Counterfactual Explanations as Interventions in Latent Space

Riccardo Crupi, Alessandro Castelnovo, Daniele Regoli,
Beatriz San Miguel Gonzalez

Imperial College London
14 Feb 2022

Content

1. Project description and objectives
2. Counterfactual Explanations
3. Proposed method: CEILS (Counterfactual Explanations as Interventions on Latent Space)
4. Evaluation
5. Conclusions

Project description and objectives

Joint collaboration to research on **Trustworthy AI in the financial domain.**

Goal

The **goal** is to generate more feasible counterfactual recommendations and explanations.

Contribution

CEILS (Counterfactual Explanations as Interventions on Latent Space) has the advantage of leveraging the underlying causal relations by design and it can be set on top of standard counterfactual generators.

Intuition

The explanations are found in the latent space of variables defined by the residuals of an Additive Noise Model over the input space variables and its Structural Causal Model.

Counterfactual definition

why counterfactuals are useful

«A set of recommendations to communicate end users what should change in order to obtain a desired result (e.g. a loan)».

Consider a Classifier $C : X \rightarrow Y$ defining whether a profile x will have a desired result or not ($y = 1$ *or* $y = 0$).

The counterfactual x_{cf} of x_0 is such that

$$C(x_{\text{cf}}) \neq C(x_0)$$

main problem with counterfactuals

Usually, x_{cf} is generated based on the proximity to x_0 → This produces unfeasible recommendations such as *"reduce your age and increase your credit score"*

Counterfactual explanations and recourse

General formulation

In a "static" world, the action a could correspond to $x_{\text{cf}} - x_0$

In general this is not true, due to interdependence of variables.
A more general formulation is*:

$$\begin{cases} a^* = \arg\min_{a \in \mathcal{F}_{\mathcal{A}}} \; cost(a, x^0), \\ x^{0,\text{cf}} = S(x^0, a) \in \mathcal{P}_{\mathcal{X}}, \\ \mathcal{C}(x^{0,\text{cf}}) \neq \mathcal{C}(x^0); \end{cases}$$

Recourse

what actions would have led me to reach such profile?

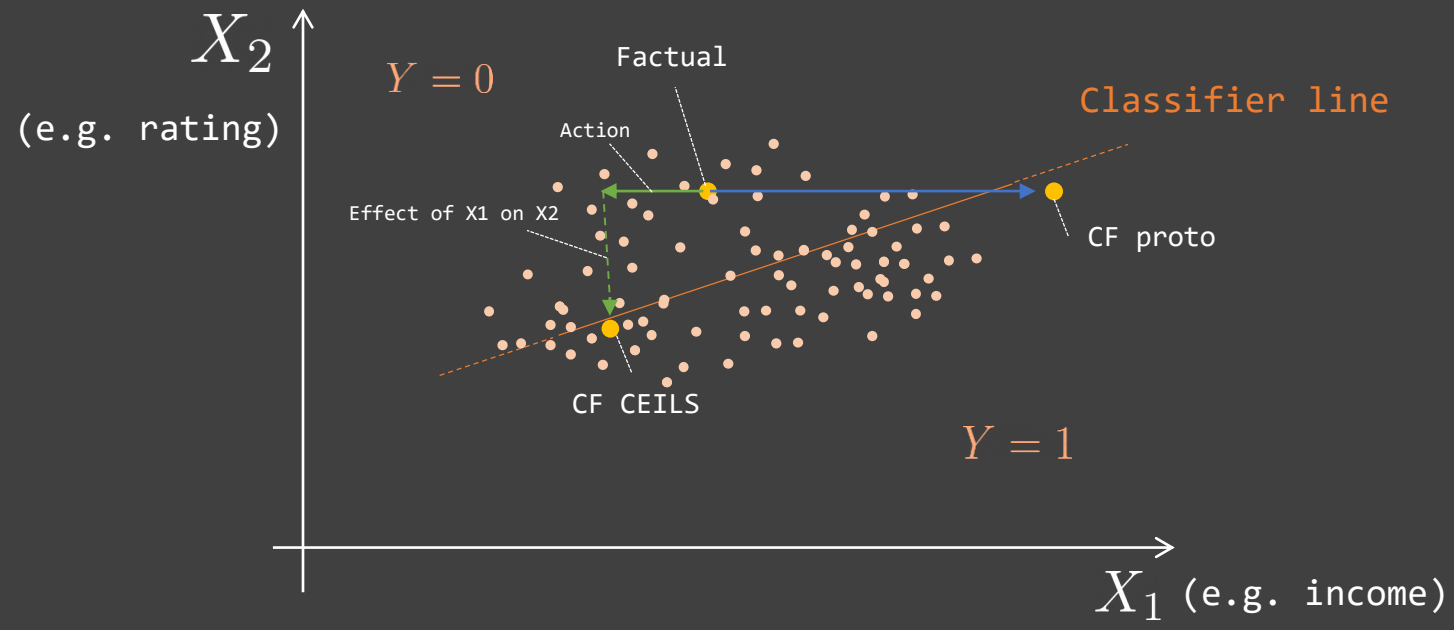
*Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. arXiv preprint arXiv:2010.04050, 2021a.

