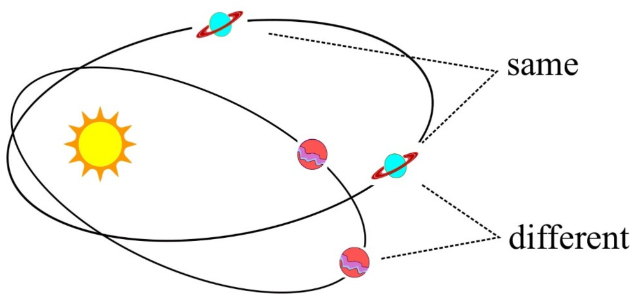
Test Picture

(Does not need to be in training set)

Siamese Neural Networks
Particle in Gravitational Potential

Problem:

- Given two observations of positions and velocities, do they belong to the same particle trajectory?



SNN Solution:

- Prepare Dataset of positive data where the pair is connected by solving the equations of motion

$$((x, y, v_x, v_y), (x', y', v_x', v_y'))$$

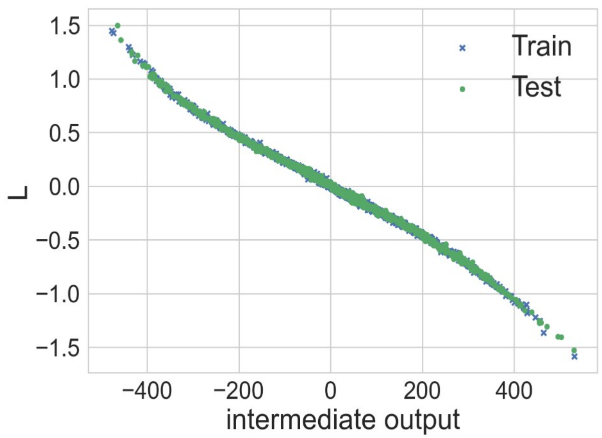
- Prepare Negative Dataset by permuting positive dataset
- Train SNN to distinguish between positive and negative pairs

Siamese Neural Networks
Particle in Gravitational Potential

Results:

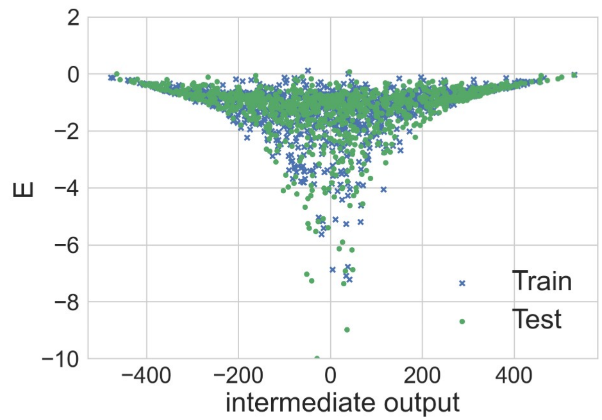
Training accuracy : 98%

Test accuracy : 97%



- Interpretation by polynomial regression on latent representation:

$$\begin{aligned} f(\mathbf{x}) \approx & -403.71xv_y - 4.85x - 0.58xy \\ & -0.17xv_x - 0.02v_y^2 - 0.01v_xv_y \\ & +0.00v_y^2 + 0.01v_y + 0.02v_x \\ & +0.45x^2 + 0.66y^2 + 0.74 \\ & +0.99yv_y + 1.24y + 402.44yv_x \\ \approx & -403\underbrace{(xv_y - yv_x)}_{=L_z} \end{aligned}$$

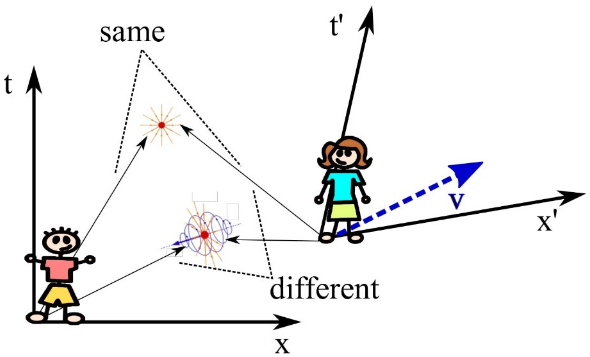


- Network has learned the angular momentum to infer its prediction.

