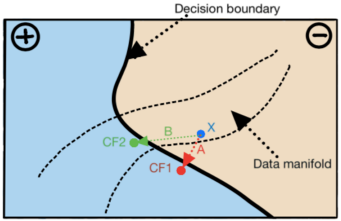
Interpretable Machine Learning

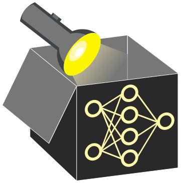
Counterfactual Explanations (CEs) Motivation



Learning goals

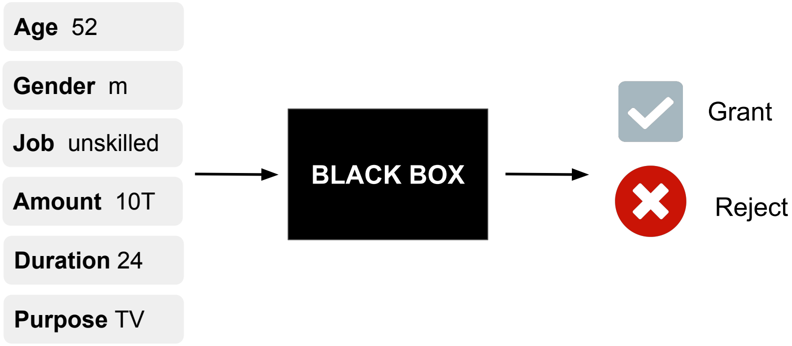
- Understand the motivation behind CEs
- Know why and how CEs are used
- Recognize the philosophical foundations of counterfactual reasoning

MOTIVATING EXAMPLE: CREDIT RISK & CE



x: customer and credit information

y: grant or reject credit



Age 52

Gender m

Job unskilled

Amount 10T

Duration 24

Purpose TV

→ BLACK BOX →

✓ Grant

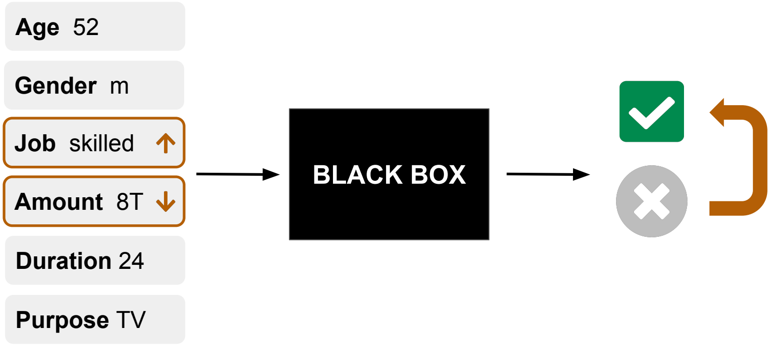
✗ Reject

Potential questions:

- Why was the credit rejected?
- Is this decision fair compared with similar applicants?
- **How should x be changed so that the credit is accepted?**

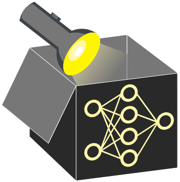
MOTIVATING EXAMPLE: CREDIT RISK & CE

CEs provide answers in the form of "What-If"-scenarios.



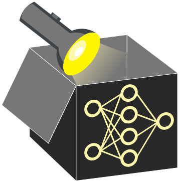
“If the applicant had higher skills and the credit amount had been reduced to $8.000, the loan would have been granted."

CORE DEFINITION AND PURPOSE OF CE



- **Counterfactual explanation (CE):** Hypothetical input $\mathbf{x}'$ close to the data point of interest $\mathbf{x}$ whose prediction equals a user-defined desired outcome y'

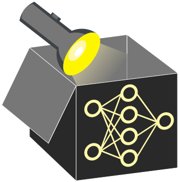
CORE DEFINITION AND PURPOSE OF CE



- **Counterfactual explanation (CE):** Hypothetical input $\mathbf{x}'$ close to the data point of interest $\mathbf{x}$ whose prediction equals a user-defined desired outcome y'
- **Proximity constraint:**

  Find $\mathbf{x}' \approx \mathbf{x}$ such that $f(\mathbf{x}') = y'$ and distance $d(\mathbf{x}, \mathbf{x}')$ is minimal

CORE DEFINITION AND PURPOSE OF CE

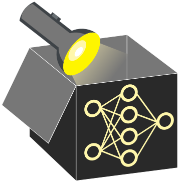


- **Counterfactual explanation (CE):** Hypothetical input $\mathbf{x}'$ close to the data point of interest $\mathbf{x}$ whose prediction equals a user-defined desired outcome y'
- **Proximity constraint:**

  Find $\mathbf{x}' \approx \mathbf{x}$ such that $f(\mathbf{x}') = y'$ and distance $d(\mathbf{x}, \mathbf{x}')$ is minimal

- **Minimal actionable changes:** Difference $\mathbf{x}' - \mathbf{x}$ shows the smallest feature change a user could realize in practice

CORE DEFINITION AND PURPOSE OF CE

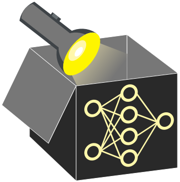


- **Counterfactual explanation (CE):** Hypothetical input $\mathbf{x}'$ close to the data point of interest $\mathbf{x}$ whose prediction equals a user-defined desired outcome y'
- **Proximity constraint:**