Counterfactual definition

Plausibility & feasibility

The plausible set *P* can be the original distribution of the data.

Feasibility instead concerns constraints on actions, in the image X_2 can't be actionable directly.
In a "static" world *a=d*.



Generation of Structural Equations

Assumptions

Causal graph provided for X.

SCM: additive noise model.

$$X_v = f_v(\mathbf{pa}(X_v)) + U_v, \quad v = 1, \ldots, d.$$

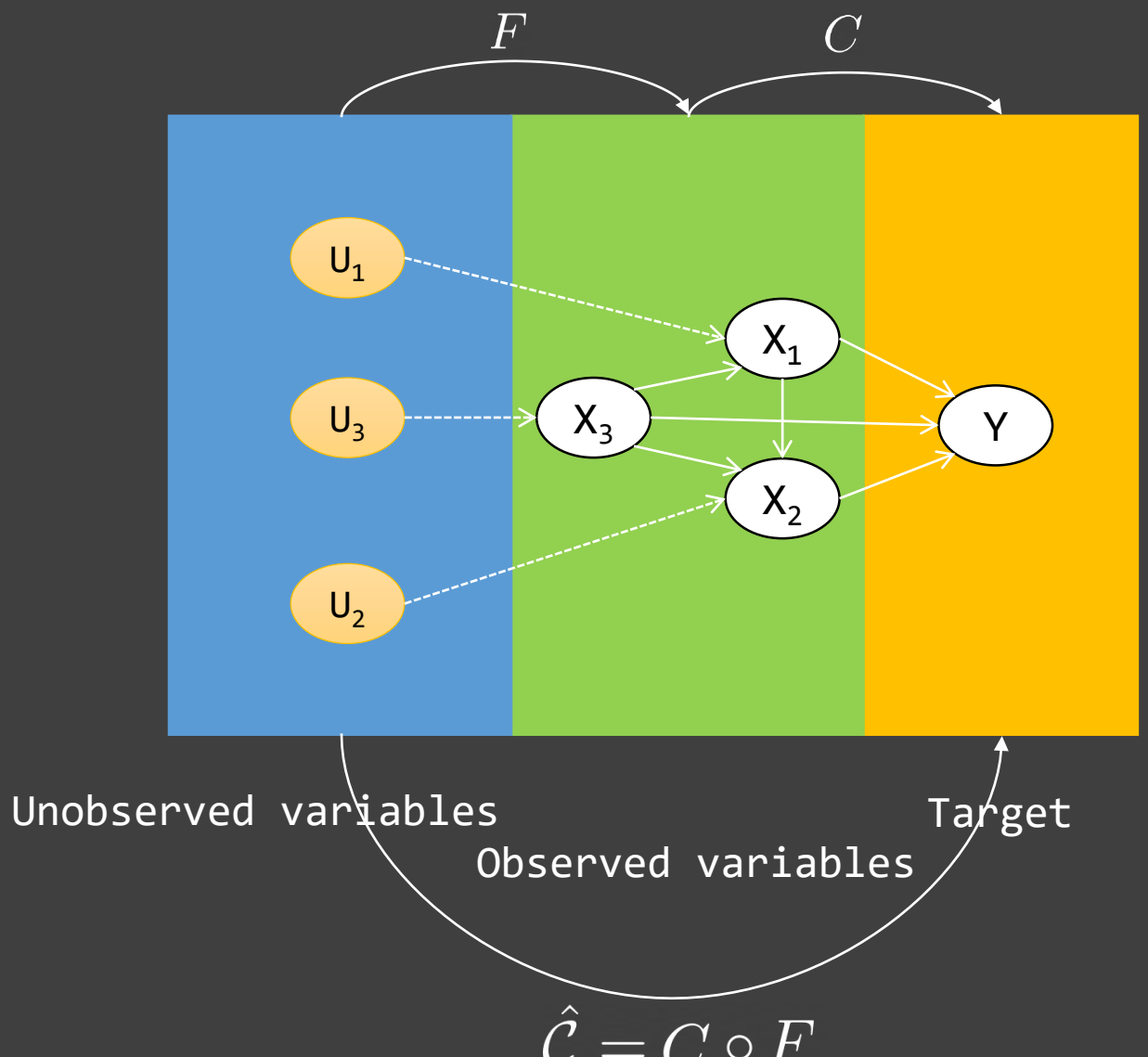
f_v - neural network regressors
U_v - residual errors.

