Title: Vulnerability of quantum systems to adversarial perturbations

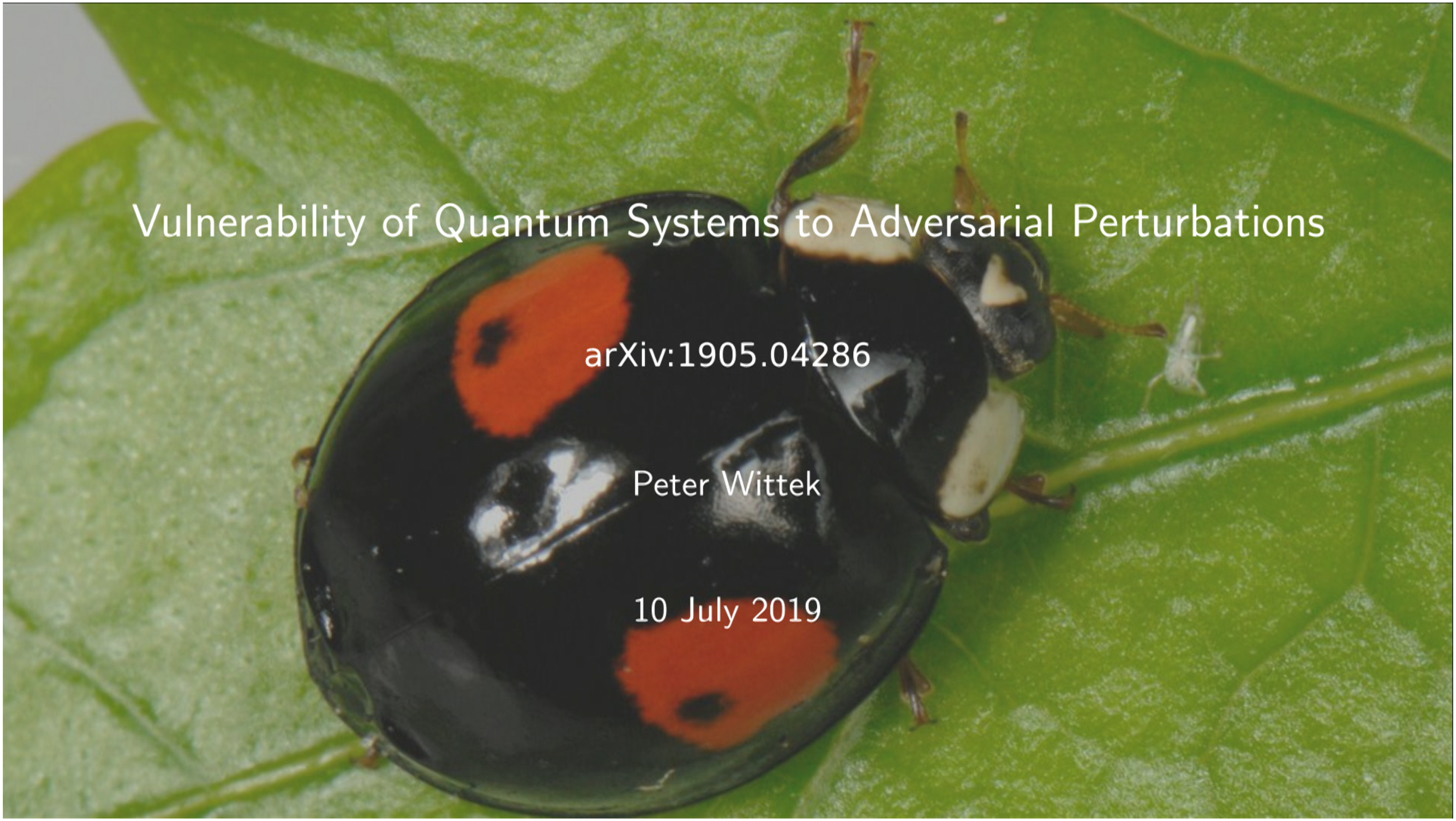
Speakers: Peter Wittek

Collection: Machine Learning for Quantum Design

Date: July 10, 2019 - 11:30 AM

URL: http://pirsa.org/19070009

Abstract: High-dimensional quantum systems are vital for quantum technologies and are essential in demonstrating practical quantum advantage in quantum computing, simulation and sensing. Since dimensionality grows exponentially with the number of qubits, the potential power of noisy intermediate-scale quantum (NISQ) devices over classical resources also stems from entangled states in high dimensions. An important family of quantum protocols that can take advantage of high-dimensional Hilbert space are classification tasks. These include quantum machine learning algorithms, witnesses in quantum information processing and certain decision problems. However, due to counter-intuitive geometrical properties emergent in high dimensions, classification problems are vulnerable to adversarial attacks. We demonstrate that the amount of perturbation needed for an adversary to induce a misclassification scales inversely with dimensionality. This is shown to be a fundamental feature independent of the details of the classification protocol. Furthermore, this leads to a trade-off between the security of the classification algorithm against adversarial attacks and quantum advantages we expect for high-dimensional problems. In fact, protection against these adversarial attacks require extra resources that scale at least polynomially with the Hilbert space dimension of the system, which can erase any significant quantum advantage that we might expect from a quantum protocol. This has wide-ranging implications in the use of both near-term and future quantum technologies for classification.



