Reliable AI: Successes, Challenges, and Limitations

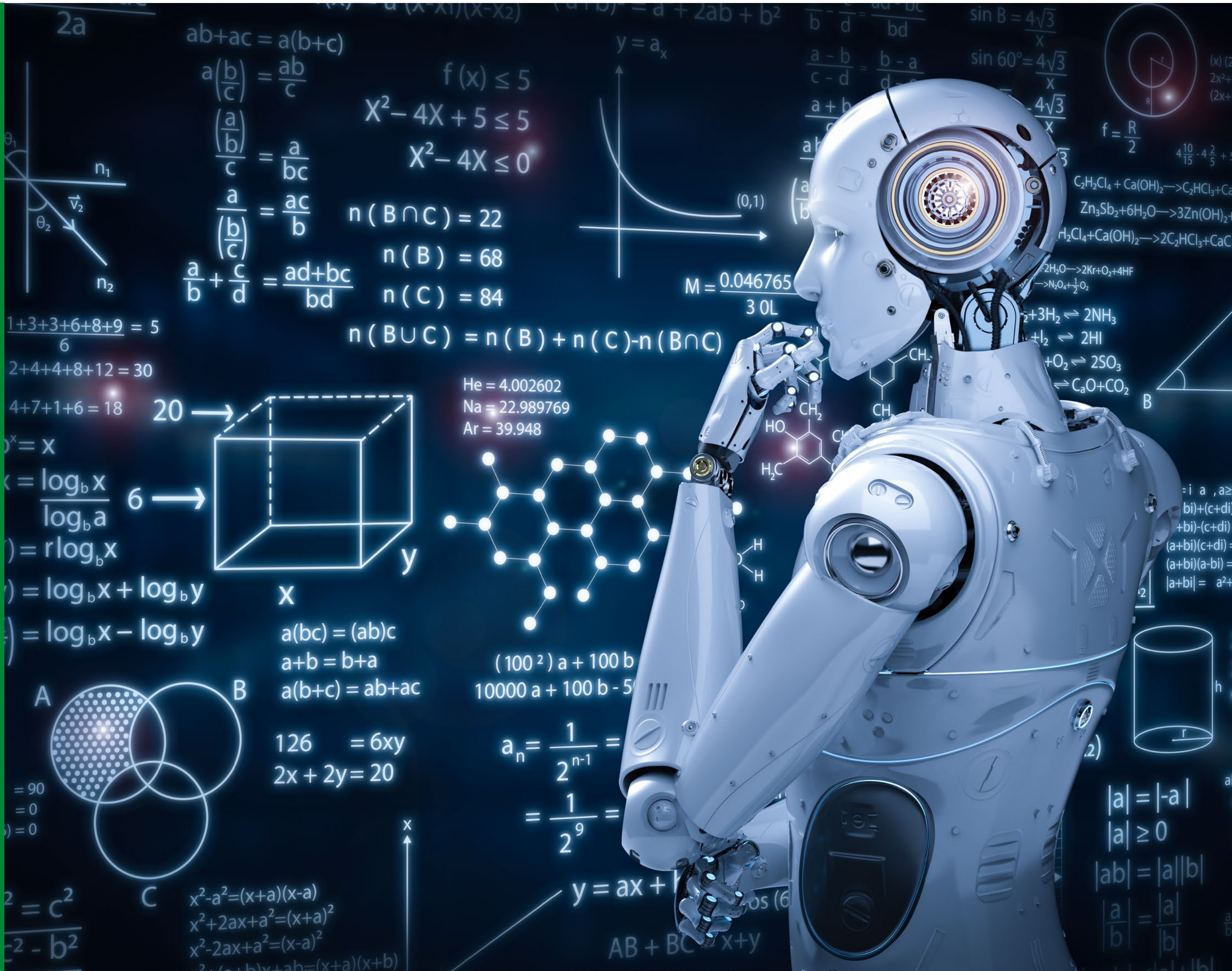
Gitta Kutyniok

LMU Munich

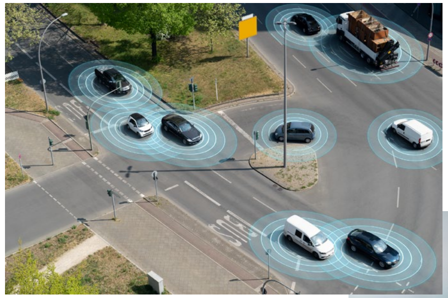
(also University of Tromsø, Norway & DLR – German Aerospace Center)

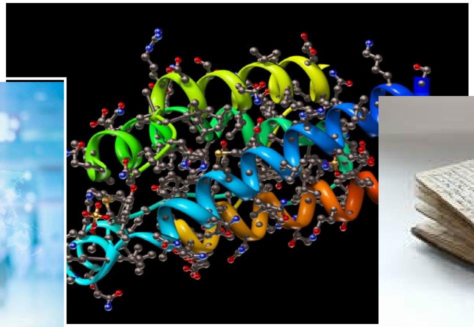
FedCSIS 2024

Belgrad, Serbia, 8-11 September, 2024



Fourth Industrial Revolution by Artificial Intelligence





Radical Change of our Society in its Full Breadth!

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Artificial Intelligence = Alchemy?

AAAS | Science

AI researchers allege that machine learning is alchemy

By **Matthew Hutson** | May. 3, 2018 , 11:15 AM

Ali Rahimi, a researcher in artificial intelligence (AI) at Google in San Francisco, California, took a swipe at his field last December—and received a 40-second ovation for it. Speaking at an AI conference, Rahimi charged that machine learning algorithms, in which computers learn through trial and error, **have become a form of "alchemy."** Researchers, he said, do not know why some algorithms work and others don't, nor do they have rigorous criteria for choosing one AI architecture over another. Now, in a paper presented on 30 April at the International Conference on Learning Representations in Vancouver, Canada, Rahimi and his collaborators **document examples** of what they see as the alchemy problem and offer prescriptions for bolstering AI's rigor.

MATH 4AI

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

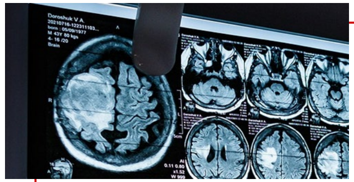
Challenges in Reliable AI

Problems with Safety

Example:
Accidents involving robots

Problems with Security

Example:
Risks of hacking into AI systems



Problems with Privacy

Example:
Privacy violations of health data

Problems with Responsibility

Example:
Black-box and biased decisions

Current major problem worldwide:

Lack of reliability of AI technology!

MATH 4AI

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Strong Requirements for Reliability

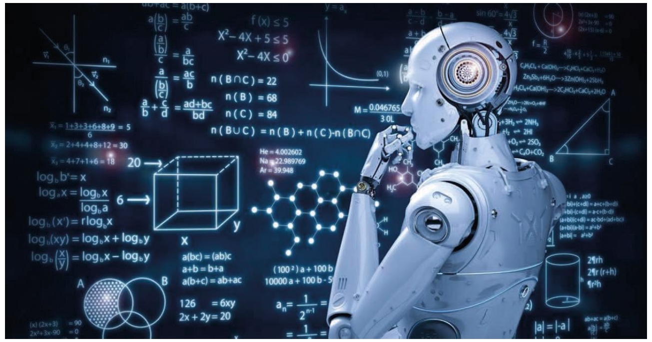
International Position concerning Reliable AI:

→ AI Act of the European Union

→ G7 Hiroshima AI Process

Major Challenge:

Derive a profound mathematical understanding!



A Mathematical Perspective on Reliability

Deep Neural Networks

Deep neural networks are a work horse for artificial intelligence!

Key Goal of McCulloch and Pitts (1943):

→ Introduce *artificial Intelligence!*

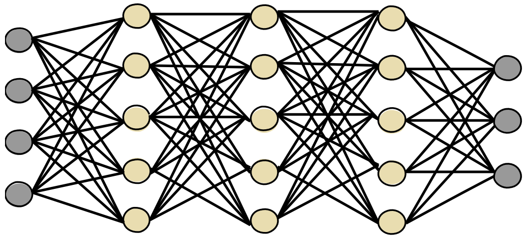
Artificial Neurons:

$$f(x_1, ..., x_n) = \rho\left(\sum_{i=1}^{n} x_i w_i - b\right)$$

Definition of a Neural Network:

A *deep neural network* is a function $\Phi : \mathbb{R}^d \to \mathbb{R}^{N_L}$ of the form

$$\Phi(x) = T_L \rho(T_{L-1}\rho(\ldots\rho(T_1(x))), \quad x \in \mathbb{R}^d,$$



with

$$T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, \ \ell = 1, \ldots, L, \text{ where } T_\ell x = W^{(\ell)} x + b^{(\ell)}$$

A Mathematical Understanding of Deep Learning

Expressivity:

→ Which *aspects of a neural network architecture* affect the performance of deep learning?

Applied Harmonic Analysis, Approximation Theory, ...

Learning:

→ Why does *stochastic gradient descent* converge to good local minima despite the non-convexity of the problem?

Algebraic/Differential Geometry, Optimal Control, Optimization, ...

Generalization:

→ Can we derive overall *success guarantees* (on the test data set)?

Learning Theory, Probability Theory, Statistics, ...

Explainability:

→ Why did a trained deep neural network *reach a certain decision*?

Information Theory, Uncertainty Quantification, ...

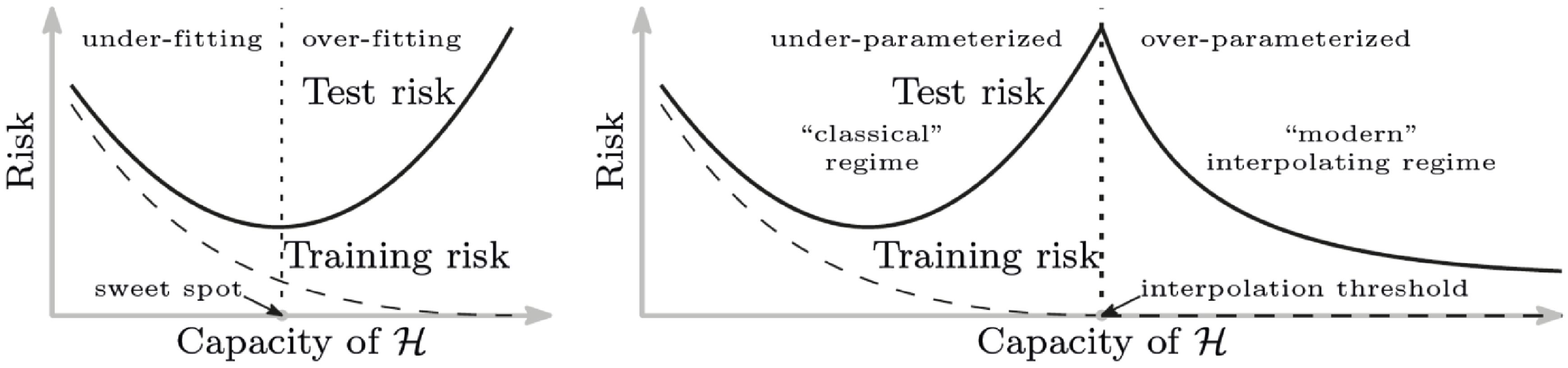
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Generalization: Mathematical Success Guarantees

Understanding the Amazing Generalization Ability of Deep Neural Networks

Why do neural networks perform that well in the high-parameter regime?

Can we estimate the generalization error?



