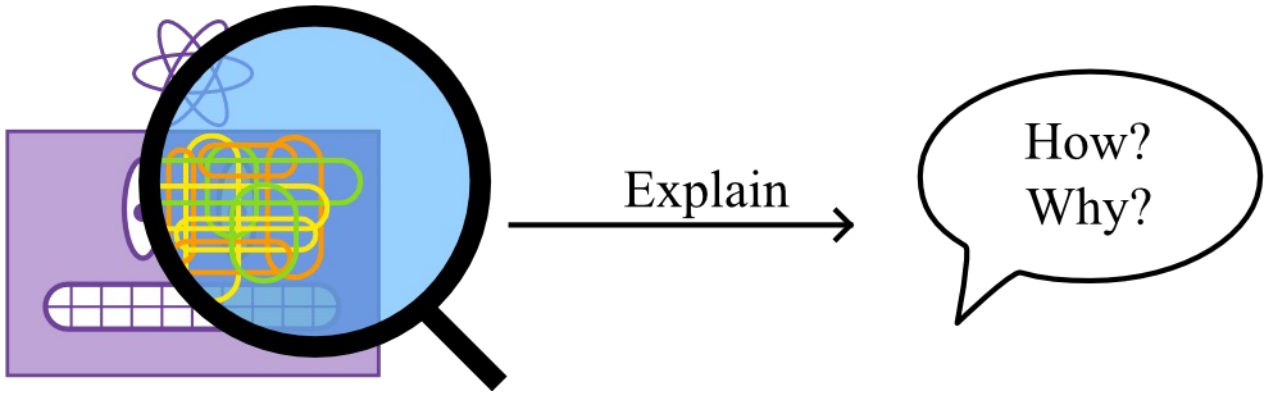
# OPPORTUNITIES AND LIMITATIONS OF EXPLAINING QUANTUM MACHINE LEARNING

Jonas Naujoks @ QTML 2024




Elies Gil-Fuster, **Jonas R. Naujoks**, G. Montavon, T. Wiegand, W. Samek, Jens Eisert

# What to expect from this talk?

- Explore the application of explanation techniques
  - What is explainability?
  - Why should we care?
  - Challenges

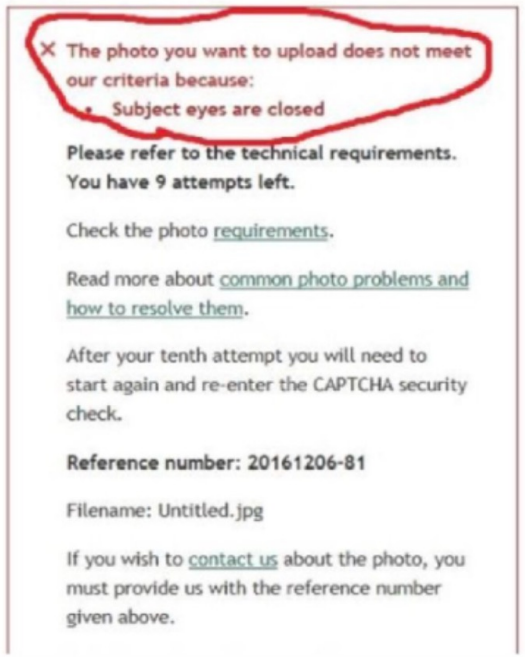
- We do not claim
  - Any sort of quantum advantage
  - That our approach works best

