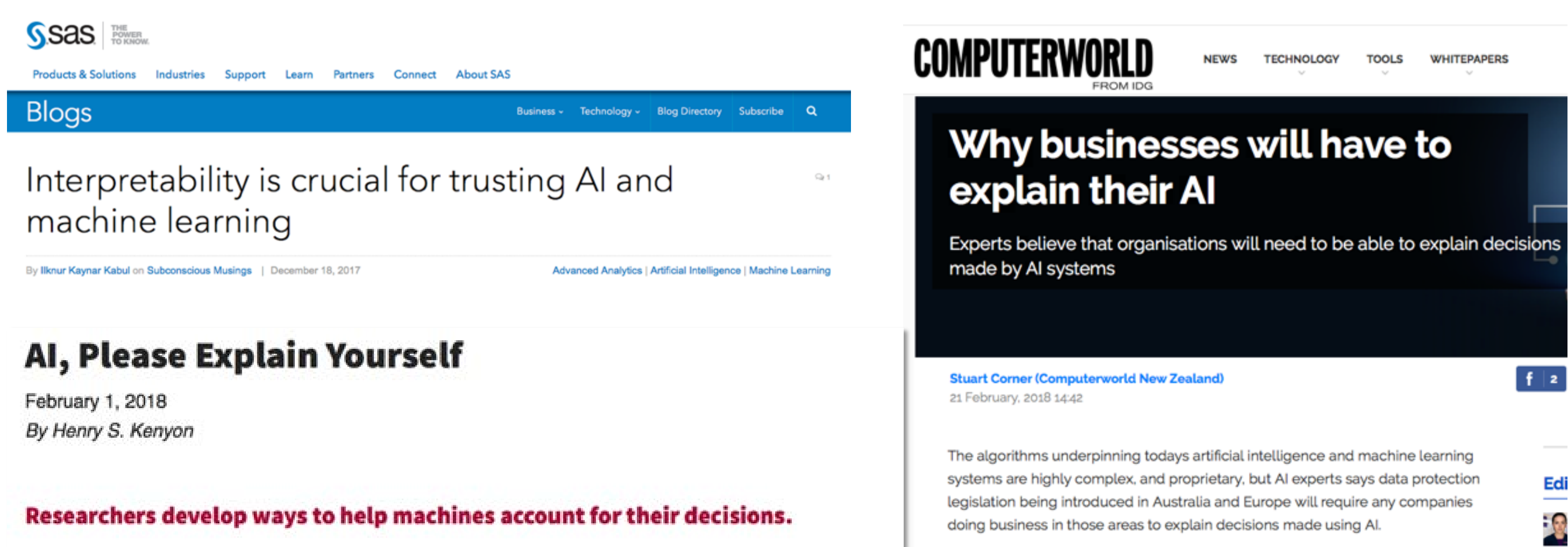


Motivation

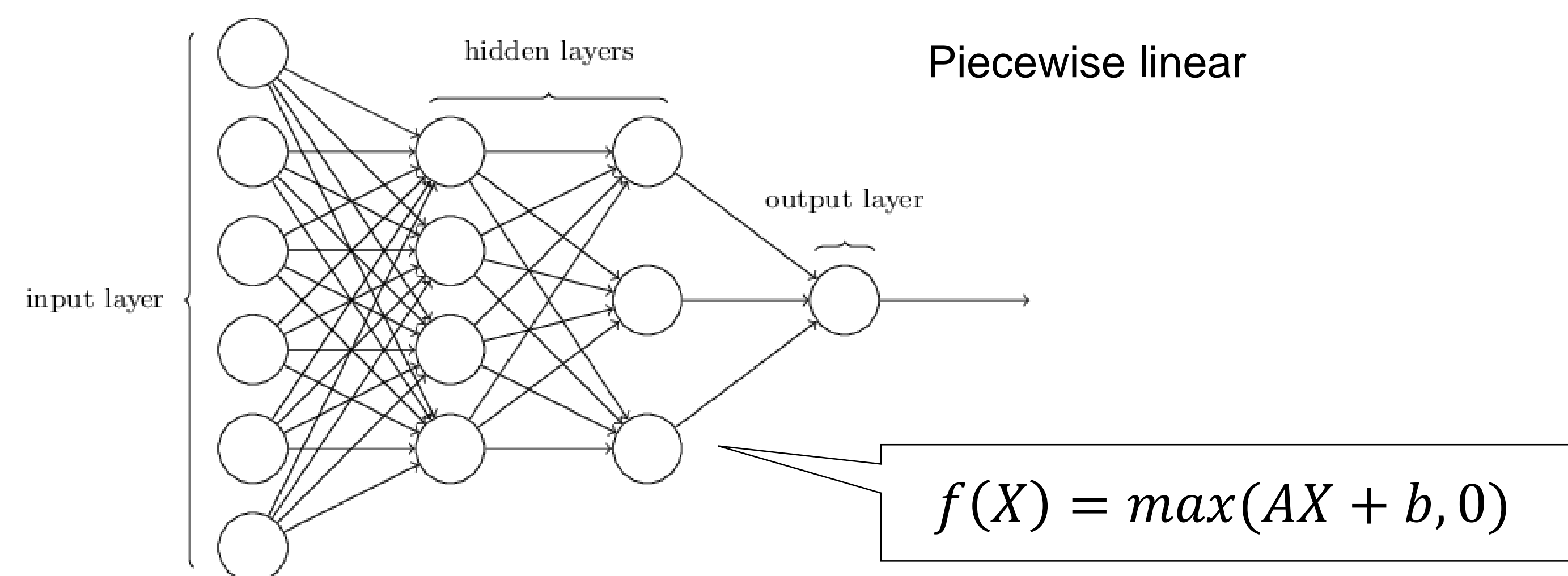


General Data Protection Regulation (enacted 2016, taking effect 2018)

