Title: Phase Detection with Neural Networks: Interpreting the Black Box

Speakers: Anna Dawid-Å•Ä™kowska

Series: Machine Learning Initiative

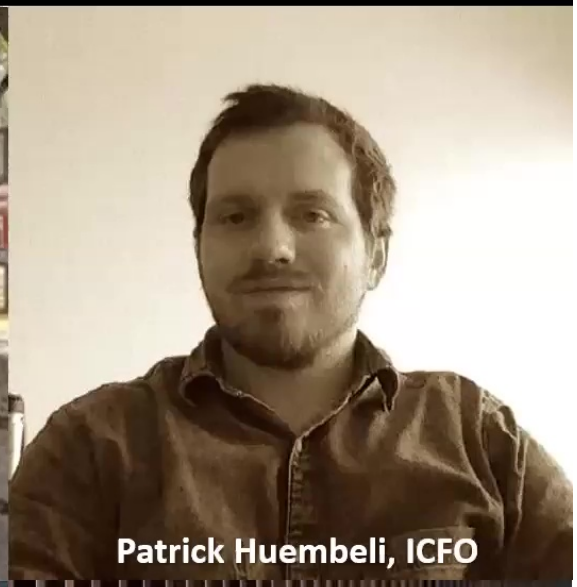
Date: June 23, 2020 - 11:00 AM

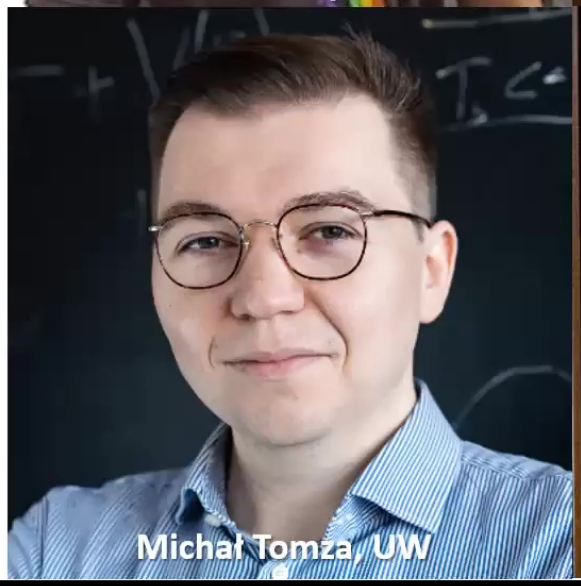
URL: http://pirsa.org/20060050

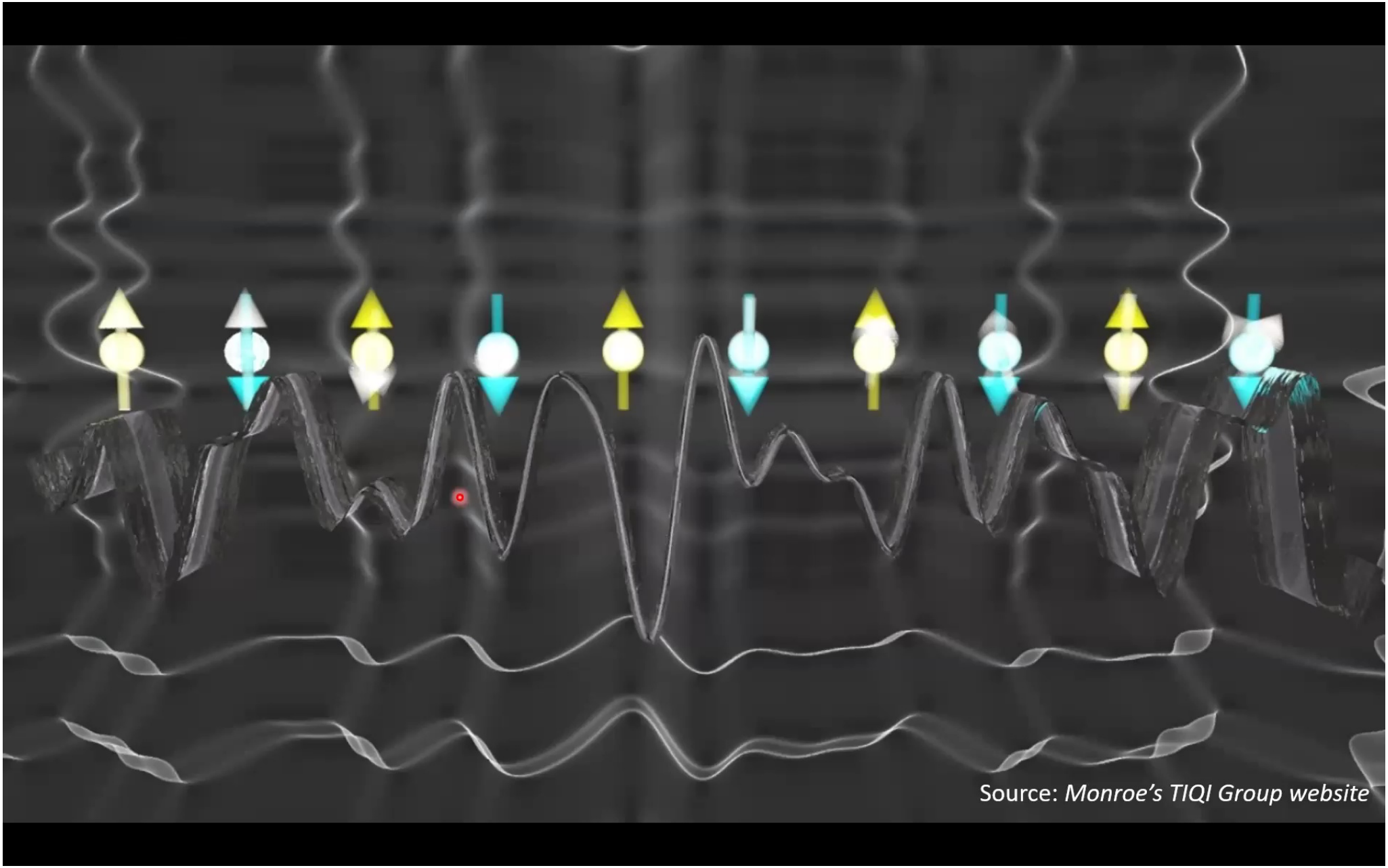
Abstract: Neural networks (NNs) normally do not allow any insight into the reasoning behind their predictions. We demonstrate how inï¬,uence functions can unravel the black box of NN when trained to predict the phases of the one-dimensional extended spinless Fermi-Hubbard model at half-ï¬•lling. Results provide strong evidence that the NN correctly learns an order parameter describing the quantum transition. Moreover, we demonstrate that inï¬,uence functions not only allow to check that the network, trained to recognize known quantum phases, can predict new unknown ones but even guide physicists in understanding patterns responsible for the phase transition. This method requires no a priori knowledge on the order parameter, the system itself, or even the architecture of the ML model.

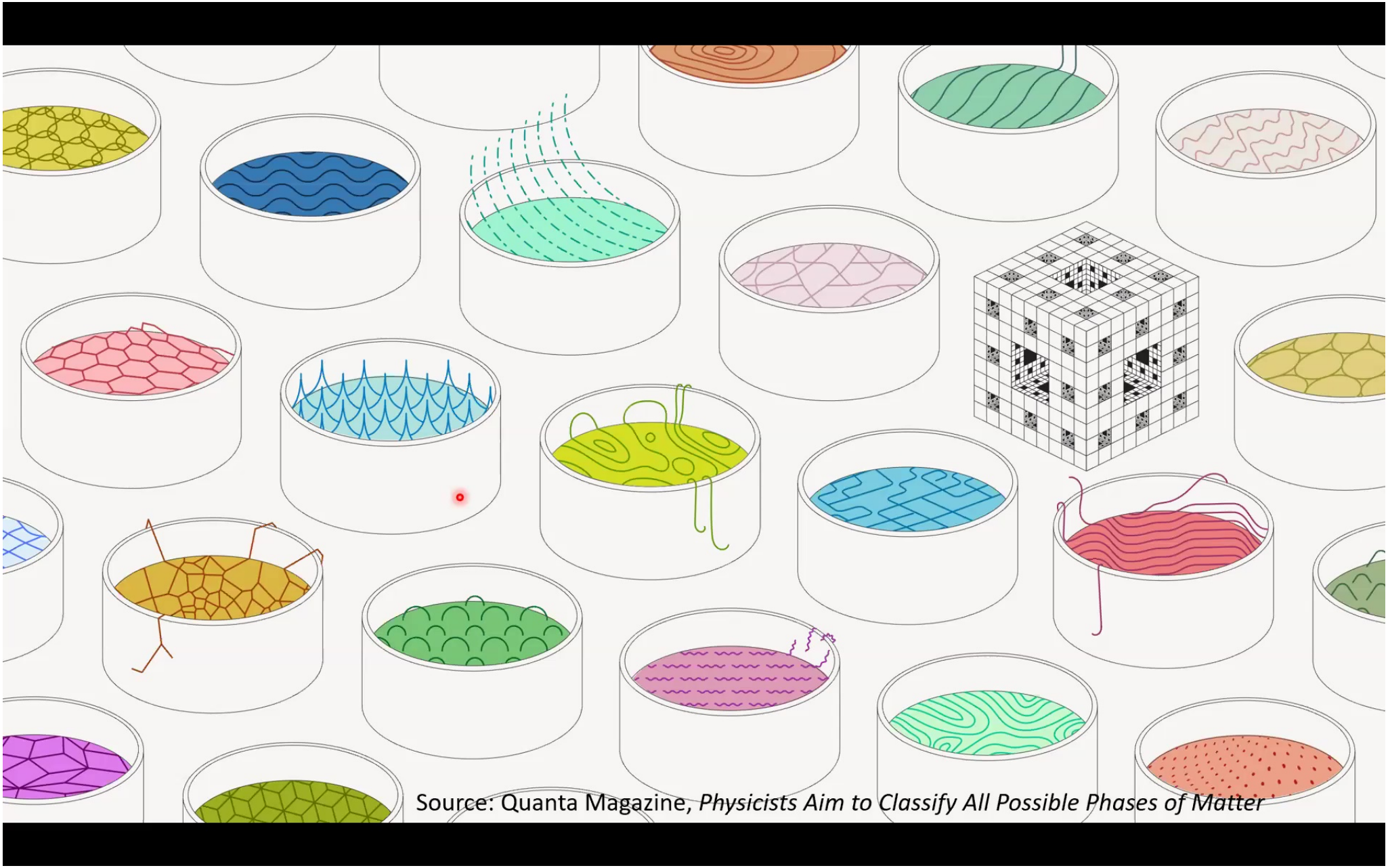
ICFO
FACULTY OF PHYSICS
UNIVERSITY OF WARSAW
Phase detection with neural networks: interpreting the black box
arXiv:2004.04711
code sample: DOI 5281/zenodo.3759432

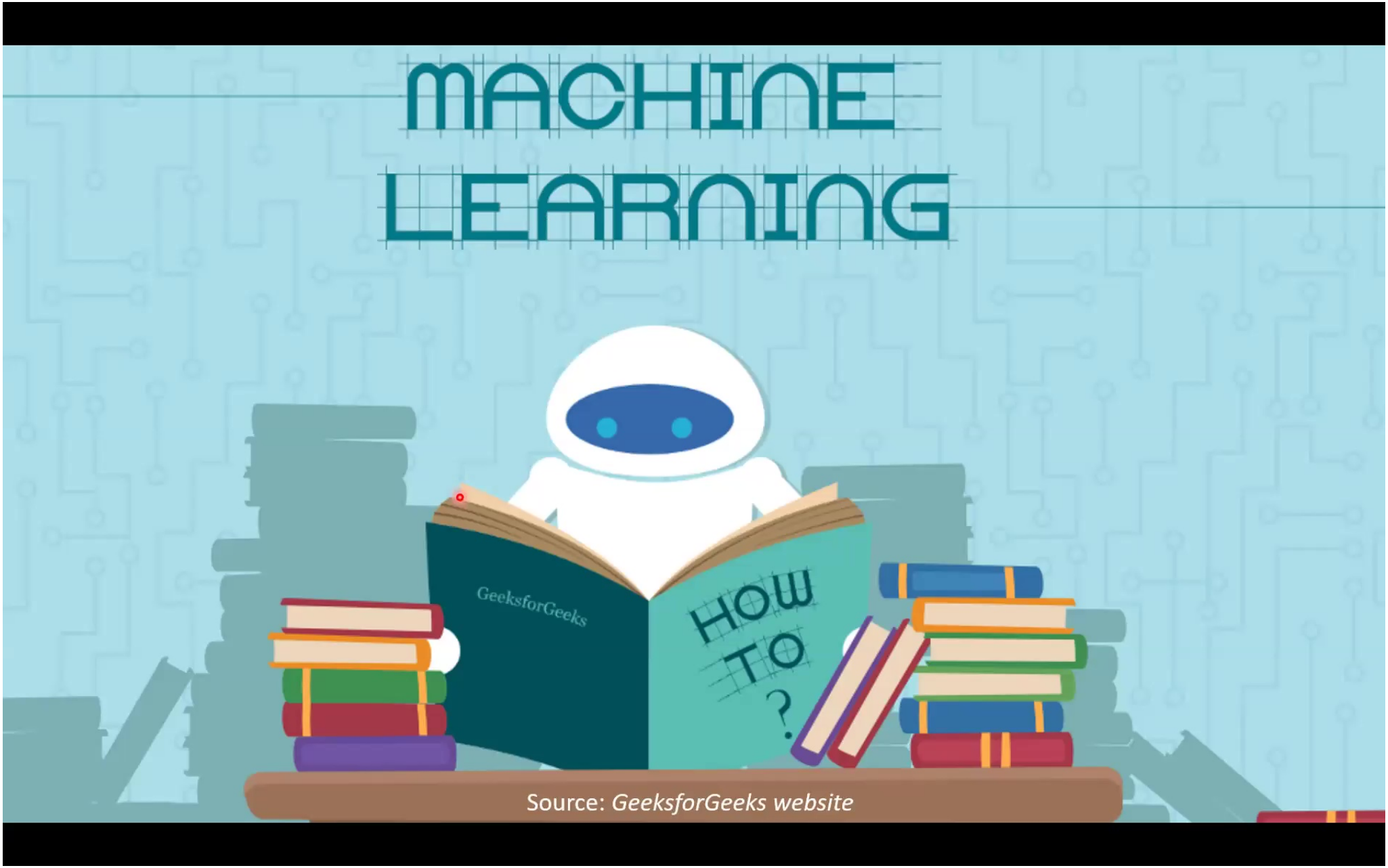
# It was a teamwork!