# The Problem of the Black Box

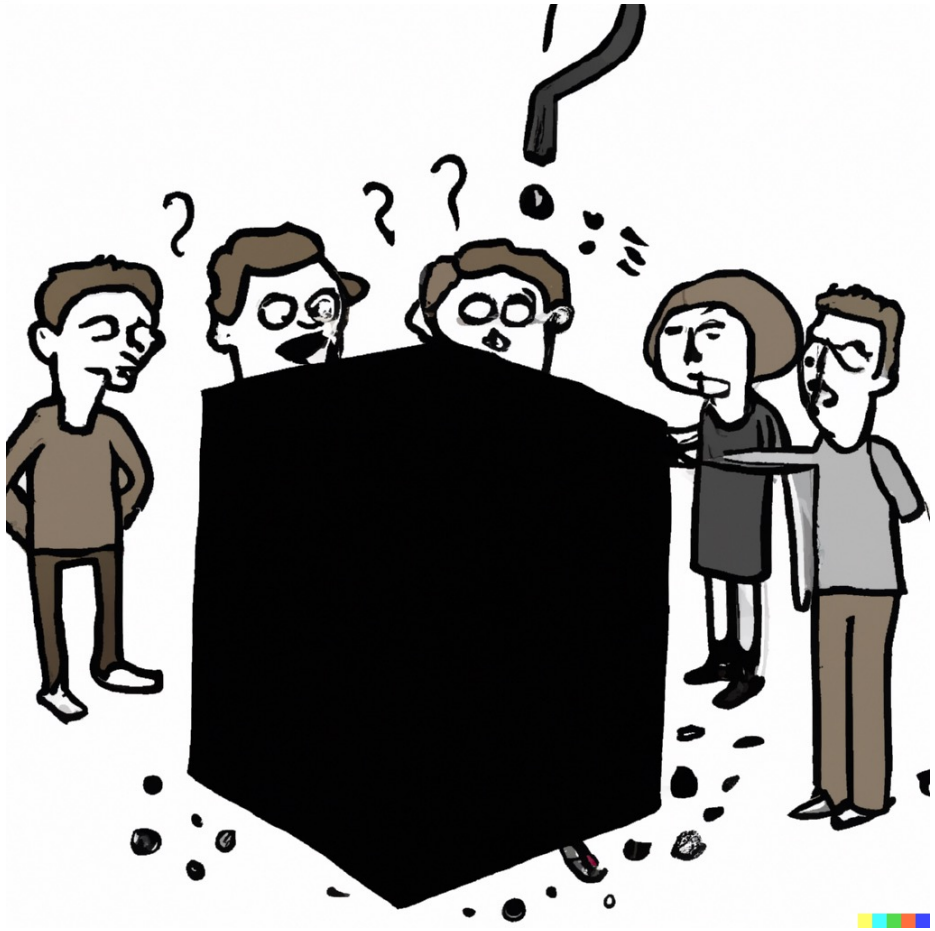
## New Zealand passport robot thinks this Asian man's eyes are closed