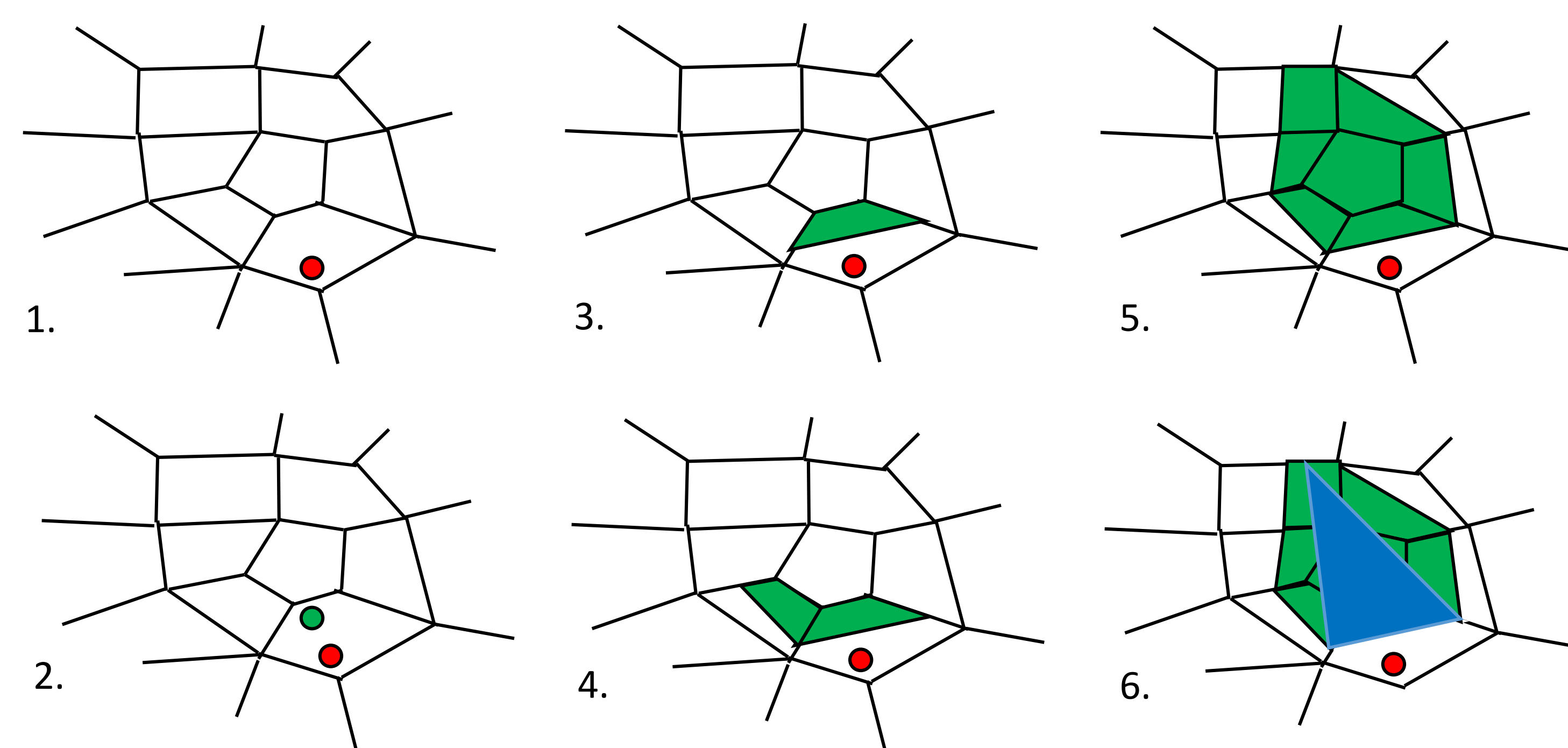
- Grants consumers a "right to explanation"
- Fine for violations: up to 10 million EURO

