GENERATING ALGORITHMIC RECOURSE USING USER PREFERENCE

SOUMYA SARKAR

UNDER THE GUIDANCE OF

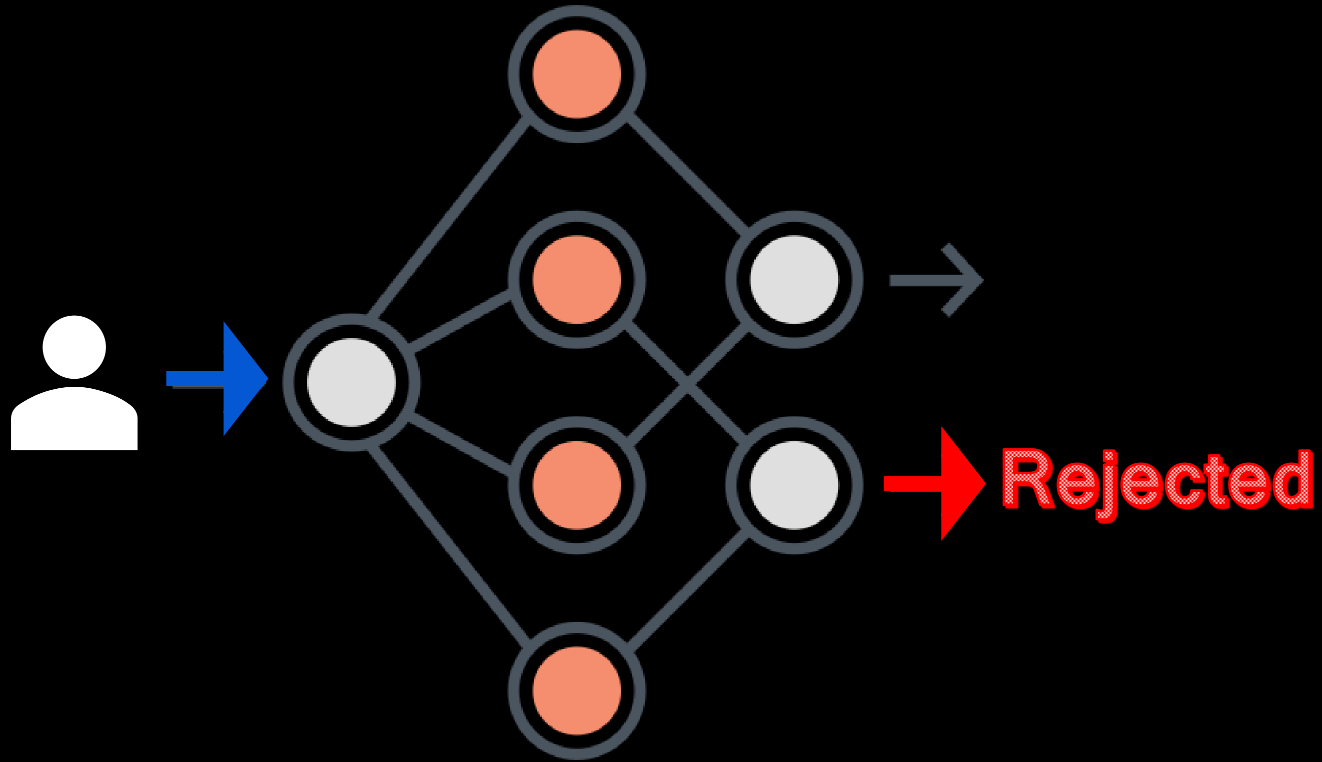
DR. SHWETA JAIN

भारतीय प्रौद्योगिकी संस्थान रोपड़
INDIAN INSTITUTE OF TECHNOLOGY ROPAR
धियो यो नः प्रचोदयात्

THE WHAT?

THE WHY?

THE HOW?



Why does my model show that the shortest path is through a river?

Why was this investment of mine flagged for being a potential loss?

Why does this model show a lower expected range for women job-seekers as compared to men?

Why am I being recommended this song over and over again?

Is this even fair? Is this model discriminatory?

How smart is my model and how well does it generalize?

WHY WAS MY LOAN REJECTED?

Why did the model say I am a potential criminal?

Why does my smart band say I might potentially have a terminal disease?

How well does this algorithm truly work over the expected data?

How is the model calculating my age of marriage as 55?

Why am I not eligible for that prestigious scholarship?

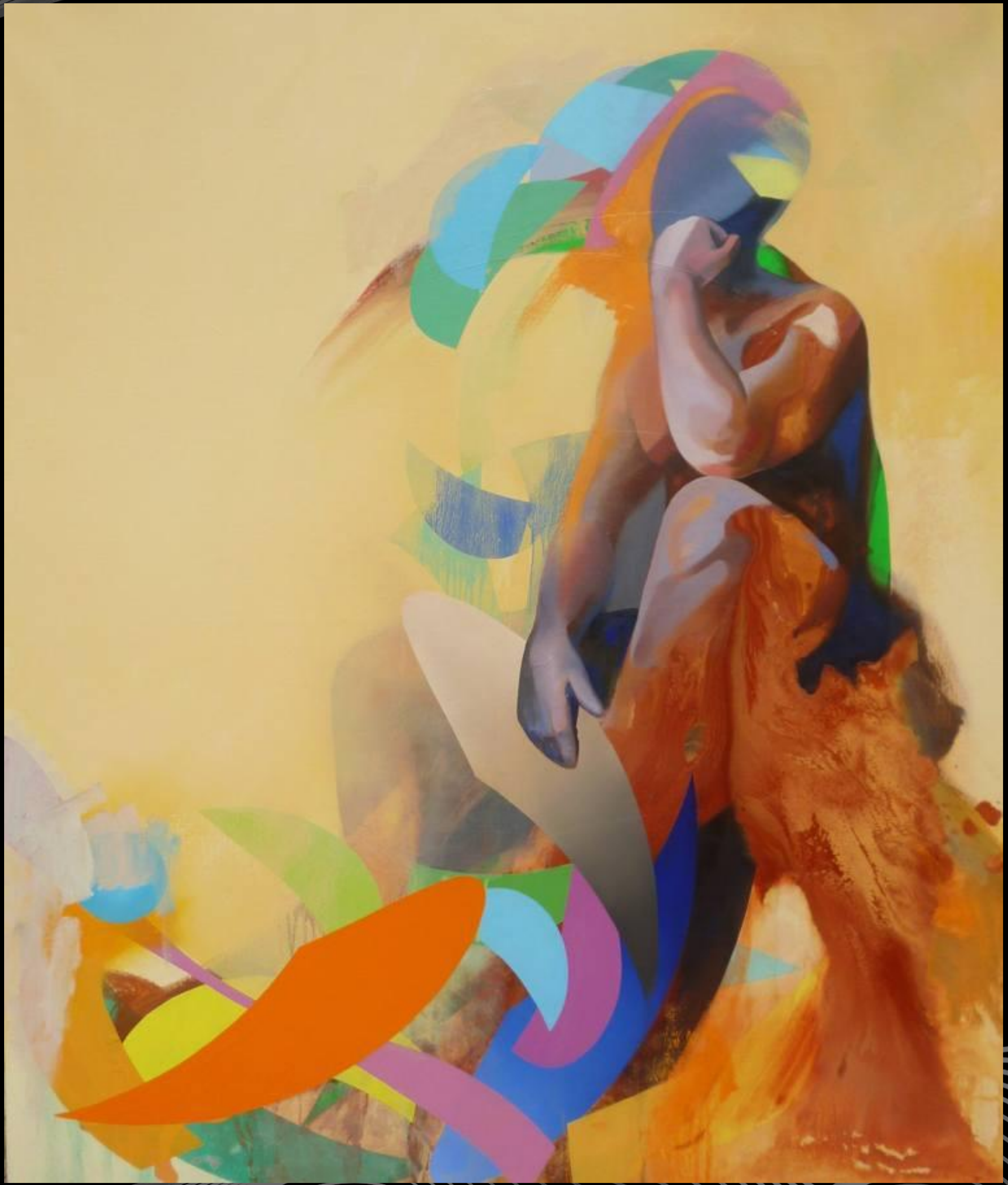
So how do I ensure that this time, I *do* pass the GATE cutoff?