(Source: Belkin, Hsu, Ma, Mandal; 2019)

Some Common Approaches:

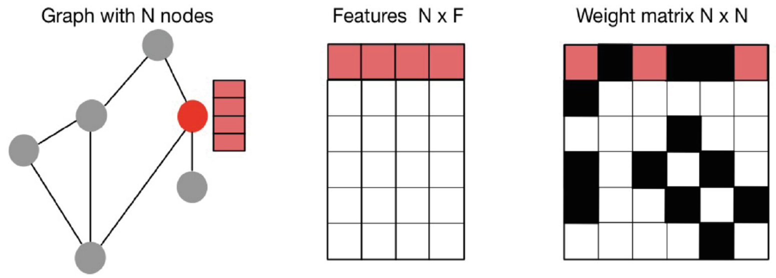
- → VC dimension
- → Rademacher complexity
- → Neural tangent kernels

Goal: Error Bounds for the performance on unseen data!

Graph Neural Networks

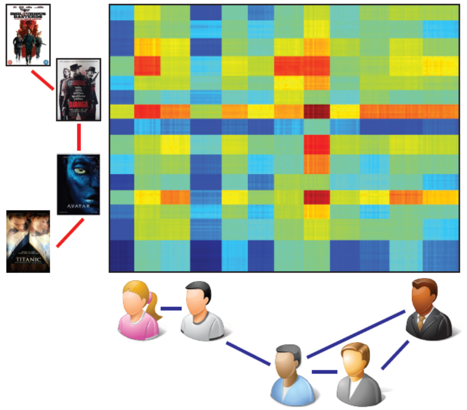
Graph neural networks generalize classical neural networks to signals over graph domains.

Graph signal:

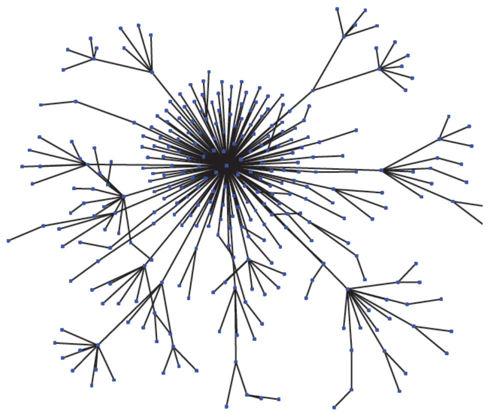


$$s: \text{ graph nodes } \to \mathbb{R}^c$$

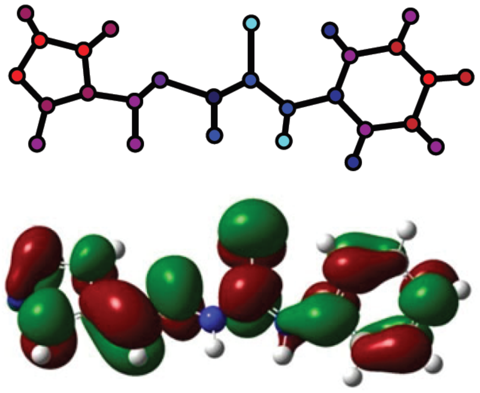
Exemplary Applications:



Recommender system



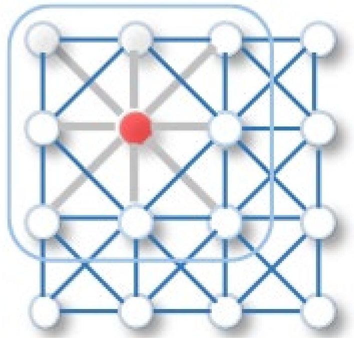
Fake news detection

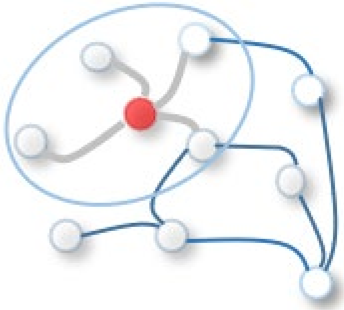


Chemistry

Graph (Convolutional) Neural Networks

Convolution:





Spatial Approaches:

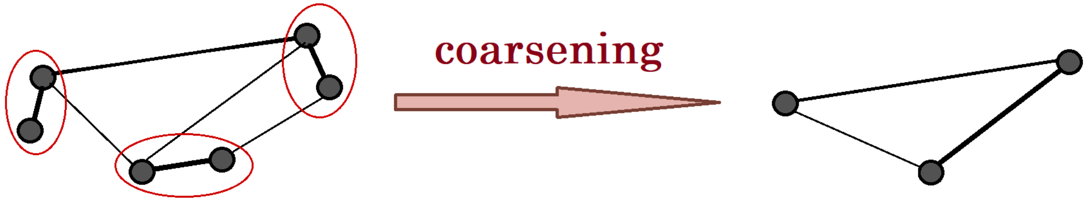
- Sliding window
- Aggregating feature information from the neighbors of each node

Spectral Approaches:

- Convolution theorem
- Defined in frequency domain
- Filter = multiplication in the frequency domain

Activation Function: …similar

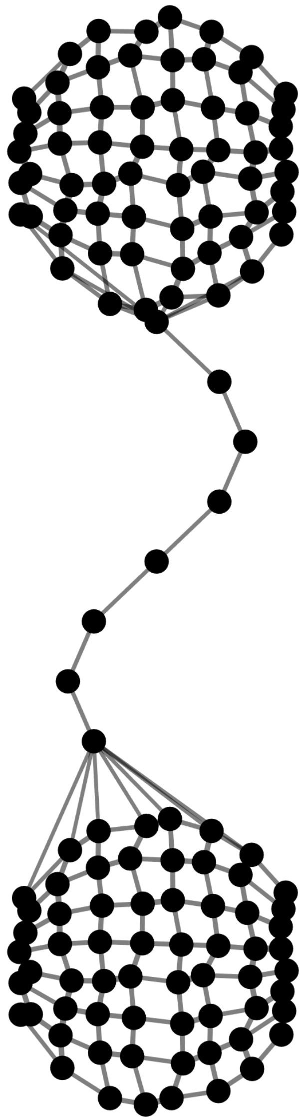
Pooling:

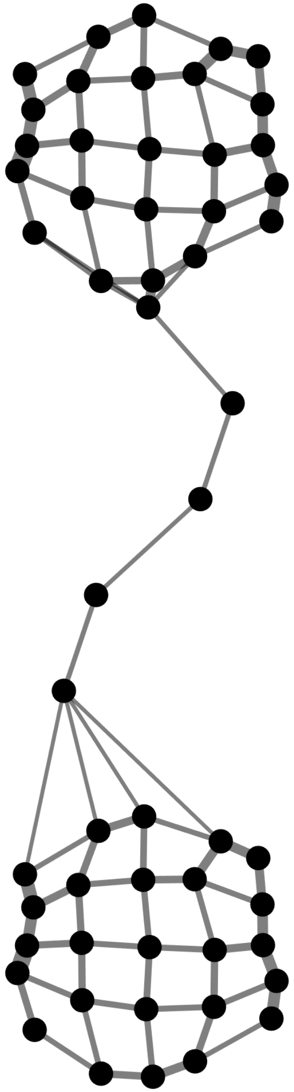


A Special Form of Generalization Capability

General Form of Generalization:

Graph neural networks should *generalize* to graphs and signals unseen in the training set.





A Special Form of Generalization Capability

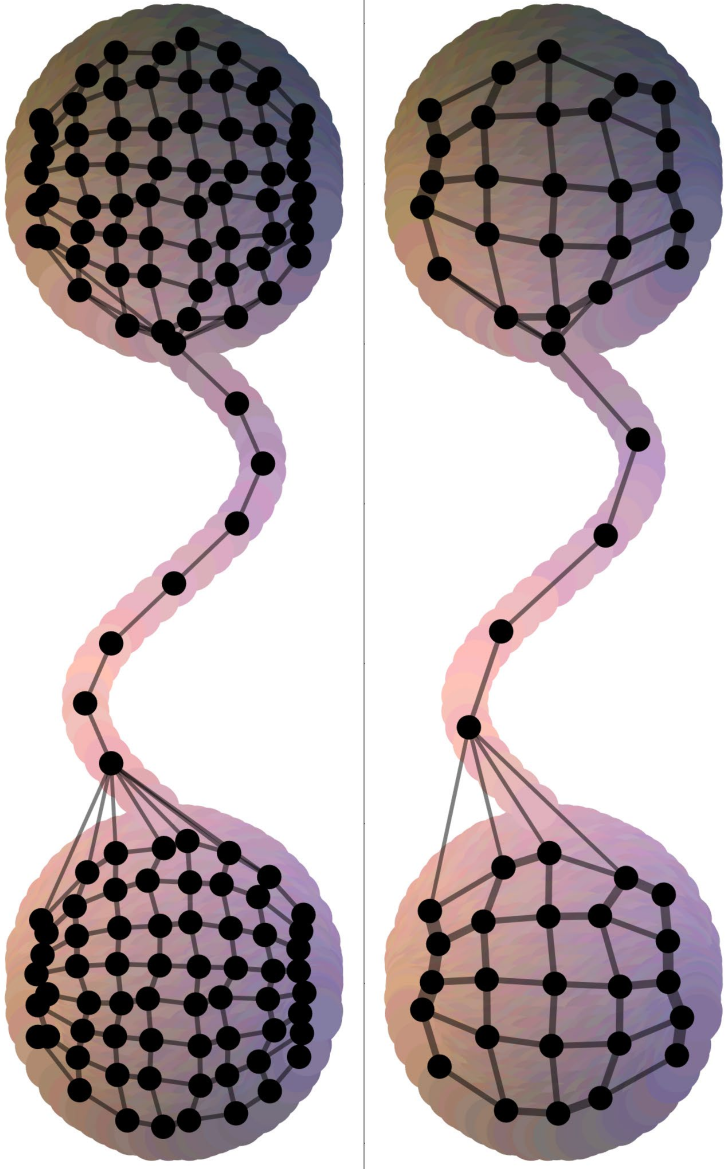
General Form of Generalization:

Graph neural networks should *generalize* to graphs and signals unseen in the training set.

The Concept of Transferability:

If two graphs *model the same phenomenon*, a trained graph neural network should have approximately the *same repercussion on both graphs*.

We will derive a complete analysis of this subproblem of generalization!



Graph Laplacian: Oscillations on Graphs

Definition: Let D be the degree matrix and W the adjacency matrix. Then the *unnormalized Graph Laplacian* is defined by

$$\Delta_u = D - W$$

and the *normalized Graph Laplacian* is given by

$$\Delta_n = D^{-1/2} \Delta_u D^{-1/2}.$$

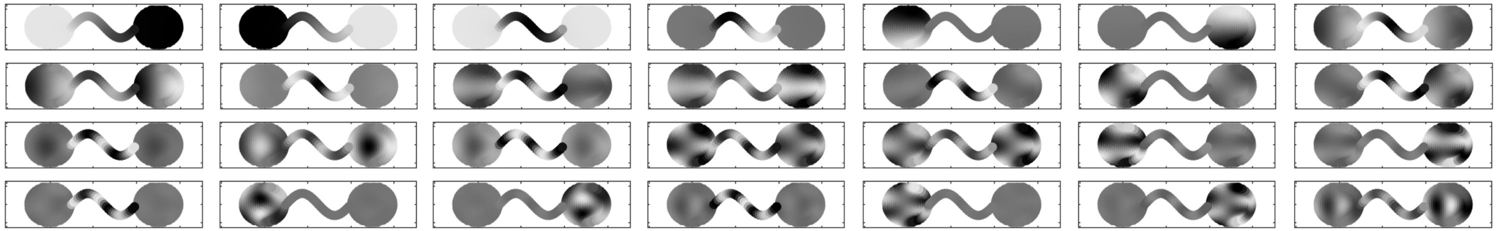
Remark: The Graph Laplacian Δ is self-adjoint. We will denote its

→ eigenvalues by $\{\lambda_j\}_j$ *(Frequencies),*

→ eigenvectors by $\{u_j\}_j$ *(Fourier modes).*

The graph Laplacian encapsulates the geometry of the graph!

