Robust Explainable AI: the Case of Counterfactual Explanations

27 October 2023

Francesco Leofante
f.leofante@imperial.ac.uk

About me

Imperial College Research Fellow

Centre for Explainable AI

Contacts:

- f.leofante@imperial.ac.uk
- https://fraleo.github.io/
- @fraleofante

SCAN ME

Agenda

- Explainable AI
- Counterfactual explanations and recourse
- Robustness
  - **what** does it mean?
  - **why** is it needed?
  - **how** can we achieve it?

Explainable AI (XAI)

XAI methods span a wide range of topics within AI and beyond, e.g.

- automated planning
- machine learning
- human computer interaction

Explainable AI (XAI)

XAI methods span a wide range of topics within AI and beyond, e.g.

- automated planning
- machine learning
- human computer interaction

Today we will focus on **explaining deep neural networks (DNNs)**

- **high-level** concepts rather than specific algorithms
- **fictional** use case and explanations

Supervised learning

Training set

- Age: 25
- Amount: £40K
- Duration: 36M

denied

- Age: 32
- Amount: £20K
- Duration: 24M

accepted

- Age: 82
- Amount: £26K
- Duration: 34M

denied

- Age: 54
- Amount: £14K
- Duration: 24M

accepted

Supervised learning

Training set

- Age: 25
- Amount: £40K
- Duration: 36M

denied

- Age: 32
- Amount: £20K
- Duration: 24M

accepted

- Age: 82
- Amount: £26K
- Duration: 34M

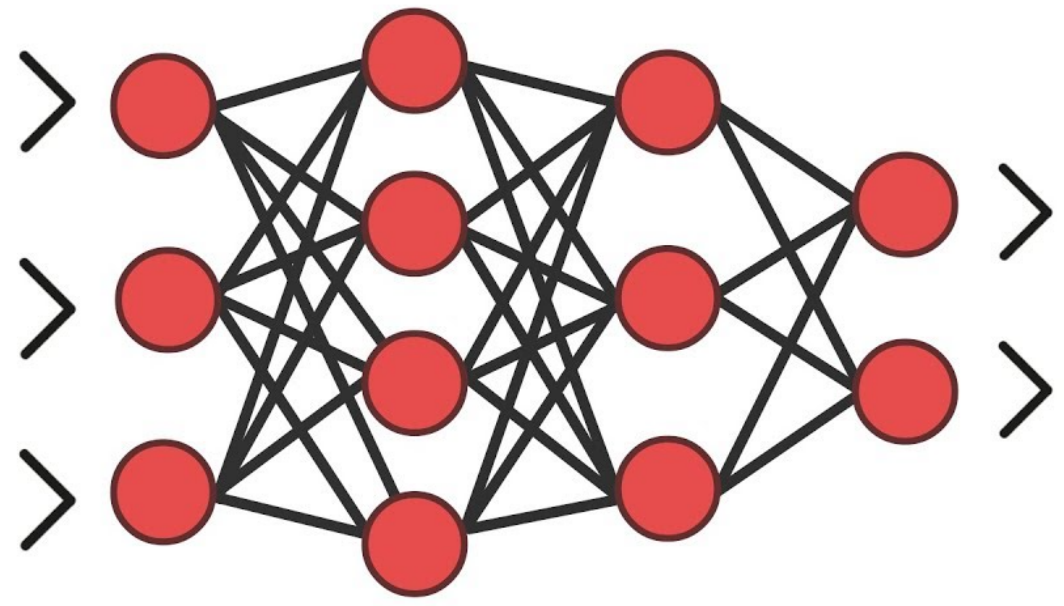
denied

- Age: 54
- Amount: £14K
- Duration: 24M

accepted

Deep neural network

(using your favourite algorithm)



Supervised learning

Training set

- Age: 25
- Amount: £40K
- Duration: 36M

denied

- Age: 32
- Amount: £20K
- Duration: 24M

accepted

- Age: 82
- Amount: £26K
- Duration: 34M

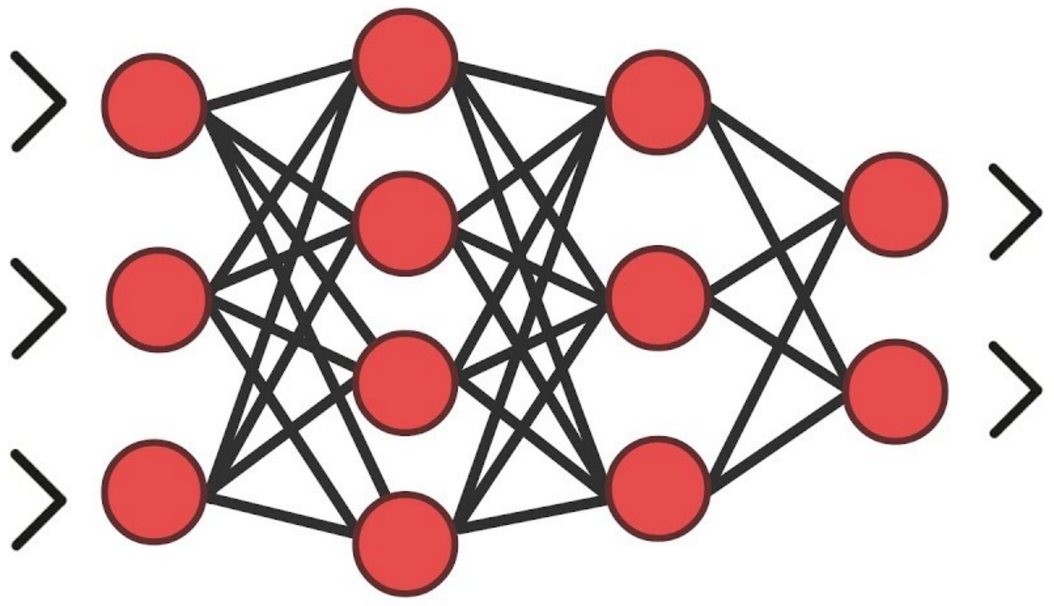
denied

- Age: 54
- Amount: £14K
- Duration: 24M

accepted

Deep neural network

(using your favourite algorithm)



New instance

Predicted class: **denied**

Supervised learning

Training set

- Age: 25
- Amount: £40K
- Duration: 36M

denied

- Age: 32
- Amount: £20K
- Duration: 24M

accepted

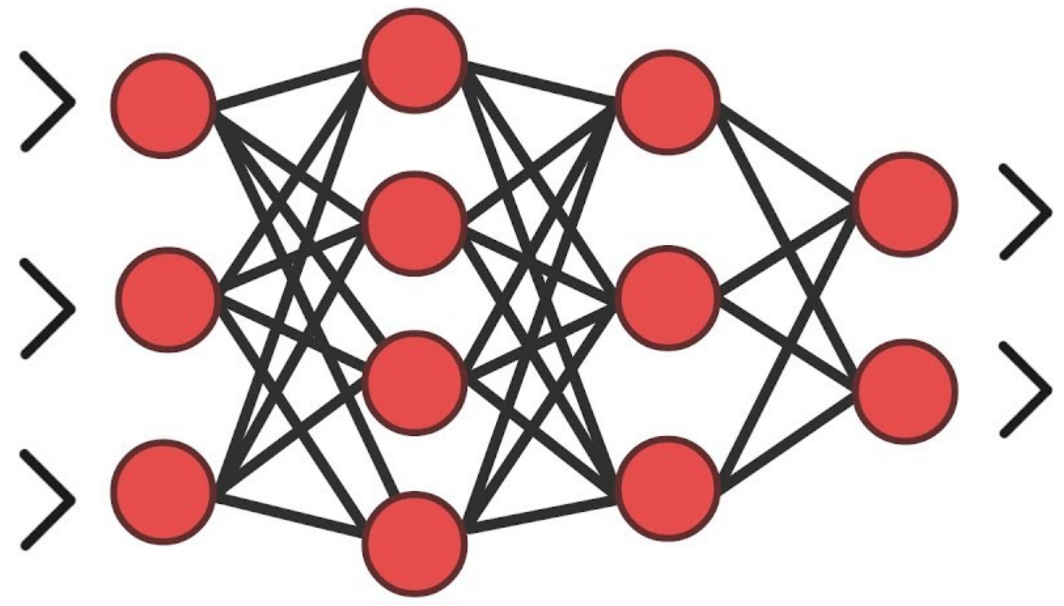
- Age: 82
- Amount: £26K
- Duration: 34M

denied

- Age: 54
- Amount: £14K
- Duration: 24M

accepted

Focus: explaining model predictions



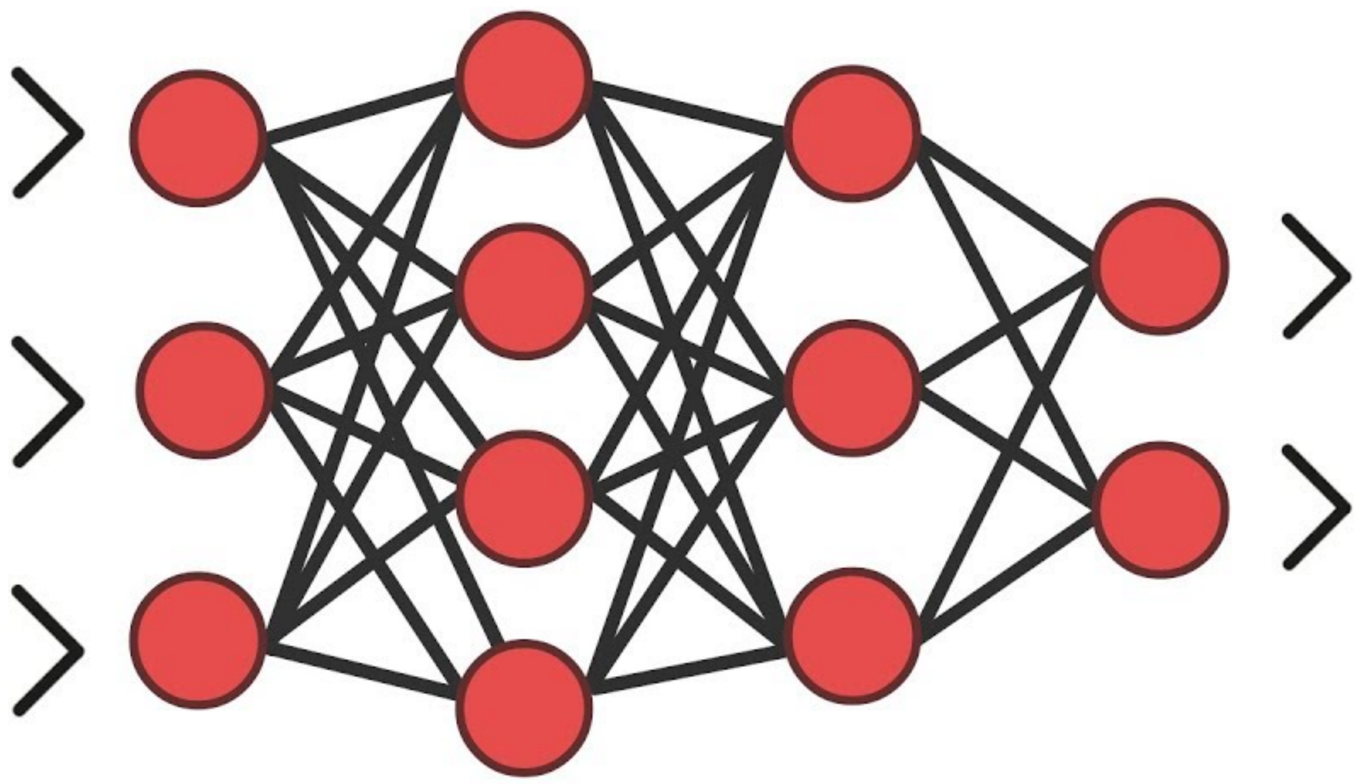
Predicted class: **denied**

New instance

- Why is it denied?
- Why not accepted?
- How do I get accepted?
- And many more questions…

Challenge

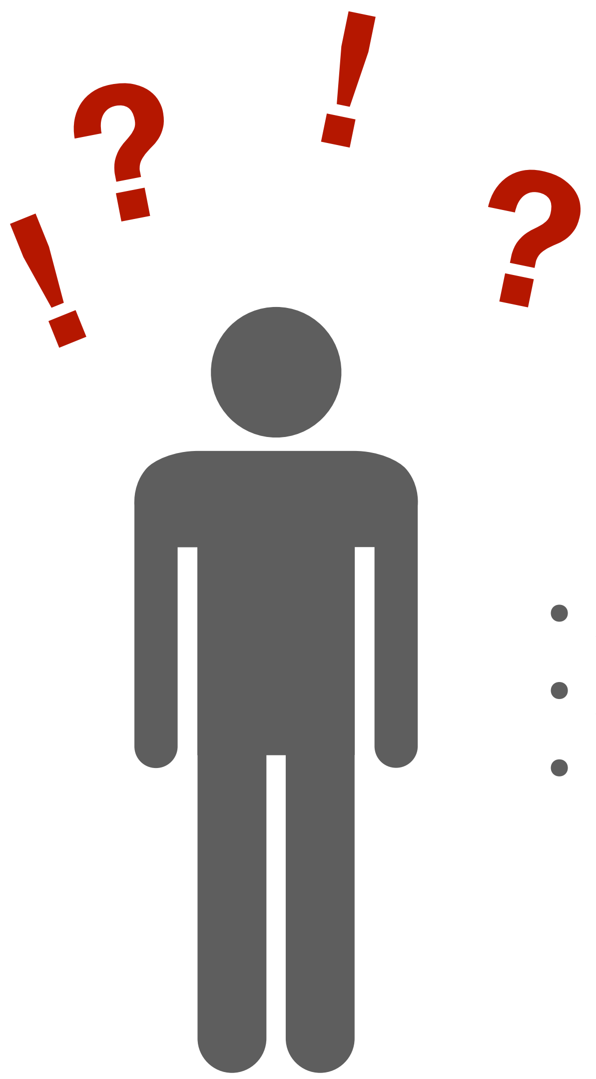
- Age: 30
- Amount: £15K
- Duration: 24M



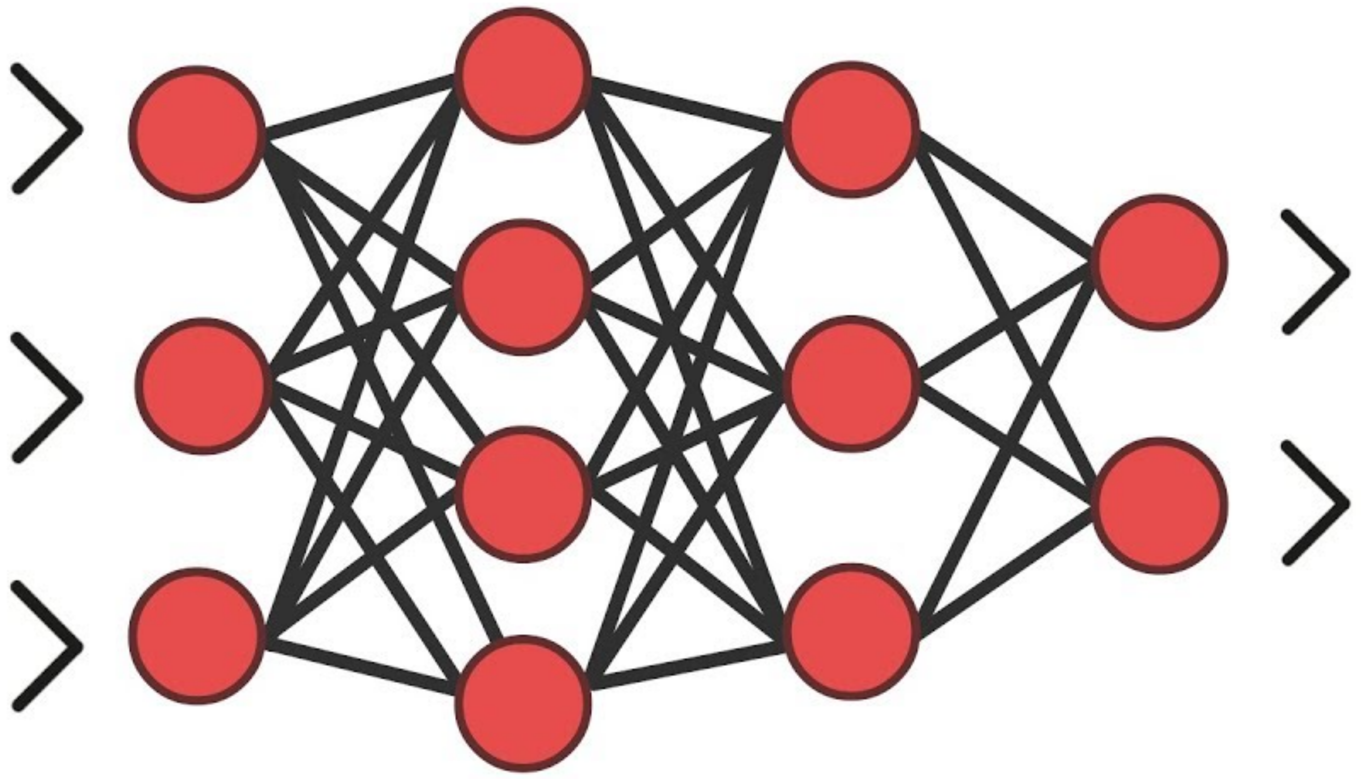
Loan denied

DNNs are black boxes!

Challenge



- Age: 30
- Amount: £15K
- Duration: 24M



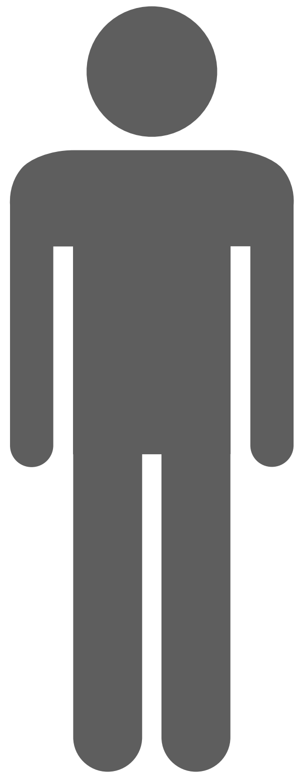
Loan denied

DNNs are black boxes!