Causal sufficiency (no hidden confounders).

Variables & Functions

CEILS

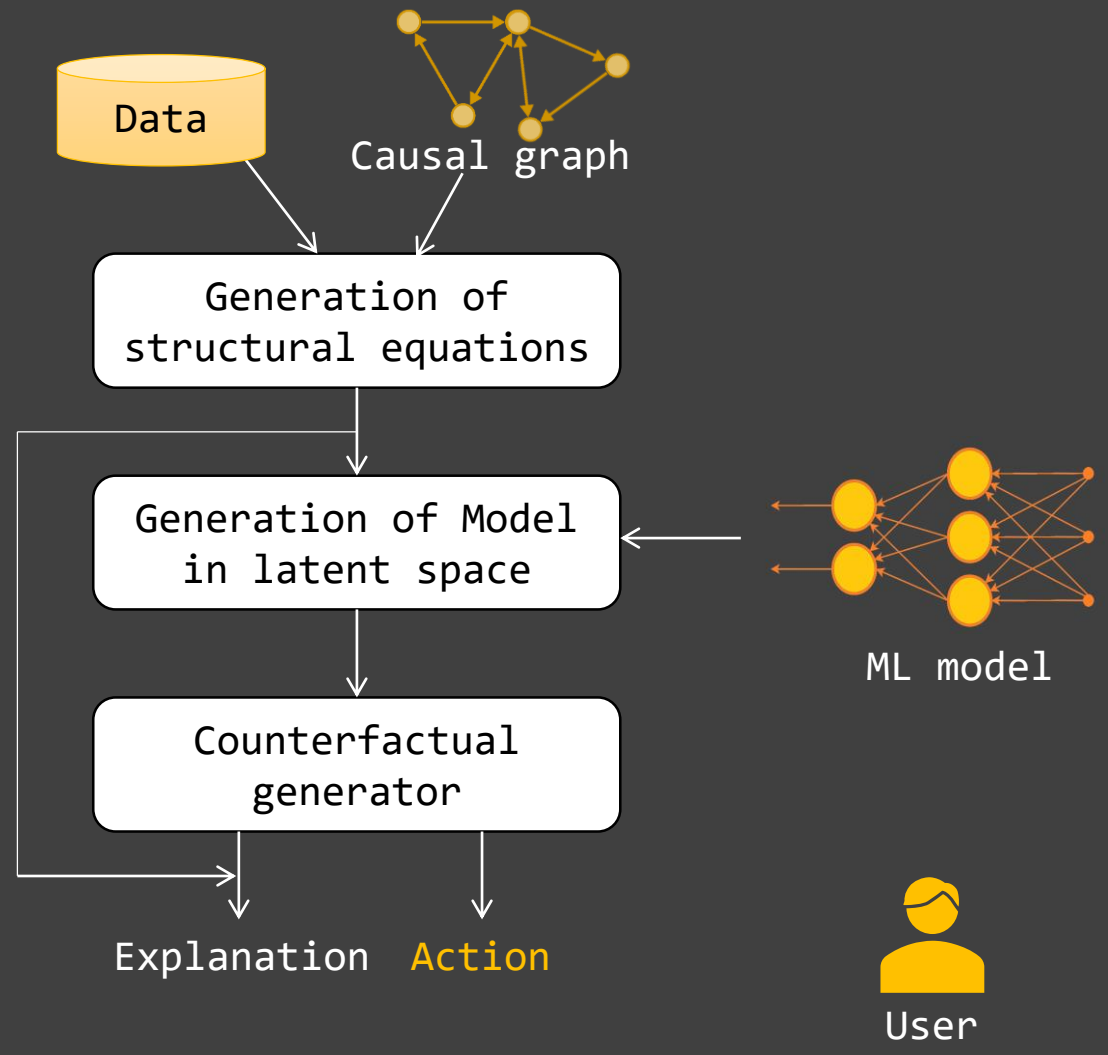
Model in the Latent Space



Workflow

CEILS

A new methodology to generate counterfactual explanations focused on the production of more feasible actions