Source: *Monroe's TIQI Group website*


Source: Quanta Magazine, *Physicists Aim to Classify All Possible Phases of Matter*


MACHINE
LEARNING
GeeksforGeeks
HOW
TO
?
Source: GeeksforGeeks website

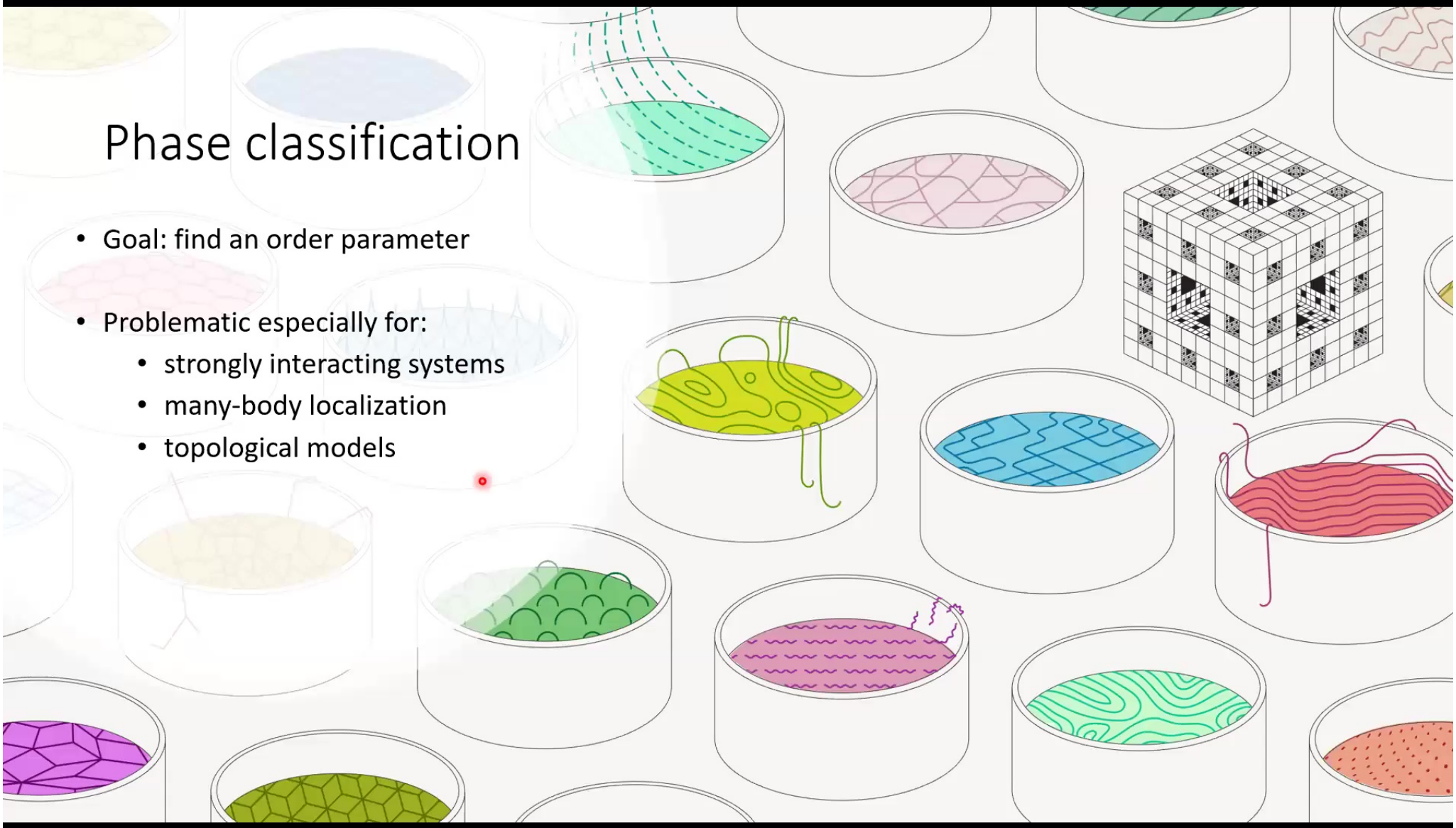
But how?
ACHINE
EARNING
HOW
TO

# Outline

1. Machine learning in phase classification
   *aim and motivation*
2. Interpretability of machine learning
   *methods*
3. ML model
   *methods*
4. Physical model
   *methods*
5. What the machine learns in a phase classification problem?
   *results*

# Phase classification

- Goal: find an order parameter
- Problematic especially for:
  - strongly interacting systems
  - many-body localization
  - topological models



# Machine learning

Machine learning (ML) studies algorithms whose performance **improves with data** (“learning from experience”)

Data $\boldsymbol{X}$ → → Model $\hat{f}$

Mitchell, McGraw Hill, 1997

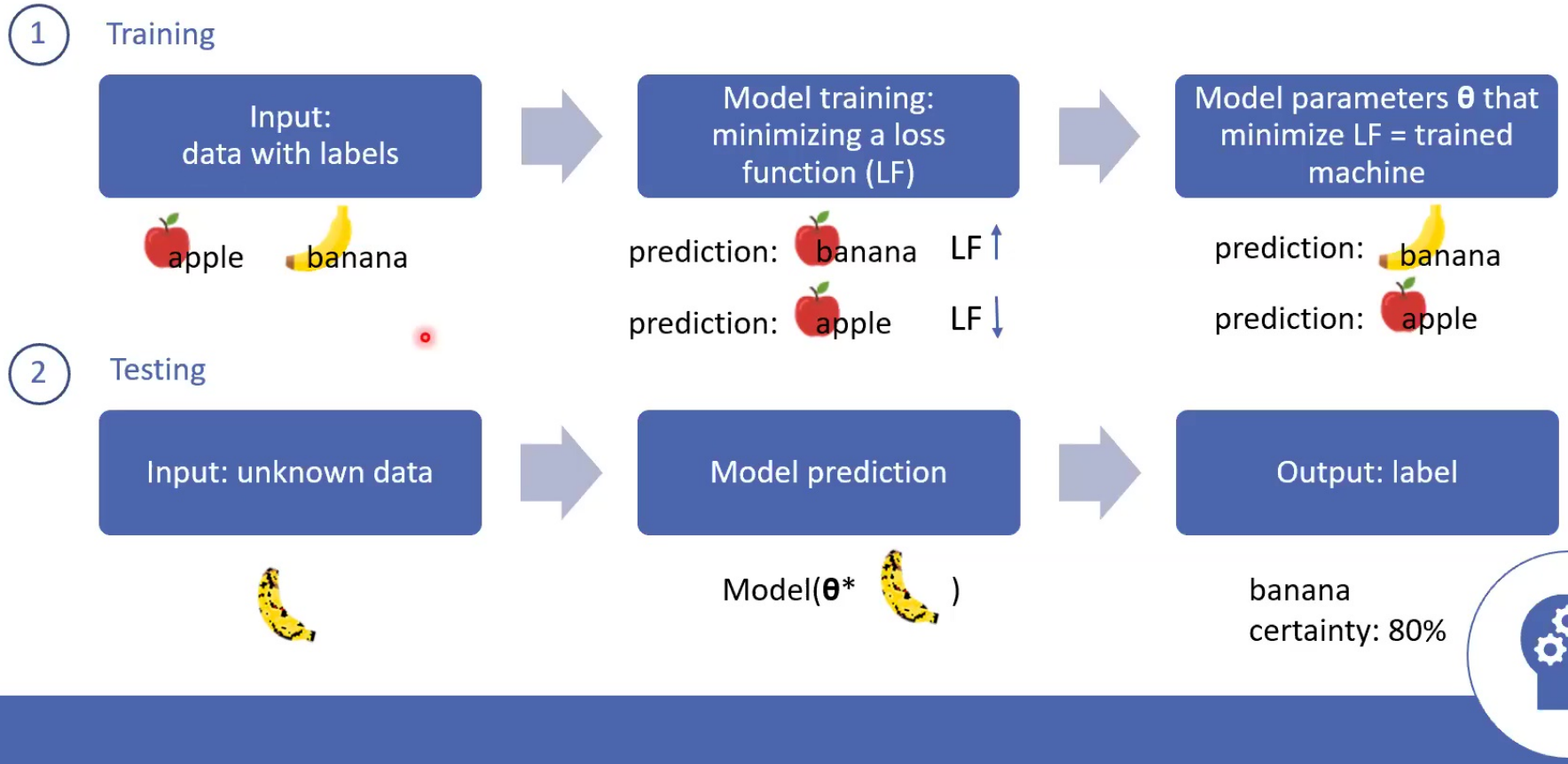
# Types of machine learning

**Supervised** – machine learns on pairs of input and output data

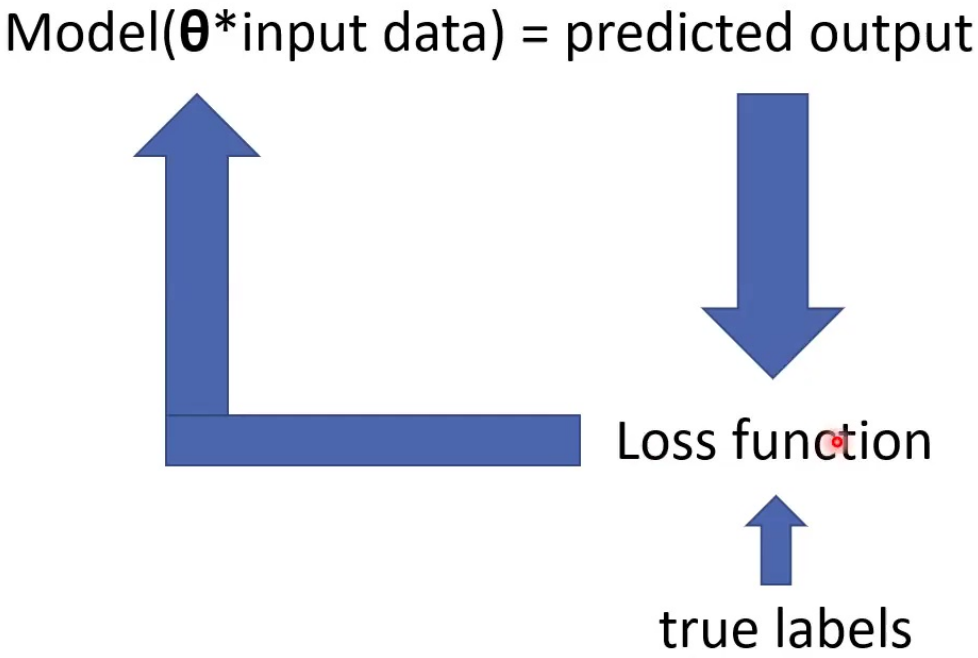
**Unsupervised** – machine groups and interprets basing just on the input data

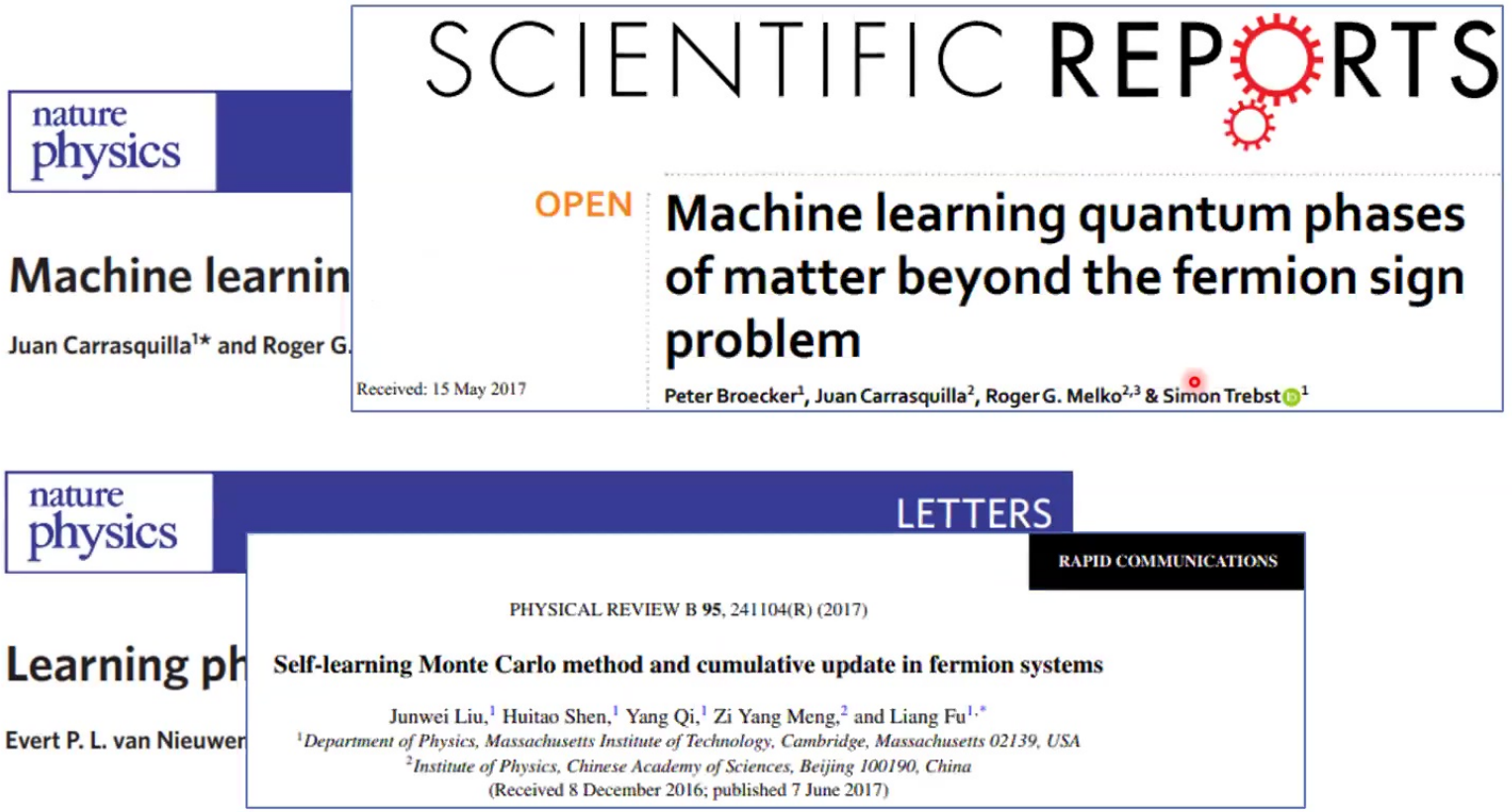
**Reinforcement** – algorithm learns to react to an environment

# Supervised machine learning



# Training


Model(θ*input data) = predicted output
Loss function
true labels



nature physics

Machine learnin

Juan Carrasquilla[1]* and Roger G.

