# Counterfactual Explanations for Machine Learning

2024 MT4H International Workshop – Valencia, Spain

January 11-13, 2024

**Gabriele Tolomei**

Department of Computer Science

Sapienza University of Rome

tolomei@di.uniroma1.it

# Who Am I?

# Who Am I?

UniPI
(1999-2005)

# Who Am I?

UniPI
(1999-2005)

UniVE
(2008-2013)

# Who Am I?

UniPI
(1999-2005)

UniVE
(2008-2013)

Yahoo! Labs
(2014-2017)

# Who Am I?

UniPI
(1999-2005)

UniVE
(2008-2013)

Yahoo! Labs
(2014-2017)

UniPD
(2017-2019)

# Who Am I?

UniPI
(1999-2005)

UniVE
(2008-2013)

Yahoo! Labs
(2014-2017)

UniPD
(2017-2019)

Sapienza
(2019-)

# My Research Group

HERCOLE Lab

# My Research Group

Human-Explainable

HERCOLE Lab

# My Research Group

Robust

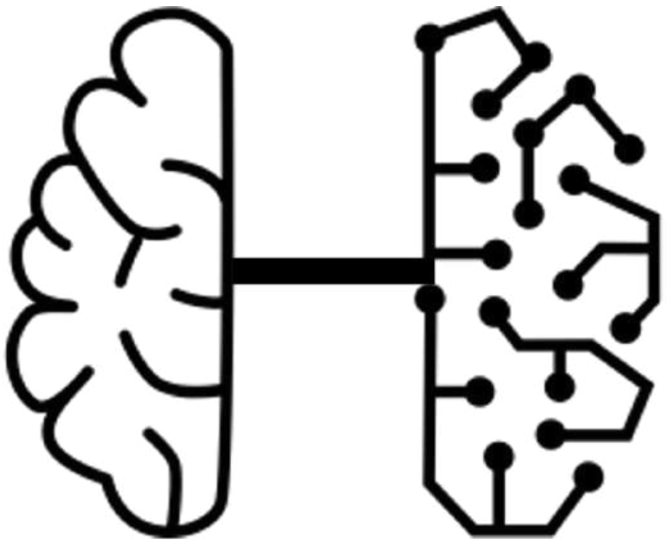
Human-Explainable

HERCOLE Lab

# My Research Group

Robust

Human-Explainable



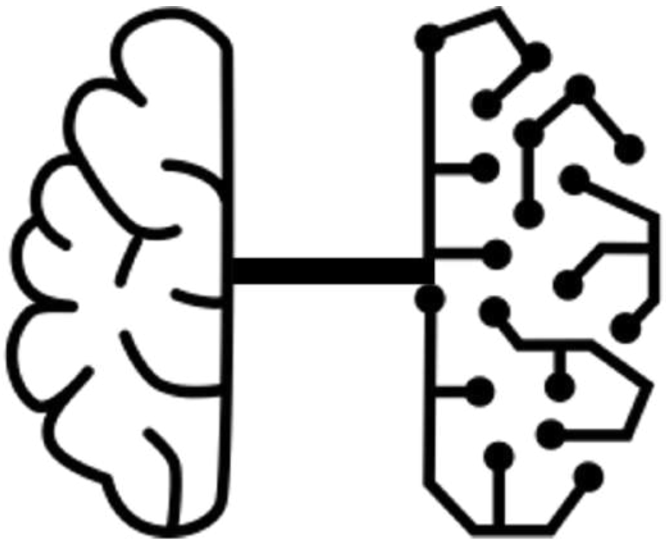
COllaborative

HERCOLE Lab

# My Research Group

Robust

Human-Explainable



COllaborative

HERCOLE Lab

LEarning

# My Research Group

Sounds cool?

HERCOLE Lab

Check out the lab's [home page](home page)

(still under construction, sic!)

# My Research Group: People

HERCOLE Lab

## PhD Students

### Cesare Campagnano
Sapienza University of Rome

PhD Student in Computer Science

### Edoardo Gabrielli
Sapienza University of Rome

PhD Student in Cybersecurity

### Flavio Giorgi
Sapienza University of Rome

PhD Student in Computer Science

### Vittoria Vineis
Sapienza University of Rome

PhD Student in Data Science

## Collaborators

### Fabio Pinelli
IMT School for Advanced Studies Lucca

Assistant Professor of Computer Science

### Fabrizio Silvestri
Sapienza University of Rome

Full Professor of Computer Science

### Federico Siciliano
Sapienza University of Rome

PhD Student in Data Science

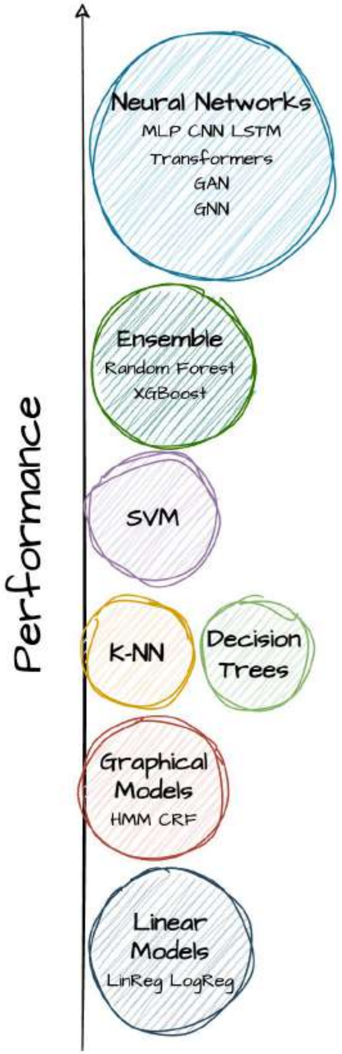
### Giovanni Trappolini
Sapienza University of Rome

Postdoctoral Researcher

### Ziheng Chen
Walmart Labs, Sunnyvale, CA, USA

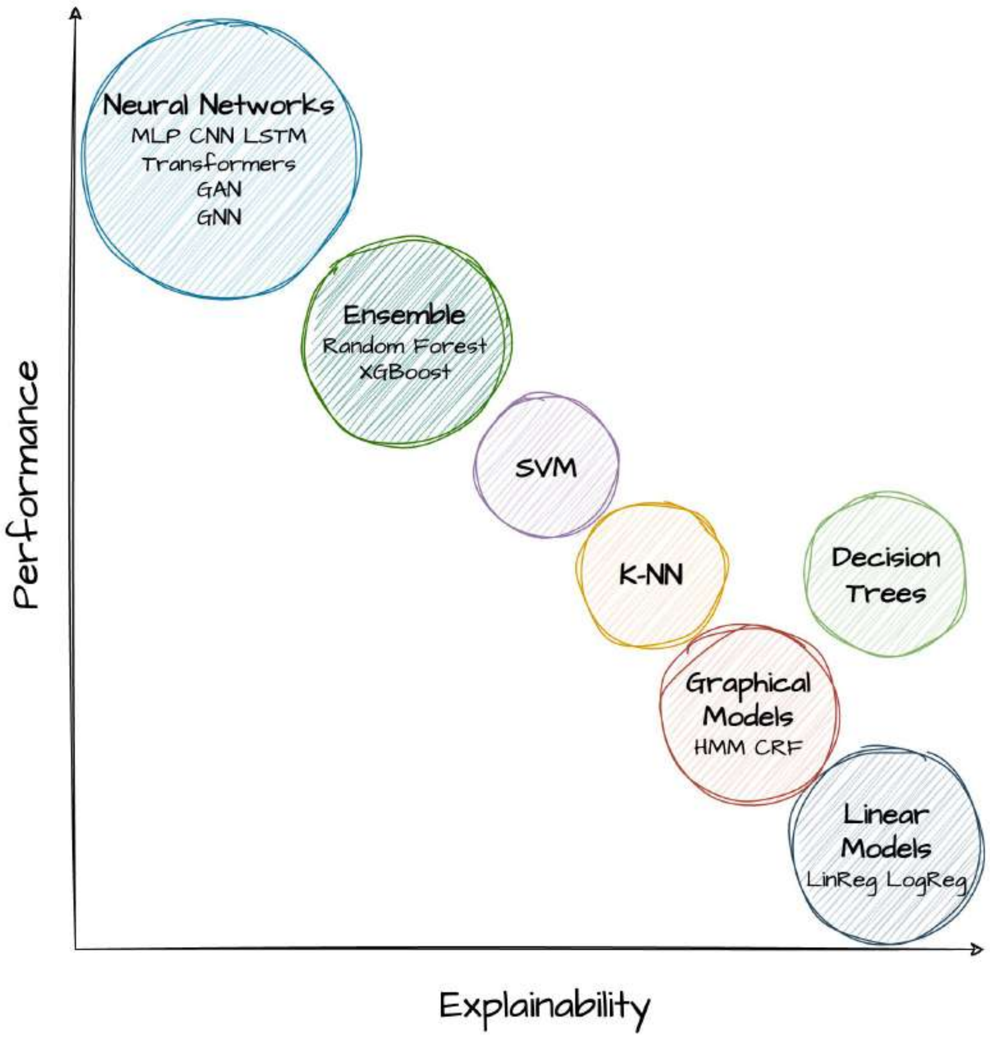
Research Scientist

# Performance vs. Explainability Trade-Off




There has been a trend for AI/ML models to get more powerful

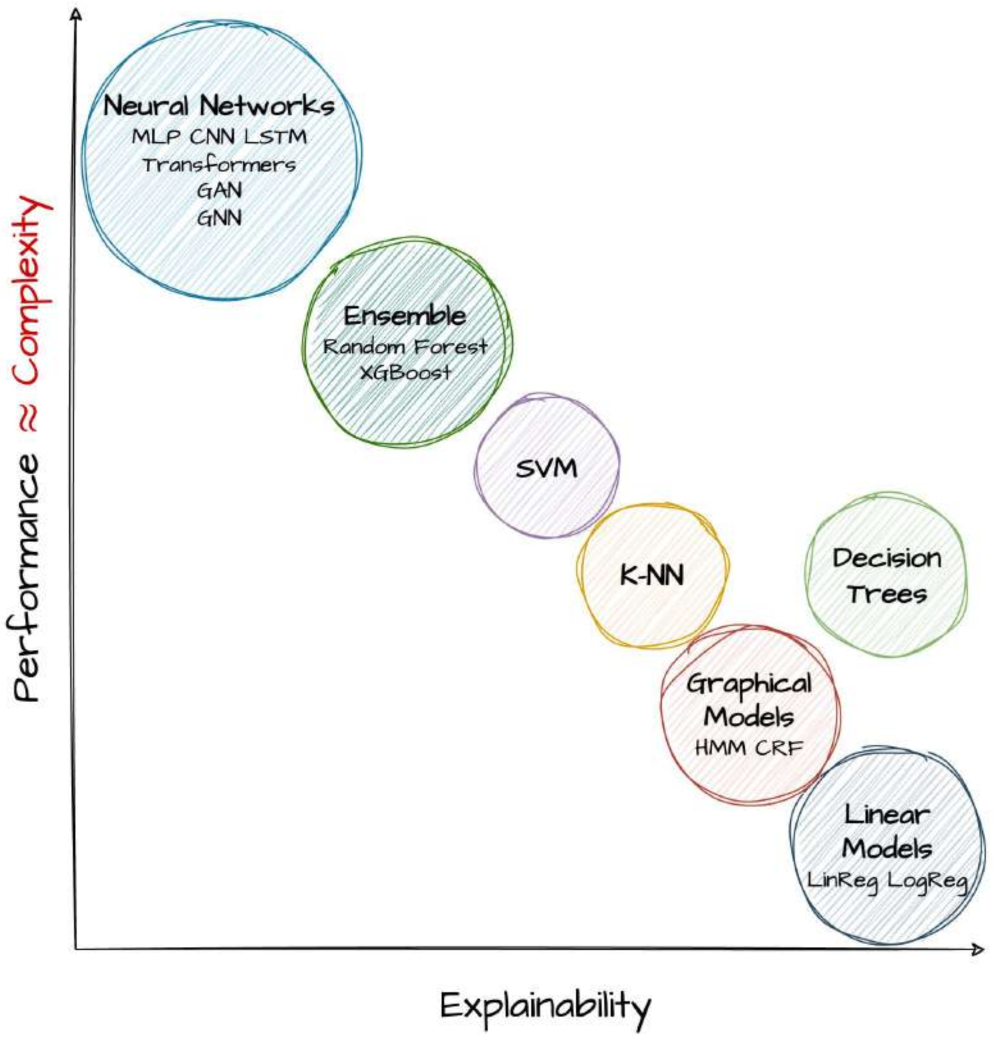
# Performance vs. Explainability Trade-Off




