Harnessing Transport Theory for PostHoc *Explanations of Machine Learning*

Lei You, PhD
Assistant Professor in Applied Mathematics
Technical University of Denmark (DTU)

Lei You

Assistant Professor in Applied Mathematics | Data Science

Experience

2022-Now

Assistant Professor
DTU - Technical University of Denmark · Full-time
Dec 2022 - Present · 1 yr 11 mos
Copenhagen Metropolitan Area

(Tenure-track) Assistant Professor in Applied Mathematics
Teaching: Applied Machine Learning, Data Visualization and Analysis, Deve ...see more

2021-2022

Data Scientist, Logistics Optimization
Wolt · Full-time
Nov 2021 - Nov 2022 · 1 yr 1 mo
Stockholm, Stockholm County, Sweden

Doordash is an on-demand food/grocery delivery platform leading global markets. I have been working in the Logistics Optimization team in the brand Wolt. F ...see more

2019-2021

Senior Data Scientist
Bolt · Full-time
Jul 2019 - Nov 2021 · 2 yrs 5 mos
Stockholm, Sweden

Bolt is a unicorn ride-hailing company and one of the leaders in European markets. The scope of my work consists of works on each stage in the data science ...see more

2018-2019

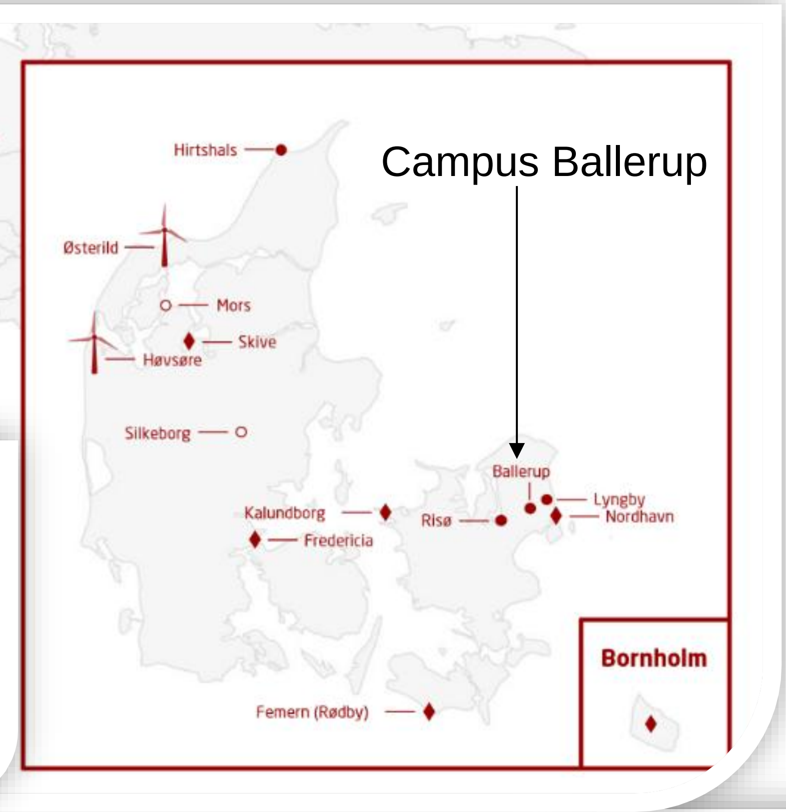
Visiting Data Scientist
Boston Consulting Group (BCG) · Internship
Nov 2018 - Jan 2019 · 3 mos
Stockholm, Sweden

Worked for an international brand of clothing business. Responsible for deriving machine learning models for demand forecast and implementation of bacl ...see more

Education

Uppsala University **PhD, 2015-2019**
Doctor of Philosophy (Ph.D.), Computer Science
2015 - 2019

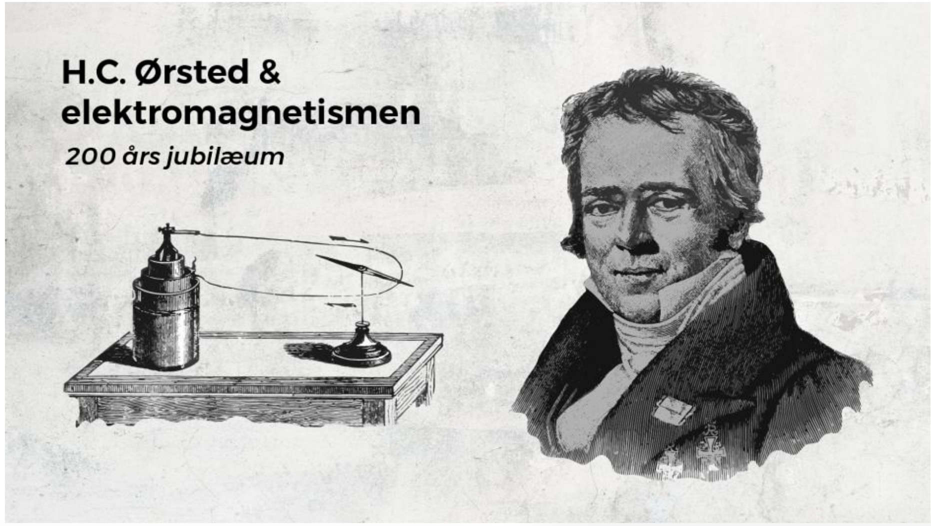
Dissertation Title: *Network Optimization of Evolving Mobile Systems with Presence of Interference Coupling*
Best Dissertation Award in INFORMS, Telecommunications & Network Analytics, 2020



1829

H.C. Ørsted

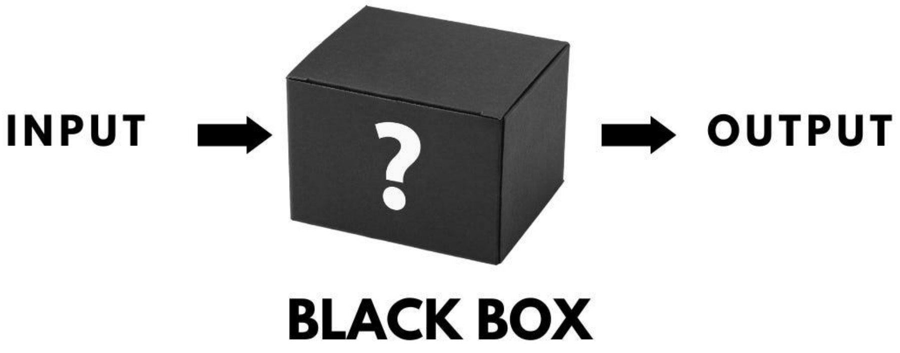
The father of electromagnetism—who founded the university



What Is a Good Explanation?

Explanations are contrastive.

Humans usually do not ask why a certain prediction was made, but **why the prediction is made instead of another prediction**.



Explanations are selected.

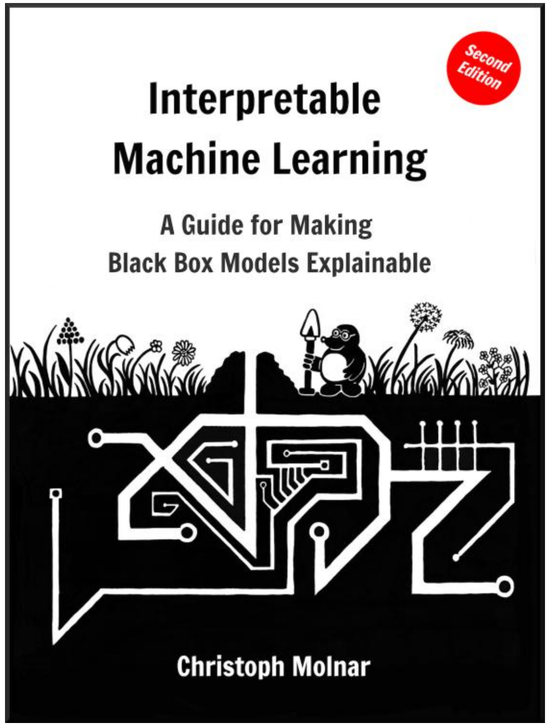
We are used to **selecting one or two causes rather than a variety** of possible causes the THE explanations.

Explanations are instructive.

We are looking for explanations **that can provide practical guidance** to enhance model building, business operations, or individual decision-making.