By James Griffiths, CNN

Updated 1:46 AM ET, Fri December 9, 2016

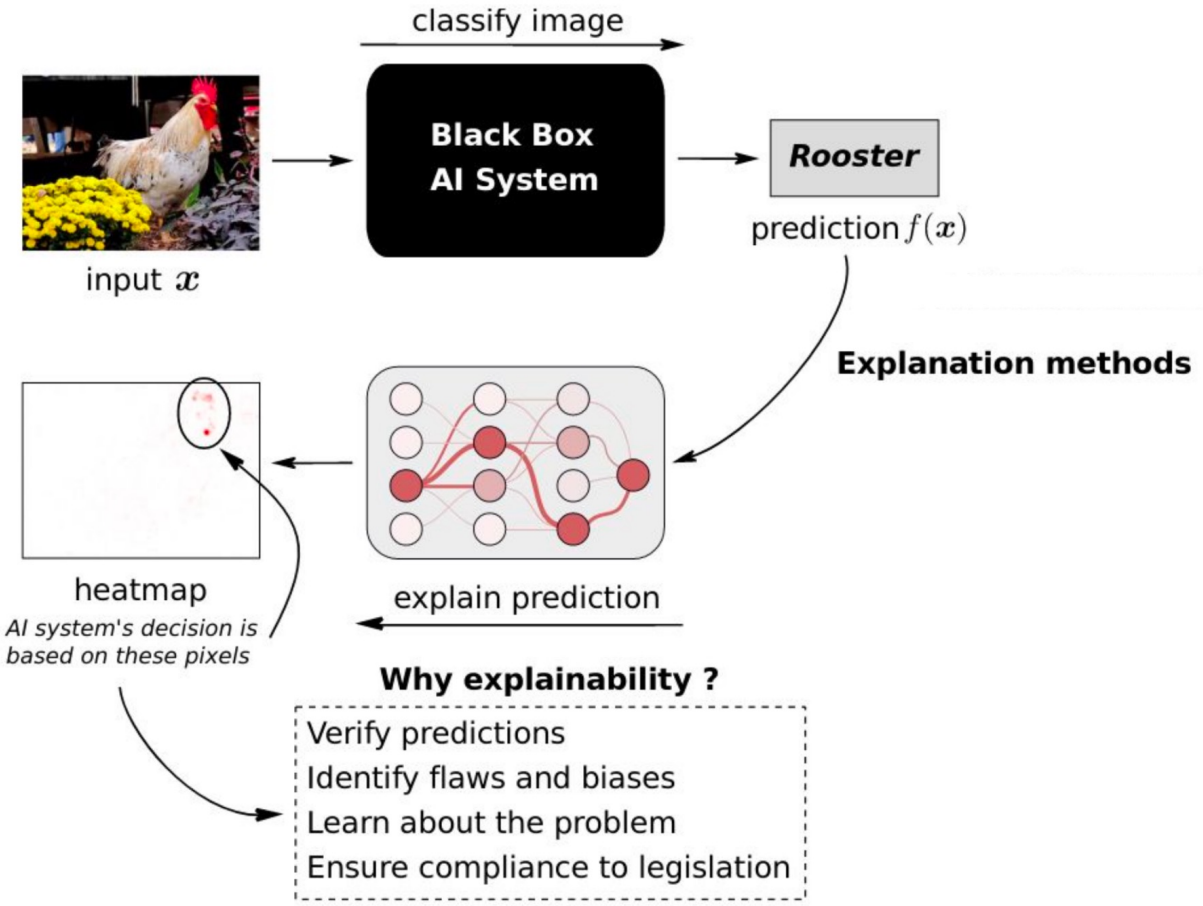



New Zealand's online passport application system couldn't recognize Richard Lee's open eyes.



# What is the Aim of Explainable AI?

- To make black-box AI models interpretable!

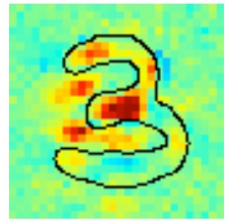


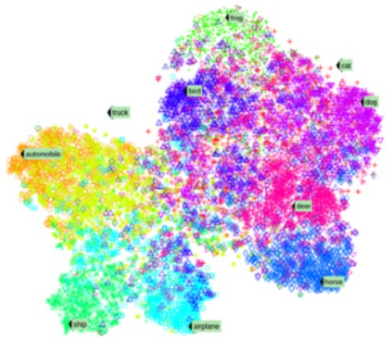

Samek et al. (2017)

# What is to be “explained”?

prediction

*“Explain why a certain pattern x has been classified in a certain way f(x).”*





model

*“What concept does a particular neural encode?"*

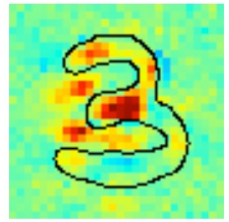
data

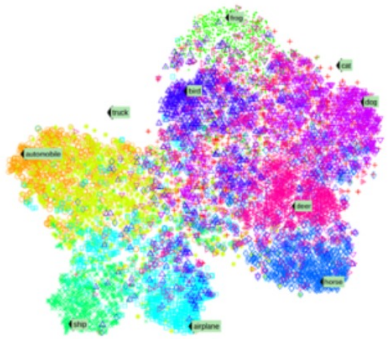
*“Which samples / dimensions of the data are most relevant for the task / training.”*