Graph Laplacian Eigenvectors



Spectral Graph Convolution

Definition: Letting $\{u_j\}_j$ denote the eigenvectors of the graph Laplacian, we define the *spectral graph convolution operator* by

$$Cf = \sum_j c_j \langle f, u_j \rangle \, u_j .$$

Problem with the Implementation:

→ *Computationally demanding*

- Eigendecomposition is slow.
- No general FFT for graphs.

→ *Not transferable*

- The eigendecomposition is not stable to graph perturbations.
- A fixed filter has different repercussions on similar graphs.

Solution: Implement convolution using functional calculus!

Functional Calculus

Definition: Let T be a self-adjoint operator with discrete spectrum

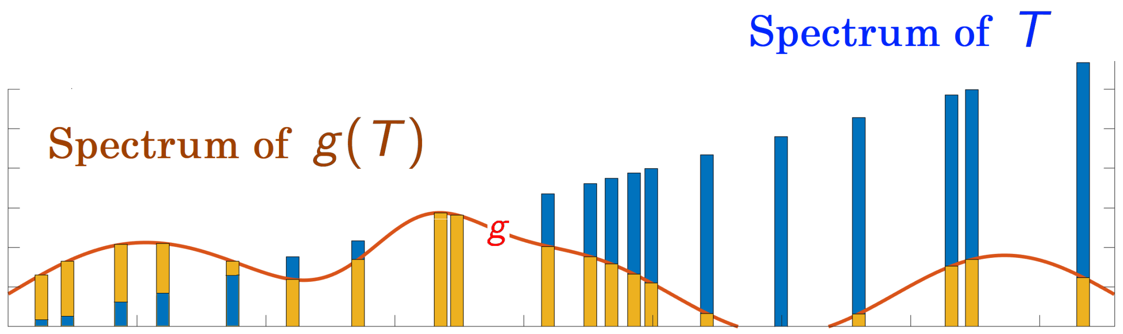
$$Tv = \sum_j \lambda_j \langle v, u_j \rangle u_j .$$

A function $g : \mathbb{R} \to \mathbb{C}$ of T is then defined via

$$g(T)v = \sum_j g(\lambda_j) \langle v, u_j \rangle u_j .$$

Remark:

If $g(\lambda) = \frac{\sum_{l=0}^{L} c_l \lambda^l}{\sum_{l=0}^{L} d_l \lambda^l}$, then $g(T) = \left(\sum_{l=0}^{L} c_l T^l\right) \left(\sum_{l=0}^{L} d_l T^l\right)^{-1}$.



Spectral Filtering using Functional Calculus

Functional Calculus Filters:

The functional calculus for $g : \mathbb{R} \to \mathbb{C}$ applied to the graph Laplacian yields

$$g(\Delta)f = \sum_j g(\lambda_j) \langle f, u_j \rangle \, u_j .$$

Recall:

The previous implementation used

$$Cf = \sum_j c_j \langle f, u_j \rangle \, u_j .$$

Advantages of Functional Calculus Viewpoint:

This approach...

→ *...solves the instability problem* (Levie, Isufi, Kutyniok; 2019).

→ *...solves the computational problem*, if g is a rational function.

Three Approaches to Transferability

Stability under Perturbation (Levie, Isufi, K; 2019), (Kenlay, Thanou, Dong; 2021):

→ Two graphs which are *small perturbations* of each other.

Topological Space Sampling (Keriven, Bietti, Vaiter; 2020), (Levie, Huang, Bucci, Bronstein, K; 2020):

→ Two graphs which sample the *same underlying continuous space*.

Graphon Approach (Ruiz, Chamon, Ribeiro; 2020):

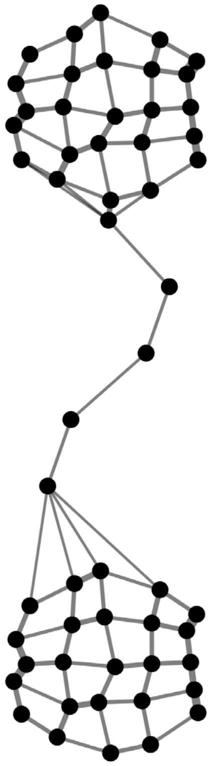
→ Two graphs that come from the *same sequence that converges to a graphon* in a homomorphism density sense.

Graphs Modeling the Same Phenomenon

Interpretation:

→ *Weighted graphs:*

- Points and strength of correspondence between pairs of points.



Graphs Modeling the Same Phenomenon

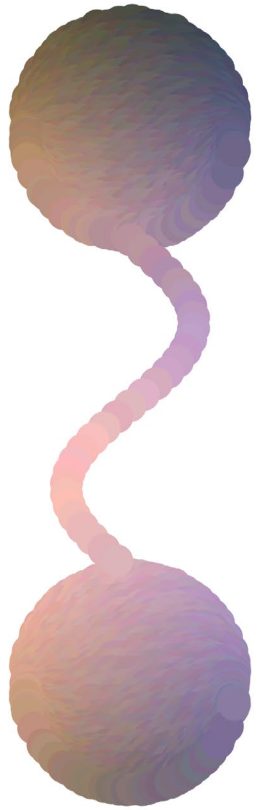
Interpretation:

→ *Weighted graphs:*

- Points and strength of correspondence between pairs of points.

→ *Metric spaces:*

- Points and distances.



Graphs Modeling the Same Phenomenon

Interpretation:

→ *Weighted graphs:*

- Points and strength of correspondence between pairs of points.

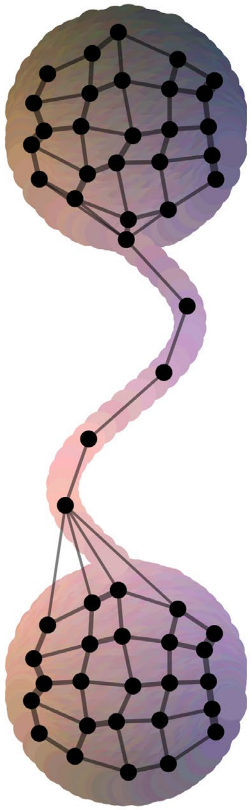
→ *Metric spaces:*

- Points and distances.

Our Viewpoint:

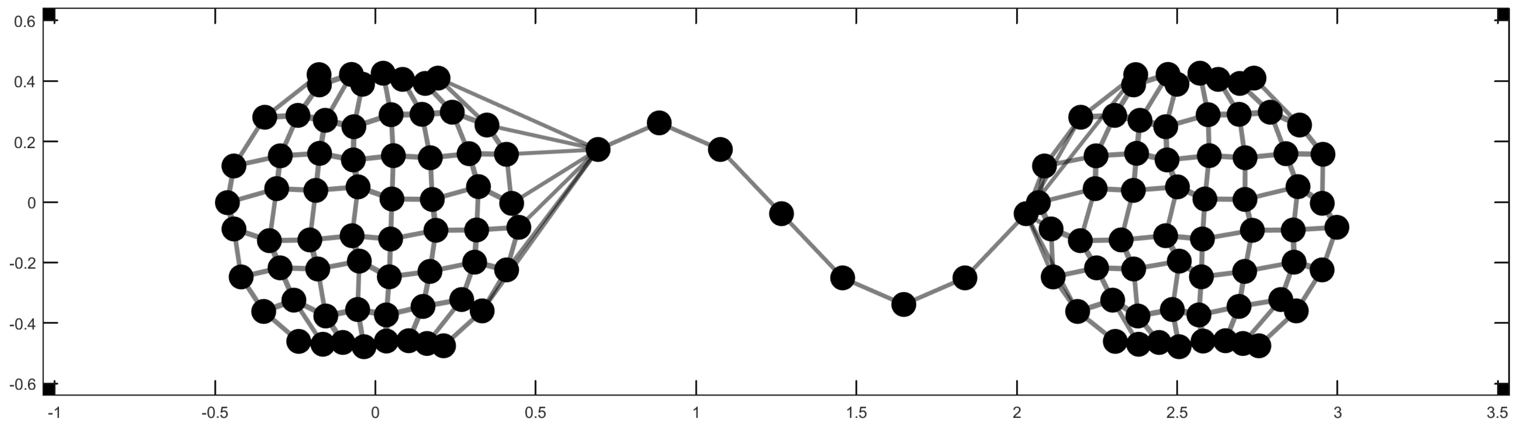
Think of graphs as discretizations of metric spaces:

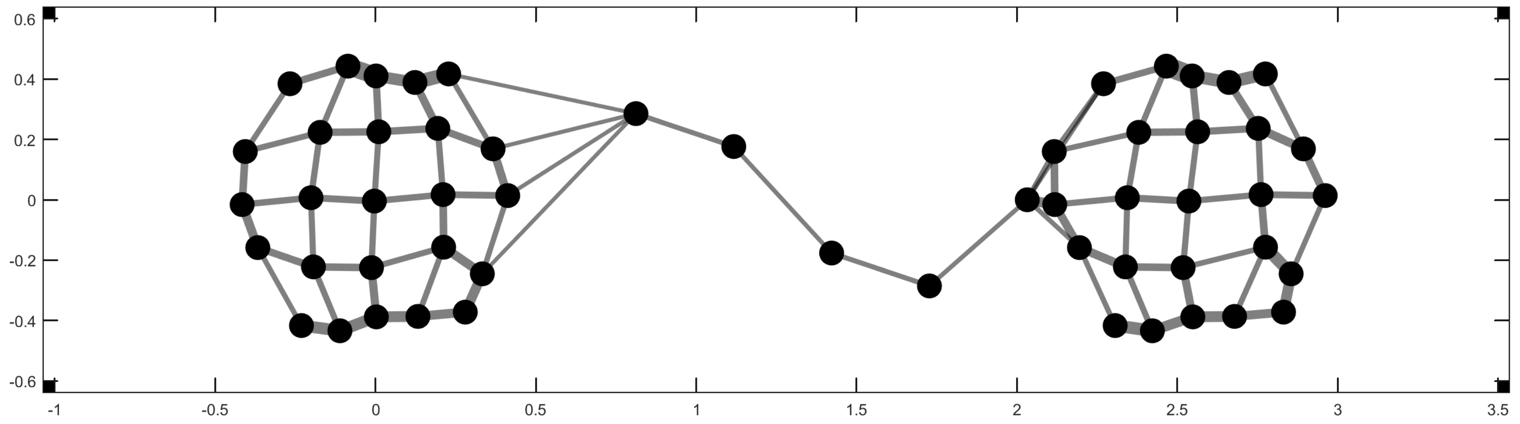
$$\text{Distance} \nearrow \quad \Longleftrightarrow \quad \text{edge weight} \searrow$$



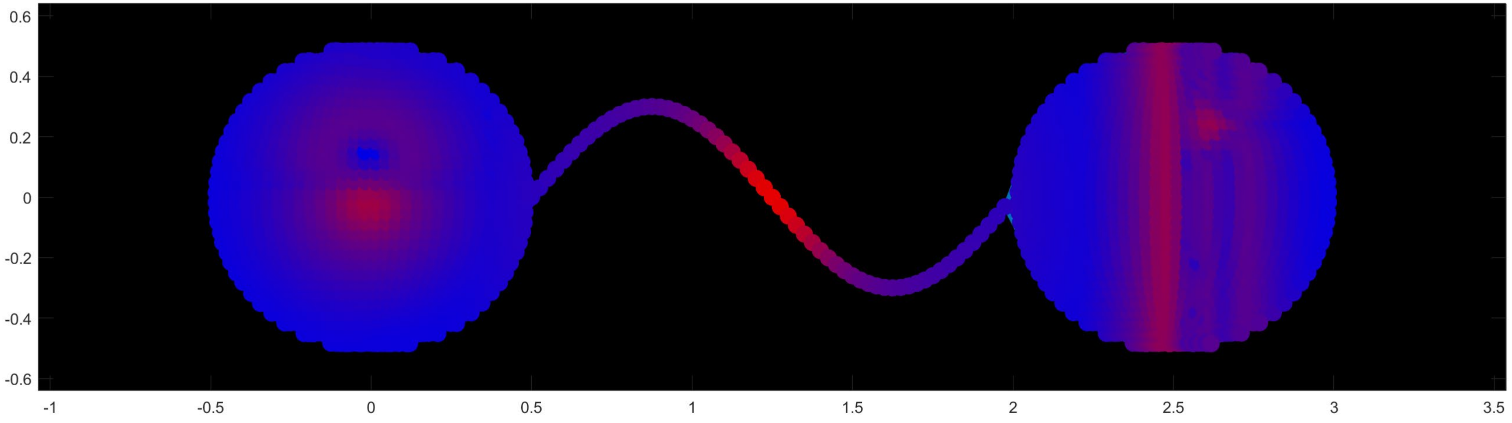
Graphs that represent the same phenomenon are discretizations of the same metric space!

