Introduction to Machine Learning Methods in Condensed Matter Physics

LECTURE 4, FALL 2021

Yi Zhang (张亿)

International Center for Quantum Materials, School of Physics
Peking University, Beijing, 100871, China

Email: frankzhangyi@pku.edu.cn

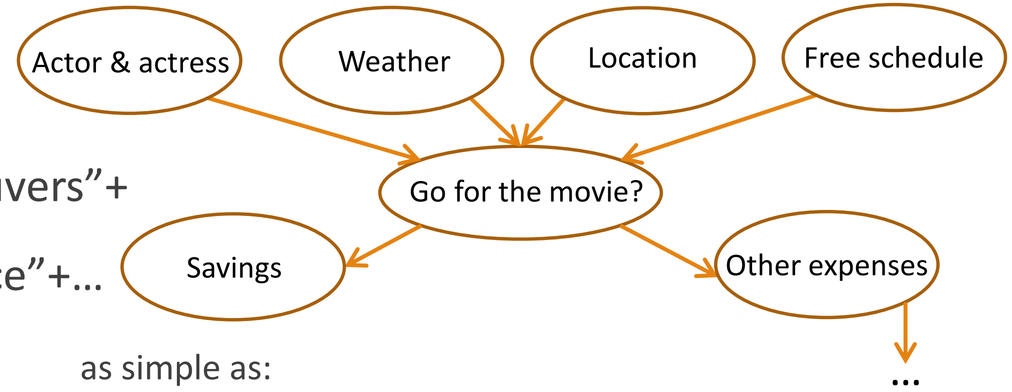
Deep neural network for the better

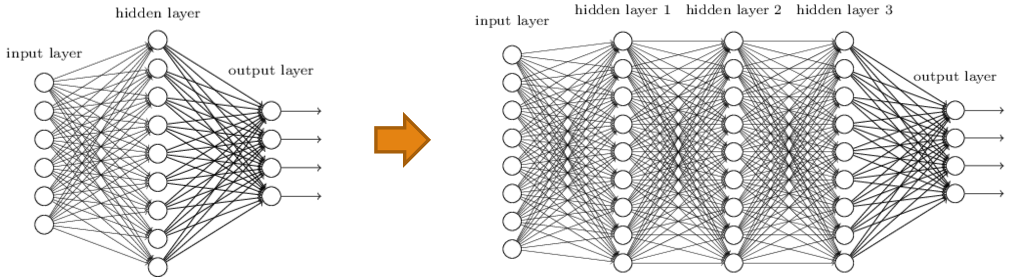
- Deep neural network for more logical layers

  Breaking a difficult task into smaller, easier tasks

  “Driving a car” = “Parking ”+“Traffic rules” +“Maneuvers”+

  “Car control ”+“Situation awareness” +“Maintenance”+…





as simple as:

```
net = network2.Network([784, 30, 30, 30, 10])
```

- Back propagation also works for deep feed-forward neural networks:

$$\delta_j^L = \frac{\partial C}{\partial z_j^L} = \sum_k \frac{\partial C}{\partial a_k^L}\frac{\partial a_k^L}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L}\sigma'(z_j^L)$$

$$\delta_j^l = \frac{\partial C}{\partial z_j^l} = \sum_k \frac{\partial C}{\partial z_k^{l+1}}\frac{\partial z_k^{l+1}}{\partial z_j^l} = \sum_k \delta_k^{l+1} w_{kj}^{l+1}\sigma'(z_j^l)$$

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l$$

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1}\delta_j^l$$

- Same regularization and initialization methods, cross entropy, Softmax layer, etc.
- “Going deep” can be “exponentially” more helpful than “going wide”.

Problem with deep neural network

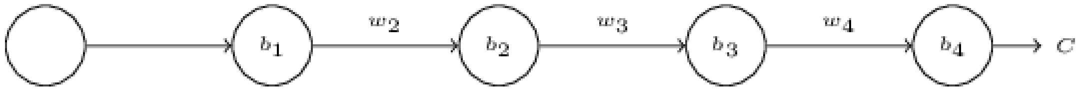
- However, the accuracy on MNIST data peaks at few layers.

  Deep neural networks are harder to train:

  Earlier layers see exponentially smaller gradient.