  Find $\quad \mathbf{x}' \approx \mathbf{x} \quad$ such that $\quad f(\mathbf{x}') = y' \quad$ and distance $\quad d(\mathbf{x}, \mathbf{x}') \quad$ is minimal

- **Minimal actionable changes:** Difference $\mathbf{x}' - \mathbf{x}$ shows the smallest feature change a user could realize in practice
- **Primary audience:**
  - Individuals aiming to alter model predictions
  - ML engineers exploring model behavior under adversarial conditions $\rightsquigarrow$ how small text changes in an email flip the prediction from "spam" to "no spam"

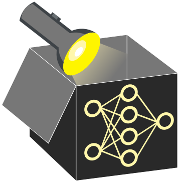
INTERPRETIVE AIMS & ROLES



CEs can serve various purposes; the user can decide what to learn from them, e.g.:

“If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

INTERPRETIVE AIMS & ROLES

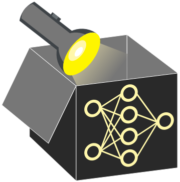


CEs can serve various purposes; the user can decide what to learn from them, e.g.:

“If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.

INTERPRETIVE AIMS & ROLES



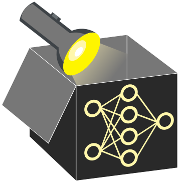
CEs can serve various purposes; the user can decide what to learn from them, e.g.:

“If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.

- **Provide reasons:**
  Interesting, I did not know that age plays a role in loan applications.

INTERPRETIVE AIMS & ROLES



CEs can serve various purposes; the user can decide what to learn from them, e.g.:

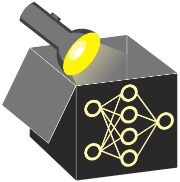
“If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.

- **Provide reasons:**
  Interesting, I did not know that age plays a role in loan applications.

- **Provide grounds to contest the decision:**
  How dare you, I won’t accept age-based discrimination in my application.

INTERPRETIVE AIMS & ROLES



CEs can serve various purposes; the user can decide what to learn from them, e.g.:

“If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

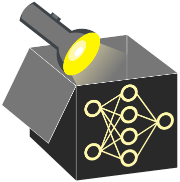
- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.

- **Provide reasons:**
  Interesting, I did not know that age plays a role in loan applications.

- **Provide grounds to contest the decision:**
  How dare you, I won’t accept age-based discrimination in my application.

- **Detect model biases:**
  There is a bug, an increase in amount should not increase approval rates.

PHILOSOPHICAL FOUNDATIONS ▸ “Lewis” 1973

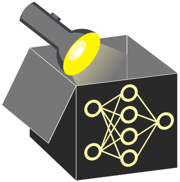


Counterfactuals have a long tradition in analytic philosophy
⇝ A **counterfactual conditional** takes the form:

"If S had occurred, Q would have occurred."

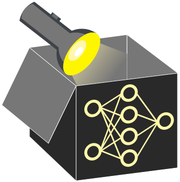
- *S*: past event that never happened ⇝ CE run contrary to fact
- Statement is true iff *Q* holds in all **closest** worlds where *S* is true
- Closest worlds preserve laws and change as few facts as possible (related to *S*)

PHILOSOPHICAL FOUNDATIONS



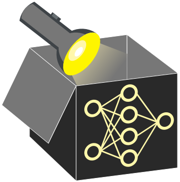
- CEs have largely been studied to explain causal dependence
- **Causal dependence**: Q depends on $S \Leftrightarrow$ without S, no Q
  $\leadsto$ Good CEs reveal critical causal factors that drove algorithmic decision
  $\leadsto$ **CE objective**: find $\mathbf{x}' \approx \mathbf{x}$ with $f(\mathbf{x}') = y'$ to expose causal features

PHILOSOPHICAL FOUNDATIONS



- CEs have largely been studied to explain causal dependence
- **Causal dependence**: Q depends on $S \Leftrightarrow$ without S, no Q
  $\rightsquigarrow$ Good CEs reveal critical causal factors that drove algorithmic decision
  $\rightsquigarrow$ **CE objective**: find $\mathbf{x}' \approx \mathbf{x}$ with $f(\mathbf{x}') = y'$ to expose causal features
- Relaxing closeness may add causally irrelevant edits to the explanation
  $\rightsquigarrow$ e.g., suggest to lower loan *and* increase age (but only loan matters)

PHILOSOPHICAL FOUNDATIONS



- CEs have largely been studied to explain causal dependence
- **Causal dependence**: Q depends on $S \Leftrightarrow$ without S, no Q
  $\rightsquigarrow$ Good CEs reveal critical causal factors that drove algorithmic decision
  $\rightsquigarrow$ **CE objective**: find $\mathbf{x}' \approx \mathbf{x}$ with $f(\mathbf{x}') = y'$ to expose causal features
- Relaxing closeness may add causally irrelevant edits to the explanation
  $\rightsquigarrow$ e.g., suggest to lower loan *and* increase age (but only loan matters)
- CEs are contrastive: Explain a decision by comparing it to a different case
  $\rightsquigarrow$ If age were 30 (not 60), loan would have been $9k (not rejected)
  $\rightsquigarrow$ Answers contrastive question:
  “Why Q’ instead of Q?” (preferred by humans)