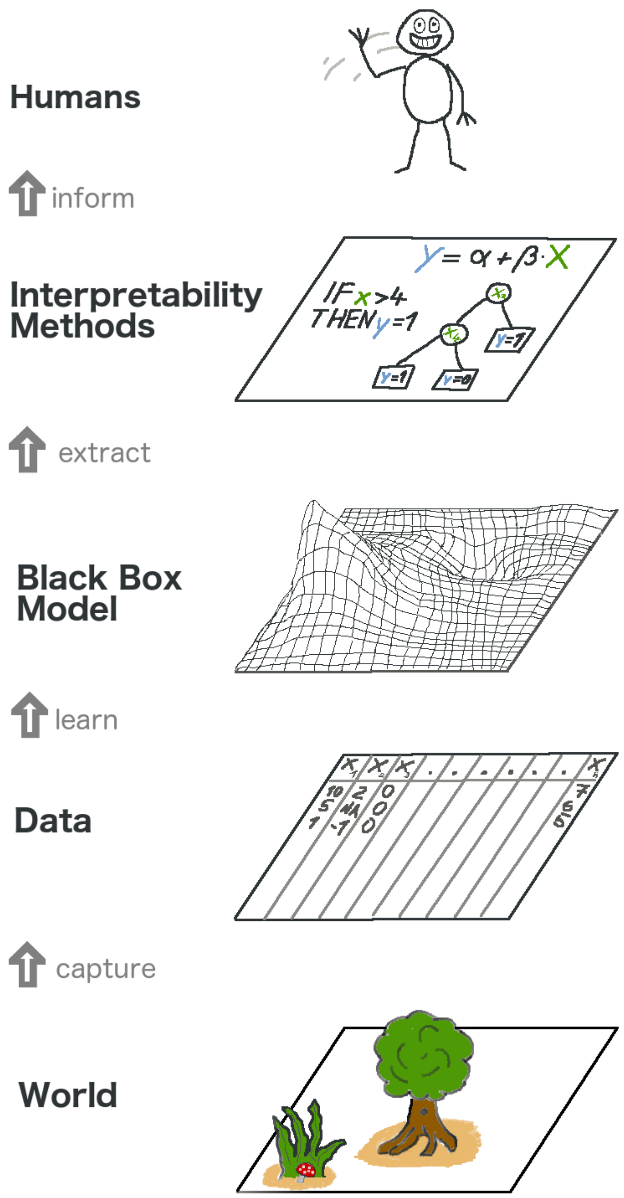
Interpretable Machine Learning

Making Black Box Models Explainable

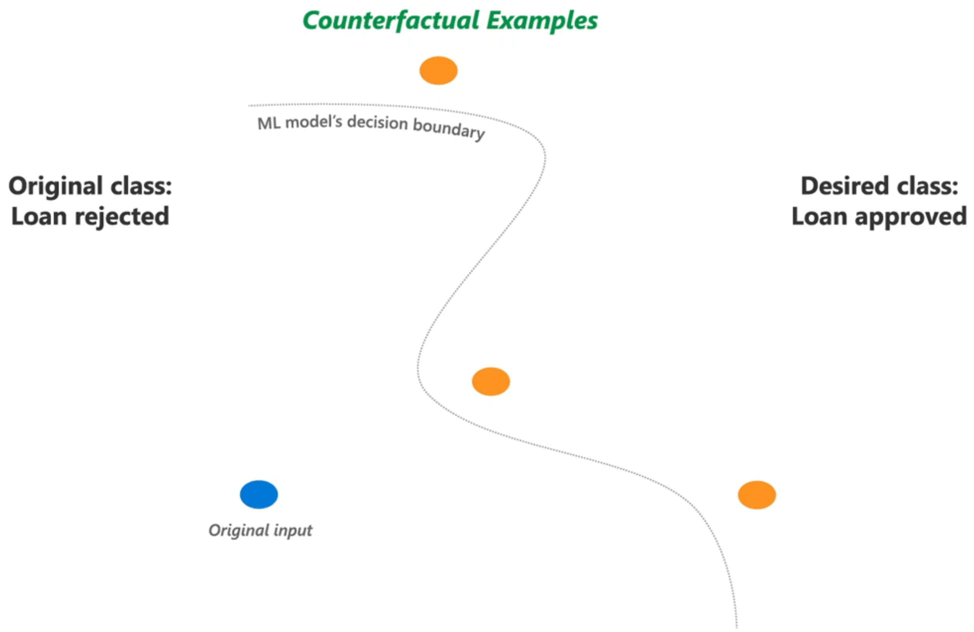


Capturing local/global behavior of Black Box Model

Some Other Model



Counterfactual Explanations (CE)



Factual

$$\mathbf{x}' \rightarrow y'$$

Reality

We expect to find a new data point showing that **small input difference leads to large output difference**

Generating explanations means generating data

Counterfactual

$$\underbrace{\mathbf{x}' + \boldsymbol{\Delta}}_{\text{Hypothetical Reality}} \rightarrow f(\mathbf{x}' + \boldsymbol{\Delta})$$

$$\mathbf{x}' \rightarrow f(\mathbf{x}') \rightarrow y'$$

"causal" relationship

Pioneering Research of CE

Watcher et al., minimizing $\mathcal{L}(x)$

$$\mathcal{L}(x, x', y^*, \lambda) = \lambda(f(x) - y^*)^2 + \|x - x'\|^2$$

Counterfactual output reaches a desired target y^*

Counterfactual resembles the factual

Factual

Some observation in our interest

Counterfactual

To be found by optimization

Solved by gradient desent

$$f(x) = \text{Softmax}(\cdots \text{Relu}(\cdots \text{Relu}(\theta x + b) \cdots) \cdots)$$

Usually $y^* \neq f(x')$

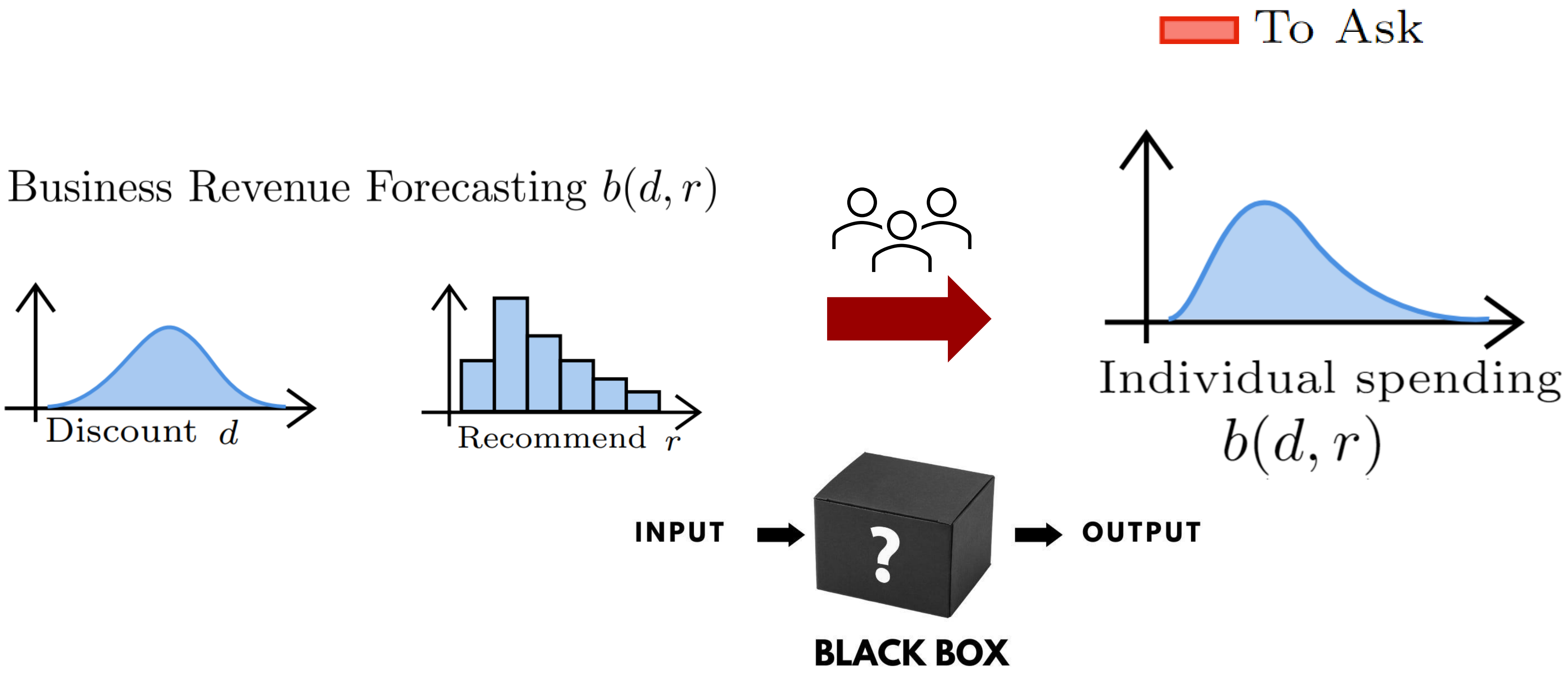
$\frac{\partial f}{\partial x}$ Back-Propagation for x

CE is first introduced

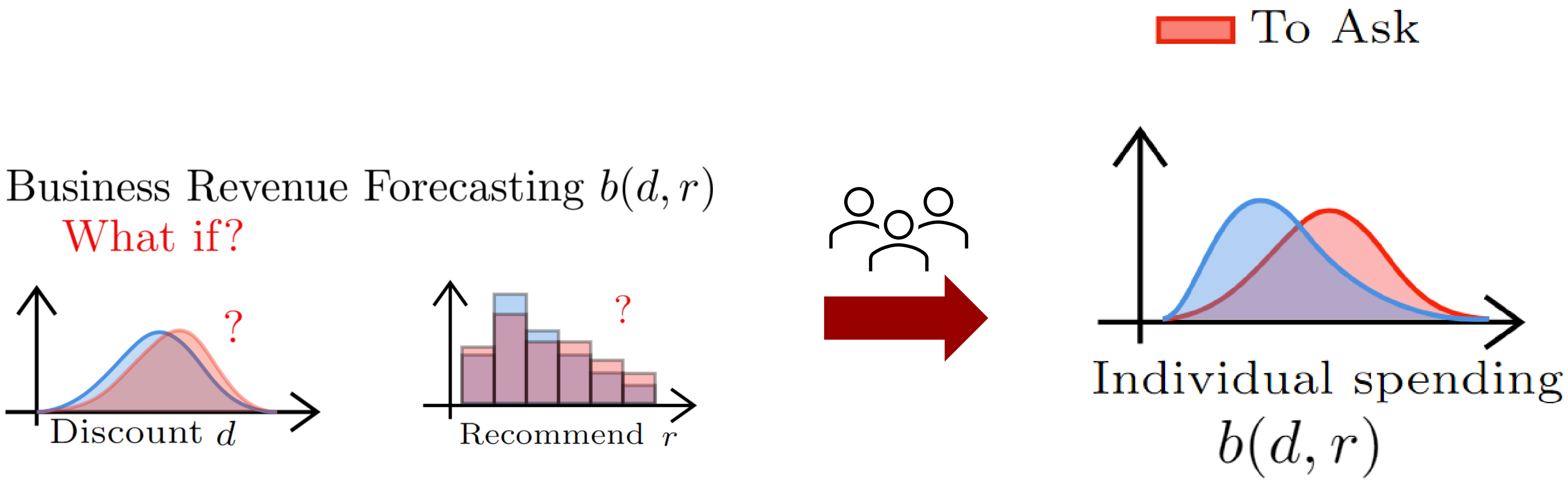
CE is found by solving optimization problem

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." *Harv. JL & Tech.* 31 (2017): 841.