SCIENTIFIC REPORTS

OPEN

Machine learning quantum phases of matter beyond the fermion sign problem

Received: 15 May 2017

Peter Broecker[1], Juan Carrasquilla[2], Roger G. Melko[2,3] & Simon Trebst[1]

nature physics

LETTERS

Learning ph

Evert P. L. van Nieuwer

RAPID COMMUNICATIONS

PHYSICAL REVIEW B **95**, 241104(R) (2017)

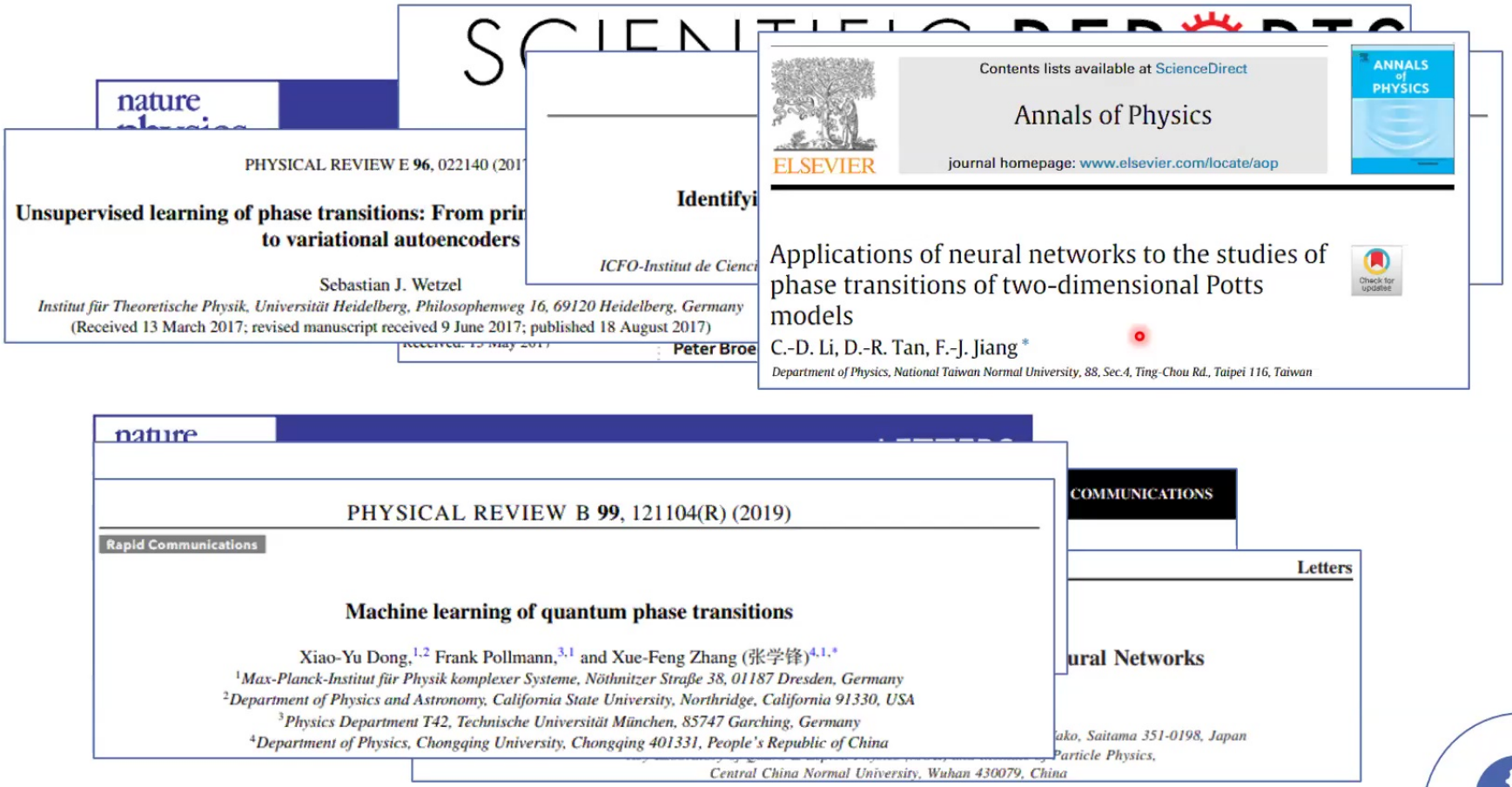
**Self-learning Monte Carlo method and cumulative update in fermion systems**

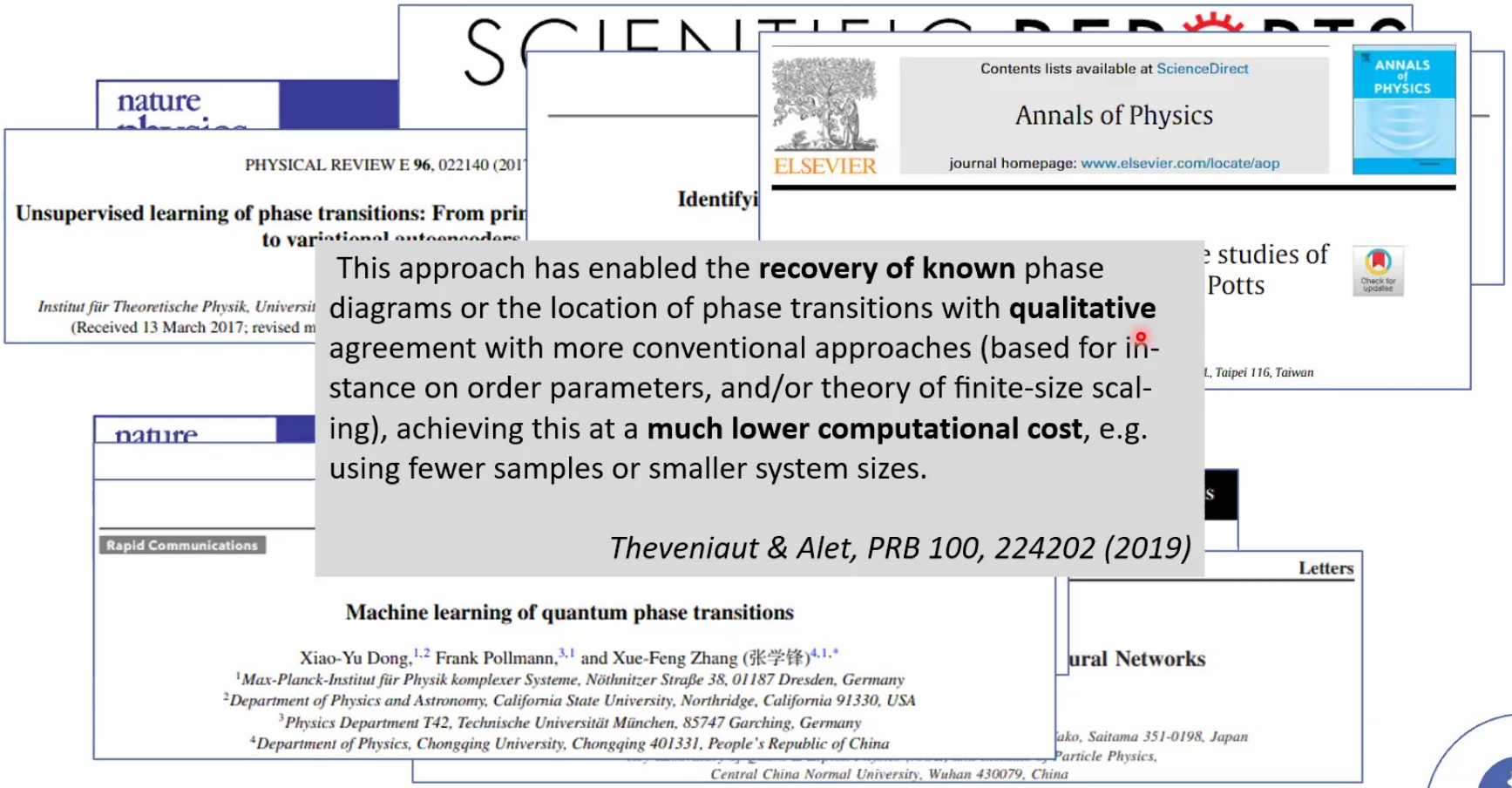
Junwei Liu,[1] Huitao Shen,[1] Yang Qi,[1] Zi Yang Meng,[2] and Liang Fu[1,*]

[1]*Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

[2]*Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China*

(Received 8 December 2016; published 7 June 2017)


nature
PHYSICAL REVIEW E 96, 022140 (201
Unsupervised learning of phase transitions: From prin
to variational autoencoders
Sebastian J. Wetzel
Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 16, 69120 Heidelberg, Germany
(Received 13 March 2017; revised manuscript received 9 June 2017; published 18 August 2017)
Identifyi
ICFO-Institut de Cienci
Peter Broe
Contents lists available at ScienceDirect
Annals of Physics
journal homepage: www.elsevier.com/locate/aop
ELSEVIER
ANNALS of PHYSICS
Applications of neural networks to the studies of phase transitions of two-dimensional Potts models
C.-D. Li, D.-R. Tan, F.-J. Jiang *
Department of Physics, National Taiwan Normal University, 88, Sec.4, Ting-Chou Rd., Taipei 116, Taiwan
nature
PHYSICAL REVIEW B 99, 121104(R) (2019)
Rapid Communications
Machine learning of quantum phase transitions
Xiao-Yu Dong,[1,2] Frank Pollmann,[3,1] and Xue-Feng Zhang (张学锋)[4,1,*]
[1]Max-Planck-Institut für Physik komplexer Systeme, Nöthnitzer Straße 38, 01187 Dresden, Germany
[2]Department of Physics and Astronomy, California State University, Northridge, California 91330, USA
[3]Physics Department T42, Technische Universität München, 85747 Garching, Germany
[4]Department of Physics, Chongqing University, Chongqing 401331, People's Republic of China
COMMUNICATIONS
Letters
ural Networks
ako, Saitama 351-0198, Japan
Particle Physics,
Central China Normal University, Wuhan 430079, China




This approach has enabled the **recovery of known** phase diagrams or the location of phase transitions with **qualitative** agreement with more conventional approaches (based for instance on order parameters, and/or theory of finite-size scaling), achieving this at a **much lower computational cost**, e.g. using fewer samples or smaller system sizes.

*Theveniaut & Alet, PRB 100, 224202 (2019)*

# Machines in phase classification – open problems

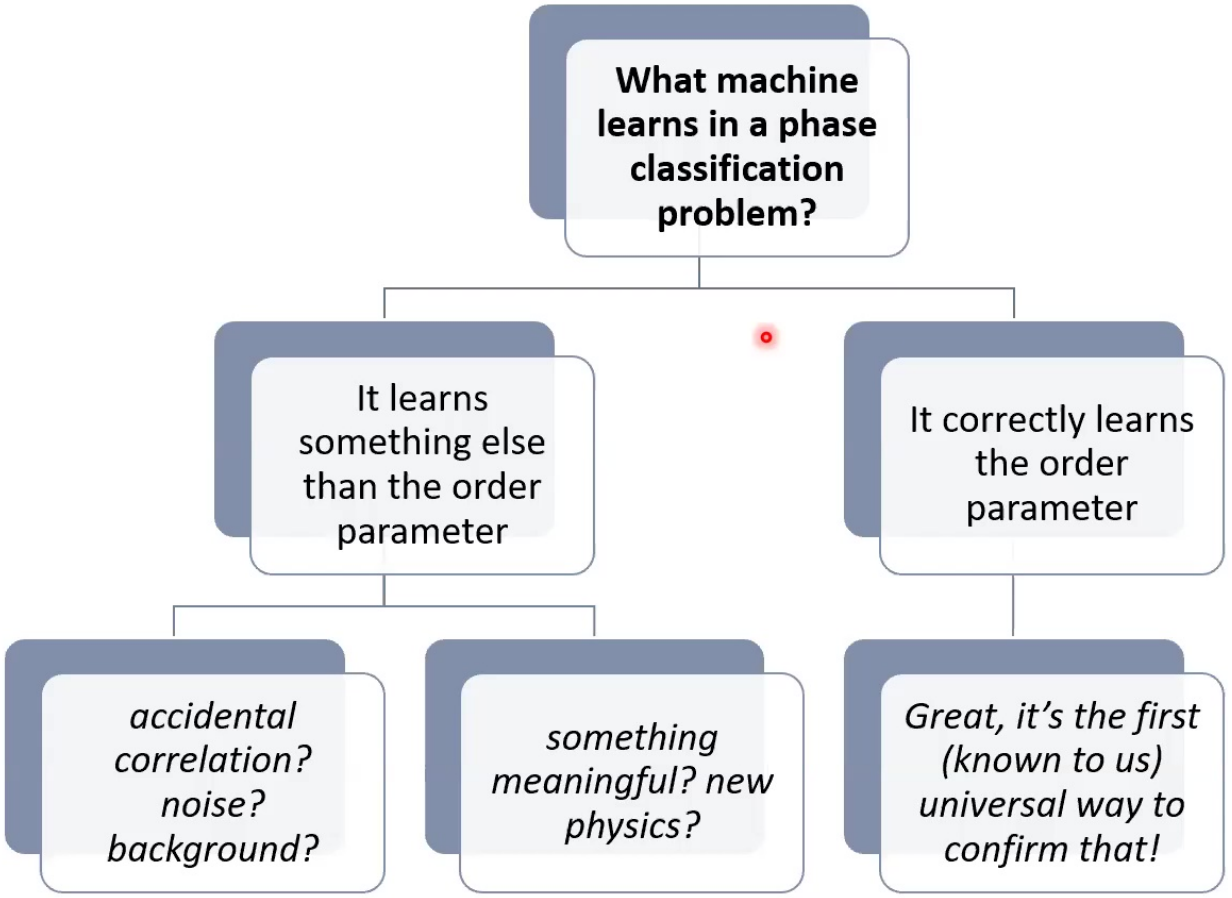
**quantum many-body localization**

- disagreement of predicted critical exponents
- high sensitivity to hyperparameters describing the training process

**topological phases of matter**

- learning schemes trained on raw Monte Carlo configurations were found to be not effective
- pre-engineered features are needed

# Aim and motivation



Taken from: xkcd, A Webcomic of Romance, Sarcasm, Math, And Language, https://xkcd.com/1838/

*People worry that the computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world.*

Pedro Domingos "The Master Algorithm"

! even small invisible changes or a different background context can completely derail predictions

NNs can generate a plausible caption for an image without ever looking at that image !

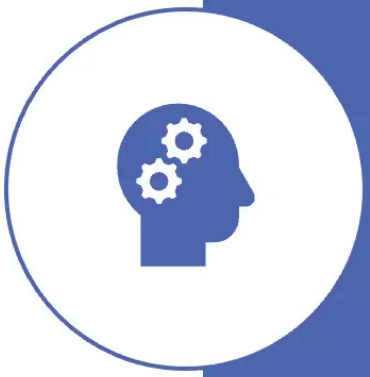
! high error rates for faces from minority groups

the algorithm's hiring and insurance decisions are biased towards selecting men and white people !