Post-hoc explainability: **counterfactual explanations**

Counterfactual explanations (CXs)

Original instance

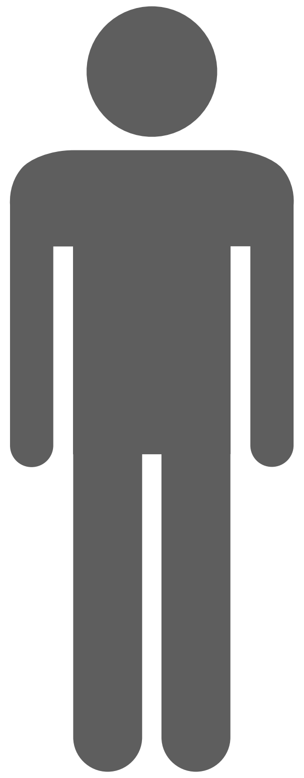


- Age: 30
- Amount: £15K
- Duration: 24M

Loan denied

Counterfactual explanations (CXs)

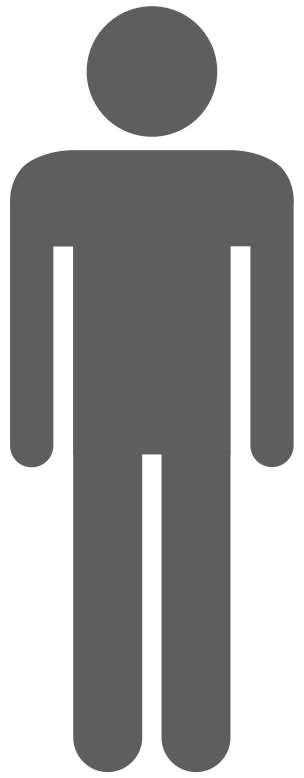
Original instance



- Age: 30
- Amount: £15K
- Duration: 24M

Loan denied

Counterfactual explanation



- Age: 30
- Amount: **£10K**
- Duration: 24M

The application would have been accepted **had you asked for £10K instead of £15K**

Computing a CX

- Given an input x_F and a binary classifier $\mathscr{M}$ such that $\mathscr{M}(x_F) = c$
- A distance function d

Computing a CX

- Given an input x_F and a binary classifier $\mathscr{M}$ such that $\mathscr{M}(x_F) = c$
- A distance function d

A **counterfactual explanation** x is computed as:

$$\arg\min_{x} d(x_F, x)$$

$$\text{subject to } \mathscr{M}(x) = 1 - c$$

Computing a CX

Most approaches solve relaxation defined as:

$$\arg\min_{x} \ \ell(\mathcal{M}(x), 1 - c) + \lambda \cdot d(x_F, x)$$

Counterfactual explanations without opening the black box: automated decisions and the GDPR. Wachter et al, Harvard Journal of Law & Technology 2018.

Computing a CX

Most approaches solve relaxation defined as:

$$\arg\min_{x} \boxed{\ell(\mathcal{M}(x), 1 - c)} + \lambda \cdot d(x_F, x)$$

where:

- ℓ is a differentiable loss function which minimises the gap between current and desired prediction

Counterfactual explanations without opening the black box: automated decisions and the GDPR. Wachter et al, Harvard Journal of Law & Technology 2018.

Computing a CX

Most approaches solve relaxation defined as:

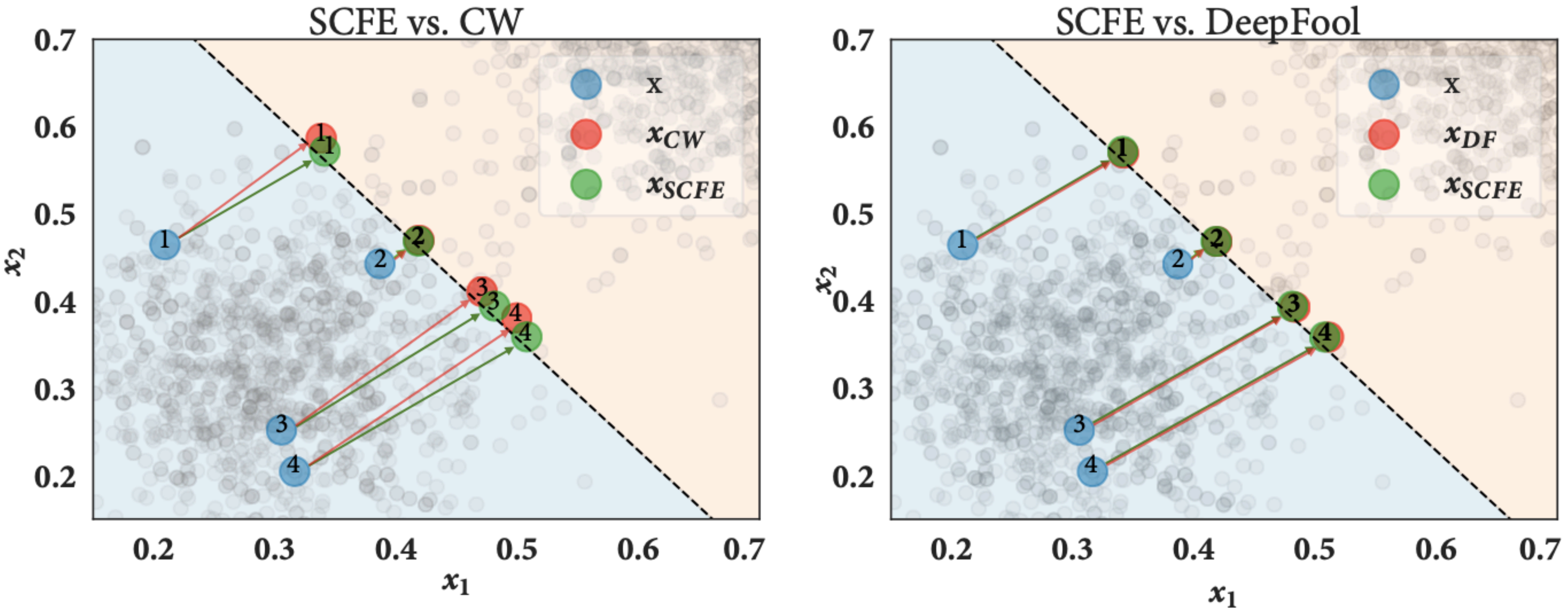
$$\arg\min_{x} \ell(\mathcal{M}(x), 1 - c) + \boxed{\lambda \cdot d(x_F, x)}$$

where:

- ℓ is a differentiable loss function which minimises the gap between current and desired prediction
- λ controls distance trade-off

Counterfactual explanations without opening the black box: automated decisions and the GDPR. Wachter et al, Harvard Journal of Law & Technology 2018.

Is minimising distance always good?



CXs are often **indistinguishable** from **adversarial examples**!

Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis. Pawelczyk et al, AISTATS 2022.

Brittle explanations ahead!

Threats

1. Model perturbations
2. Model multiplicity
3. Noisy execution

Robust XAI

Threats

1. Model perturbations
2. Model multiplicity
3. Noisy execution

Rethinking CX algos to mitigate these risks.

Brittle explanations ahead!