Comparing the Repercussion of a Filter on Two Graphs

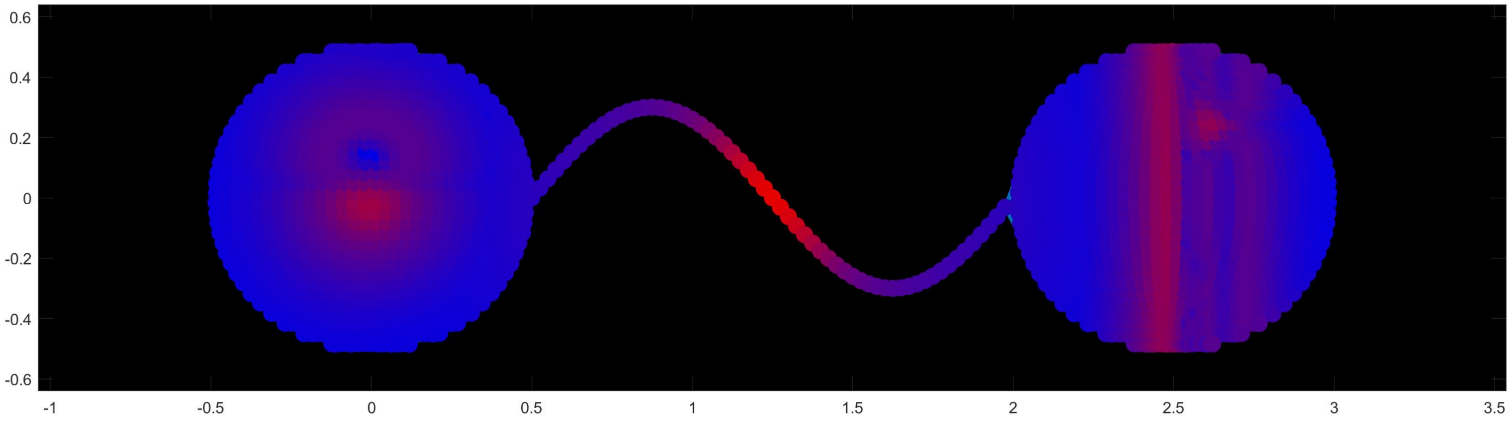




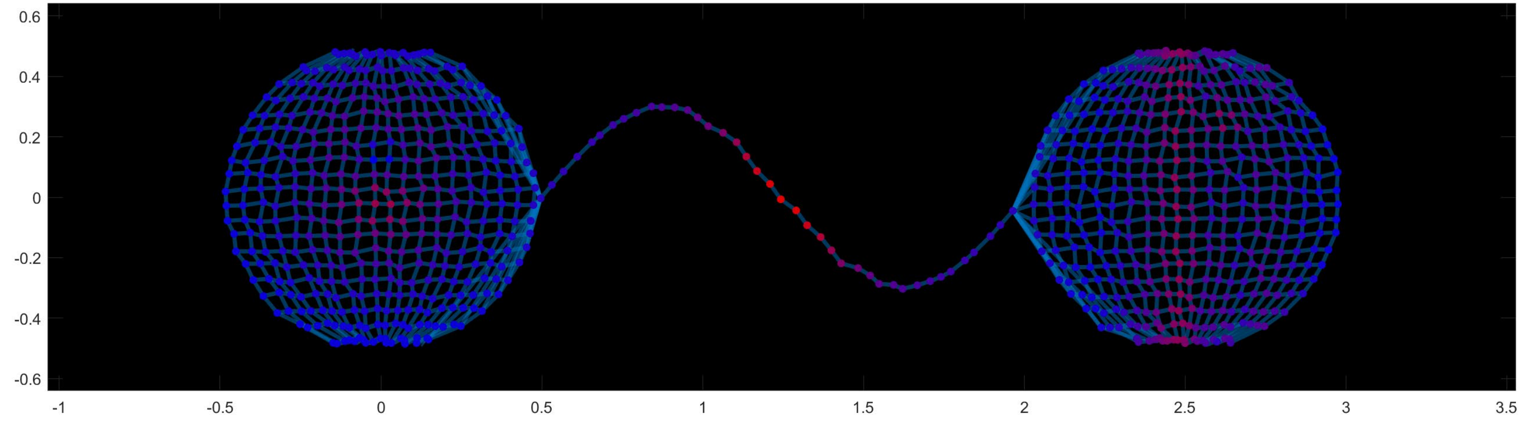
Comparing the Repercussion of a Filter on Two Graphs



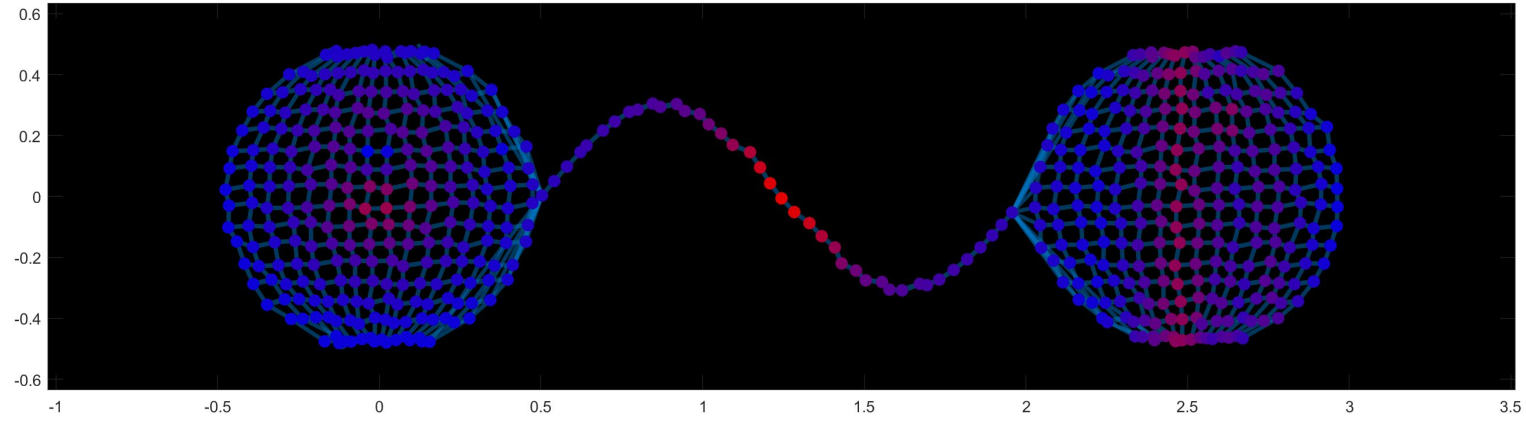
Take a generic signal $f : \mathcal{M} \to \mathbb{C}$



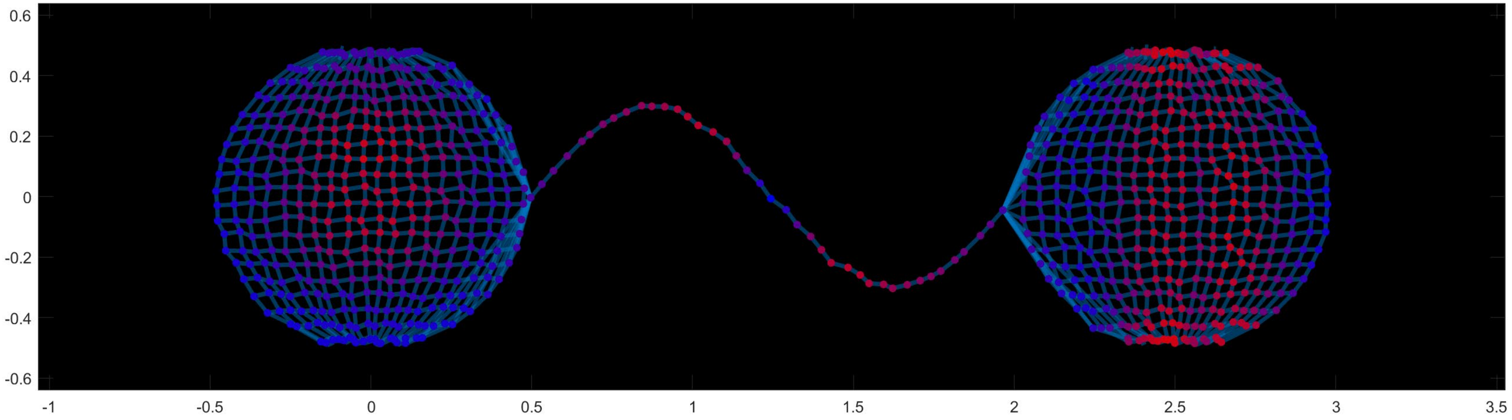
Comparing the Repercussion of a Filter on Two Graphs



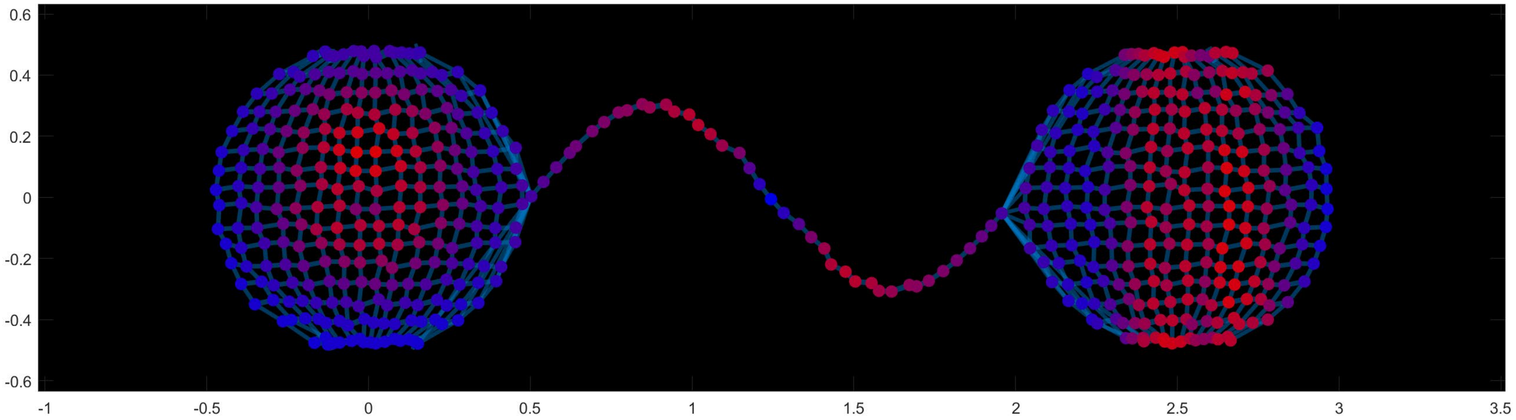
Sample to both graphs $\quad S_1 f : G_1 \to \mathbb{C}, \quad S_2 f : G_2 \to \mathbb{C}$