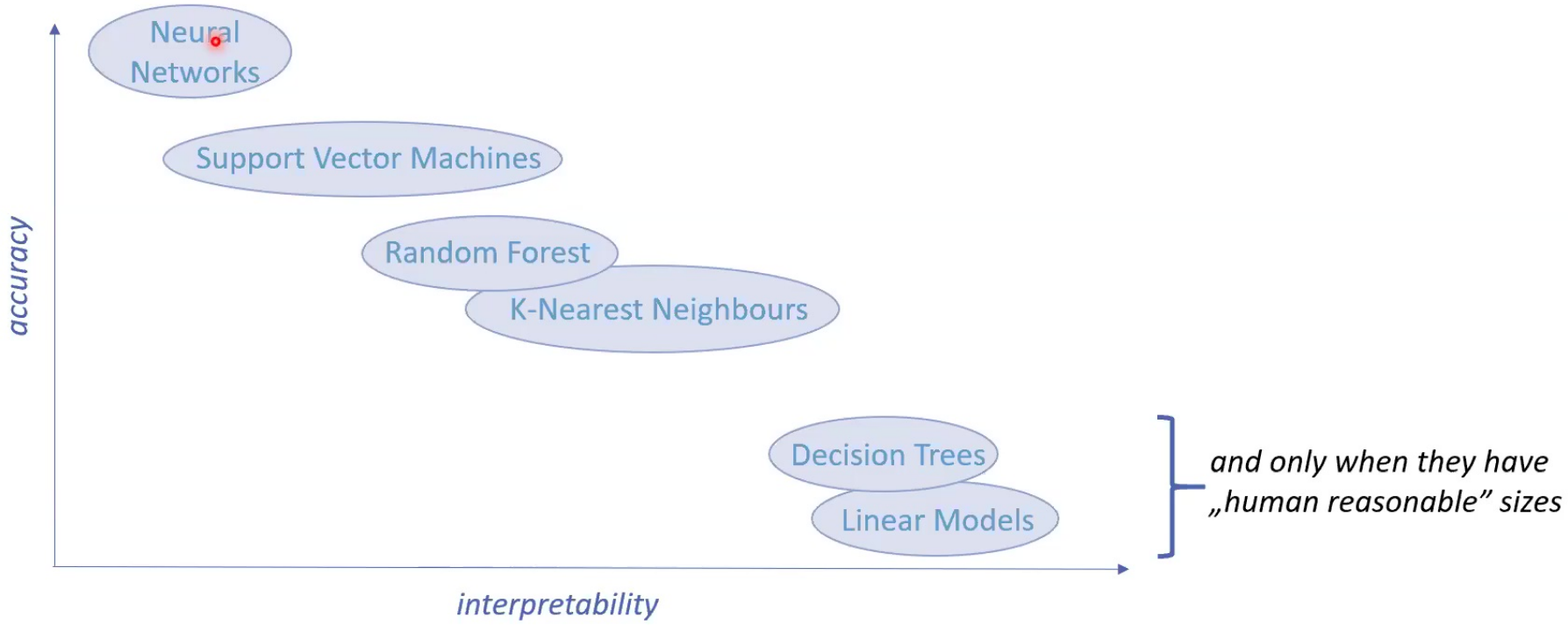
# Some definitions

***Black boxes***
Systems that hide their internal logic to the user
(either the internals are unknown or uninterpretable to humans)

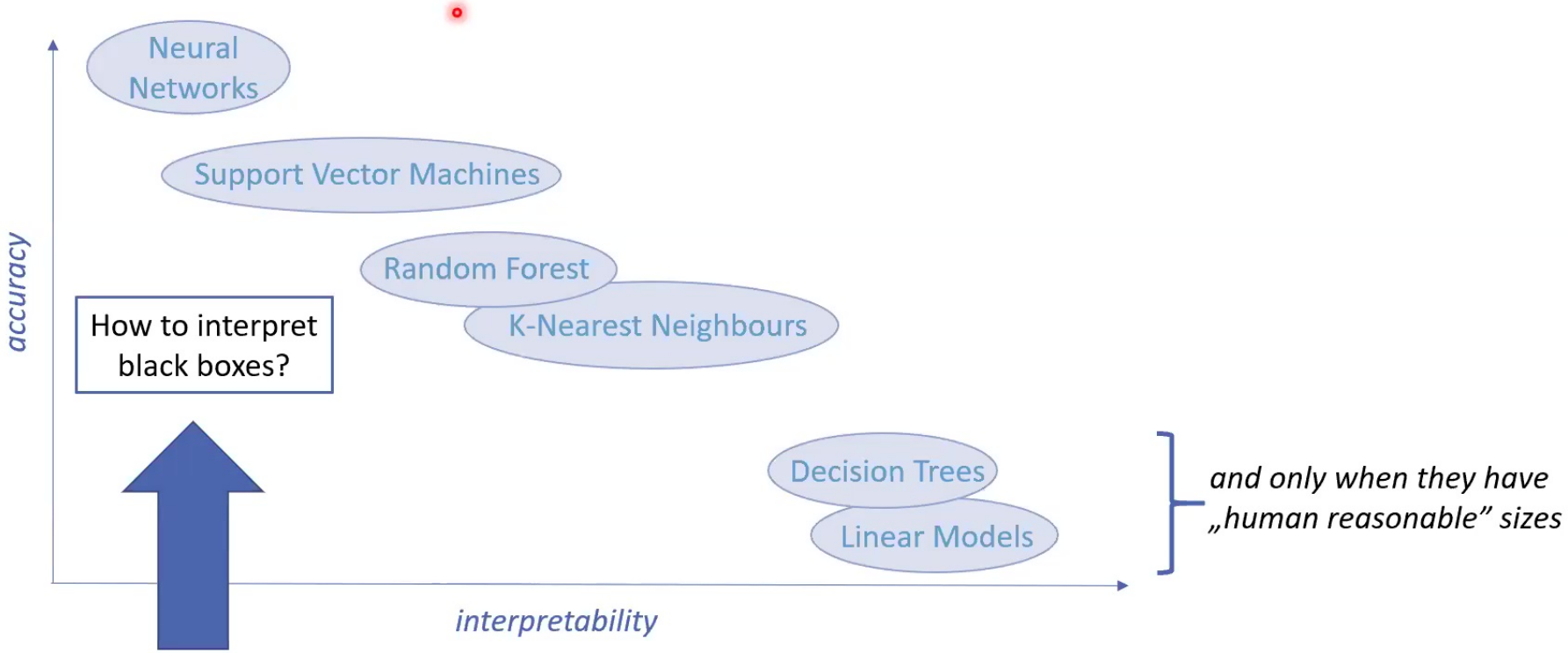
***Interpretability (*also: *comprehensibility)***
Ability to explain or to present in understandable terms to a human



# Trade-off between complexity and interpretability



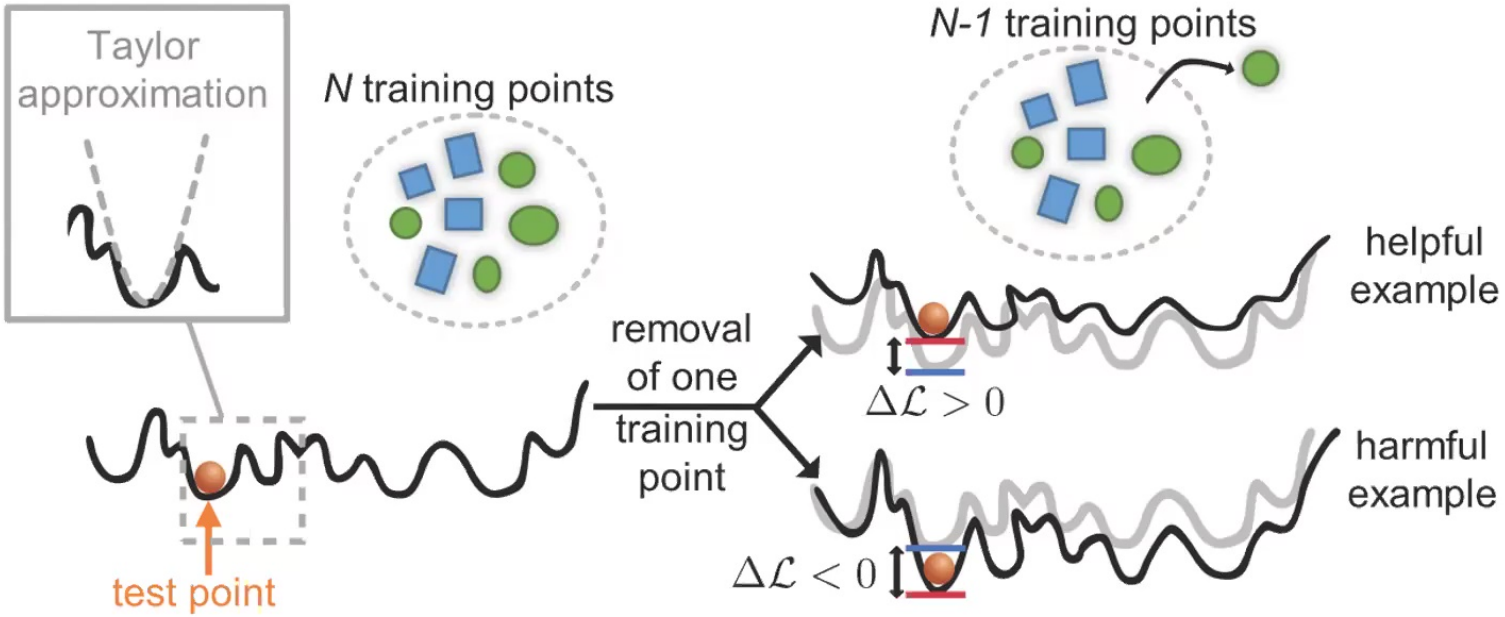
# Trade-off between complexity and interpretability



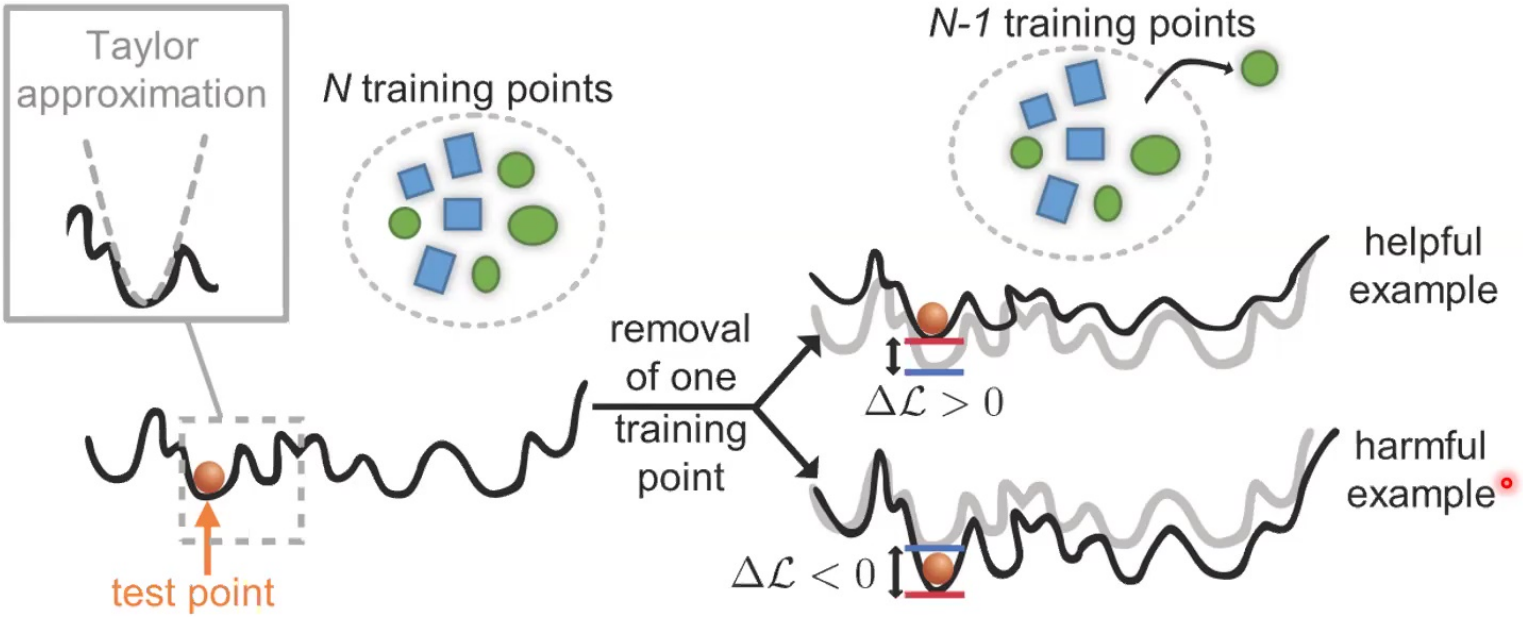
# Interpretable machine learning

Influence functions!
Approximation of…

# Leave-one-out training

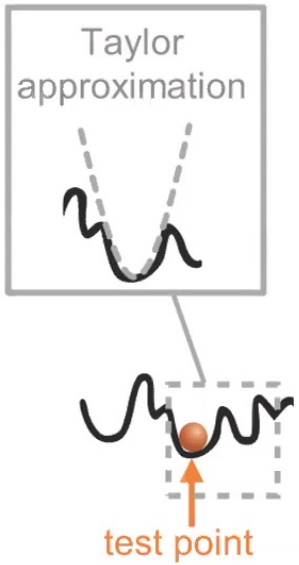


# Leave-one-out training




prohibitively expensive!

# Influence functions




**Analytical approximation for leave-one-out training**

$$\mathcal{I}(z_{\mathrm{r}}, z_{\mathrm{test}}) = \frac{1}{n} \nabla_\theta \mathcal{L}(z_{\mathrm{test}}, \hat{\theta})^T H_\theta^{-1}(\hat{\theta}) \nabla_\theta \mathcal{L}(z_{\mathrm{r}}, \hat{\theta})$$

$\hat{\theta}$ - set of parameters corresponding to the minimum of the training loss function

Assumption: Hessian is positive-definite.