A Stakeholder's (Business Operator's) View



A Stakeholder's (Business Operator's) View



Model sanity check

Purpose: Using counterfactual explanations to understand the model's behavior.

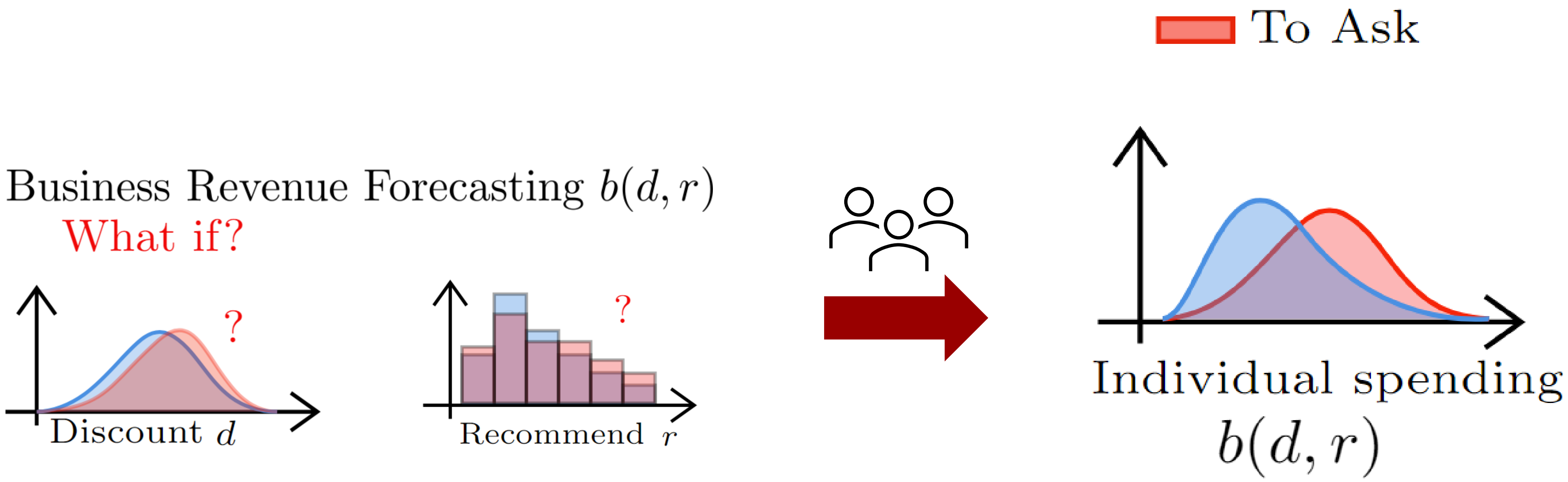
Example: Has the model learned correct business logics?

Business operation

Purpose: If we believe in the model, then use it to adjust the business operation strategy.

Example: How to launch a successful campaign?

A Stakeholder's (Business Operator's) View



The counterfactual distribution needs to resemble the originally observed.

We are finding distributions as counterfactuals.

Transportation Theory

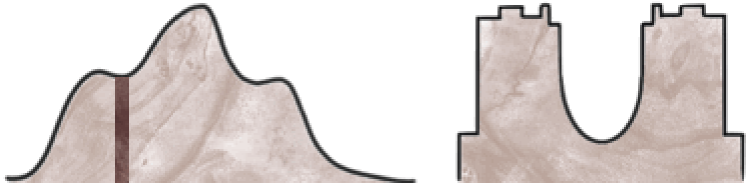
$$\mathcal{W}^2(\mathbf{x}, \mathbf{x}') \triangleq \min_{\boldsymbol{\pi} \geq \mathbf{0}} \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \|\mathbf{x}^{(i)} - \mathbf{x}'^{(j)}\|^2$$

$$\text{s.t.} \quad \sum_{j=1}^{n} \pi_{ij} = \frac{1}{n}$$

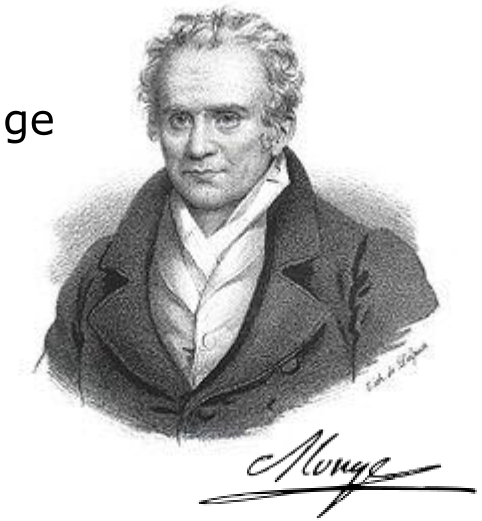
$$\sum_{i=1}^{n} \pi_{ij} = \frac{1}{m}$$

Also named ”**Wasserstein** Distance”

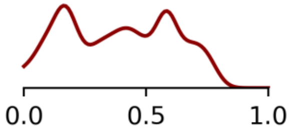
The problem was formalized by the French mathematician Gaspard Monge in 1781



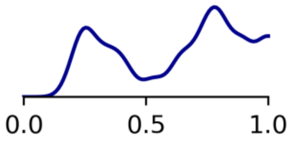
“Physical Movement”

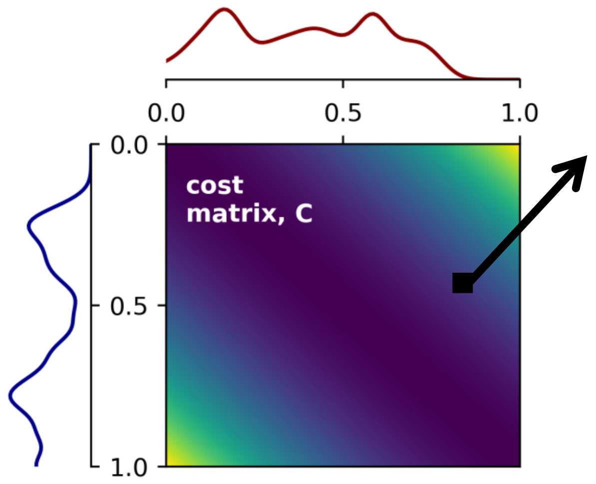


$$\mathbf{x} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\}$$



$$\mathbf{x}' = \{\mathbf{x}'^{(1)}, \mathbf{x}'^{(2)}, \ldots, \mathbf{x}'^{(m)}\}$$





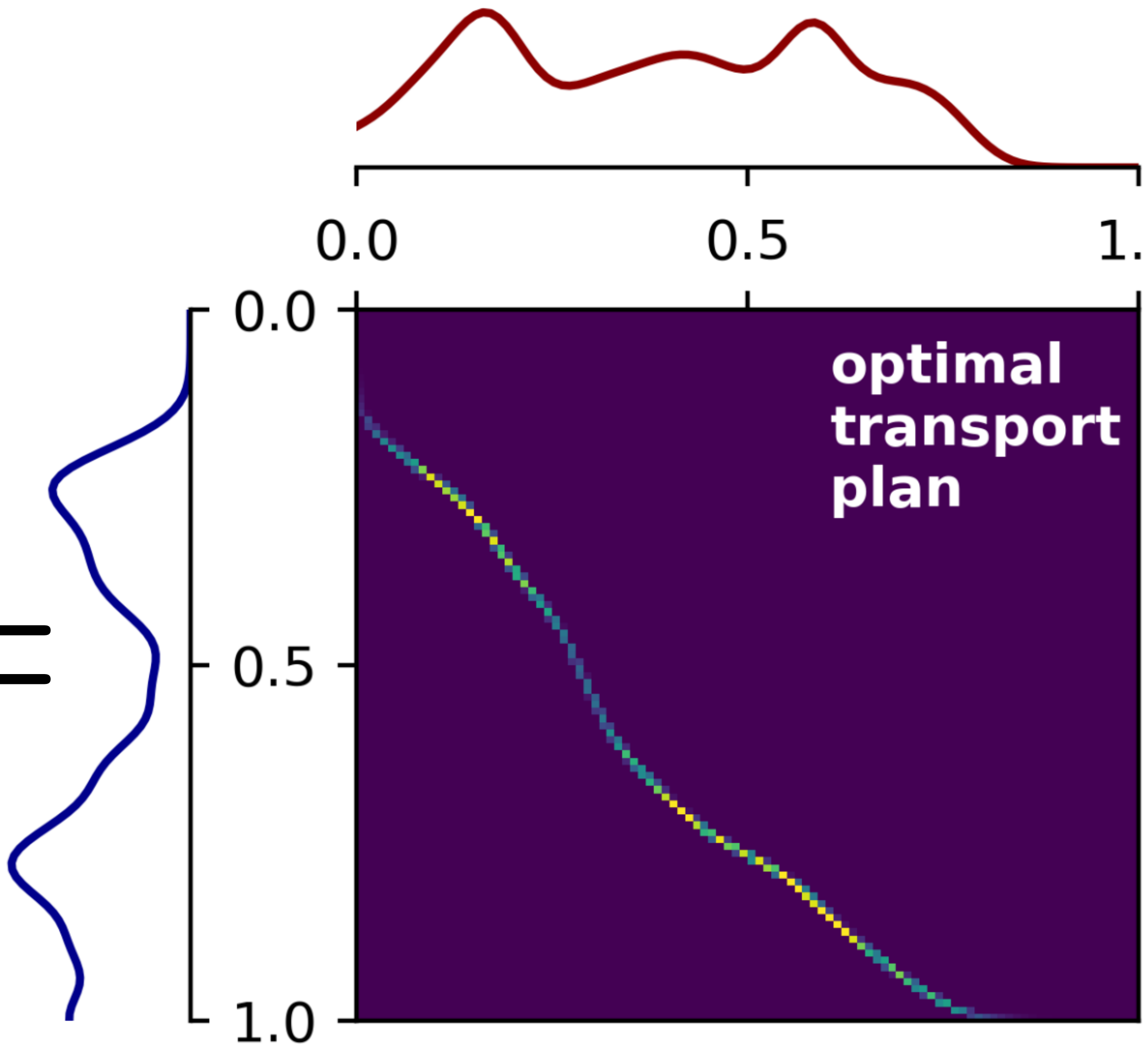
$$\|\mathbf{x}^{(i)} - \mathbf{x}'^{(j)}\|^2$$