There has been a trend for AI/ML models to get more powerful

Performance improvements often come at a cost of **compromised explainability**

# Performance vs. Explainability Trade-Off

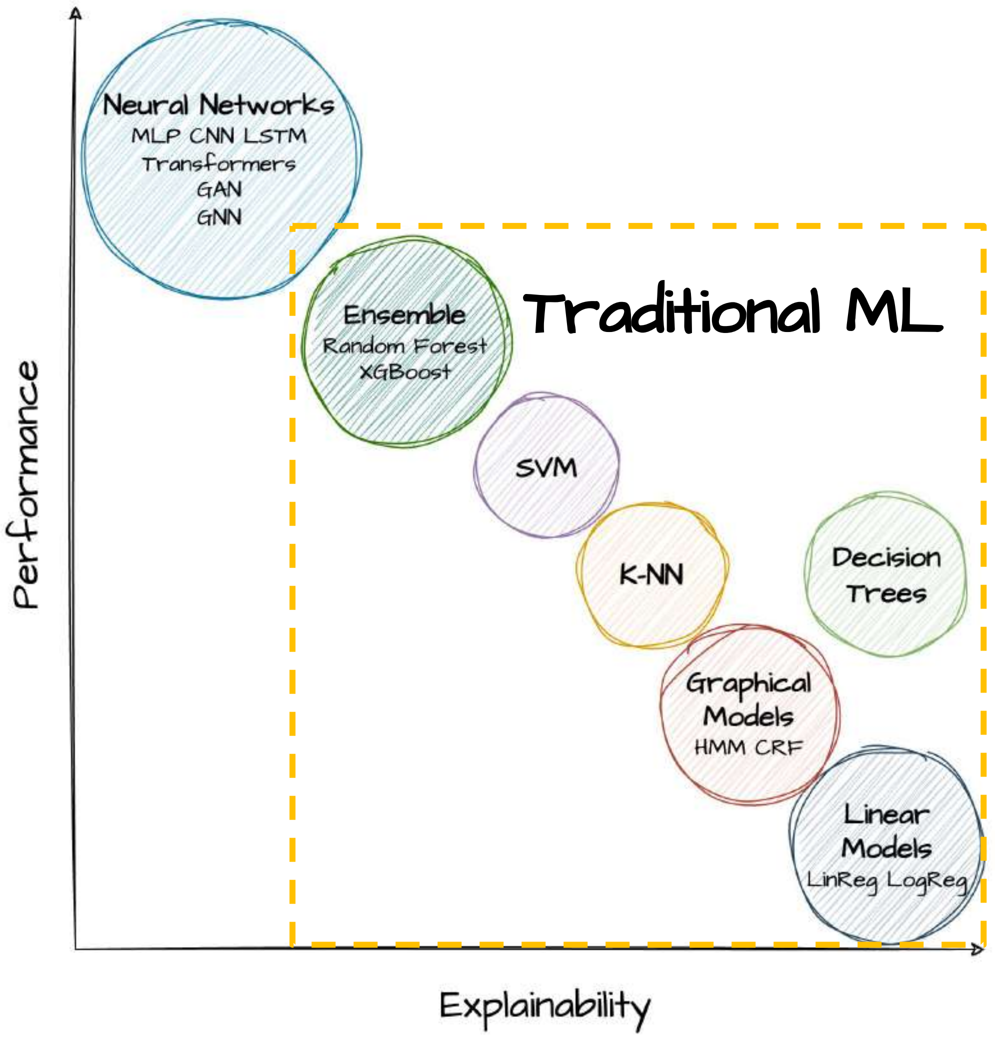



There has been a trend for AI/ML models to get more powerful

Performance improvements often come at a cost of **compromised explainability**

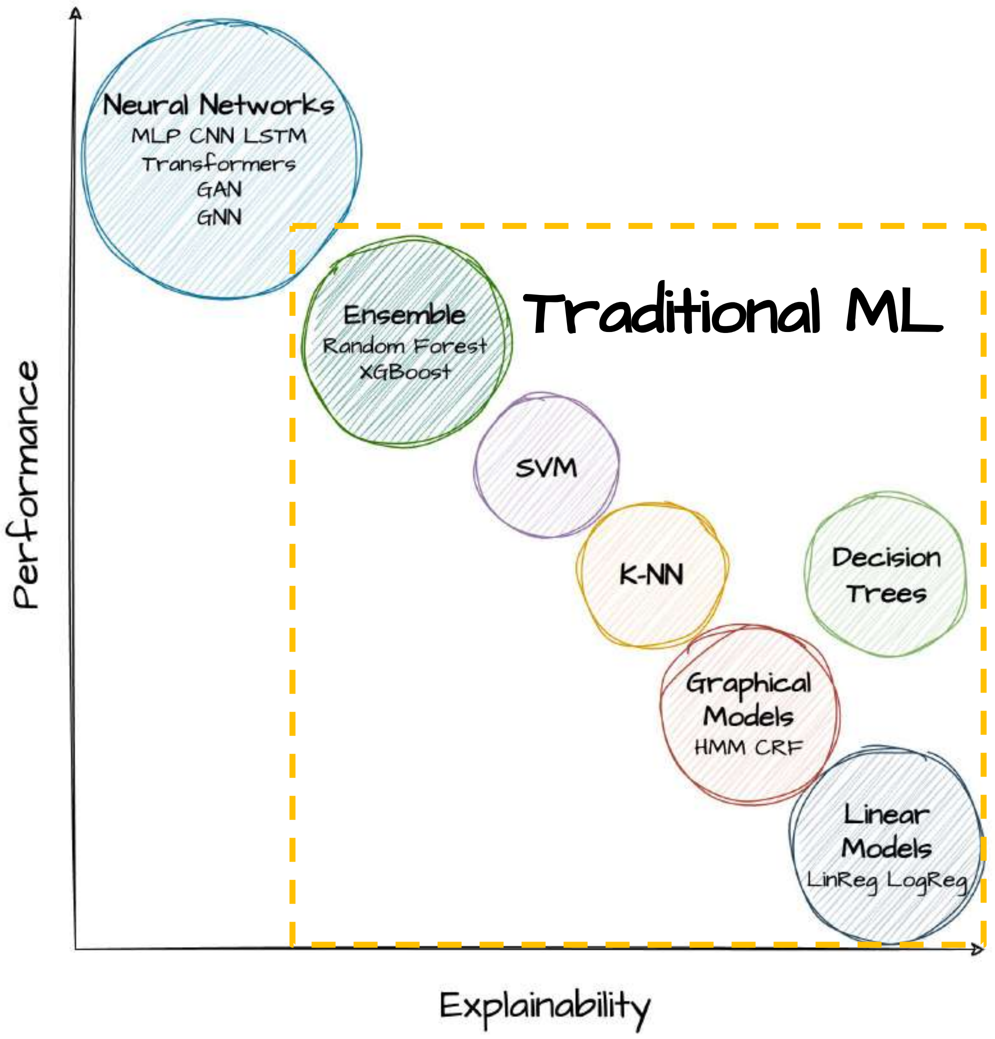
This is partly due to the **increasing model complexity** (e.g., number of parameters, deep network architectures)

# Performance vs. Explainability Trade-Off

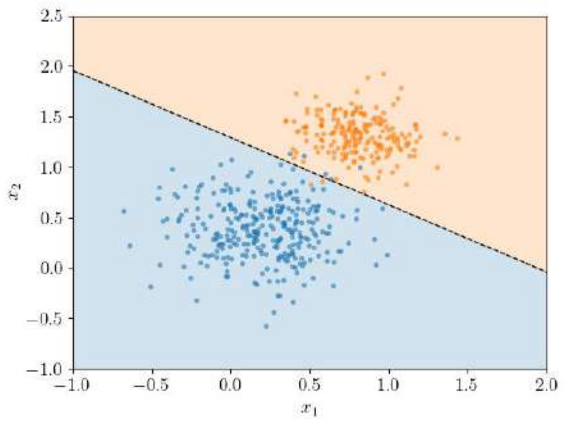



Simpler models may be **less accurate** but **more explainable**
e.g., linear/logistic regression coefficients are interpretable *by design*

# Performance vs. Explainability Trade-Off

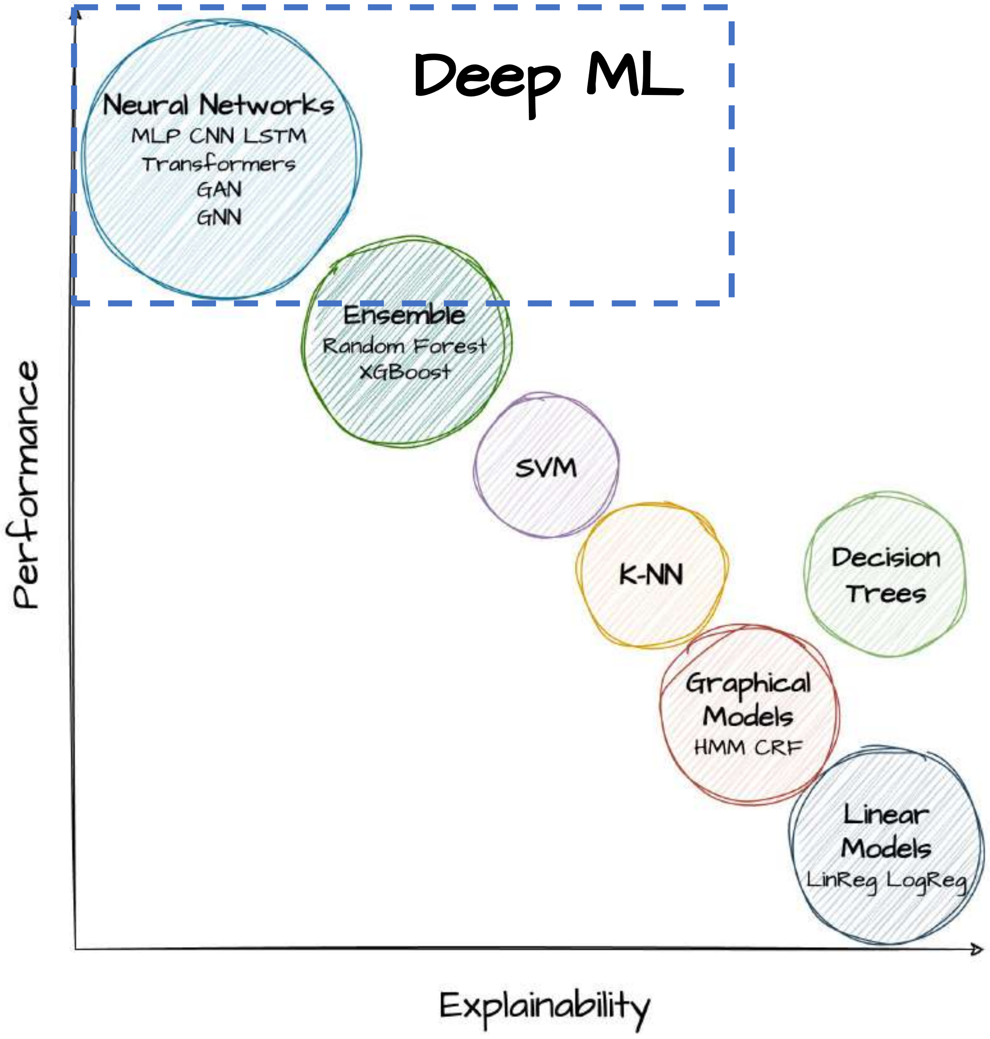



Simpler models may be **less accurate** but **more explainable**
e.g., linear/logistic regression coefficients are interpretable *by design*



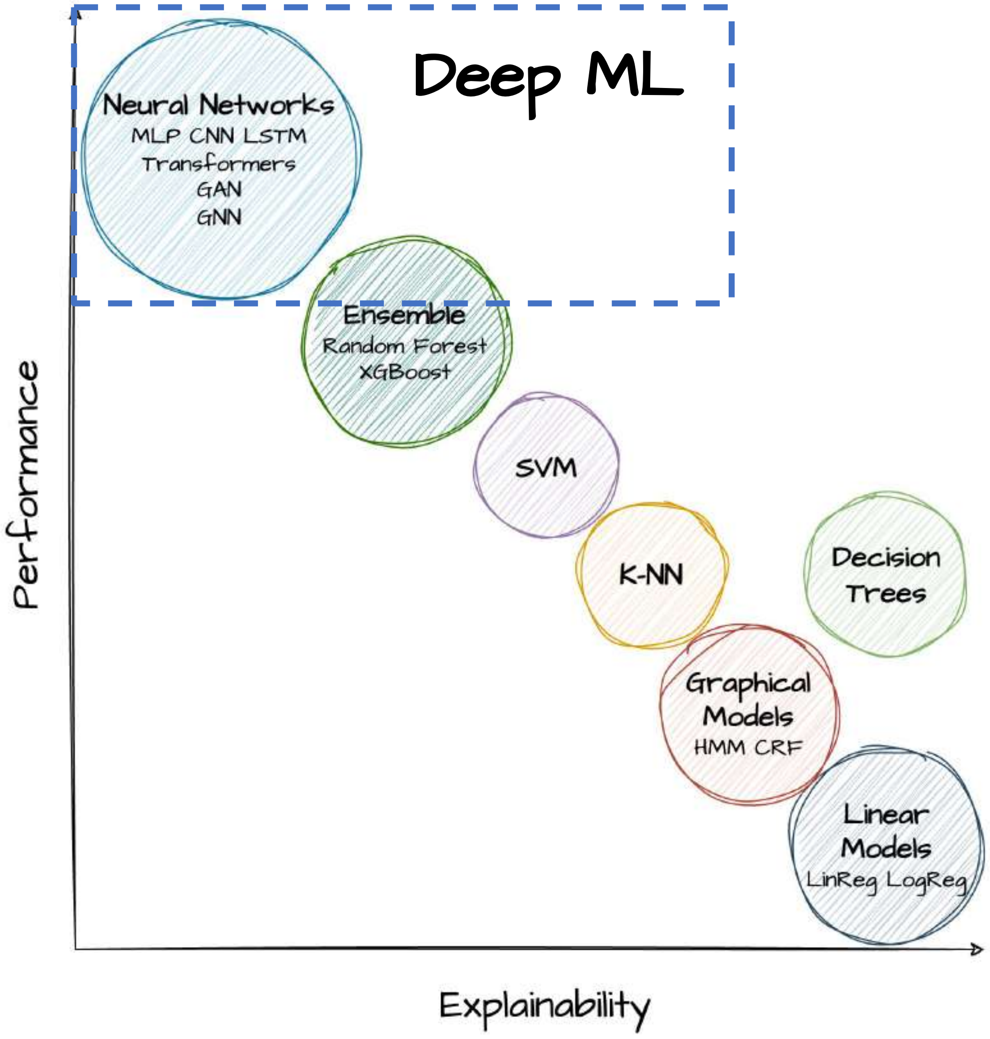

**Simple Decision Boundary Surface**

# Performance vs. Explainability Trade-Off

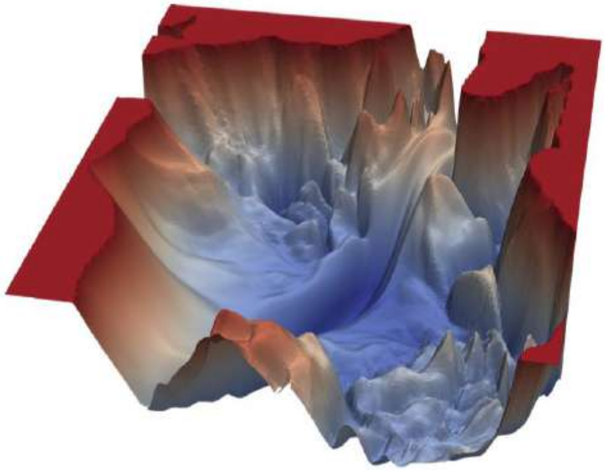



Complex models are **more expressive** but **opaque**
e.g., multi-billion parameter NNs

# Performance vs. Explainability Trade-Off




Complex models are **more expressive** but **opaque**
e.g., multi-billion parameter NNs



**Convoluted Decision Boundary Surface**

# The Need for Explainable AI (XAI)


https://medium.com/@BonsaiAI/what-do-we-want-from-explainable-ai-5ed12cb36c07

# The Need for Explainable AI (XAI)

- AI/ML systems are widely deployed to support decision-making processes in several application contexts

# The Need for Explainable AI (XAI)

- AI/ML systems are widely deployed to support decision-making processes in several application contexts
- In many domains, **highly accurate predictions are not enough!**

# The Need for Explainable AI (XAI)

- AI/ML systems are widely deployed to support decision-making processes in several application contexts
- In many domains, **highly accurate predictions are not enough!**
  - **Healthcare:** A physician must be able to tell their patient the rationale behind an AI/ML-based diagnosis