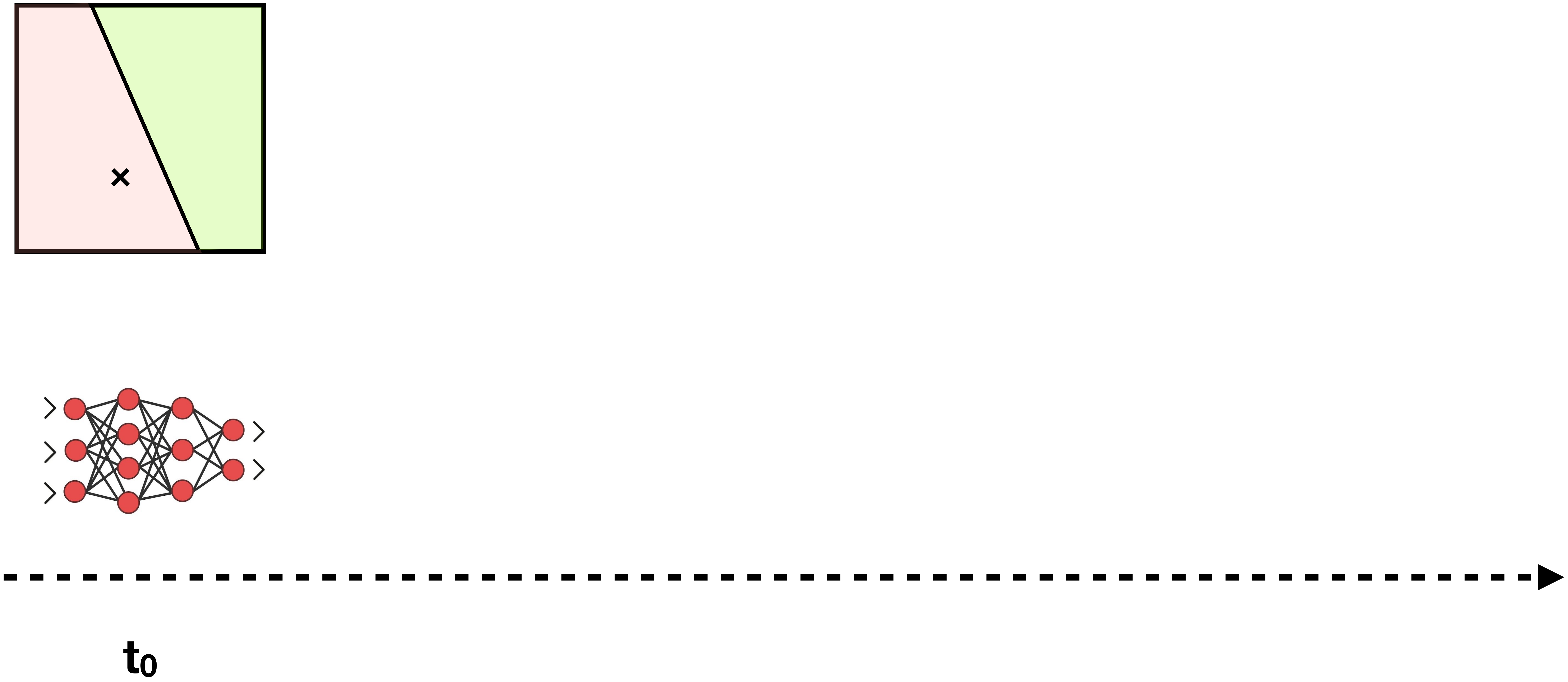
Threats

1. **Model perturbations**
2. Model multiplicity
3. Noisy execution

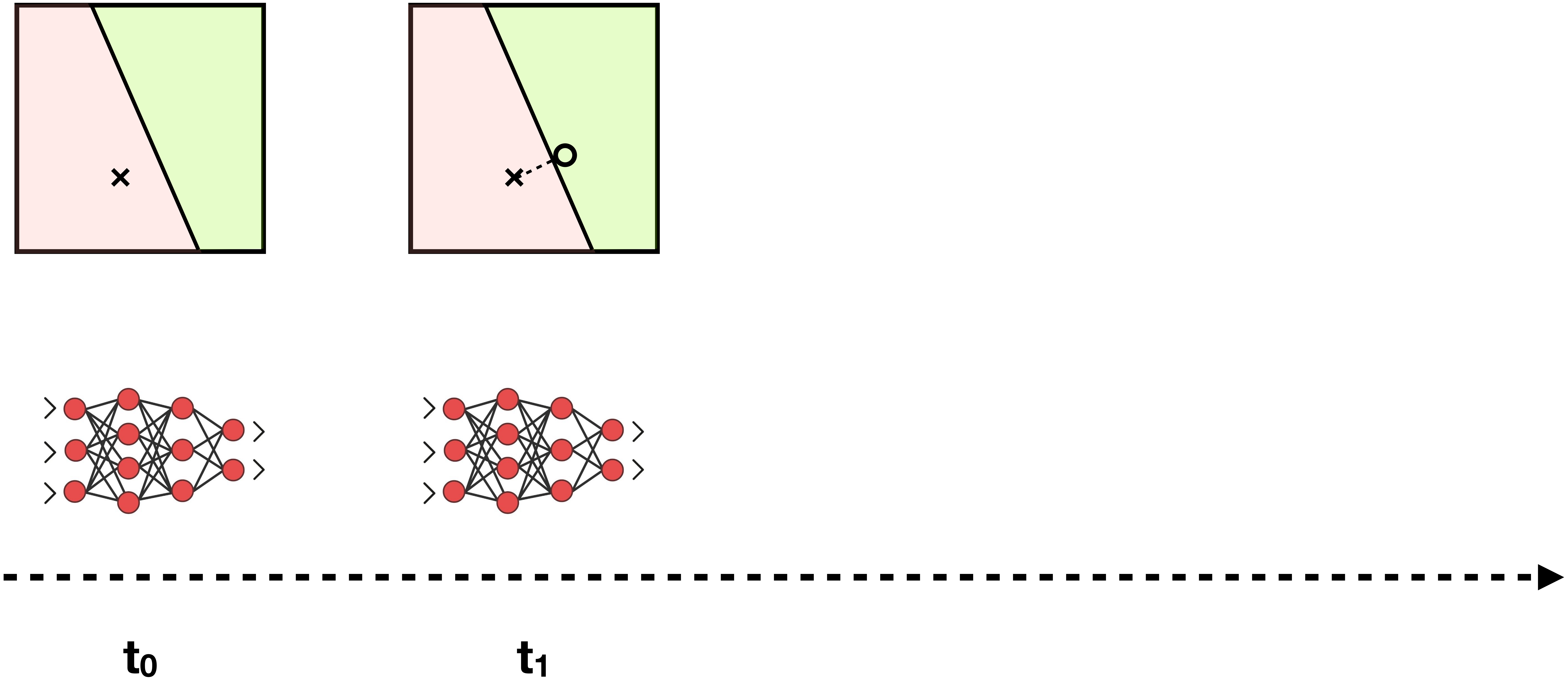
Model perturbations

t_0

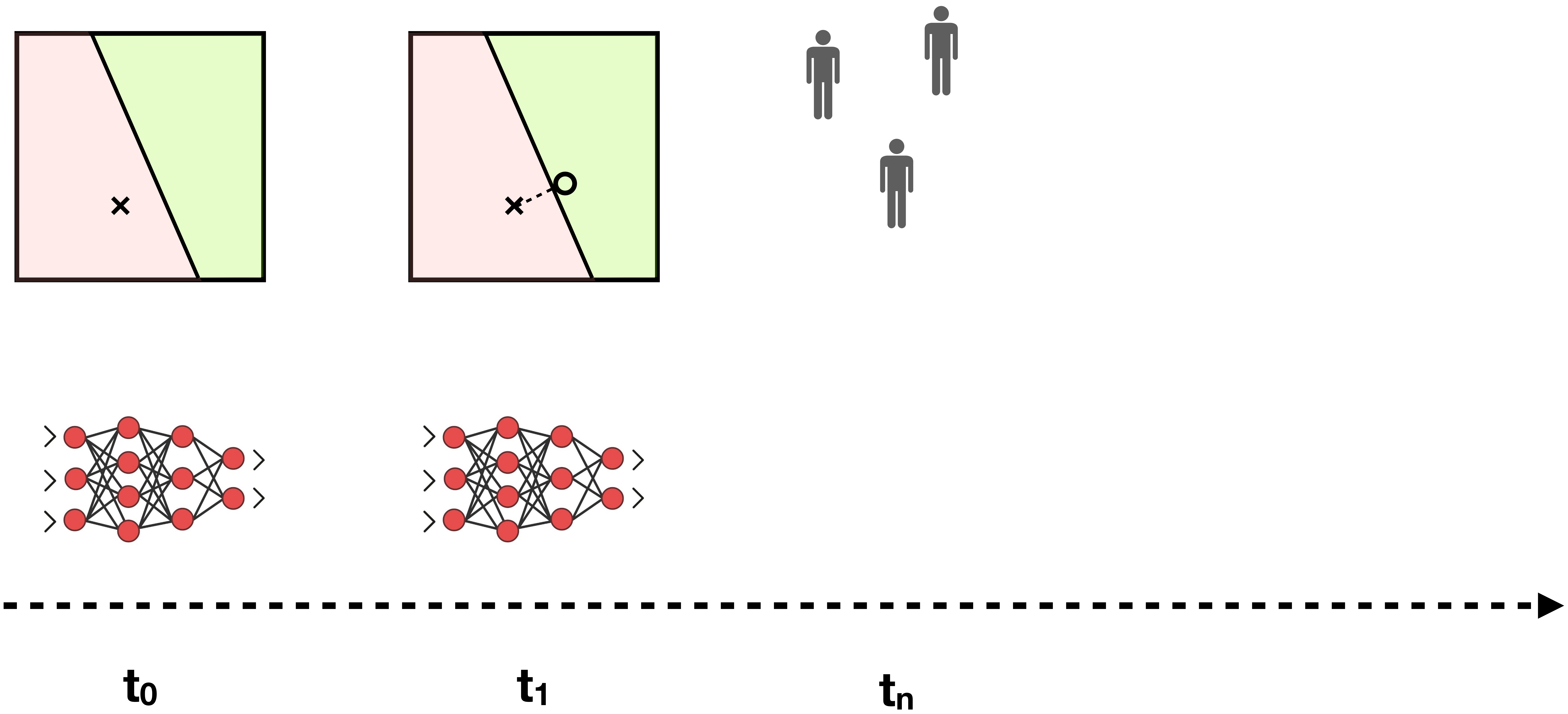
Model perturbations



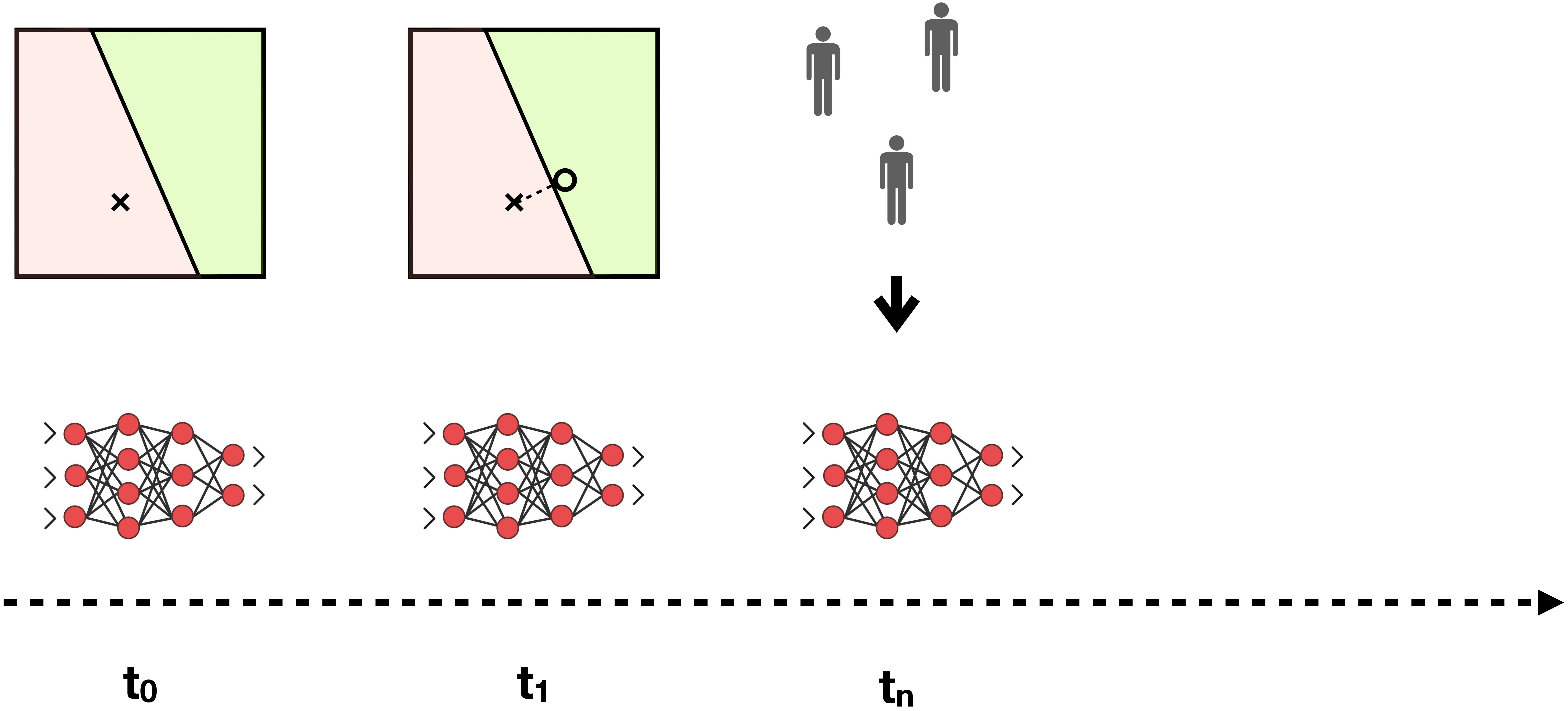
Model perturbations



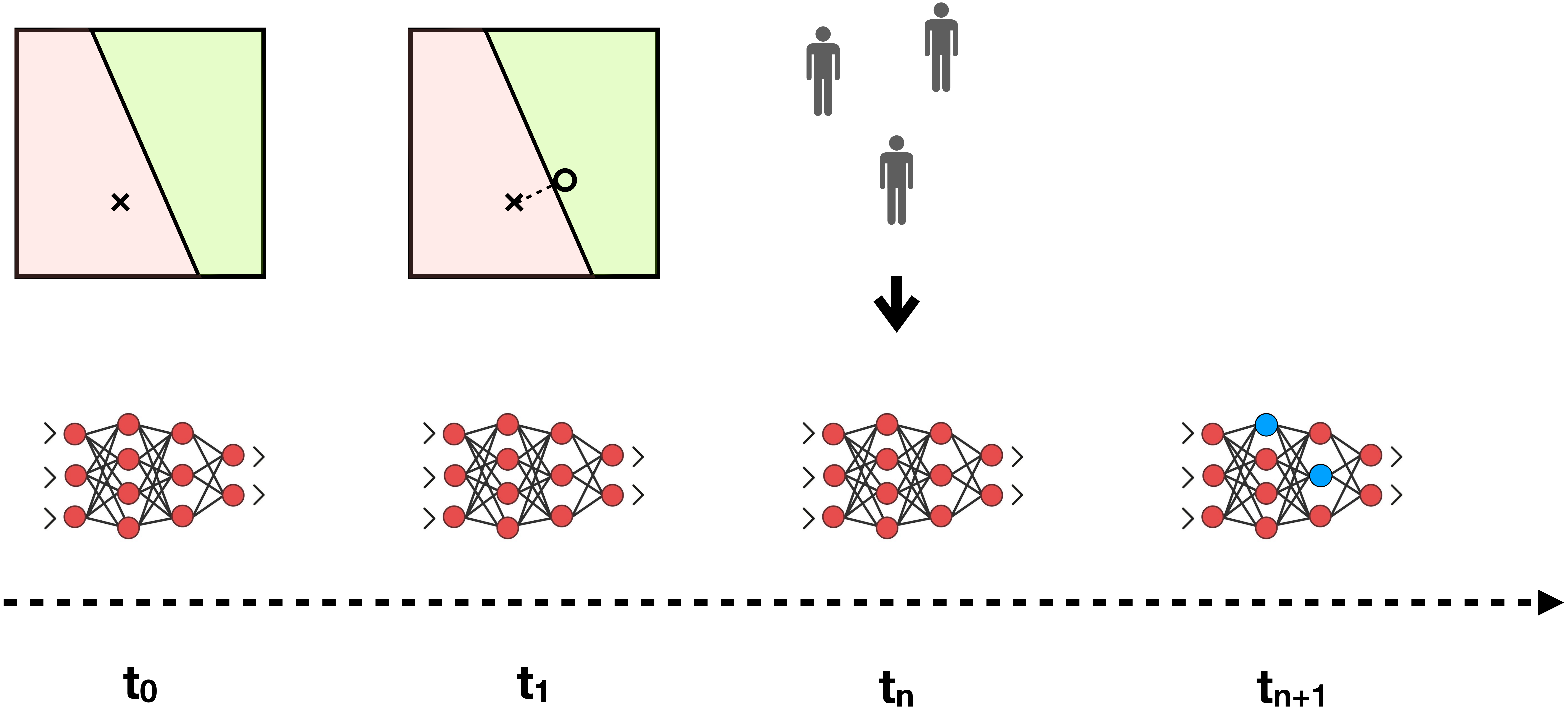
Model perturbations



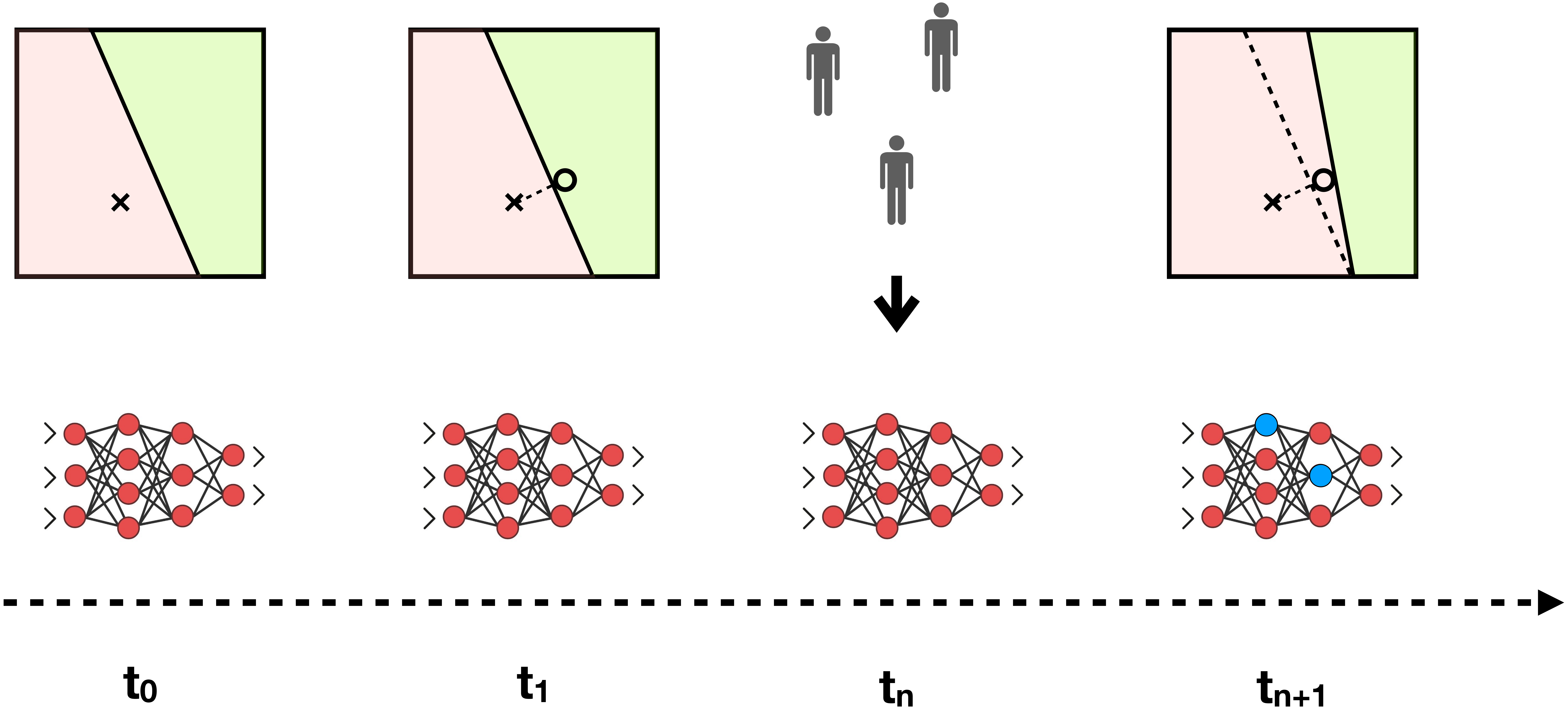
Model perturbations



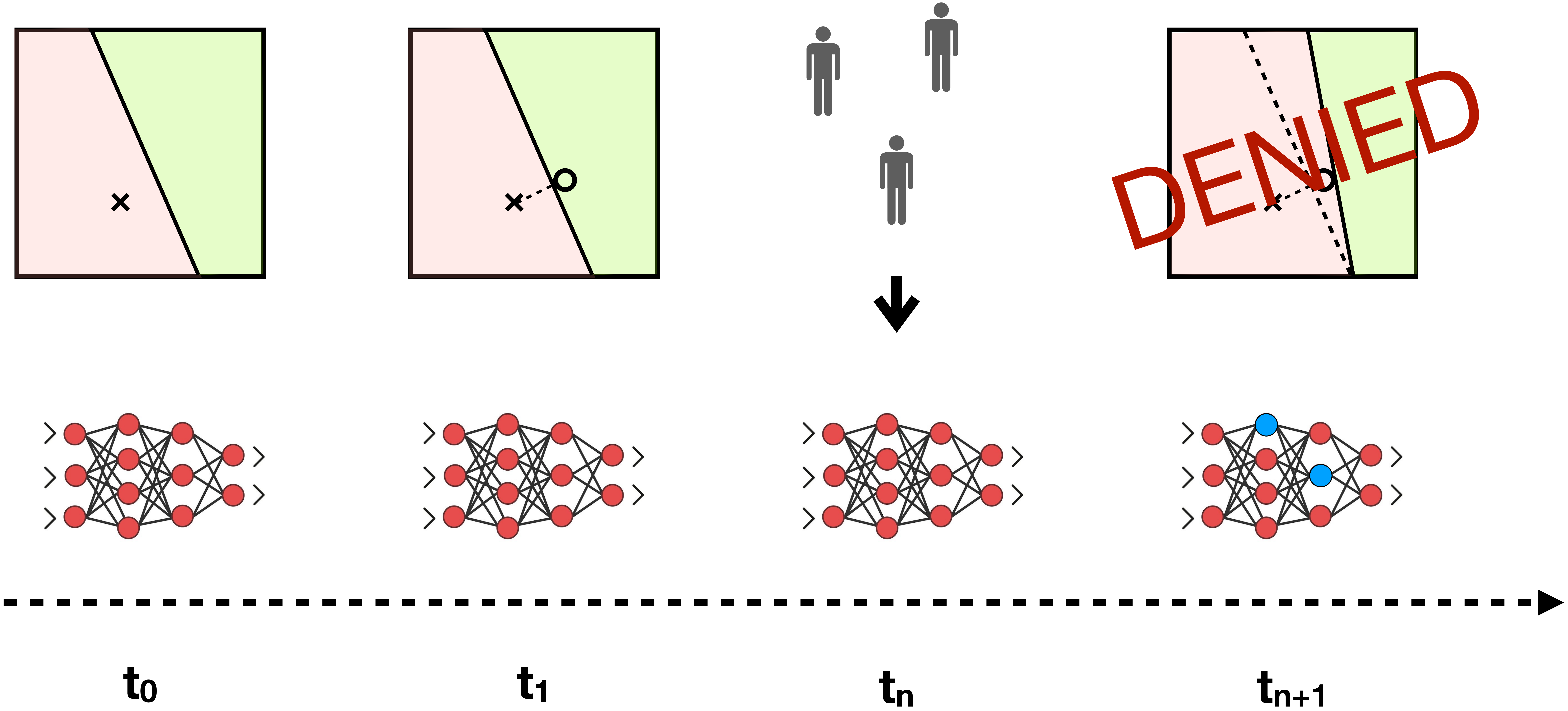
Model perturbations



Model perturbations



Model perturbations



Implications

Model shifts may occur as a result of data shifts

Implications

Model shifts may occur as a result of data shifts

Dilemma

Implications

Model shifts may occur as a result of data shifts

Dilemma

- **Trust** the old CX, although possibly contradicted by new data

accepted

Implications

Model shifts may occur as a result of data shifts

Dilemma

- **Trust** the old CX, although possibly contradicted by new data
- **Trash** the old CX, possibly upsetting end users

Our solution

We use interval abstractions to obtain formal robustness guarantees.

Formalising the Robustness of Counterfactual Explanations for Neural Networks. Jiang et al, AAAI 2023.

Our solution

We use interval abstractions to obtain formal robustness guarantees.

A **model shift** S is a function mapping an DNN into another one s.t.

- the two DNNs have same topology and,
- their differences (in parameter space) are bounded.

Our solution

We use interval abstractions to obtain formal robustness guarantees.

A **model shift** S is a function mapping an DNN into another one s.t.

- the two DNNs have same topology and,
- their differences (in parameter space) are bounded.

Define set of **plausible model shifts** as:

$$\Delta = \{S \mid ||\mathcal{M} - S(\mathcal{M})|| \leq \delta\}$$

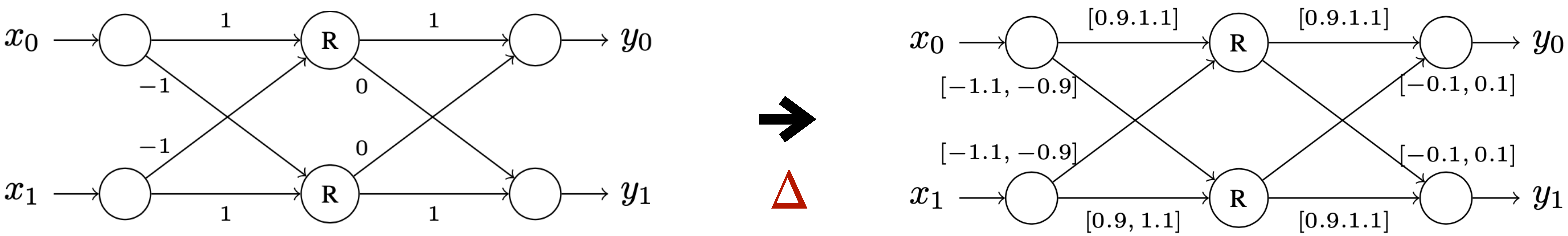
Our solution

- Plausible model shifts induce a family of DNNs…
- Need a way to reason about them concisely!

Our solution

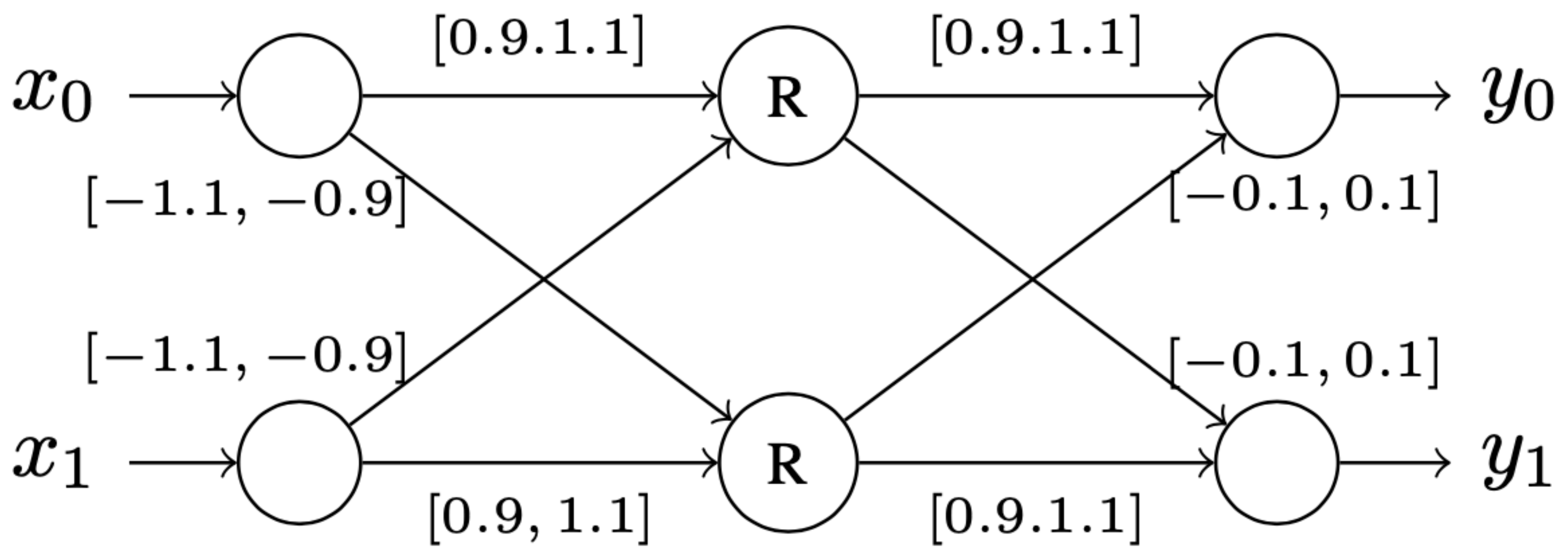
- Plausible model shifts induce a family of DNNs…
- Need a way to reason about them concisely!

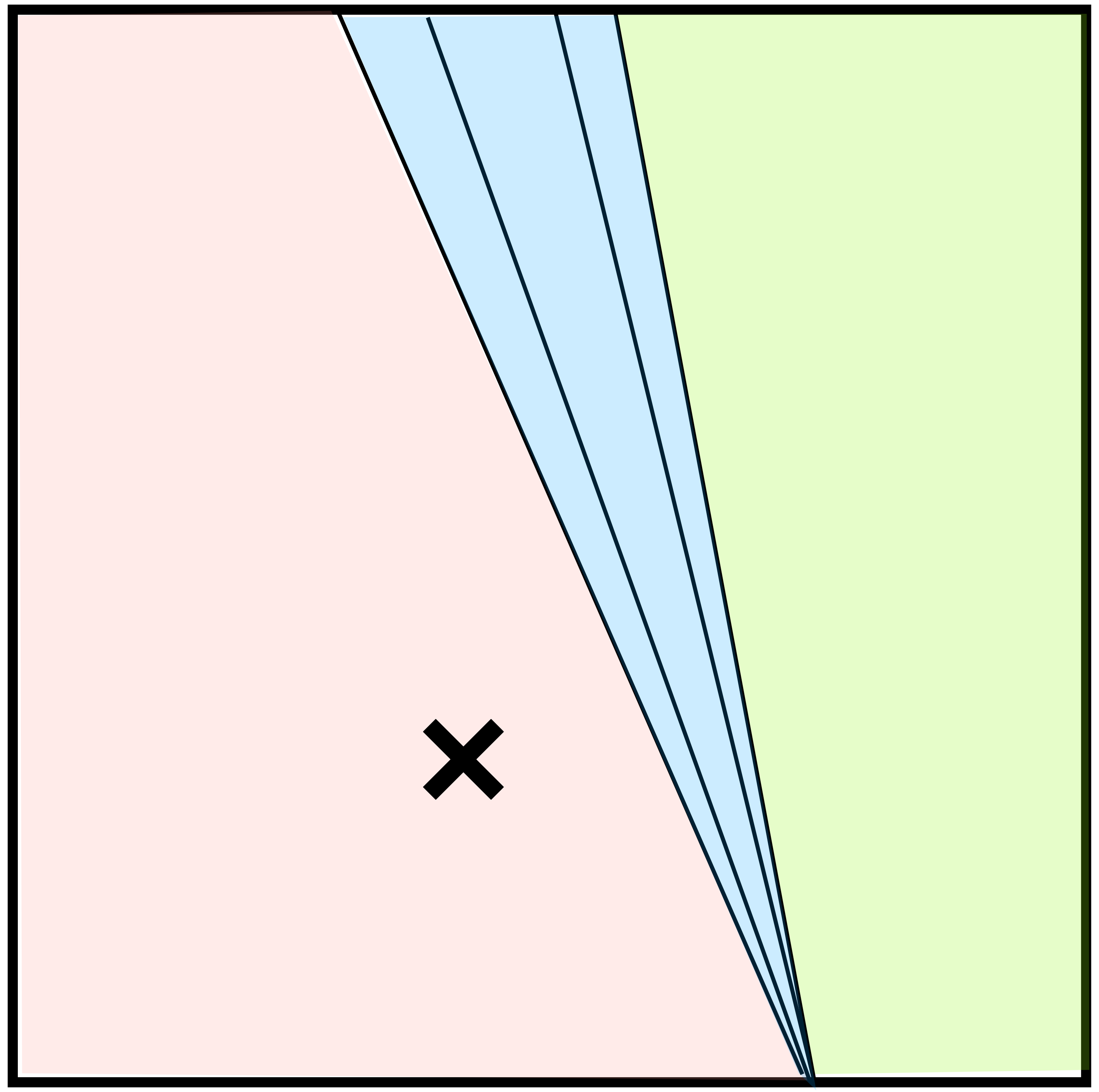
Enter the **interval neural network** $\mathscr{I}$