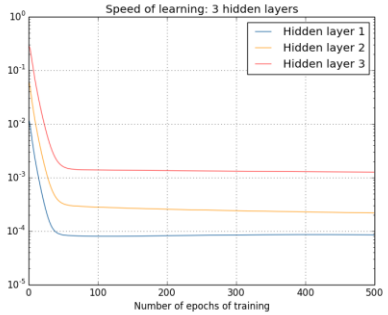
  Example:

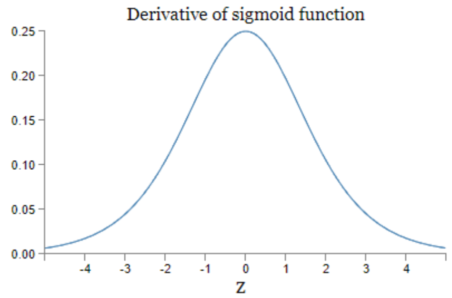
$$\frac{\partial C}{\partial b_1} = \sigma'(z_1) \times \overbrace{w_2 \times \sigma'(z_2)}^{<\frac{1}{4}} \times \overbrace{w_3 \times \sigma'(z_3)}^{<\frac{1}{4}} \times w_4 \times \sigma'(z_4) \times \frac{\partial C}{\partial a_4}$$

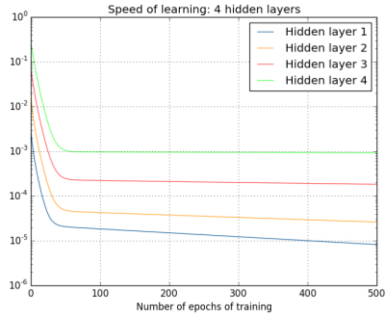


$$z_j = w_j a_{j-1} + b_j$$

$$\|\delta^1\| = 0.003$$
$$\|\delta^2\| = 0.017$$
$$\|\delta^3\| = 0.070$$
$$\|\delta^4\| = 0.285$$







  In general, the condition to satisfy $|w\sigma'(wa+b)| \geq 1$ is quite strict.

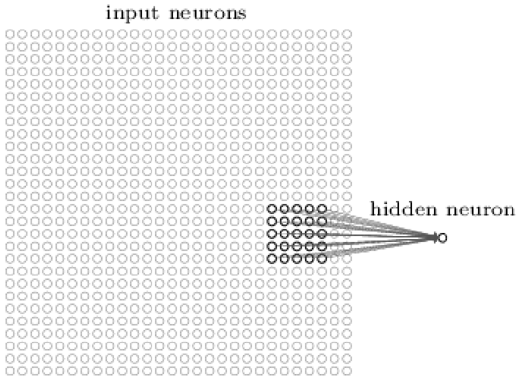
  The *vanishing gradient problem:* deep neural networks' gradient is unstable. ⇨ the second "AI winter"

- Other issues: over-training, lower interpretability, initialization dependence, etc.

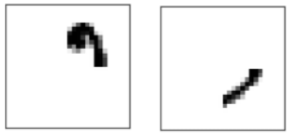
Convolutional neural network

- Unused information on image recognition with a fully connected ANN: *locality* and *translation symmetry*
- Local receptive field with shared weights and biases:

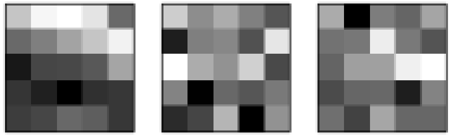
$$\sigma\left(b+\sum_{l=0}^{4}\sum_{m=0}^{4} w_{l,m}\, a_{j+l,k+m}\right)$$



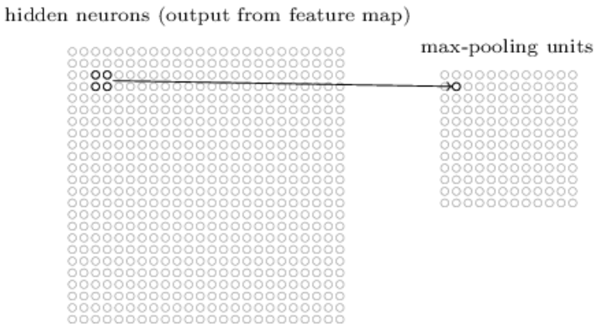
detect local features: example



in reality, hard to interpret:

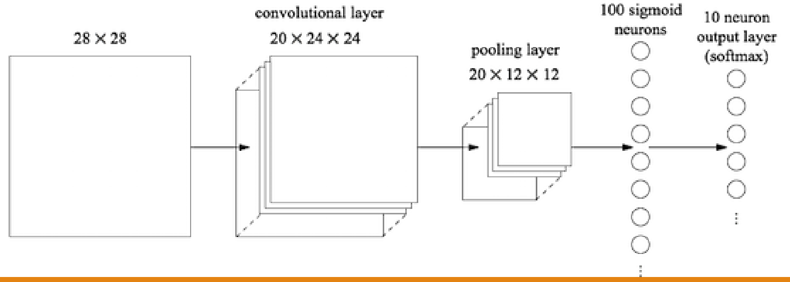


- Pooling layers: e.g., max-pooling: output maximum in a 2x2 region



Greatly reduce the number of ANN parameters and concentrate gradients

- Overall, with convolution architecture:



Back propagation still works.

Convolutional neural network for image recognition

- With two convolutional layers, two fully connected layers with dropout:

  also: L2 regularization, expanded training data, RELU activation

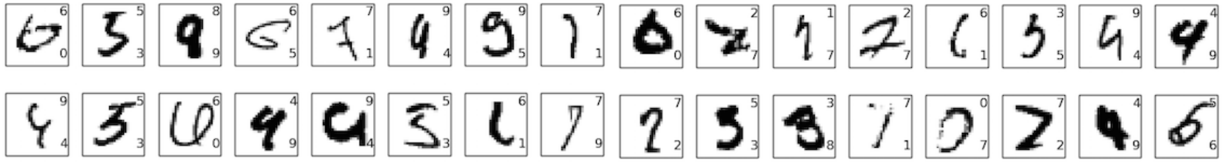
  also x2: multiple CNNs with a majority vote

  note: each idea may or may not improve. Every little bit helps.

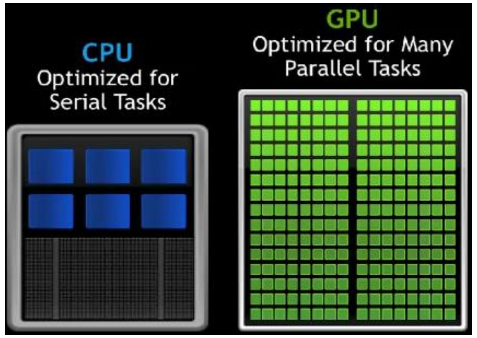
  also note: do not lose the bigger picture. Deeper ≠ better.

```
net = Network([
    ConvPoolLayer(image_shape=(mini_batch_size, 1, 28, 28),
                  filter_shape=(20, 1, 5, 5),
                  poolsize=(2, 2),
                  activation_fn=ReLU),
    ConvPoolLayer(image_shape=(mini_batch_size, 20, 12, 12),
                  filter_shape=(40, 20, 5, 5),
                  poolsize=(2, 2),
                  activation_fn=ReLU),
    FullyConnectedLayer(
        n_in=40*4*4, n_out=1000, activation_fn=ReLU, p_dropout=0.5),
    FullyConnectedLayer(
        n_in=1000, n_out=1000, activation_fn=ReLU, p_dropout=0.5),
    SoftmaxLayer(n_in=1000, n_out=10, p_dropout=0.5)],
    mini_batch_size)
```

- The remaining "errors" are rather understandable!



- The power of GPU: parallel computing, matrix processing, etc.



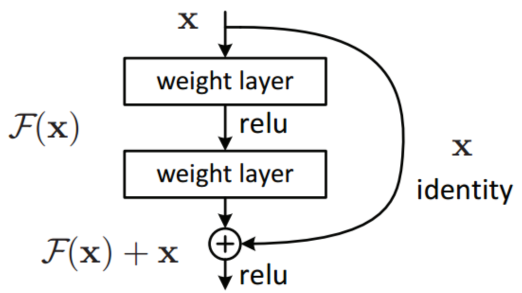
Thousands of CUDA (Compute Unified Device Architecture) in a single GPU