# The Need for Explainable AI (XAI)

- AI/ML systems are widely deployed to support decision-making processes in several application contexts
- In many domains, **highly accurate predictions are not enough!**
  - **Healthcare:** A physician must be able to tell their patient the rationale behind an AI/ML-based diagnosis
  - **Finance:** A banker must be able to tell their customer why they won't grant them a loan

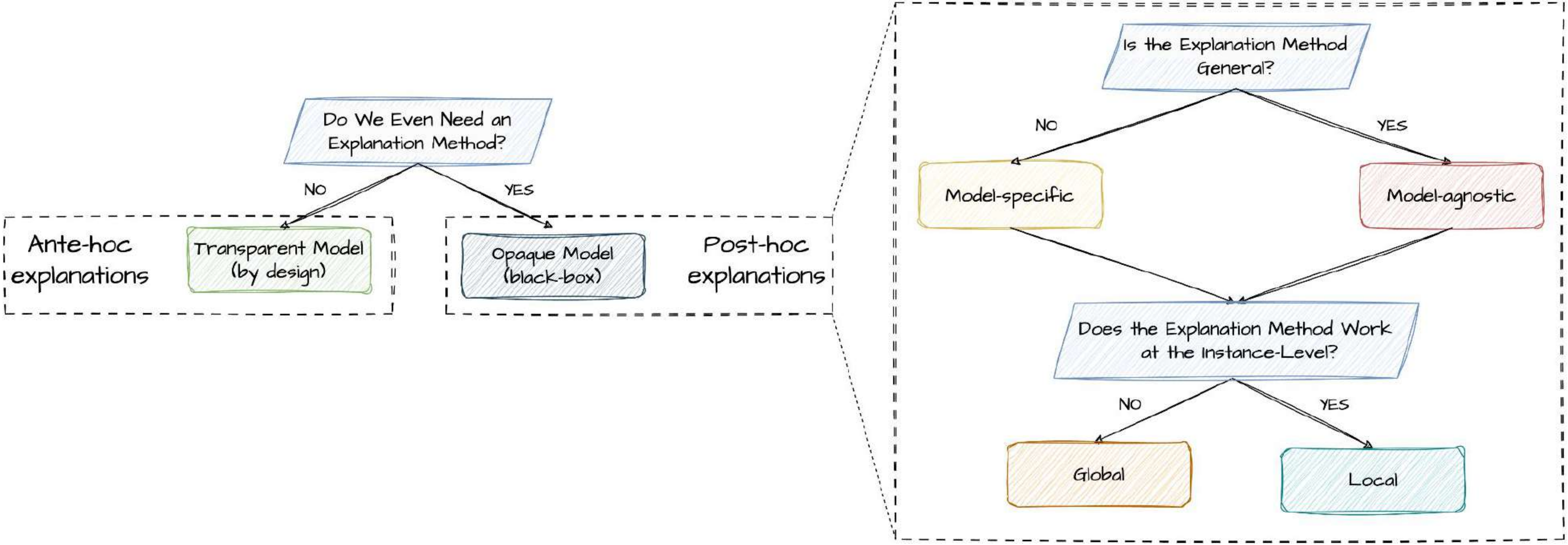
# The Need for Explainable AI (XAI)

- AI/ML systems are widely deployed to support decision-making processes in several application contexts
- In many domains, **highly accurate predictions are not enough!**
  - **Healthcare:** A physician must be able to tell their patient the rationale behind an AI/ML-based diagnosis
  - **Finance:** A banker must be able to tell their customer why they won't grant them a loan
- AI/ML-based predictions should be comprehensible to every stakeholder (including non-experts)

# The Need for Explainable AI (XAI)

- Several attempts have been made to promote XAI as part of broader **data privacy regulation initiatives**
  - EU GDPR (General Data Protection Regulation)
  - HIPAA (Health Insurance Portability and Accountability Act) Privacy Rule
  - CCPA (California Consumer Privacy Act)
  - PCI DSS (Payment Card Industry Data Security Standard)
  - NIST AI Risk Management Framework
  - …

# Taxonomy of XAI Methods



# Counterfactual Explanations: Intuition

- Post-hoc local explanation method to interpret predictions of individual instances

# Counterfactual Explanations: Intuition

- Post-hoc local explanation method to interpret predictions of individual instances
- Search for modified versions of input samples that result in alternative output responses from the predictive model

# Counterfactual Explanations: Intuition

- Post-hoc local explanation method to interpret predictions of individual instances
- Search for modified versions of input samples that result in alternative output responses from the predictive model
- Explanations take the following form:

  "***If A had been different, B would not have occurred***"

# Counterfactual Explanations: Example

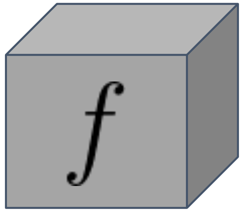
Will I have diabetes?

# Counterfactual Explanations: Example

Will I have diabetes?

Yes/No + Explanation

AI/ML model

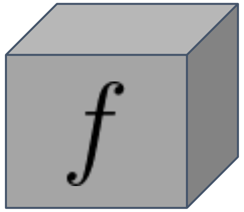


# Counterfactual Explanations: Example

Will I have diabetes?

Yes/No + Explanation

AI/ML model




**Factual**

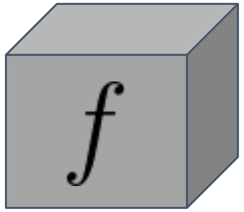
| Age | Gender | Exercise Level | Fat Level |
|---|---|---|---|
| 45 | M | Low | High |

# Counterfactual Explanations: Example

Will I have diabetes?

Yes/No + Explanation

AI/ML model




**Factual**

| **Age** | **Gender** | **Exercise Level** | **Fat Level** |
|---|---|---|---|
| 45 | M | Low | High |

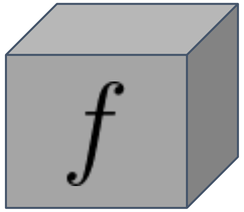
$f\Big($  $\Big) =$ **Yes**

# Counterfactual Explanations: Example

Will I have diabetes?

Yes/No + Explanation

AI/ML model




**Factual**

| Age | Gender | Exercise Level | Fat Level |
|---|---|---|---|
| 45 | M | Low | High |

$f($$) =$ **Yes**

***Counterfactual***

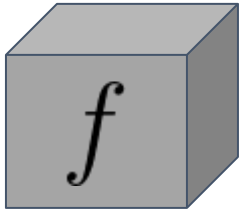
| Age | Gender | Exercise Level | Fat Level |
|---|---|---|---|
| 45 | M | **Medium** | **Low** |

# Counterfactual Explanations: Example

Will I have diabetes?

Yes/No + Explanation

AI/ML model



**Factual**

| **Age** | **Gender** | **Exercise Level** | **Fat Level** |
|---|---|---|---|
| 45 | M | Low | High |

$f($  $) =$ **Yes**

***Counterfactual***

| **Age** | **Gender** | **Exercise Level** | **Fat Level** |
|---|---|---|---|
| 45 | M | **Medium** | **Low** |

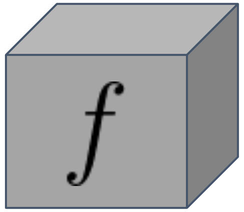
$f($  $) =$ **No**

# Counterfactual Explanations: Example

Will I have diabetes?

Yes/No + Explanation

AI/ML model



**Factual**

| **Age** | **Gender** | **Exercise Level** | **Fat Level** |
|---|---|---|---|
| 45 | M | Low | High |

$f($  $)=$ **Yes**

***Counterfactual***

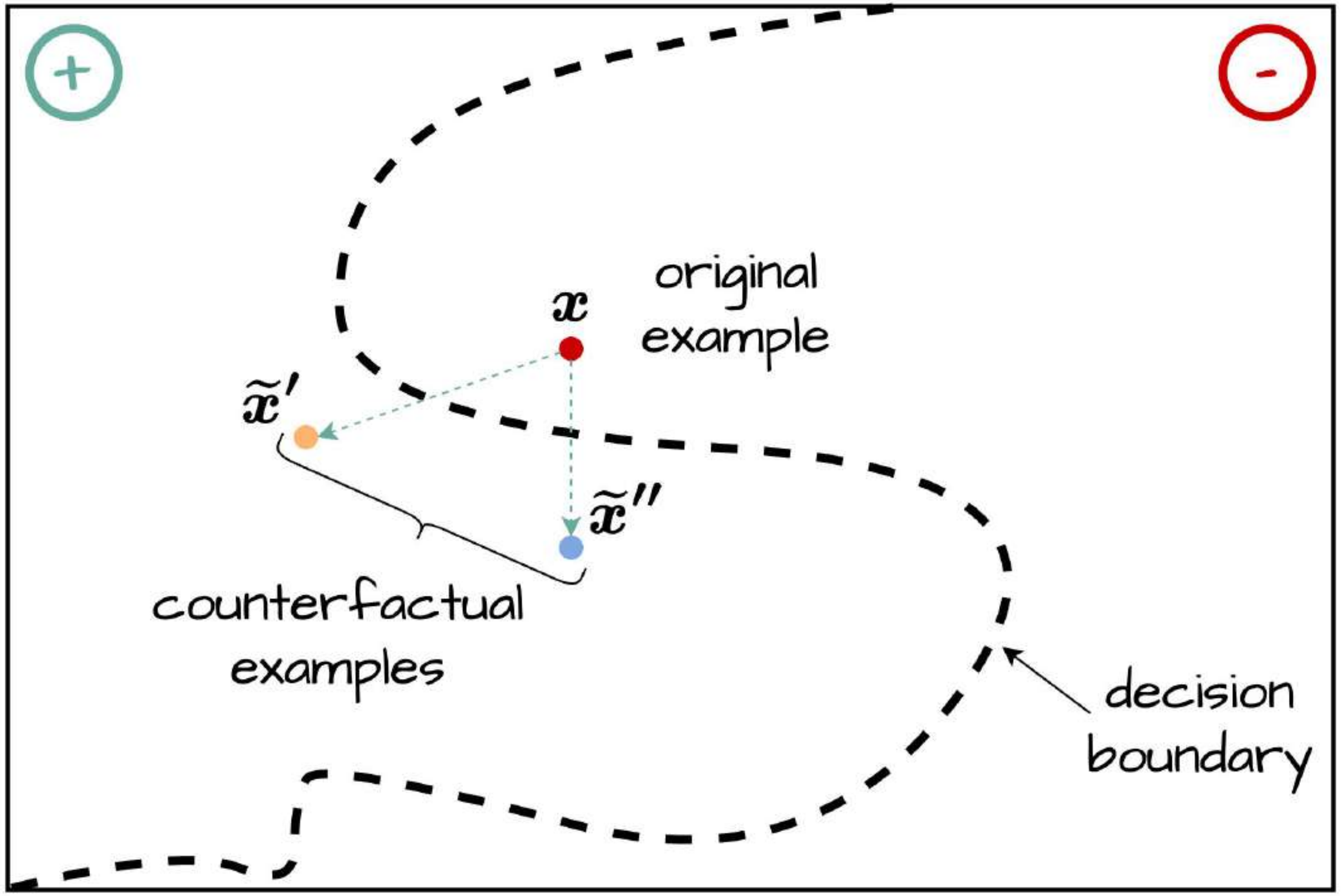
| **Age** | **Gender** | **Exercise Level** | **Fat Level** |
|---|---|---|---|
| 45 | M | **Medium** | **Low** |

$f($  $)=$ **No**

**Explanation:**

You will ***not*** develop diabetes if you ***increase your exercise level*** and ***lower your fat level***

# Finding Counterfactual Examples (CFs)




Given an input sample ***x***, there may be (infinitely?) many counterfactual examples

We need to restrict our search to “some” of them!

# Finding the "Optimal" CF (for a given *x*)

$$\tilde{\boldsymbol{x}}^* = \operatorname{argmin}_{\tilde{\boldsymbol{x}}}\{\boxed{\ell_{\mathrm{CF}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}; f)} + \lambda \ell_{\mathrm{dist}}(\boldsymbol{x}, \tilde{\boldsymbol{x}})\}$$

**counterfactual loss**
penalizes if the CF goal
is **not** met

CF goal: $f(\tilde{\boldsymbol{x}}) \neq f(\boldsymbol{x})$

# Finding the “Optimal” CF (for a given ***x***)

$$\tilde{\boldsymbol{x}}^* = \operatorname{argmin}_{\tilde{\boldsymbol{x}}}\{\ell_{\mathrm{CF}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}; f) + \lambda \boxed{\ell_{\mathrm{dist}}(\boldsymbol{x}, \tilde{\boldsymbol{x}})}\}$$

**distance loss**
discourages the CF to be too far away from the original input ***x***

e.g., *L1*-norm $|\tilde{\boldsymbol{x}} - \boldsymbol{x}|$

# Finding the "Optimal" CF (for a given ***x***)

$$\tilde{\boldsymbol{x}}^* = \mathrm{argmin}_{\tilde{\boldsymbol{x}}}\{\ell_{\mathrm{CF}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}; f) + \lambda \ell_{\mathrm{dist}}(\boldsymbol{x}, \tilde{\boldsymbol{x}})\}$$

$$\text{s.t.: } 1 \leq p_{\max} \leq m, \text{ where } 1 \leq m \leq |\mathcal{F}| \leq n$$

Limit on the number of "actionable" features to change

Set of "actionable" features

# Evaluation Metrics for CFs

**Validity (1-Fidelity)**
Measures the ratio of generated CFs that actually meet the counterfactual goal (**the higher the better**)

# Evaluation Metrics for CFs

**Validity (1-Fidelity)**
Measures the ratio of generated CFs that actually meet the counterfactual goal (**the higher the better**)

**Proximity**
Computes the distance between a (valid) CF and the original input sample (**the lower the better**)

*L1*-norm or *L2*-norm

# Evaluation Metrics for CFs

**Validity (1-Fidelity)**
Measures the ratio of generated CFs that actually meet the counterfactual goal
(**the higher the better**)