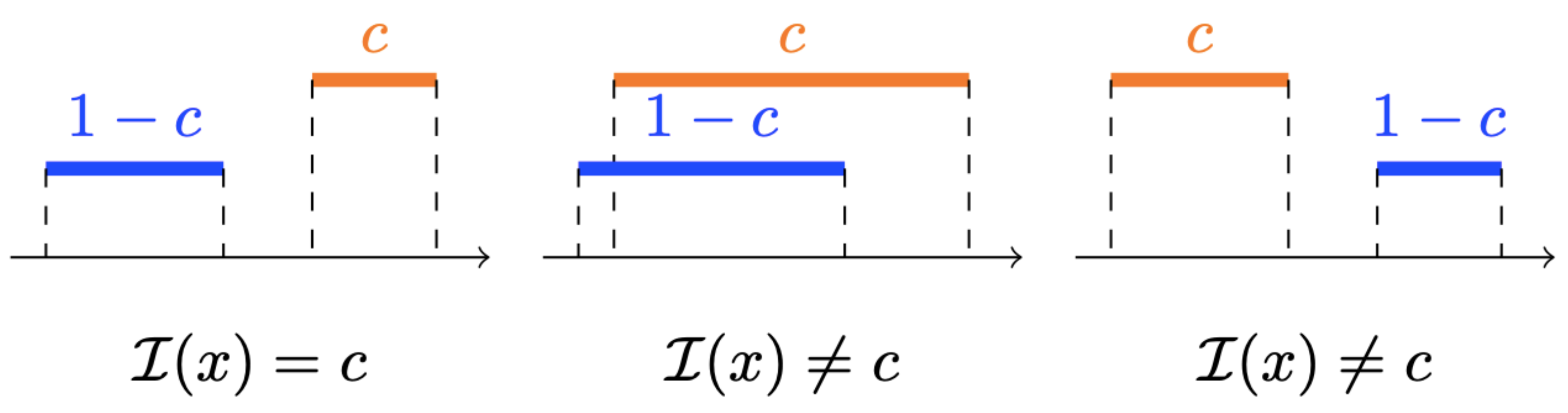


Abstraction based Output Range Analysis for Neural Networks, Prabhakar and Afzal, NeurIPS 2019.

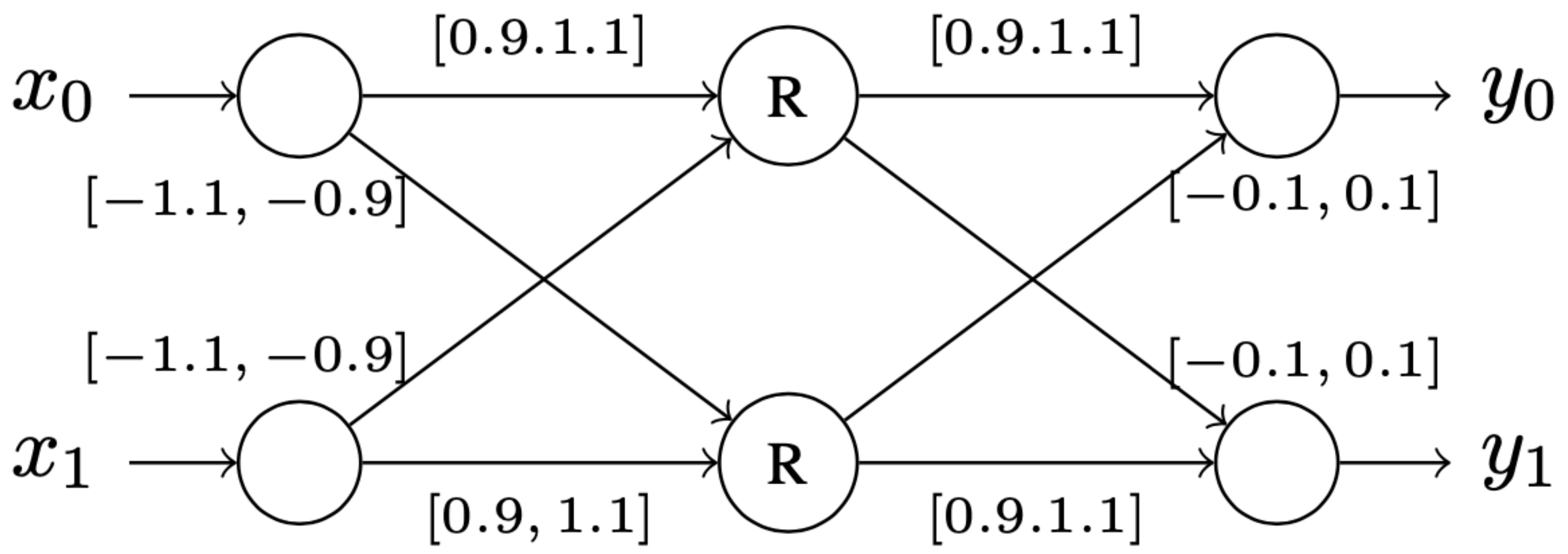
Our solution

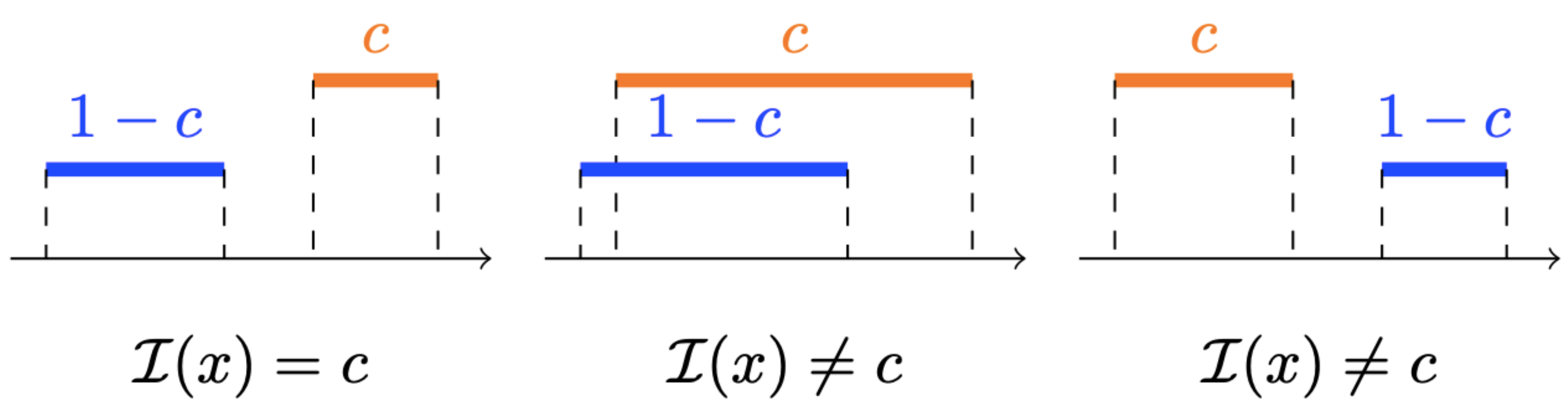


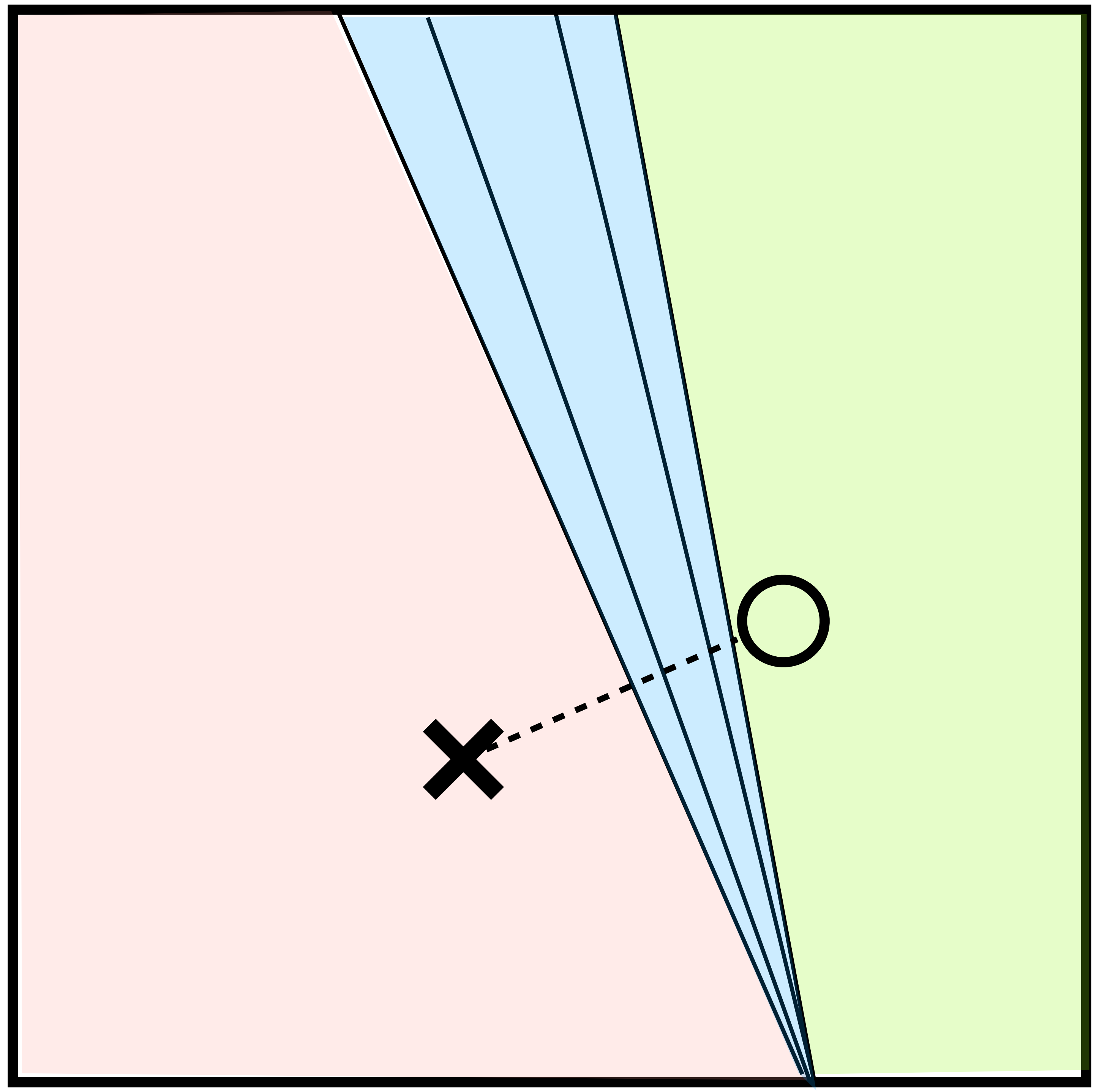




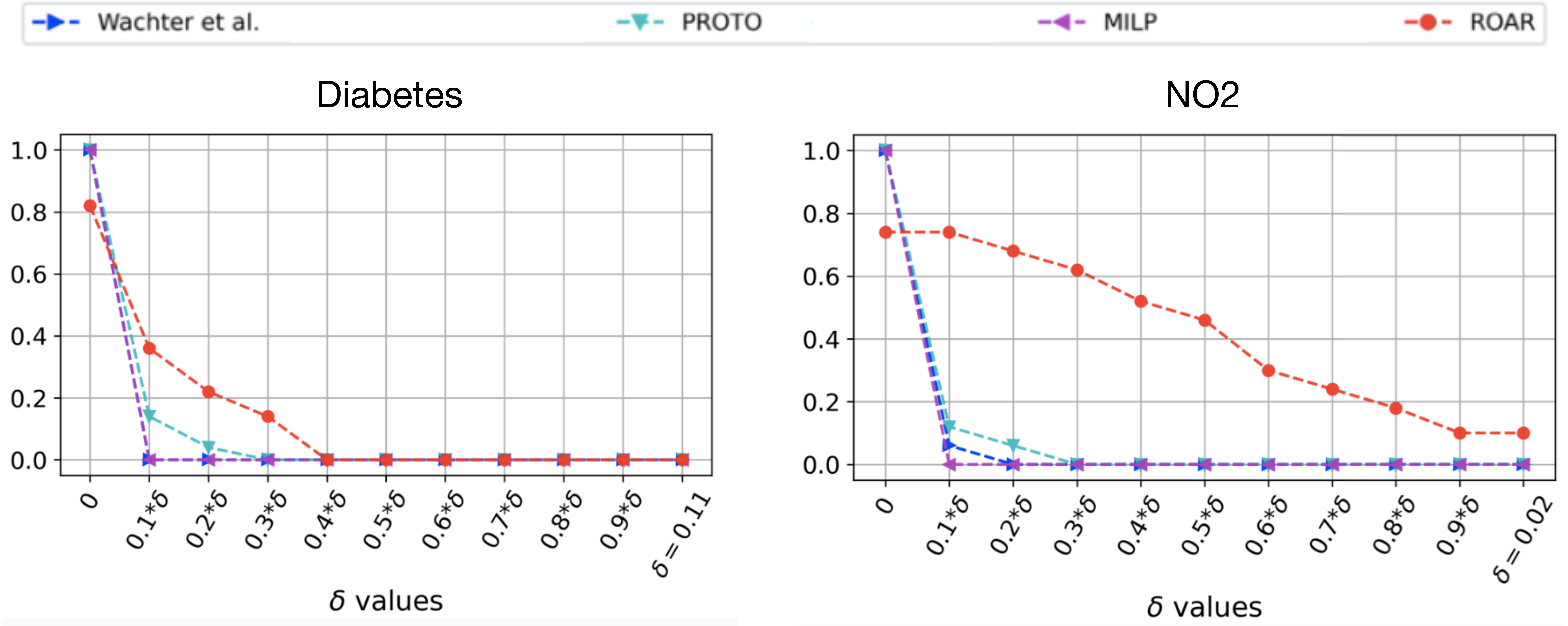
Our solution





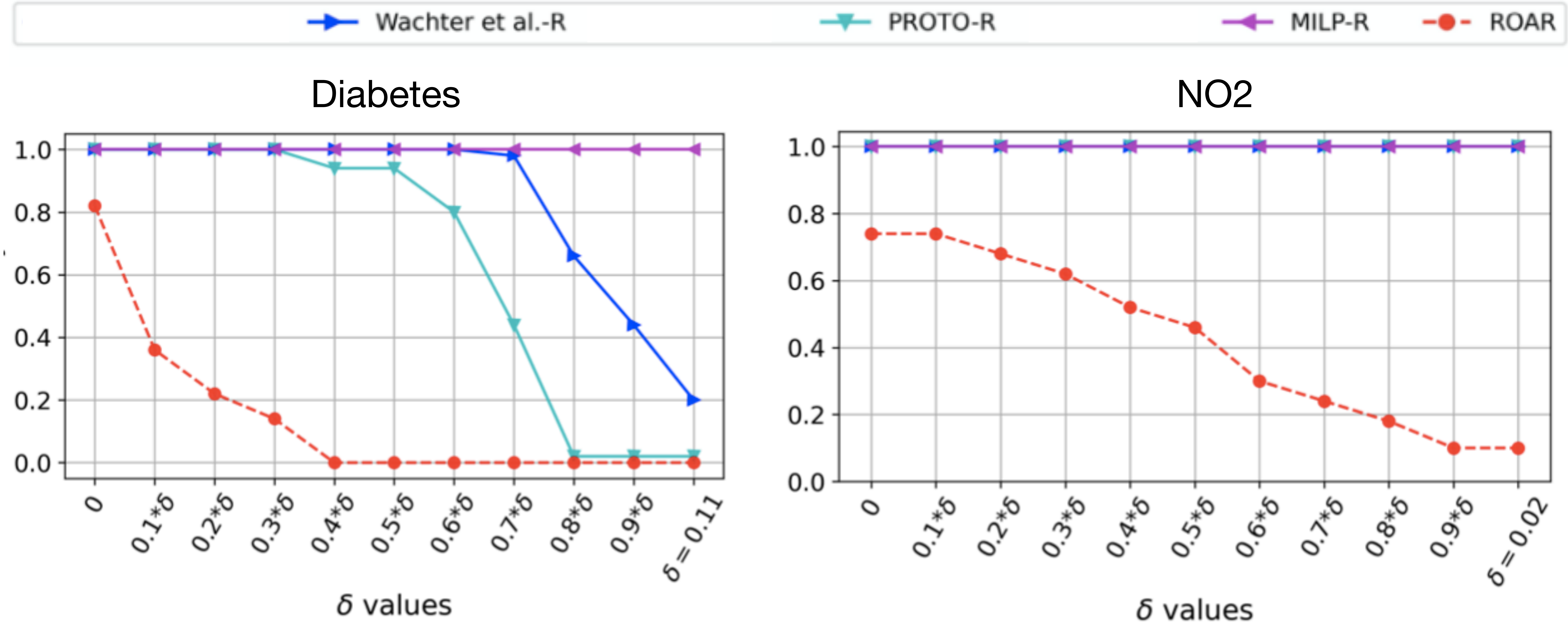


Our solution



Robustness decreases with shift magnitude - **for robust methods as well**!

Our solution



Robustness of base methods increased - **100% in some cases**.

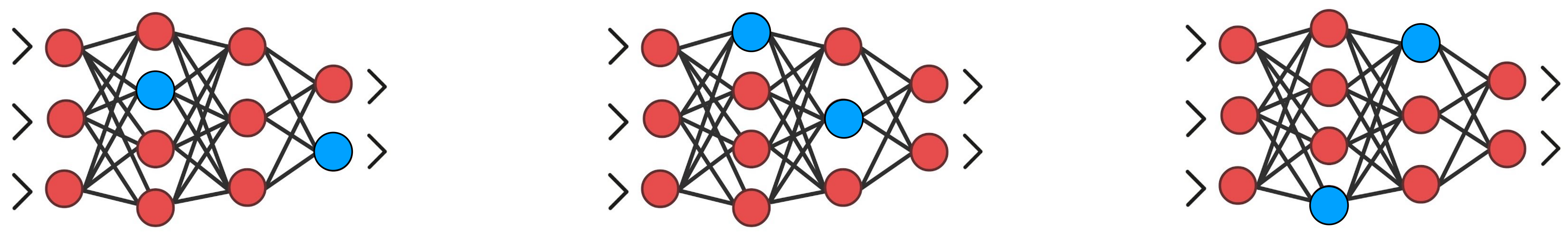
Brittle explanations ahead!