Why was my driver's license denied?

How was my resume sorted by the automated system and why was I out?

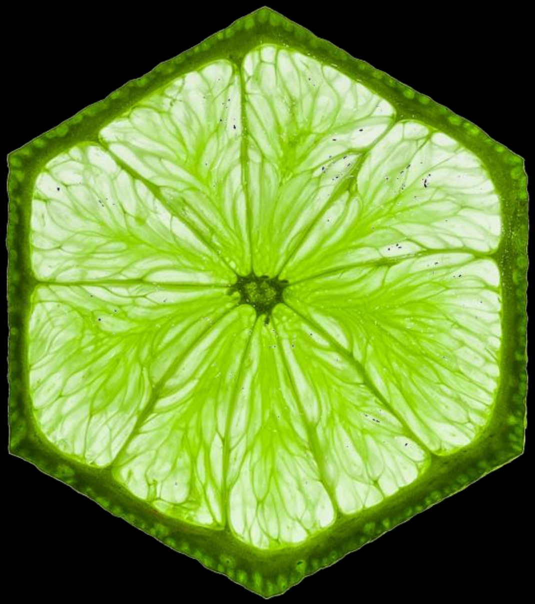
How well does this algorithm truly work over the expected data?

THE WHAT?
THE WHY?
THE HOW?