Generalization to non-convex models was done by Koh & Liang: arXiv:1703.04730, ICML 2017's best paper

# Influence functions

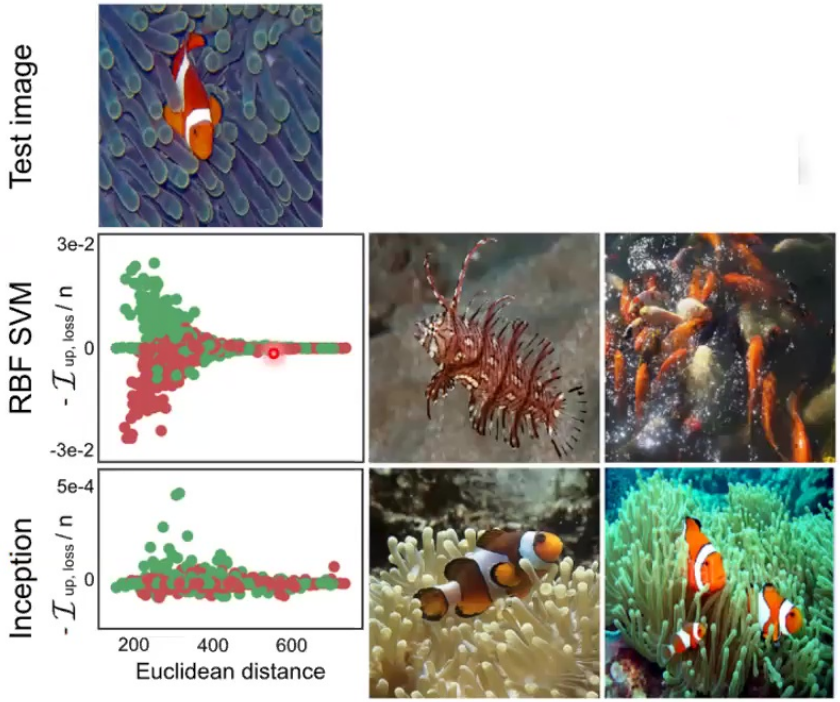



*Figure 4.* **Inception vs. RBF SVM. Bottom left:** $-\mathcal{I}_{\text{up,loss}}(z, z_{\text{test}})$ vs. $\|z - z_{\text{test}}\|_2^2$. Green dots are fish and red dots are dogs. **Bottom right:** The two most helpful training images, for each model, on the test.

Koh & Liang, arXiv:1703.04730, ICML 2017

PHYSICAL REVIEW RESEARCH **2**, 023283 (2020)

**Interpreting machine learning of topological quantum phase transitions**

Yi Zhang,[1,2,*] Paul Ginsparg,[1,†] and Eun-Ah Kim[1,‡]
[1]*Department of Physics, Cornell University, Ithaca, New York 14853, USA*
[2]*International Center for Quantum Materials, Peking University, Beijing 100871, China*

PHYSICAL REVIEW B **97**, 174435 (2018)

**Parameter diagnostics of phases and phase transition learning by neural networks**

Philippe Suchsland and Stefan Wessel
*Institut für Theoretische Festkörperphysik, JARA-FIT and JARA-HPC, RWTH Aachen University, 52056 Aachen, Germany*

PHYSICAL REVIEW B **96**, 184410 (2017)

**Machine learning of explicit order parameters: From the Ising model to SU(2) lattice gauge theory**

Sebastian J. Wetzel and Manuel Scherzer
*Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 16, 69120 Heidelberg, Germany*
(Received 27 May 2017; revised manuscript received 27 October 2017; published 8 November 2017)

PHYSICAL REVIEW E **98**, 022138 (2018)

**Smallest neural network to learn the Ising criticality**

Dongkyu Kim and Dong-Hee Kim*
*Department of Physics and Photon Science, School of Physics and Chemistry, Gwangju Institute of Science and Technology, Gwangju 61005, Korea*

**Discovering Symmetry Invariants and Conserved Quantities by Interpreting Siamese Neural Networks**

Sebastian J. Wetzel,[1] Roger Melko,[1,2] Joseph Scott,[3] Maysum Panju,[3] and Vijay Ganesh[3]
[1]*Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada*
[2]*Department of Physics and Astronomy, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada*
[3]*Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada*

PHYSICAL REVIEW B **96**, 205146 (2017)

**Kernel methods for interpretable machine learning of order parameters**

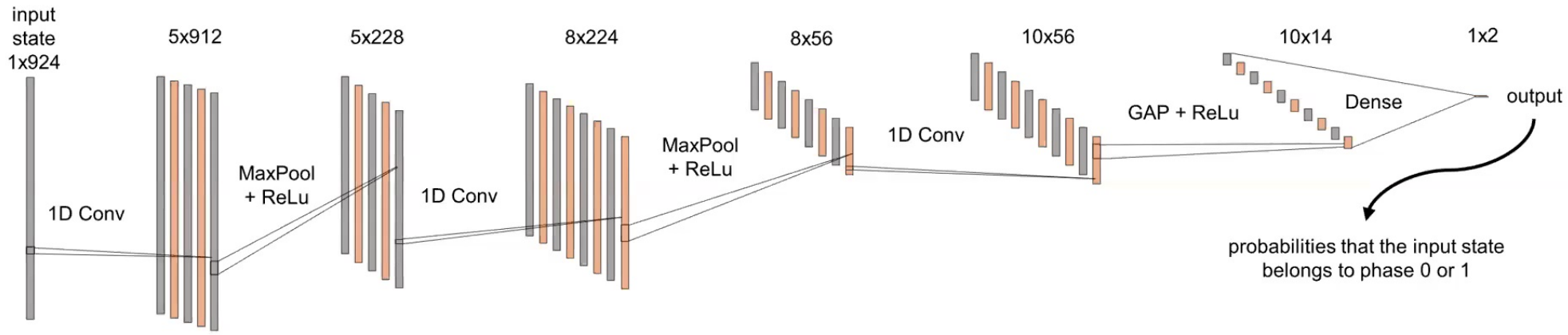
Pedro Ponte and Roger G. Melko
*Department of Physics and Astronomy, University of Waterloo, Ontario N2L 3G1, Canada and Perimeter Institute of Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada*
(Received 18 May 2017; published 27 November 2017)

# Interpretable ML in quantum physics

# Outline

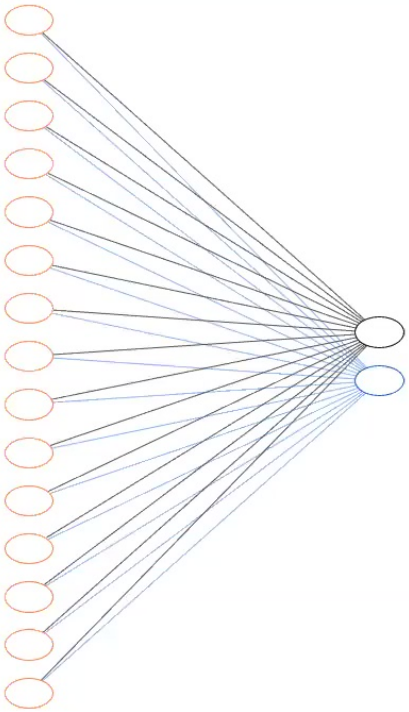
1. Machine learning in phase classification
   *aim and motivation*
2. Interpretability of machine learning
   *methods*
3. ML model
   *methods*
4. Physical model
   *methods*
5. What the machine learns in a phase classification problem?
   *results*

# ML model we use: 1D CNN

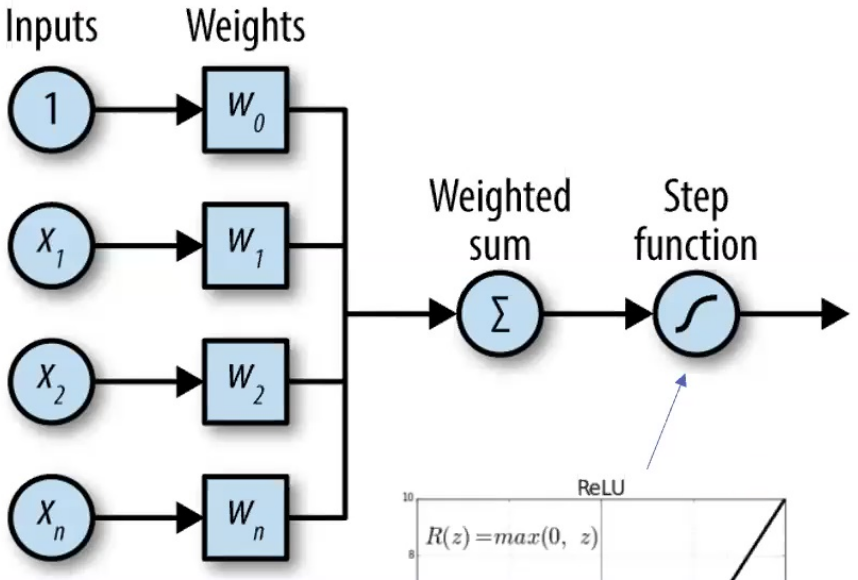


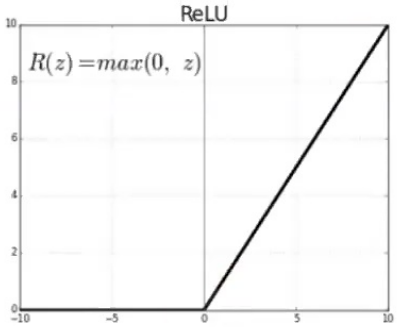

- Convolutional layers
- Max pooling
- ReLU = Rectified Linear Unit
- GAP = Global Average Pooling
- Dense = Fully-connected layer

# ML model: fully-connected layer



input output



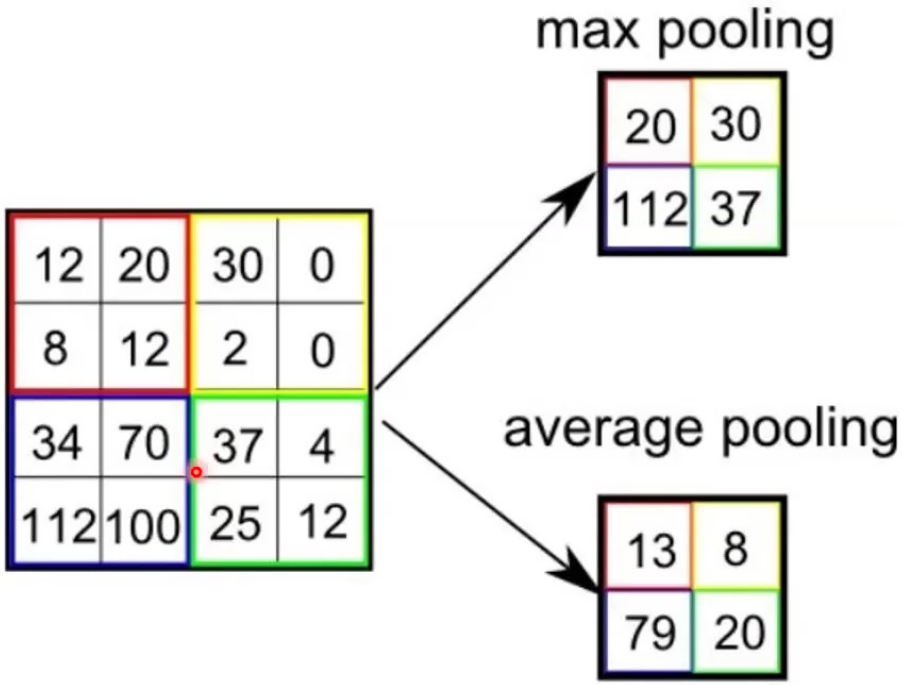
# ML model: convolutional layers, how they operate

Image

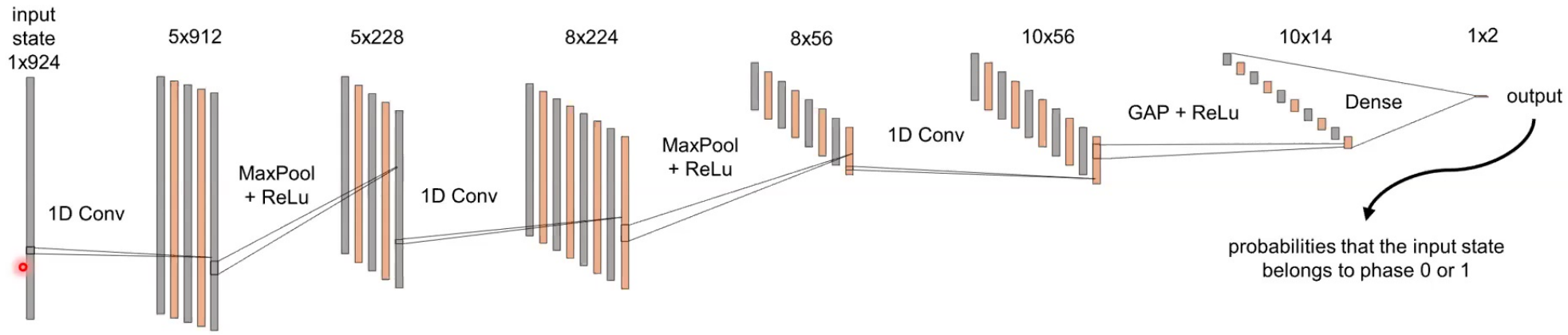
| | | |
|---|---|---|
| 4 | 3 | 4 |
| 2 | 4 | 3 |
| 2 | 3 | 4 |

Convolved Feature

# ML model: pooling layers



# ML model we use: 1D CNN



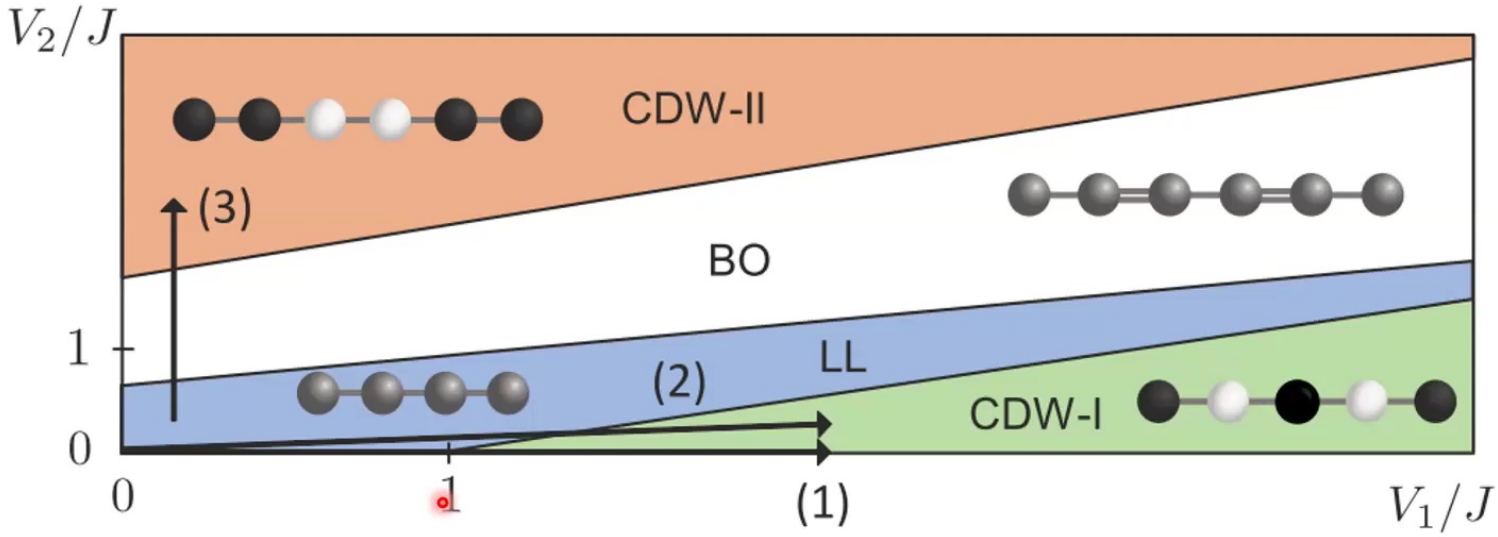

- Convolutional layers
- Max pooling
- ReLU = Rectified Linear Unit
- GAP = Global Average Pooling
- Dense = Fully-connected layer

