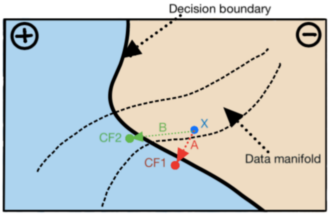
Interpretable Machine Learning

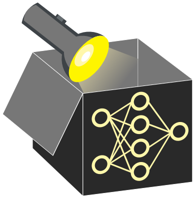
Counterfactual Explanations



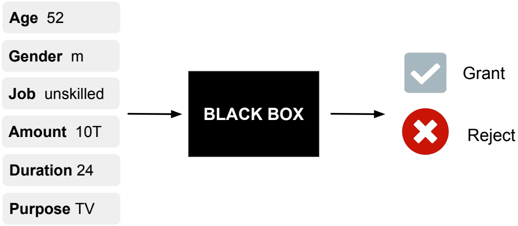
Learning goals

- Understand the motivation behind CEs
- See the mathematical foundation of CEs

EXAMPLE: CREDIT RISK APPLICATION



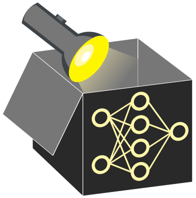
- **x**: customer and credit information
- *y*: grant or reject credit



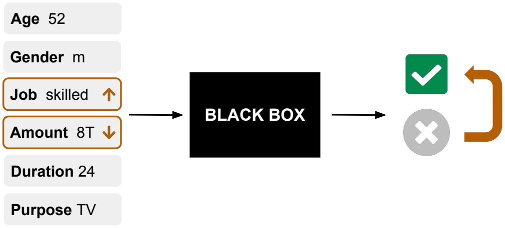
Questions:

- Why was the credit rejected?
- Is it a fair decision?
- **How should x be changed so that the credit is accepted?**

EXAMPLE: CREDIT RISK APPLICATION

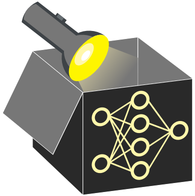


Counterfactual Explanations provide answers in the form of "What-If"-scenarios.



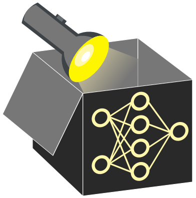
“If the person was more skilled and the credit amount had been reduced to $8.000, the credit would have been granted."

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA



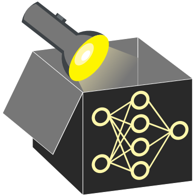
- Counterfactual explanations == counterfactuals == CEs

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA



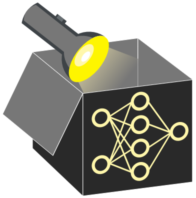
- Counterfactual explanations == counterfactuals == CEs
- Explain particular predictions of an ML model by presenting an alternative input whose prediction equals a desired outcome

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA



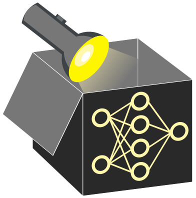
- Counterfactual explanations == counterfactuals == CEs
- Explain particular predictions of an ML model by presenting an alternative input whose prediction equals a desired outcome
- Represent **close neighbors** of a data point we are interested in, but belonging to the **desired outcome**

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA



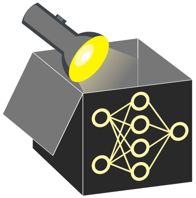
- Counterfactual explanations == counterfactuals == CEs
- Explain particular predictions of an ML model by presenting an alternative input whose prediction equals a desired outcome
- Represent **close neighbors** of a data point we are interested in, but belonging to the **desired outcome**
- Reveal which minimal changes to the input are sufficient to receive a different outcome
  $\rightsquigarrow$ Useful if there is a chance to change the input features (e.g., by changing behaviour)

