Deep Learning (Explainable AI)

why? → open up black box

Activation maximization framework:

$$\max_{x \in X} \log p(w_c | x) - \lambda \|x\|^2$$

↳ class prototype

- find x with highest score
- $\lambda \|x\|^2$: regularization

In code space: $\max_{z \in Z} \log p(w_c | g(z)) - \lambda \|z\|^2$

$x \to g(z)$

Ante-hoc: choose interpretable model and train it

Post-hoc: choose working model and make interpretable afterwards

↳ possible since 2015

Layer-wise relevance propagation (LRP)

- deep Taylor decomposition
- applicable to all NNs with monotonous activation fct.
- shows which pixels contribute to classification