# What is to be “explained”?

prediction

*“Explain why a certain pattern x has been classified in a certain way f(x).”*





model

*“What concept does a particular neural encode?"*

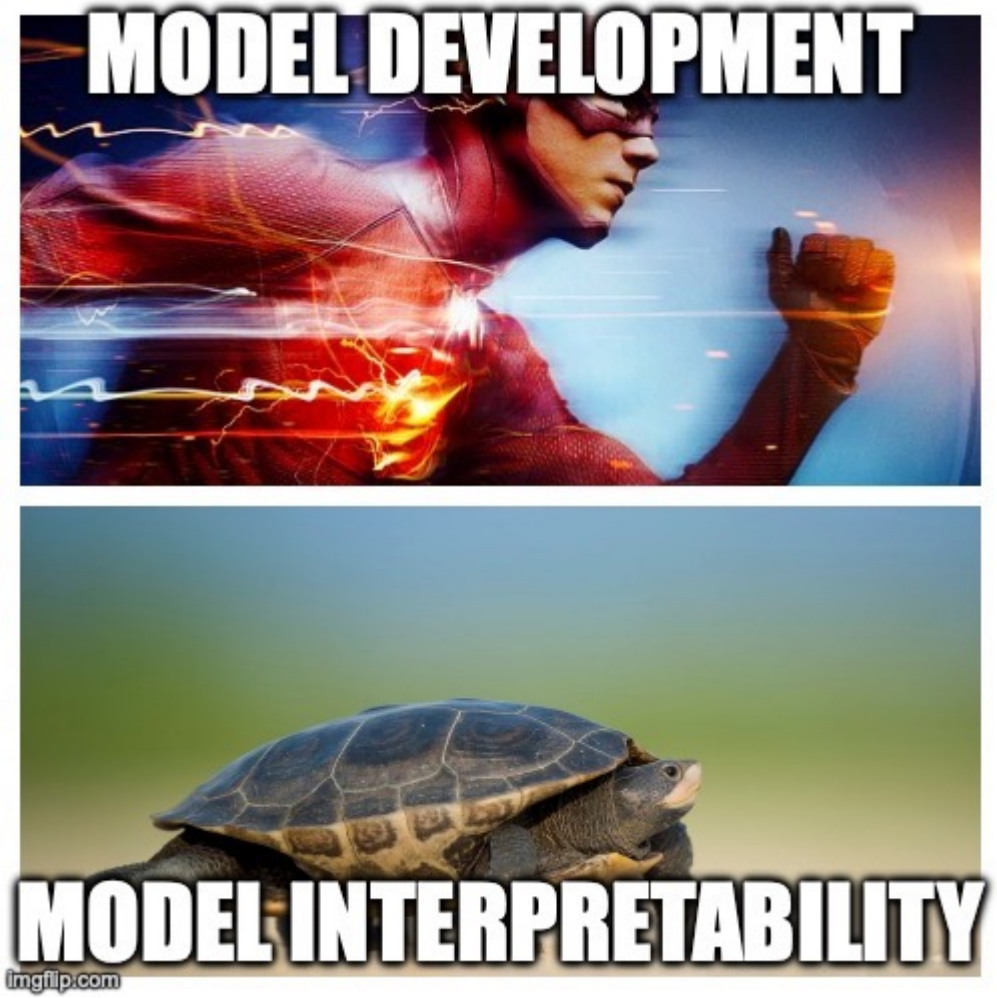
data

*“Which samples / dimensions of the data are most relevant for the task / training.”*