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA



- Counterfactual explanations == counterfactuals == CEs
- Explain particular predictions of an ML model by presenting an alternative input whose prediction equals a desired outcome
- Represent **close neighbors** of a data point we are interested in, but belonging to the **desired outcome**
- Reveal which minimal changes to the input are sufficient to receive a different outcome
  $\rightsquigarrow$ Useful if there is a chance to change the input features (e.g., by changing behaviour)
- The targeted audience of CEs are often end-users

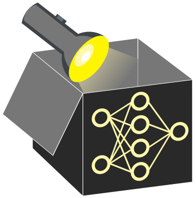
AIMS & ROLES



CEs can serve various purposes; the user can decide what to learn from them. For example:

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

AIMS & ROLES

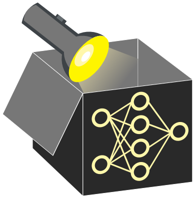


CEs can serve various purposes; the user can decide what to learn from them. For example:

“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.

AIMS & ROLES



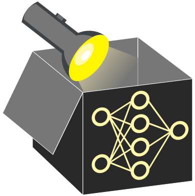
CEs can serve various purposes; the user can decide what to learn from them. For example:

“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.

AIMS & ROLES



CEs can serve various purposes; the user can decide what to learn from them. For example:

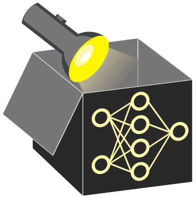
“If the person had been **one year older** and the **credit amount had been increased** to $12.000,
the credit would have been granted."

- **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 do not want to be discriminated for my age in an application.

AIMS & ROLES



CEs can serve various purposes; the user can decide what to learn from them. For example:

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

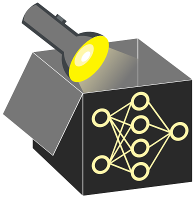
- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.

- **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 do not want to be discriminated for my age in an application.

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

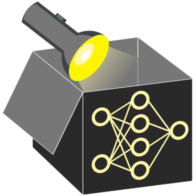
PHILOSOPHICAL BASIS



Counterfactuals have a long-standing tradition in analytic philosophy
⇝ Accoding to ▸ Lewis (1973), a **counterfactual conditional** is a statement of the form:

"If S was the case, Q would have been the case."

PHILOSOPHICAL BASIS

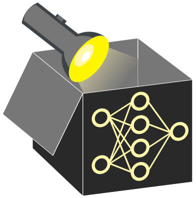


Counterfactuals have a long-standing tradition in analytic philosophy
⇝ Accoding to ▸ Lewis (1973), a **counterfactual conditional** is a statement of the form:

"If *S* was the case, *Q* would have been the case."

- *S* is an event that must relate to a past event that didn't occur
  ⇝ counterfactuals run **contrary** to the **facts**

PHILOSOPHICAL BASIS

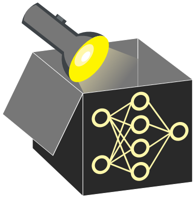


Counterfactuals have a long-standing tradition in analytic philosophy
$\rightsquigarrow$ Accoding to ▸ Lewis (1973), a **counterfactual conditional** is a statement of the form:

"If S was the case, Q would have been the case."

- S is an event that must relate to a past event that didn't occur
  $\rightsquigarrow$ counterfactuals run **contrary** to the **facts**
- Above statement is true, if in all possible worlds most similar to the actual world where S had been the case, Q would have been the case

PHILOSOPHICAL BASIS

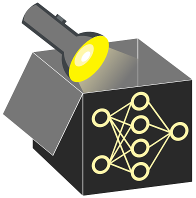


Counterfactuals have a long-standing tradition in analytic philosophy
$\rightsquigarrow$ According to ▸ Lewis (1973), a **counterfactual conditional** is a statement of the form:

“If *S* was the case, *Q* would have been the case."