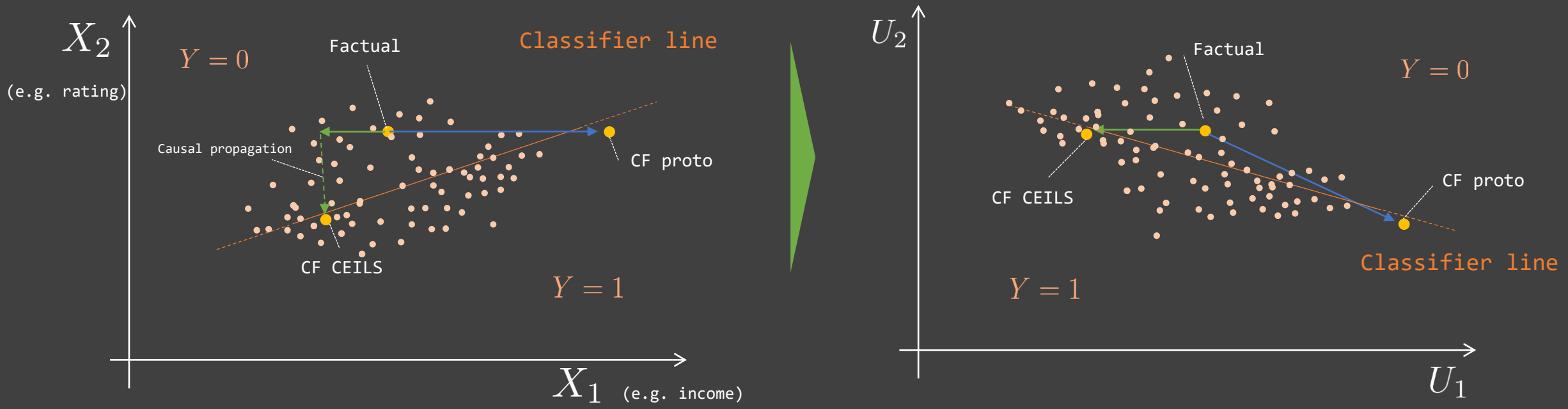


Causal modelling

Feasibility constrains

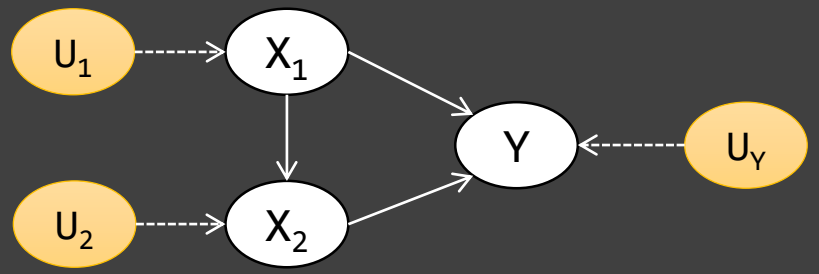
The feasibility constraints live in the *U* space.
The variables are classified as:

- immutable
- mutable but non-actionable
- actionable



Synthetic Dataset

Experiments on a synthetic dataset



$$\begin{cases} X_1 = U_1, \quad U_1 \sim \mathcal{N}(-1,1) \\ X_2 = X_1 + U_2, \quad U_2 \sim \mathcal{N}(5,1) \\ Y = \mathbf{1}_{(3X_2 - X_1 + U_Y) > t}, \quad U_Y \sim \mathcal{N}(0,1) \end{cases}$$

100,000 samples

Generate counterfactual explanations with:

- Alibi Prototype (correlation keeping)
- CEILS built up on Alibi Prototype

X_1- actionable
X_2- mutable but non-actionable

Alibi Prototype ignores the impact of X_1 on X_2

Evaluation - Experiment

Counterfactual Generation

1,000 random instances of the dataset are used as observations to be explained

Generate counterfactual explanations with:

- Baseline approach (Alibi Prototype)*
- CEILS built up on Alibi Prototype

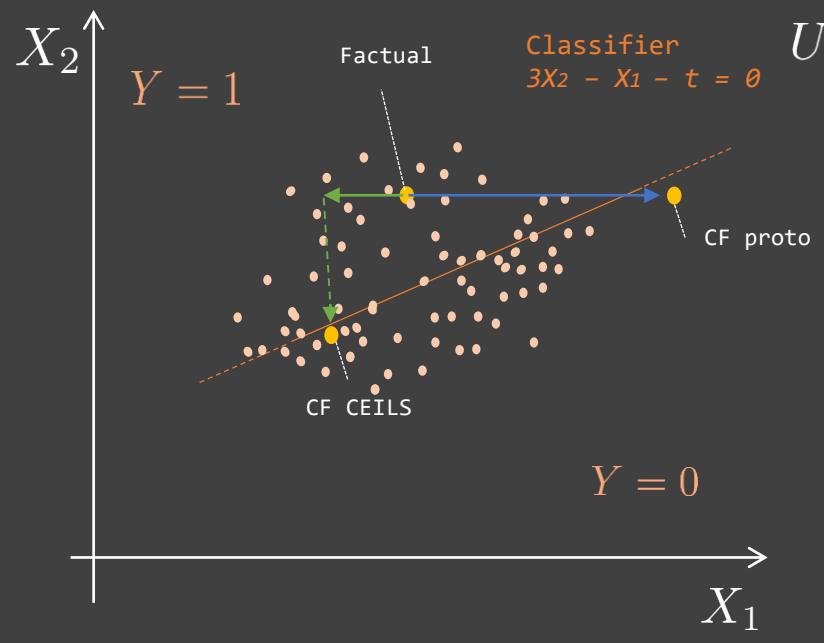
*Klaise, J., Van Looveren, A., Vacanti, G., & Coca, A. (2021). Alibi Explain: algorithms for explaining machine learning models. *Journal of Machine Learning Research*, *22*(181), 1-7.

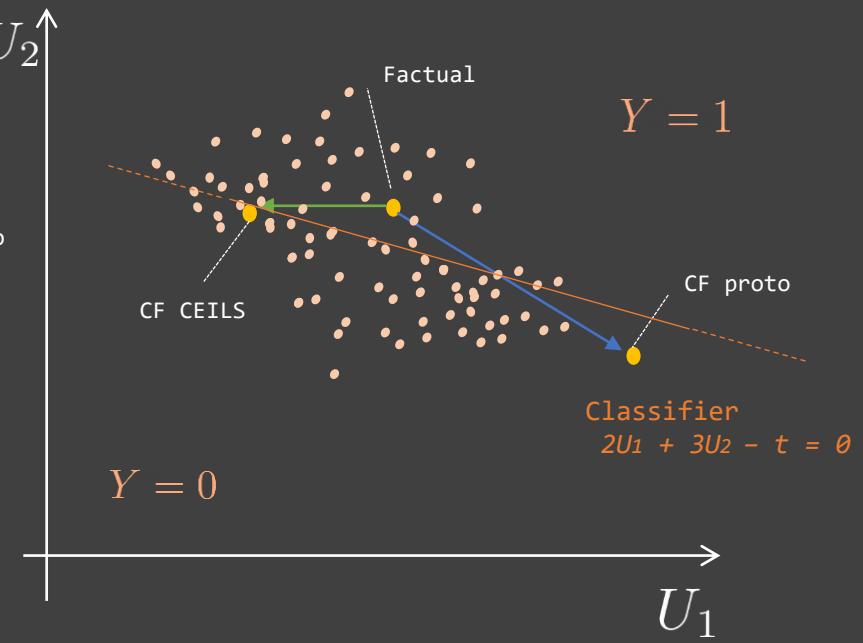
Examples

Experiments on a synthetic dataset

$$\begin{cases} X_1 = U_1, \quad U_1 \sim \mathcal{N}(-1, 1) \\ X_2 = X_1 + U_2, \quad U_2 \sim \mathcal{N}(5, 1) \\ Y = \mathbf{1}_{(3X_2 - X_1 + U_Y) > t}, \quad U_Y \sim \mathcal{N}(0, 1) \end{cases}$$

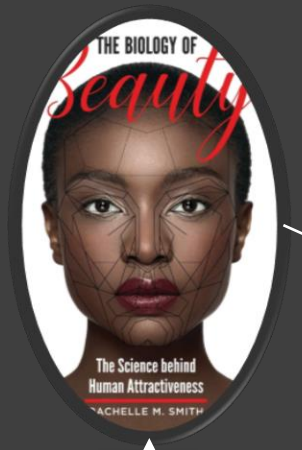
	Variable	Factual	counterfactuals				
			Prototype	CEILS	ΔPrototype	ΔCEILS	Action
Example 1	Y	0	1	1			
	X_1	-1.098	-1.364	-0.995	-0.266	0.103	0.103
	X_2	3.896	3.896	4.006	0	0.110	0
Example 2	Y	1	0	0			
	X_1	-0.886	-0.726	-0.973	0.159	-0.087	-0.087
	X_2	4.080	4.080	3.999	0	-0.081	0





Valentine's Day Experiment

Attractiveness



Fun

They call me Dirichlet because all my potential is latent and awaiting for allocation*