# Physical model

$$\hat{H} = -J \sum_{\langle i,j \rangle} c_i^\dagger c_j + V_1 \sum_{\langle i,j \rangle} n_i n_j + V_2 \sum_{\langle\langle i,j \rangle\rangle} n_i n_j$$

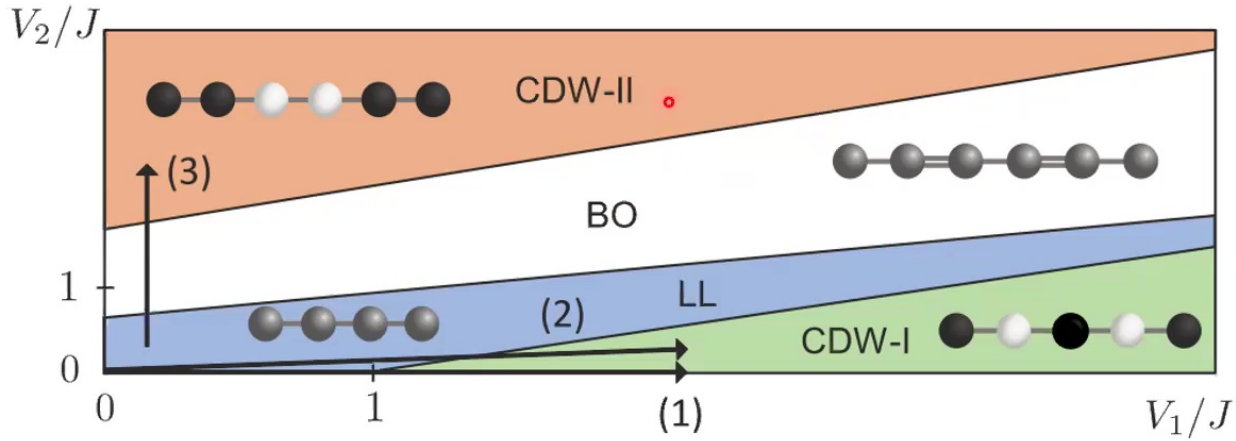
- spinless Fermi-Hubbard model at half-filling
- 1D chain with 12 sites
- periodic boundary conditions
- exact diagonalization with eigenfunctions represented in Fock basis
- fixed gauge (global sign)

# Phase diagram




LL - Luttinger liquid
BO - bond order phase
CDW-I and II - charge density wave type I and II

# Phase diagram



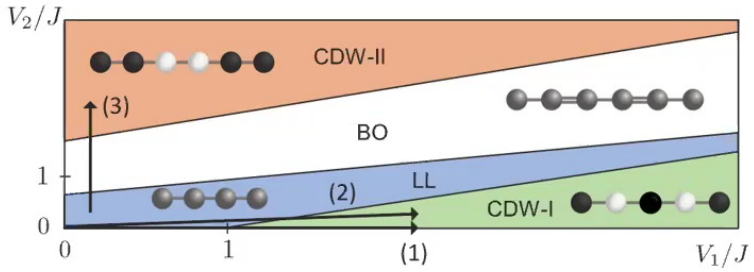

LL - Luttinger liquid
BO - bond order phase
CDW-I and II - charge density wave type I and II

$$O_{\text{CDW-I}} = \frac{1}{L}\sum_{\langle i,j\rangle} |n_i - n_j|$$

$$O_{\text{CDW-II}} = \frac{1}{L}\sum_{\langle\langle i,j\rangle\rangle} |n_i - n_j|$$

$$O_{\text{BO}} = \frac{1}{L}\sum_i (-1)^i B_i, \text{ where } B_i = \left\langle c_i^\dagger c_{i+1} + c_{i+1}^\dagger c_i \right\rangle$$

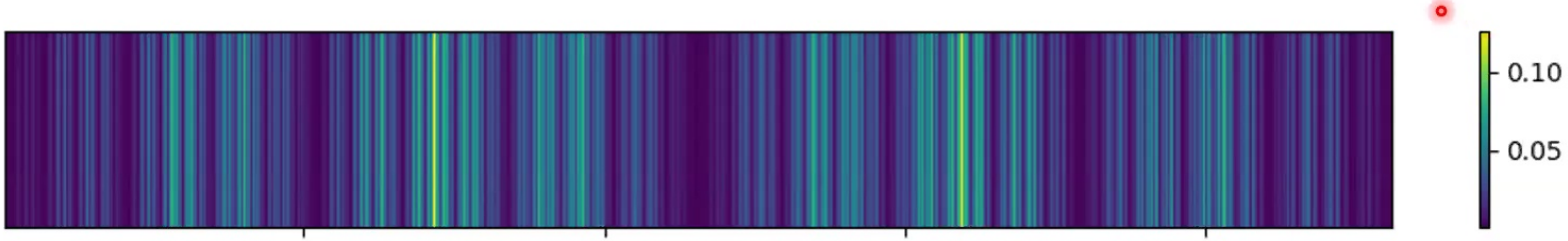
# Phase diagram




LL - Luttinger liquid
BO - bond order phase
CDW-I and II - charge density wave type I and II

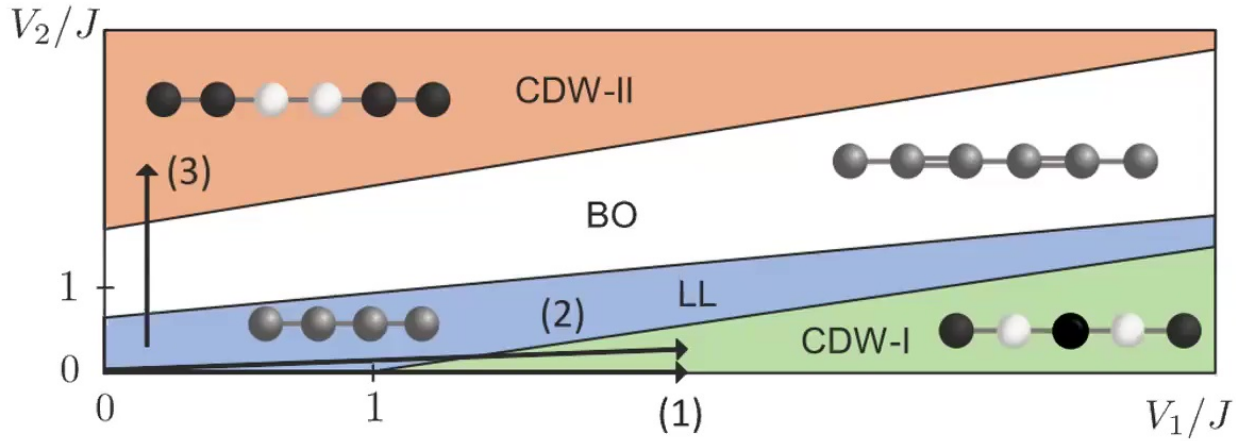
$$O_{\text{CDW-I}} = \frac{1}{L}\sum_{\langle i,j\rangle} |n_i - n_j|$$
$$O_{\text{CDW-II}} = \frac{1}{L}\sum_{\langle\langle i,j\rangle\rangle} |n_i - n_j|$$
$$O_{\text{BO}} = \frac{1}{L}\sum_i (-1)^i B_i, \text{ where } B_i = \left\langle c_i^\dagger c_{i+1} + c_{i+1}^\dagger c_i \right\rangle$$

What machine gets




+ label: e.g., 0 for LL, 1 for CDW-I

# Phase diagram



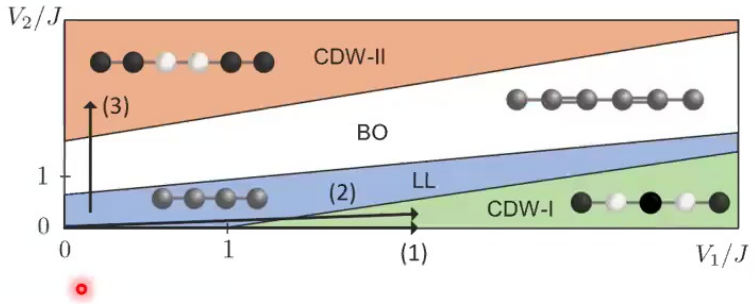

LL - Luttinger liquid
BO - bond order phase
CDW-I and II - charge density wave type I and II

$$O_{\text{CDW-I}} = \frac{1}{L}\sum_{\langle i,j\rangle} |n_i - n_j|$$

$$O_{\text{CDW-II}} = \frac{1}{L}\sum_{\langle\langle i,j\rangle\rangle} |n_i - n_j|$$

$$O_{\text{BO}} = \frac{1}{L}\sum_i (-1)^i B_i, \text{ where } B_i = \left\langle c_i^\dagger c_{i+1} + c_{i+1}^\dagger c_i \right\rangle$$

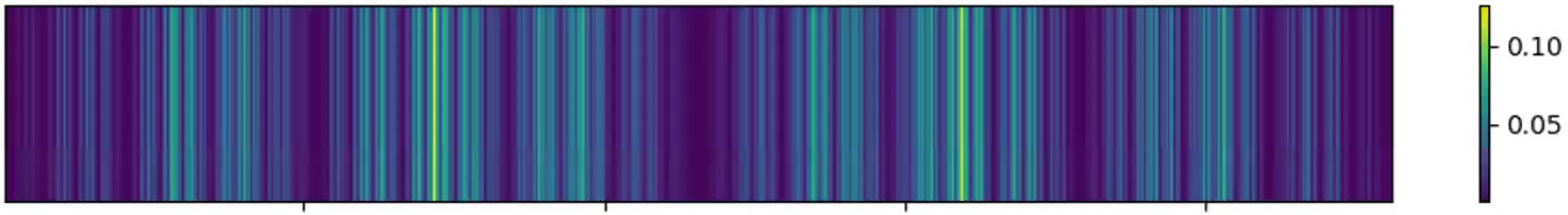
# Phase diagram




LL - Luttinger liquid
BO - bond order phase
CDW-I and II - charge density wave type I and II

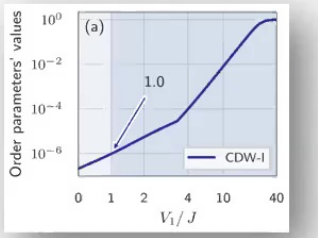
$$O_{\text{CDW-I}} = \frac{1}{L}\sum_{\langle i,j\rangle} |n_i - n_j|$$
$$O_{\text{CDW-II}} = \frac{1}{L}\sum_{\langle\langle i,j\rangle\rangle} |n_i - n_j|$$
$$O_{\text{BO}} = \frac{1}{L}\sum_i (-1)^i B_i, \text{ where } B_i = \left\langle c_i^\dagger c_{i+1} + c_{i+1}^\dagger c_i \right\rangle$$