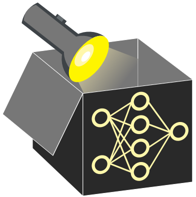
- *S* is an event that must relate to a past event that didn't occur
  $\rightsquigarrow$ counterfactuals run **contrary** to the **facts**
- Above statement is true, if in all possible worlds most similar to the actual world where *S* had been the case, *Q* would have been the case
- A world is similar to another if laws are maximally preserved between the worlds and only a few facts are changed

PHILOSOPHICAL BASIS



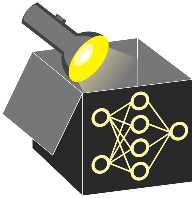
- Counterfactuals have largely been studied to explain causal dependence

PHILOSOPHICAL BASIS



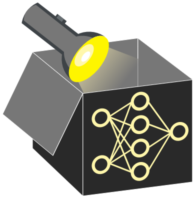
- Counterfactuals have largely been studied to explain causal dependence
- Causal dependence underlies the explanatory power
  $\rightsquigarrow$ good CEs point to critical causal factors that drove the algorithmic decision

PHILOSOPHICAL BASIS



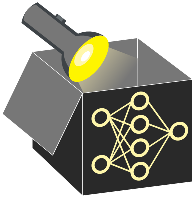
- Counterfactuals have largely been studied to explain causal dependence
- Causal dependence underlies the explanatory power
  $\rightsquigarrow$ good CEs point to critical causal factors that drove the algorithmic decision
- If maximal closeness is relaxed, causally irrelevant factors can become part of the explanation
  $\rightsquigarrow$ e.g., decreasing loan amount by $20.000 and being one year older is recommended by the explainer although only loan amount might be causally relevant

PHILOSOPHICAL BASIS



- Counterfactuals have largely been studied to explain causal dependence
- Causal dependence underlies the explanatory power
  $\rightsquigarrow$ good CEs point to critical causal factors that drove the algorithmic decision
- If maximal closeness is relaxed, causally irrelevant factors can become part of the explanation
  $\rightsquigarrow$ e.g., decreasing loan amount by $20.000 and being one year older is recommended by the explainer although only loan amount might be causally relevant
- CEs are often contrastive, i.e., they explain a decision by referring to an alternative outcome
  $\rightsquigarrow$ e.g., if the loan applicant was 30 instead of 60 years old, the approved loan would have been over $100.000 instead of $40.000

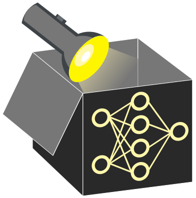
MATHEMATICAL PERSPECTIVE



Terminology:

- $\mathbf{x}$: original/factual datapoint whose prediction we want to explain
- $y' \subset \mathbb{R}^g$: desired prediction ($y' = 1000$ or $y' =$ “grant credit") or interval ($y' = [1000, \infty[$)

MATHEMATICAL PERSPECTIVE



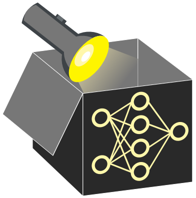
Terminology:

- **x**: original/factual datapoint whose prediction we want to explain
- $y' \subset \mathbb{R}^g$: desired prediction ($y' = 1000$ or $y' =$ “grant credit") or interval ($y' = [1000, \infty[$)

A **valid** counterfactual $\mathbf{x}'$ is a datapoint:

1. whose prediction $\hat{f}(\mathbf{x}')$ is equal to the desired prediction y'
2. that is maximally close to the original datapoint **x**

MATHEMATICAL PERSPECTIVE



Terminology:

- $\mathbf{x}$: original/factual datapoint whose prediction we want to explain
- $y' \subset \mathbb{R}^g$: desired prediction ($y' = 1000$ or $y' =$ “grant credit") or interval ($y' = [1000, \infty[$)

A **valid** counterfactual $\mathbf{x}'$ is a datapoint:

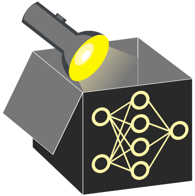
1. whose prediction $\hat{f}(\mathbf{x}')$ is equal to the desired prediction y'
2. that is maximally close to the original datapoint $\mathbf{x}$