Threats

1. Model perturbations
2. **Model multiplicity**
3. Noisy execution

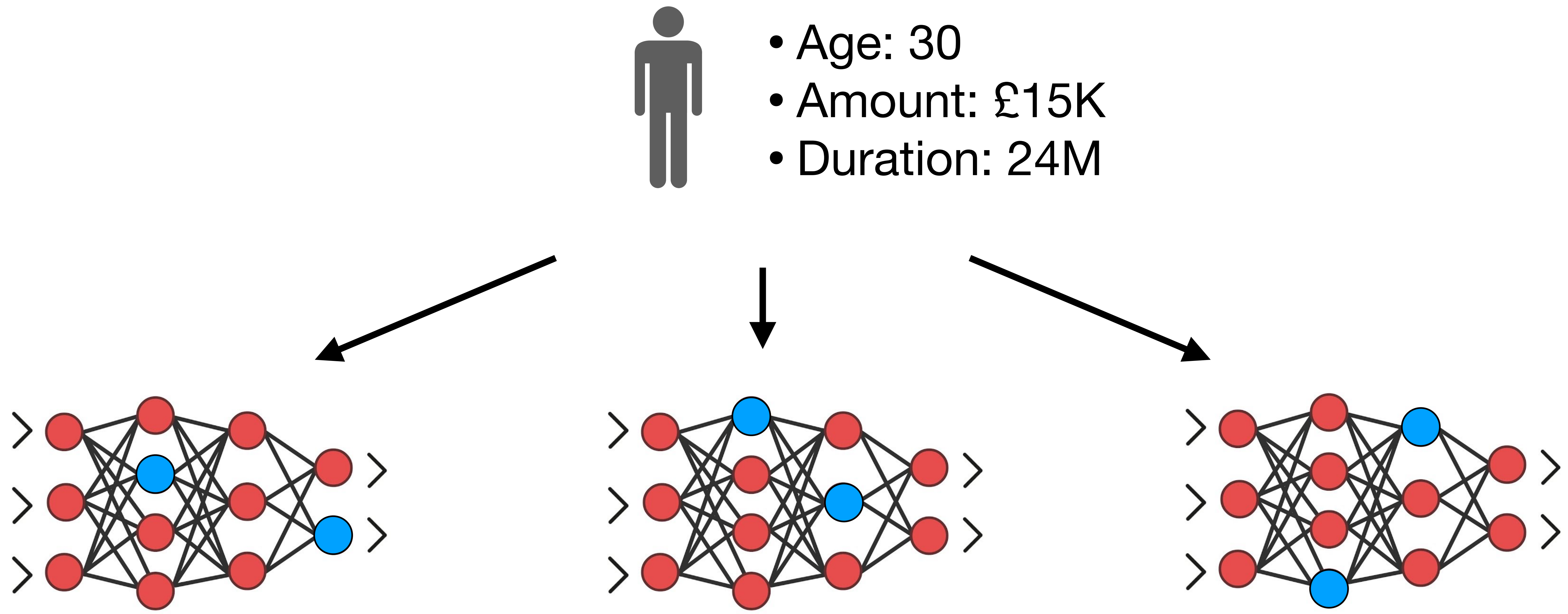
Model multiplicity

Situation where models of equal accuracy differ in the process by which they reach a given prediction

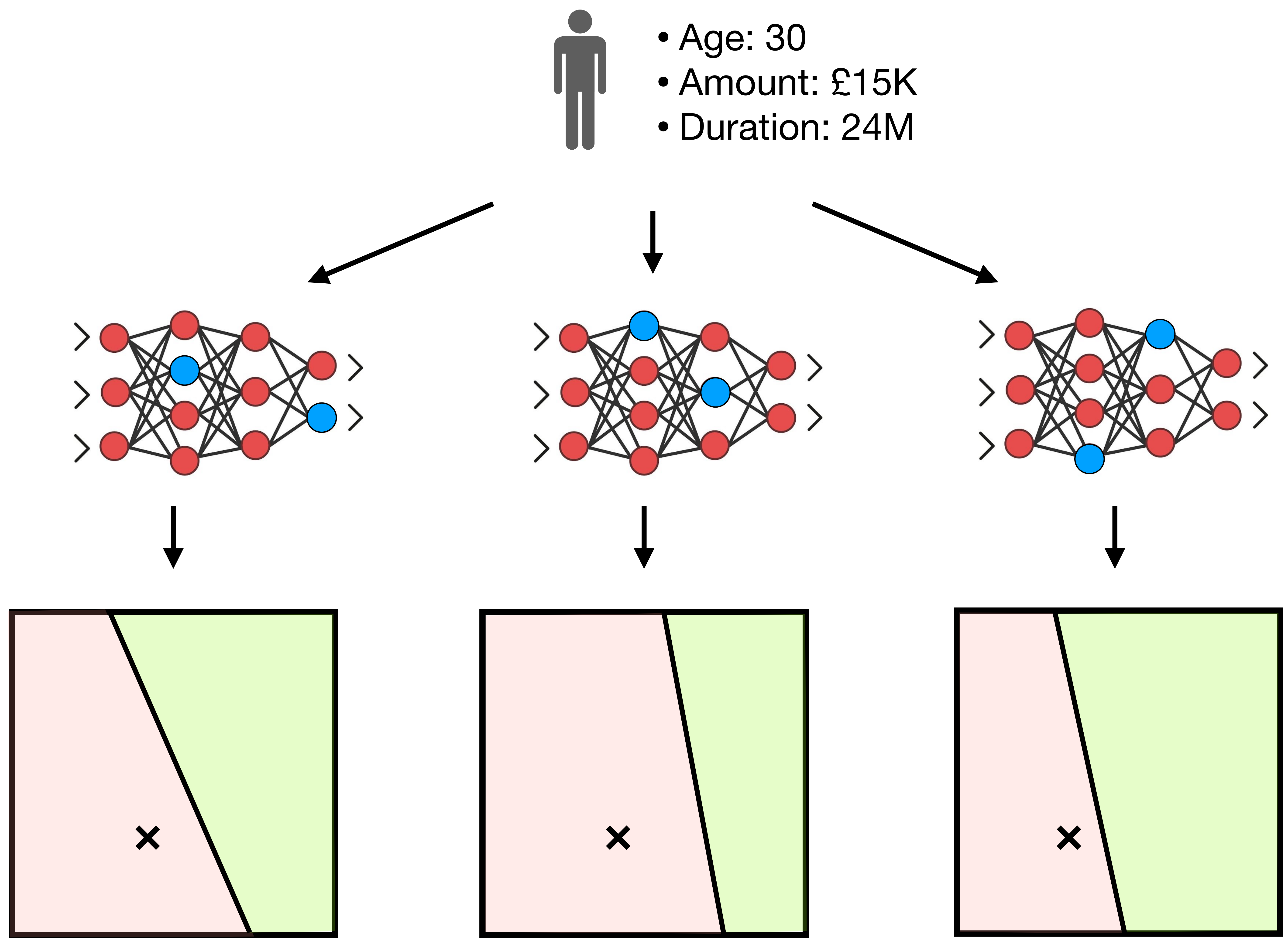


Model Multiplicity: Opportunities, Concerns, and Solutions. Black et al, ACM FAccT'22.