Siamese Neural Networks
Lorentz Transformation of Electromagnetic Fields

Problem:

- Given two field configurations, can they be transformed into each other by a Lorentz transformation?



SNN Solution:

- Prepare Dataset of positive data where the pair is connected by a Lorentz Transformation

$$((E_x, E_y, E_z, B_x, B_y, B_z), (E'_x, E'_y, E'_z, B'_x, B'_y, B'_z))$$

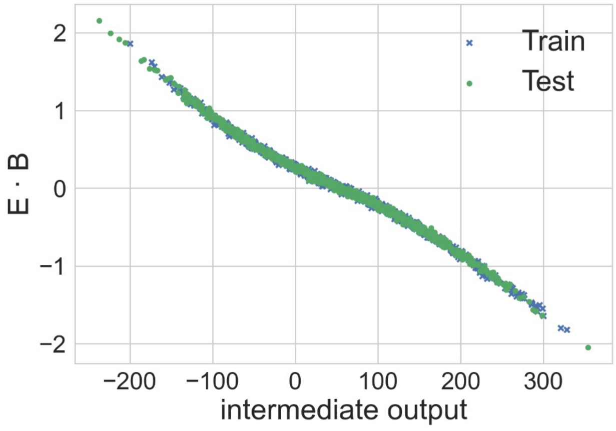
- Prepare Negative Dataset by permuting poitive dataset
- Train SNN to distinguish between positive and negative pairs

Siamese Neural Networks
Lorentz Transformation of Electromagnetic Fields

Results:

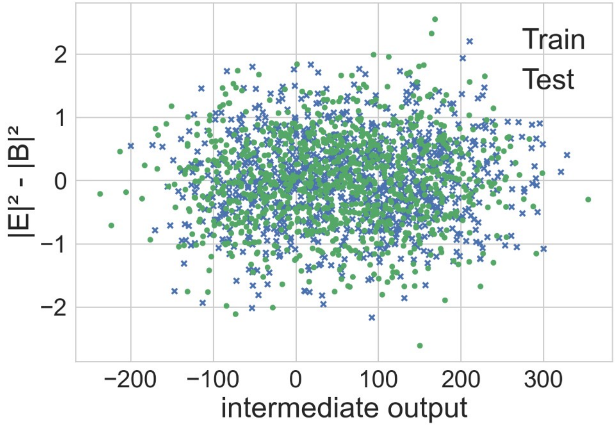
Training accuracy : 95%

Test accuracy : 94%



- Interpretation by polynomial regression on latent representation:

$$
\begin{aligned}
f(\mathbf{x}) &\approx -170.53 E_2 B_2 - 170.22 E_1 B_1 - 170.20 E_3 B_3 \\
&\quad - 4.13 B_3^2 + \ldots + 4.92 E_2^2 + 53.43 \\
&\approx -170 \underbrace{\left(E_1 B_1 + E_2 B_2 + E_3 B_3\right)}_{=E \cdot B} + 53
\end{aligned}
$$



- Network has learned the determinant of the field strength tensor to infer its prediction.