**Proximity**
Computes the distance between a (valid) CF and the original input sample
(**the lower the better**)

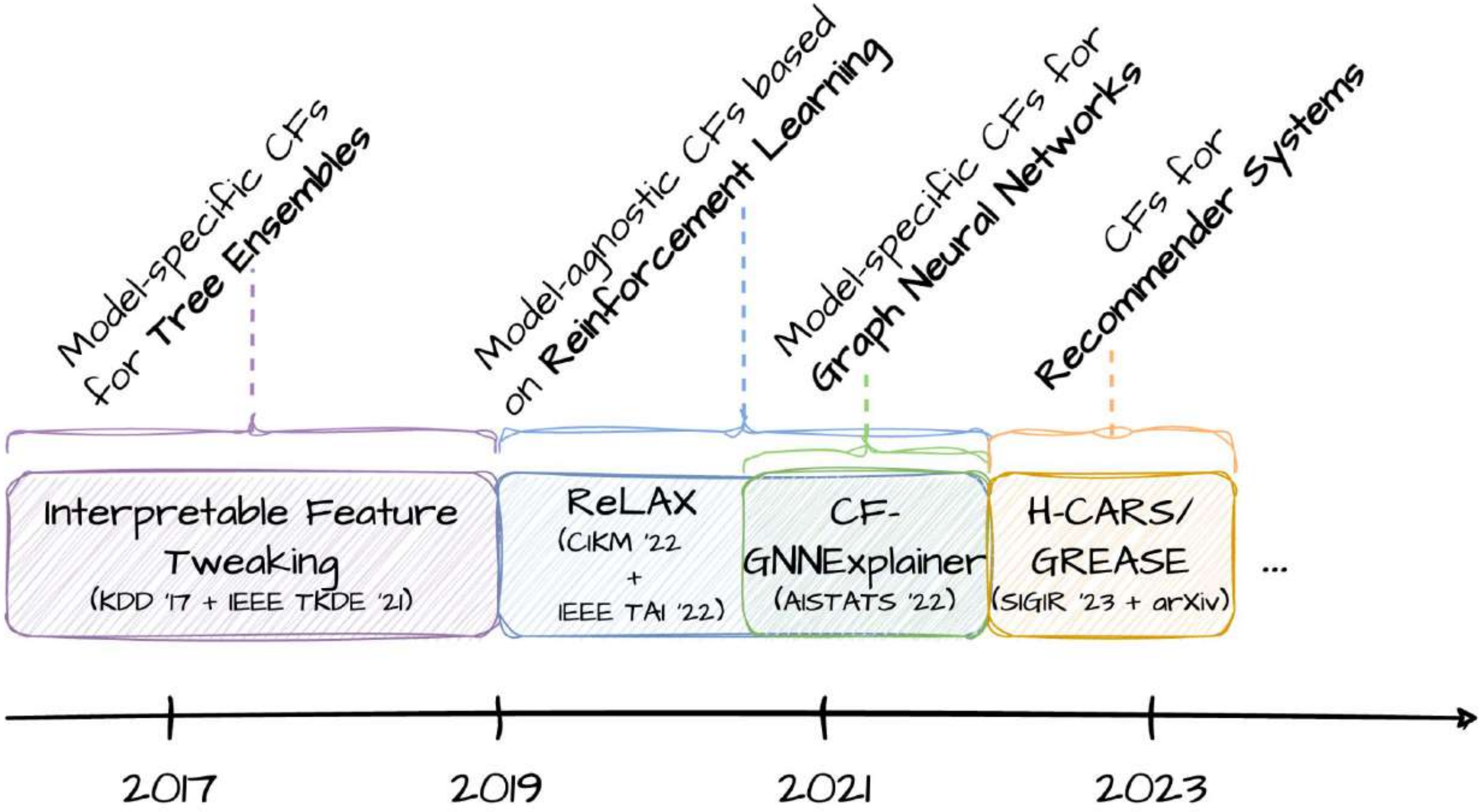
*L1*-norm or *L2*-norm

**Sparsity**
Indicates the number of features modified to obtain the CF
(**the lower the better**)

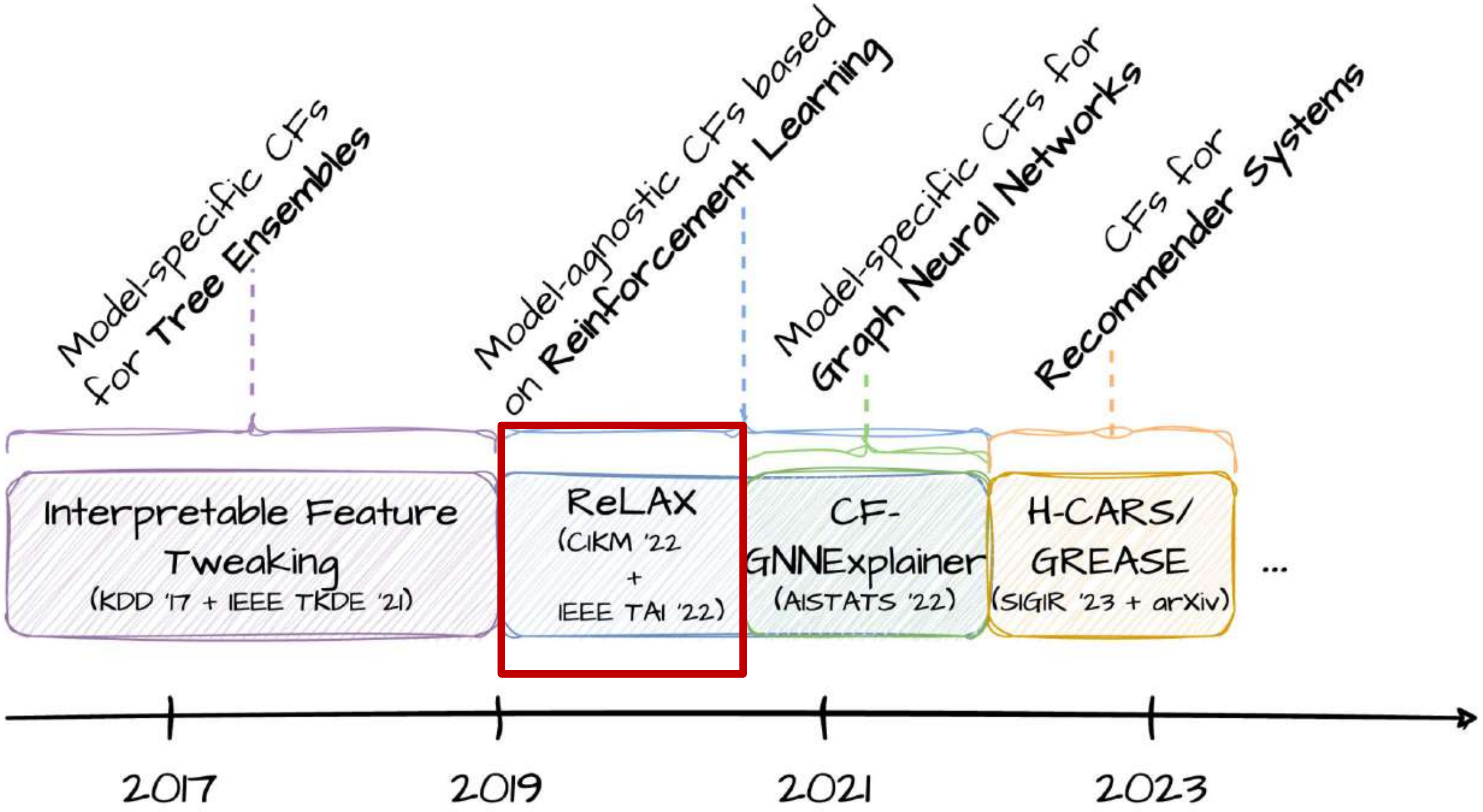
*L0*-norm

Sahil, V., Dickerson, J. and Hines, K., 2022. Counterfactual Explanations for Machine Learning: A Review. arXiv:2010.10596.

# Our Contributions to CF Explanations



# Our Contributions to CF Explanations

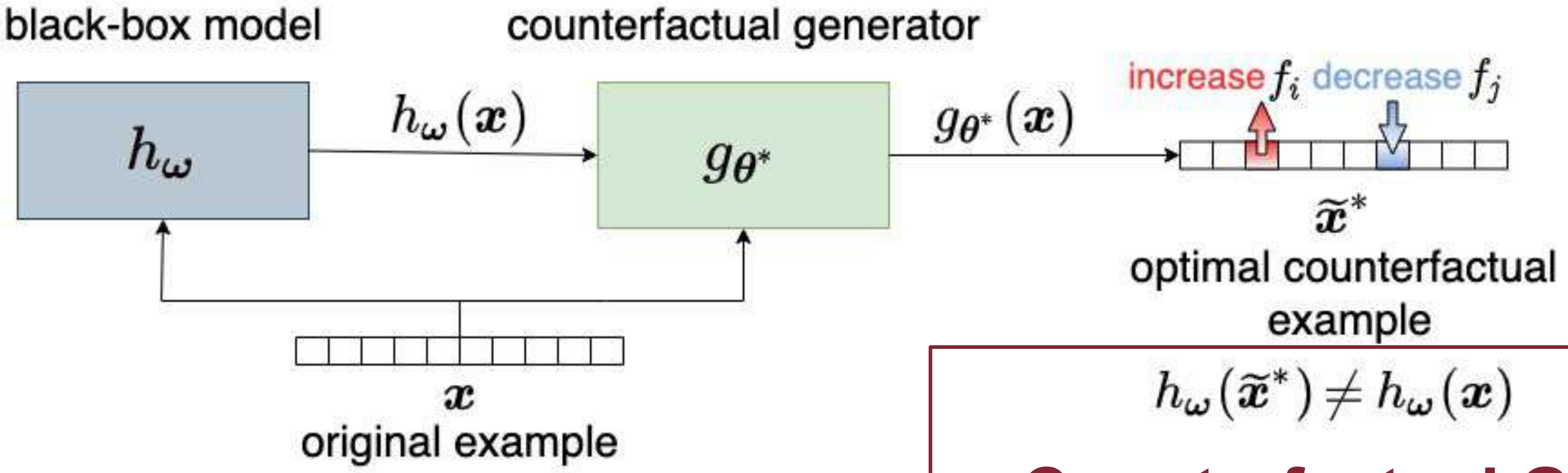


# ReLAX:
# Reinforcement Learning Agent Explainer for Arbitrary Predictive Models

Chen, Z., Silvestri, F., Wang, J., Zhu, H., Ahn, H. and Tolomei, G., 2022, October. ReLAX: Reinforcement Learning Agent Explainer for Arbitrary Predictive Models. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (pp. 252-261).

Chen, Z., Silvestri, F., Tolomei, G., Wang, J., Zhu, H. and Ahn, H., 2022. Explain the Explainer: Interpreting Model-Agnostic Counterfactual Explanations of a Deep Reinforcement Learning Agent. IEEE Transactions on Artificial Intelligence.

# Finding the “Optimal” CF Generator

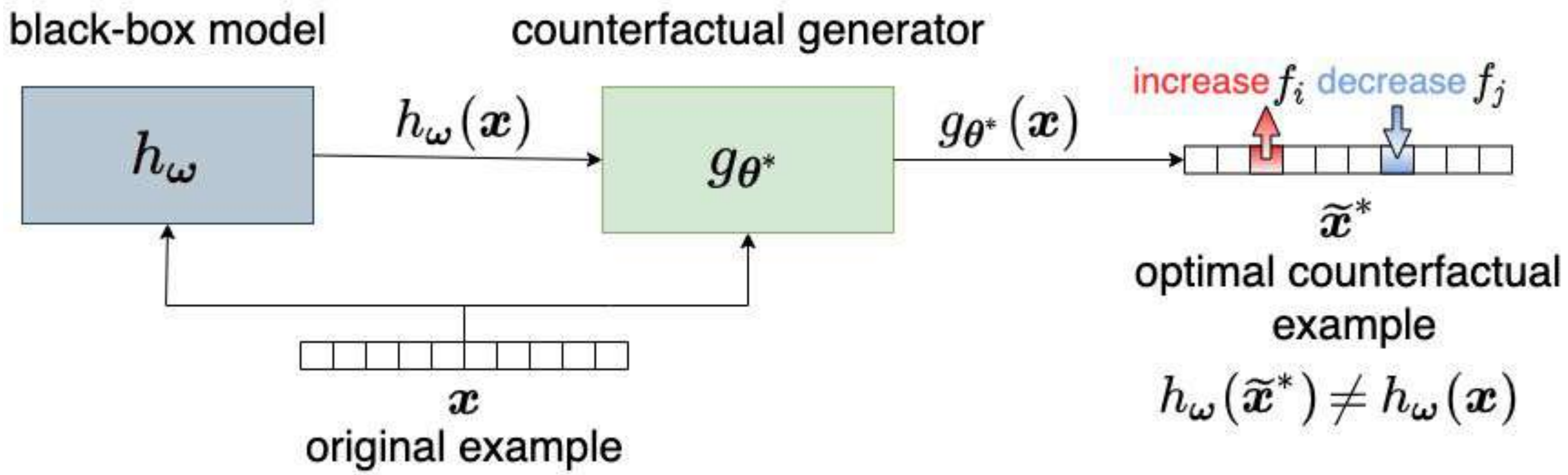



$$h_\omega(\tilde{\boldsymbol{x}}^*) \neq h_\omega(\boldsymbol{x})$$

**Counterfactual Goal**

works both for classification and regression tasks

# Finding the "Optimal" CF Generator




**How Do We Find $g_{\theta^*}$?**

# Finding the "Optimal" CF Generator

$$\boldsymbol{\theta}^* = \mathrm{argmin}_{\boldsymbol{\theta}} \Big\{ \mathcal{L}(g_{\boldsymbol{\theta}}; \mathcal{D}, h_{\boldsymbol{\omega}}) \Big\}$$
$$\text{s.t.: } p_{\max} \leq m$$

$$\mathcal{L}(g_{\boldsymbol{\theta}}; \mathcal{D}, h_{\boldsymbol{\omega}}) = \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \ell_{\mathrm{CF}}(\boldsymbol{x}, g_{\boldsymbol{\theta}}(\boldsymbol{x}); h_{\boldsymbol{\omega}}) + \lambda \ell_{\mathrm{dist}}(\boldsymbol{x}, g_{\boldsymbol{\theta}}(\boldsymbol{x}))$$

from *instance*-level (**local**) to *dataset*-level (**global**) explanations

# Finding the "Optimal" CF Generator

$$\boldsymbol{\theta}^* = \mathrm{argmin}_{\boldsymbol{\theta}} \Big\{ \mathcal{L}(g_{\boldsymbol{\theta}}; \mathcal{D}, h_{\boldsymbol{\omega}}) \Big\}$$
$$\text{s.t.: } p_{\max} \leq m$$

$$\mathcal{L}(g_{\boldsymbol{\theta}}; \mathcal{D}, h_{\boldsymbol{\omega}}) = \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \ell_{\mathrm{CF}}(\boldsymbol{x}, g_{\boldsymbol{\theta}}(\boldsymbol{x}); h_{\boldsymbol{\omega}}) + \lambda \ell_{\mathrm{dist}}(\boldsymbol{x}, g_{\boldsymbol{\theta}}(\boldsymbol{x}))$$