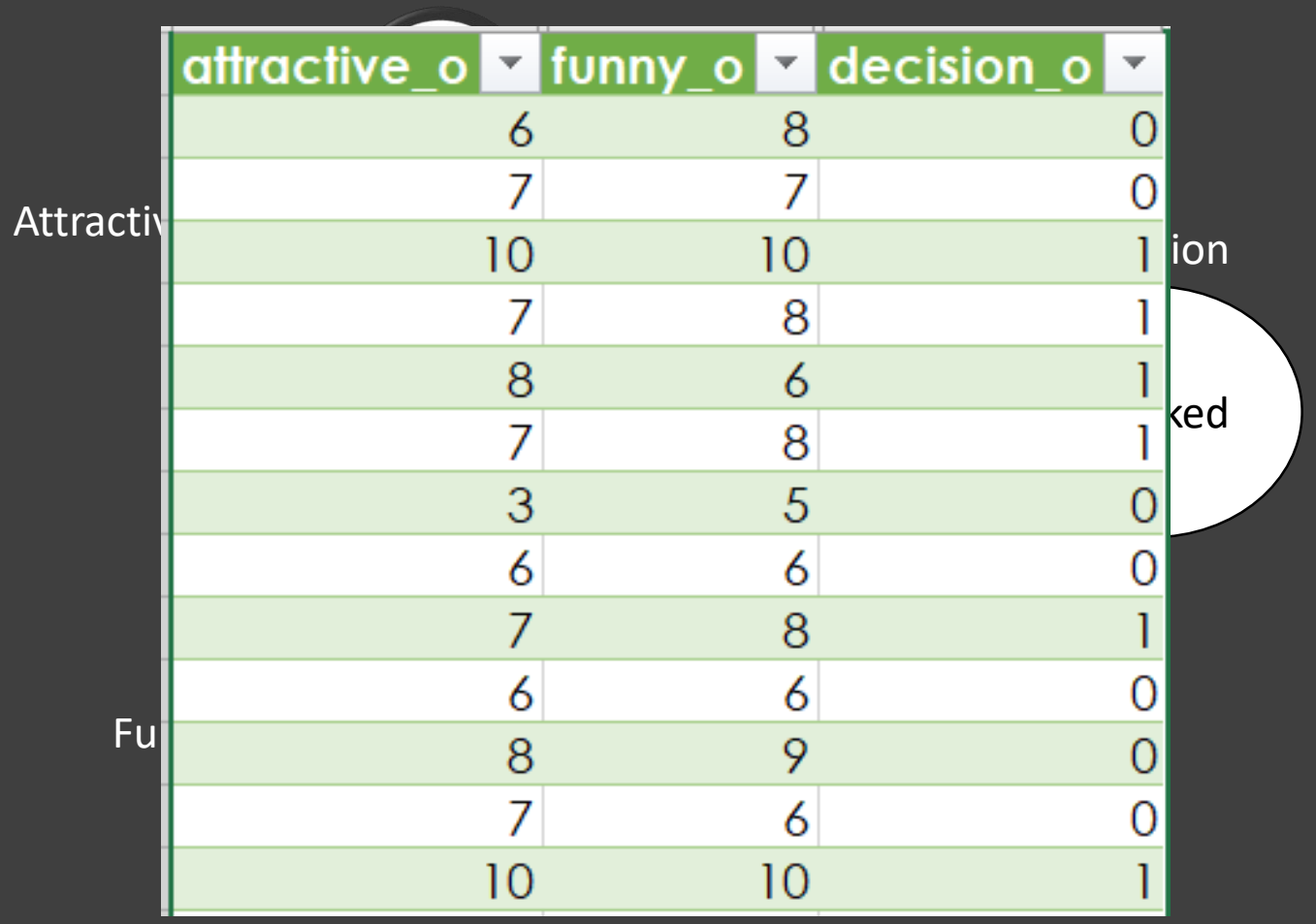
Decision

Be Liked

*https://twitter.com/ML_Hipster/status/381914112227147777?s=20&t=OQzGgZWyfMsBwCkbD2DnBg

Valentine's Day Experiment

https://datahub.io/machine-learning/speed-dating



attractive_o	funny_o	decision_o
6	8	0
7	7	0
10	10	1
7	8	1
8	6	1
7	8	1
3	5	0
6	6	0
7	8	1
6	6	0
8	9	0
7	6	0
10	10	1

Valentine's Day Experiment*

Variable	Factual	Prototype	CEILS	ΔProto	ΔCeils	Action
Y	0	1	1			
F	6.000	7.321	6.585	1.321	0.585	0.585
A	7.000	7.000	7.384	0.000	0.384	0.000

*https://github.com/FLE-ISP/CEILS/tree/main/experiments_run/speedDate_experiments

Evaluation in the Finance Domain

Private Dataset

Use case: credit lending

220,304 applications

8 features to determine whether the credit application is accepted or rejected.