Algorithm

We focus on **neural networks with ReLU activations**:



Our algorithm is inspired by **concolic testing**:



Problem Definition

Judgement Problem: special binary classification problem where one result is more desirable than the other.



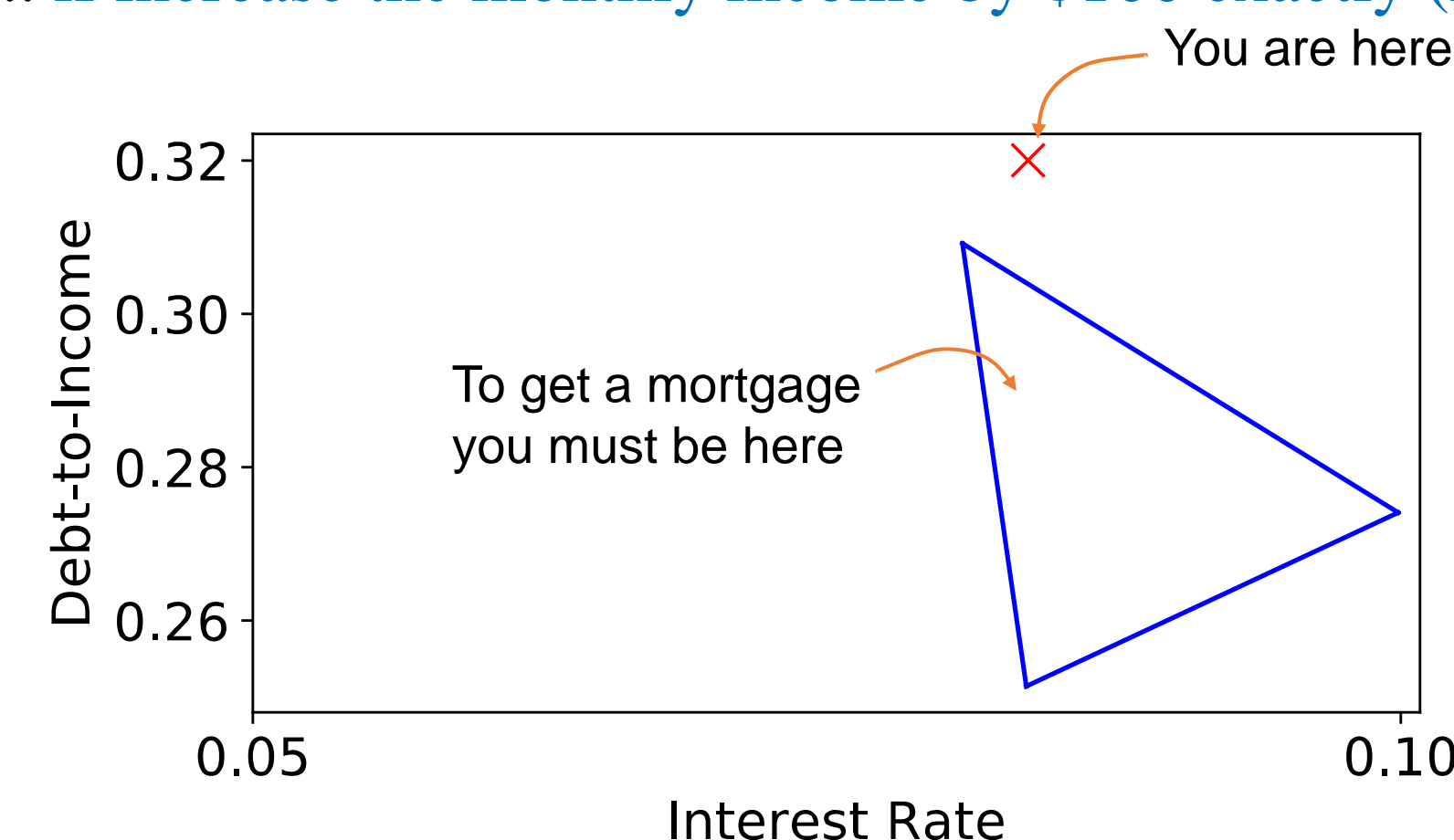
Why?