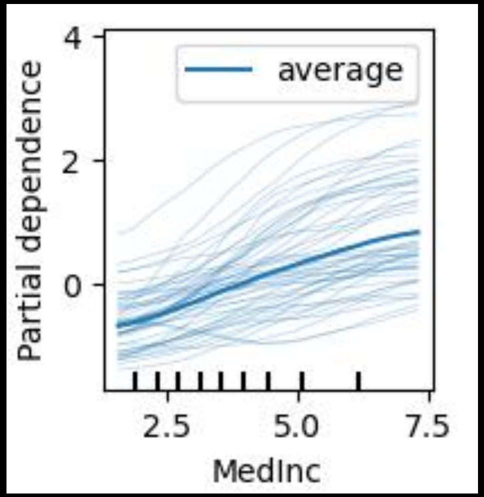


ENTER: EXPLAINABLE AI

SHAP - A global and local explanation technique

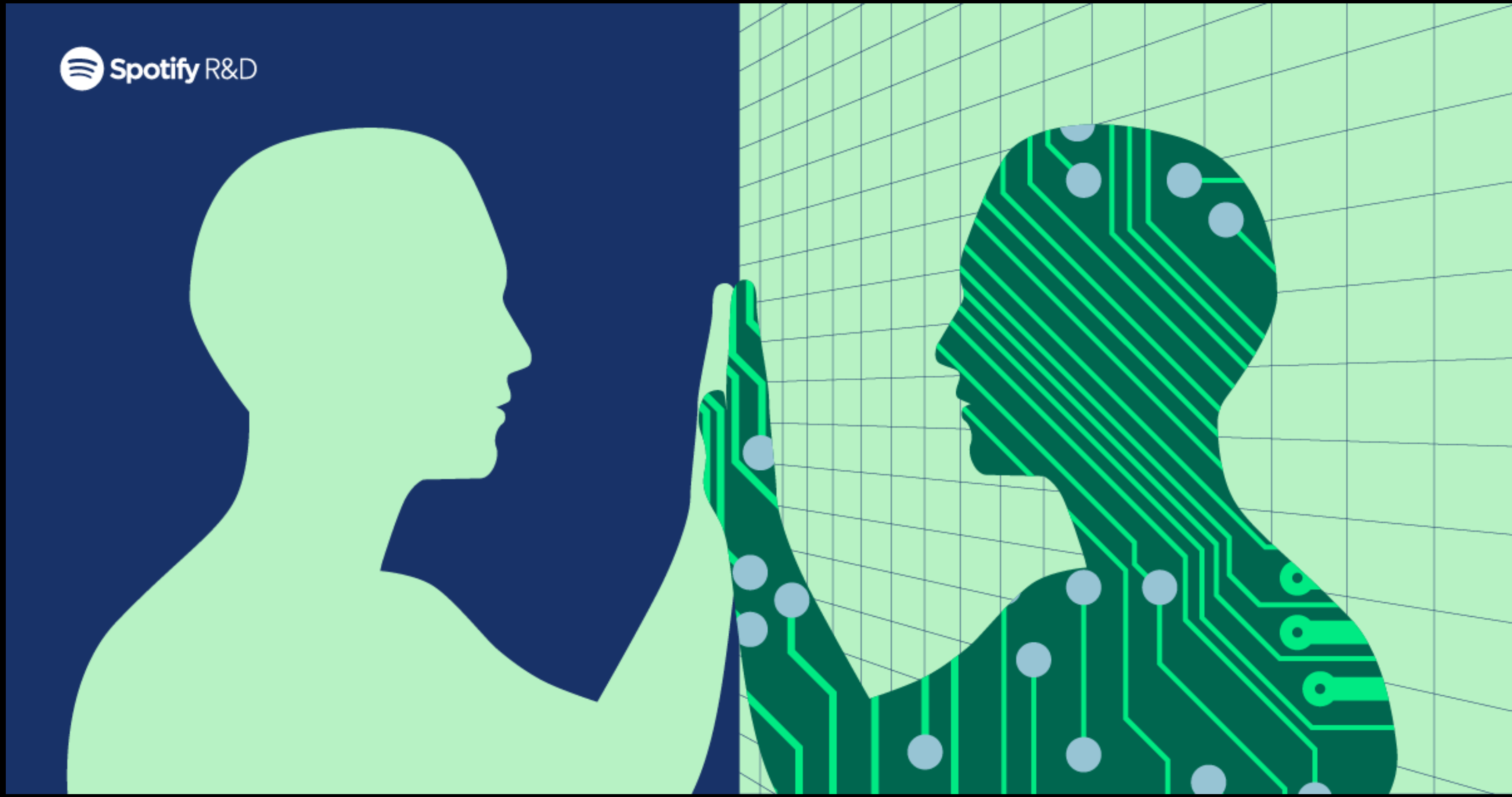


LIME - A local explanation technique

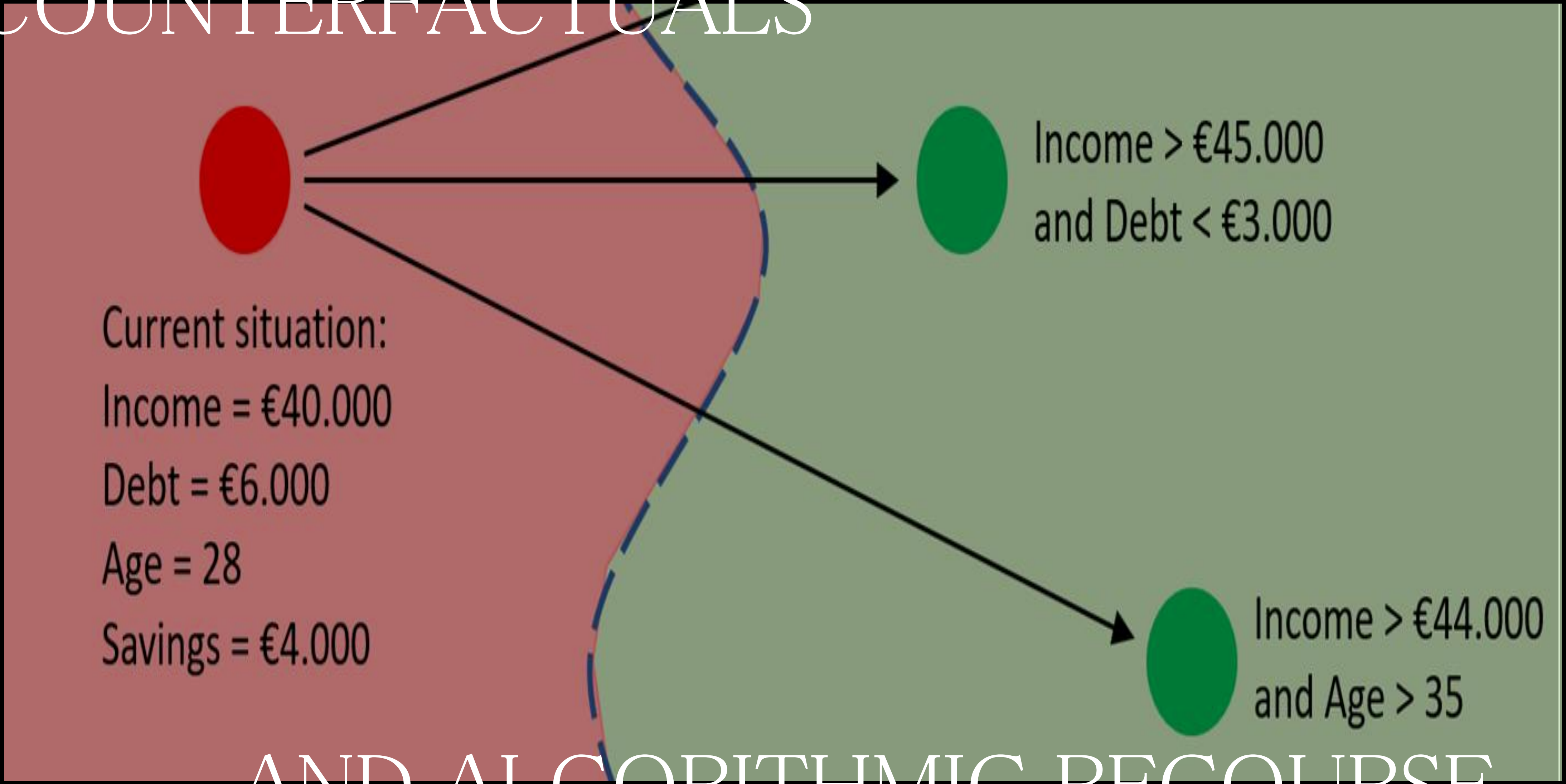


PDP - A global explanation technique

THE WHAT? THE WHY? THE HOW?



COUNTERFACTUALS



AND ALGORITHMIC RECOURSE

COUNTER FACTUALS

A lot of use cases are cropping up

The Economist | Menu | Weekly edition | The world in brief | Search

Finance & economics | Might-have-beens

A Lloyd's report urges insurers to ask "what if?"

Counterfactual risk analysis might improve underwriting

THE Times Higher Education | Home | News | Rankings | Jobs | Students | Events | Resources

Home | Latest | Opinion | In-depth | Leadership

Counterfactual history: why what didn't happen matters

The counterfactual approach can open up fascinating new perspectives and give a voice to the neglected 'losers' of history, professors say

November 28, 2021

scientific reports

Explore content | About the journal | Publish with us

nature > scientific reports > articles > article

Article | Open access | Published: 18 October 2023

Evaluating the COVID-19 vaccination program in Japan, 2021 using the counterfactual reproduction number

Taishi Kayano, Yura Ko, Kanako Otani, ...

Scientific Reports 13, Article number: ...

33k Accesses | 3135 Altmetric | Me...

scientific reports

Explore content | About the journal | Publish with us

nature > scientific reports > articles > article

Article | Open access | Published: 04 September 2023

Counterfactual scenarios reveal historical impact of cropland management on soil organic carbon stocks in the United States

Stephen M. Ogle, F. Jay Breidt, Stephen Del Grosso, Ram Gurung, Ernie Marx, Shannon Spencer,

KPMG | Themen | Br...

Counterfactual Explanations: The What-Ifs of AI Decision Making

Counterfactuals: Demystifying AI decision-making for greater clarity.

MIT Technology Review | Featured | Topics | Newsletters | Events | Podcasts | SIGN IN | SUBSCRIBE

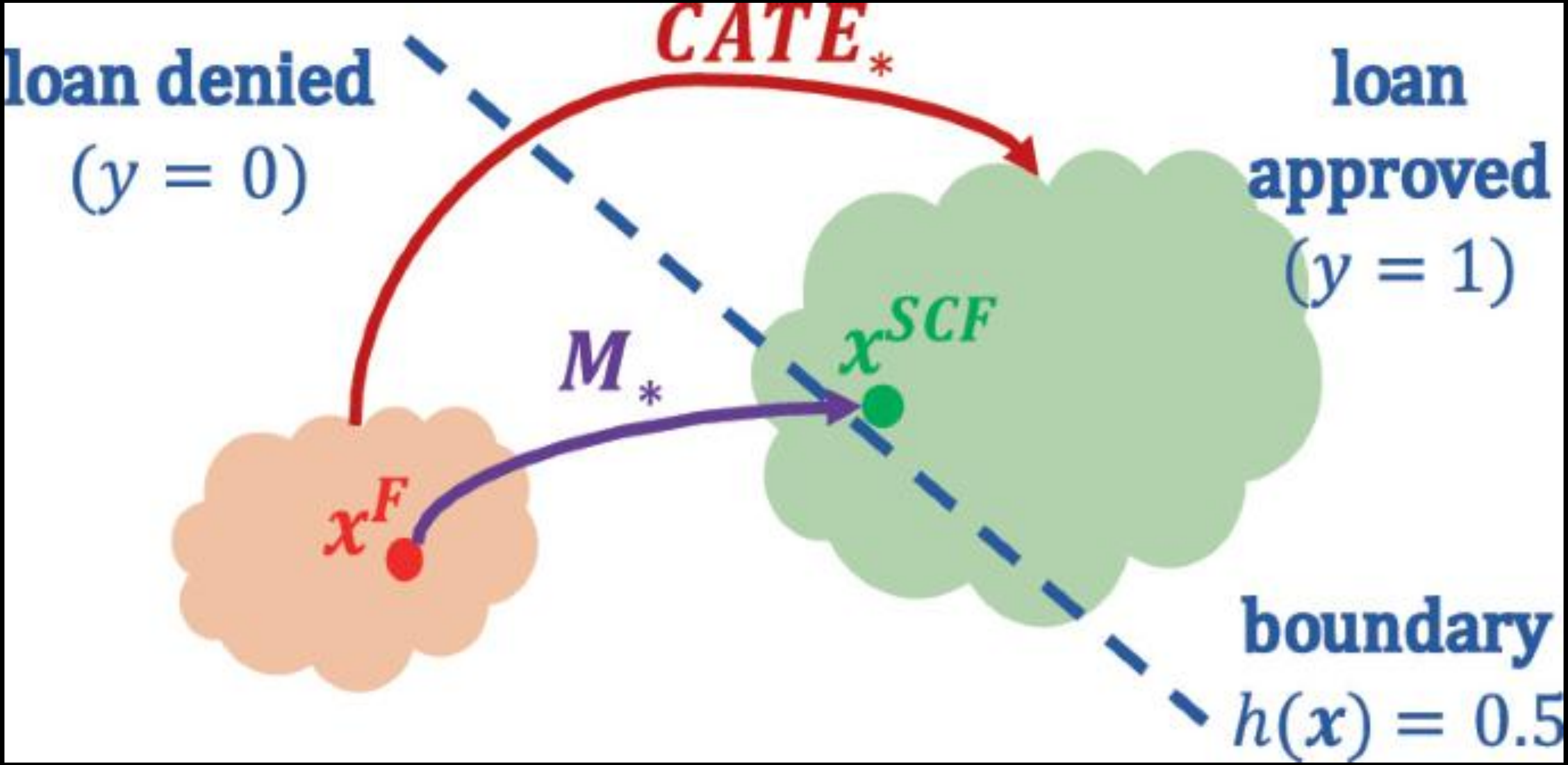
ARTIFICIAL INTELLIGENCE

The complex math of counterfactuals could help Spotify pick your next favorite song

A new kind of machine-learning model is set to improve automated decision making in finance, health care, ad targeting, and more.