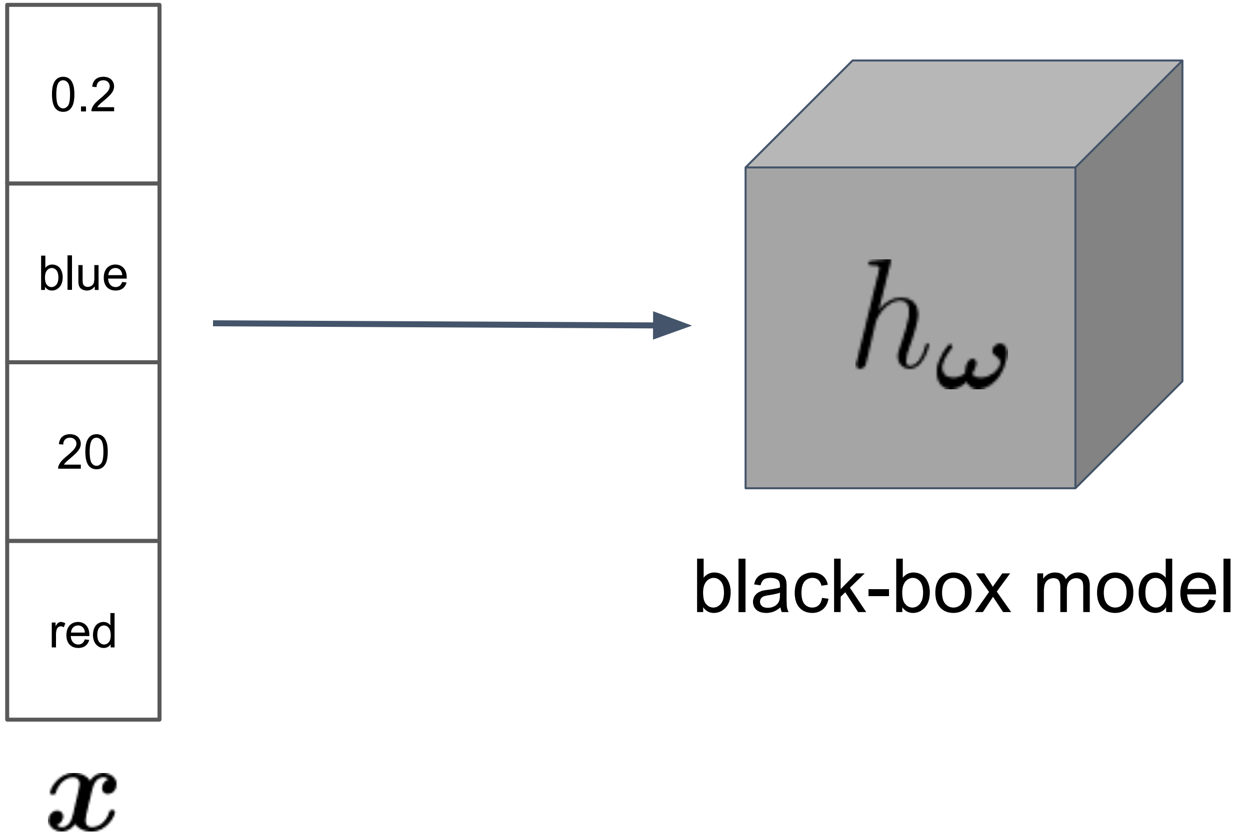
from **optimizing** to **learning**

# ReLAX: Intuition

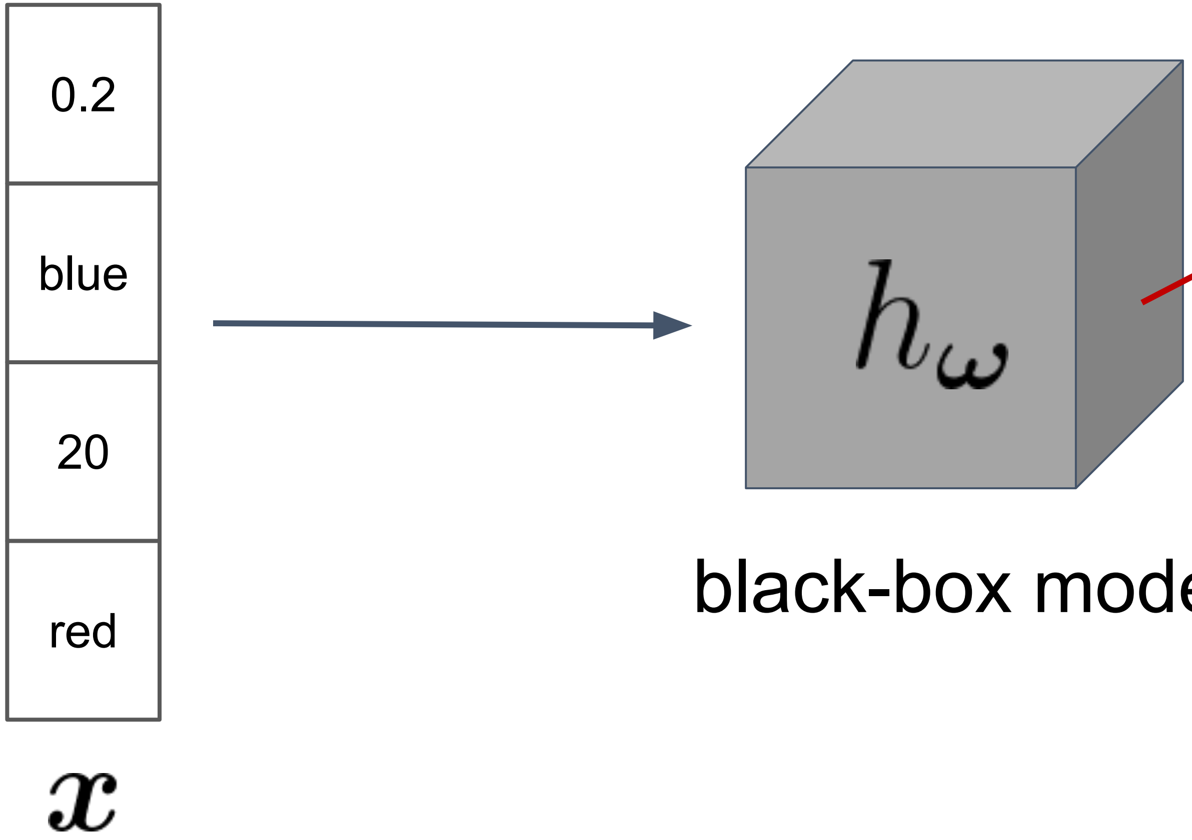
| |
|---|
| 0.2 |
| blue |
| 20 |
| red |

$x$

# ReLAX: Intuition



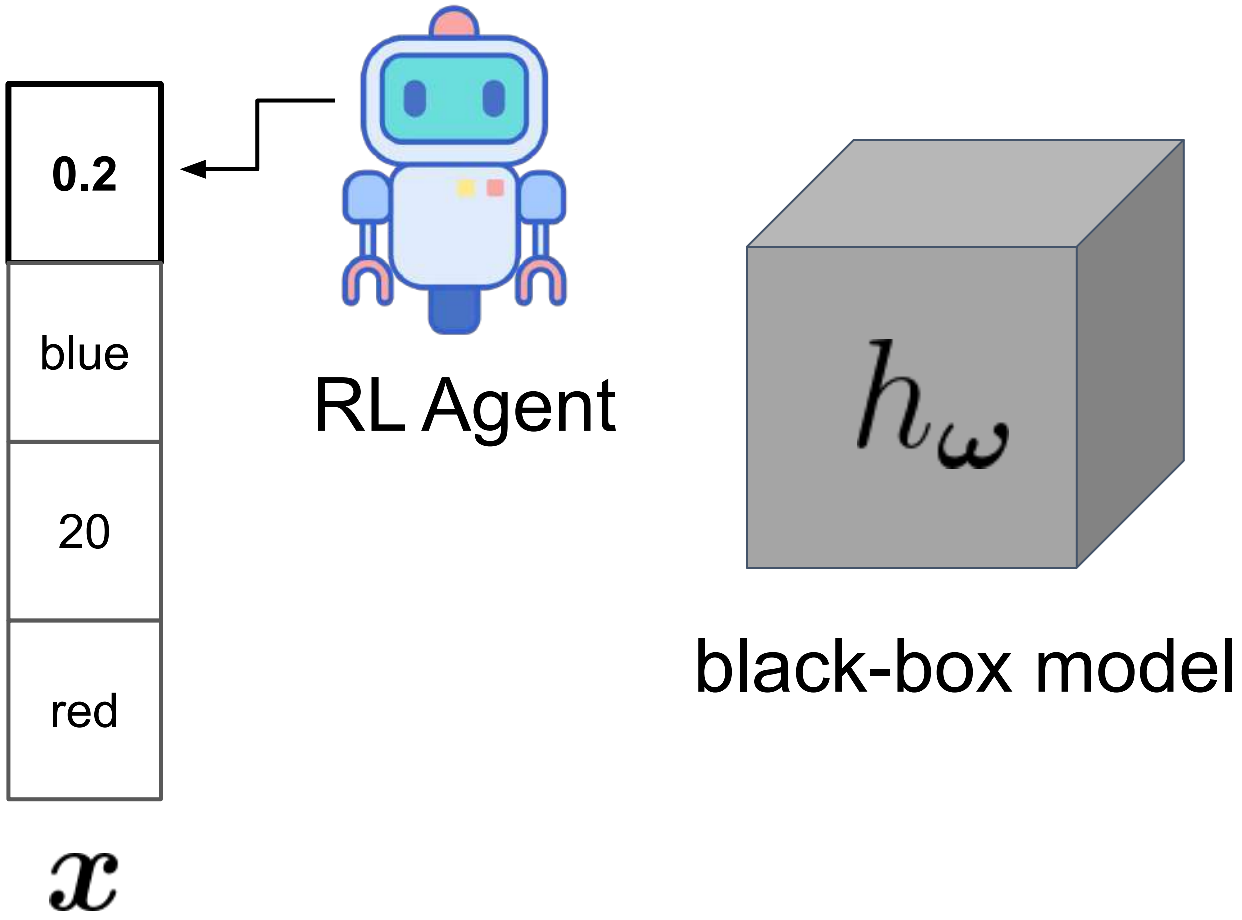
# ReLAX: Intuition




black-box model

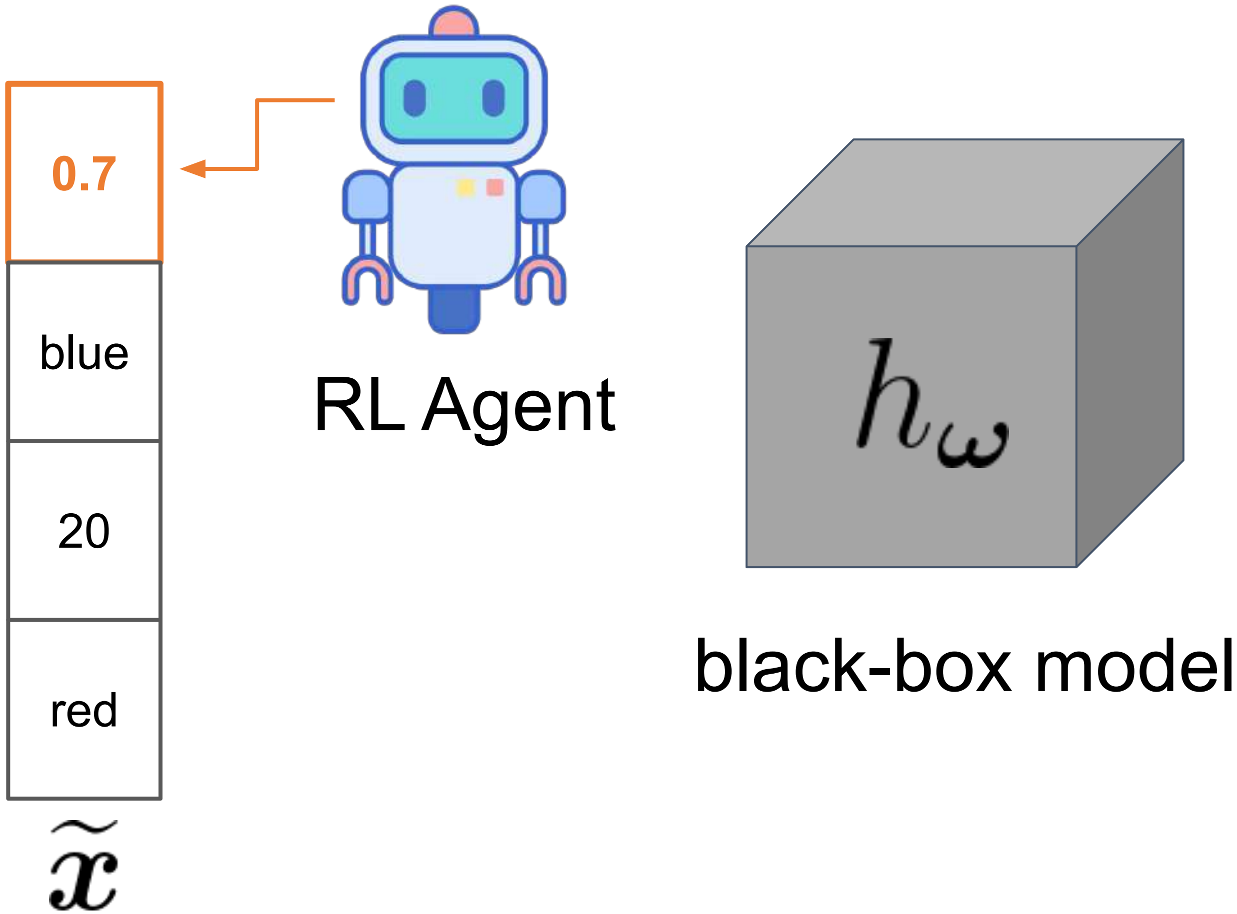
$$\boxed{h_\omega(\boldsymbol{x}) = -1}$$