Corrections as interpretations:

- Increase your salary
- Ask for less money
- Improve your credit score
- Look at a different type of house
- Lower your current debt
- ...

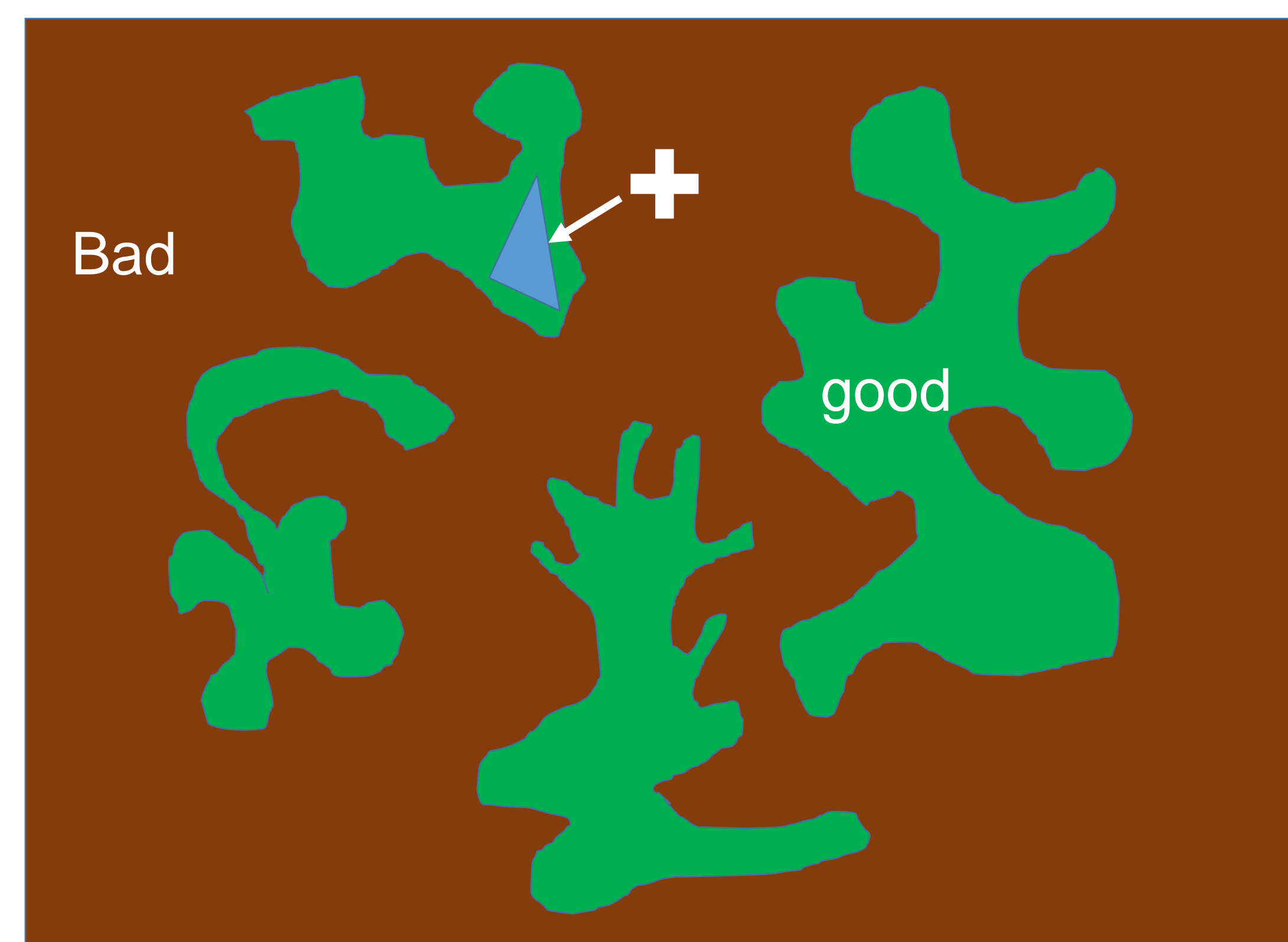
Desired properties of corrections:

- Minimal**
 - ✓ Your application would be approved if you increase the monthly income by \$100
 - ✗ ... if you become a millionaire living in Montana
- Stable**
 - ✓ ... if you increase the monthly income by \$100 (+/- is fine)
 - ✗ ... if increase the monthly income by \$100 exactly (not even \$99 or \$101)



Symbolic

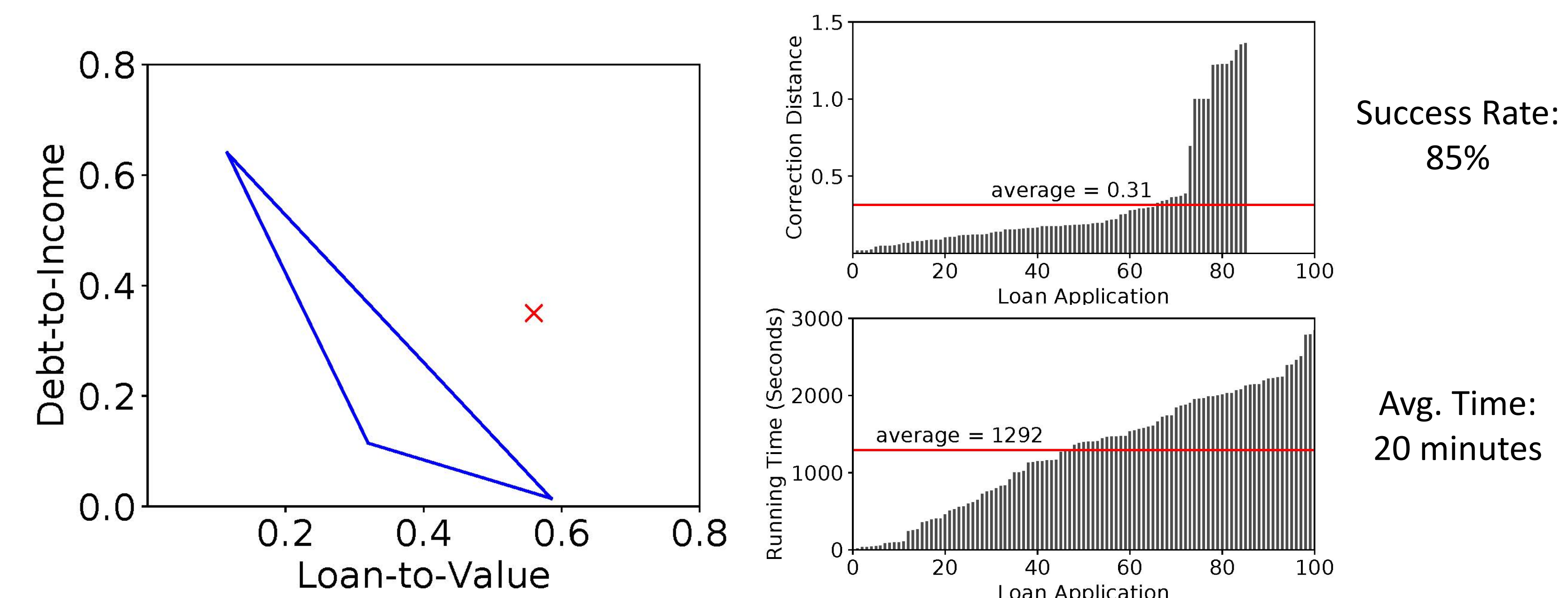
Our corrections are **sound underapproximations** that are **local** to the input