Deep neural network with residual blocks

- Residue block gives the neural network the choice to short circuit selected layers:

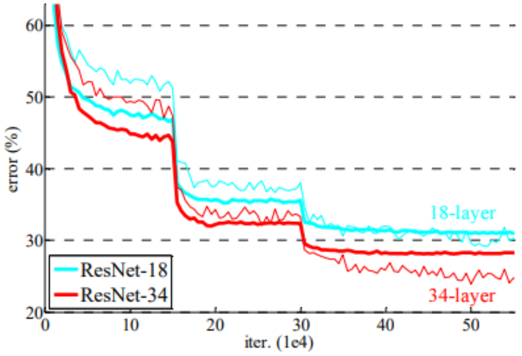
  The layers are trying to learn the residue $F(x) = H(x) - x$

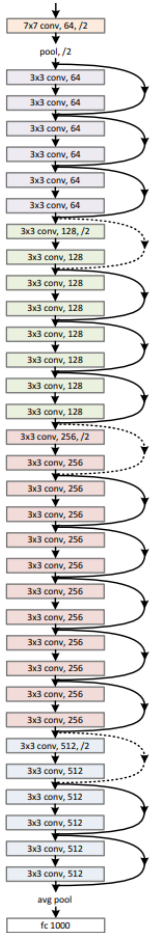
  between input and output.



  *The ANN is actually not efficient at learning an identity function.

- With residue blocks, we can propagate larger gradients to earlier layers to make them learn as fast as the later layers and less likely to over train.





- Resnet gives us the ability to train deeper networks.

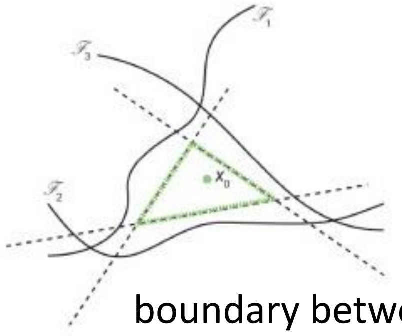
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, arXiv:1512.03385

Adversarial attacks and defenses

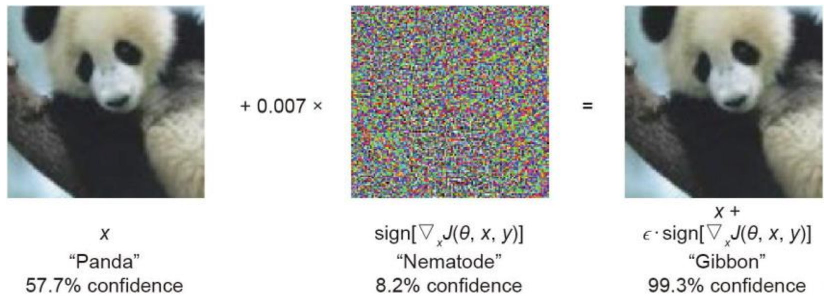
- Given an ANN, perform gradient descent with respect to the input x to maximize output in a certain (wrong) category:

$$x' = x - \epsilon \cdot \mathrm{sign}\left[\nabla_x J\left(\theta, x, y'\right)\right]$$

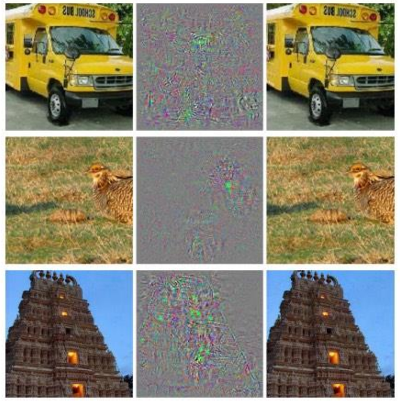
Easy, local maximum is sufficient.