# ReLAX: Intuition



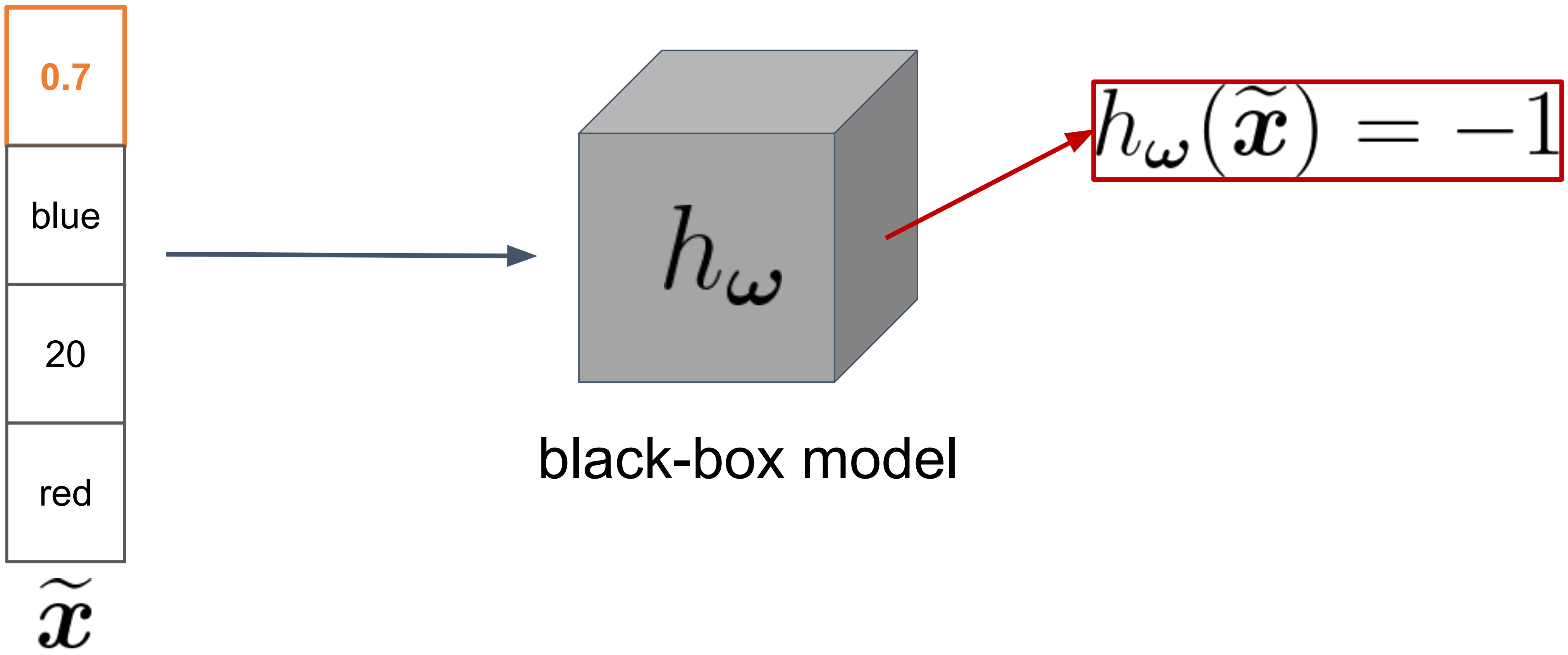

**1)** The RL Agent picks a feature to modify

# ReLAX: Intuition

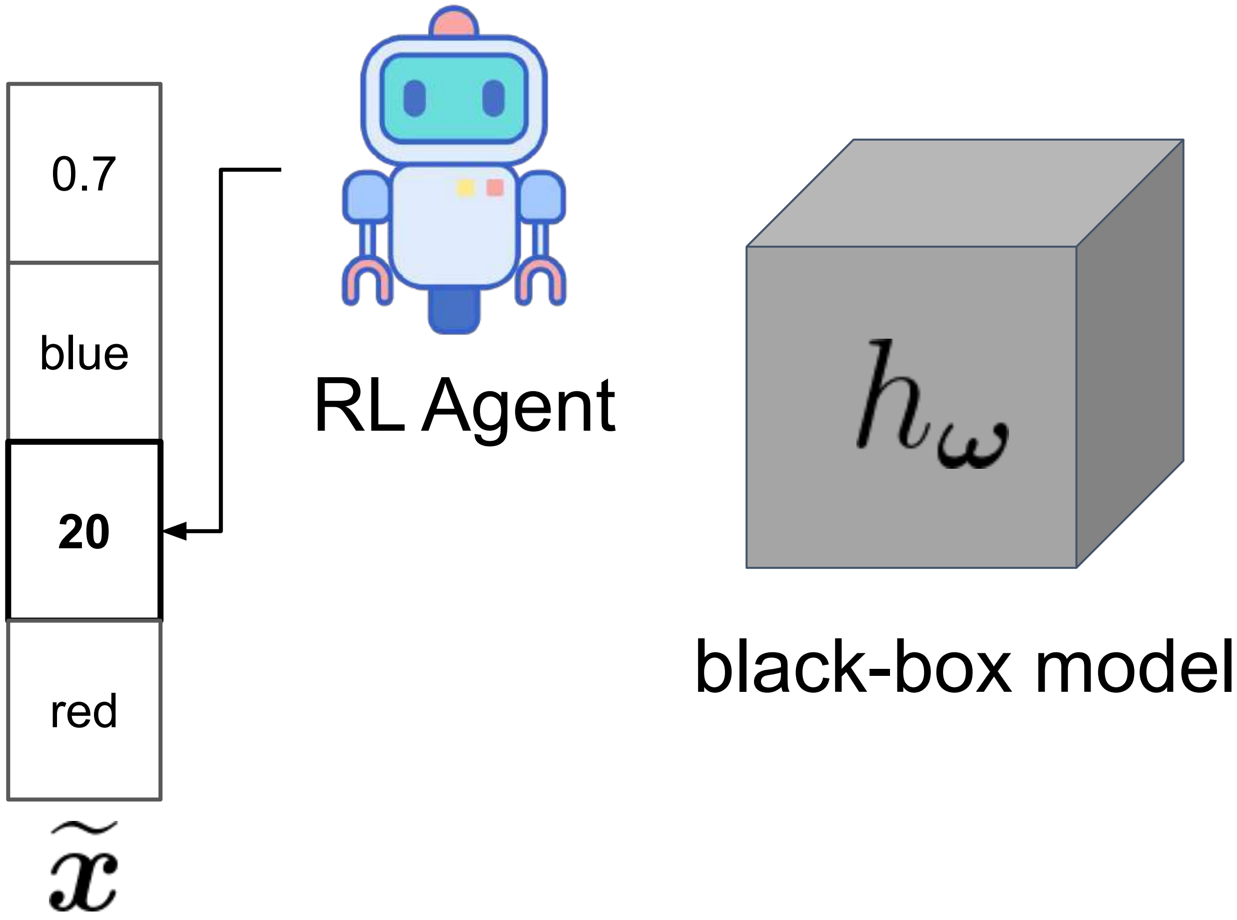



**2)** The RL Agent chooses the magnitude of the feature change

# ReLAX: Intuition

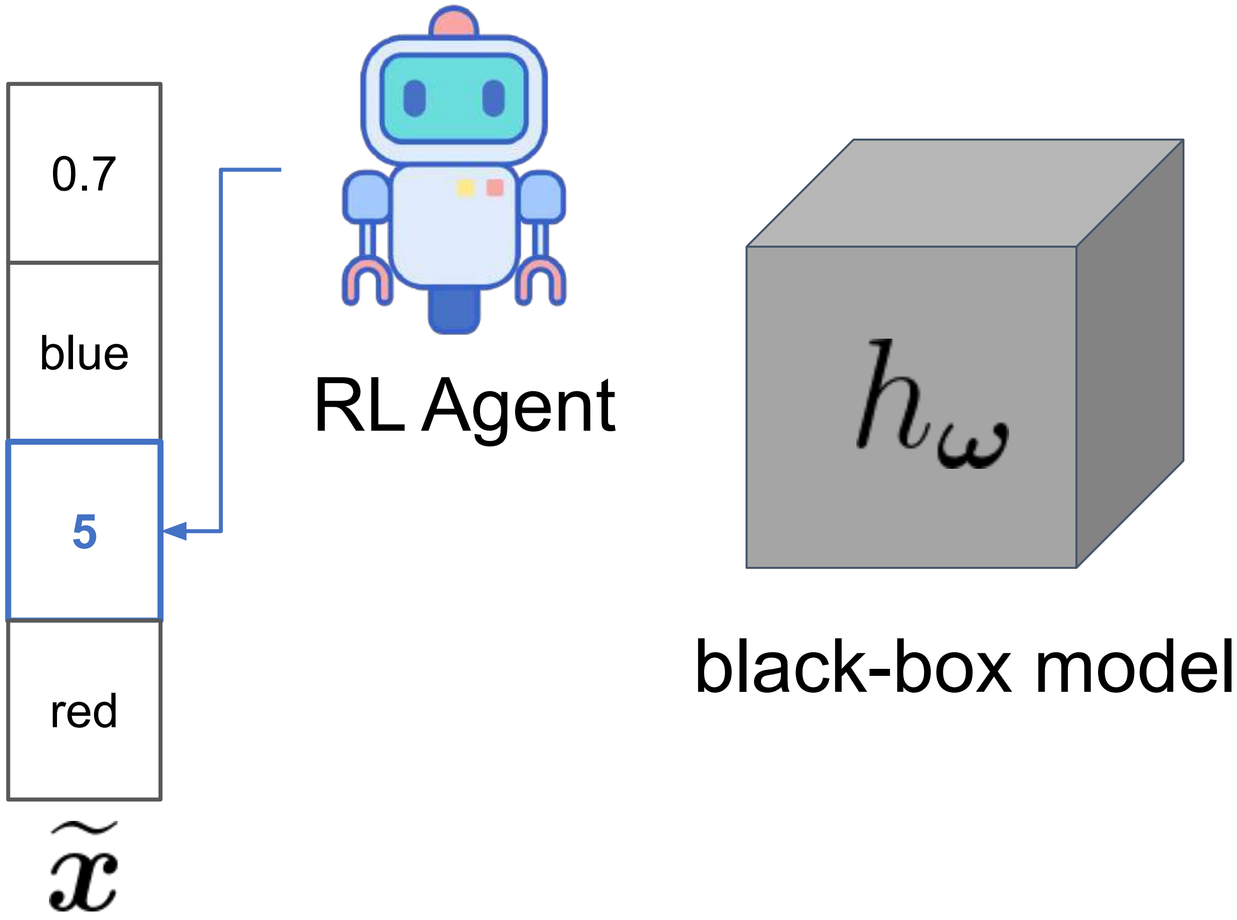


# ReLAX: Intuition



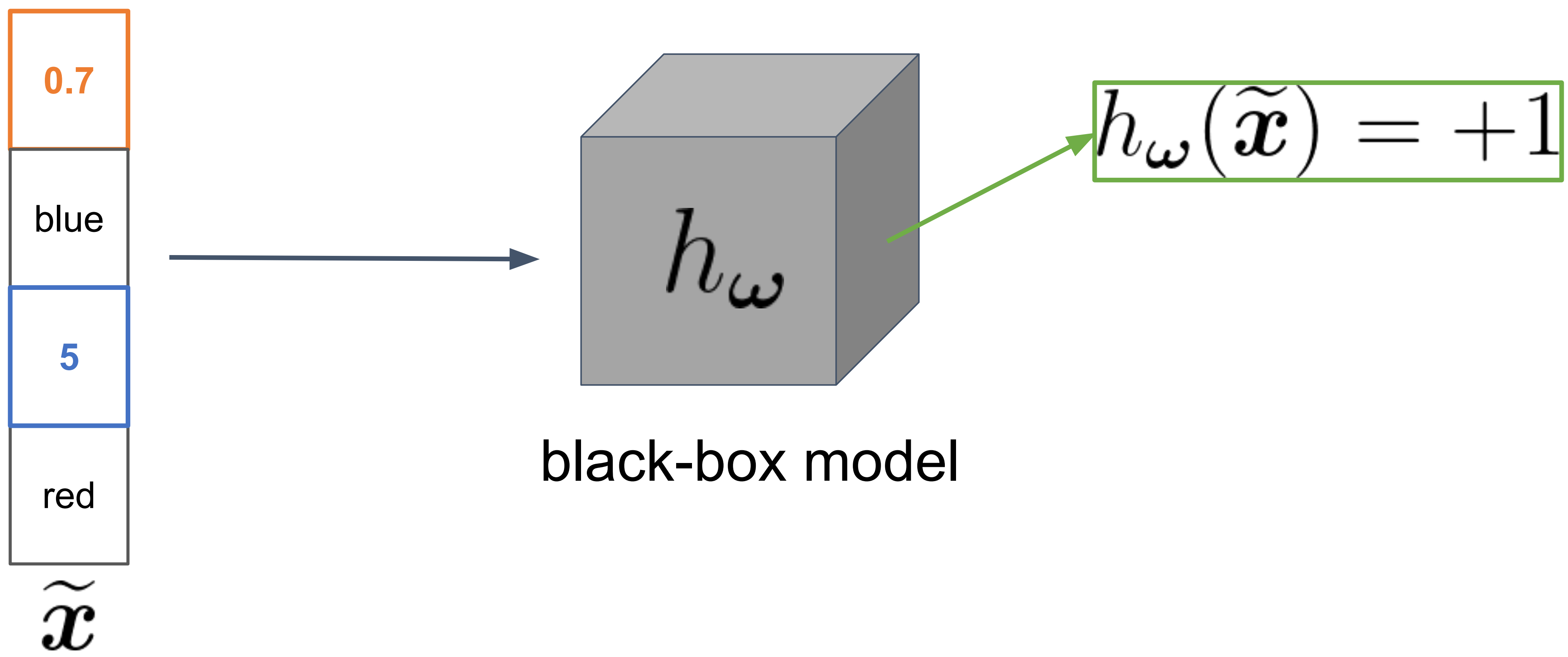

**1)** The RL Agent picks a feature to modify

# ReLAX: Intuition

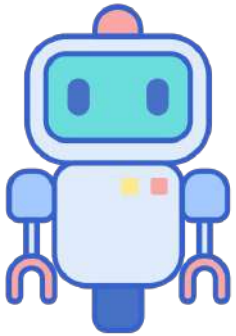



**2)** The RL Agent chooses the magnitude of the feature change

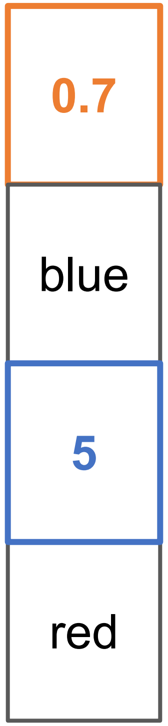
# ReLAX: Intuition



# ReLAX: Intuition



RL Agent




$\widetilde{x}$

The RL Agent terminates when the CF goal is met!