Vulnerability of Quantum Systems to Adversarial Perturbations

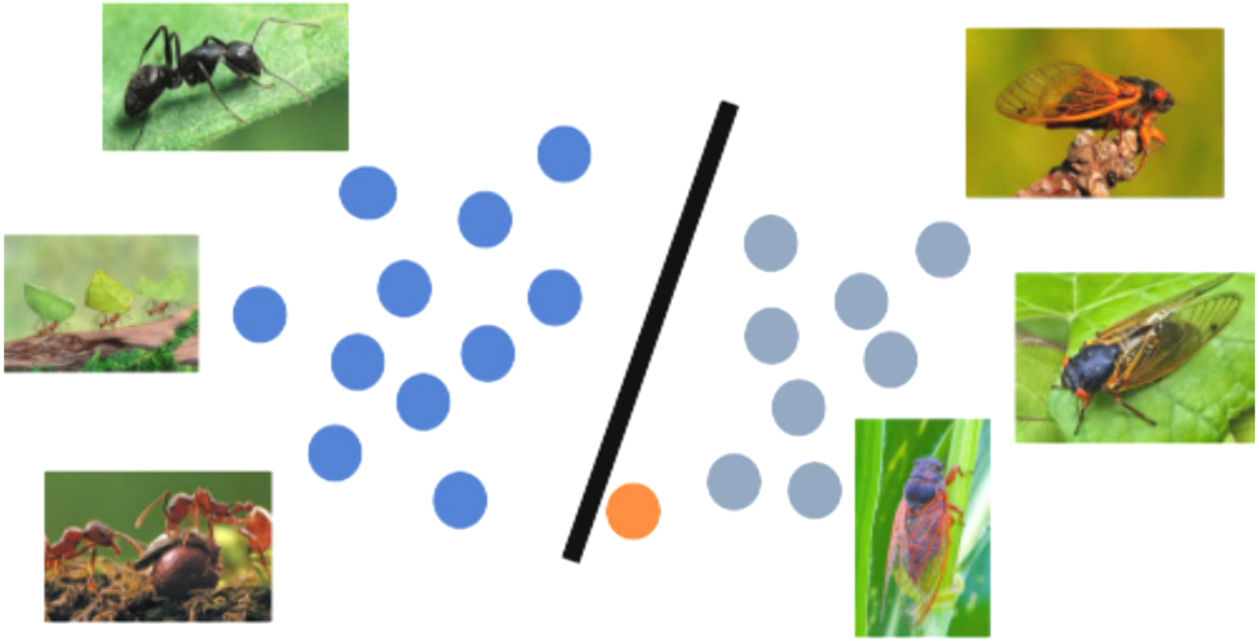
arXiv:1905.04286

Peter Wittek

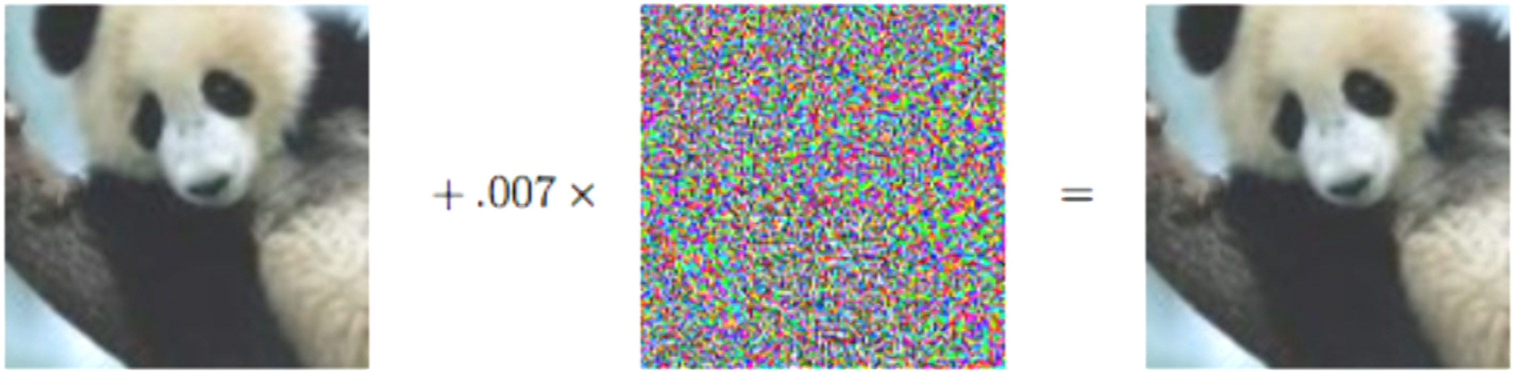
10 July 2019

Machine Learning: Differentiable Architectures and Thresholding

Learning model: $h(x) : \mathbb{R}^d \mapsto \mathbb{Z}$.
Usually a composition: $h(x) = \mathrm{threshold}(v(x))$, where $v(x) : \mathbb{R}^d \mapsto \mathbb{R}$ is differentiable.
Typical question: generalization performance.

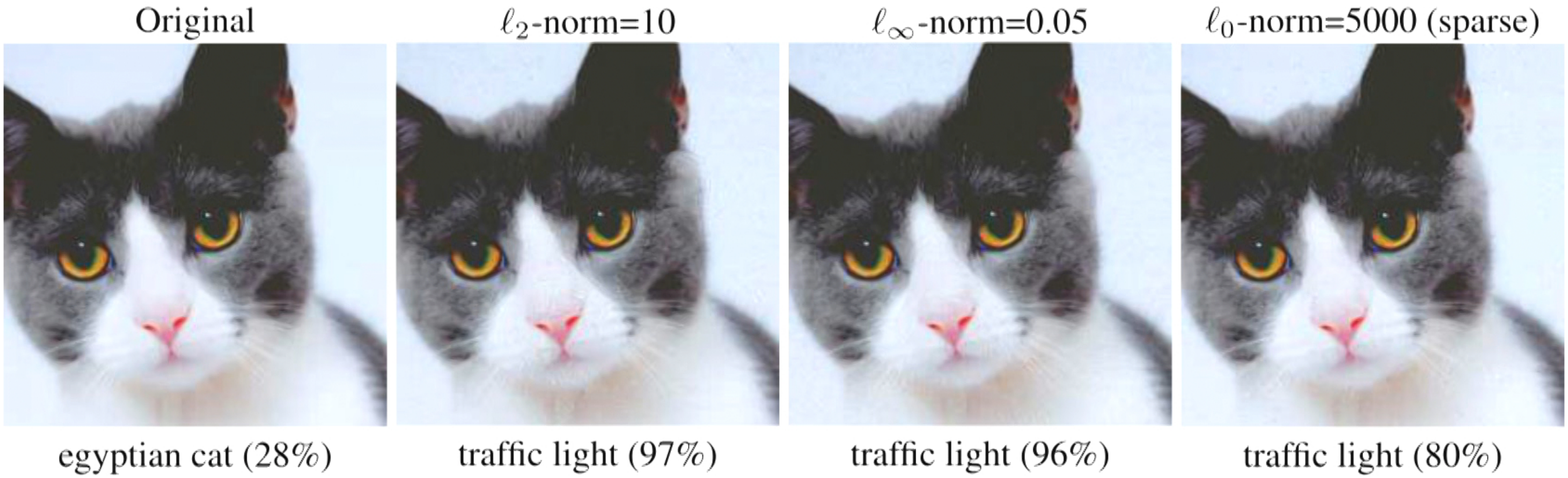


Machine Learning Has a Problem



Machine Learning Has a Problem

It generalizes to any norm:

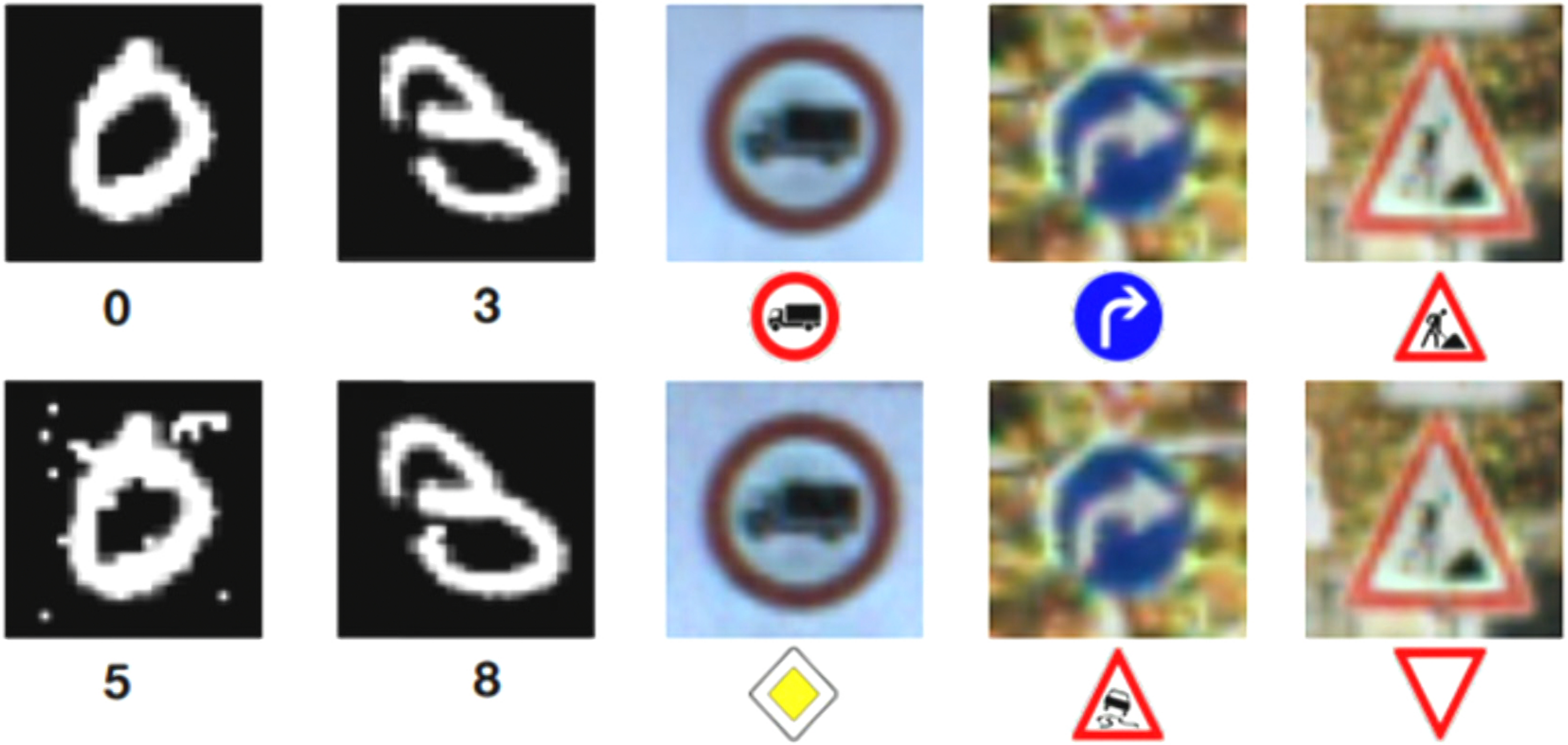


Source: arXiv:1809.02104

Early Literature on Adversarial Attacks

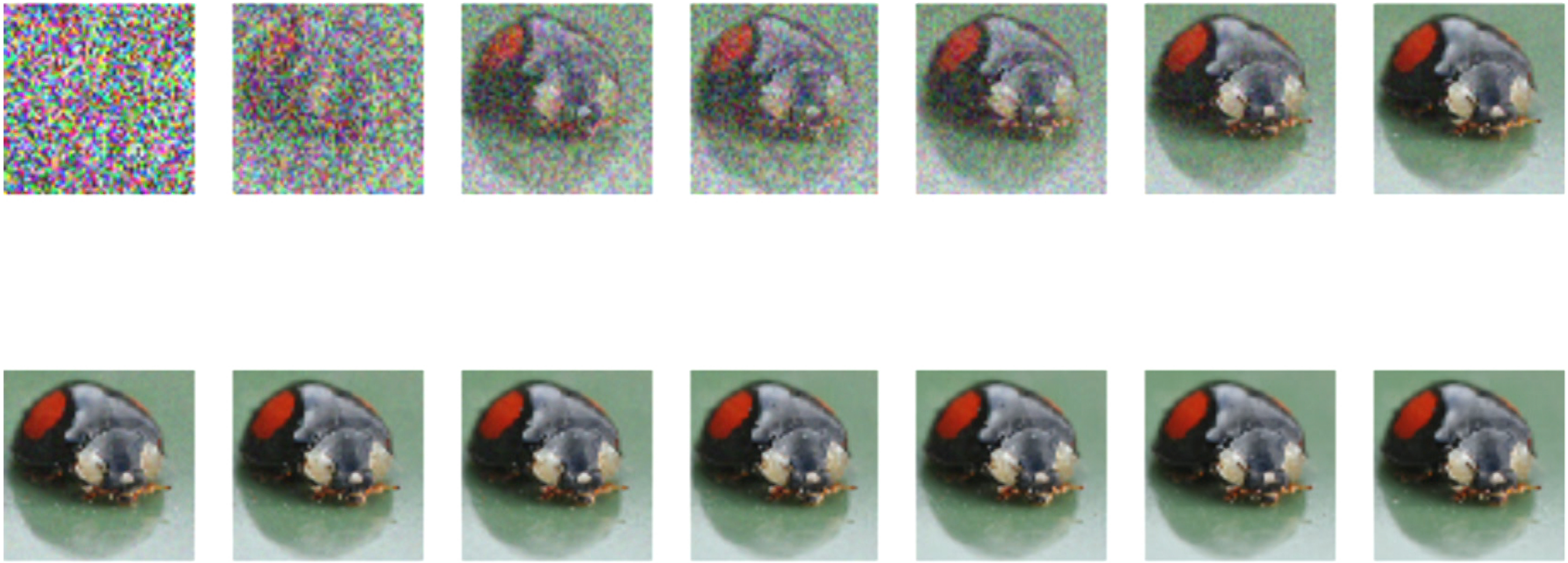
- Discovered in 2013 (arXiv:1312.6199).
- Sensitivity to adversarial perturbations.
- Distinct from random noise: generalization performance?
- White-box attack: access to gradient.
- Initially thought of as a property of deep learning networks.