What machine gets

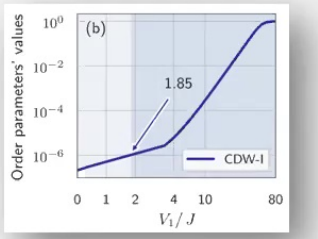



+ label: e.g., 0 for LL, 1 for CDW-I

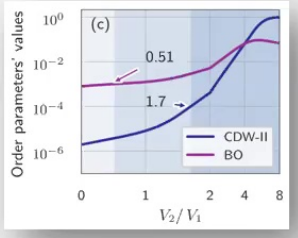
Three set-ups




LL – CDW(I) transition line,
$V_2$=0

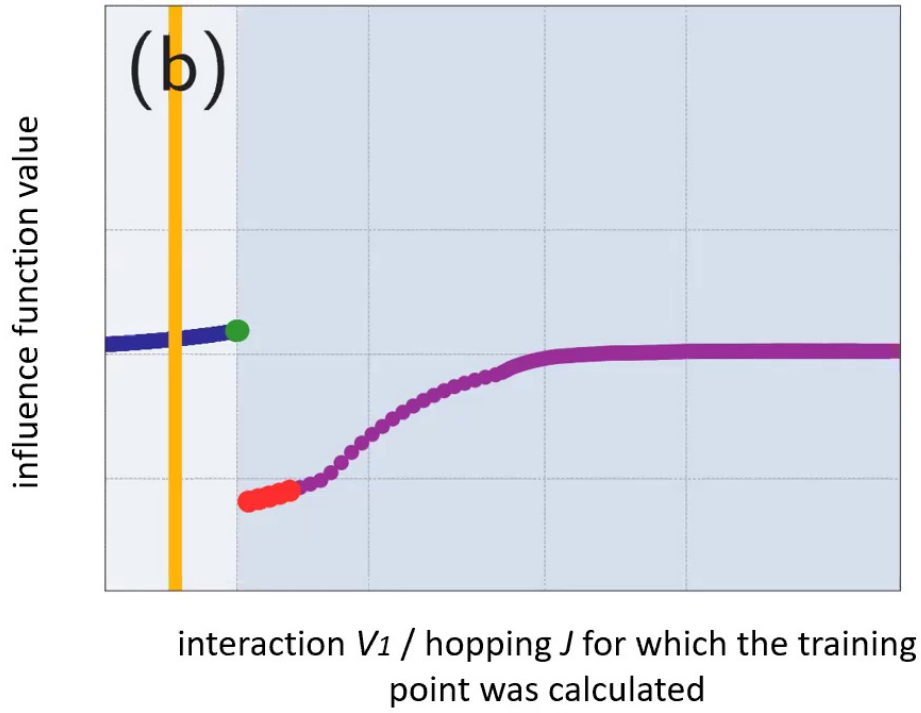



Transfer learning:
LL – CDW(I) transition lines
$V_2$=0 and $V_2$=0.25 $V_1$

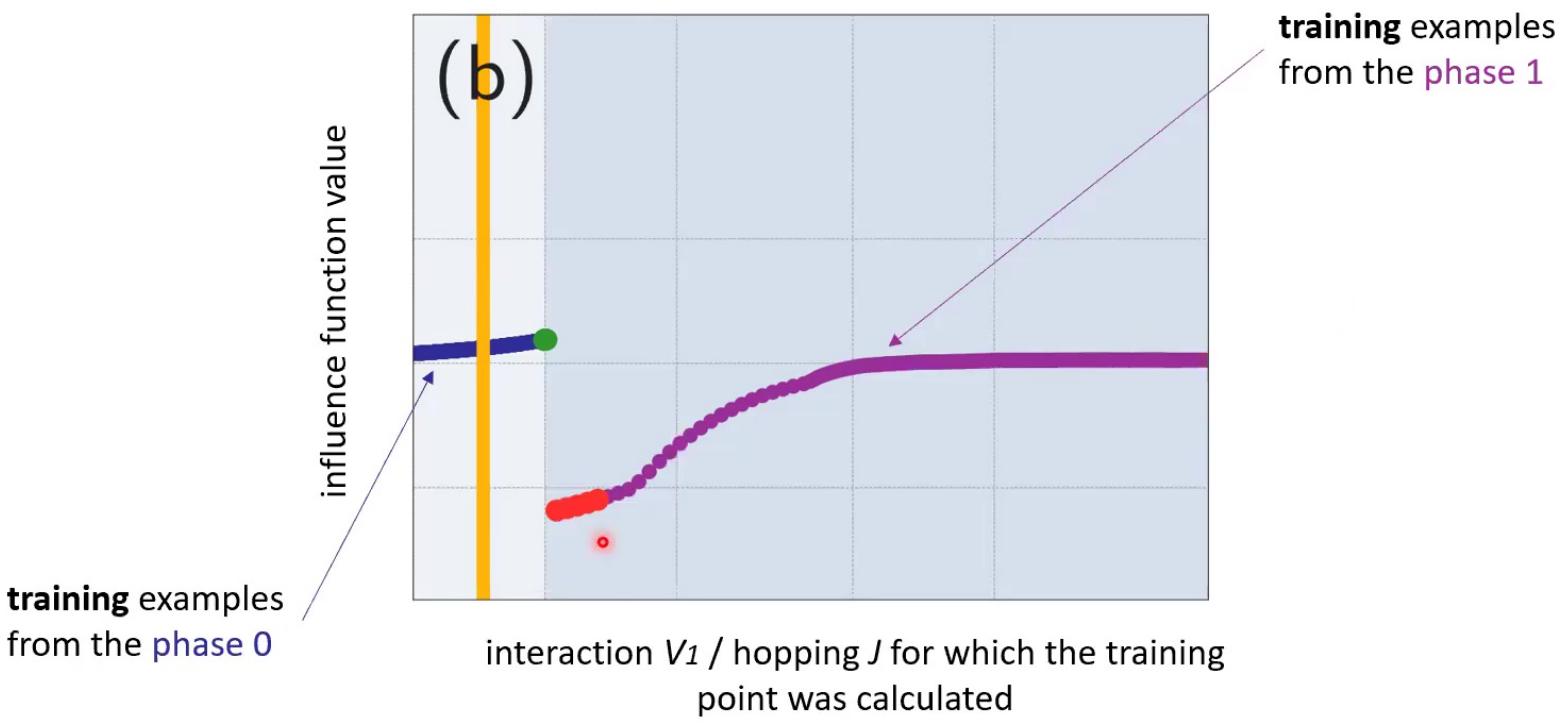



LL-BO-CDW(II) transition line

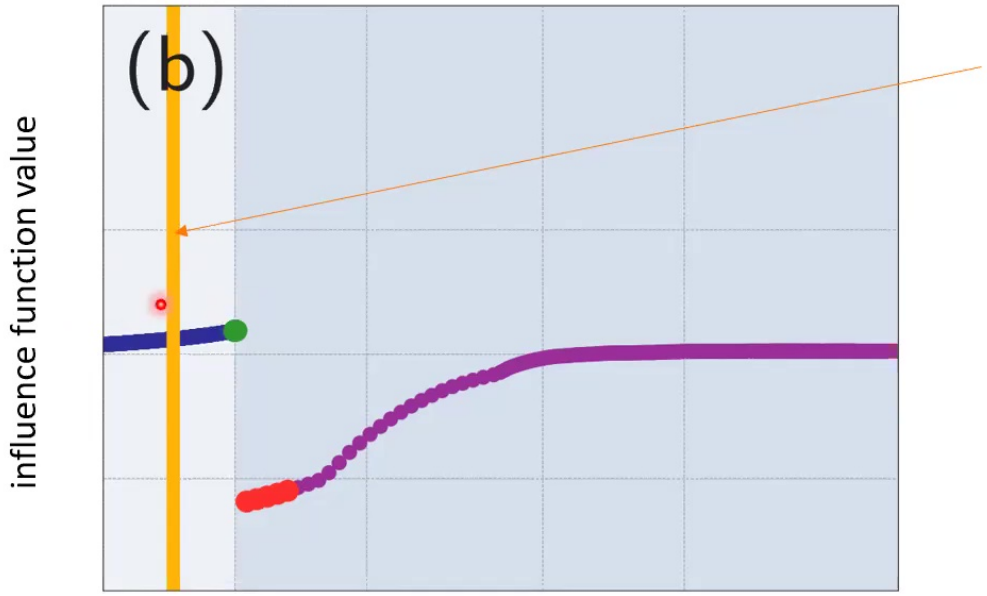
## Prelude: What do we see on the figure?



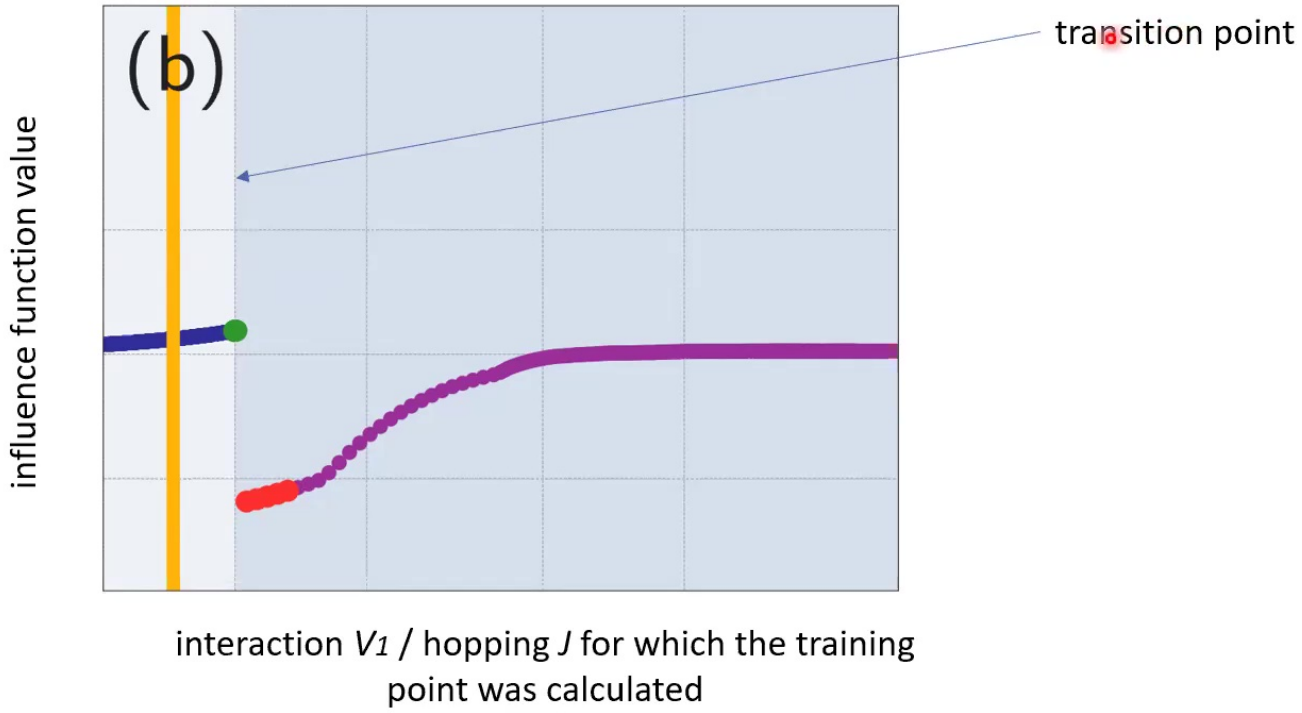
## Prelude: What do we see on the figure?



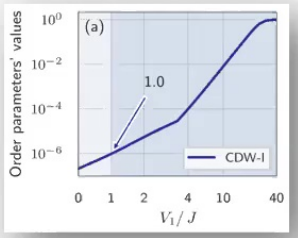
## Prelude: What do we see on the figure?



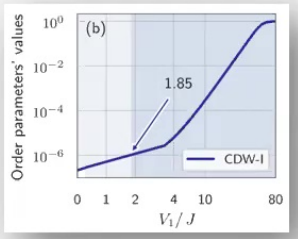
## Prelude: What do we see on the figure?



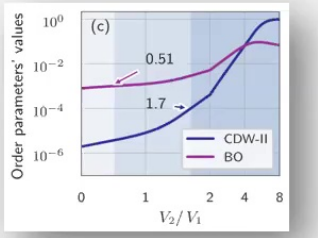
## Three set-ups




LL – CDW(I) transition line, $V_2$=0

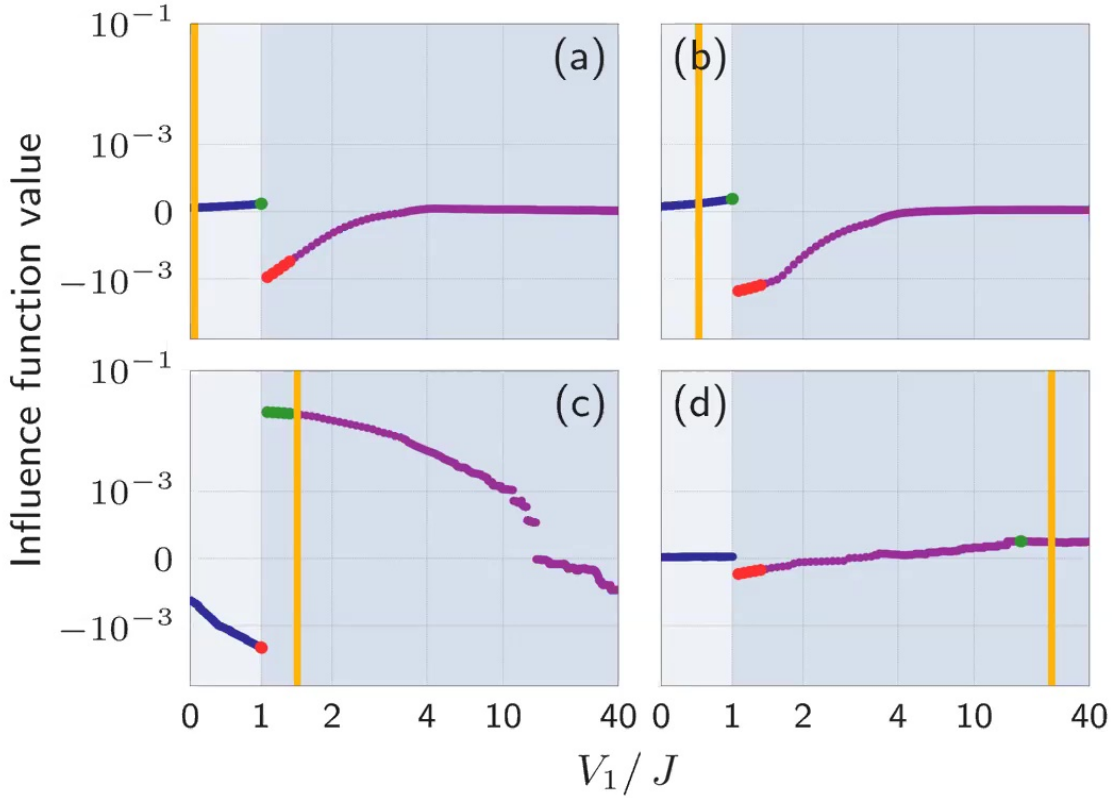


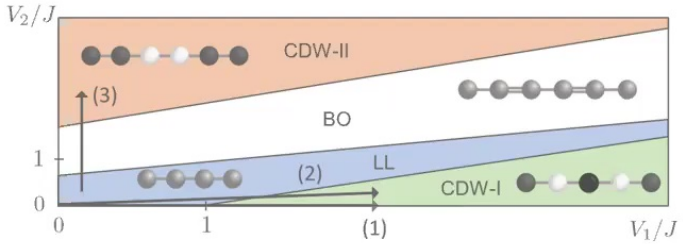

Transfer learning:
LL – CDW(I) transition lines
$V_2$=0 and $V_2$=0.25 $V_1$

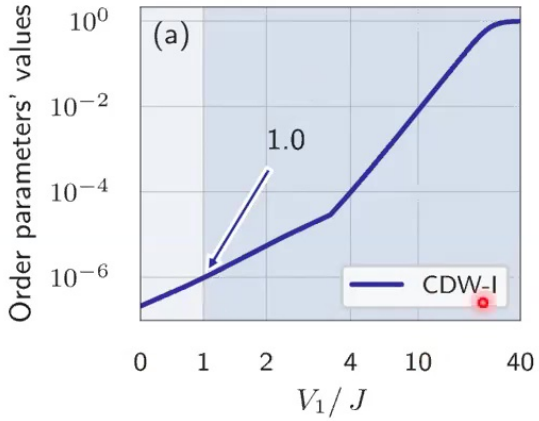



LL-BO-CDW(II) transition line

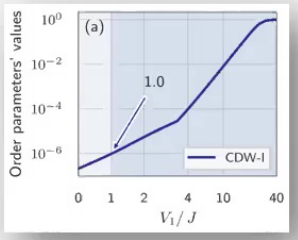
## LL – CDW(I) transition line, $V_2$=0 – *arrow no. 1*