By Will Douglas Heaven

April 4, 2023



ALGORITHMIC RECOURSE

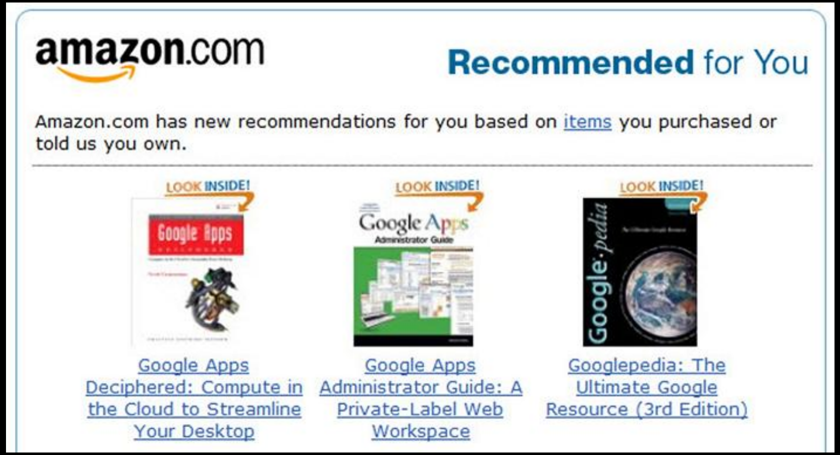
Now your loan too can be approved!

A BIT ABOUT HUMAN PREFERENCES

EVERYONE HAS IT, BUT NOBODY CAN QUANTIFY IT

What is the best way to ask users for preferences, especially their inherent resistance towards changes?

Which ice-cream would you choose?



Amazon recommendations

Voting

CAUSALITY

CORRELATION IS NOT CAUSATION!

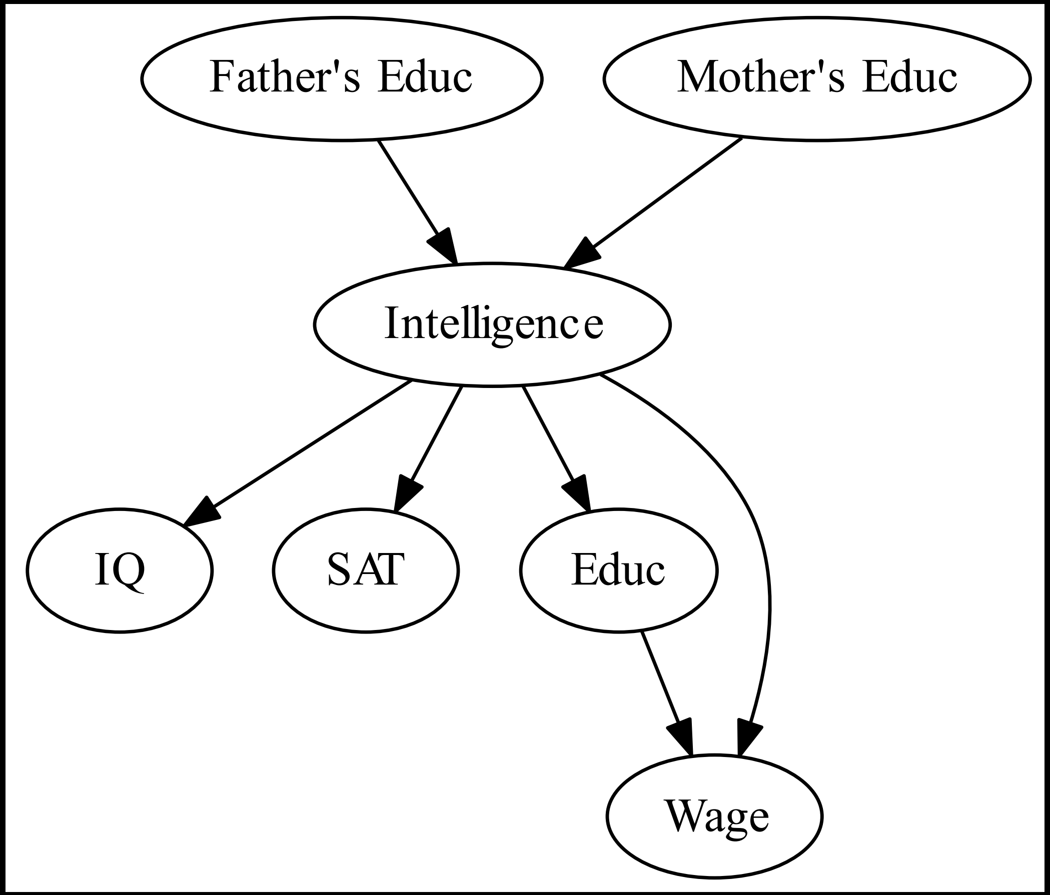
But what *is* **causation**?

So does more ice cream being sold mean more soccer games being played?

CAUSALITY IS IMPORTANT

Effects of your actions can trickle downstream.

Ignoring it has been proven to lead to suboptimal recourse [1].



A causal graph

- [1] - Karimi, A. H., Schölkopf, B., & Valera, I. (2021, March). Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 353-362).

LITERATURE SURVEY

COUNTER FACTUAL GENERATION

Plenty of work, but several assumptions needed

Algorithmic Recourse based on User's Feature-order Preference

Manan Singh*
IIT Palakkad, India, India
142214003@smail.iitpkd.ac.in

Sai Srinivas Kancheti*
IIT Hyderabad, India, India
cs21resch01004@iith.ac.in

Shivam Gupta*
IIT Ropar, India, India
shivam.20csz0004@iitrpr.ac.in

Ganesh Ghalme
IIT Hyderabad, India, India
ganeshghalme@ai.iith.ac.in

Shweta Jain
IIT Ropar, India, India

Narayanan C. Krishnan
IIT Palakkad, India, India

ABSTRACT

The state-of-the-art recourse generation
the user's profile (feature vector). Howev
same profile may still have different prefe
recourse generated from a single profile m
peal to both the users. For example, one rej

Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses

Synthesizing explainable counterfactual policies for algorithmic recourse with program synthesis

Giovanni De Toni[1,2] · Bruno Lepri[1] · Andrea Passerini[2]

Algorithmic Recourse: from Counterfactual Explanations to Interventions

Amir-Hossein Karimi
MPI-IS, Germany
ETH Zürich, Switzerland

Bernhard Schölkopf
MPI-IS, Germany

Isabel Valera
MPI-IS, Germany
Saarland University, Germany

COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR

Sandra Wachter,* Brent Mittelstadt,** & Chris Russell***

Consequence-aware Sequential Counterfactual Generation

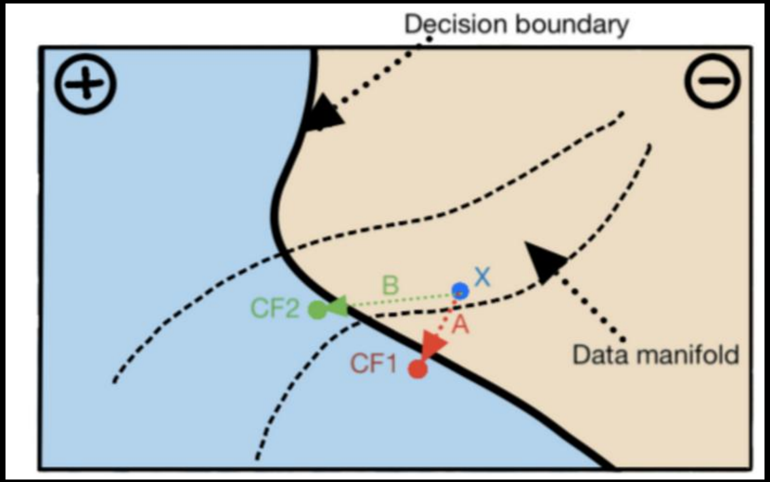
Philip Naumann[1,2] (✉) and Eirini Ntoutsi[1,2]