Arms Race



Source: arXiv:1602.02697

This Bug is Fundamental

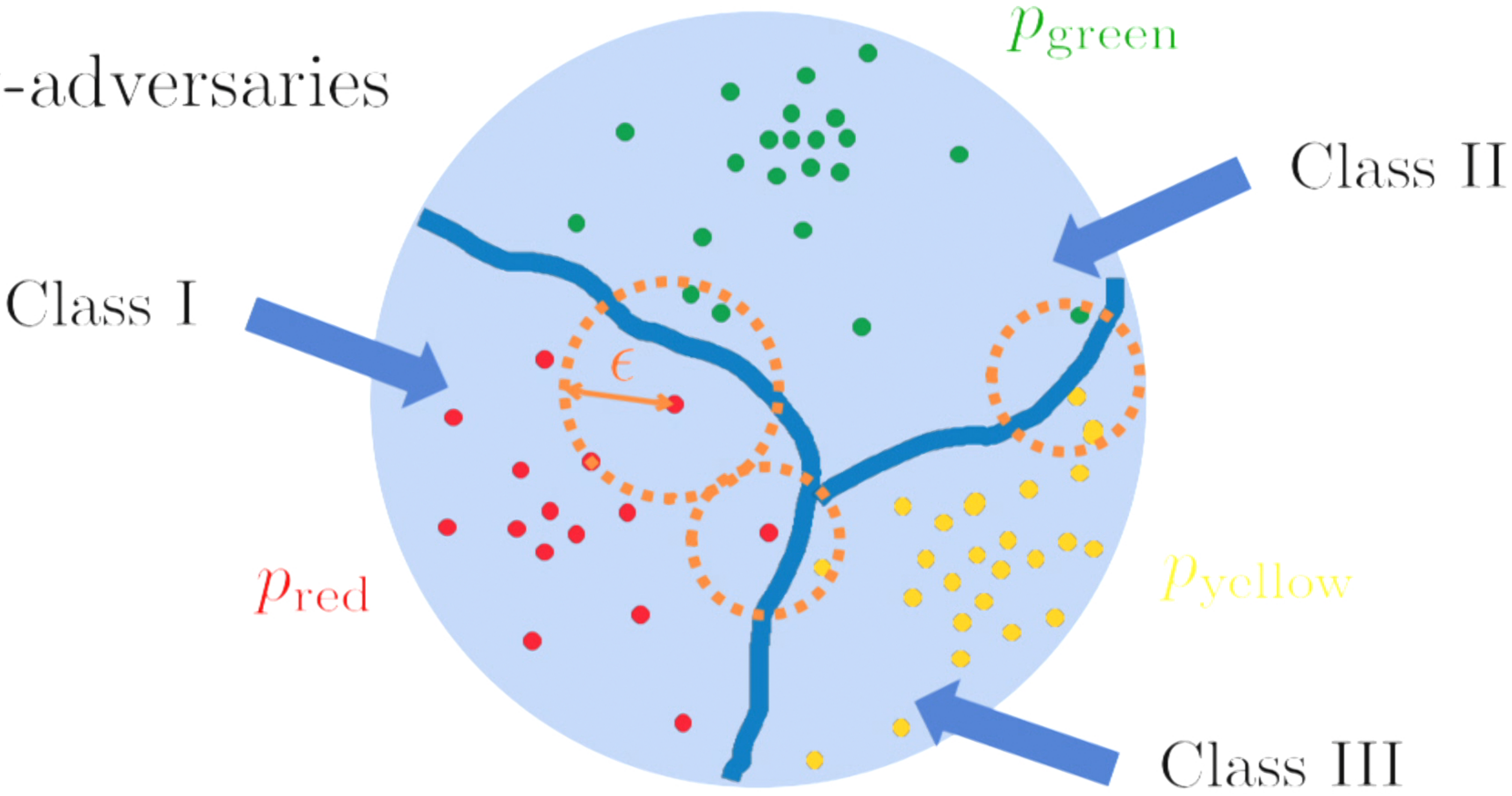


Source: arXiv:1712.04248

It's a black-box attack!

The Problem Is the Nature of High-Dimensional Geometry

ϵ-adversaries



p_{green}

Class II

Class I

ϵ

p_{red}

p_{yellow}

Class III

That Implies that Quantum Computing Has a Problem

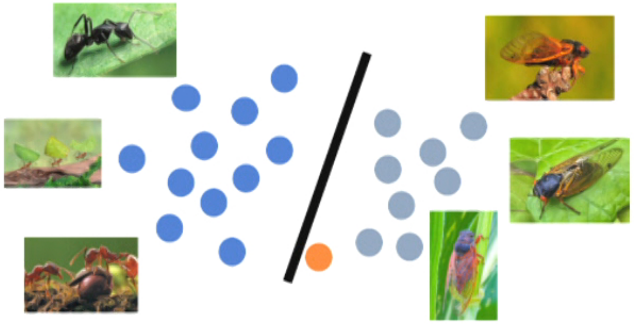
Machine Learning

Learning model: $h(x) : \mathbb{R}^d \mapsto \mathbb{Z}$.