Reformulate these two objectives (denoted by o_1 and o_2) as optimization problem:

$$\arg\min_{\mathbf{x}'} \lambda_1 o_p(\hat{f}(\mathbf{x}'), y') + \lambda_2 o_f(\mathbf{x}', \mathbf{x})$$

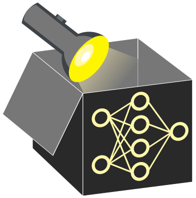
- λ_1 and λ_2 balance the two objectives
- Choice of o_p (distance on prediction space) and of o_f (distance on feature space) is crucial

MATHEMATICAL PERSPECTIVE ▸ Dandl et al. (2020)



- Regression: o_p could be the L_1-distance $o_p(\hat{f}(\mathbf{x}'), y') = |\hat{f}(\mathbf{x}') - y'|$
- Classification: L_1-distance for scores and 0-1 Loss for labels, e.g., $o_p(\hat{f}(\mathbf{x}'), y') = \mathcal{I}_{\{\hat{f}(\mathbf{x}') \neq y'\}}$

MATHEMATICAL PERSPECTIVE ▸ Dandl et al. (2020)



- Regression: o_p could be the L$_1$-distance $o_p(\hat{f}(\mathbf{x}'), y') = |\hat{f}(\mathbf{x}') - y'|$
- Classification: L$_1$-distance for scores and 0-1 Loss for labels, e.g., $o_p(\hat{f}(\mathbf{x}'), y') = \mathcal{I}_{\{\hat{f}(\mathbf{x}') \neq y'\}}$
- o_f could be the Gower distance (suitable for mixed feature space):

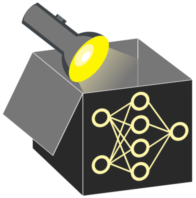
$$o_f(\mathbf{x}', \mathbf{x}) = d_G(\mathbf{x}', \mathbf{x}) = \frac{1}{p} \sum_{j=1}^{p} \delta_G(x_j', x_j) \in [0, 1]$$

The value of δ_G depends on the feature type (numerical or categorical):

$$\delta_G(x_j', x_j) = \begin{cases} \frac{1}{\widehat{R}_j} |x_j' - x_j| & \text{if } x_j \text{ is numerical} \\ \mathcal{I}_{\{x_j' \neq x_j\}} & \text{if } x_j \text{ is categorical} \end{cases}$$

with $\widehat{R}_j$ as the value range of feature j in the training dataset (to ensure that $\delta_G(x_j', x_j) \in [0, 1]$)

FURTHER OBJECTIVES

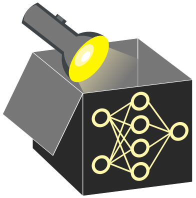


Additional constraints can improve the explanation quality of the corresponding CEs
$\rightsquigarrow$ popular constraints include sparsity and plausibility

Sparsity:

- End-users often prefer short over long explanations
  $\rightsquigarrow$ counterfactuals should be **sparse**

FURTHER OBJECTIVES

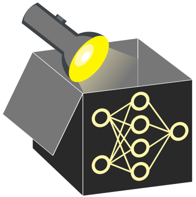


Additional constraints can improve the explanation quality of the corresponding CEs
$\rightsquigarrow$ popular constraints include sparsity and plausibility

Sparsity:

- End-users often prefer short over long explanations
  $\rightsquigarrow$ counterfactuals should be **sparse**
- Objective o_f can take the number of changed features into account (but does not have to)
  $\rightsquigarrow$ e.g., the L_0- and the L_1-norm (similar to LASSO) can do this

FURTHER OBJECTIVES



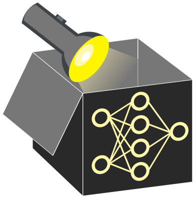
Additional constraints can improve the explanation quality of the corresponding CEs
$\rightsquigarrow$ popular constraints include sparsity and plausibility

Sparsity:

- End-users often prefer short over long explanations
  $\rightsquigarrow$ counterfactuals should be **sparse**
- Objective o_f can take the number of changed features into account (but does not have to)
  $\rightsquigarrow$ e.g., the L_0- and the L_1-norm (similar to LASSO) can do this
- Independently from o_f, sparsity in the changes can be additionally considered by another objective that counts the number of changed features via the L0-norm:

$$o_s(\mathbf{x}', \mathbf{x}) = \sum_{j=1}^{p} \mathcal{I}_{\{x'_j \neq x_j\}}$$

FURTHER OBJECTIVES



Plausibility:

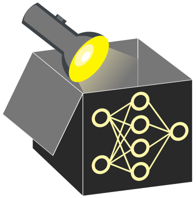
- CEs should suggest plausible alternatives
  $\rightsquigarrow$ e.g., not plausible to suggest to raise your income and get unemployed at the same time

FURTHER OBJECTIVES

Plausibility:

- CEs should suggest plausible alternatives
  $\rightsquigarrow$ e.g., not plausible to suggest to raise your income and get unemployed at the same time
- CEs should be realistic and adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ avoid unrealistic combinations of feature values

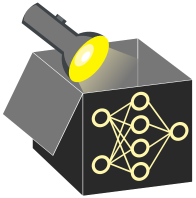
FURTHER OBJECTIVES



Plausibility:

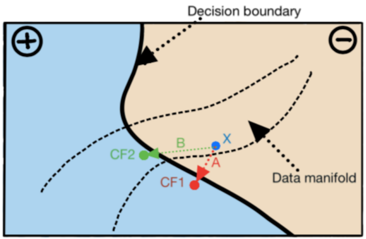
- CEs should suggest plausible alternatives
  $\rightsquigarrow$ e.g., not plausible to suggest to raise your income and get unemployed at the same time
- CEs should be realistic and adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ avoid unrealistic combinations of feature values
- Estimating joint distribution of training data is complex, especially for mixed feature spaces
  $\rightsquigarrow$ Proxy: ensure that $\mathbf{x}'$ is close to training data $\mathbf{X}$

FURTHER OBJECTIVES



Plausibility:

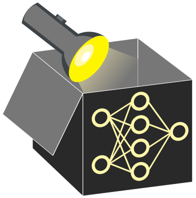
- CEs should suggest plausible alternatives
  $\rightsquigarrow$ e.g., not plausible to suggest to raise your income and get unemployed at the same time
- CEs should be realistic and adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ avoid unrealistic combinations of feature values
- Estimating joint distribution of training data is complex, especially for mixed feature spaces
  $\rightsquigarrow$ Proxy: ensure that $\mathbf{x}'$ is close to training data $\mathbf{X}$



Example from Verma et al. (2020)

- Two possible paths for $\mathbf{x}$, originally classified to $\ominus$
- Two valid CEs in class $\oplus$: CF1 and CF2
- Path **A** for CF1 is shorter
- Path **B** for CF2 is longer but adheres to data manifold

FURTHER OBJECTIVES



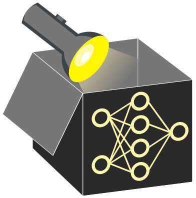
To ensure plausibility, o_4 could, e.g., be the Gower distance of $\mathbf{x}'$ to its nearest data point of the training dataset which we denote $\mathbf{x}^{[1]}$:

$$o_4(\mathbf{x}', \mathbf{X}) = d_G(\mathbf{x}', \mathbf{x}^{[1]}) = \frac{1}{p} \sum_{j=1}^{p} \delta_G(x'_j, x_j^{[1]})$$

We can extend the previous optimization problem by adding o_s (for sparsity) and o_4 (for plausibility):

$$\arg\min_{\mathbf{x}'} \lambda_1 o_p(\hat{f}(\mathbf{x}'), y') + \lambda_2 o_f(\mathbf{x}', \mathbf{x}) + \lambda_3 o_s(\mathbf{x}', \mathbf{x}) + \lambda_4 o_4(\mathbf{x}', \mathbf{X})$$