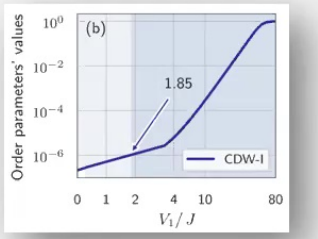



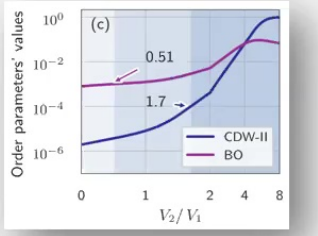
## Three set-ups




LL – CDW(I) transition line,
$V_2$=0

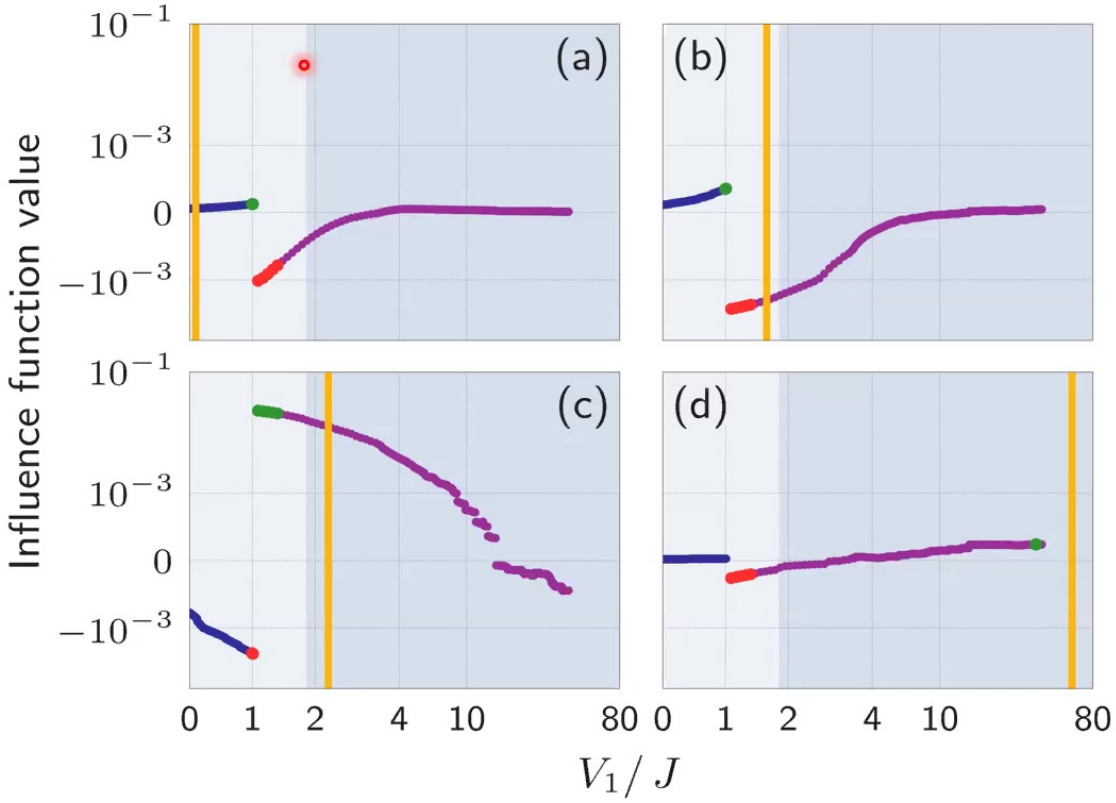


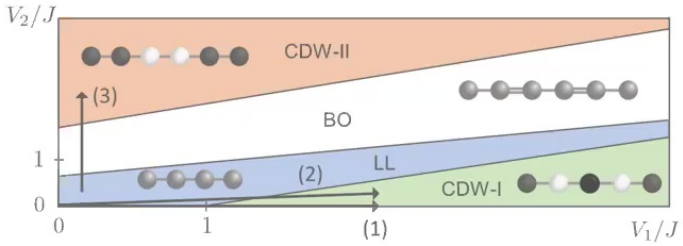

Transfer learning:
LL – CDW(I) transition lines
$V_2$=0 and $V_2$=0.25 $V_1$

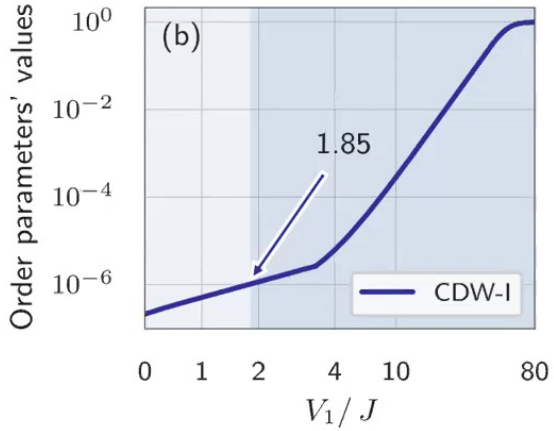



LL-BO-CDW(II) transition line

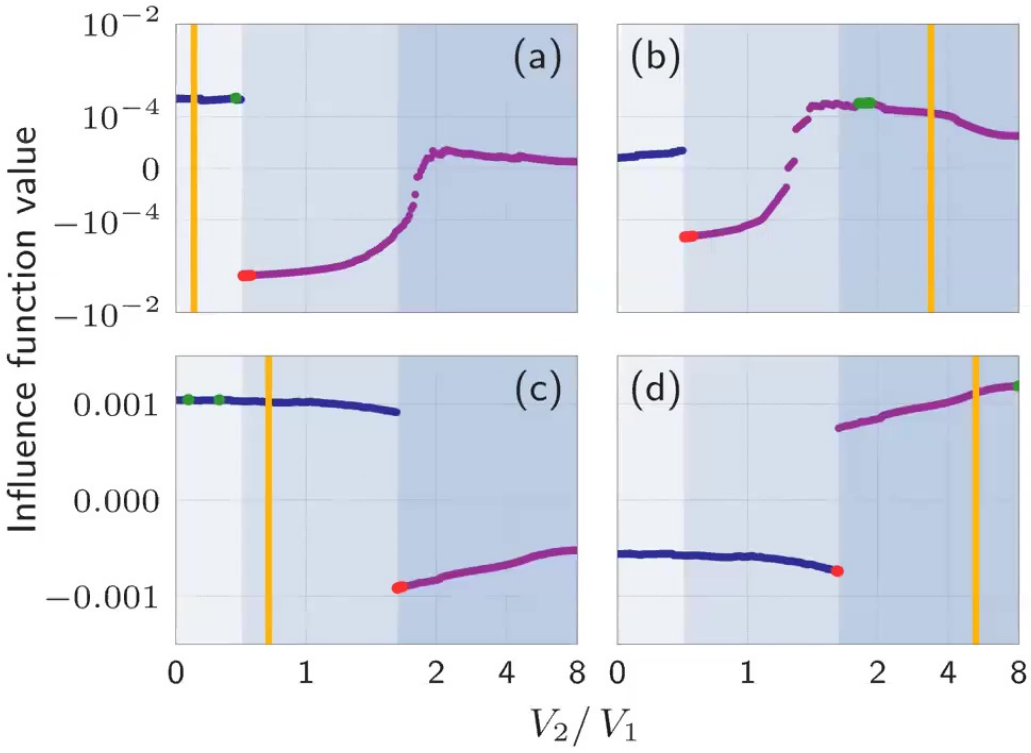
## Transfer learning: LL – CDW(I) transition lines, $V_2$=0 and $V_2$=0.25 $V_1$ – *arrows no. 1 and 2*

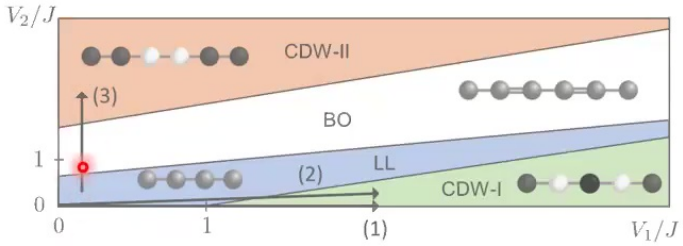


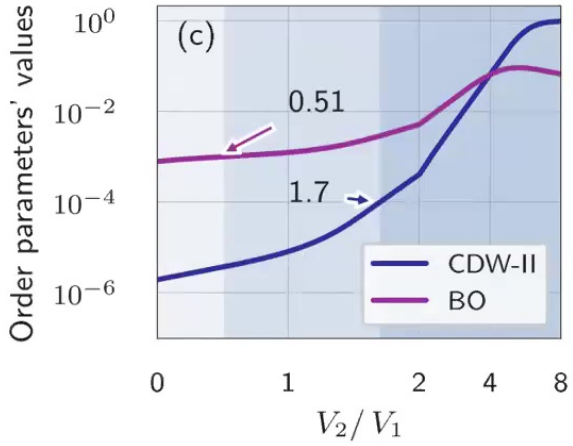



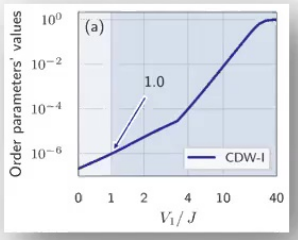
## LL-BO-CDW(II) transition line – *arrow no. 3*



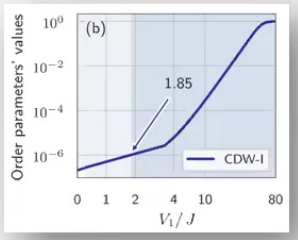



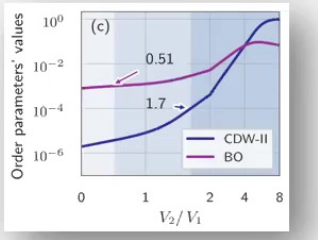
## Three set-ups




LL – CDW(I) transition line,
$V_2$=0

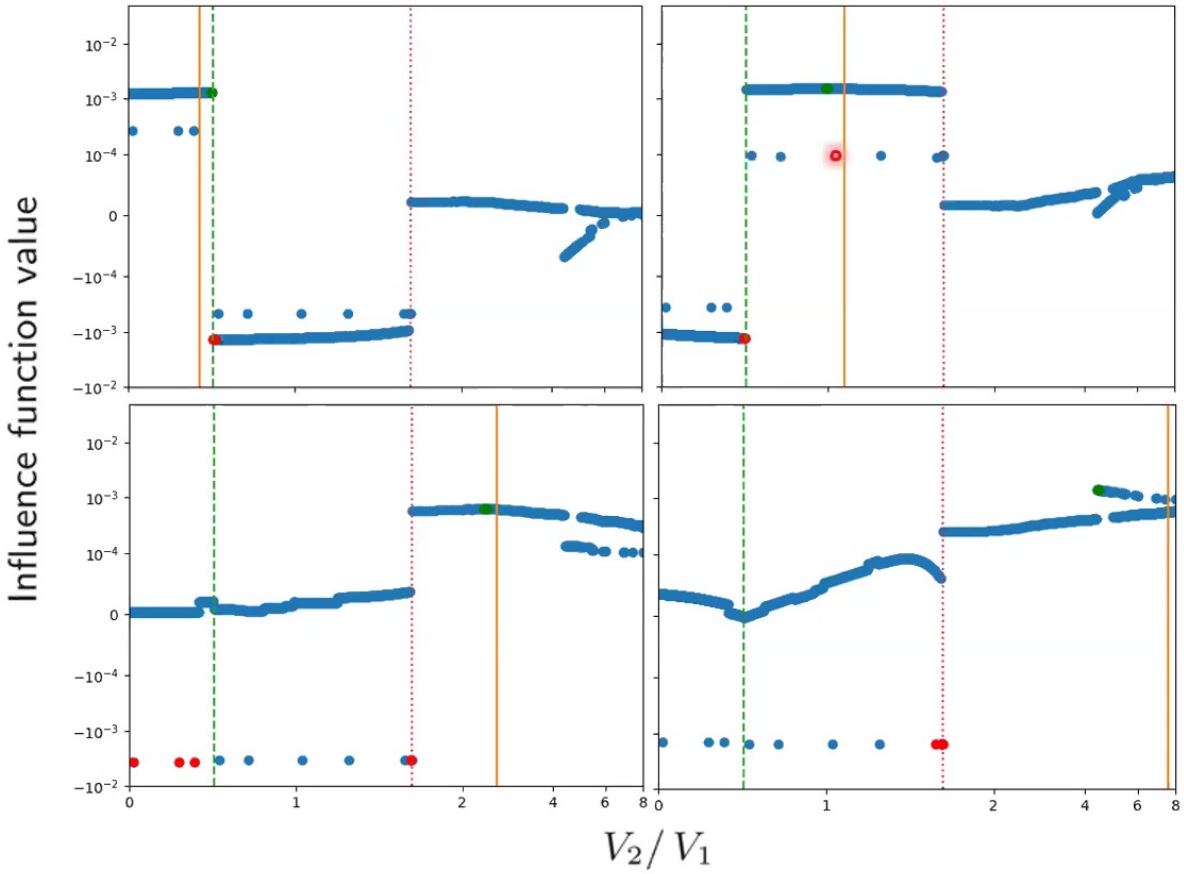


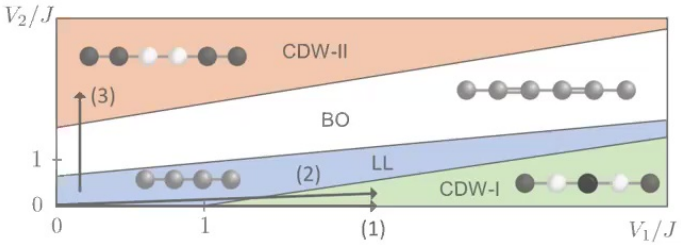

Transfer learning:
LL – CDW(I) transition lines
$V_2$=0 and $V_2$=0.25 $V_1$

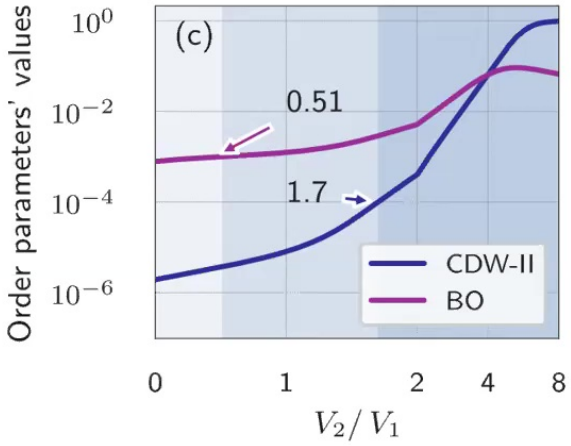



LL-BO-CDW(II) transition line

## The same LL-BO-CDW(II) transition line – *arrow no. 3 – NO GAUGE FIXING*



# Conclusions and outlook

Strong indication that the ML algorithm indeed learned a relevant order parameter

→ *need for common use of interpretability methods in science!*

Influence functions as a tool to guide the search for unknown order parameters, notion of similarity

→ *MBL, topological models?*

Interpretable NN can detect unknown phases!

→ *again, MBL, topological models?*

Tool to detect outliers and judge the quality of the training

→ *Distinguishing between types of phase transition?*
→ *Analysis of NNs trained to classify experimental data? What about noise?*

# Thank you for your attention!

https://arxiv.org/abs/2004.04711

code sample:
DOI 5281/zenodo.3759432