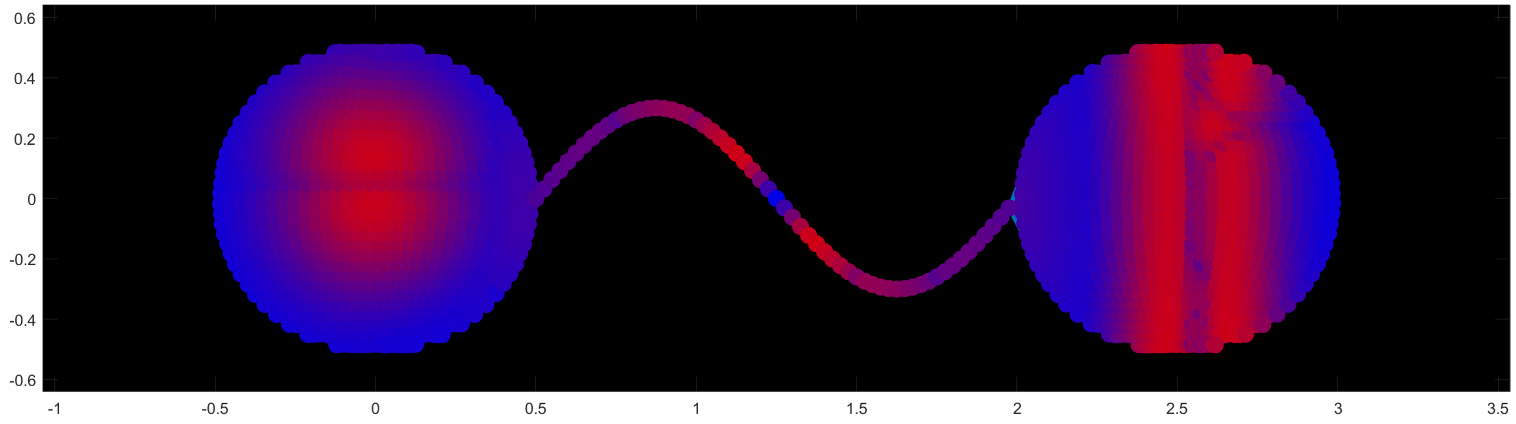
Comparing the Repercussion of a Filter on Two Graphs



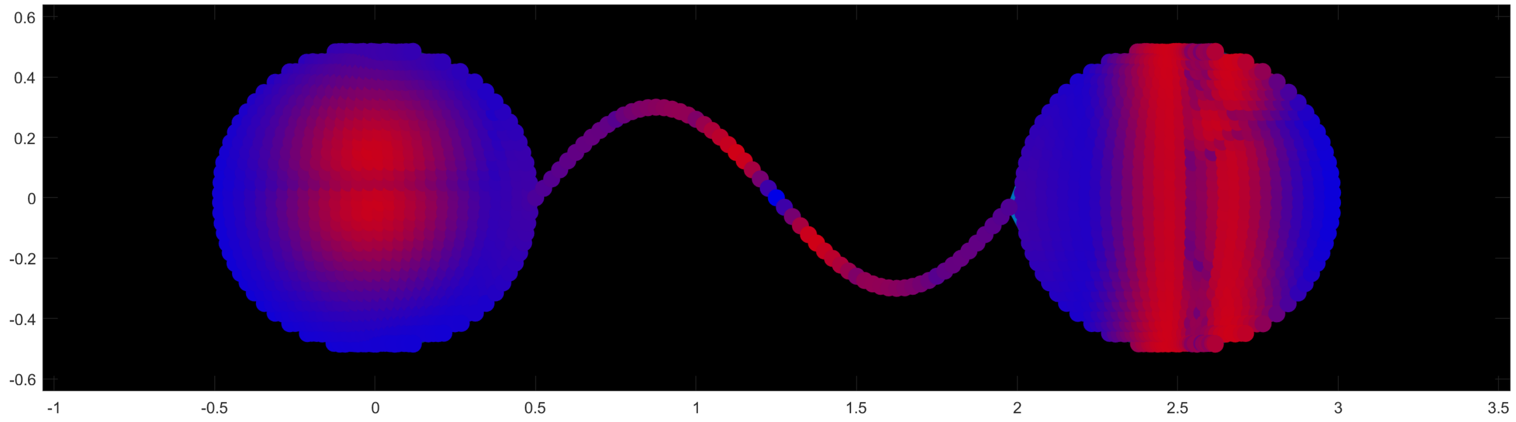
Apply both graph filters $g(\mathbf{\Delta}_1)S_1 f, \quad g(\mathbf{\Delta}_2)S_2 f$



Comparing the Repercussion of a Filter on Two Graphs



Interpolate back to $L^2(\mathcal{M})$ to get $\|R_1 g(\boldsymbol{\Delta}_1) S_1 f - R_2 g(\boldsymbol{\Delta}_2) S_2 f\| \approx 0$



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Main Result

Theorem (Levie, Huang, Bucci, Bronstein, Kutyniok; 2021):

“Transferability of graph (convolutional) neural network

$\leq$ Transferability of graph Laplacian + Consistency error"

Theorem (Levie, Huang, Bucci, Bronstein, K; 2021):
Consider two graphs G_j, $j = 1, 2$ and two graph Laplacians Δ_j, $j = 1, 2$, approximating the same Laplacian $\mathcal{L}$ in $\mathcal{M}$, and consider a ReLU graph CNN with Lipschitz filters. Further, let $G_{j,l}$ be the graph in layer l with graph Laplacians $\Delta_{j,l}$. Also, assume that, for all layers l, bands λ_l, and $j = 1, 2$,

$$\|S_{j,l}^{\lambda_l}\mathcal{L}P(\lambda_l) - \Delta_{j,l}S_{j,l}^{\lambda_l}P(\lambda_l)\| \leq \delta$$

and

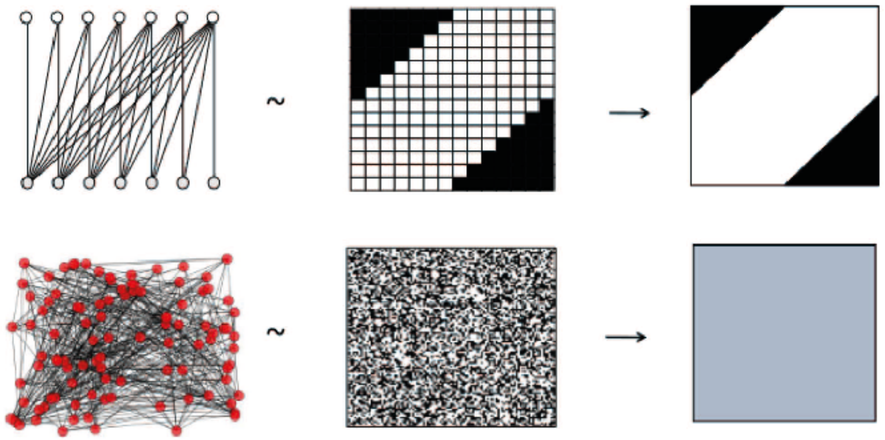
$$\|P(\lambda_L) - R_{j,L}^{\lambda_L}S_{j,L}^{\lambda_L}P(\lambda_L)\| \leq \delta$$

for some $0 < \delta < 1$. Then, for all output-channels k and mappings $\Phi_{j,L}^k$ given by the graph CNN,

$$\|R_{1,L}^{\lambda_L}\Phi_{1,L}^k S_{1,1}^{\lambda_0}P(\lambda_0) - R_{2,L}^{\lambda_L}\Phi_{2,L}^k S_{2,1}^{\lambda_0}P(\lambda_0)\|$$
$$\leq 2\left(LD\sqrt{\dim(PW(\lambda))} + L + 1\right)\delta.$$

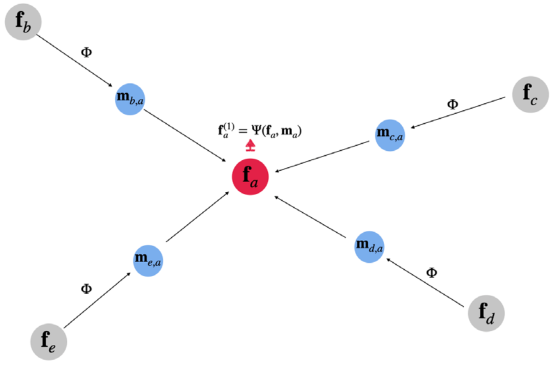
Further Results on Generalization Ability of GNNs

Graph Convolutional Neural Networks:



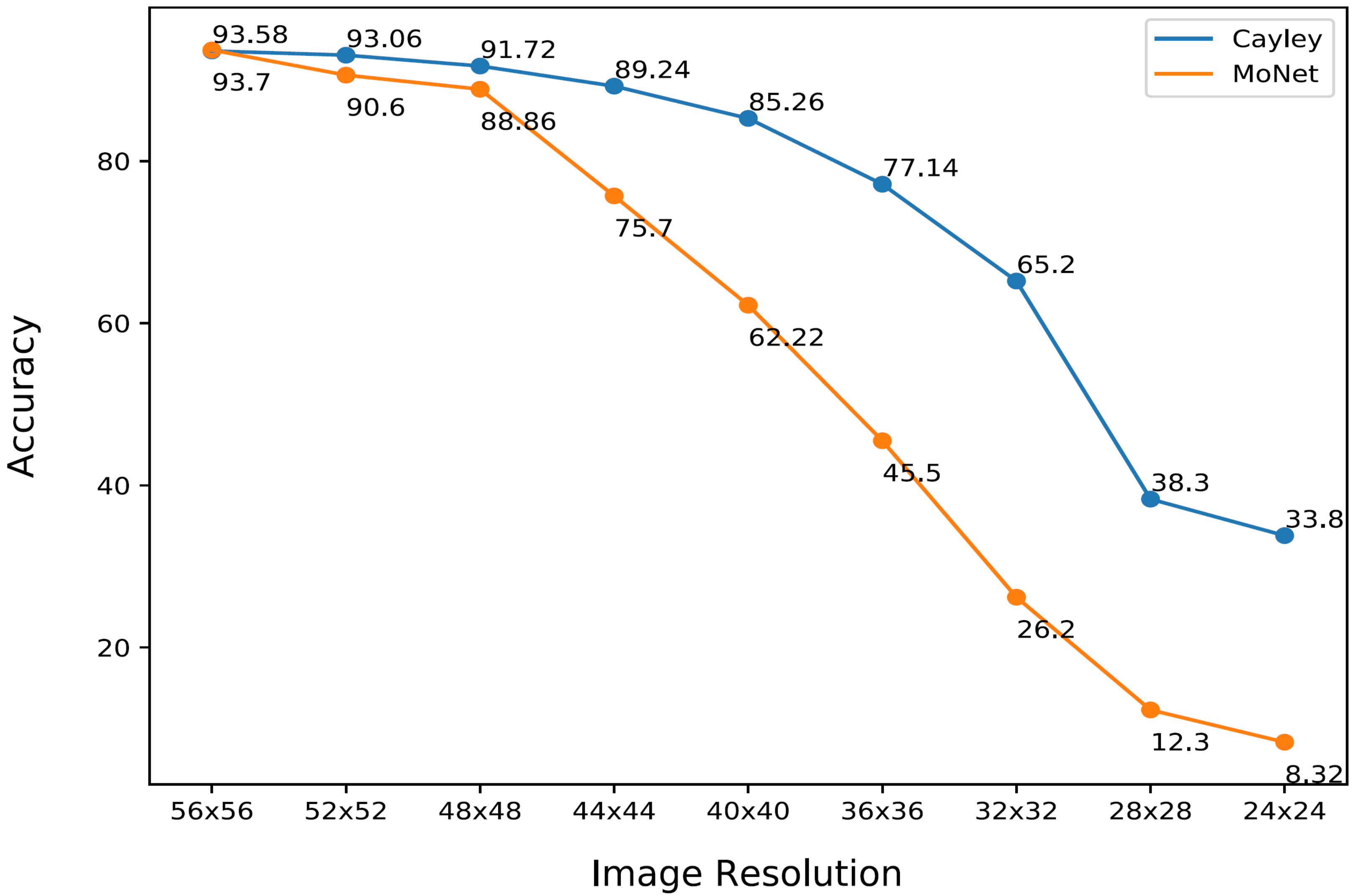
- *Similar results on transferability* for the *graphon* setting (Maskey, Levie, Kutyniok; 2022).
- This builds on (Ruiz, Wang, Ribeiro; 2021).