Experiment Results

Mortgage Underwriting

Network structure: 5-layer feedforward network with 1000 neurons
Dataset: applications and performance of 36 million single-family loans

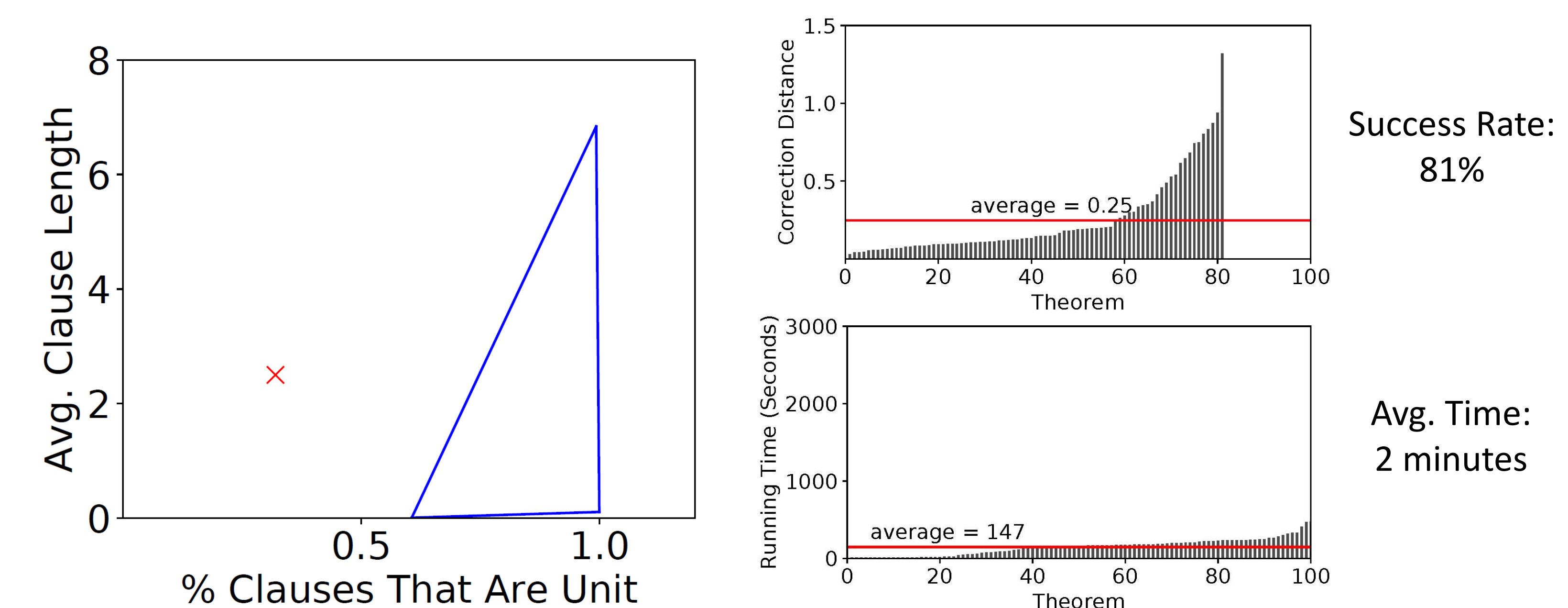


Success Rate: 85%

Avg. Time: 20 minutes

Solver Performance Prediction

Network structure: 8-layer feedforward network with 800 neurons
Dataset: statistics of 6k first-order theorems and whether they can be solved by a solver