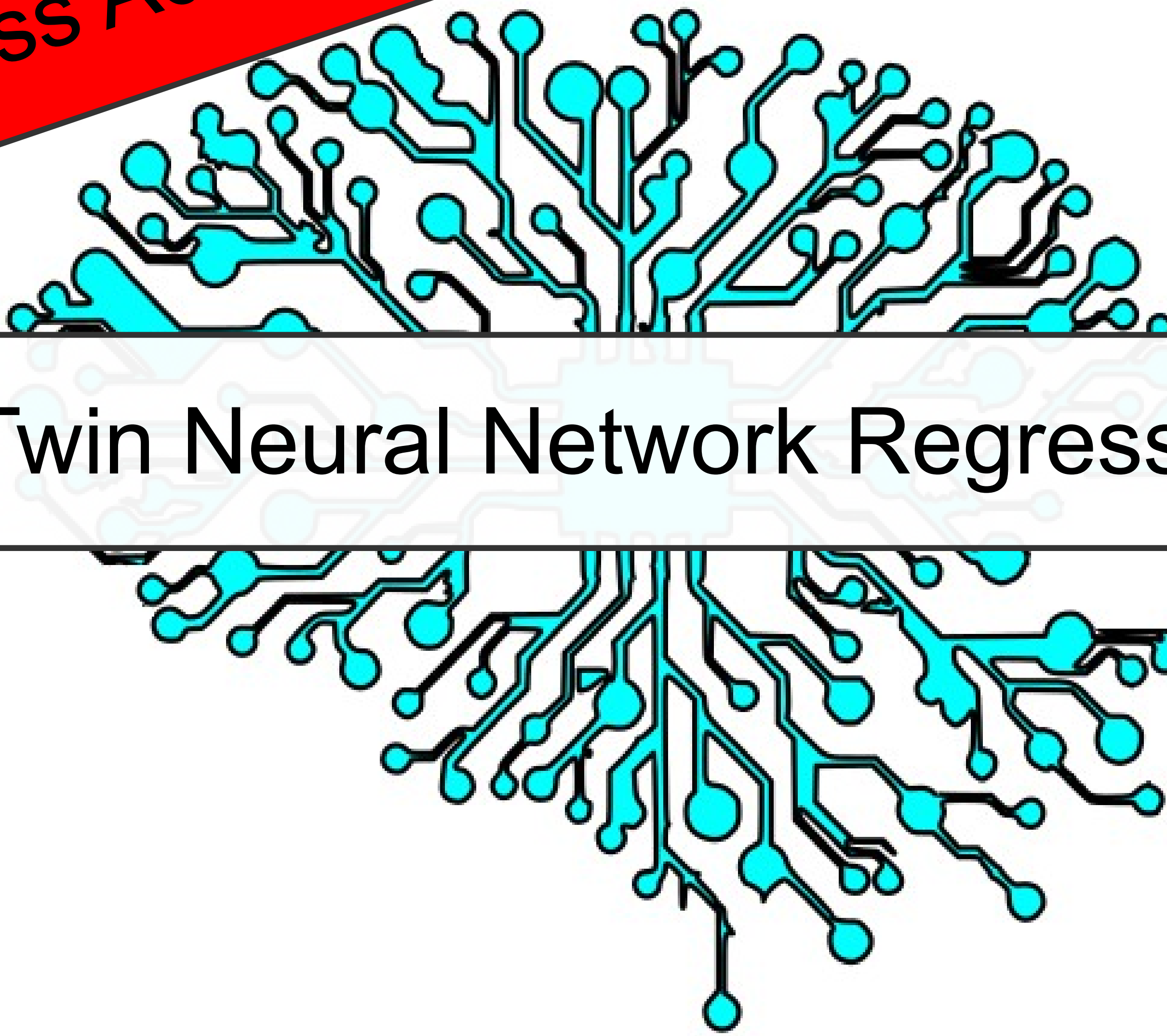
Summary

- ✗ Interpretation of Artificial Neural Networks is hard because information is distributed among many layers and neurons
- ✗ Interpretation is possible by identifying bottlenecks and performing regression
- ✗ Interpretation is constructive and can give insight into the underlying physics:

  Neural Networks applied to phase recognition learn order parameters or energies

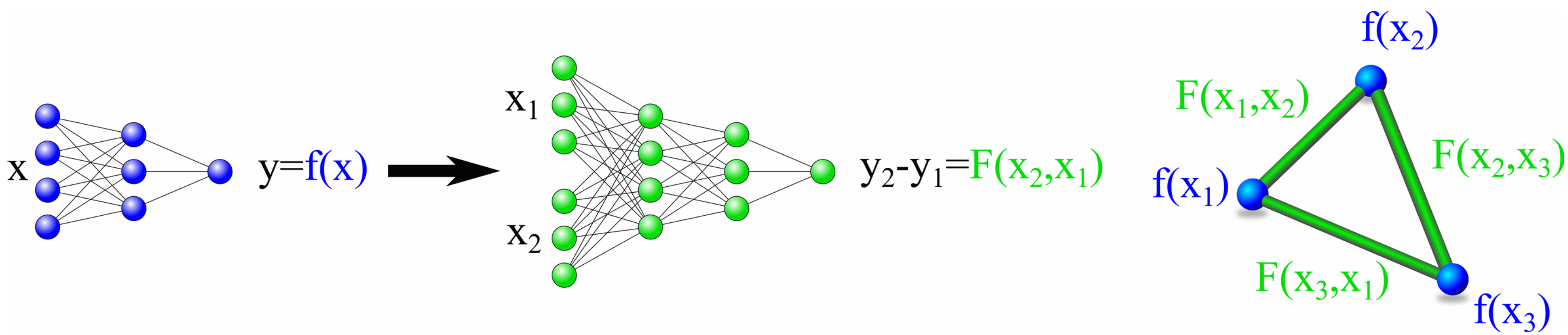
  Siamese Networks for similarity detection learn invariants or conserved quantities

Shameless Advertisement



Twin Neural Network Regression

Twin Neural Network Regression



- Solution of the Original Regression Problem:

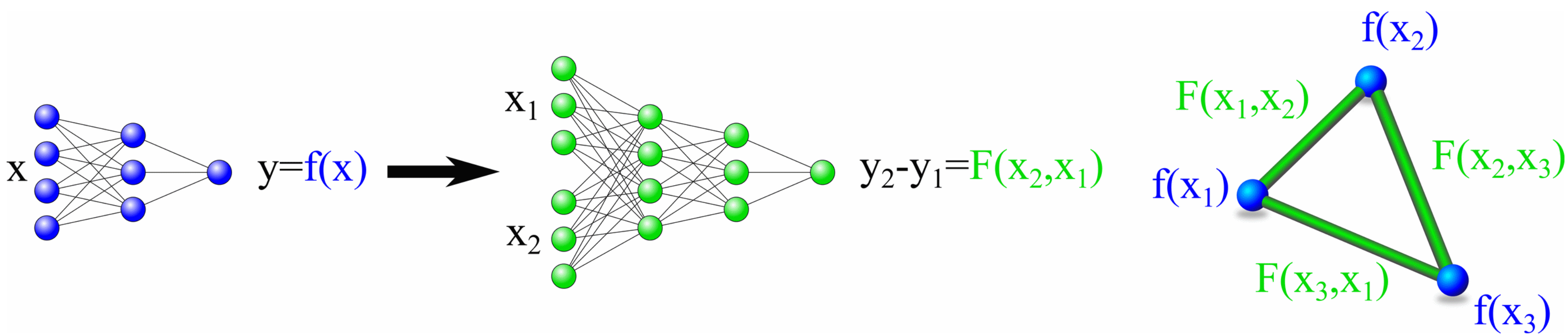
$$y_2 = F(x_2, x_1) + y_1$$

Solution to original problem

Unknown data point

Known training data point

Bias-Variance Tradeoff



TNN implicit ensemble

$$y_i^{pred} = \frac{1}{m}\sum_{j=1}^{m} F(x_i, x_j^{train}) + y_j^{train} = \frac{1}{m}\sum_{j=1}^{m} \frac{1}{2}F(x_i, x_j^{train}) - \frac{1}{2}F(x_j^{train}, x_i) + y_j^{train}$$