Optimal Transportation: A Joint Probability

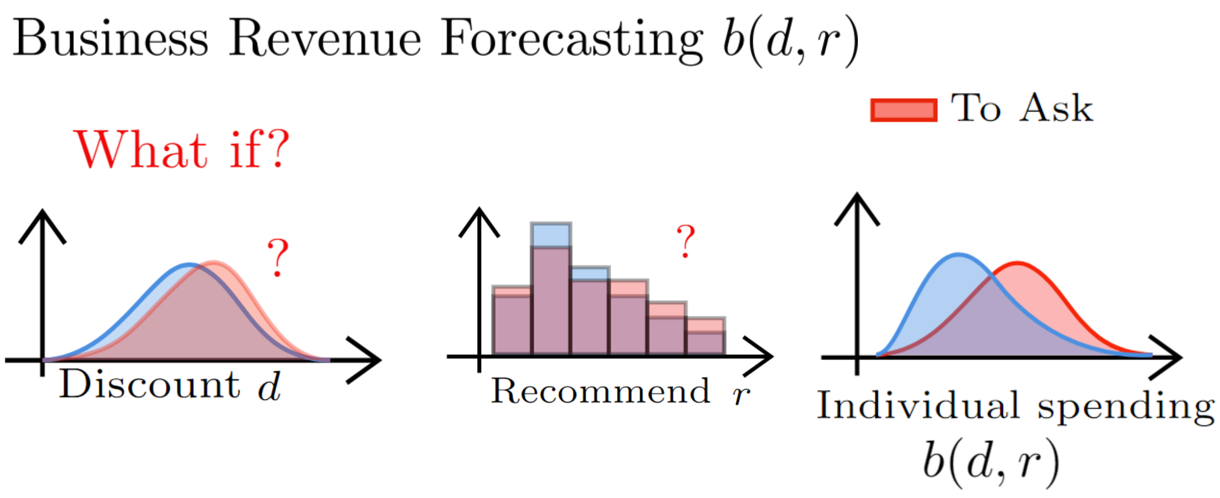
High Dimension x?

$$\mathcal{SW}^2 \triangleq \int_{\mathbb{S}^{d-1}} \mathcal{W}^2(\boldsymbol{\theta}^\top \mathbf{x}, \boldsymbol{\theta}^\top \mathbf{x}') \;\; \mathrm{d}\boldsymbol{\theta}$$

$$\pi =$$



Distributional Counterfactual Explanations (DCE)



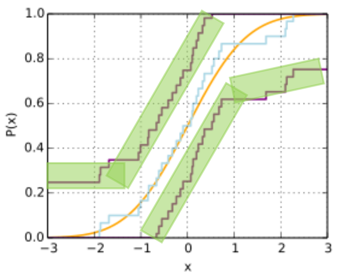
$$
\begin{aligned}
&\max_{\mathbf{x}, P} P \\
\text{s.t.}\quad & P \leq \mathbb{P}\left[\mathcal{SW}^2(\mathbf{x}, \mathbf{x}') < U_x\right] \\
& P \leq \mathbb{P}\left[\mathcal{W}^2(b(\mathbf{x}), y^*) < U_y\right] \\
& P \geq 1 - \frac{\alpha}{2}
\end{aligned}
$$

L. You, L. Cao, M. Nilsson, B. Zhao, and L. Lei, "Distributional Counterfactual Explanation With Optimal Transport", *International Conference on Artificial Intelligence and Statistics (AISTATS) 2025,* accepted **(Oral, top 2%)**.

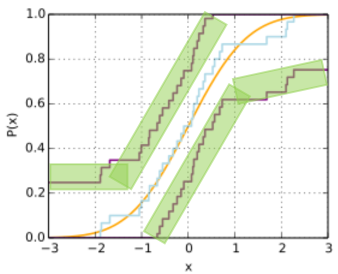
Distributional Counterfactual Explanations (DCE)

Dvoretzky–Kiefer–Wolfowitz–Massart inequality (**DKW inequality**) provides a bound on the worst-case distance

$\boldsymbol{\theta}^\top \mathbf{x}$



$\boldsymbol{\theta}^\top \mathbf{x}'$




```
def _dkw(x, u, alpha):
    """DKW lower and upper (1-alpha/2)-confidence bands for the
    u-quantiles of a distribution, based on a sample x."""

    n = len(x)
    gam = np.sqrt((1 / (2 * n)) * np.log(4 / alpha))  # 4 instead of 2.
    lower = _sample_quantile(x, u - gam)
    upper = _sample_quantile(x, u + gam)

    return lower, upper
```

$$\max_{\mathbf{x}, P} P$$

$$\text{s.t.} \quad P \leq \mathbb{P}\left[\mathcal{SW}^2(\mathbf{x}, \mathbf{x}') < U_x\right]$$

$$P \leq \mathbb{P}\left[\mathcal{W}^2(b(\mathbf{x}), y^*) < U_y\right]$$

$$P \geq 1 - \frac{\alpha}{2}$$

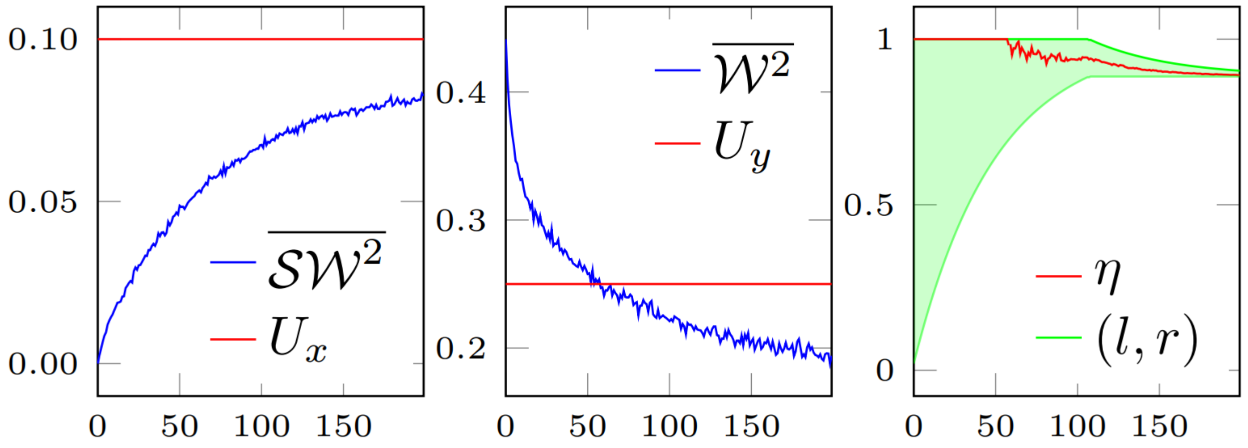
$$\mathbb{P}\left[\mathcal{W}^2(b(\mathbf{x}), y^*) \leq \underbrace{\frac{1}{1-2\delta}\int_\delta^{1-\delta} D(u)\,\mathrm{d}u}_{\leq U_y}\right] \geq 1 - \frac{\alpha}{2},$$

$$\mathbb{P}\left[\mathcal{SW}^2(\mathbf{x}, \mathbf{x}') \leq \underbrace{\frac{1}{1-2\delta}\int_{\mathbb{S}^{d-1}}\int_\delta^{1-\delta} D_{\boldsymbol{\theta},N}(u)\,\mathrm{d}u\,\mathrm{d}\sigma_N(\boldsymbol{\theta})}_{\leq U_x}\right] \geq 1 - \frac{\alpha}{2}$$

Manole, T., Balakrishnan, S., and Wasserman, L. (2022). Minimax confidence intervals for the sliced wasserstein distance. *Electronic Journal of Statistics*, 16(1):2252–2345.