Gender and citizenship: are constrained to be immutable features (they cannot change in any way)

Age and bank seniority: can only increase

Rating: feature non-actionable but that can vary due to changes in other variables.

Evaluation – Metrics

Metrics

Feature Space: metrics on counterfactual **explanations**

- **Validity:** the fraction of generated explanations that are valid counterfactuals, i.e. that are given a different outcome y with respect to the factual instance
- **Proximity:** the distance between the original instance and the counterfactual explanation (categorical and continuous variables)
- **Sparsity:** the number of features that need to change with respect to the original input
- **Distance:** L1 distance between counterfactual and factual observations

Latent Space: metrics on recommended **actions**

- **Cost:** L1 norm of the action that has to be done in order to reach a counterfactual explanation
- **Feasibility:** the percentage of actions that are compatible with the feasibility constraints over features

Comparison Results:

Evaluation – Results

	baseline	CEILS
validity	22%	82%
continuous proximity	289.57 ± 830.79	43.23 ± 109.46
categorical proximity	0.0 ± 0.0	0.09 ± 0.15
sparsity	2.86 ± 0.95	2.83 ± 1.17
sparsity action	-	2.28 ± 1.04
distance	2.16 ± 1.1	1.72 ± 0.87
cost	2.51 ± 1.24	1.35 ± 0.81
feasibility	0.064	1.0

- Validity: the fraction of generated explanations that are valid counterfactuals (i.e. that are given a different outcome y with respect to the factual instance).
  → baseline: rating is effectively immutable
  → CEILS: rating is non-actionable but mutable (can change due to changes in other features)

Conclusions

Counterfactual Explanations as Interventions in Latent Space (CEILS) pursues a twofold goal:

1. take into account causality in generating counterfactual explanations and to employ them to provide feasible recommendations
2. having the important characteristic of being a methodology easily adaptable on top of existing counterfactual generator engines.

The experimental results show that there are cases in which the baseline generator would recommend explanation completely unfeasible with respect to the underlying causal structure.

This is a first attempt in the direction of the ambitious target of providing to end users with realistic explanations, feasible recommendations and with less effort to gain the desired output in automatic decision making processes.

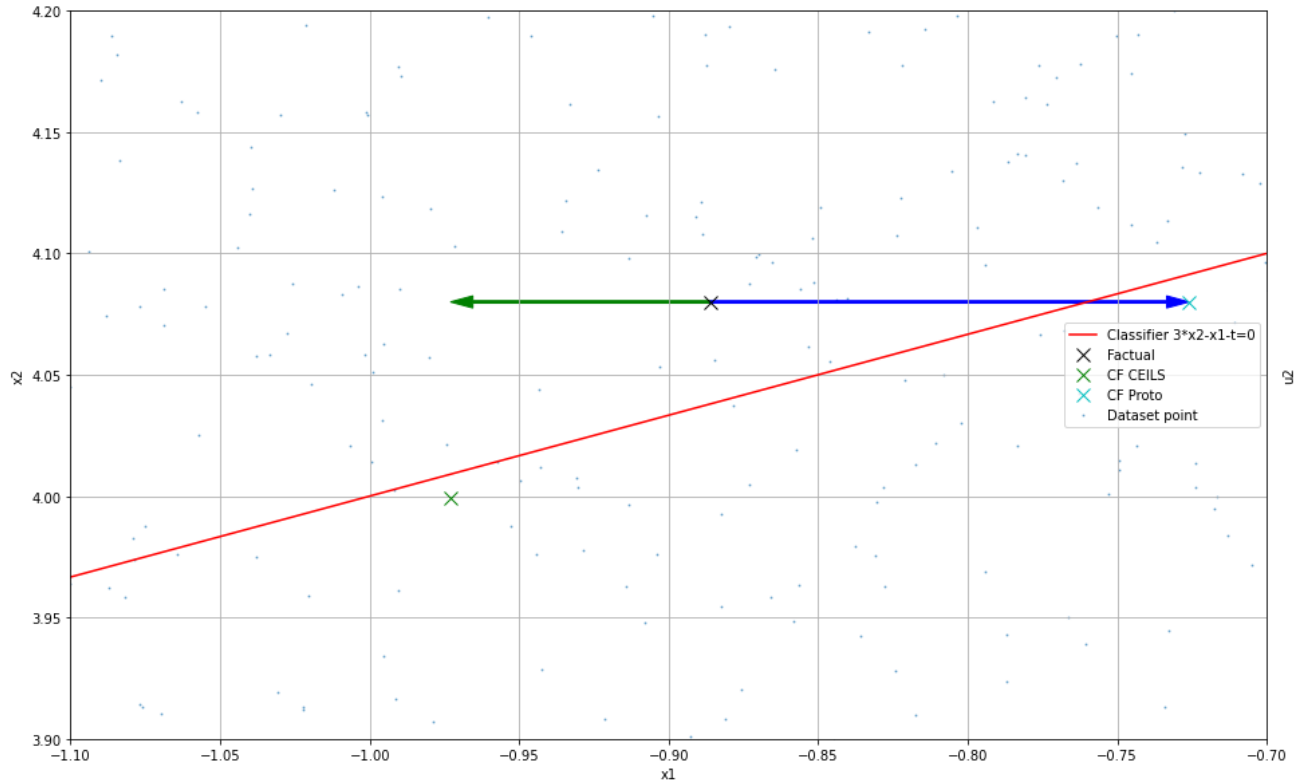
Counterfactual Explanations as Interventions in Latent Space