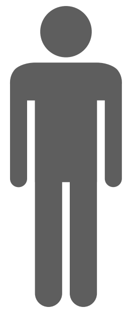
Model multiplicity



Model multiplicity

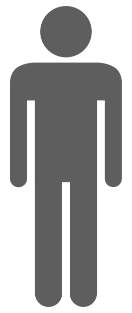


Model multiplicity

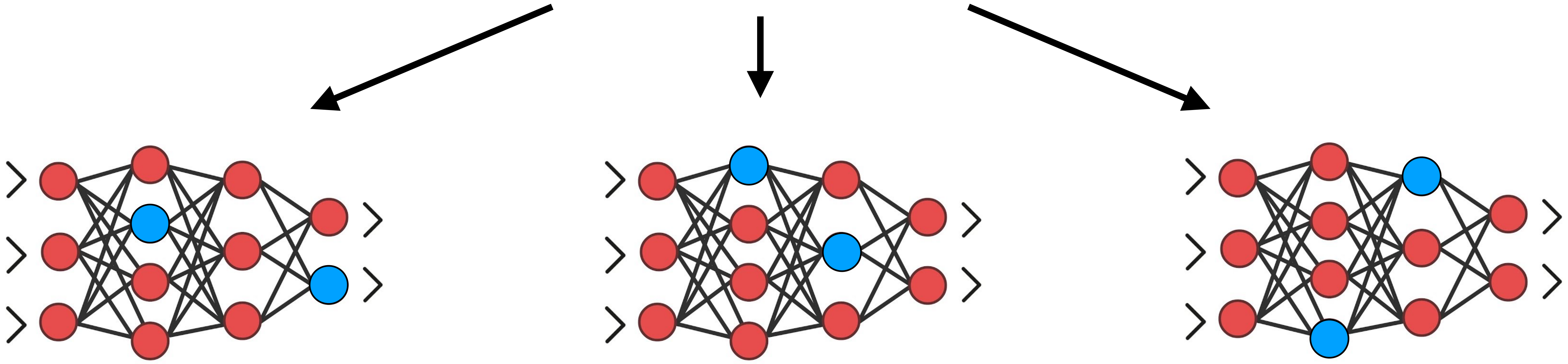


- Age: 30
- Amount: **£10K**
- Duration: 24M

Model multiplicity

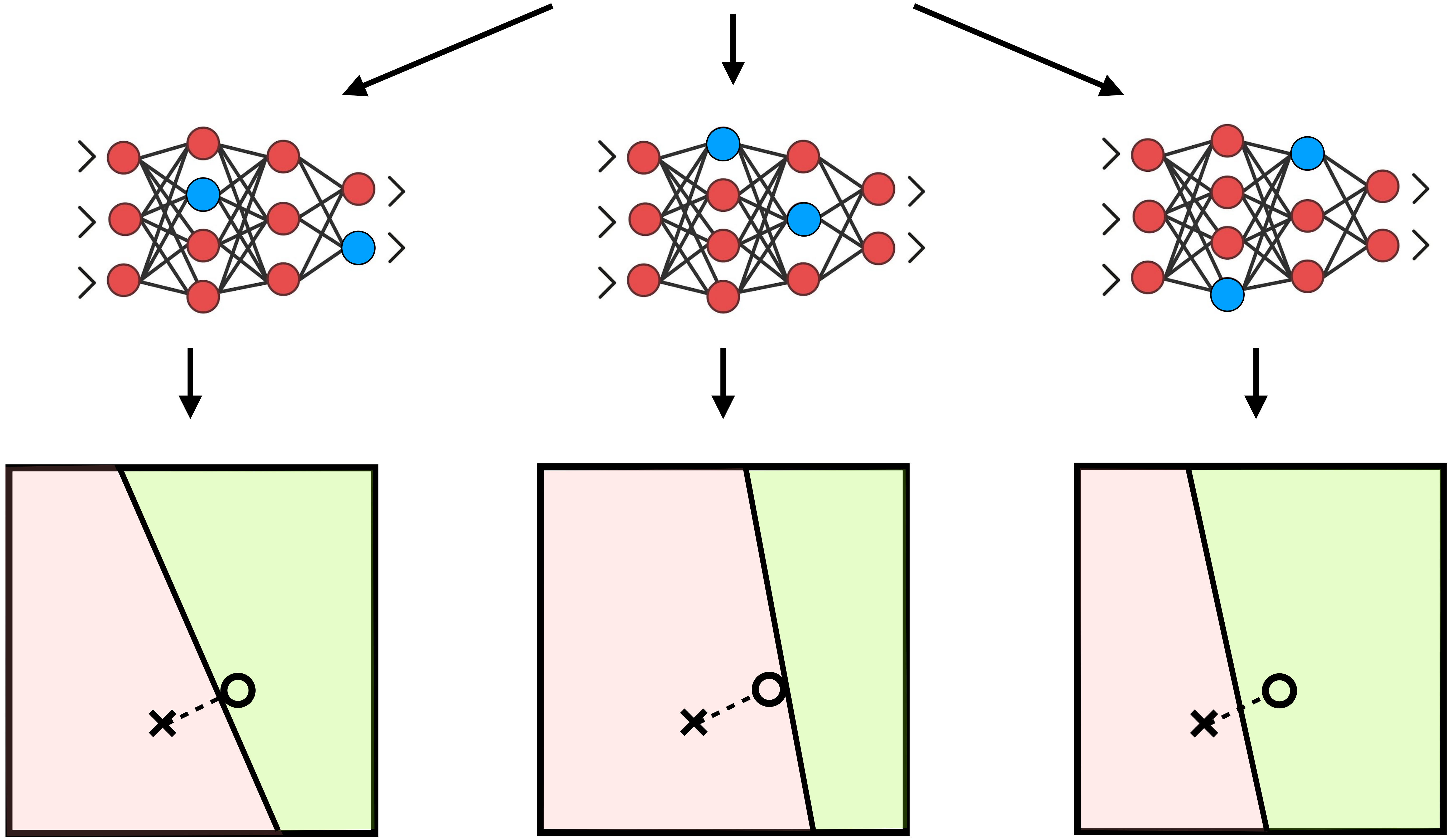


- Age: 30
- Amount: **£10K**
- Duration: 24M

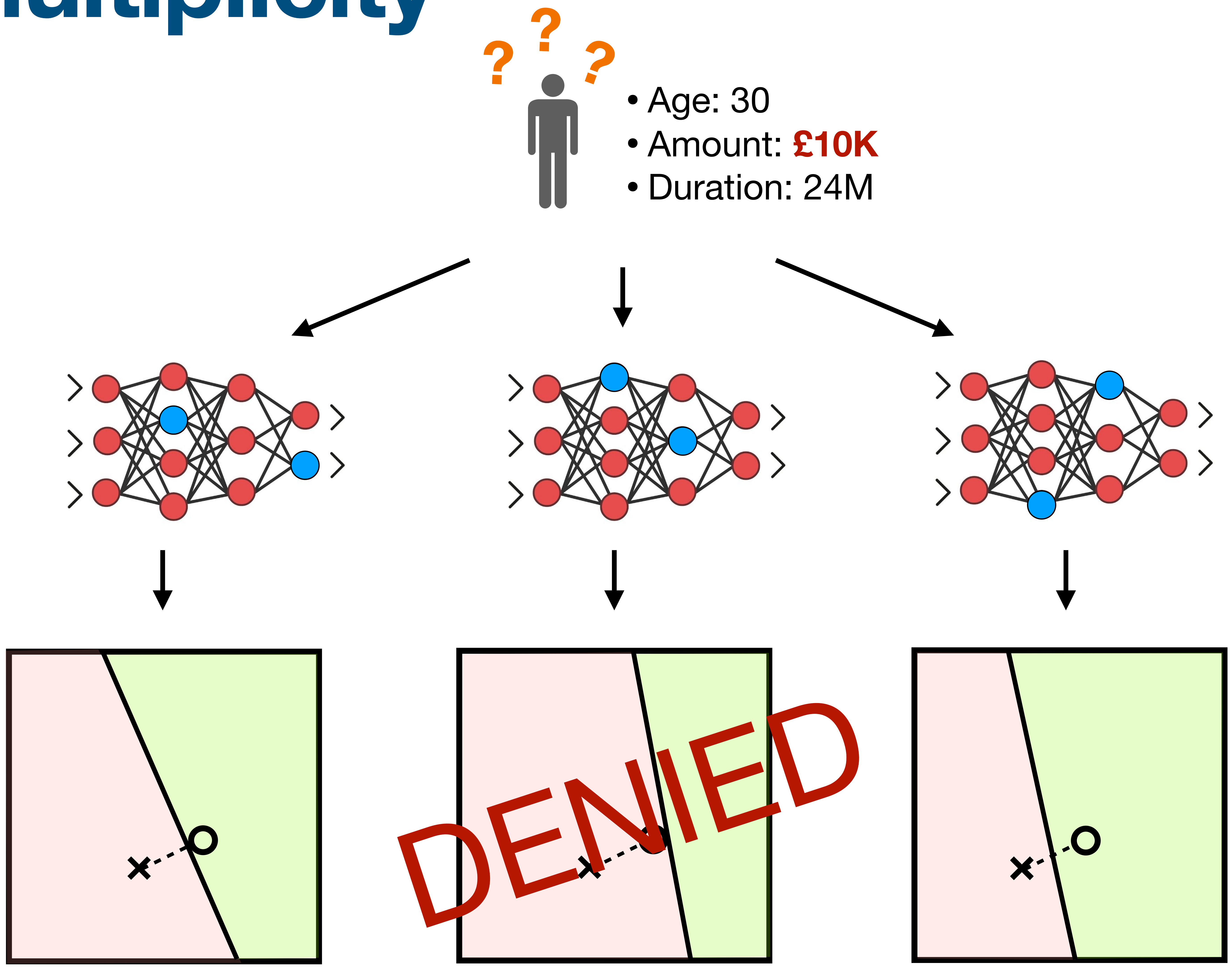


Model multiplicity

- Age: 30
- Amount: **£10K**
- Duration: 24M

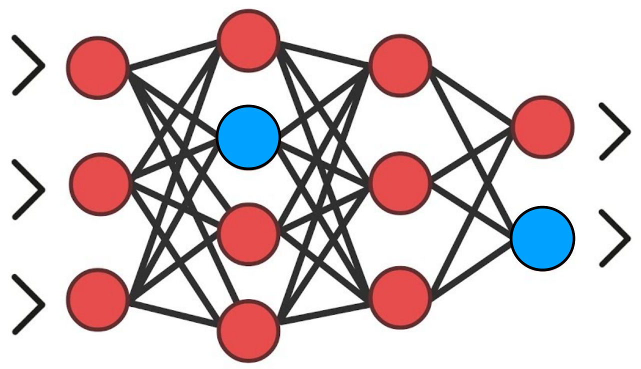


Model multiplicity

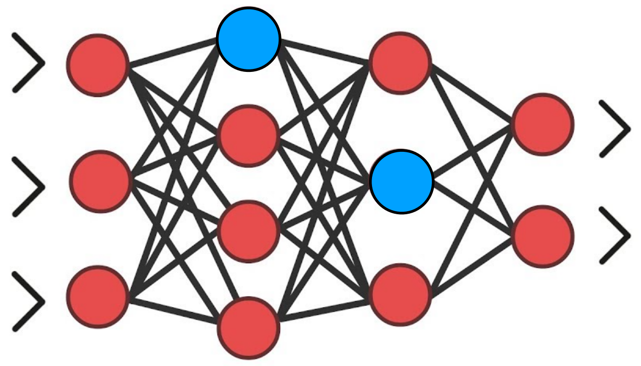


Implications

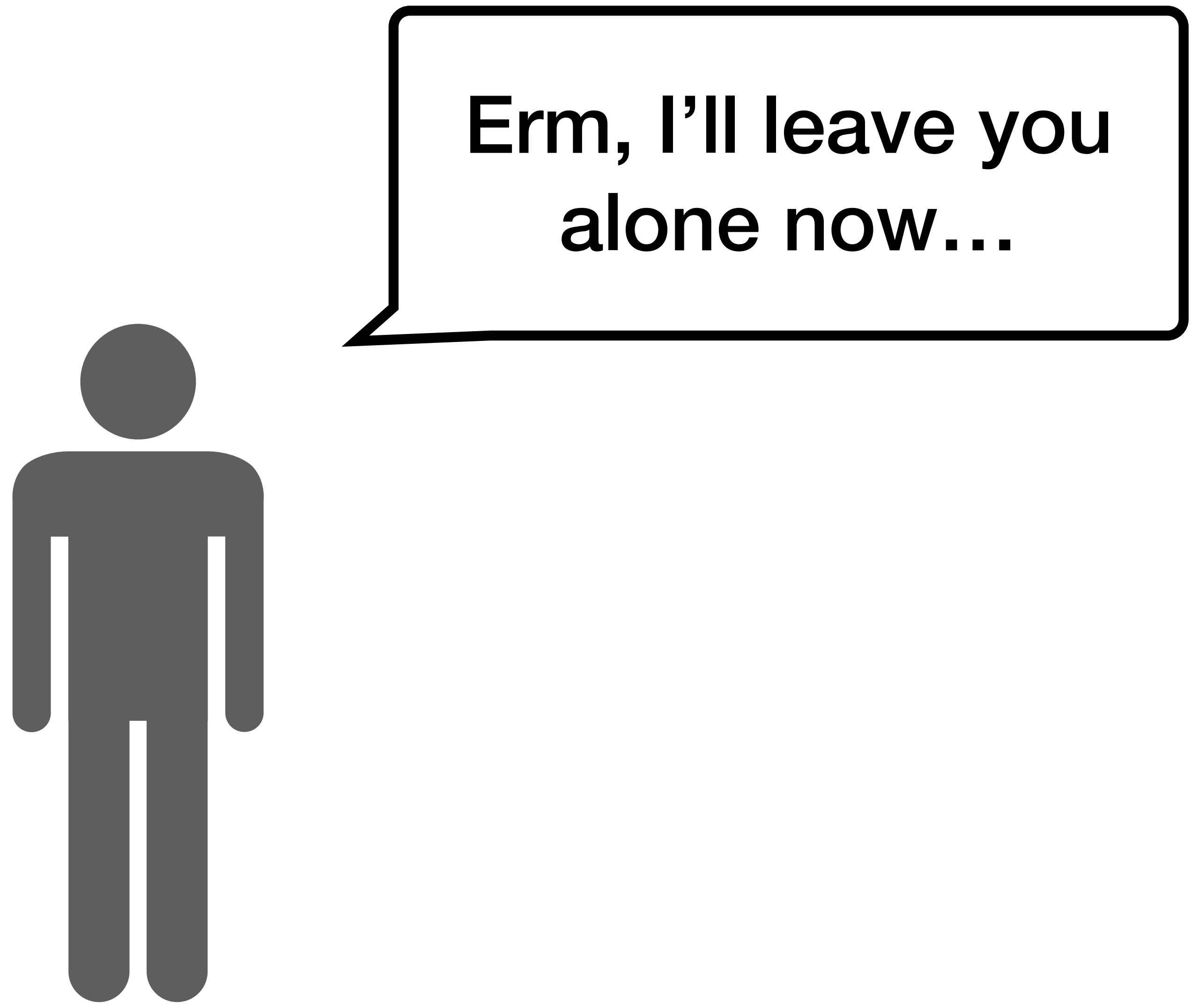
- Disagreeing models might raise concerns about the **justifiability** of CXs
- Different models might offer **better/worse recourse** options



Increase by £50



That’s not enough!



Our solution

We use tools from **relational verification**.

- Introduce a **novel product construction** tailored for the problem.
- Use this construction to **study the complexity** of generating robust CFXs under model multiplicity.
- Propose an approach to **generate robust CFXs** via MILP.

Counterfactual Explanations and Model Multiplicity: a Relational Verification View. Leofante et al, KR 2023.

Sequential products

$i := 0;$
while $(i < N)$ do
  $j := N-1;$
  while $(j > i)$ do
    if $(a[j-1] > a[j])$ then
      $swap(a, j, j-1);$
    j--
  i++

Program c

*Example taken from: Relational Verification Using Product Programs. Barthe et al, FM'11.

Sequential products

```
i := 0;
while (i < N) do
  j := N-1;
  while (j > i) do
    if (a[j-1] > a[j]) then
      swap(a, j, j-1);
    j--
  i++
```

Program c

×

```
i := 0;
while (i < N) do
  j := N-1;
  while (j > i) do
    if (a[j-1] > a[j]) then
      swap(a, j, j-1);
    j--
  i++
```

Program c’

=

```
i := 0; i' := 0;
while (i < N) do
  j := N-1; j' := N-1;
  while (j > i) do
    if (a[j-1] > a[j]) then
      swap(a, j, j-1);
    if (a'[j'-1] > a'[j']) then
      swap(a', j', j'-1);
    j--; j'--
  i++; i'++
```

Product program P

*Example taken from: Relational Verification Using Product Programs. Barthe et al, FM’11.

Sequential products

```
i := 0;
while (i < N) do
  j := N−1;
  while (j > i) do
    if (a[j−1] > a[j]) then
      swap(a, j, j−1);
    j−−
  i++
```

×

```
i := 0;
while (i < N) do
  j := N−1;
  while (j > i) do
    if (a[j−1] > a[j]) then
      swap(a, j, j−1);
    j−−
  i++
```

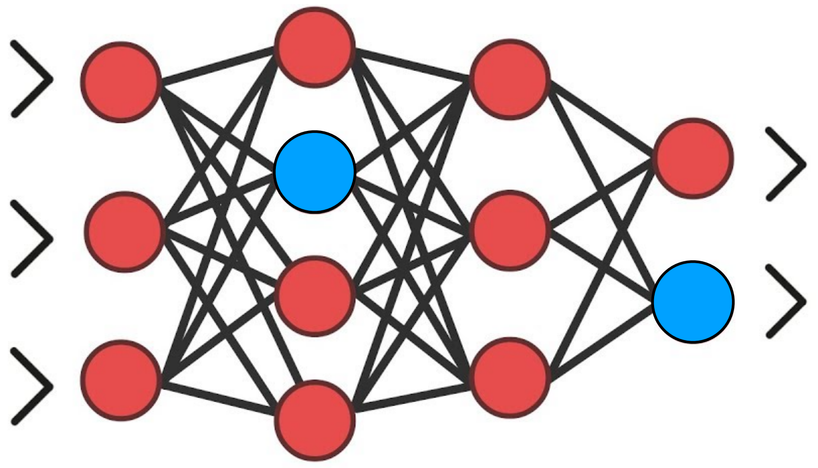
=

```
i := 0; i′ := 0;
while (i < N) do
  j := N−1; j′ := N−1;
  while (j > i) do
    if (a[j−1] > a[j]) then
      swap(a, j, j−1);
    if (a′[j′−1] > a′[j′]) then
      swap(a′, j′, j′−1);
    j−−; j′−−
  i++; i′++
```

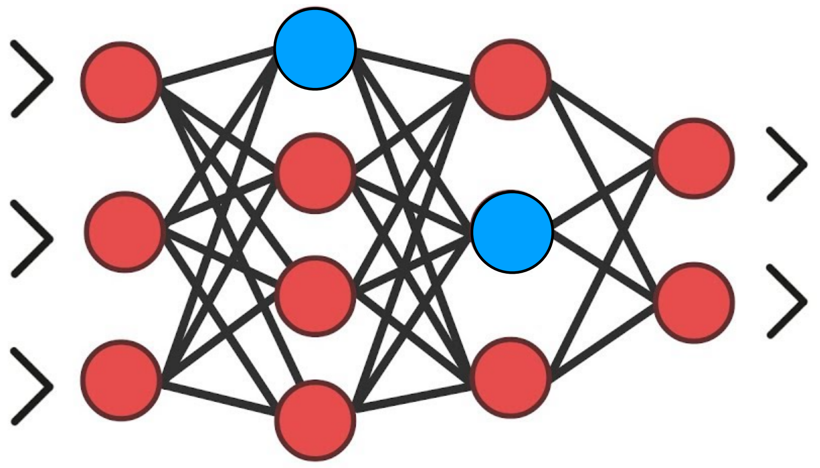
Program c

Program c’

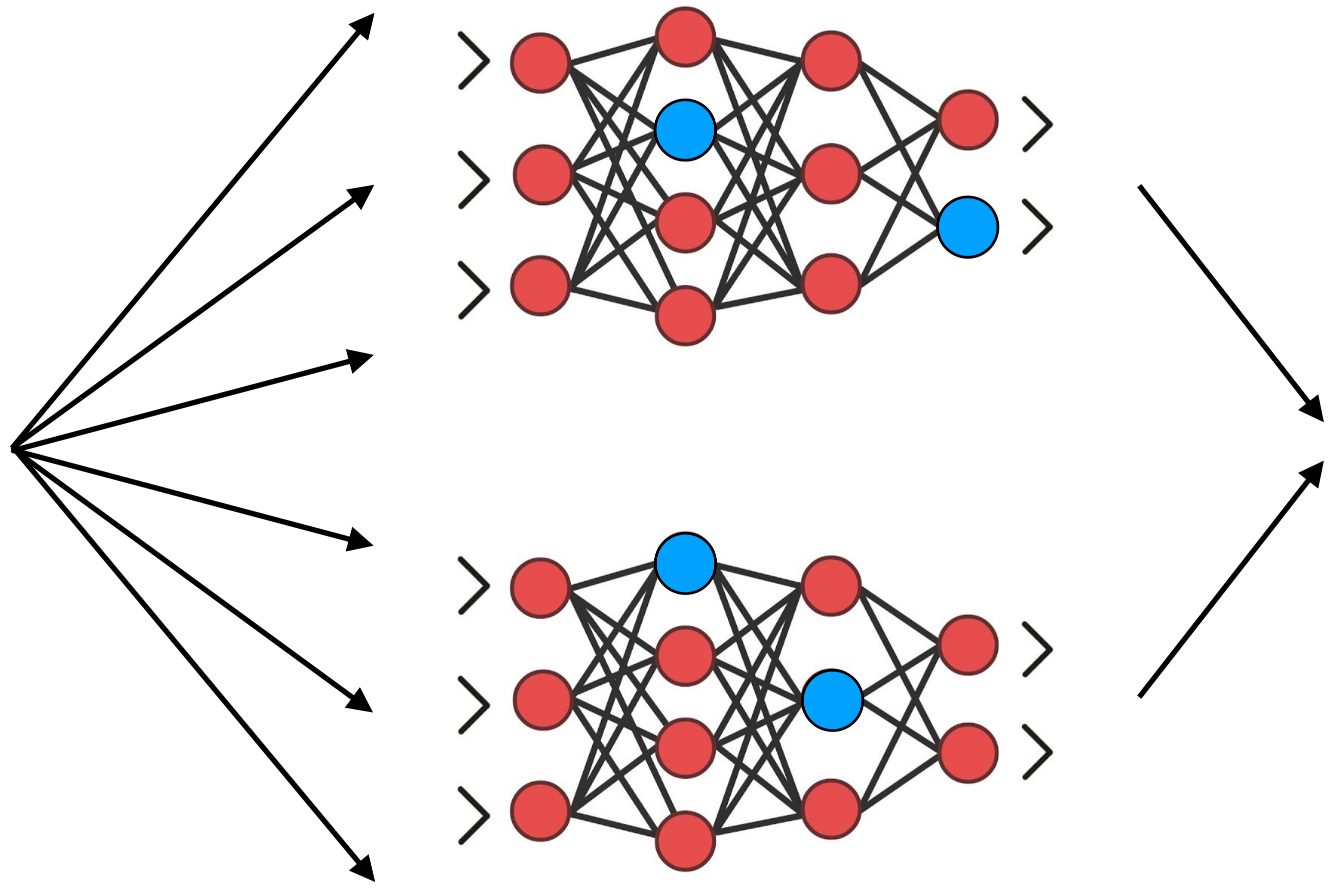
Product program P



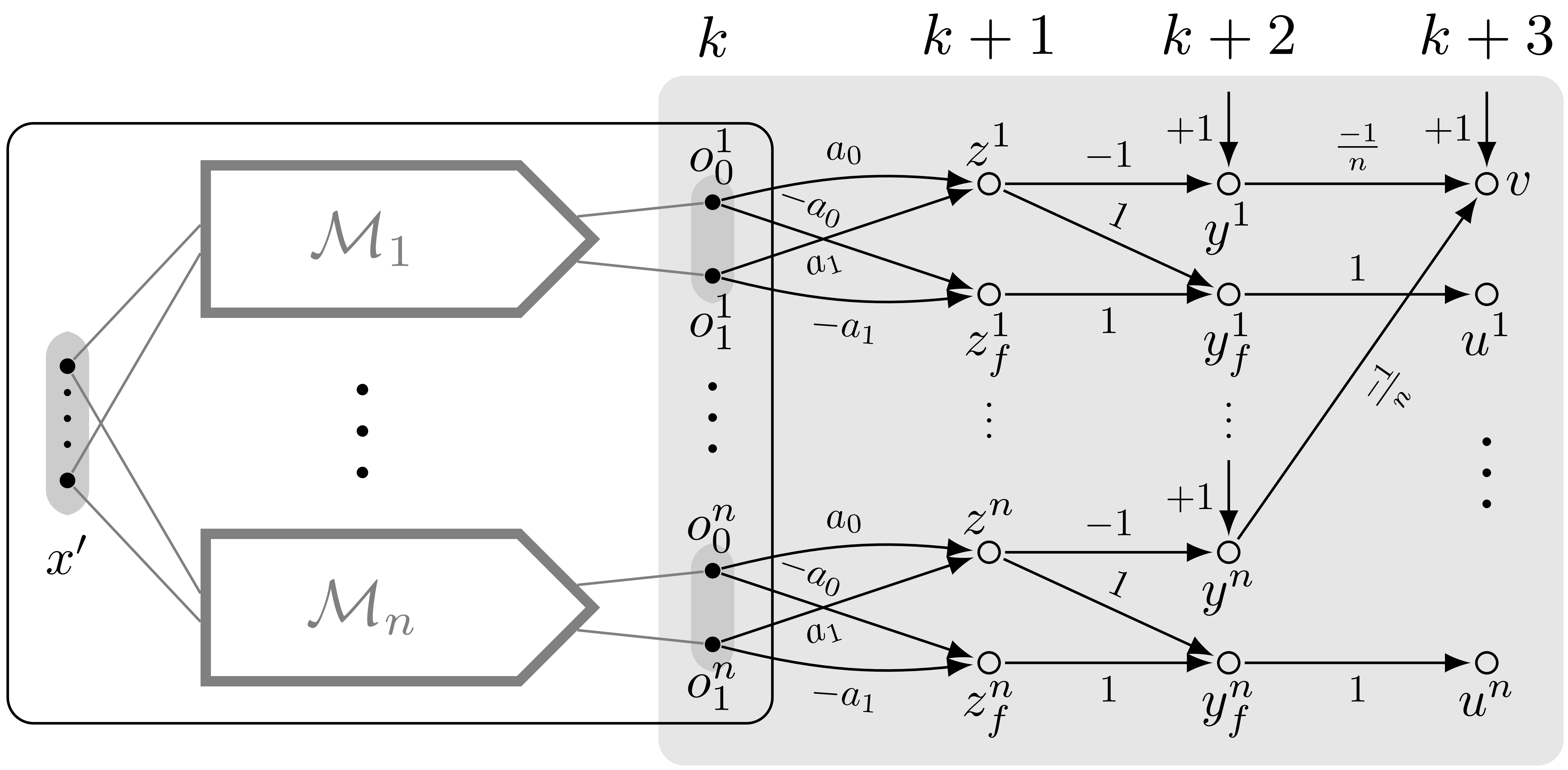
×



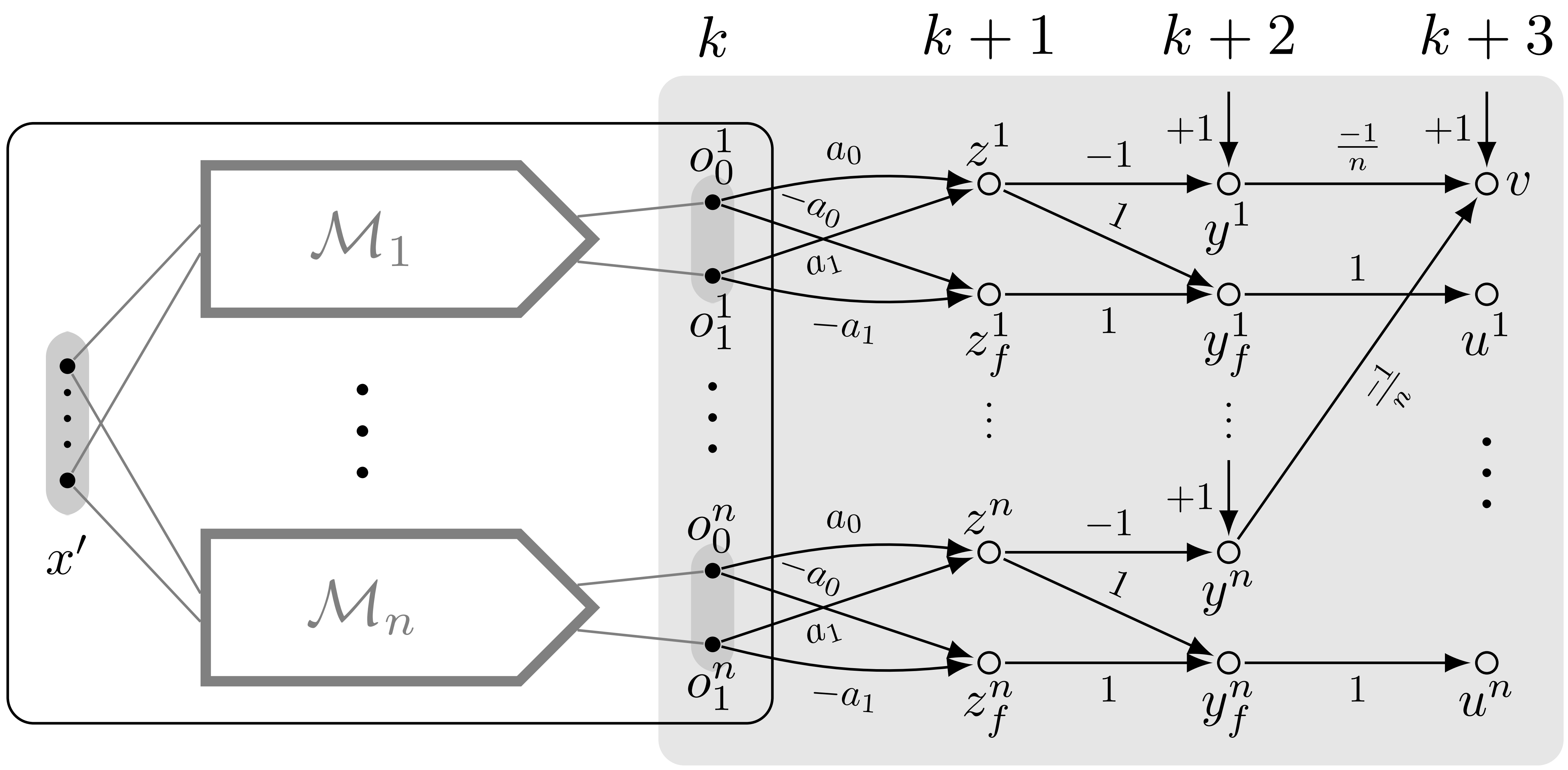
=



Our solution



Our solution



Property of the product

(P1) $v = 0$ and $u^j > 0$ for all $j \in \{1, \dots, n\}$

$\updownarrow$

(P2) *x' is a robust counterfactual for x across $\mathcal{M}$.*

Our solution

Result #1:

Thm. Determining whether there exists a robust counterfactual for a set of structurally equivalent piece-wise linear models is NP-complete.