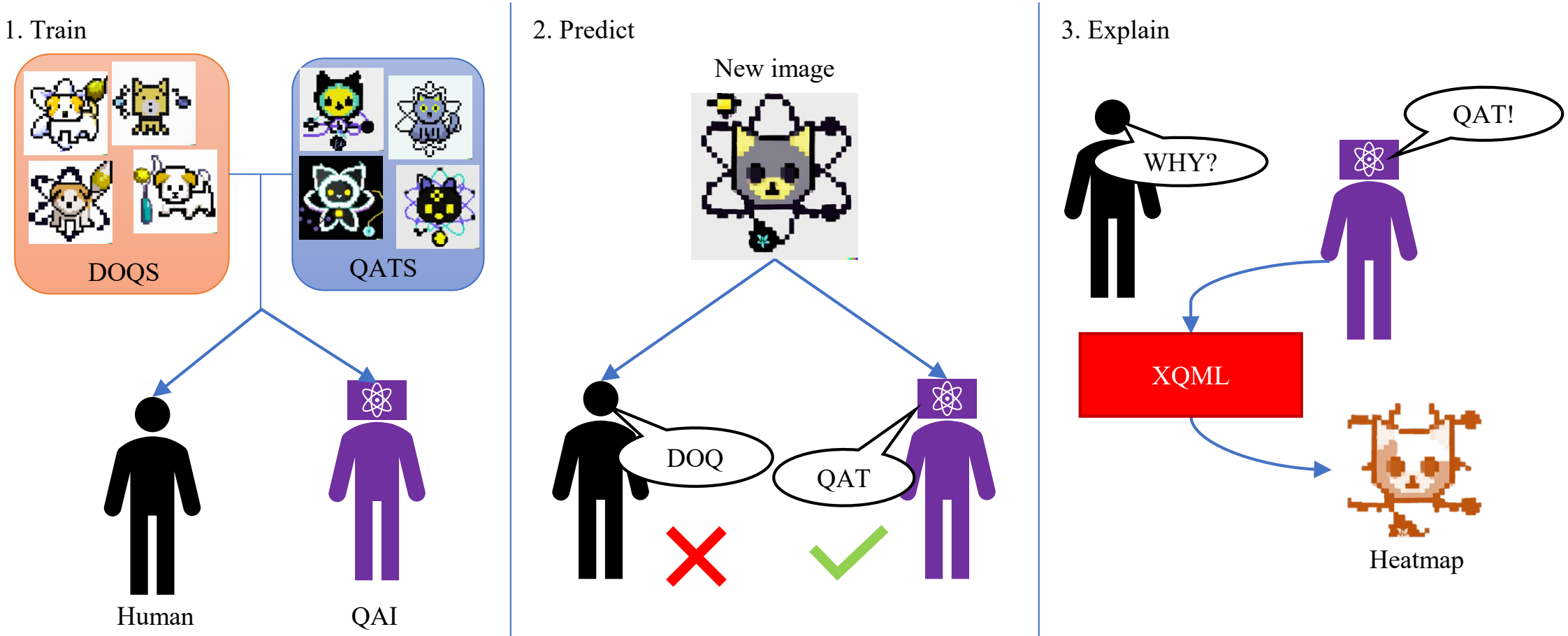
# Why should we care?

- Classical ML is becoming ever more powerful
- Explainability and intepretability research lags behind model development
- **This doesn’t have to be the case for QML!**