Message Passing Graph Neural Networks:

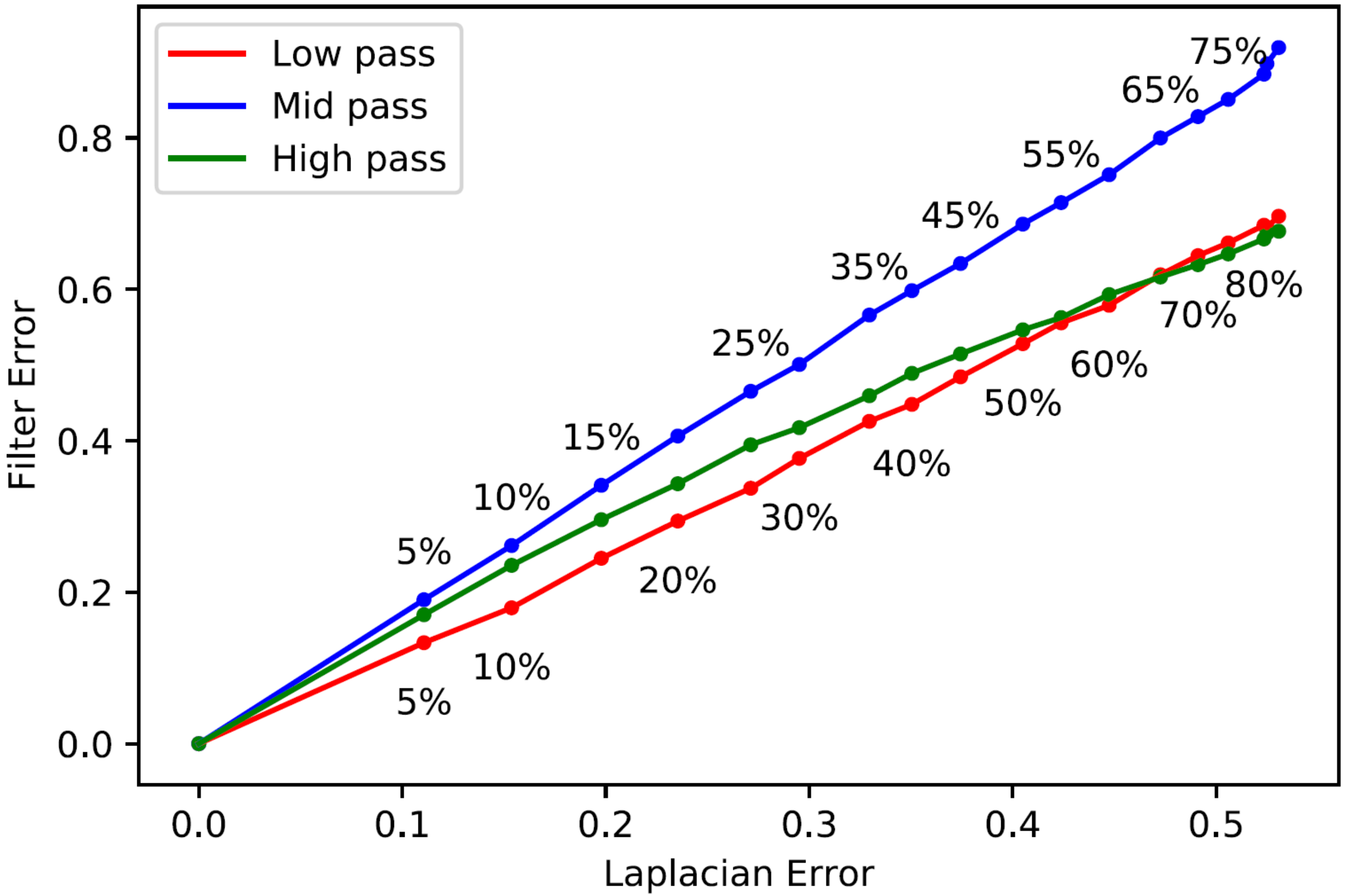


- *Non-asymptotic generalization bounds*, only depending on the regularity of the network and space (Maskey, Levie, Lee, Kutyniok; 2023).
- This builds on (Garg, Jegelka, Jaakkola; 2020), (Verma, Zhang; 2019), (Yehudai, Fetaya, Meirom, Chechik, Maron; 2022).

Spectral versus Spatial Methods



Transferability under Graph Perturbation (Randomly Removing Edges)



A Mathematical Understanding of Deep Learning

MATH
4AI

Expressivity:

→ Which *aspects of a neural network architecture* affect the performance of deep learning?

Applied Harmonic Analysis, Approximation Theory, ...

Learning:

→ Why does *stochastic gradient descent* converge to good local minima despite the non-convexity of the problem?

Algebraic/Differential Geometry, Optimal Control, Optimization, ...

Generalization:

→ Can we derive overall *success guarantees* (on the test data set)?

Learning Theory, Probability Theory, Statistics, ...

Explainability:

→ Why did a trained deep neural network *reach a certain decision*?

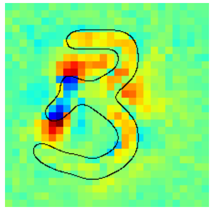
Information Theory, Uncertainty Quantification, ...

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Explainability:
A Mathematical Approach

Some General Thoughts about Explainability

Main Goal: We aim to *understand* decisions of ``black-box'' predictors!



Selected Questions:

→ What *exactly* is relevance in a mathematical sense?

→ Can we develop a theory for *optimal relevance maps*?

→ How to extend to *challenging modalities*?

→ Can we derive *higher level explanations*?

Vision:

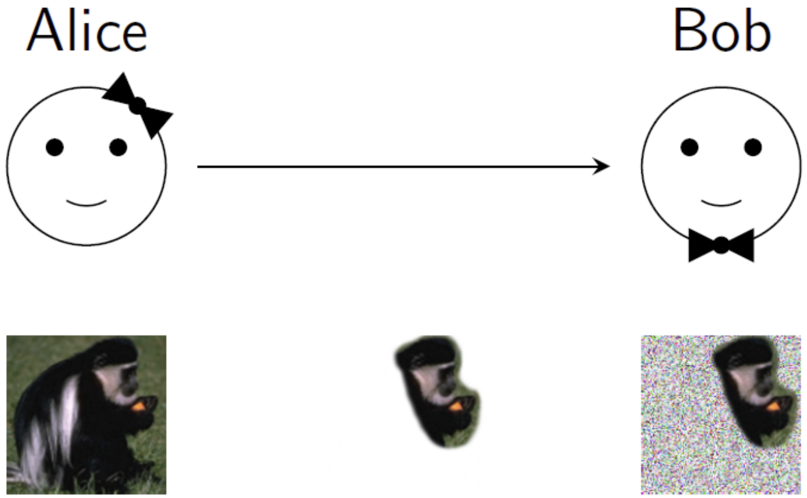
Questioning the AI as a human about the reason for a decision!

The explainability approach itself needs to be reliable!

Information Theory: Rate-Distortion Viewpoint

The Setting: Let

- $\Phi\colon [0,1]^d \to [0,1]$ be a *classification function*,
- $x \in [0,1]^d$ be an *input signal*.



$\Phi(x) = 0.97$ "Monkey"

Original image x Partial image S Random completion y

$\Phi(y) = 0.91$ "Monkey"

Expected Distortion:

$$D(S) = D(\Phi, x, S) = \mathbb{E}\left[\frac{1}{2}\left(\Phi(x) - \Phi(y)\right)^2\right]$$

Rate-Distortion Explanation (RDE)

MATH ^{4}AI

Rate-Distortion Function:

$$R(\epsilon) = \min_{S \subseteq \{1,\ldots,d\}} \{|S| \,:\, D(S) \leq \epsilon\}$$

Use this viewpoint for the definition of a relevance map!

Theorem (Wäldchen, Macdonald, Hauch, Kutyniok, 2020):

Finding a minimizer of $R(\epsilon)$ is very hard!

Computable Variant of RDE (Macdonald, Wäldchen, Hauch, Kutyniok, 2020):

$$\text{minimize} \quad D(s) + \lambda \|s\|_1 \quad \text{subject to} \quad s \in [0,1]^d$$

…allows rigorous mathematical performance analysis!

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

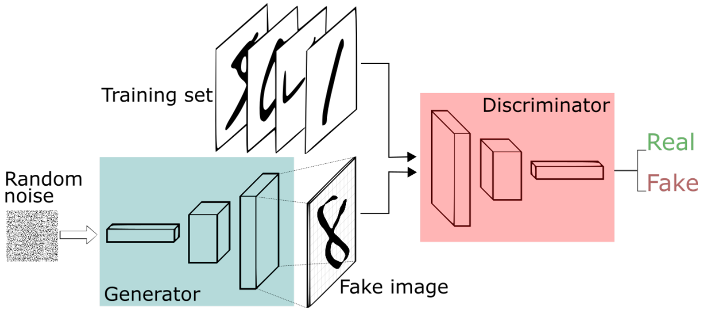
Going Beyond….

Extending to More Realistic Scenarios?

Extension 1 (Heiß, Levie, Resnick, Kutyniok, Bruna; 2020):

→ Choose the obfuscations more natural

→ Example: Apply an inpainting GAN



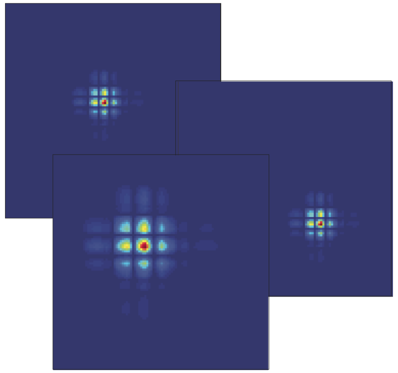
Obtaining Higher-Level Explanations?