Our solution

Result #1:

Thm. Determining whether there exists a robust counterfactual for a set of structurally equivalent piece-wise linear models is NP-complete.

Result #2:

Thm. Determining whether there exists a robust counterfactual for a set of piece-wise linear models is NP-complete.

Our solution

Result #1:

Thm. Determining whether there exists a robust counterfactual for a set of structurally equivalent piece-wise linear models is NP-complete.

Result #2:

Thm. Determining whether there exists a robust counterfactual for a set of piece-wise linear models is NP-complete.

Result #3:

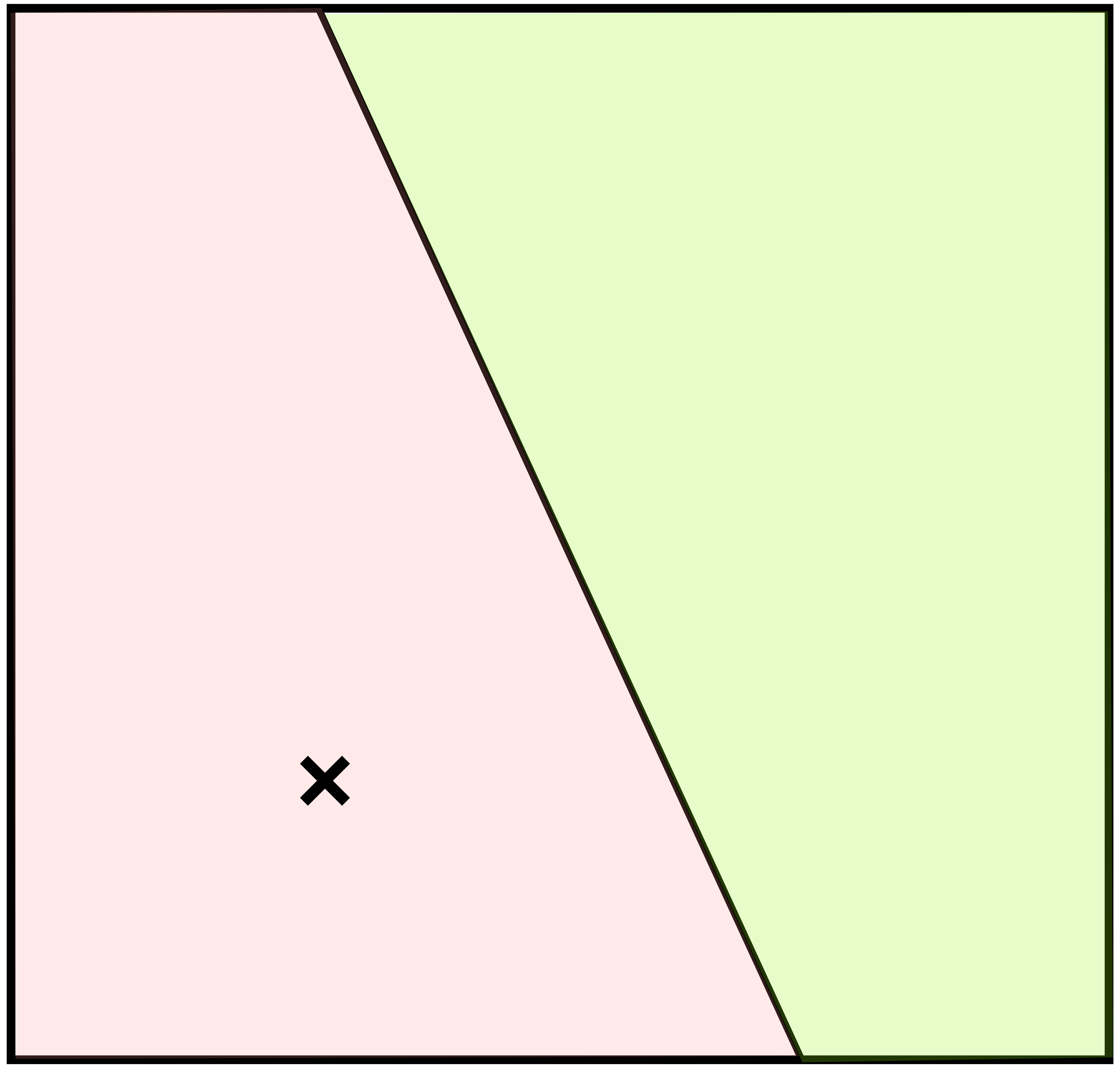
- The product network is itself a neural network
- We extend standard MILP encodings for CFX computation to generate robust CFXs under model multiplicity.

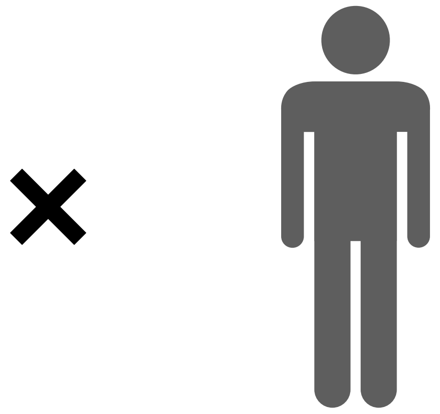
Brittle explanations ahead!

Threats

1. Model perturbations
2. Model multiplicity
3. **Noisy execution**

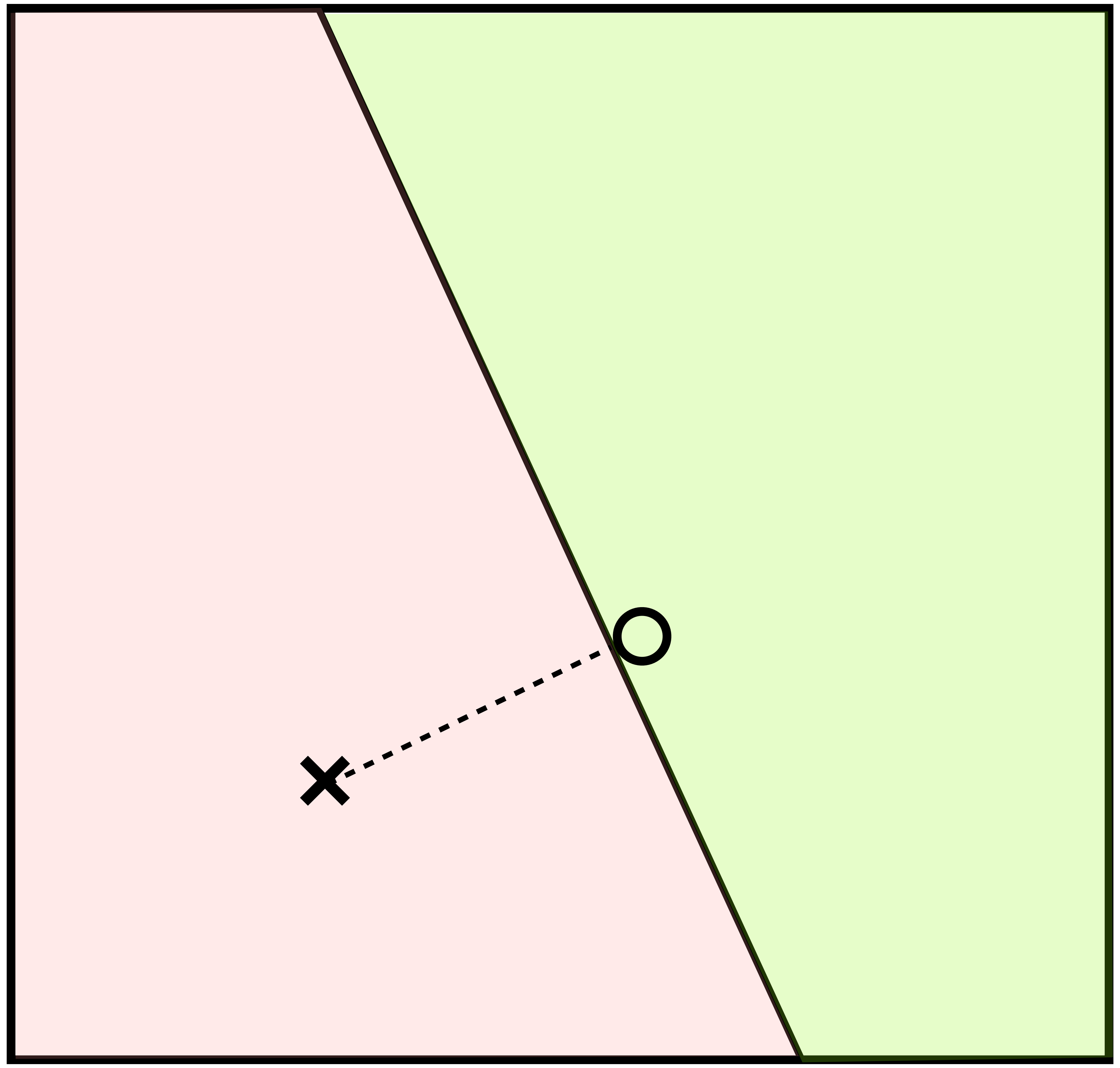
Noisy execution





- Age: 30
- Amount: **£15K**
- Duration: 24M

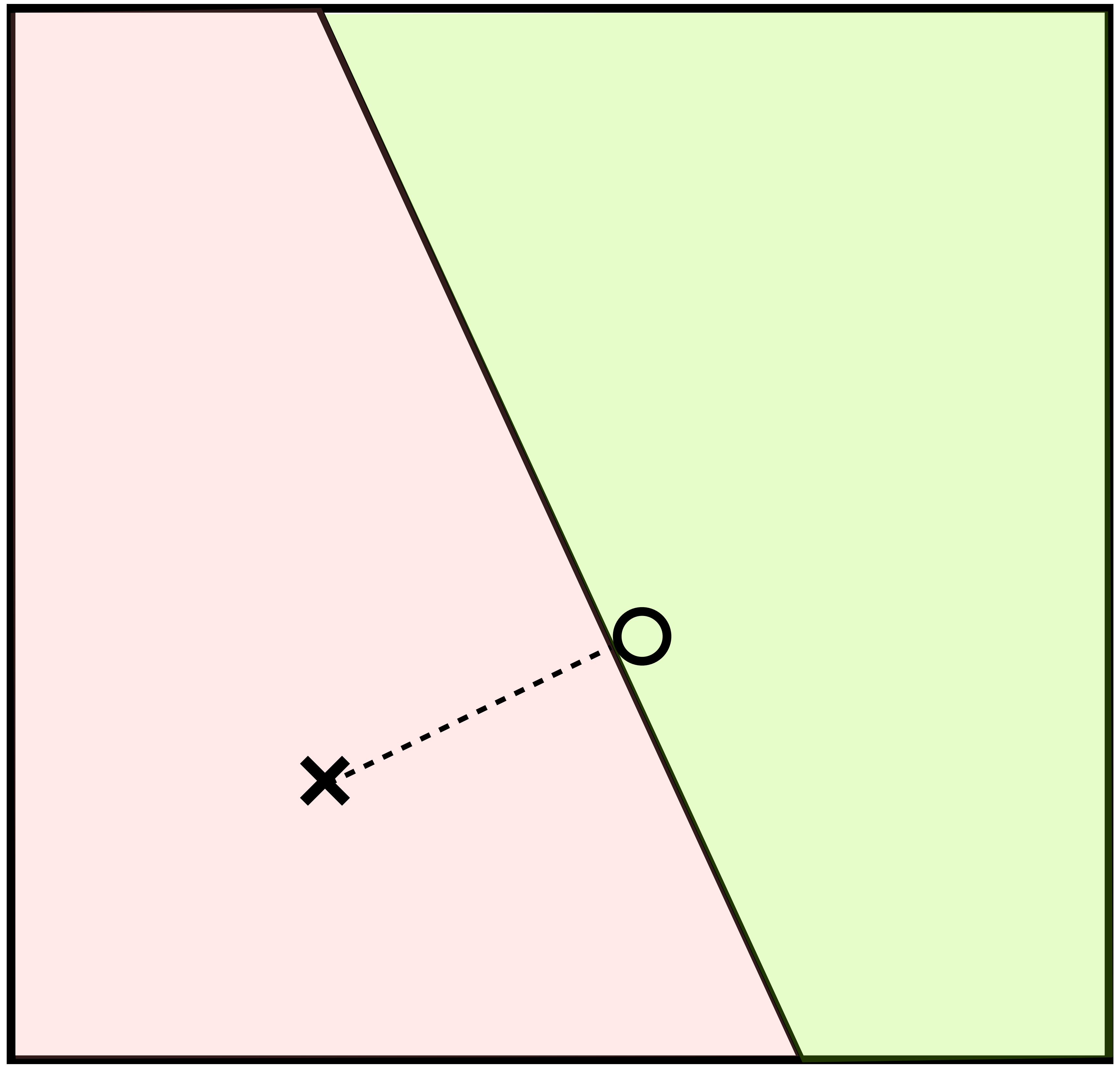
Noisy execution



- Age: 30
- Amount: **£15K**
- Duration: 24M

- Age: 30
- Amount: **£10K**
- Duration: 24M

Noisy execution

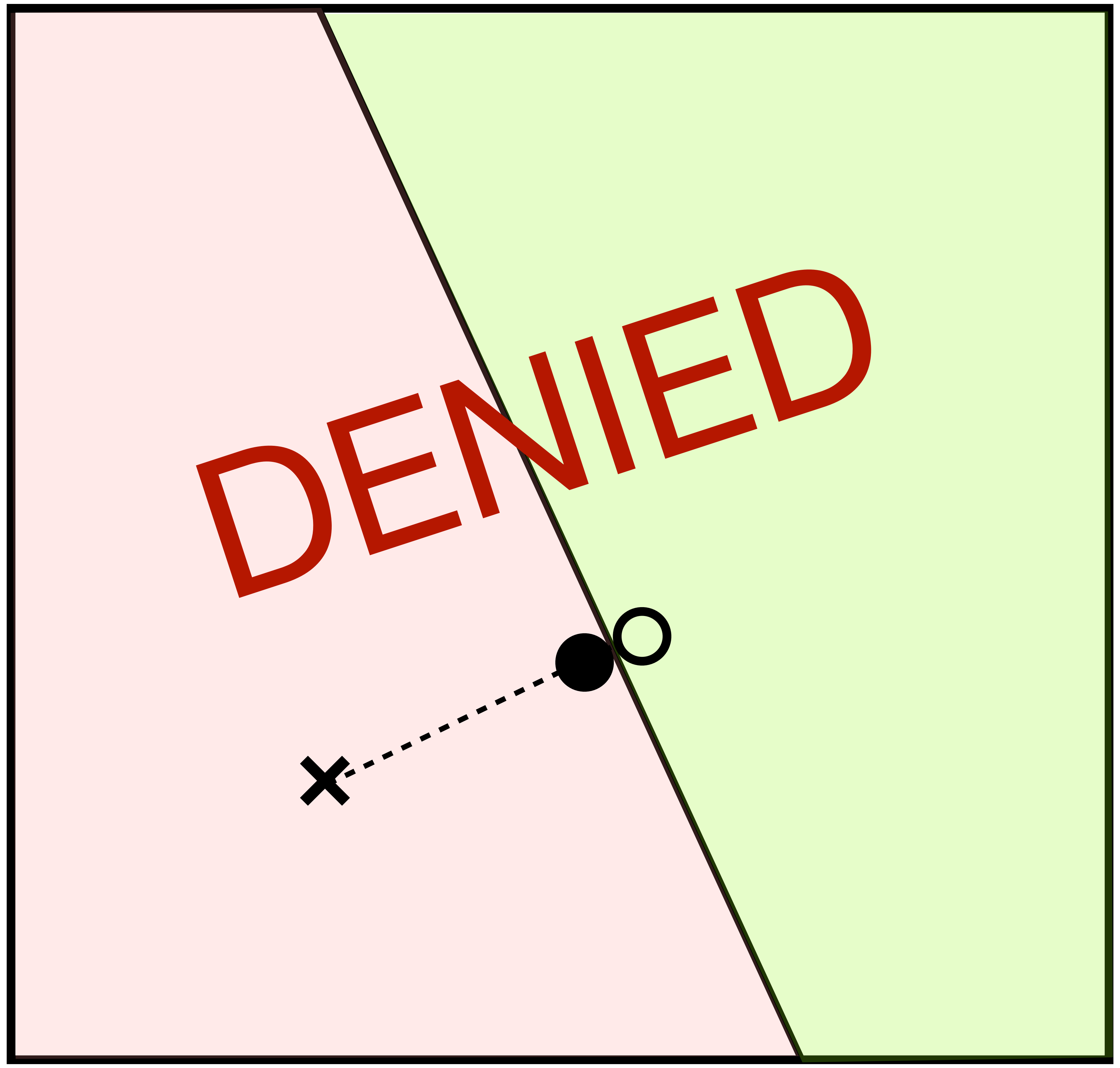


×
- Age: 30
- Amount: **£15K**
- Duration: 24M

○
- Age: 30
- Amount: **£10K**
- Duration: 24M

●
- Age: 30
- Amount: **£9.9K**
- Duration: 24M

Noisy execution



- Age: 30
- Amount: **£15K**
- Duration: 24M

- Age: 30
- Amount: **£10K**
- Duration: 24M

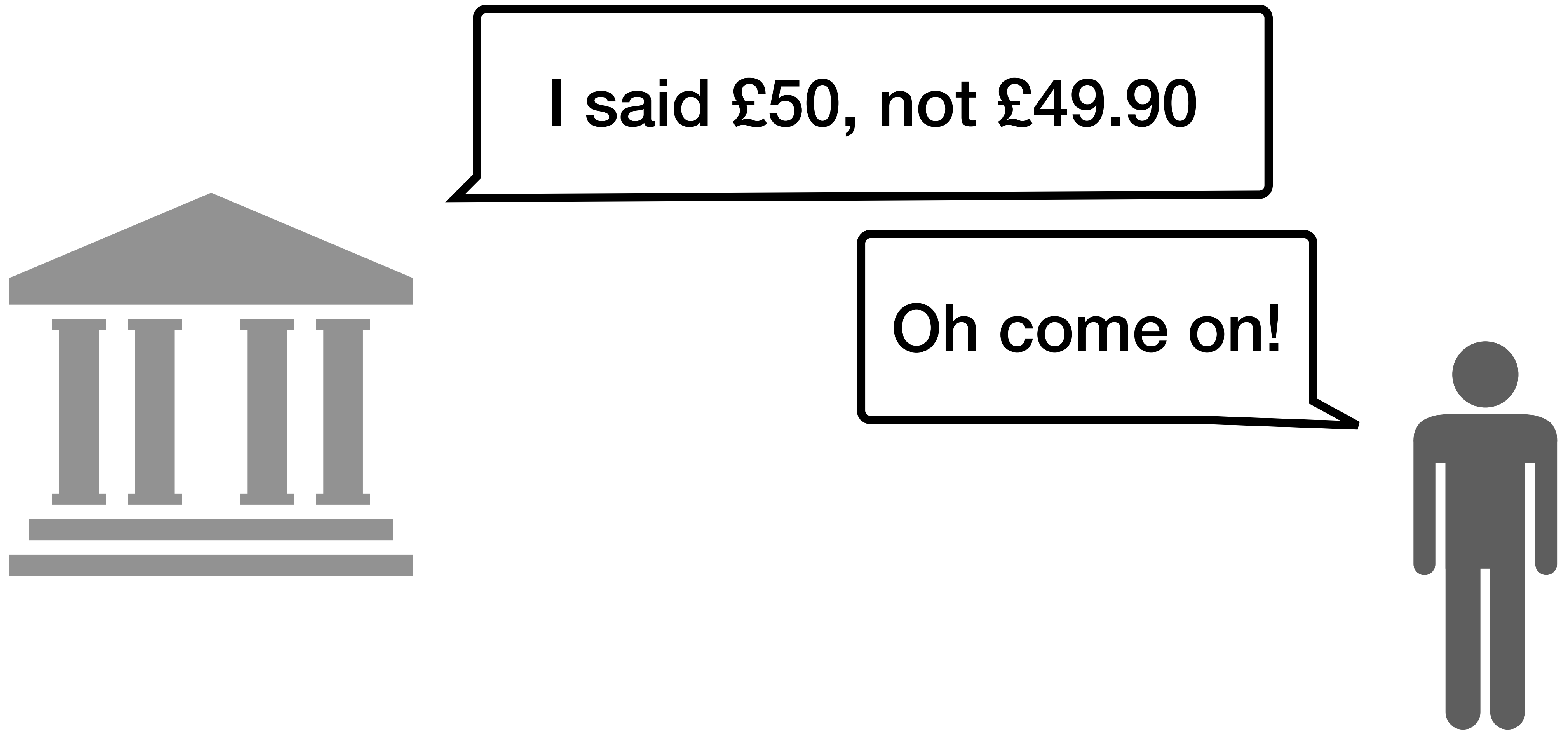
? ? ?