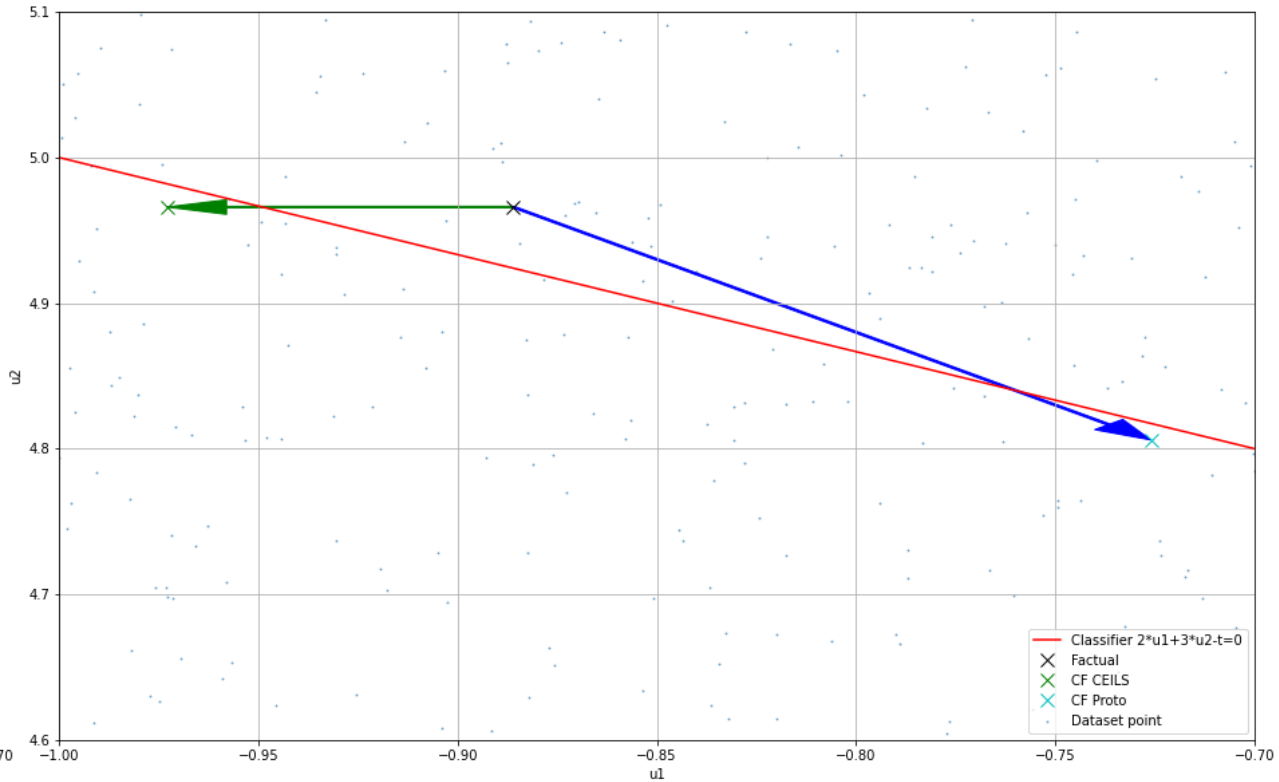
Thank you for your intervention!

Riccardo Crupi, Alessandro Castelnovo, Daniele Regoli, Beatriz San Miguel Gonzalez

Imperial College London

14 Feb 2022


Classifier 3*x2-x1-t=0
Factual
CF CEILS
CF Proto
Dataset point
x1
x2


Classifier 2*u1+3*u2-t=0
Factual
CF CEILS
CF Proto
Dataset point
u1
u2