Usually a composition:

$h(x) = \mathrm{threshold}(v(x))$

$v(x) : \mathbb{R}^d \mapsto \mathbb{R}$ is differentiable.



Quantum Computing and Quantum Protocols

$U : \mathbb{C}^d \mapsto \mathbb{C}^d$

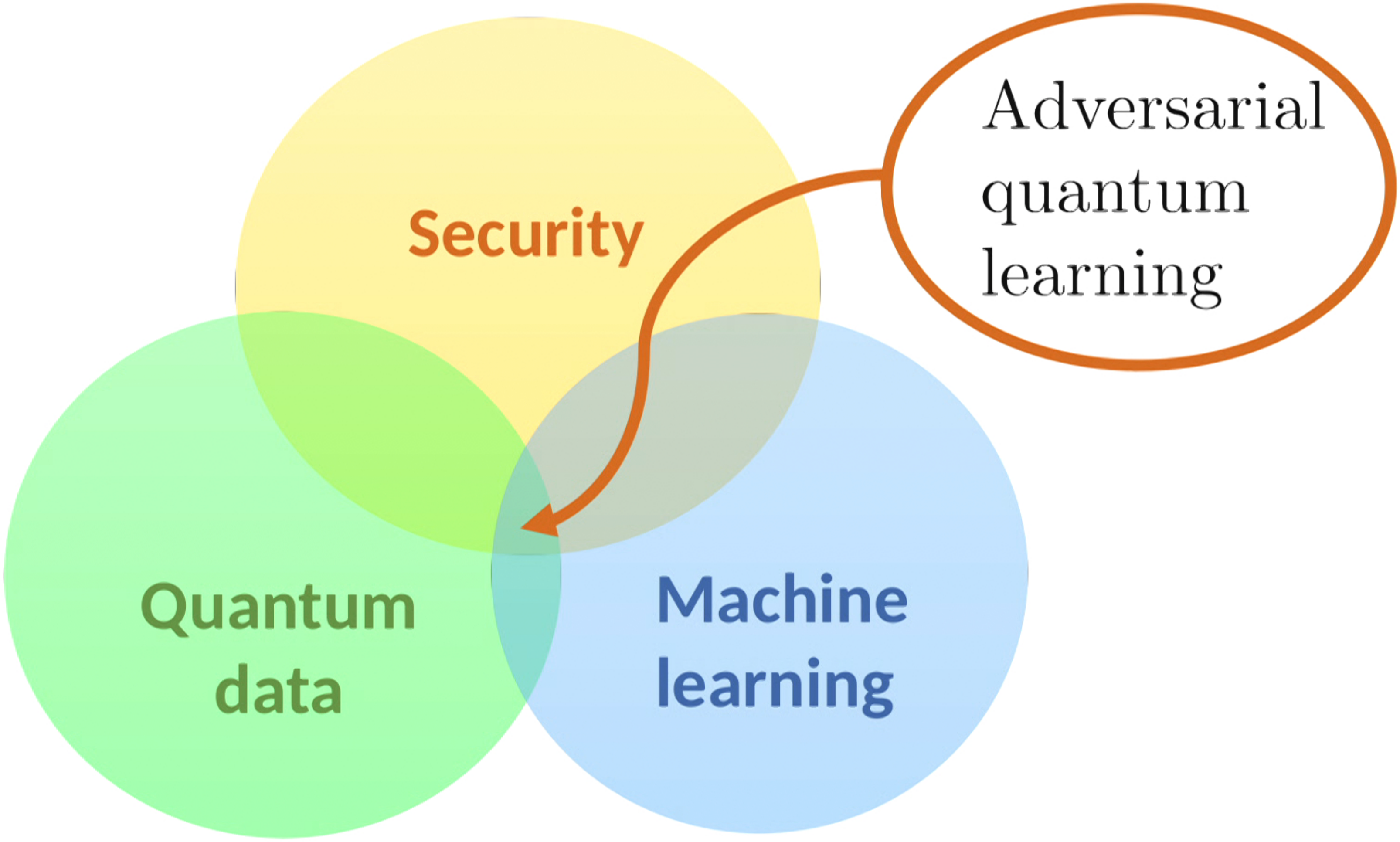
But: often measured and thresholded!

$h(|\psi\rangle) = \mathrm{threshold}(\mathrm{measure}(U|\psi\rangle))$

So:

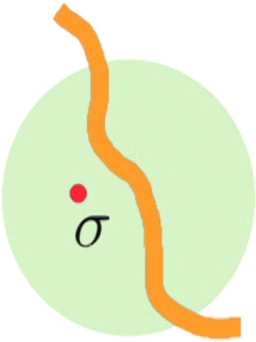
$h(|\psi\rangle) : \mathbb{C}^d \mapsto \mathbb{Z}$.

Setting


Security
Quantum data
Machine learning
Adversarial quantum learning

It's Not Just about Learning Protocols: Classification in General

Is σ entangled or not?



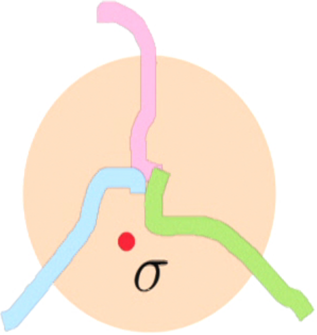
$f(\sigma) = \mathrm{Tr}(\mathcal{W}_{\mathrm{ent}}\sigma)$

Entanglement witness $\mathcal{W}_{\mathrm{ent}}$ learned

YC Ma, MH Yung, arXiv:1705.00813

Phase transitions

Which phase is σ in?

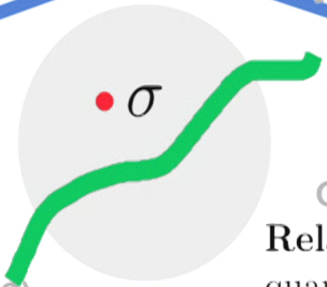


J. Carrasquilla, R. Melko, Nature Physics (2017)

P. Huembelli et al, PRB (2018)

Is σ an unusual state?