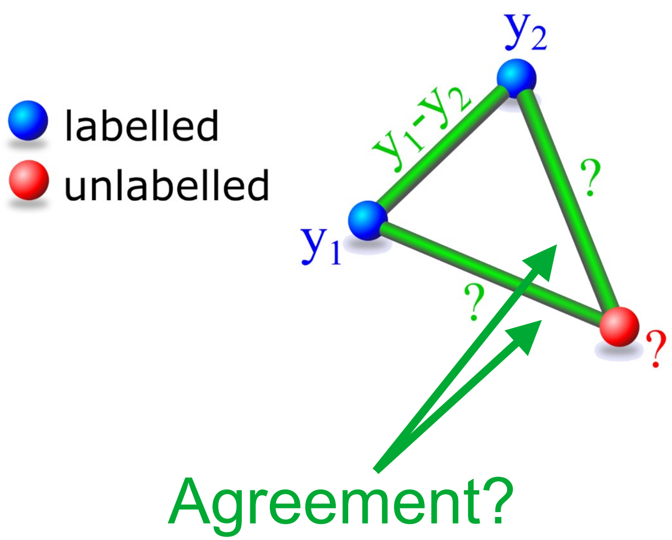
- Get huge ensemble of twice the training data set size
- Ensemble is relatively uncorrelated, since the predicted differences are different by construction

Uncertainty Signal

Do ensemble members agree?

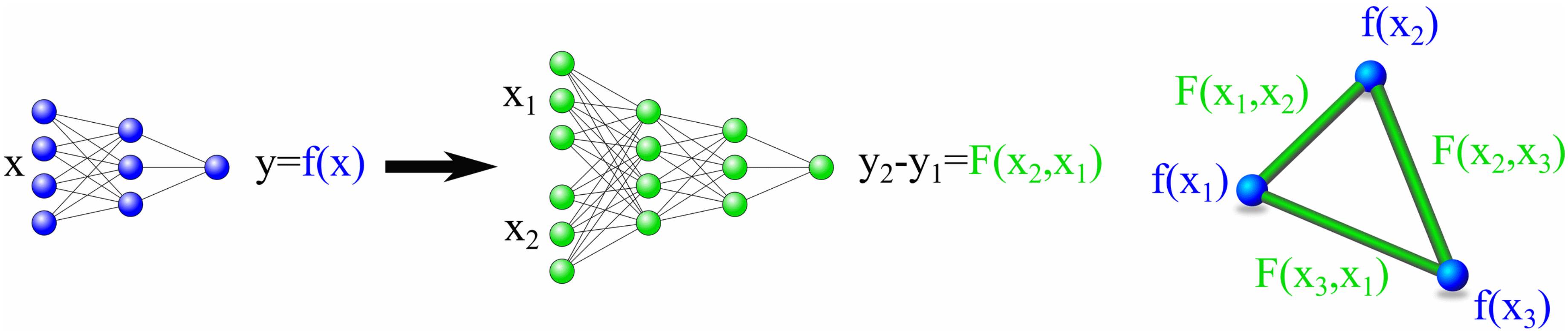
$$y_i^{pred} = \frac{1}{m}\sum_{j=1}^{m} F(x_i, x_j^{train}) + y_j^{train} = \frac{1}{m}\sum_{j=1}^{m} \frac{1}{2}F(x_i, x_j^{train}) - \frac{1}{2}F(x_j^{train}, x_i) + y_j^{train}$$

- Uncorrelated predictions make different mistakes
- Measure ensemble standard deviation



(additional uncertainty signal based on loop consistencies)

Semi-Supervised Learning



- Train to enforce loop consistency during training
- Loops can be used as training data even if the data points within them are unlabelled

$$0 = F(x_i, x_j) + F(x_j, x_k) + F(x_k, x_i)$$

- It can be viewed as two predictions provide a suggested label for the third.