Distributional Counterfactual Explanations (DCE)

$$Q(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\nu}, \eta) \triangleq (1-\eta) \cdot \underbrace{Q_x(\mathbf{x}, \boldsymbol{\mu})}_{\mathcal{SW}^2(\mathbf{x}, \mathbf{x}')} + \eta \cdot \underbrace{Q_y(\mathbf{x}, \boldsymbol{\nu})}_{\mathcal{W}^2(b(\mathbf{x}), y^*)}$$



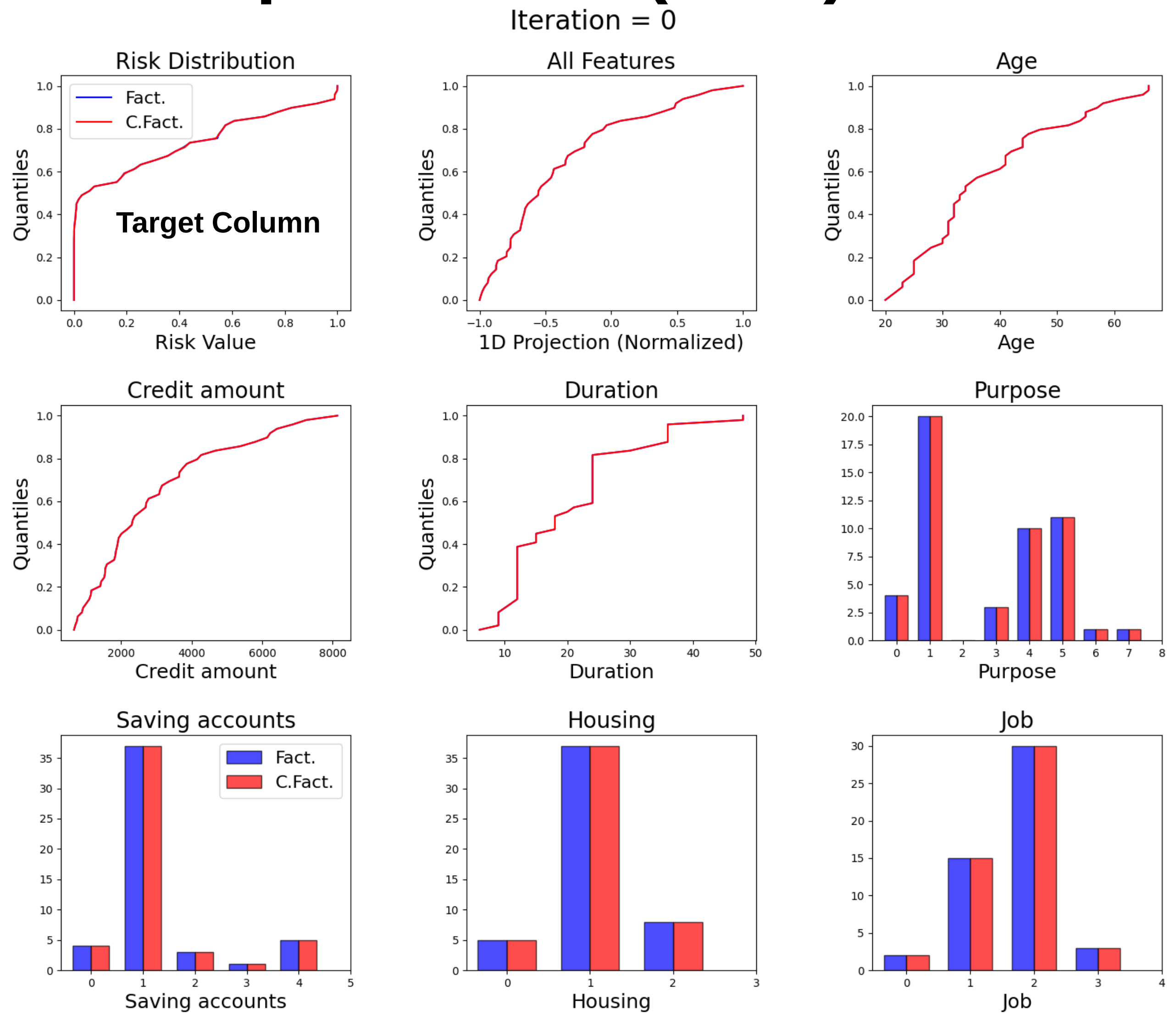
Algorithm 1 Distributional counterfactual

Require: $\mathbf{x}$, $\mathbf{y}^*$, model b, projections Θ, bounds U_x, U_y and significance level α.

Ensure: Counterfactual $\mathbf{x}$ or $\varnothing$.

```
x^0 ← x' + σ; t ← 0
repeat
    μ^t ← arg min_μ Q_x(x^t, μ)
    ν^t ← arg min_ν Q_y(x^t, ν)
    W̄² ← Eq. (10)
    S̄W̄² ← Eq. (11)
    η^t ← Algorithm 2 (or 3 in Appendix D)
    ∇̃Q ← ∇̃_x Q (x, μ^t, ν^t, η^t)
    x^{t+1} ← Retr(−τ∇̃Q)
    t ← t + 1
until ||x^{t+1} − x^t|| ≤ ε
if S̄W̄² ≤ U_x and W̄² ≤ U_y then
    return x^{t+1}
end if
return ∅
```

Distributional Counterfactual Explanations (DCE)



$$\max_{\mathbf{x},P} P$$

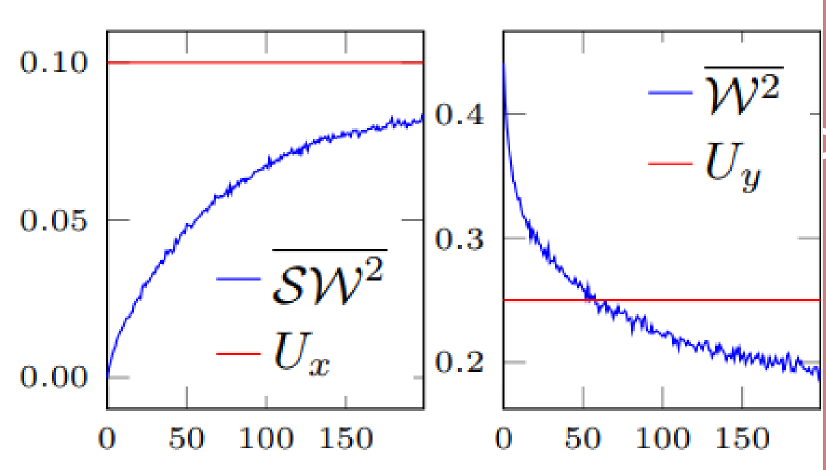
$$\text{s.t.} \quad P \le \mathbb{P}\left[\mathcal{SW}^2(\mathbf{x}, \mathbf{x}') < U_x\right]$$

$$P \le \mathbb{P}\left[\mathcal{W}^2(b(\mathbf{x}), y^*) < U_y\right]$$

$$P \ge 1 - \frac{\alpha}{2}$$

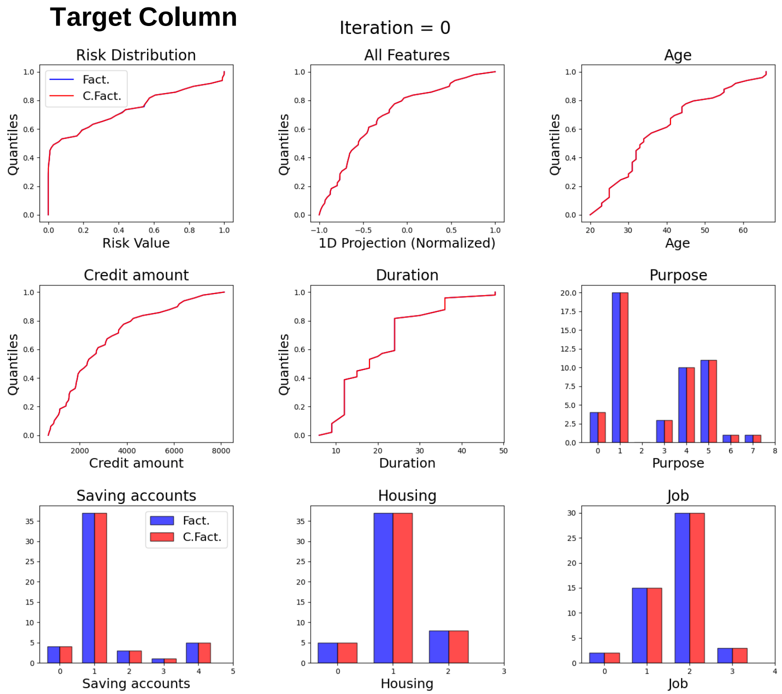
The first work on CE with distributional setup