[1] Freie Universität Berlin, Germany
[2] L3S Research Center, Leibniz Universität Hannover, Germany
{philip.naumann, eirini.ntoutsi}@fu-berlin.de

Personalized Algorithmic Recourse with Preference Elicitation

THREE MODULES

DoWhy



Preference taken from user using a novel duelling bandits-based algorithm

Causal Discovery using the causal-learn and dowhy packages for python

Counterfactual generation based on causality, user preference and other restrictions

THE BRADLEY–TERRY MODEL

A SIMPLE HEURISTIC SATISFYING SEVERAL KEY PROPERTIES IN A BANDIT PERSPECTIVE

- Strong stochastic transitivity
- Stochastic triangle inequality

$$P(b_i > b_j) = \frac{\mu_i}{\mu_i + \mu_j}$$

THE COST OF INTERVENTIONS

HOW COST IS CALCULATED

SMALLEST COST IN ALL SEQUENCE OF ACTIONS

- L2 cost
- Classification loss (BCE)
- A Reduction Factor
- ~~Cost of children of features being intervened on.~~

$$\mathcal{I}^* = \arg\min_{\mathcal{I}} \mathbb{C}(\mathcal{I}, x) \quad \text{such that } h(\mathcal{I}^*(x)) \neq h(x) \tag{3.3}$$

where we define the cost function $\mathbb{C}(\mathcal{I}, x)$ as:

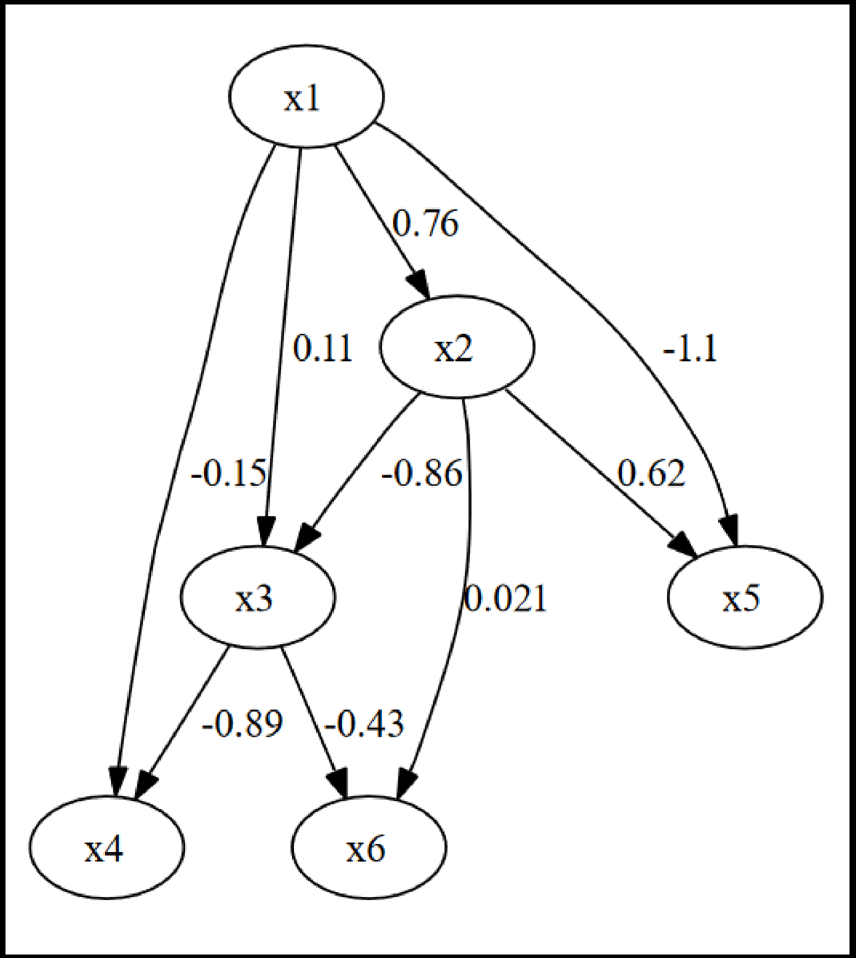
$$\mathbb{C}(\mathcal{I}, x) = \sum_{a_i \in \mathcal{I}} C(a_i, x) * R_{factor} \tag{3.4}$$

$$\text{where } C(a_i, x) = \lambda \|a_i(x_i) - x_i\|_2 + L_{classification}(a_i(x))$$