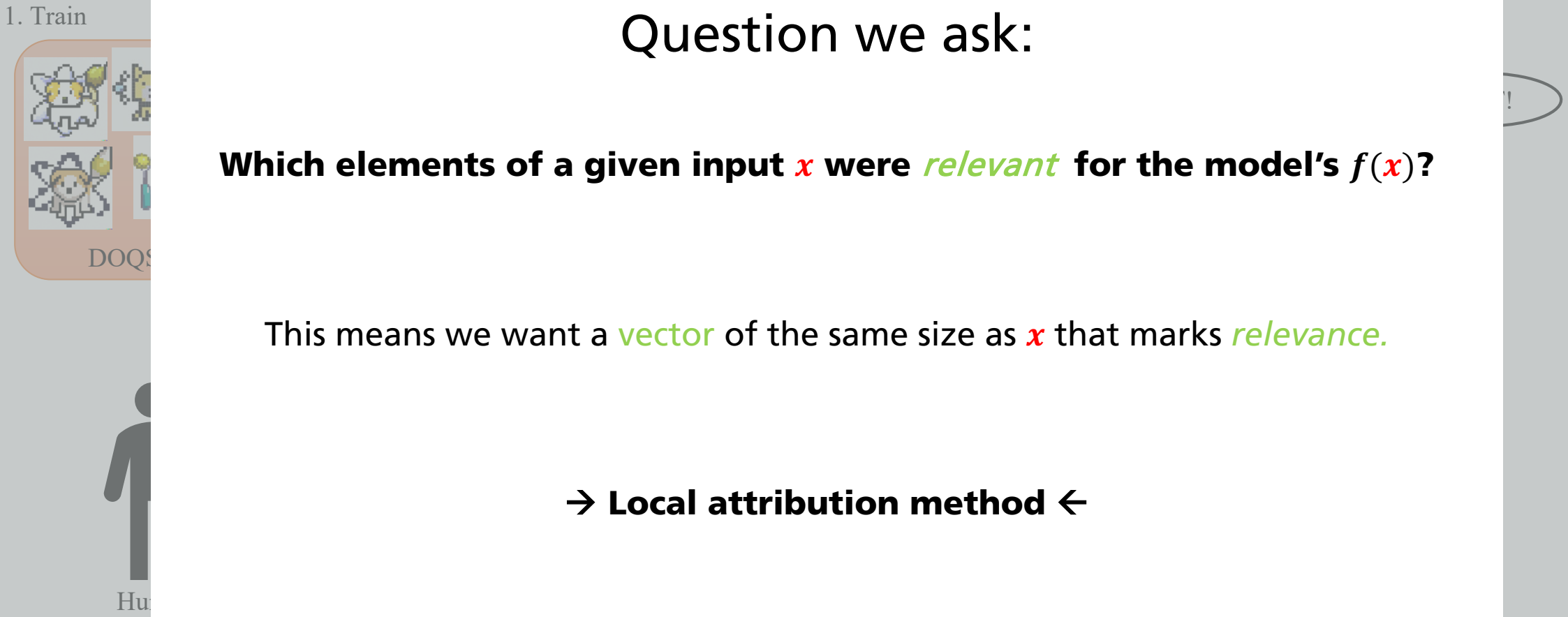
## Classical ML



# Explaining Quantum Machine Learning Models (PQCs)



# Explaining Quantum Machine Learning Models (PQCs)




## Question we ask:

**Which elements of a given input $x$ were *relevant* for the model's $f(x)$?**

This means we want a vector of the same size as $x$ that marks *relevance.*

**→ Local attribution method ←**

# Explaining Quantum Machine Learning Models (PQCs)

- Consider functions from single qubit encoding PQCs

$$f_v(x) = \sum_{\omega \in \Omega} (a_\omega(\hat{v}) \cos(\omega, x) + b_\omega(\hat{v}) \sin\langle \omega, x \rangle)$$