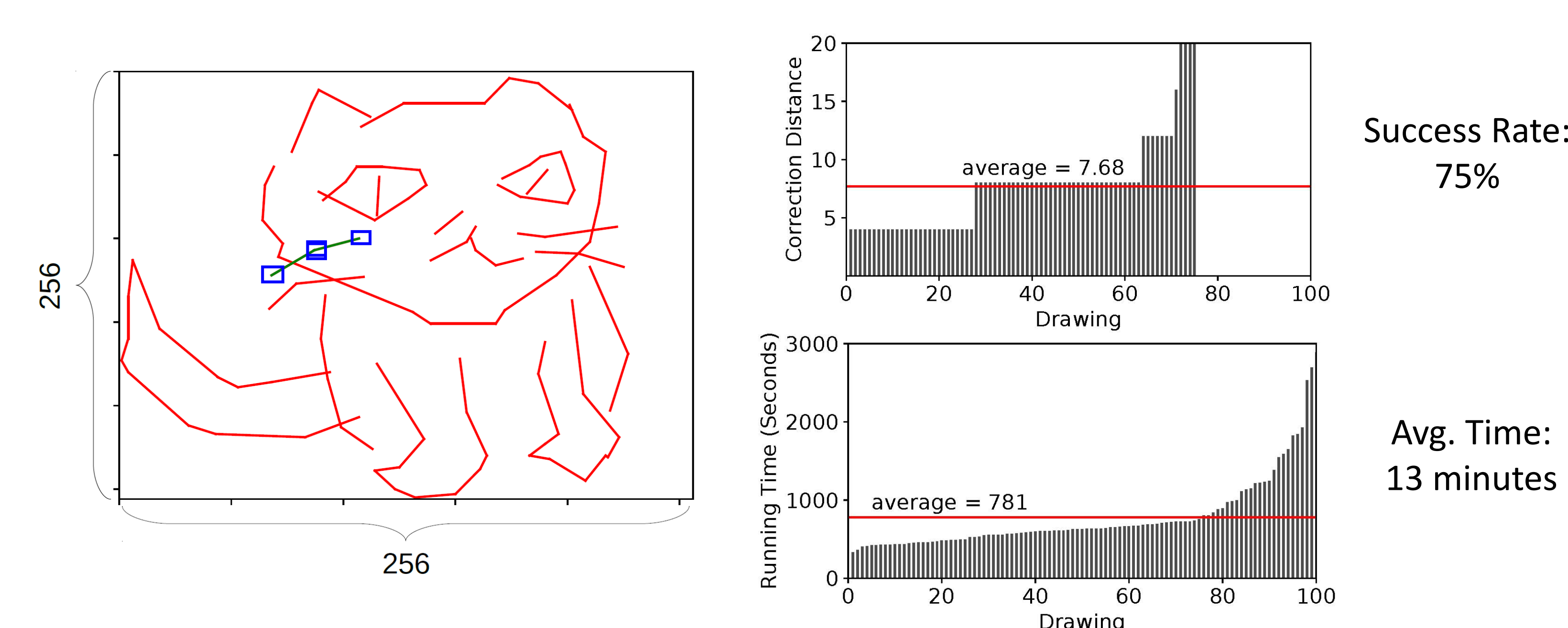


Success Rate: 81%

Avg. Time: 2 minutes

Drawing Tutorial

Network structure: convolutional network with 4096 neurons
Dataset: 0.12 million variants of a canonical cat drawing and 0.12 million other cat drawings



Success Rate: 75%

Avg. Time: 13 minutes