Anomaly detection



N. Liu, P. Rebentrost, PRA (2018)

G. Sentis et al, PRL (2016)

Related: quantum change point, quantum template-matching

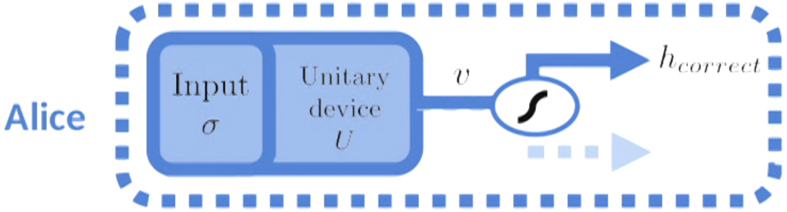
What's Safe?

Our scenario does not cover regression problems, e.g., a unitary transformation $U : \mathbb{C}^d \mapsto \mathbb{C}^d$.

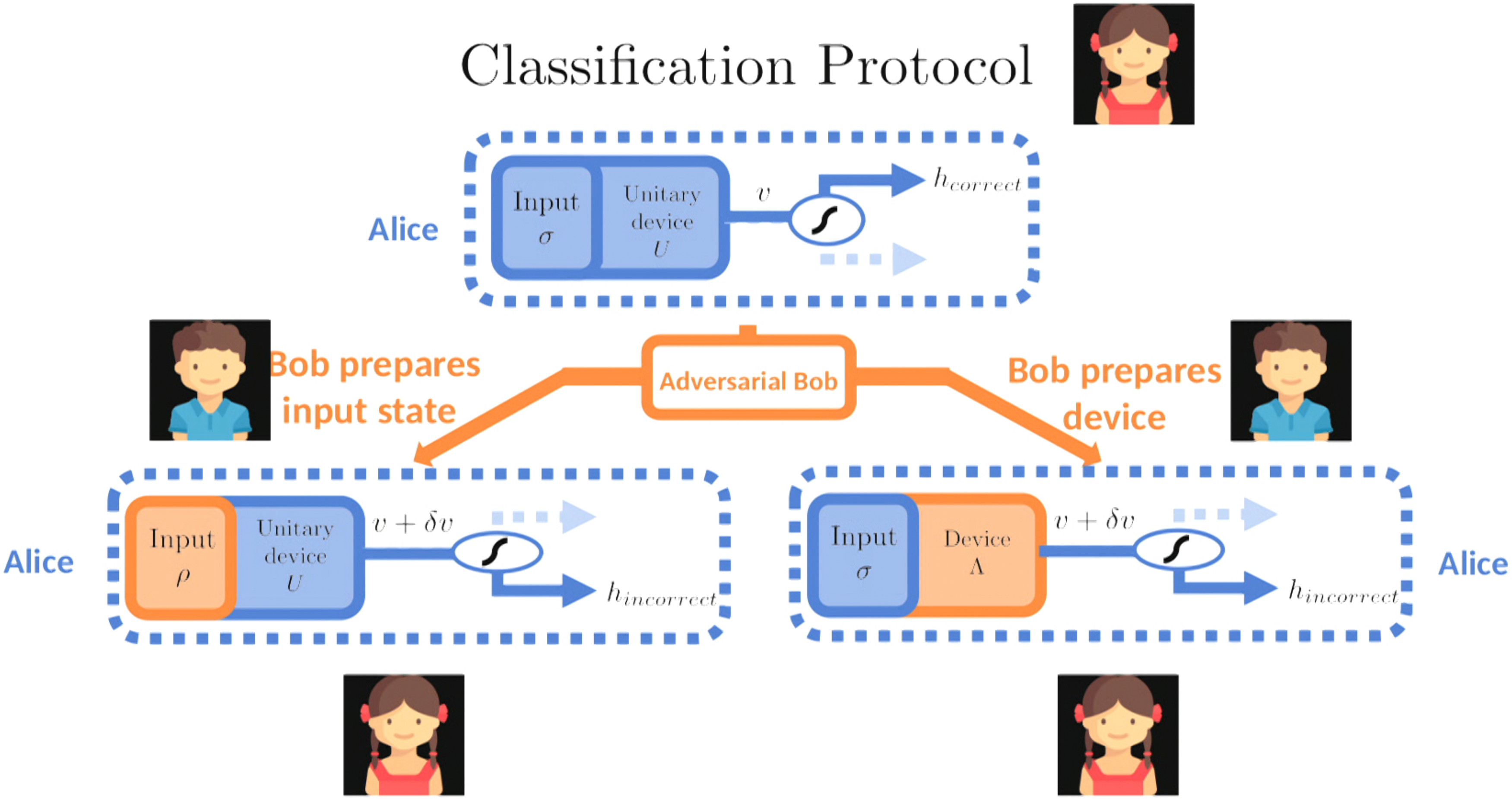
Uhlmann's theorem guarantees that the fidelity does not get worse after the transformation.