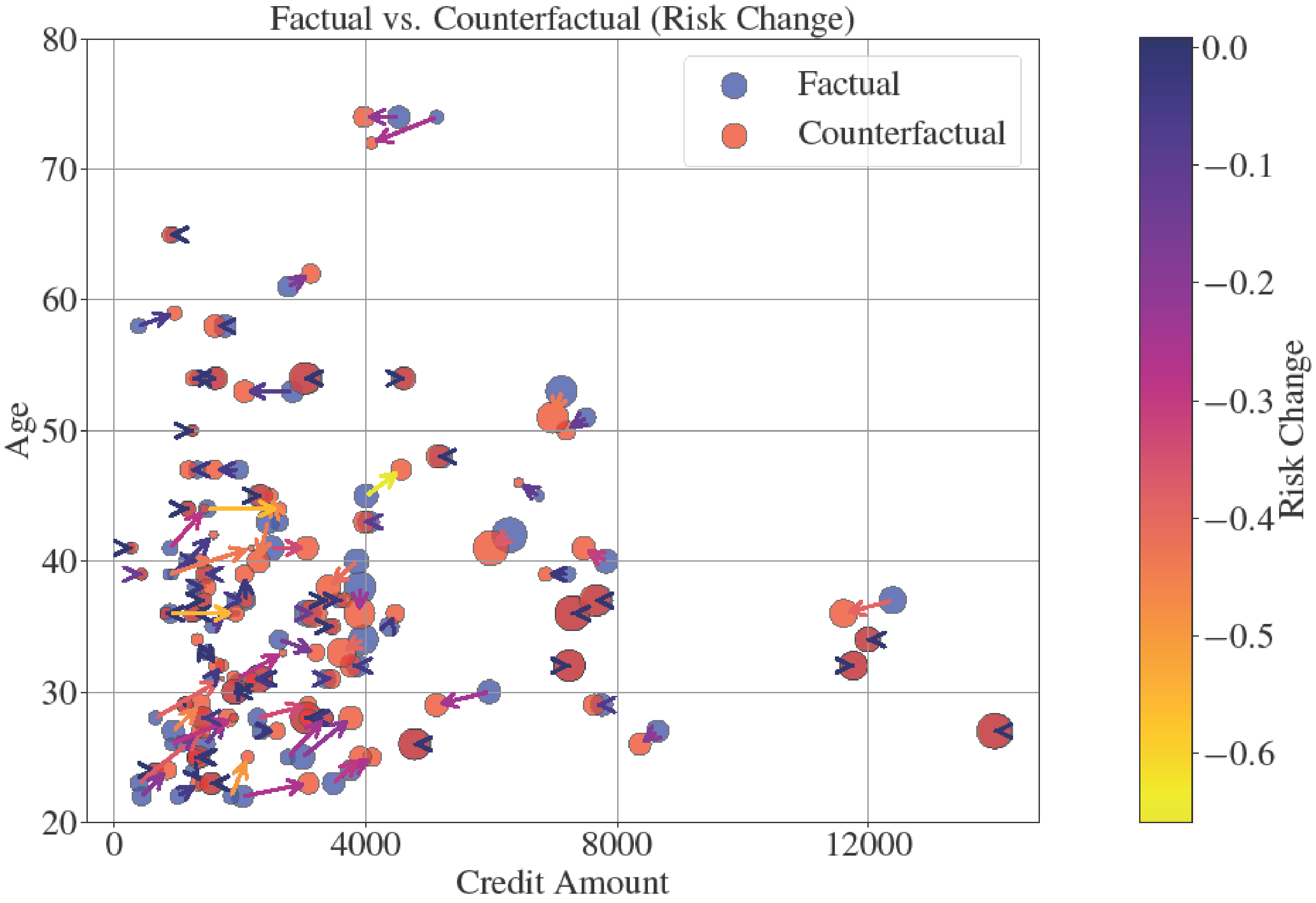
With rigorous statistcal guarantee for counterfactual validity and counterfactual proximity



d to …

Distributional Counterfactual Explanations (DCE)





Conciseness in CE

Example: User engagement on an e-commerce platform

Factual (Observation)

x

		®
200	5	No
150	3	No
100	2	No
150	6	No

Counterfactual 1

z′

		®
250	8	**Yes**
150	3	No
350	9	**Yes**
150	6	No

Counterfactual 2

z″

		®
200	5	No
150	7	**Yes**
100	2	No
350	6	**Yes**

Desired Outcome (Full Engagement)

$\mathbf{y}^*$

®
Yes
Yes
Yes
Yes

Scientific Problem

Given a (group of) factual instance(s), how can we devise an action plan that requires the least feature modifications to achieve a desired counterfactual outcome?

Factual

$\mathbf{x}$

		®
200	5	No
150	3	No
100	2	No
150	6	No

Counterfactual

$\mathbf{z}'$

		®
250	8	**Yes**
150	3	No
350	9	**Yes**
150	6	No

$\mathbf{z}''$

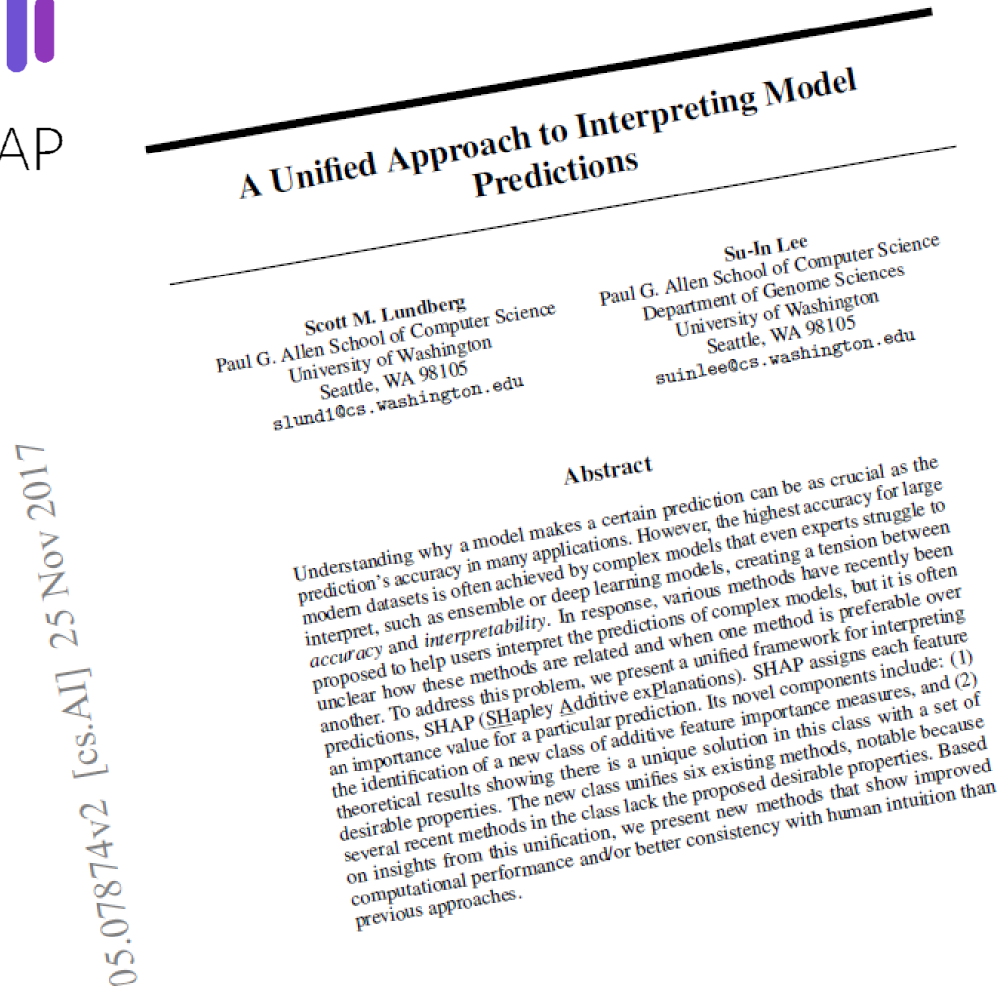
		®
200	5	No
150	7	**Yes**
100	2	No
350	6	**Yes**

$\mathbf{y}^*$

®
Yes
Yes
Yes
Yes

Feature Attribution With Shapley Values

SHAP



arXiv:1705.07874v2 [cs.AI] 25 Nov 2017

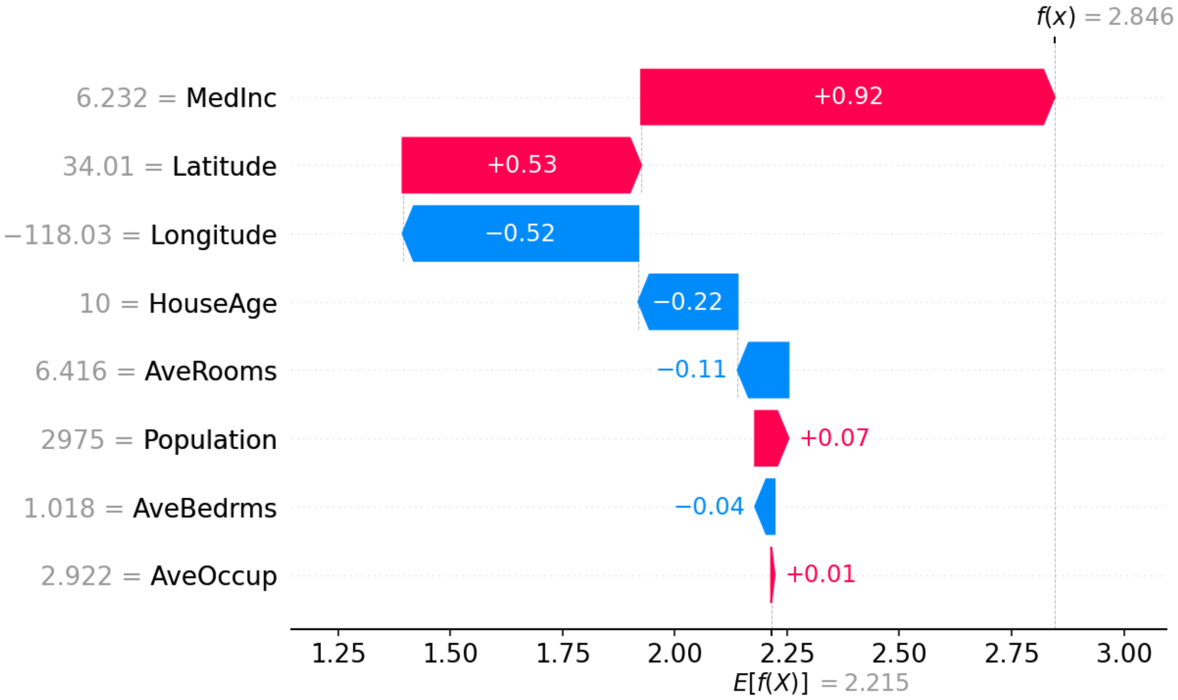
A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