$$\text{and } R_{factor} = 0.85(1 - 0.3 * y_{pred})$$

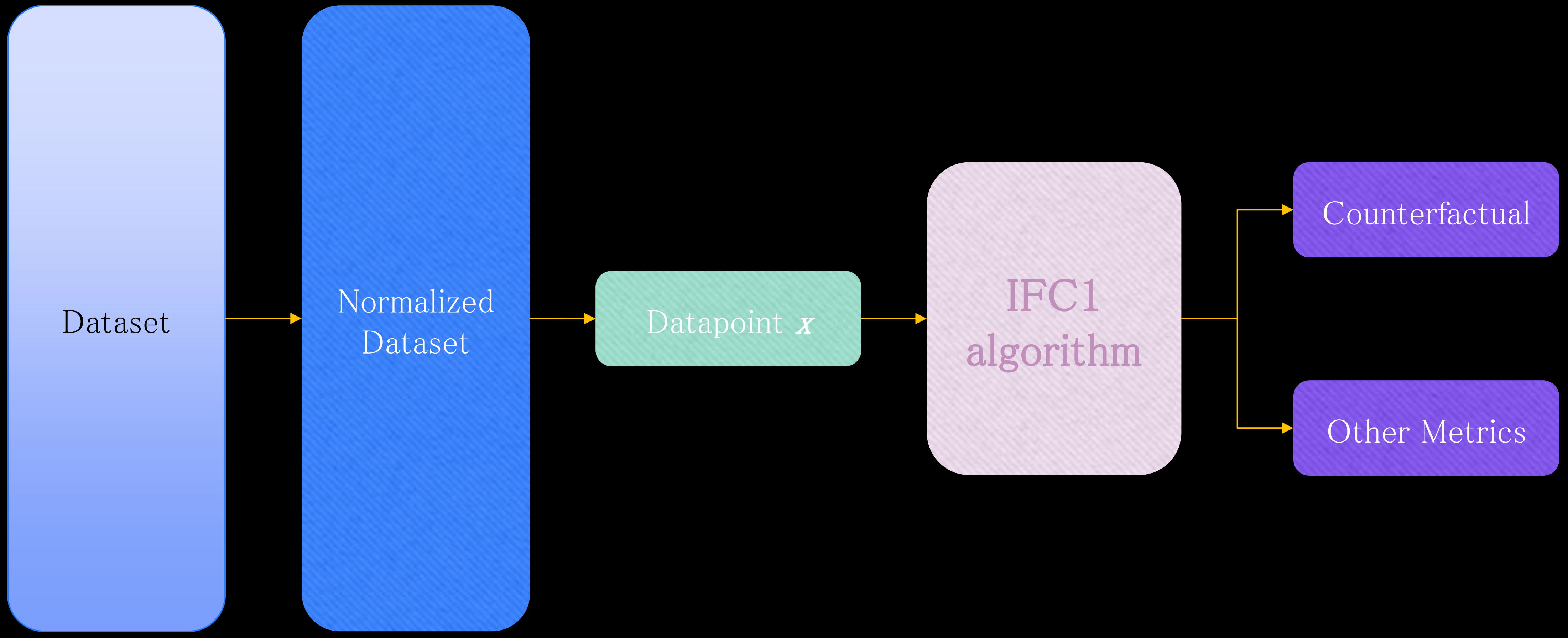
CAUSAL INFERENCE

WE USE ICA-BASED LINGAM [2] FOR ALL INFERENCE PURPOSES

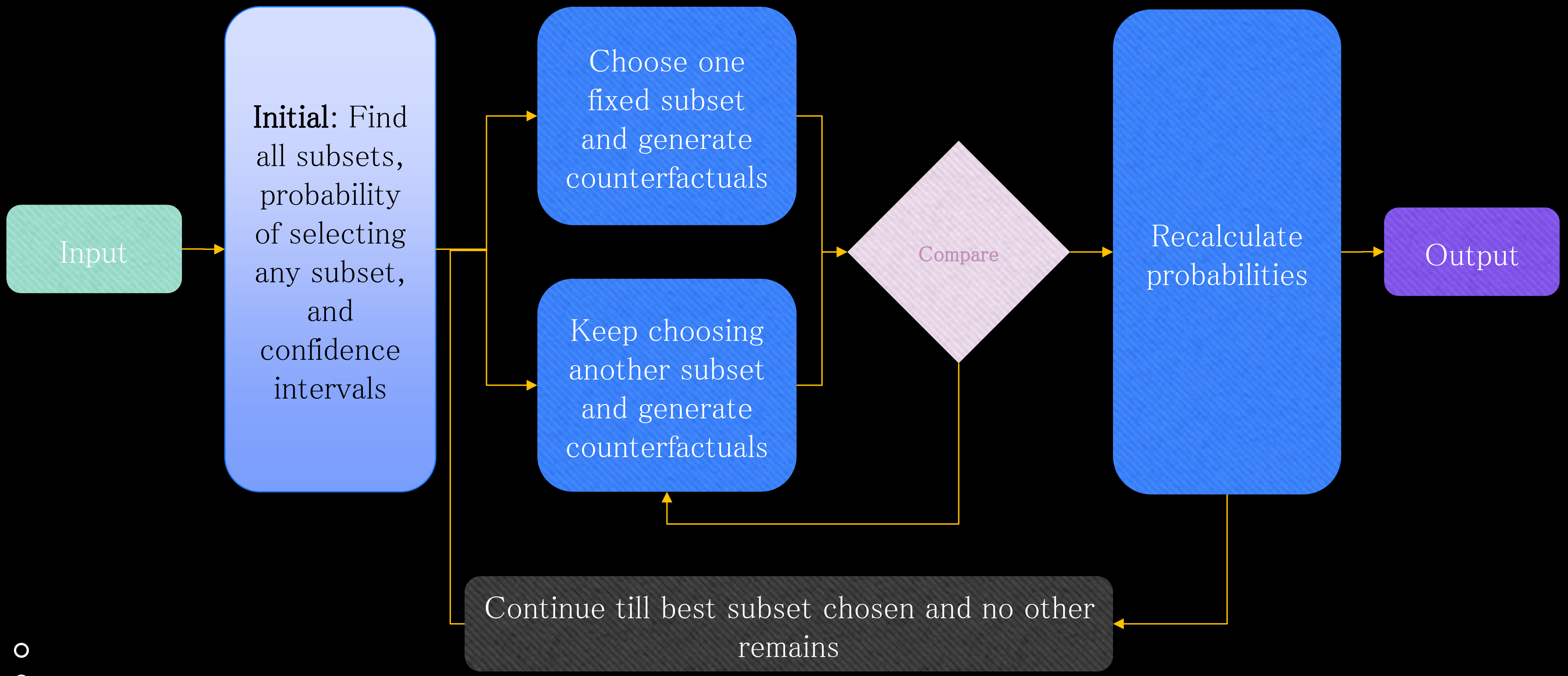


[2] - Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., & Jordan, M. (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7(10).

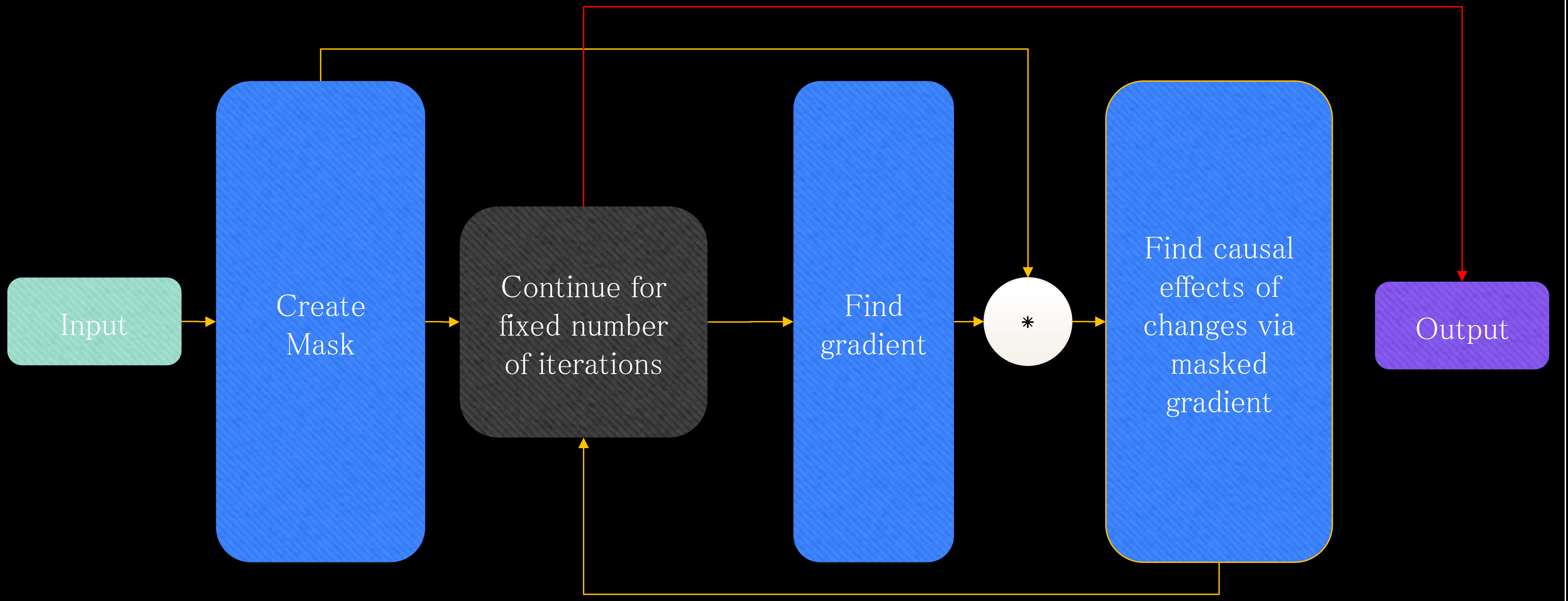
METHODOLOGY



IFC1 ALGORITHM



COUNTERFACTUAL GENERATION ALGORITHM



EXPERIMENTAL RESULTS

- Experimental Results on Synthetic Dataset
- Experimental Results on Real World Dataset (Give Me Some Credit)
- For Bradley-Terry model, $\mu_i = 1, \mu_j = 1$

Hyperparamaters	**Custom Dataset**	**Give Me Credit**
Number of users	100	50
Size of dataset sampled	1000	5000
Learning Rate	10^{-3}	10^{-3}
Normalization	Z-score	Z-score
Maximum number of permutations of subset	3	3
(Custom 1 only) $\sum \alpha$	1000	1000
(Custom 1 only) α multiplier	100	100