Scenarios

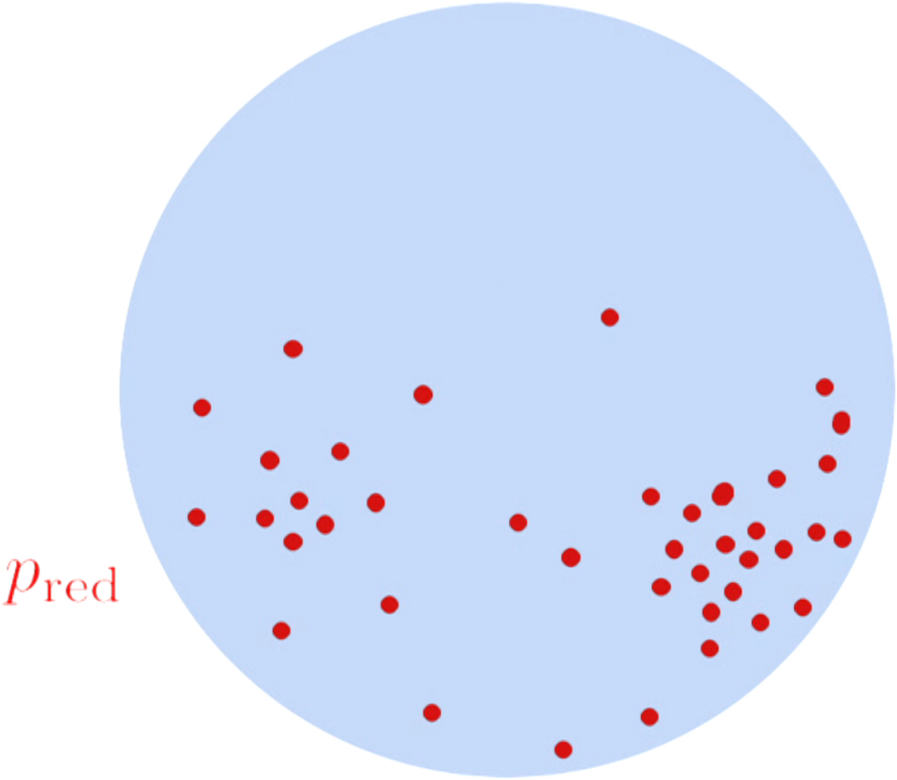
Classification Protocol



Scenarios

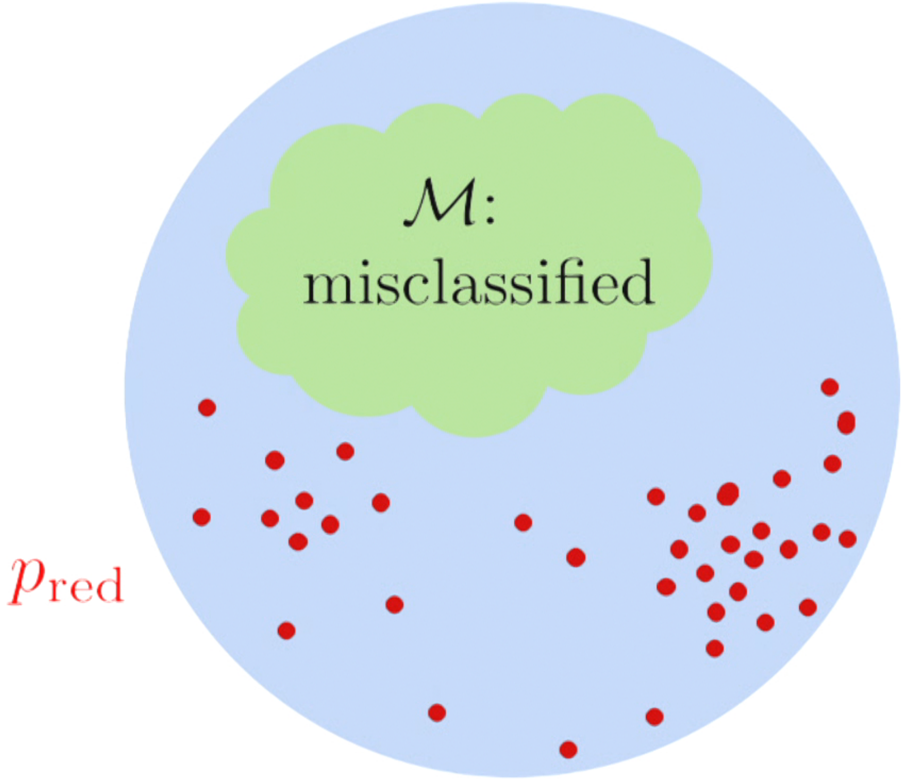

Classification Protocol
Alice
Input
σ
Unitary device
U
v
$h_{correct}$
Adversarial Bob
Bob prepares input state
Bob prepares device
Alice
Input
ρ
Unitary device
U
$v + \delta v$
$h_{incorrect}$
Input
σ
Device
Λ
$v + \delta v$
$h_{incorrect}$
Alice

The Geometry of the High-Dimensional Space



A class is distributed according to some probability distribution p.

The Geometry of the High-Dimensional Space

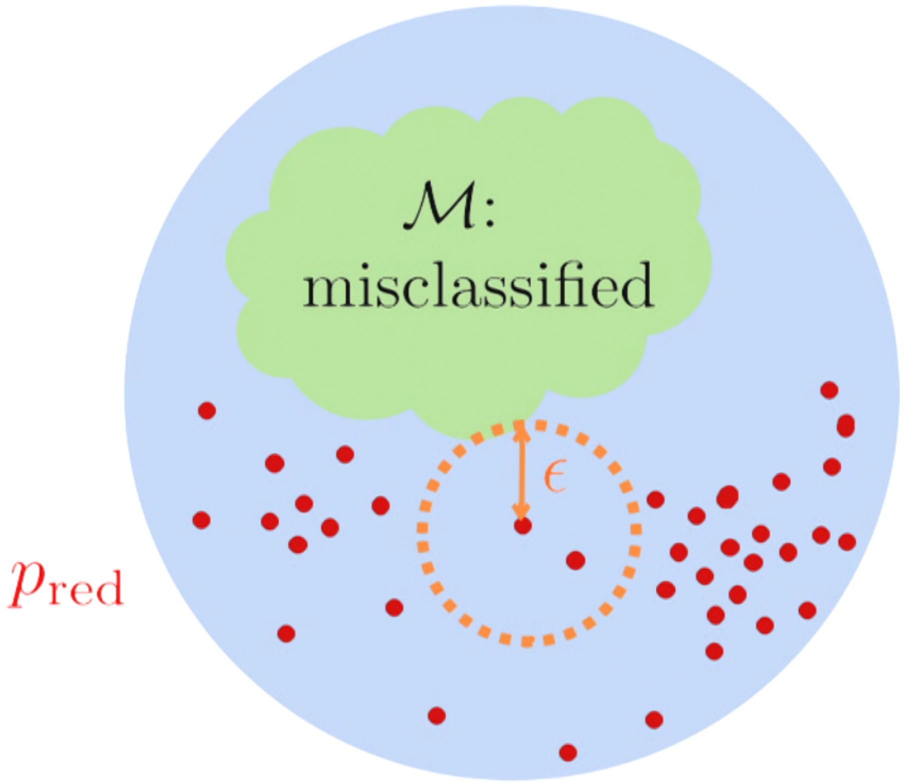


Assumption: A model trained on it has nonzero error rate on p.

Probability of misclassifying: $\mu(\mathcal{M})$.

Note: this is not the same as empirical error!