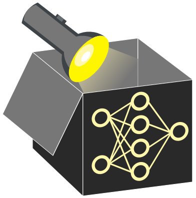
REMARKS: THE RASHOMON EFFECT



Issue (Rashomon effect):

- Solution to the optimization problem might not be unique
- Many equally close CE might exist that obtain the desired prediction
  $\Rightarrow$ Many different equally good explanations for the same decision exist

REMARKS: THE RASHOMON EFFECT



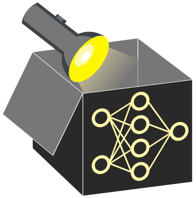
Issue (Rashomon effect):

- Solution to the optimization problem might not be unique
- Many equally close CE might exist that obtain the desired prediction
  $\Rightarrow$ Many different equally good explanations for the same decision exist

Possible solutions:

- Present all CEs for a given $\mathbf{x}$ (but: time and human processing capacity is limited)
- Focus on one or few CEs (but: by which criterion should they be selected?)

REMARKS: THE RASHOMON EFFECT



Issue (Rashomon effect):

- Solution to the optimization problem might not be unique
- Many equally close CE might exist that obtain the desired prediction
  $\Rightarrow$ Many different equally good explanations for the same decision exist

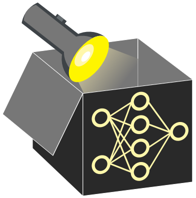
Possible solutions:

- Present all CEs for a given $\mathbf{x}$ (but: time and human processing capacity is limited)
- Focus on one or few CEs (but: by which criterion should they be selected?)

Note:

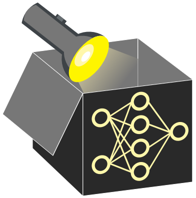
- As the model is generally non-linear, inconsistent and diverse CEs can arise e.g. suggesting either an increase or decrease in credit duration (confuses the explainee)
- How to deal with the Rashomon effect is considered an open problem in IML

REMARKS: MODEL OR REAL-WORLD



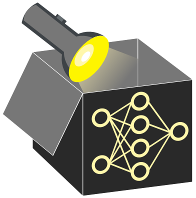
- Most CEs provide explanations of model predictions, but CEs might appear to explain the real-world for end-users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted

REMARKS: MODEL OR REAL-WORLD



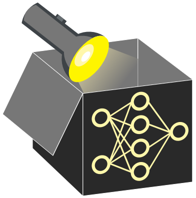
- Most CEs provide explanations of model predictions, but CEs might appear to explain the real-world for end-users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- Consider a CE that proposes to increase the feature age by 5 to obtain the loan
  $\rightsquigarrow$ a loan applicant takes this information and applies 5 years later for the loan

REMARKS: MODEL OR REAL-WORLD



- Most CEs provide explanations of model predictions, but CEs might appear to explain the real-world for end-users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- Consider a CE that proposes to increase the feature age by 5 to obtain the loan
  $\rightsquigarrow$ a loan applicant takes this information and applies 5 years later for the loan
- However, by then, many other feature values might have changed
  $\rightsquigarrow$ not only age, also other causally dependent features e.g. job status might have changed
  $\rightsquigarrow$ ▸ Karimi et al. (2020) avoid this by considering causal dependencies between features

REMARKS: MODEL OR REAL-WORLD



- Most CEs provide explanations of model predictions, but CEs might appear to explain the real-world for end-users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- Consider a CE that proposes to increase the feature age by 5 to obtain the loan
  $\rightsquigarrow$ a loan applicant takes this information and applies 5 years later for the loan
- However, by then, many other feature values might have changed
  $\rightsquigarrow$ not only age, also other causally dependent features e.g. job status might have changed
  $\rightsquigarrow$ Karimi et al. (2020) avoid this by considering causal dependencies between features
- Also, the bank's algorithm might change and previous CEs are not applicable anymore