EXPERIMENTAL RESULTS: CUSTOM 1 AND CUSTOM 2

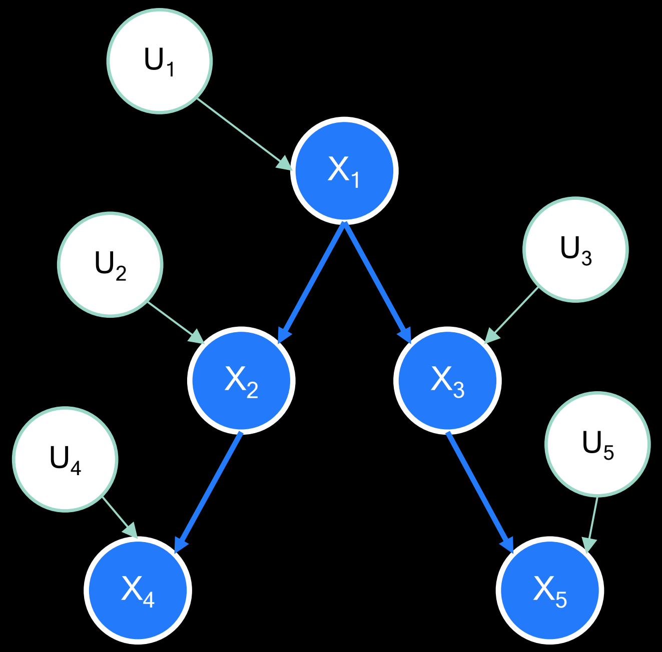
$$\arg\min_{x'}\max_{\lambda} \lambda \mathbb{L}(f_w(x') - y') + d(x_i, x')$$

$$d(x_i, x') = \sum_{i=0} \alpha_i \|x_i - x\|_1 \tag{3.7}$$

where $\alpha_{i+1} = c.\alpha_i$, and $\alpha_1 = \frac{\beta_0(c-1)}{c^n-1}$. Here c is known as the alpha multiplier, and $\sum_i \alpha_i = \beta_0$ is the sum of all alphas.

SYNTHETIC DATASET

$$X_1 := U_1$$
$$X_2 := 2X_1 + U_2$$
$$X_3 := 3X_1 + U_3$$
$$X_4 := X_2 - 2X_3 + U_4$$
$$X_5 := 2X_3 - 2X_1 + U_5$$



$$y = \begin{cases} 1 \;\; if \;\; sigmoid(x_1 + x_2 + x_3 + x_4 + x_5) \geq 0.5 \\ 0 \quad otherwise \end{cases}$$

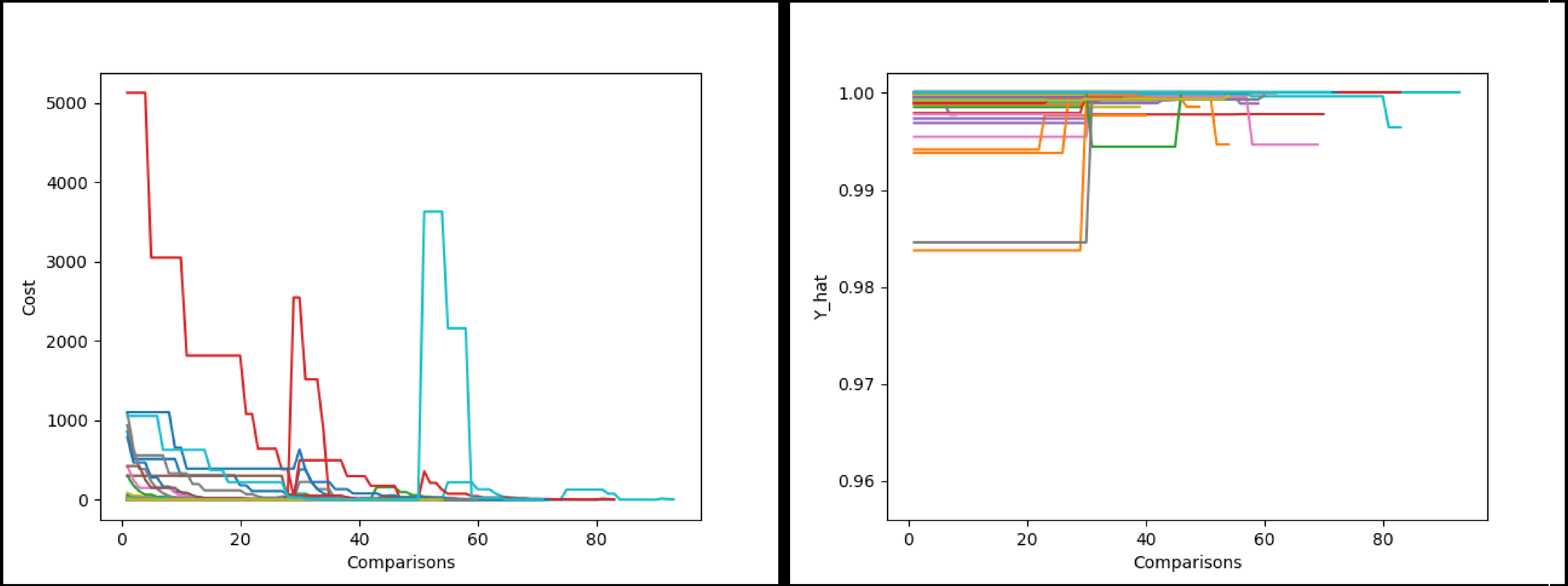
EXPERIMENTS ON SYNTHETIC DATASET

Paramater	**Manan**	**Manan with Causality**	**Custom 1**	**Custom 2**	**Our Method**
Average Number of Features Changed	1.08	1	5	1	3.42
Average L_2 Cost of Counterfactuals	14.007	896.582	1.14	21.54	6.34
Validity Percentage	100	100	80	100	50
Average Time to Execute (s)	1.222	1.331	0.771	9.002	116.037

EXPERIMENTS ON SYNTHETIC DATASET

Paramater	**Custom 2**	**Our Method**
Average number of comparisons	40.94	45.78
Average percentage of cases it failed to detect user preference/changed fixed features	100	4

EXPERIMENTS ON SYNTHETIC DATASET



EXPERIMENTS ON SYNTHETIC DATASET

```
2024-05-08 02:33:38.927548: y_pred = tensor([0.]) at index 106 with datapoint [-1.5935743 -2.250385  -1.6981801  1.0659246 -1.8776345].

2024-05-08 02:33:38.927548: Final y_pred = tensor([0.]) at index 106.

2024-05-08 02:34:23.670241: Original data point is [-1.5935743 -2.250385  -1.6981801  1.0659246 -1.8776345]
2024-05-08 02:34:23.670241: The counterfactual is given by [-1.5935743 -2.250385  -1.6981801  1.0659246  7.097441 ] with number of features changed = 3, and cost = 2.501935391262246
2024-05-08 02:34:23.670241: Subset chosen by algorithm is ['X1', 'X3', 'X5'], whereas actual subset chosen by user is ['X1', 'X5']. (Fixed features = [])
2024-05-08 02:34:23.670241: Time taken 36.9246928691864 seconds
2024-05-08 02:34:23.670241: Prediction = 1.0
```

We get successful recourse, and subset prediction which is at least a superset (thereby not missing out on any features in the user's preference order).