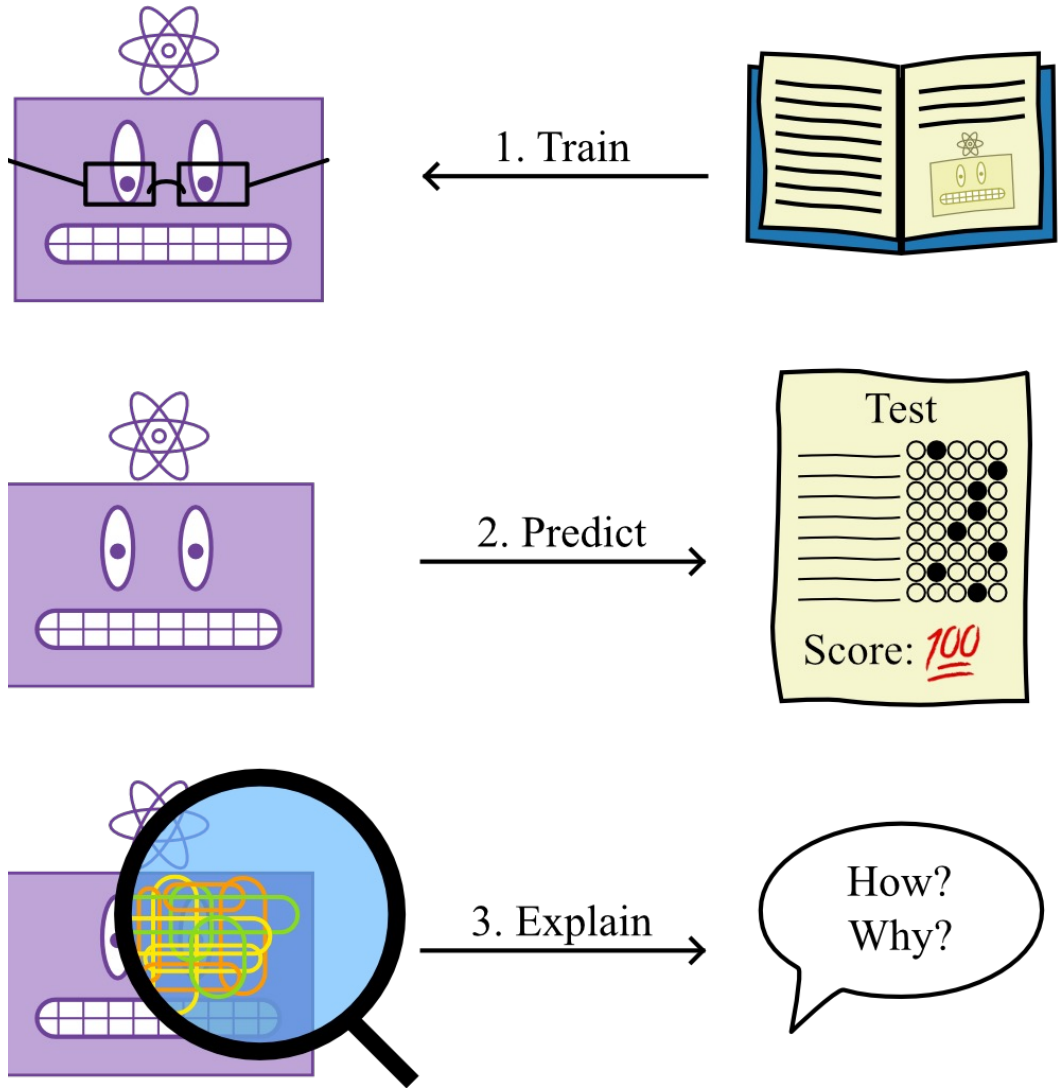
- Some classical methods can applied directly

- Taylor decomposition + trigonometry gives our attribution methods

# Numerical Results

- Devise an easy proof-of-principle task + pipeline
  - Toy classification task
  - 2 new + some classic methods
  - Explanations need to be evaluated!

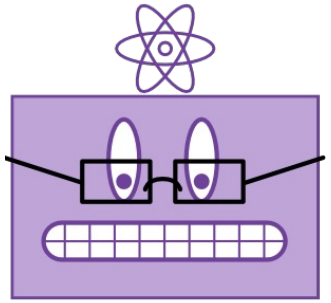
- **Result**: Comparable classical and new explainability methods



# Challenges of Explaining QML

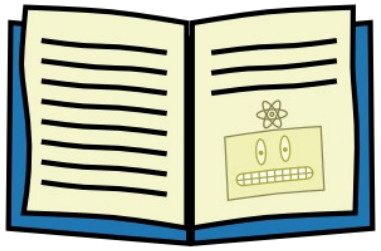
- Core features of QM render impossible direct application of classical methods:
  - No-cloning-theorem, exponential state space, concentration phenomena…
  - Single non-linearity, different structure
  - Underlying quantum computing hardware

- Scalabity of our proposed methods:
  - Involve several steps that requires exponential resources



# Thank you for your attention!

Expect the paper on ArXiv soon!



jonas.naujoks@hhi.fraunhofer.de