# Markov Decision Process Formulation

We formulate the problem of finding the optimal CF generator $g_{\boldsymbol{\theta}^*}$ as an MDP

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, p_0, r, \gamma\}$$

# Markov Decision Process Formulation

We formulate the problem of finding the optimal CF generator $g_{\theta^*}$ as an MDP

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, p_0, r, \gamma\}$$

***States*** $\mathcal{S}$ The current modified sample along with the features changed so far

# Markov Decision Process Formulation

We formulate the problem of finding the optimal CF generator $g_{\theta^*}$ as an MDP

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, p_0, r, \gamma\}$$

***States*** $\mathcal{S}$ The current modified sample along with the features changed so far

***Actions*** $\mathcal{A}$ Discrete-Continuous Hybrid Actions: *Which* feature, *What* change

# Markov Decision Process Formulation

We formulate the problem of finding the optimal CF generator $g_{\theta^*}$ as an MDP

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, p_0, r, \gamma\}$$

***States*** $\mathcal{S}$ The current modified sample along with the features changed so far

***Actions*** $\mathcal{A}$ Discrete-Continuous Hybrid Actions: *Which* feature, *What* change

***Transitions*** $\mathcal{T}$ Deterministic function moving from one state to another

# Markov Decision Process Formulation

We formulate the problem of finding the optimal CF generator $g_{\theta^*}$ as an MDP

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, p_0, r, \gamma\}$$

***States*** $\mathcal{S}$ The current modified sample along with the features changed so far

***Actions*** $\mathcal{A}$ Discrete-Continuous Hybrid Actions: *Which* feature, *What* change

***Transitions*** $\mathcal{T}$ Deterministic function moving from one state to another

***Reward*** $r$ Trade-off between CF goal and the distance of the CF from the original input

# Markov Decision Process Formulation

We formulate the problem of finding the optimal CF generator $g_{\theta^*}$ as an MDP

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, p_0, r, \gamma\}$$

***States*** $\mathcal{S}$ The current modified sample along with the features changed so far

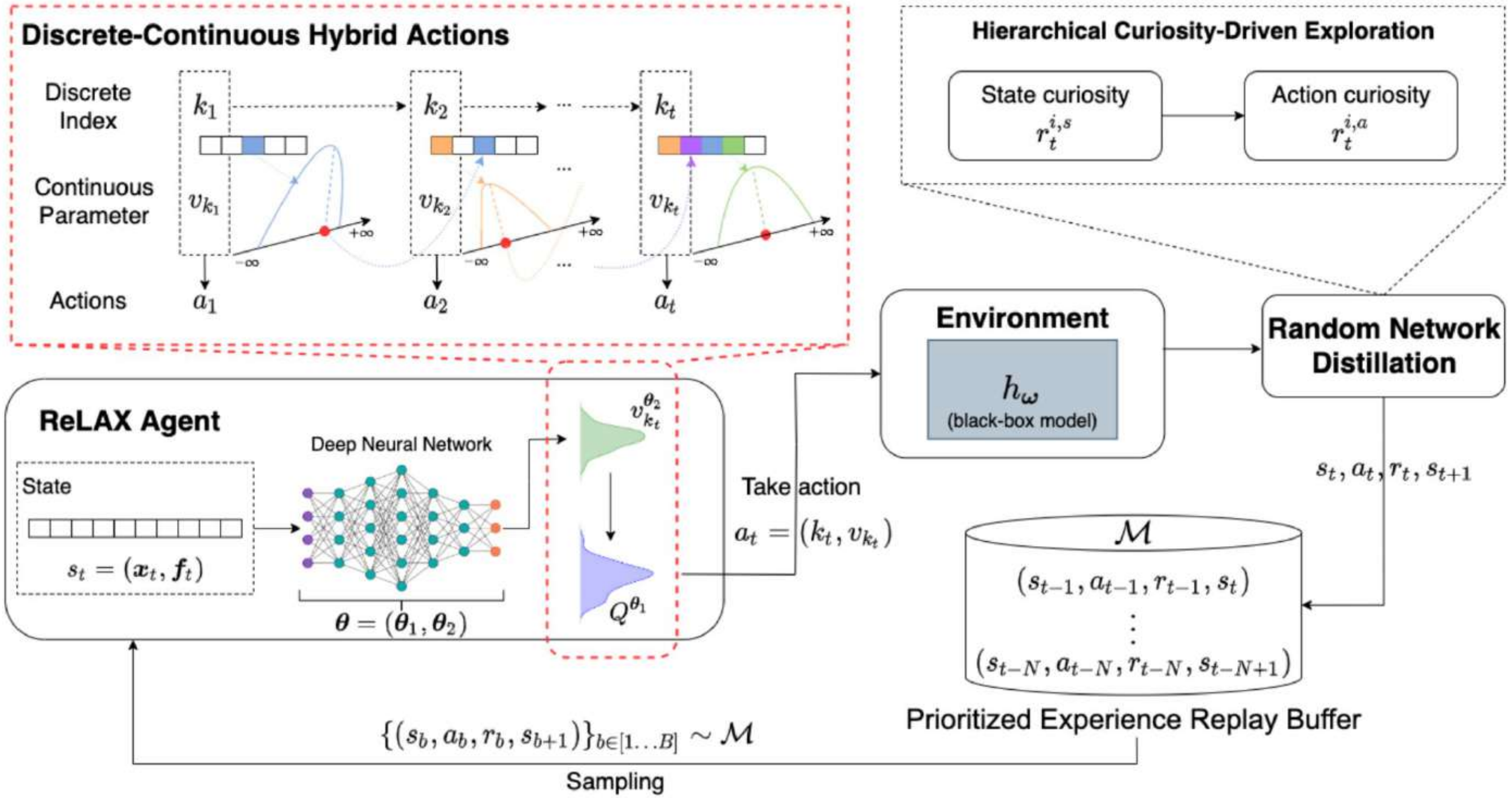
***Actions*** $\mathcal{A}$ Discrete-Continuous Hybrid Actions: *Which* feature, *What* change

***Transitions*** $\mathcal{T}$ Deterministic function moving from one state to another

***Reward*** $r$ Trade-off between CF goal and the distance of the CF from the original input

We find the **optimal policy** to apply the best sequence of actions to each input

# Our Proposed Framework



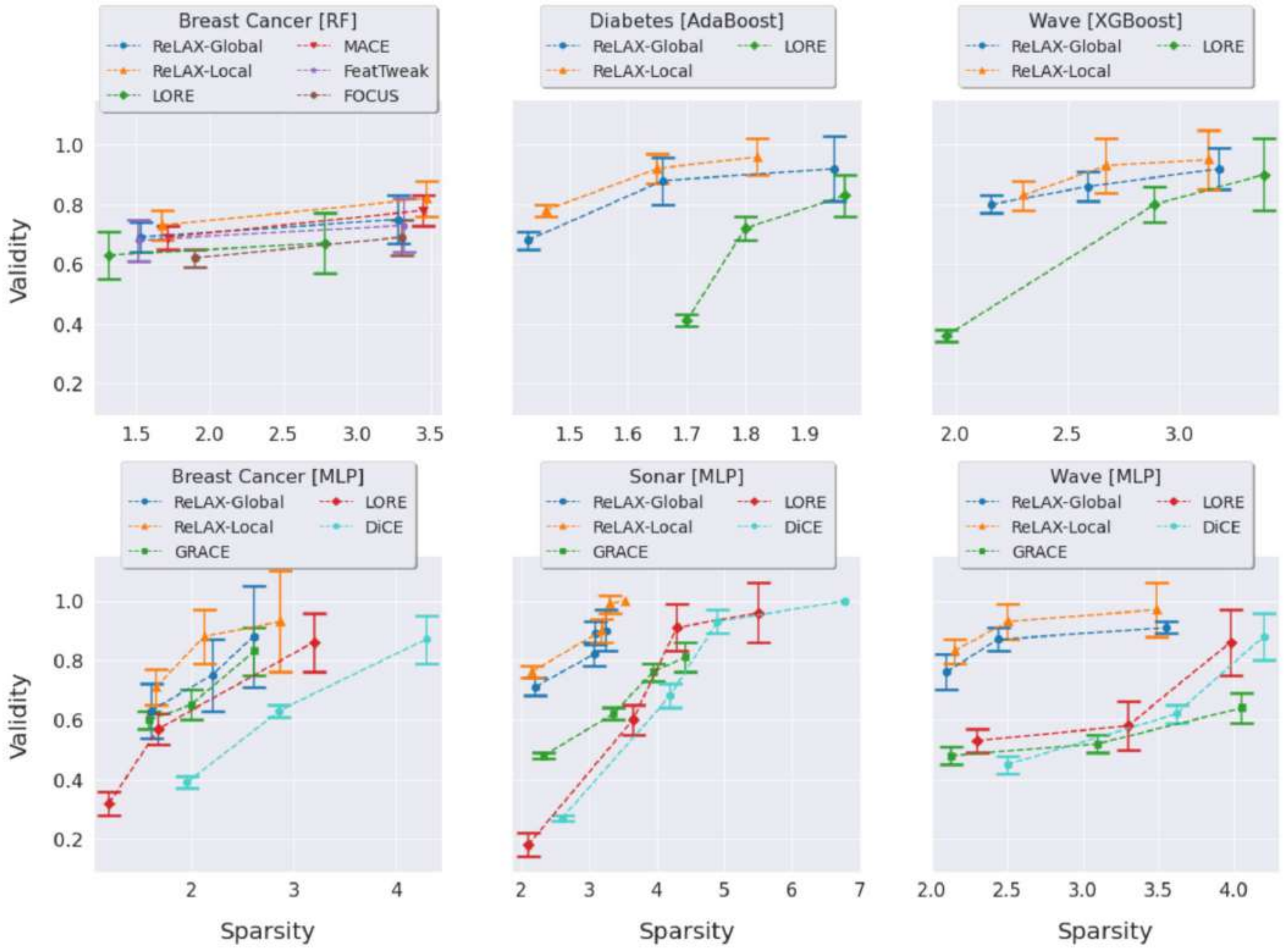
# Experiments: Datasets and Tasks

| **Dataset** | **N. of Instances** | **N. of Features** | **Task** |
|---|---|---|---|
| *Breast Cancer* [5] | 699 | 10 (numerical) | classification |
| *Diabetes* [2] | 768 | 8 (numerical) | classification |
| *Sonar* [3] | 208 | 60 (numerical) | classification |
| *Wave* [4] | 5,000 | 21 (numerical) | classification |
| *Boston Housing* [1] | 506 | 14 (mixed) | regression |

# Experiments: (Black-Box) Models

| Dataset [Best Model] | Structure | Acc. (▲)/RMSE (♦) |
|---|---|---|
| *Breast Cancer* [RF] | {#trees=100} | 0.99 (▲) |
| *Diabetes* [AdaBoost] | {#trees=100} | 0.79 (▲) |
| *Wave* [XGBoost] | {#trees=100} | 0.95 (▲) |
| *Breast Cancer* [MLP] | {#L1=64, #L2=128} | 1.00 (▲) |
| *Sonar* [MLP] | {#L1=256, #L2=256} | 0.90 (▲) |
| *Wave* [MLP] | {#L1=100, #L2=200} | 0.97 (▲) |
| *Boston Housing* [MLP-Reg] | {#L1=50, #L2=128} | 3.36 (♦) |

# Experiments: Sparsity vs. Validity




ReLAX achieves the **best trade-off** between sparsity and validity of the generated CFs

# Sparsity vs. Validity: The Boston Housing

| | *Validity (Sparsity)* | |
|---|---|---|
| **Threshold ($\delta$)** | ReLAX-Global | ReLAX-Local |
| 0.20 | $0.81 \pm 0.09$ ($\mathbf{3.02} \pm 0.17$) | $\mathbf{0.87} \pm 0.05$ ($3.10 \pm 0.18$) |
| 0.40 | $0.74 \pm 0.06$ ($\mathbf{3.09} \pm 0.16$) | $\mathbf{0.81} \pm 0.05$ ($3.18 \pm 0.16$) |
| 0.60 | $0.70 \pm 0.06$ ($\mathbf{3.21} \pm 0.12$) | $\mathbf{0.77} \pm 0.03$ ($3.28 \pm 0.09$) |

**Dataset-level Explainer**

# Sparsity vs. Validity: The Boston Housing

| | *Validity (Sparsity)* | |
|---|---|---|
| **Threshold ($\delta$)** | ReLAX-Global | ReLAX-Local |
| 0.20 | 0.81 ± 0.09 (**3.02** ± 0.17) | **0.87** ± 0.05 (3.10 ± 0.18) |
| 0.40 | 0.74 ± 0.06 (**3.09** ± 0.16) | **0.81** ± 0.05 (3.18 ± 0.16) |
| 0.60 | 0.70 ± 0.06 (**3.21** ± 0.12) | **0.77** ± 0.03 (3.28 ± 0.09) |

**Instance-level Explainer**

# Sparsity vs. Validity: The Boston Housing

| | *Validity (Sparsity)* | |
|---|---|---|
| **Threshold ($\delta$)** | ReLAX-Global | ReLAX-Local |
| 0.20 | $0.81 \pm 0.09$ (**3.02** $\pm$ 0.17) | **0.87** $\pm$ 0.05 (3.10 $\pm$ 0.18) |
| 0.40 | $0.74 \pm 0.06$ (**3.09** $\pm$ 0.16) | **0.81** $\pm$ 0.05 (3.18 $\pm$ 0.16) |
| 0.60 | $0.70 \pm 0.06$ (**3.21** $\pm$ 0.12) | **0.77** $\pm$ 0.03 (3.28 $\pm$ 0.09) |

In the case of regression task, the CF goal must be adapted with a **validity threshold** ($\delta$): $|h_{\omega}(\widetilde{\boldsymbol{x}}) - h_{\omega}(\boldsymbol{x})| \geq \delta, \ \delta \in \mathbb{R}_{>0}$

# Sparsity vs. Validity: The Boston Housing

| | *Validity (Sparsity)* | |
|---|---|---|
| **Threshold ($\delta$)** | ReLAX-Global | ReLAX-Local |
| 0.20 | $0.81 \pm 0.09$ ($\mathbf{3.02} \pm 0.17$) | $\mathbf{0.87} \pm 0.05$ ($3.10 \pm 0.18$) |
| 0.40 | $0.74 \pm 0.06$ ($\mathbf{3.09} \pm 0.16$) | $\mathbf{0.81} \pm 0.05$ ($3.18 \pm 0.16$) |
| 0.60 | $0.70 \pm 0.06$ ($\mathbf{3.21} \pm 0.12$) | $\mathbf{0.77} \pm 0.03$ ($3.28 \pm 0.09$) |

In the case of regression task, the CF goal must be adapted with a **validity threshold** ($\delta$): $|h_\omega(\widetilde{\boldsymbol{x}}) - h_\omega(\boldsymbol{x})| \geq \delta,\ \delta \in \mathbb{R}_{>0}$

The higher the threshold the harder is for ReLAX to find a valid CF

# Experiments: Proximity vs. Generation Time

| Metric | Dataset [Models] | CF Generation Methods | | | |
|---|---|---|---|---|---|
| | | ReLAX-Global | ReLAX-Local | LORE | MACE |
| ***Proximity*** | *Breast Cancer* [RF, MLP] | [4.46, 5.92] | [4.49, 5.87] | [4.63, 5.63] | [4.47, N/A] |
| | *Diabetes* [AdaBoost] | [4.41] | [4.50] | [4.76] | [N/A] |
| | *Sonar* [MLP] | [7.32] | [7.66] | [7.36] | [N/A] |
| | *Wave* [XGBoost, MLP] | [5.93, 6.38] | [6.02, 6.50] | [6.60, 6.41] | [N/A, N/A] |
| | *Boston Housing* [MLP-Reg] | [5.10] | [5.36] | [N/A] | [N/A] |
| ***Generation Time* (secs.)** | * | 1500 | **1320** | 2100 | 2280 |