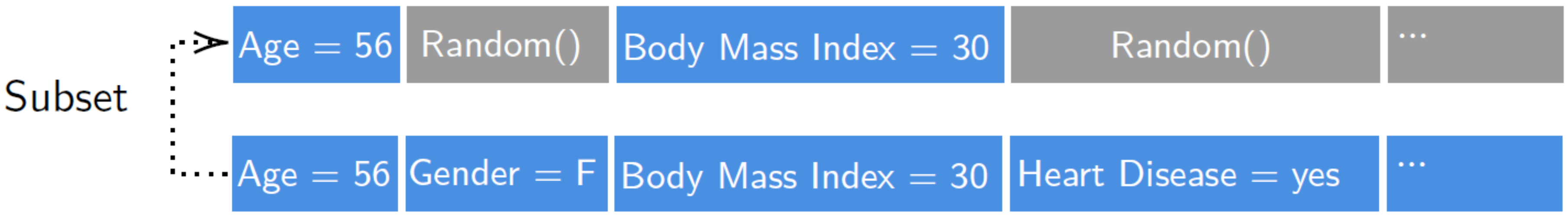
Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.



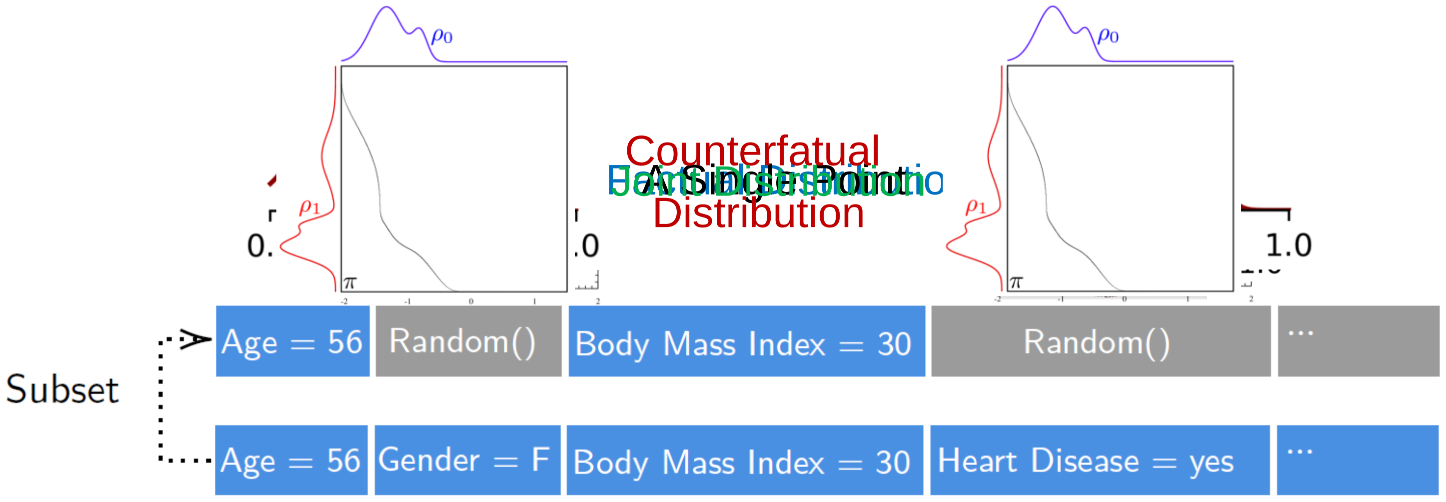
Feature Attribution With Shapley Values

$$\overbrace{\phi_i(v)}^{i\text{'s Shapley value}} = \sum_{S \subseteq D \setminus \{i\}} \underbrace{\frac{|S|!(|D|-|S|-1)!}{|D|!}}_{S\text{'s weight}} \Big(\overbrace{v(S\cup\{i\}) - v(S)}^{i\text{'s marginal contribution}} \Big)$$



Feature Attribution With Shapley Values

Missing values are simulated by a “background distribution”



Lei You, Yijun Bian, and Lele Cao

"Refining Counterfactual Explanations With Joint-Distribution-Informed Shapley Towards Actionable Minimality", *ICLR 2025* *8 (accept), 8 (accept), 6 (weak accept), 6 (weak accept) --- top 5%*

and rejected ☹

Assumed Known Conditional Distribution (General for many CE algorithms)

No Assumption on Knowledge of Distribution

Problem Formulation

$$\mathbf{y}^* = f(\mathbf{r}) \rightarrow f(\mathbf{z}) \approx \mathbf{y}^*$$

Original Counterfactuals
(Obtained from an arbitrary CE algorithm)

Refined Counterfactuals
(Supposed to be with less changes)

$$\min_{\mathbf{c},\mathbf{z}} \quad D\left(f(\mathbf{z}), \mathbf{y}^*\right)$$

$$\text{s.t.} \quad D\left(\mathbf{z}, \mathbf{x}\right) \leq \epsilon$$

$$\sum_{i=1}^{n} \sum_{k=1}^{d} c_{ik} \leq C$$

$$z_{ik} \leq M c_{ik}; \quad i = 1, \ldots n, \; k = 1, \ldots d$$

$$z_{ik} \geq -M c_{ik}; \quad i = 1, \ldots n, \; k = 1, \ldots d$$

$$c_{ik} = 1$$

COunterfactual with Limited Actions (COLA)

$f(\mathbf{z}) \approx \mathbf{y}^*$ ⟸ $\mathbf{y}^* = f(\mathbf{r})$

Step 1: Pick any 1 of the ≥100 existing CE algorithms

Obtain $\mathbf{r}$ and compute $\mathbf{p}$

$\mathbf{r} \leftarrow A_{\text{CE}}$

$$\mathbf{x} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \quad \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \\ r_{31} & r_{32} \end{bmatrix}$$

$\mathbf{p}(\mathbf{x}, \mathbf{r}) \leftarrow$ Optimal Transport

$$A_{\text{CE}}$$

Data Mining and Knowledge Discovery (2024) 38:2770–2824
https://doi.org/10.1007/s10618-022-00831-6

Check for updates

Counterfactual explanations and how to find them: literature review and benchmarking

Riccardo Guidotti[1]

Hundreds of algorithms are surveyed!

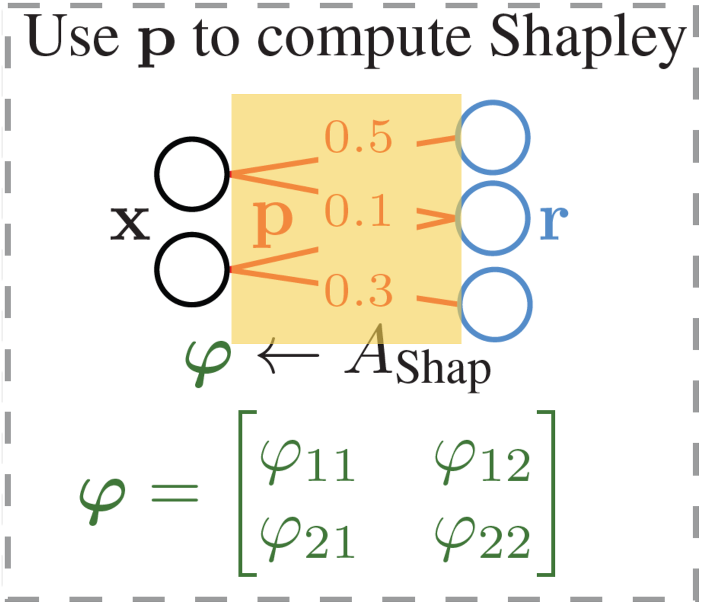
Received: 1 April 2021 / Accepted: 18 March 2022 / Published online: 28 April 2022

Abstract
Interpretable machine learning aims at unveiling the reasons behind predictions returned by uninterpretable classifiers. One of the most valuable types of explanation consists of counterfactuals. A counterfactual explanation reveals what should have been different in an instance to observe a diverse outcome. For instance, a bank customer asks for a loan that is rejected. The counterfactual explanation consists of what should have been different for the customer in order to have the loan accepted. Recently, there has been an explosion of proposals for counterfactual explainers. The aim of this work is to survey the most recent explainers returning counterfactual explanations. We categorize explainers based on the approach adopted to return the counterfactuals, and we label them according to characteristics of the method and properties of the counterfactuals returned. In addition, we visually compare the explanations, and we report quantitative benchmarking assessing minimality, actionability, stability, diversity, discriminative power, and running time. The results make evident that the current state of the art does not provide a counterfactual explainer able to guarantee all these properties simultaneously.