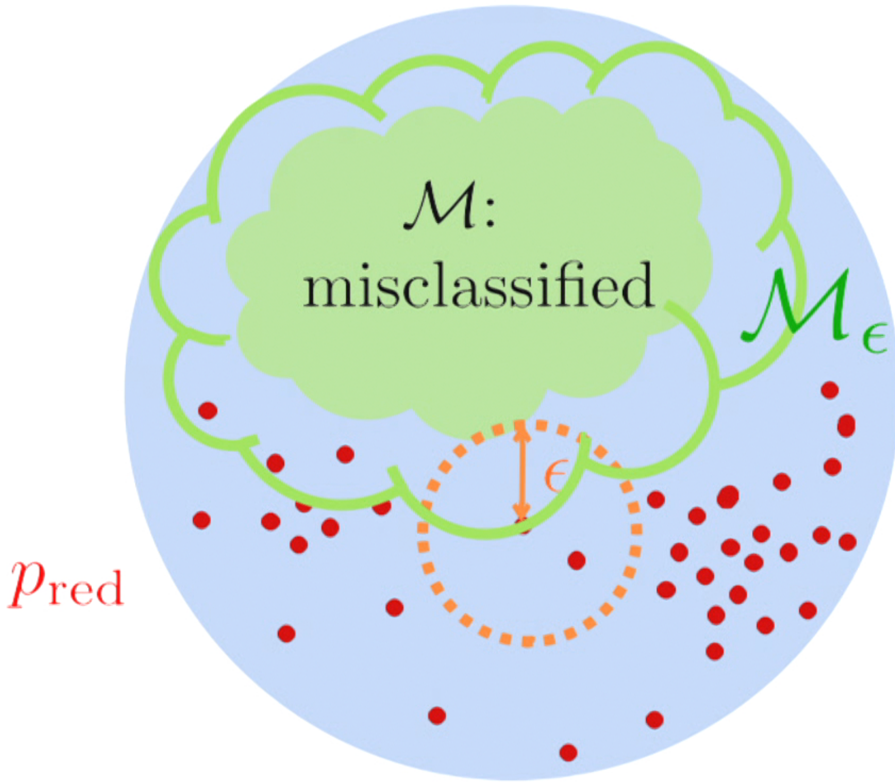
The Geometry of the High-Dimensional Space



Probability of misclassifying: $\mu(\mathcal{M}) > 0$.

ϵ-adversaries: within ϵ distance to $\mathcal{M}$.

The Geometry of the High-Dimensional Space



Probability of misclassifying: $\mu(\mathcal{M}) > 0$.

ϵ-adversaries: within ϵ distance to $\mathcal{M}$.

ϵ-expansion of $\mathcal{M}$: $\mathcal{M}_\epsilon$

Adversarial risk: $\mu(\mathcal{M}_\epsilon)$.

Concentration Function

Definition. For $\tilde{\Sigma} \subset \Sigma$ with distance measure D and probability measure μ, we define the *concentration function* $\alpha(\epsilon)$ as $\alpha(\epsilon) \equiv 1 - \inf\{\mu(\tilde{\Sigma}_\epsilon) | \mu(\tilde{\Sigma}) \geq 1/2\}$.

This does **not** mean that the error rate has to be 0.5!

Anti-concentrated space: $\alpha(\epsilon) \approx 0.5$
Concentrated space: $\alpha(\epsilon) \approx 0$

Normal Levy Family

If $\tilde{\Sigma}$ is equipped with a vector space with dimension d and

$$\alpha(\epsilon) \leq l_1 e^{-l_2 \epsilon^2 d},$$

the corresponding space belongs to the (l_1, l_2)-*normal Levy family*, where $l_1, l_2 > 0$.

Concentration Theorems Apply in $SU(d)$

Theorem. Suppose $\sigma \in SU(d)$ and a perturbation $\sigma \to \rho$ occurs, where $d_{HS}(\sigma, \rho) \le \epsilon$ and d_{HS} is the Hilbert-Schmidt distance. If the adversarial risk is bounded above by R, then ϵ^2 must be bounded above by

$$\epsilon^2 < \frac{4}{d} \ln \left(\frac{2}{\mu(\mathcal{M})(1-R)} \right).$$

Proof. By showing that $SU(d)$ equipped with the Haar measure and Hilbert-Schmidt metric belongs to the $(\sqrt{2}, 1/4)-$normal Levy family.

Consequence. It becomes more and more difficult to certify whether σ has been adversarially perturbed.

A Fundamental Trade-Off

Tension develops between

- The resources required to ensure robustness against misclassification
- The quantum advantage expected as the dimension grows.

Defense Is Very Expensive

Theorem. Alice needs $\mathcal{N}_{state}$ copies of ρ to estimate the fidelity $\mathcal{F}(\sigma, \rho)$ to a precision that is necessary to guarantee that the adversarial risk is at most R with failure probability Δ. Then she requires at least

$$\mathcal{N}_{state} \geq \frac{d^4}{g^2(\mu(\mathcal{M}), R)\Delta}$$

copies of ρ, where $g(\mu(\mathcal{M}), R) = 2\ln(2/(\mu(\mathcal{M})(1 - R)))$.

Bad for Quantum Learning

Classify states that we do *not* have the classical descriptions of:

- Quantum-template matching (Sasaki, 2002);
- Quantum anomaly detection (Liu and Rebentrost, 2018).

Control-SWAP gate serves as the key component to the classification device.

- Difficult to experimentally realise.
- Gates delegated to Bob to prepare.
- Certifying the control-SWAP gates to a constant precision η is insufficient and the minimum η to be robust against adversarial attacks must grow with scaling at least d^2

Bad for Studying Phase Transitions

Too pathways:

- Classical machine learning on classical simulations (Carrasquilla and Melko, (2017); Huembeli et al., (2018))
- Classical machine learning on data coming from quantum experiment (Uvarov et al, 2019).
- Potentially imperfect processing of experimental data from quantum device (Harris et al., (2018); King et al., (2018)).

The space is high dimensional to begin with, plus:

- Delegated state preparation.
- Delegated quantum operations.

Bad for Many Other Protocols

- Imperfect witnesses, e.g., learned entanglement witness (Ma and Yung, 2017).
- Classifying nonlocal correlations (Deng, 2017).

Conclusions

Beware of the bug!

Intrinsic problem in any quantum protocol that maps to discrete sets and has a nonzero error rate.

Beyond $SU(d)$?

What about CV systems?