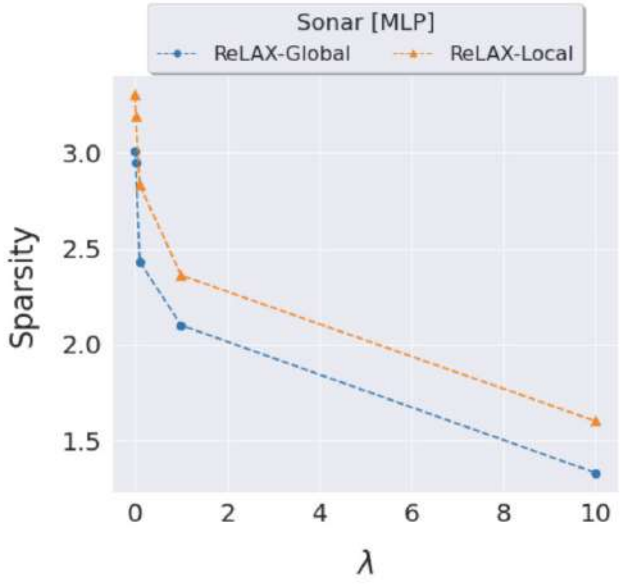
ReLAX-Global generates CFs that are closer to the original input instance but ReLAX-Local takes less time on average

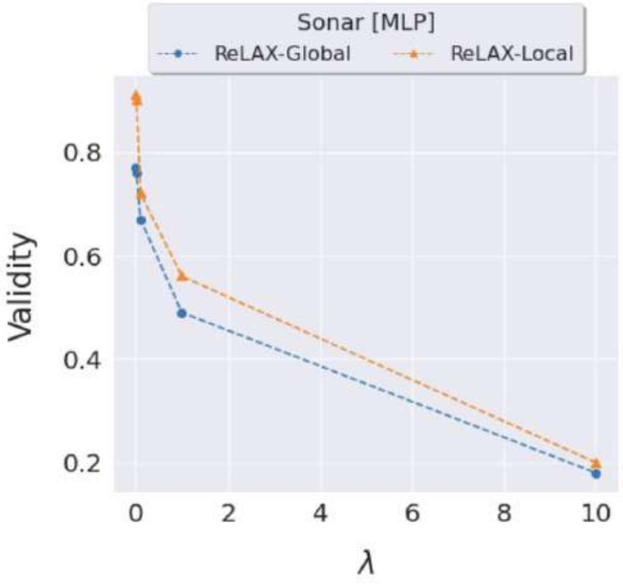
# Experiments: The Hyperparameter λ

λ controls the balance between sparsity and validity

# Experiments: The Hyperparameter λ

λ controls the balance between sparsity and validity




Larger values of λ force the agent to prefer sparser CFs at the expense of lower validity

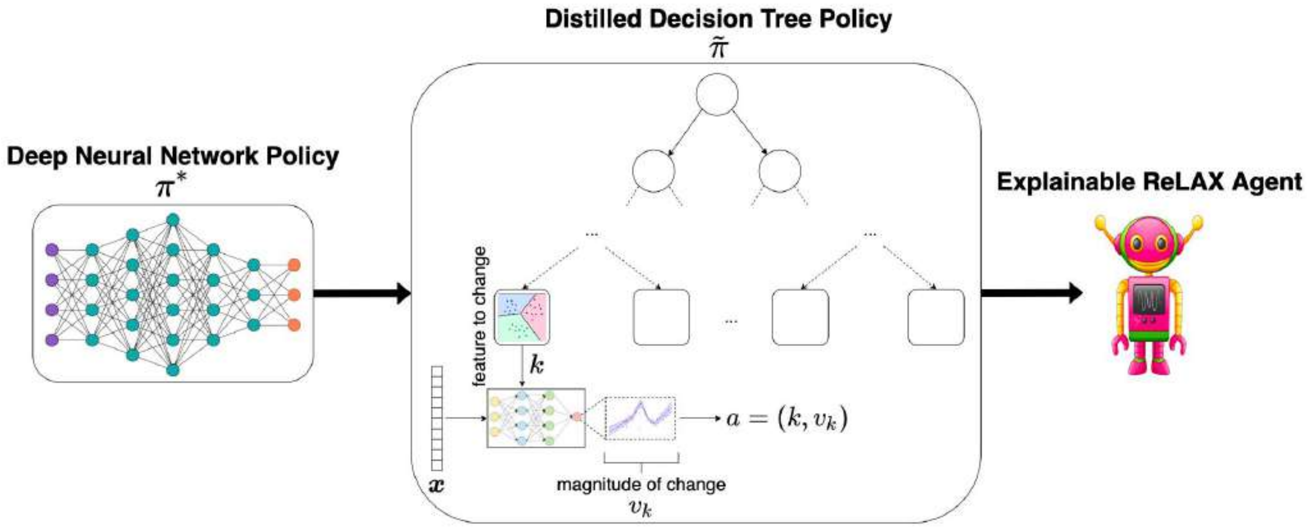
# Who Explains the Explainer?

The complex network structure of a DRL policy learned for CF generation poses a challenge for understanding the decision logic of the agent

# Who Explains the Explainer?

The complex network structure of a DRL policy learned for CF generation poses a challenge for understanding the decision logic of the agent

To explain the decision process of a learned policy, we **distill** knowledge from the policy to a naturally-interpretable **decision tree**



# Use Case: COVID-19 (Risk of Mortality)

We apply ReLAX to generate CF explanations for a binary classifier (XGBoost with 500 trees) trained to predict the risk of mortality for COVID-19

# Use Case: COVID-19 (Risk of Mortality)

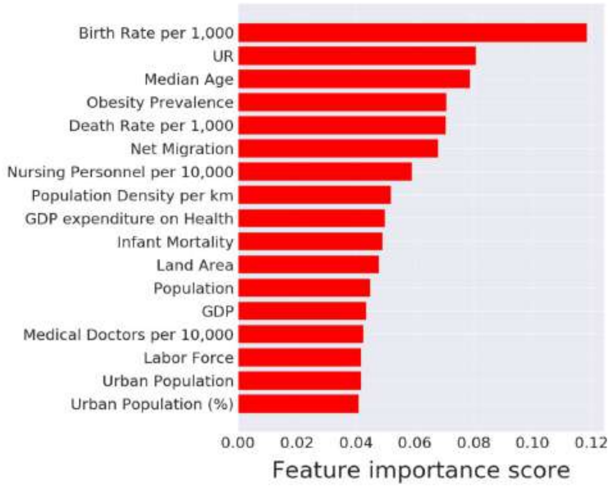
We apply ReLAX to generate CF explanations for a binary classifier (XGBoost with 500 trees) trained to predict the risk of mortality for COVID-19

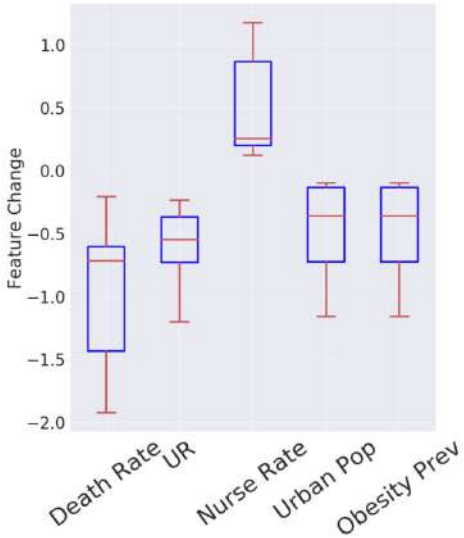
We use generated CFs to sketch an action plan to lower the risk of mortality

# Use Case: COVID-19 (Risk of Mortality)

We apply ReLAX to generate CF explanations for a binary classifier (XGBoost with 500 trees) trained to predict the risk of mortality for COVID-19

We use generated CFs to sketch an action plan to lower the risk of mortality



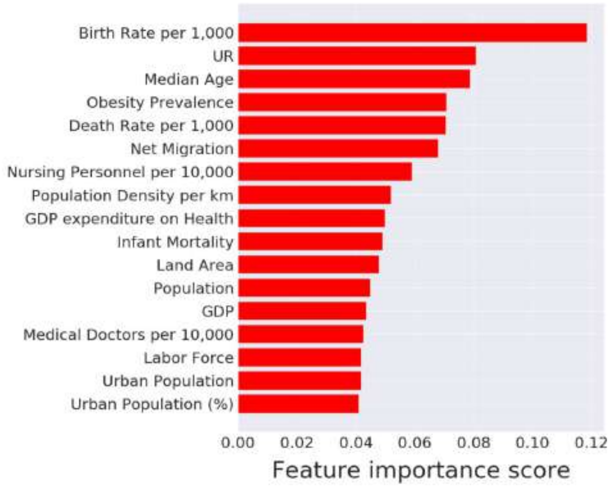



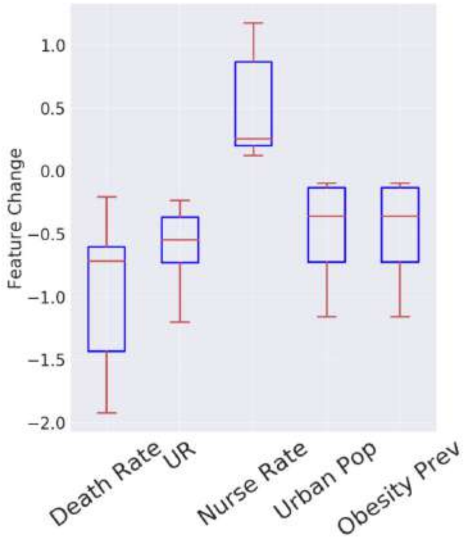

- Decrease death rate
- Decrease unemployment rate
- Increase nurse rate per 10,000 people
- Decrease urban population rate
- Decreasing obesity prevalence

# Use Case: COVID-19 (Risk of Mortality)

We apply ReLAX to generate CF explanations for a binary classifier (XGBoost with 500 trees) trained to predict the risk of mortality for COVID-19

We use generated CFs to sketch an action plan to lower the risk of mortality




- Decrease death rate
- Decrease unemployment rate
- Increase nurse rate per 10,000 people
- Decrease urban population rate
- Decreasing obesity prevalence

As obvious as they sound, many countries have suggested or enacted similar strategies to counter the COVID-19 pandemic (see here and here)

# Take-Home Message

- Attaching (human-understandable) explanations to accurate ML/AI model predictions is crucial in many critical domains

# Take-Home Message

- Attaching (human-understandable) explanations to accurate ML/AI model predictions is crucial in many critical domains
- If we don’t want to trade accuracy for explainability, we need to develop post-hoc explainers for complex, black-box models

# Take-Home Message

- Attaching (human-understandable) explanations to accurate ML/AI model predictions is crucial in many critical domains
- If we don’t want to trade accuracy for explainability, we need to develop post-hoc explainers for complex, black-box models
- Counterfactual examples (CFs) are promising tools to generate actionable explanations

# Take-Home Message

- Attaching (human-understandable) explanations to accurate ML/AI model predictions is crucial in many critical domains
- If we don't want to trade accuracy for explainability, we need to develop post-hoc explainers for complex, black-box models
- Counterfactual examples (CFs) are promising tools to generate actionable explanations
- We present a state-of-the-art CF generation method based on reinforcement learning and its application to a real use case

# So, What's Next?

- Counterfactual explanation is a very trendy research topic! A few possible open challenges are:
  - Developing **new CF generation methods** (e.g., based on/inspired by diffusion models)

# So, What’s Next?

- Counterfactual explanation is a very trendy research topic! A few possible open challenges are:
  - Developing **new CF generation methods** (e.g., based on/inspired by diffusion models)
  - Generating **CFs for new prediction settings** (e.g., sequential recommender systems, anomaly detection tools)

# So, What's Next?

- Counterfactual explanation is a very trendy research topic! A few possible open challenges are:
  - Developing **new CF generation methods** (e.g., based on/inspired by diffusion models)
  - Generating **CFs for new prediction settings** (e.g., sequential recommender systems, anomaly detection tools)
  - Incorporating **personalization into CFs** (not every actionable feature has the same weight across different input samples)

# So, What's Next?

- Counterfactual explanation is a very trendy research topic! A few possible open challenges are:
  - Developing **new CF generation methods** (e.g., based on/inspired by diffusion models)
  - Generating **CFs for new prediction settings** (e.g., sequential recommender systems, anomaly detection tools)
  - Incorporating **personalization into CFs** (not every actionable feature has the same weight across different input samples)
  - Extracting **natural language explanations** from generated CFs

# Suggested References


- Tolomei, G., Silvestri, F., Haines, A. and Lalmas, M., 2017. Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 465-474).
- Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S. and Turini, F., 2019. Factual and Counterfactual Explanations for Black Box Decision Making. IEEE Intelligent Systems, 34(6), pp.14-23.
- Le, T., Wang, S. and Lee, D., 2020. GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 238-248).
- Mothilal, R.K., Sharma, A. and Tan, C., 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607-617).
- Karimi, A.H., Barthe, G., Balle, B. and Valera, I., 2020. Model-Agnostic Counterfactual Explanations for Consequential Decisions. In International Conference on Artificial Intelligence and Statistics (pp. 895-905). PMLR.
- Lucic, A., Oosterhuis, H., Haned, H. and de Rijke, M., 2022. FOCUS: Flexible Optimizable Counterfactual Explanations for Tree Ensembles. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 5, pp. 5313-5322).
- Chen, Z., Silvestri, F., Wang, J., Zhu, H., Ahn, H. and Tolomei, G., 2022. ReLAX: Reinforcement Learning Agent Explainer for Arbitrary Predictive Models. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (pp. 252-261).
- Chen, Z., Silvestri, F., Tolomei, G., Wang, J., Zhu, H. and Ahn, H., 2022. Explain the Explainer: Interpreting Model-Agnostic Counterfactual Explanations of a Deep Reinforcement Learning Agent. IEEE Transactions on Artificial Intelligence.
- Lucic, A., Ter Hoeve, M.A., Tolomei, G., De Rijke, M. and Silvestri, F., 2022. CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. In International Conference on Artificial Intelligence and Statistics (pp. 4499-4511). PMLR.
- Chen, Z., Silvestri, F., Wang, J., Zhang, Y., Huang, Z., Ahn, H. and Tolomei, G., 2022. Grease: Generate Factual and Counterfactual Explanations for GNN-based Recommendations. arXiv preprint arXiv:2208.04222.
- Guidotti, R., 2022. Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking. Data Mining and Knowledge Discovery, pp.1-55.

Thoughts?

Ideas?

Suggestions?

**Let's Collaborate!**