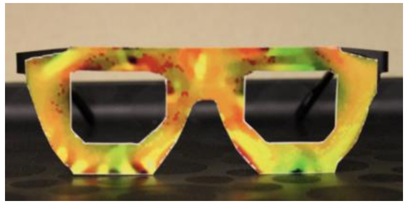
boundary between classes by nonlinear model



Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus, arXiv:1312.6199



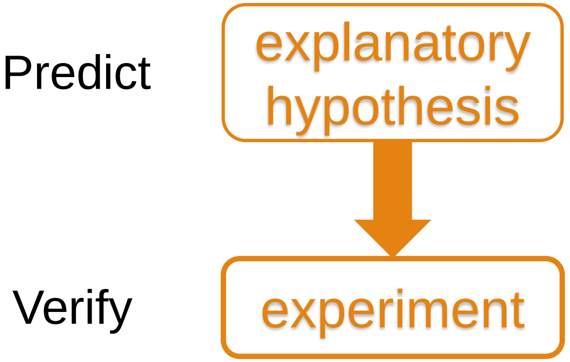
Original + noise = wrong



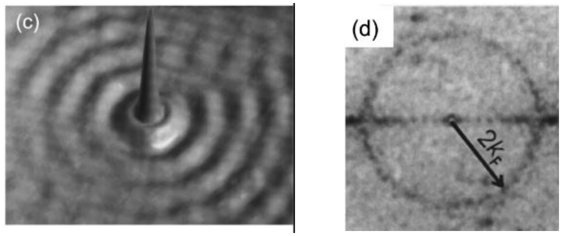
- Luckily, such "noises" are rare instances and by constructions only.
- Adversarial defenses: adversarial training, random noising, etc.
- Our understanding of ANNs are still fairly rudimentary. Artificial "intelligence", you say?

We do not yet have a fundamental understanding of our "intelligence."

Scientific protocol

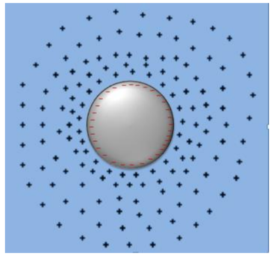


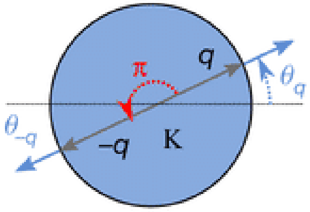
Friedel oscillations as a screening and back-scattering process around a local impurity



Pierre Mallet, et al. Comptes Rendus Physique 17, 294 (2016)

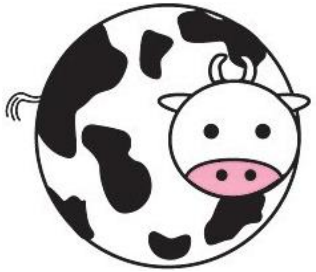
Fourier transform





Quasi-particle interference pattern as a Fermi-surface probe of electron liquids

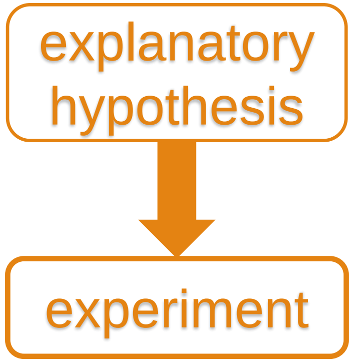
- Our theoretical capacity largely lags behind our real-world complexity:
- Consequences of controlled and uncontrolled approximations?



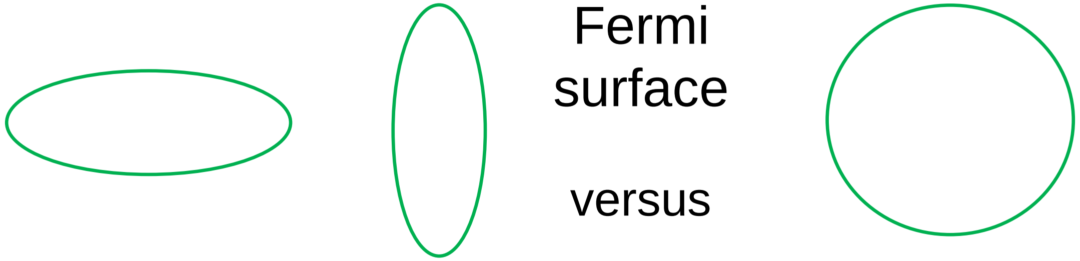
Scientific protocol

Predict

Verify

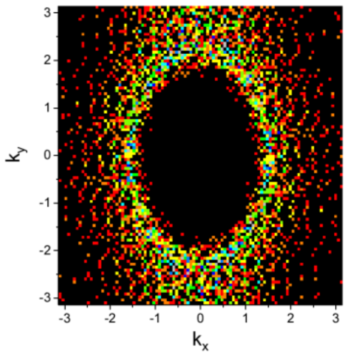


Different symmetry phases with a single impurity (dilute impurities):