- Age: 30
- Amount: **£9.9K**
- Duration: 24M

Implications

Recourses are often noisily implemented in real-world settings

- Noise may **invalidate** CX
- **Jeopardise** explanatory function
- **Reduce** trust



Manipulation-Proof Machine Learning. Björkegren et al, arxiv preprint https://arxiv.org/abs/2004.03865, 2020.

Our solution

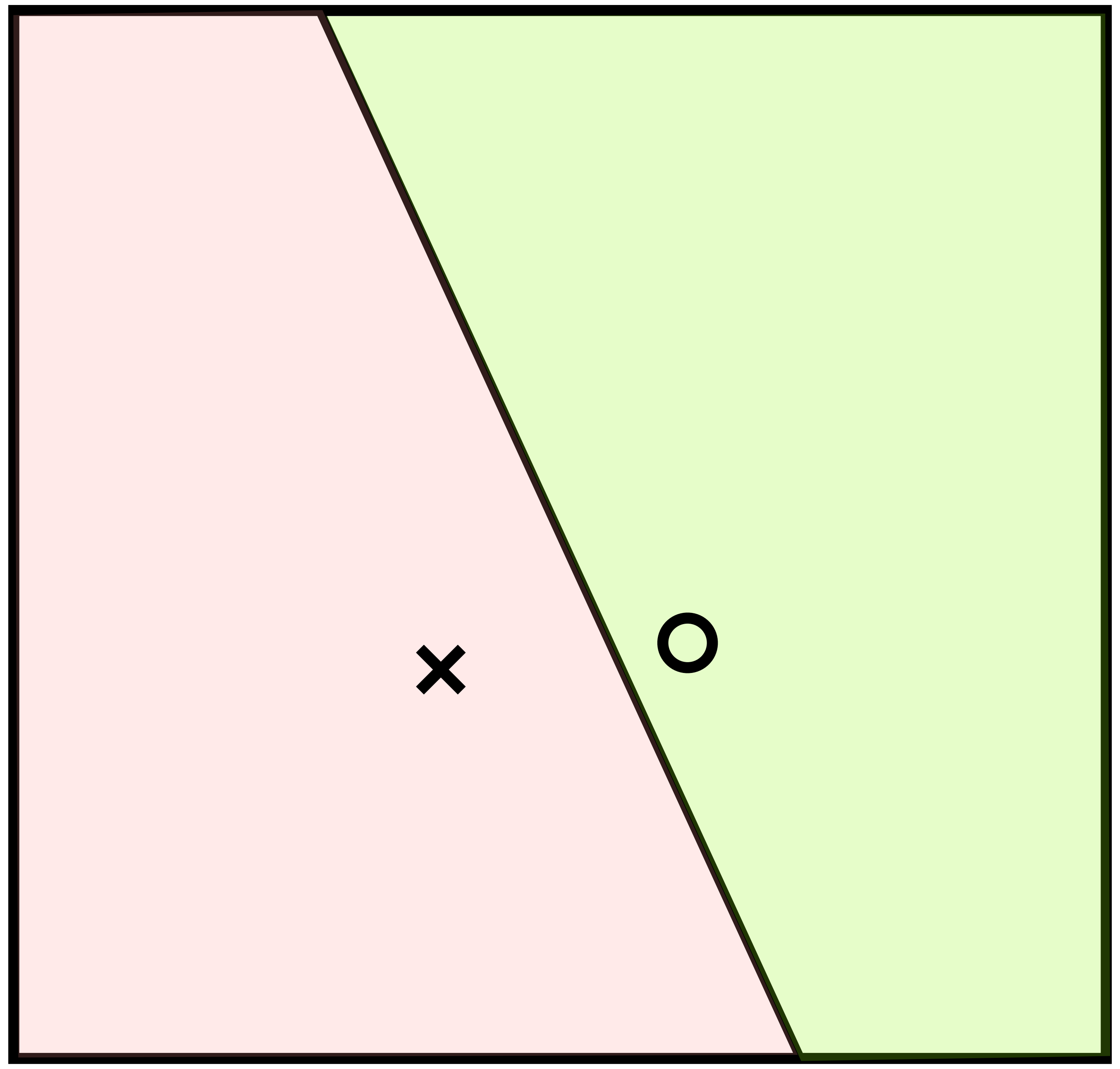
We proposed to use formal verification to identify robust CXs

Towards Robust Contrastive Explanations for Human-Neural Multi-agent Systems. Leofante and Lomuscio, AAMAS 2023.
Robust Explanations for Human-Neural Multi-agent Systems with Formal Verification. Leofante and Lomuscio, EUMAS 2023.

Our solution

We proposed to use formal verification to identify robust CXs

- Given a CX x and model $\mathscr{M}$

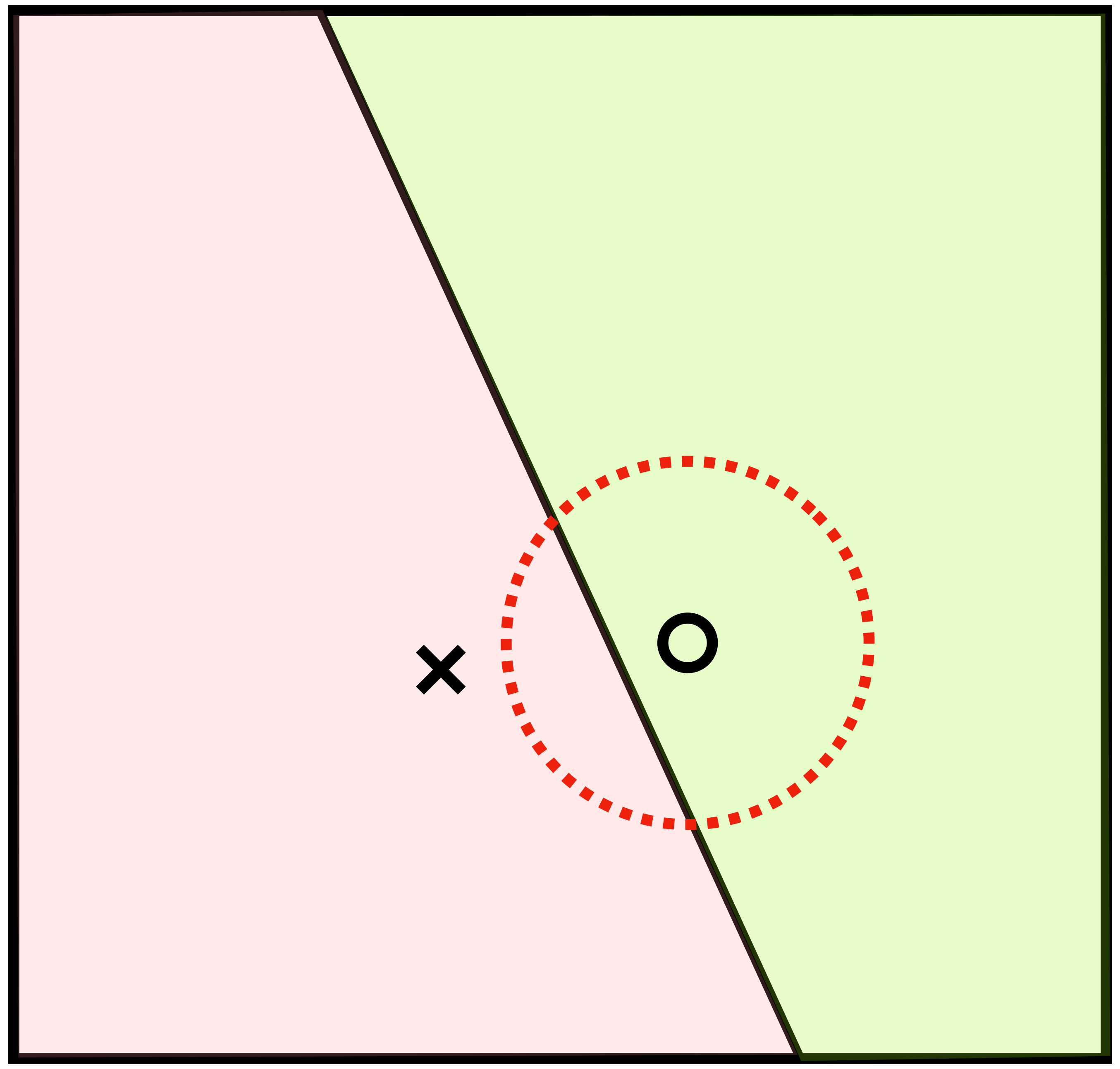


Towards Robust Contrastive Explanations for Human-Neural Multi-agent Systems. Leofante and Lomuscio, AAMAS 2023.
Robust Explanations for Human-Neural Multi-agent Systems with Formal Verification. Leofante and Lomuscio, EUMAS 2023.

Our solution

We proposed to use formal verification to identify robust CXs

- Given a CX x and model $\mathscr{M}$
- Check **local robustness** of $\mathscr{M}$ around x using verifiers

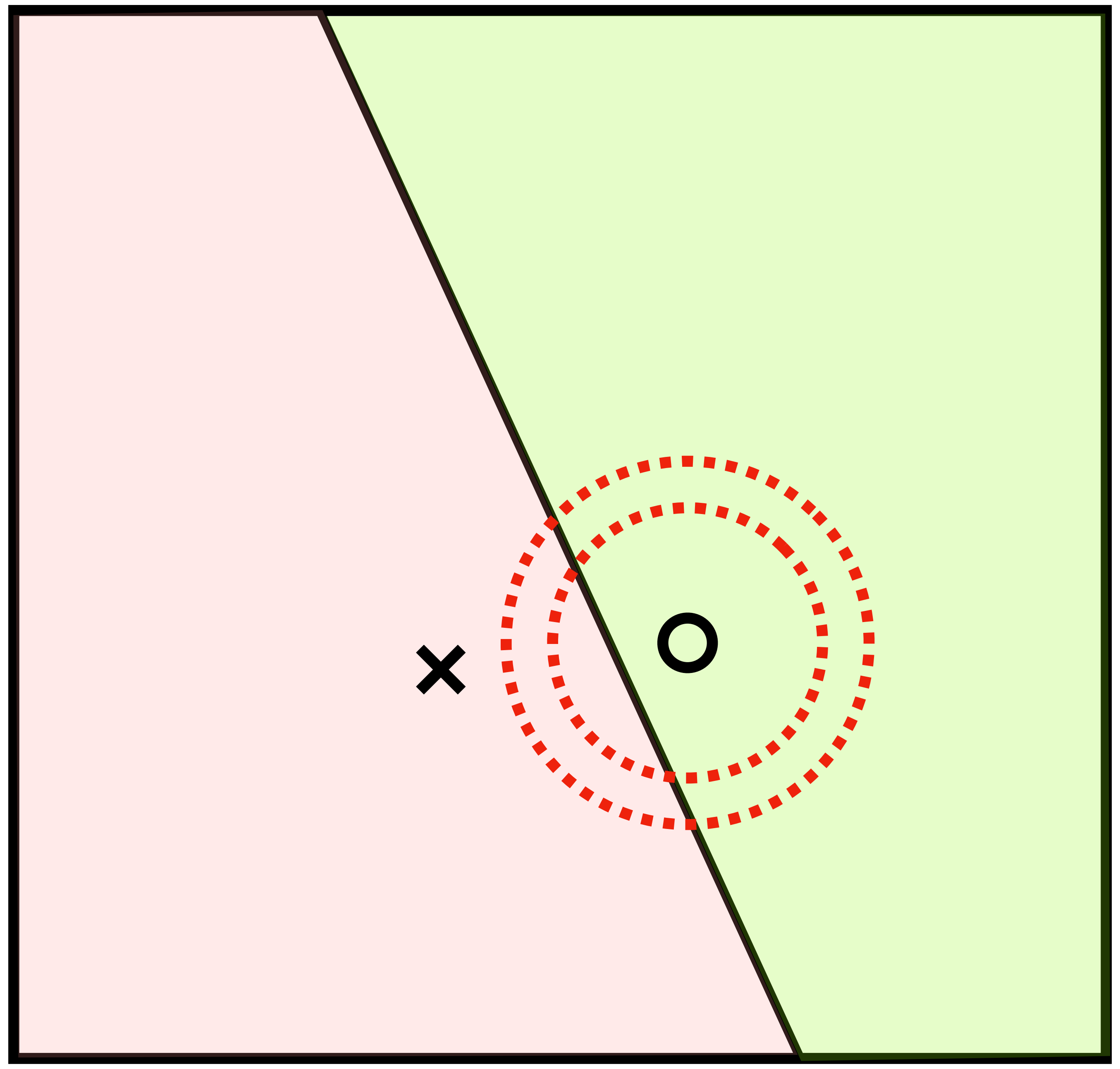


Towards Robust Contrastive Explanations for Human-Neural Multi-agent Systems. Leofante and Lomuscio, AAMAS 2023.
Robust Explanations for Human-Neural Multi-agent Systems with Formal Verification. Leofante and Lomuscio, EUMAS 2023.

Our solution

We proposed to use formal verification to identify robust CXs

- Given a CX x and model $\mathscr{M}$
- Check **local robustness** of $\mathscr{M}$ around x using verifiers

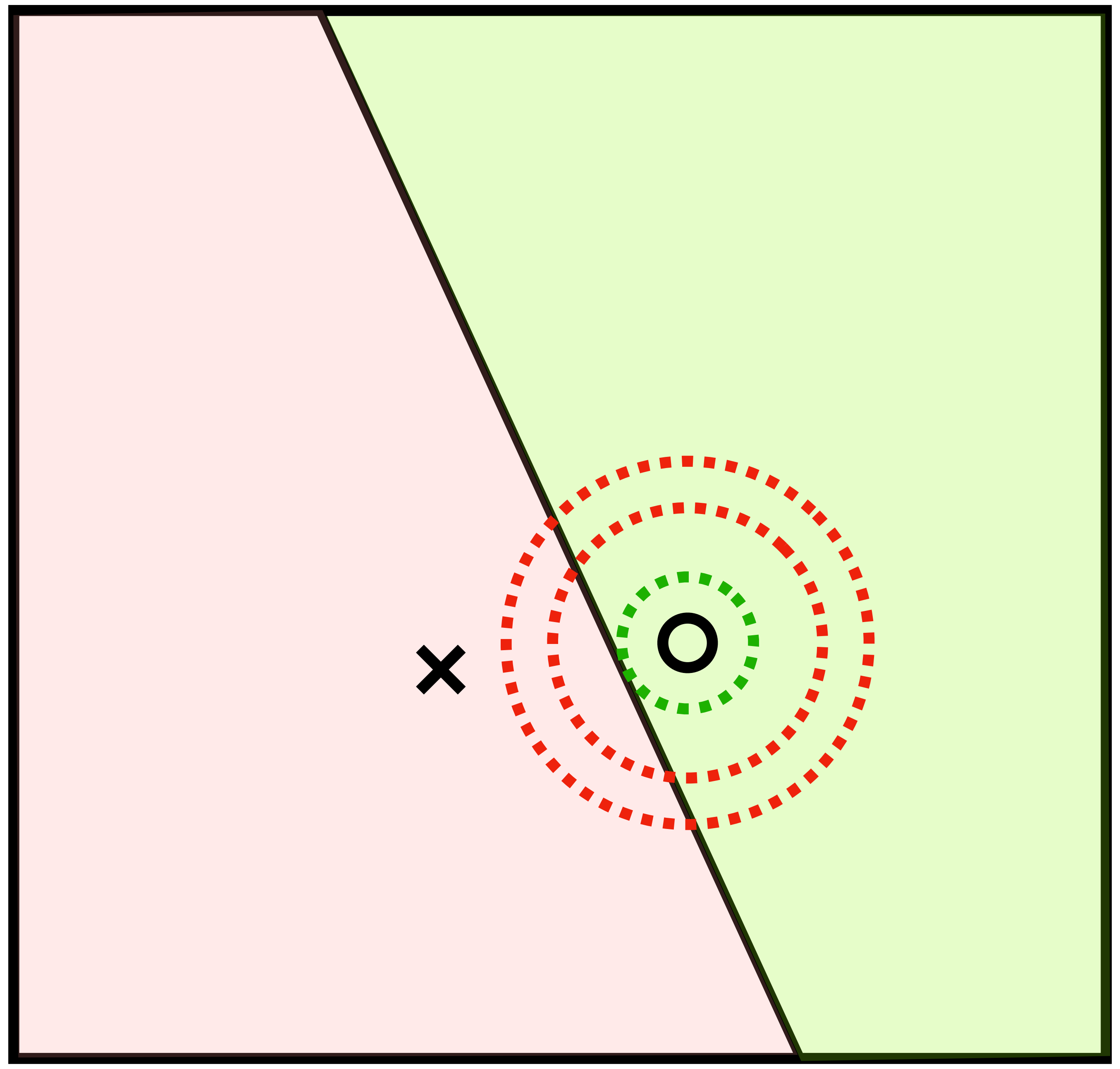


Towards Robust Contrastive Explanations for Human-Neural Multi-agent Systems. Leofante and Lomuscio, AAMAS 2023.
Robust Explanations for Human-Neural Multi-agent Systems with Formal Verification. Leofante and Lomuscio, EUMAS 2023.

Our solution

We proposed to use formal verification to identify robust CXs

- Given a CX x and model $\mathscr{M}$
- Check **local robustness** of $\mathscr{M}$ around x using verifiers
- CX **guaranteed to be robust** when safe radius identified



Towards Robust Contrastive Explanations for Human-Neural Multi-agent Systems. Leofante and Lomuscio, AAMAS 2023.
Robust Explanations for Human-Neural Multi-agent Systems with Formal Verification. Leofante and Lomuscio, EUMAS 2023.

Summing up

- CX generation methods focus on **minimising distance**
- This may result in **brittle explanations**
- We have examined **lack of robustness** in three scenarios:
  - model shifts, model multiplicity and noisy execution
- Can we borrow ideas from other areas of CS to fix this?

Thank you!

Contacts:

- f.leofante@imperial.ac.uk
- https://fraleo.github.io/
- @fraleofante

SCAN ME