Extension 2 (Kolek, Nguyen, Levie, Bruna, Kutyniok; 2021):

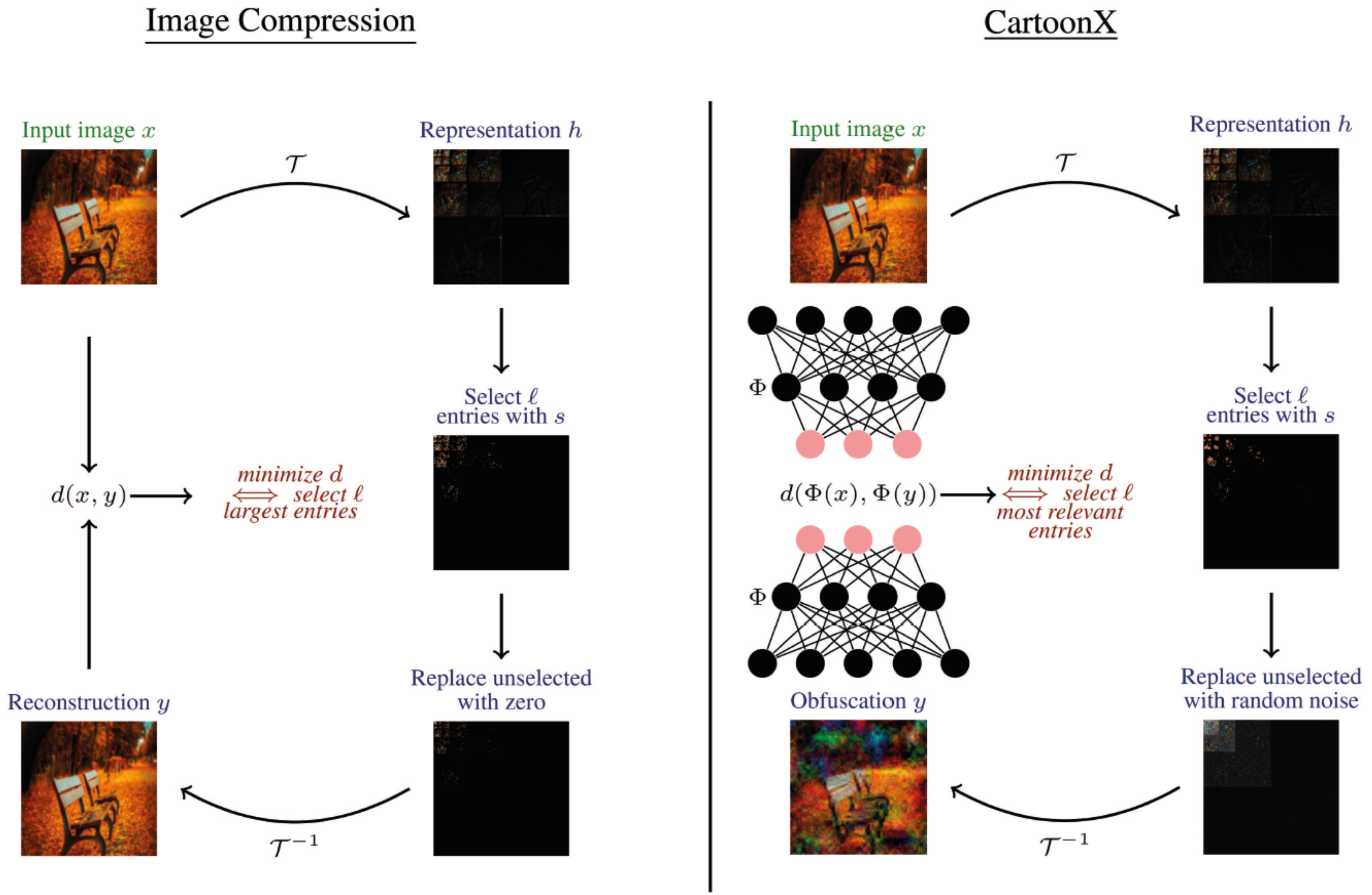
→ Apply RDE to decompositions of the data

→ Example: Take a wavelet decomposition of an image.

→ *CartoonX*

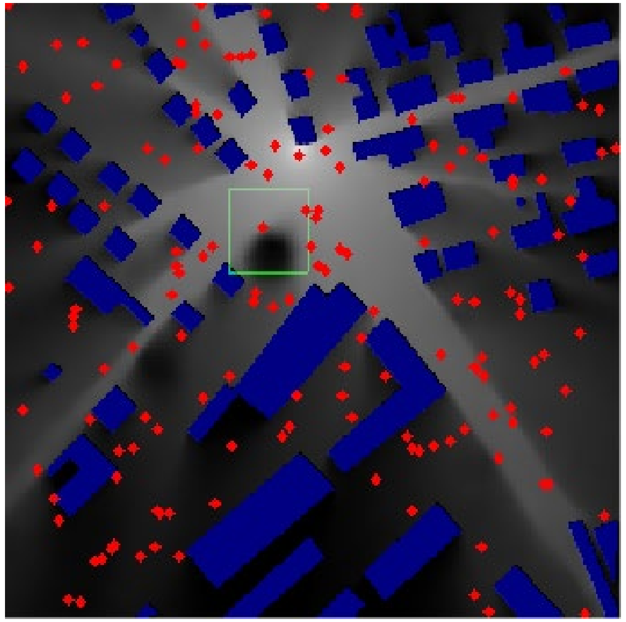


Idea of CartoonX (Kolek, Nguyen, Levie, Bruna, Kutyniok; 2022)



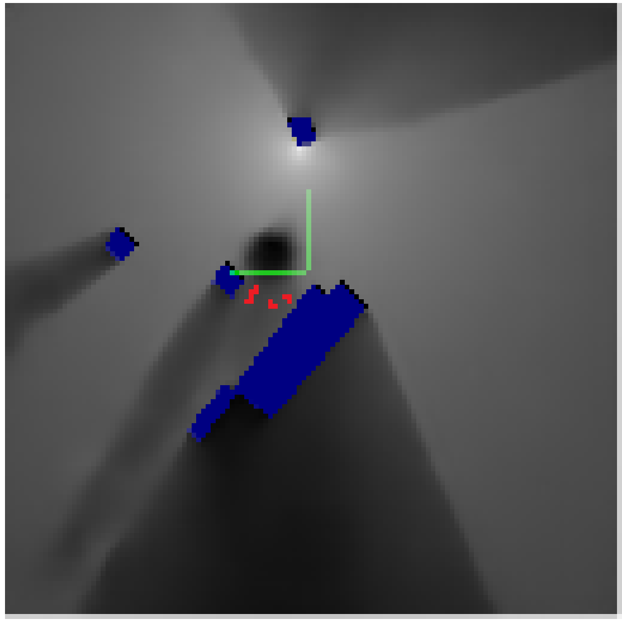
Explainability: Understanding Seemingly Wrong Decisions

Example from Telecommunication:



Estimated RadioMap via RadioUNet

(Levie, Cagkan, Kutyniok, Caire; 2020)



Rate-Distortion Explanation

(Heiß, Levie, Resnick, Kutyniok, Bruna; 2020):

Explainability: Understanding Wrong Decisions

Example from Imaging:

Wrong decision by AI:

Diaper

Wrong decision by AI:

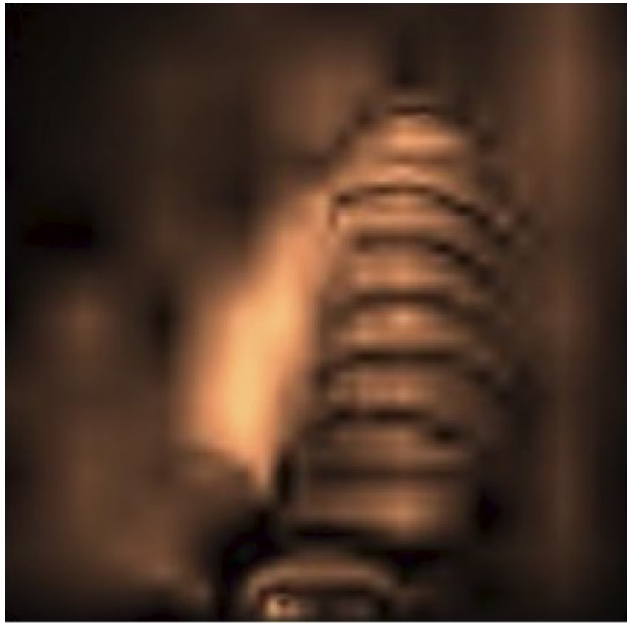
Screw

Explainability: Understanding Wrong Decisions

Example from Imaging:

Explanation by CartoonX

(Kolek, Nguyen, Levie, Bruna, Kutyniok; 2021)



Explanation by CartoonX

Extension: ShearletX (Kolek, Windesheim, Loarca, Kutyniok, Levie; 2023)!

A Mathematical Understanding of Deep Learning

Expressivity:

→ Which *aspects of a neural network architecture* affect the performance of deep learning?

Applied Harmonic Analysis, Approximation Theory, ...

Learning:

→ Why does *stochastic gradient descent* converge to good local minima despite the non-convexity of the problem?

Algebraic/Differential Geometry, Optimal Control, Optimization, ...

Generalization:

→ Can we derive overall *success guarantees* (on the test data set)?

Learning Theory, Probability Theory, Statistics, ...

Explainability:

→ Why did a trained deep neural network *reach a certain decision*?

Information Theory, Uncertainty Quantification, ...

Are there fundamental limitations?

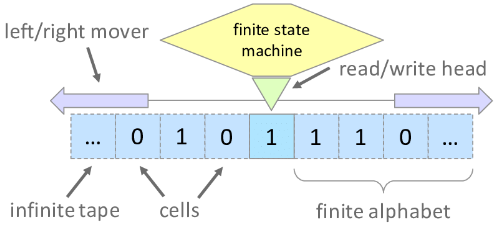
A Word of Caution:
Problems with Computability

Are There Limitations to Be Aware Of?

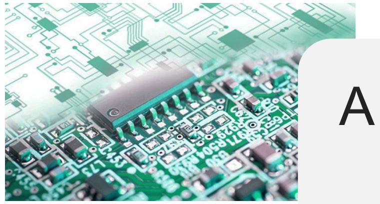
Artificial Intelligence is not a Swiss Army Knife!

More Fundamental Viewpoint:

What can actually be *computed on digital hardware*?



Turing-Machine



A *computable problem (function)* is one for which the input-output relation can be computed on a digital machine for any given accuracy.

What about Non-Computability?

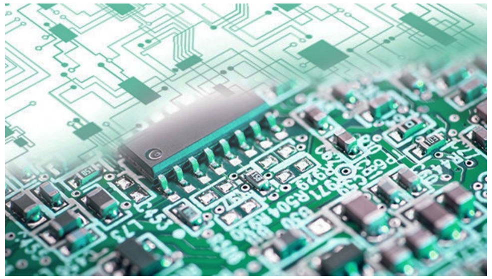
Non-computable problems can be tackled successfully in practice, if limited precision succeeds!

But we have no guarantees of correctness!

MATH 4AI

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Very Disappointing News



Theorem (Boche, Fono, Kutyniok; 2022):
The solution of a finite-dimensional inverse problem is *not (Turing-)computable* (by a deep neural network).

Solution Set:
For $A \in \mathbb{C}^{m \times N}$ and $y \in \mathbb{C}^m$ let

$$\Psi(A, y) := \arg\min_{x \in \mathbb{C}^N} \|x\|_{\ell^1} \text{ such that } \|Ax - y\|_{\ell^2} \leq \varepsilon.$$

Theorem (Boche, Fono, K; 2023):
Fix parameters $\varepsilon \in (0, \frac{1}{4})$, $N \geq 2$, and $m < N$. There does not exist a (Banach–Mazur-)computable function $\hat{\Psi} : \mathbb{C}^{m \times N} \times \mathbb{C}^m \to \mathbb{C}^N$ such that

$$\sup_{(A,y) \in \mathbb{C}^{m \times N} \times \mathbb{C}^m} \|\Psi(A, y) - \hat{\Psi}(A, y)\|_{\ell^2} < \frac{1}{4}.$$

What now?

Theory tells us…

Theorem (Boche, Fono, Kutyniok; 2023):
The solution of a finite-dimensional inverse problem is *computable* (by a deep neural network) on an *analog (Blum-Shub-Smale) machine*!

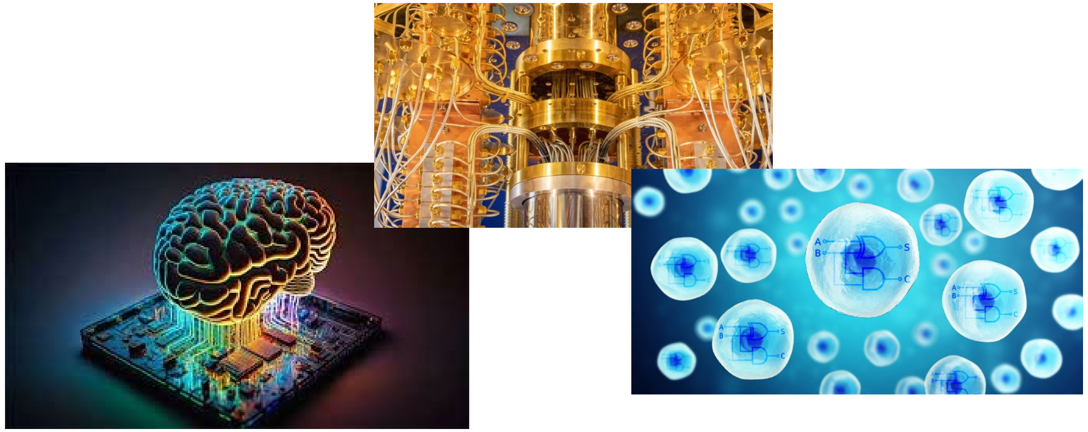
Reliability for certain problem settings requires novel hardware!