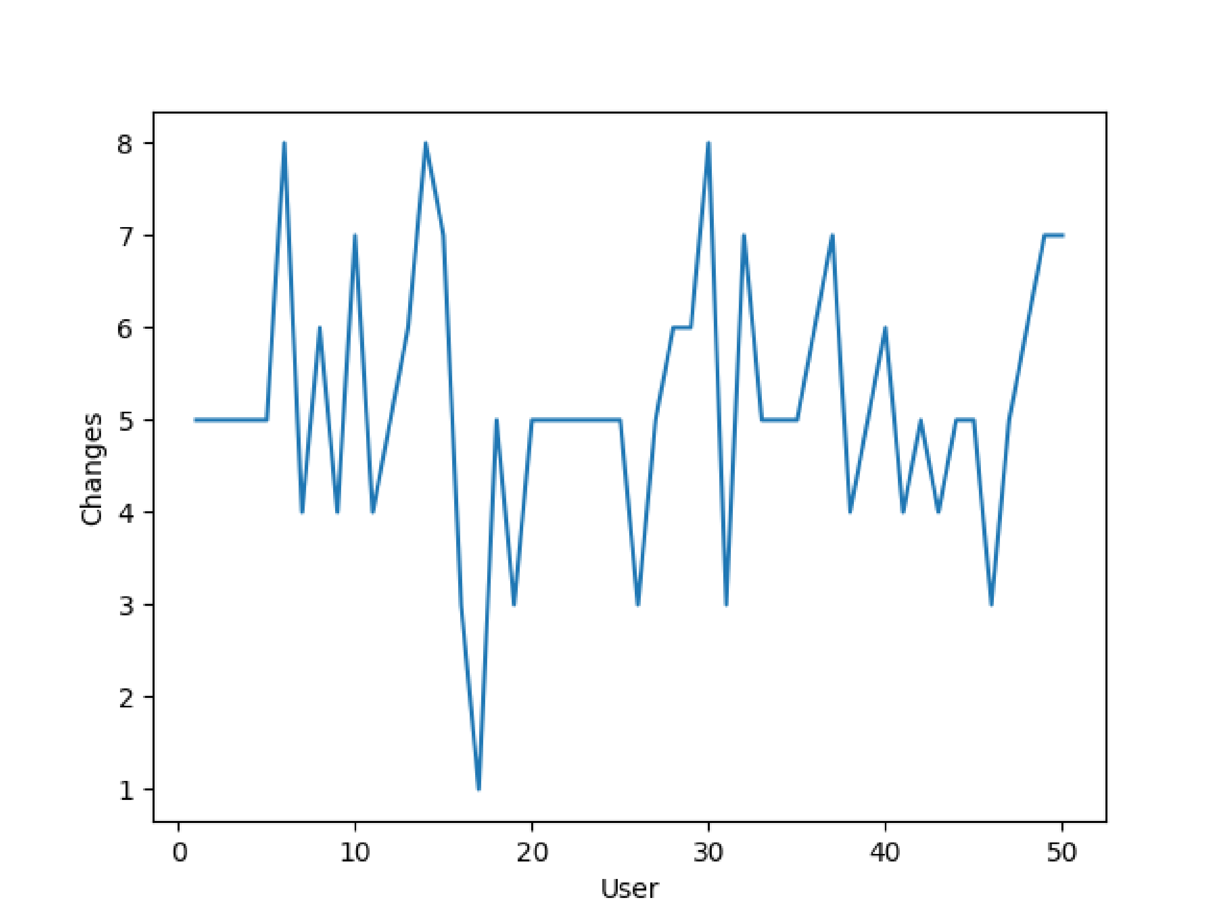
EXPERIMENTS ON GIVE ME SOME CREDIT DATASET

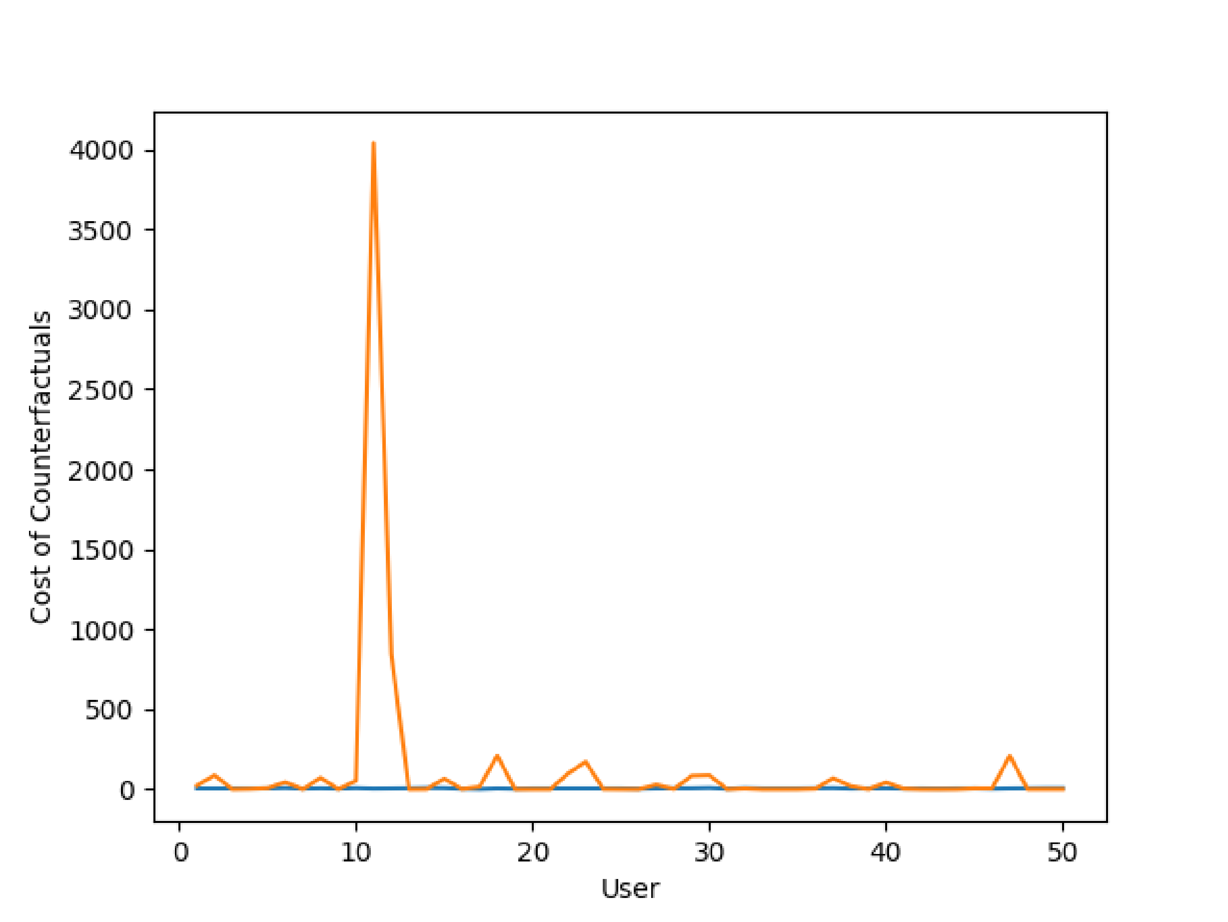
Paramater	**PEAR**	**CSCF**	**FACE**	**My Method**
Average Number of Features Changed	2.79 ± 0.42	2.51 ± 1.12	5.97 ± 0.62	5.16
Average L_2 Cost of Counterfactuals	96.04 ± 31.96	100.69 ± 120.22	327.18 ± 78.85	125.922
Validity Percentage	89	0.57 ± 0.42	0.24 ± 0.38	100

EXPERIMENTS ON GIVE ME SOME CREDIT DATASET

Paramater	**Our Method**
Average number of comparisons	477.5
Average percentage of cases it failed to detect user preference/changed fixed features	0
Average time to execute in seconds	420.583

EXPERIMENTS ON GIVE ME SOME CREDIT DATASET





EXPERIMENTS ON GIVE ME SOME CREDIT DATASET

Fixed features: ['DebtRatio']
User Preferred subset: ['age', 'NumberOfTime30-59DaysPastDueNotWorse', 'MonthlyIncome', 'NumberOfDependents']
Output subset: ['age', 'NumberOfTime30-59DaysPastDueNotWorse', 'MonthlyIncome', 'NumberOfTime60-89DaysPastDueNotWorse', 'NumberOfDependents']
Number of comparisons:750

We get successful recourse, and subset prediction which is at least a superset here as well.

CONCLUSION AND FUTURE SCOPE

- End-to-end counterfactual generation methodology
- Almost comparable to baselines at the moment, performing better on some metrics
- Can be applied in any situation involving black box models
- Special use case in critical scenarios with scope for modification
- Reducing number of comparisons

FUTURE SCOPE

- Reduction in the number of comparisons
- Modification in the cost function
- More optimization to reduce computations

THANK YOU

THE END

Hope this explained my work so far!