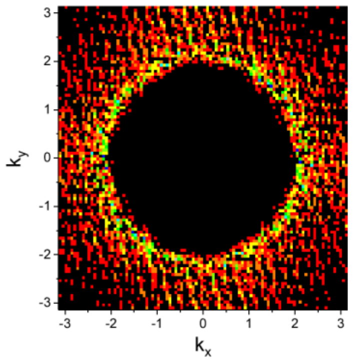


$t_x \neq t_y$ $\qquad$ $t_x = t_y$

Nematic phase



Isotropic phase



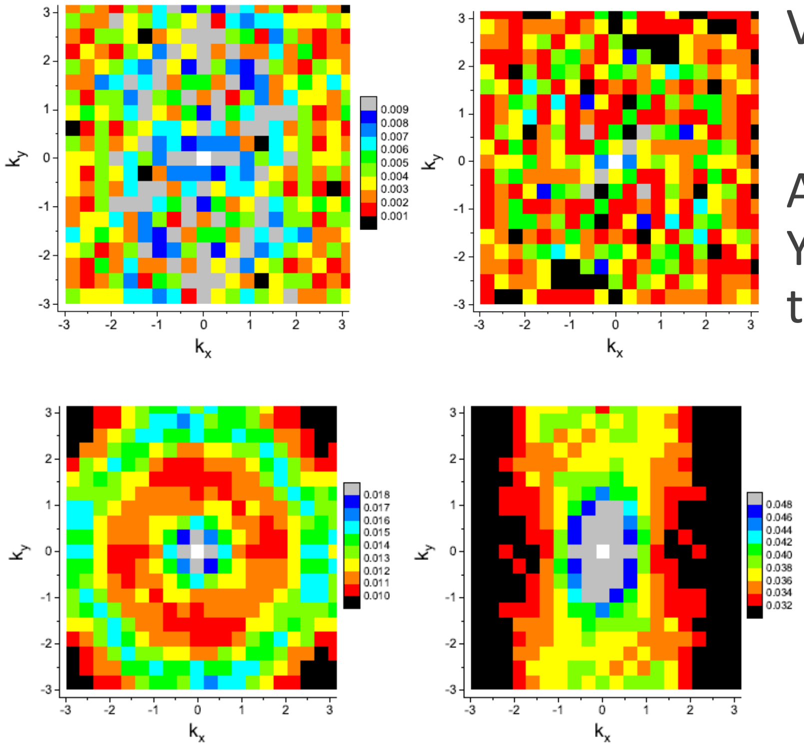
- Universality of the symmetry classes still present in this limit, indeed, its local density of states:

$$N_{\mathbf{r}}(\omega) = -2\mathrm{Im}\tilde{G}_{\mathbf{r},\mathbf{r}}(\omega)$$

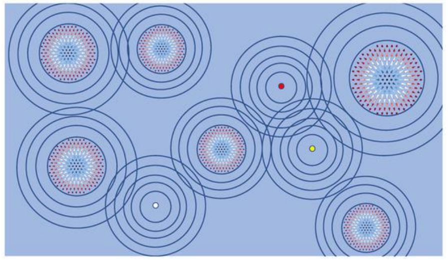
Jeremy B. Goetz, Yi Zhang, M. J. Lawler, SciPost Phys. **8**, 087 (2020); *Yi Zhang, Andrej Mesaros†, et al., Nature **570**, 484–490 (2019).*

Machine learning based scientific protocol

- Impurities break symmetries, so we know in the dense impurity limit, distinctions between symmetry classes are lost, what about intermediate impurity density?



Very noisy, is the information still present?



Average over 1000 samples to reveal:
Yes, the information is present! But we don't have a theoretical smoking-gun recipe.

Predict

e.g. data size,
complex theory,
noisy data

Verify

explanatory hypothesis

experiment

1. Generate data samples underlying different categories/hypotheses.
2. Train ANNs to classify data samples via supervised machine learning.
3. Apply the ANNs towards the target experimental or numerical dataset.