Possible Future Developments:

→ *Neuromorphic computing*

→ *Biocomputing*

→ *Quantum computing*



More Problems with Digital Hardware

Theorem (Boche, Fono, Kutyniok; 2023):
Many classification problems are also *not (Turing) computable*!

Theorem (Boche, Fono, Kutyniok; 2023):
The Pseudo Inverse is *not (Banach-Mazur) computable*!

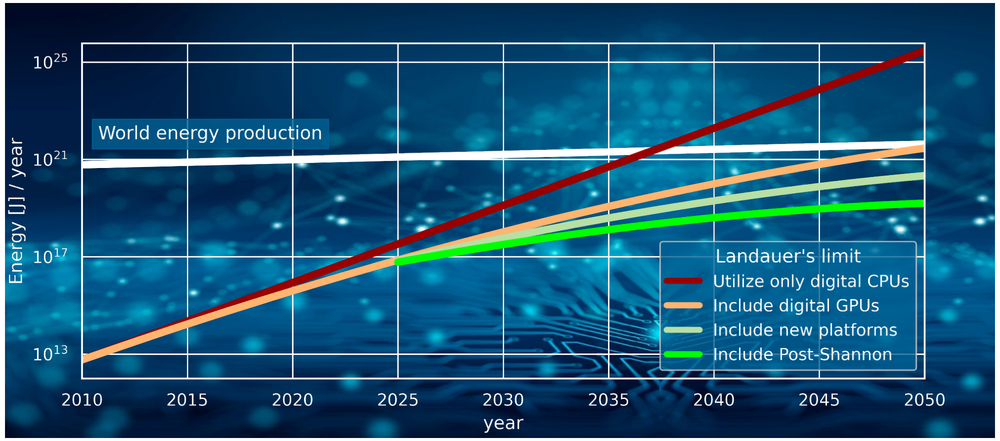
Theorem (Bacho, Boche, Kutyniok; 2023):
Computing the solutions to the Laplace and the diffusion equation on digital hardware causes a *complexity blowup*.

Theorem (Lee, Boche, Kutyniok; 2023):
Finding the solution of most optimization problems is *not (Turing-)computable*; it can *not even be approximated* by a Turing computable function!

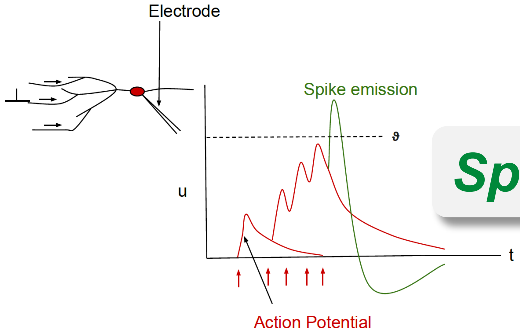
Future Perspective

Vision for the Future:

1. Provable *Computability*
2. Provable *Stability* and *Performance Guarantees*
3. Fulfillment of *Legal Requirements*
   - → Algorithmic Transparancy/Accounability
   - → Right to Explain
4. *Energy Efficiency/Sustainability*



Source:Decadal Plan of the Semiconductor Research Corporation for the Biden (US) Administration, 2021



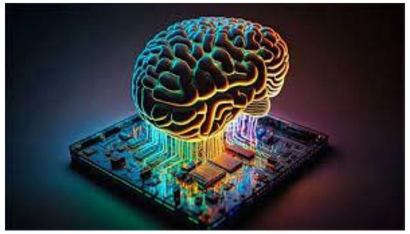
Spiking Neural Networks!

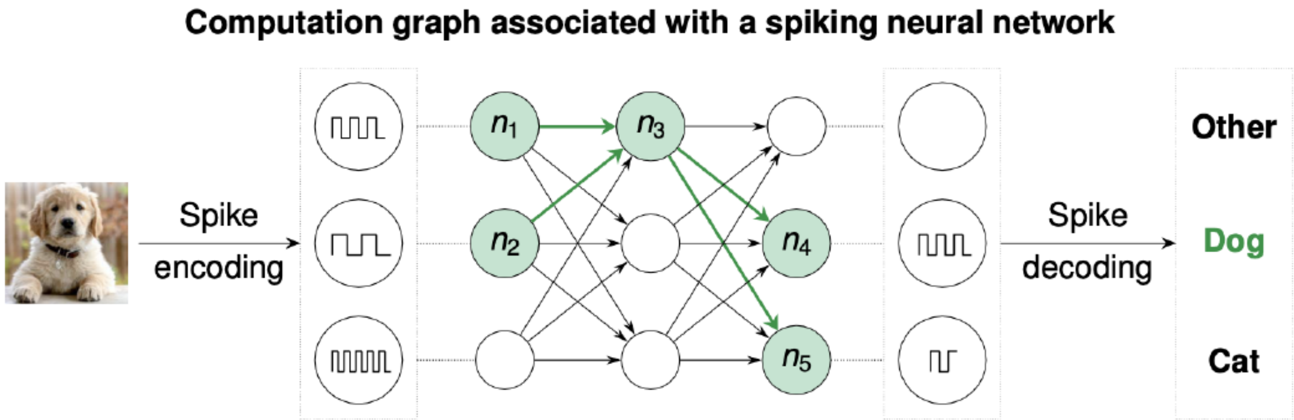
EcoLogic Computing

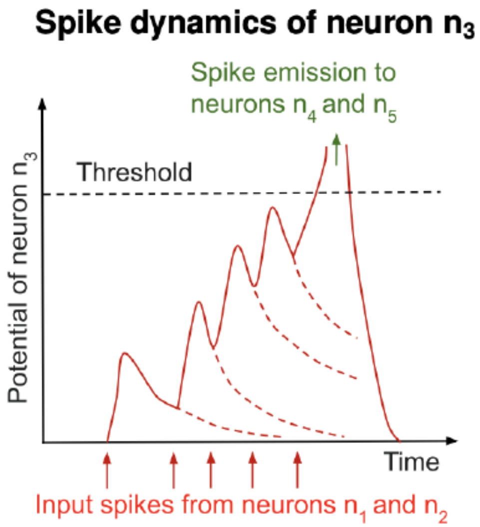
https://www.ecologic-computing.com

Truly Reliable AI … by Next Generation Computing!

A Glimpse in Spiking Neural Networks







Remarks:

→ *More biologically realistic* than first and second generation artificial neurons.

→ Information is encoded in the *timing of individual spikes.*

Meta-Theorem (Singh, Fono, Kutyniok; 2024):

"Spiking neural networks can be emulated by classical artificial ReLU-neural networks, but in certain cases, they can be shown to *perform strictly better concerning complexity*."

Future of AI Computing

MATH 4AI

Project „Next Generation AI Computing (GAIn)“

Co-PIs:

Holger Boche
(TUM)

Frank Fitzek
(TU Dresden)

Stefanie Speidel
(TU Dresden)

Funding:

Bavarian State Ministry of
Science and the Arts

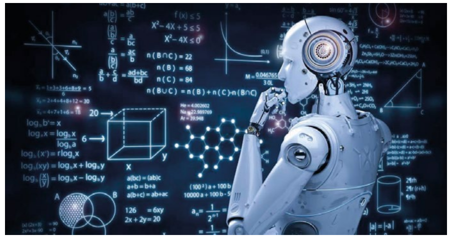
STAATSMINISTERIUM
FÜR WISSENSCHAFT
KULTUR UND TOURISMUS

Freistaat
SACHSEN

LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Conclusions

Conclusions



Artificial Intelligence:

→ *Impressive performance* in real-world applications!

→ We still have a *major problem with reliability*!

Reliability of Deep Learning from a Mathematical Perspective:

→ *Expressivity:* Optimal architectures?

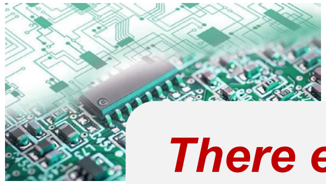
→ *Learning:* Controllable, efficient algorithms?

→ *Generalization:* Performance on test data sets?

→ *Explainability:* Explaining network decisions?

Inverse Problems:

→ *Optimal combination* of AI & Models required!



There exist serious problems for reliability of deep learning on digital hardware!

Vision for the Future:

Truly Reliable AI…by Next Generation Computing!

Konrad Zuse School of Excellence in Reliable AI

(https://zuseschoolrelai.de)

SPONSORED BY THE

Federal Ministry of Education and Research

DAAD Zuse Schools

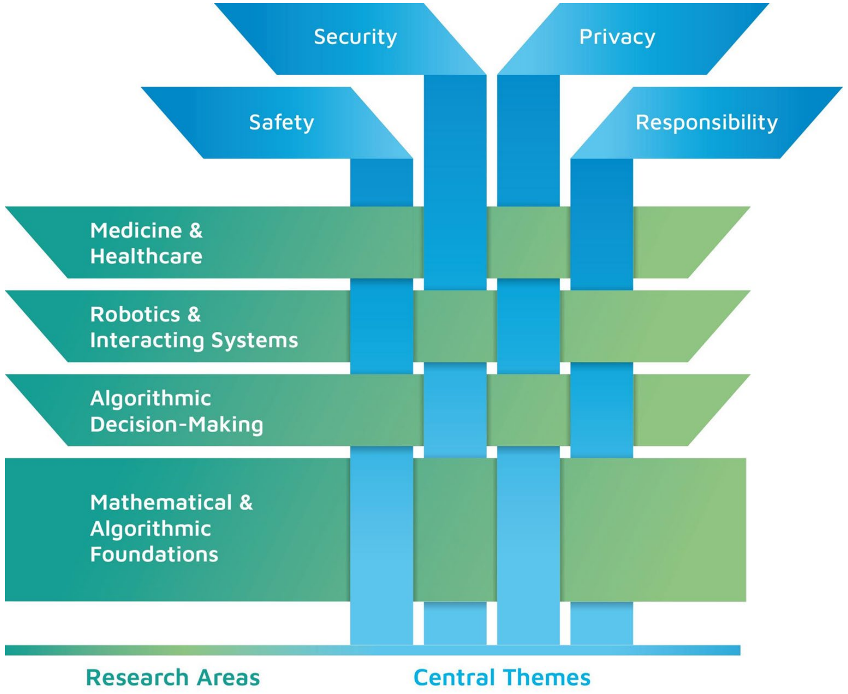
Konrad Zuse Schools of Excellence in Artificial Intelligence

MATH 4AI

Munich, Germany

relAI

Konrad Zuse School of Excellence in Reliable AI



Technische Universität München

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Mission: Train future generations of AI experts in Germany who combine technical brilliance with awareness of the importance of AI's reliability

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Thank you very much for your attention!

References available at:

www.ai.math.lmu.de/kutyniok

Survey Paper (arXiv:2105.04026):

Berner, Grohs, K, Petersen, The Modern Mathematics of Deep Learning, 2021

Related Book:

P. Grohs and G. Kutyniok, eds.,
Mathematical Aspects of Deep Learning
Cambridge University Press, 2022.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
HIGHTECH Agenda Bayern
baiosphere the bavarian ai network
MATH 4AI
DFG Deutsche Forschungsgemeinschaft
Bundesministerium für Bildung und Forschung
Bayerisches Staatsministerium für Wissenschaft und Kunst
DAAD Deutscher Akademischer Austauschdienst German Academic Exchange Service
mcml Munich Center for Machine Learning
relAI Konrad Zuse School of Excellence in Reliable AI
AI HUB@LMU
www.ai-news.lmu.de