COunterfactual with Limited Actions (COLA)

$f(\mathbf{z}) \approx \mathbf{y}^*$ ← $\mathbf{y}^* = f(\mathbf{r})$

Step 2: P-SHAP



$$A_{\text{Shap}}$$

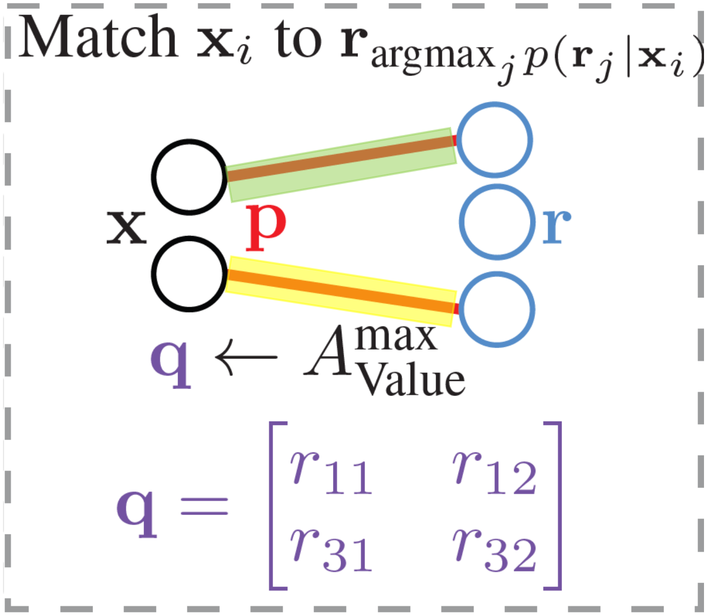
φ_{ik} tells the importance of x_{ik}

$$\begin{bmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{21} & \varphi_{22} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

COunterfactual with **L**imited **A**ctions (**COLA**)

$f(\mathbf{z}) \approx \mathbf{y}^*$ ⟵ $\mathbf{y}^* = f(\mathbf{r})$

Step 3: Computing the candidate values for revising $\mathbf{x}$ later



$$A_{\text{Value}}$$

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \\ r_{31} & r_{32} \end{bmatrix}$$

$$\mathbf{q} = \begin{bmatrix} r_{11} & r_{12} \\ r_{31} & r_{32} \end{bmatrix}$$

COunterfactual with Limited Actions (COLA)

$$f(\mathbf{z}) \approx \mathbf{y}^* \quad \Leftarrow \quad \mathbf{y}^* = f(\mathbf{r})$$

$$\|\mathbf{z} - \mathbf{x}\|_{\mathrm{F}} \leq \|\mathbf{r} - \mathbf{x}\|_{\mathrm{F}}$$

Step 4: Computing the refined counterfactaul $\mathbf{Z}$

Get $\mathbf{c}$ and $\mathbf{z}$ by $\boldsymbol{\varphi}$ and $\mathbf{q}$

$$\begin{bmatrix} \varphi_{11} = 0.2 & \varphi_{12} = 0.4 \\ \varphi_{21} = 0.3 & \varphi_{22} = 0.1 \end{bmatrix}$$

$$C = 2$$

$$\rightarrow \mathbf{c} \sim \{(1,2),(2,1)\}$$

$$\rightarrow \mathbf{z} = \begin{bmatrix} x_{11} & q_{12} \\ q_{21} & x_{22} \end{bmatrix}$$

$$\Rightarrow \quad \mathbf{q} = \begin{bmatrix} r_{11} & r_{12} \\ r_{31} & r_{32} \end{bmatrix}$$

$$\Downarrow$$

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \quad \Rightarrow \quad \mathbf{z} = \begin{bmatrix} x_{11} & q_{12} \\ q_{21} & x_{22} \end{bmatrix}$$

Results Demonstration: Individual Change

Dataset	ML Model	CE Algorithm
German Credits	LightGBM	DiCE

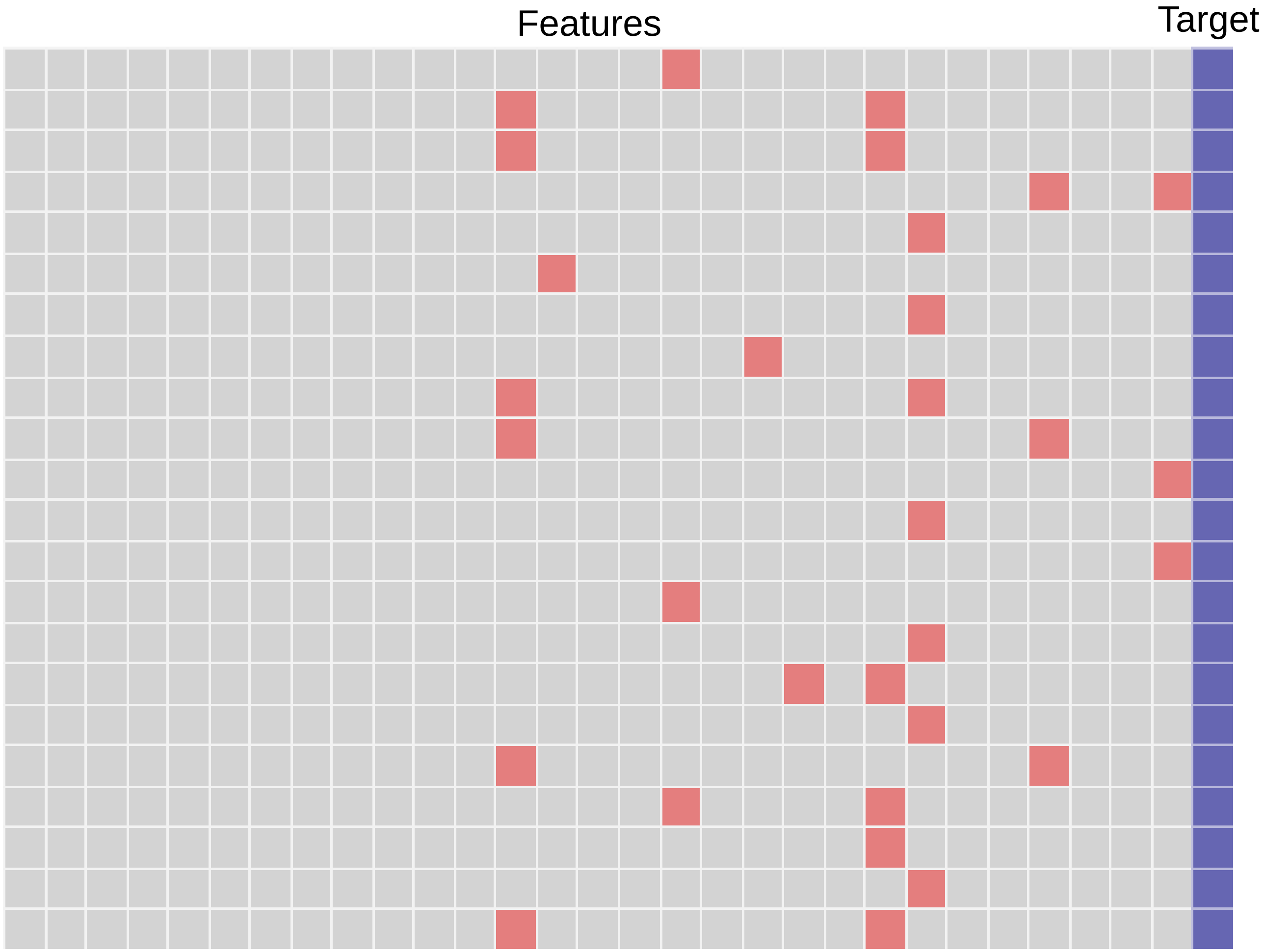
	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	24	1	2	1	2 -> 0	2	5595	72	5	1 -> 0
1	33	1	1	2	1	2	2384 -> 6536	36	6	1 -> 0
2	31	1	2	2	1	1	3161	24 -> 7	0	1 -> 0
3	23	1	0	1	0	2	14555	6 -> 71	1	1 -> 0
4	28	0	2	1	2 -> 4	2	2278	18	1	1 -> 0
5	45	1	1	1	1	1 -> 0	4006	28	1	1 -> 0
6	39	1	2	0	0	3	1271 -> 3096	15	5	1 -> 0
7	42	1	2	1	1	1 -> 3	4153	18	4	1 -> 0
8	24	0	2	1	1	2	2150	30 -> 6	1	1 -> 0
9	31	1	2	1	1	2	1935 -> 6380	24	0	1 -> 0
10	48	0	1	0	2	3	1240 -> 5706	10	1	1 -> 0
11	29	1	2	1	1	1 -> 3	6887	36	3	1 -> 0
12	37	1	1	1	0	3 -> 0	1344	24	1	1 -> 0
13	25	0	2	1	1	0	7855 -> 1340	36	1	1 -> 0
14	47	1	2	0	2	2	12612 -> 6392	36	3	1 -> 0
15	30	0	3	1	1	2	5096	48 -> 15	4	1 -> 0
16	23	0	2	2	1 -> 4	1	1442	24	1	1 -> 0
17	42	1	2	1	1	1	3446 -> 7770	36	4	1 -> 0
18	39	1	3	1	1	2 -> 0	11938	24	7	1 -> 0
19	27	1	3	0	1	1	1422 -> 3825	9	1	1 -> 0

Refined Counterfactual $\mathbf{z}$ (20 Actions)

43% Less Actions Taken

Results Demonstration: Group Change

Dataset	ML Model	CE Algorithm
Hotel Bookings	XGBoost	DiCE



Refined Counterfactual $\mathbf{z}$ (31 Actions)

72% Less Actions Taken

Overall Performance

A_{CE}	DiCE (Mothilal et al., 2020), AReS (Rawal & Lakkaraju, 2020), GlobeCE (Ley et al., 2023), KNN (Albini et al., 2022; Contardo et al.; Forel et al., 2023), Discount (You et al., 2024)
Model f	Bagging, LightGBM, Support Vector Machine (SVM), Gaussian Process (GP), Radial Basis Function Network (RBF), XGBoost, Deep Neural Network (DNN), Random Forest (RndForest), AdaBoost, Gradient Boosting (GradBoost), Logistic Regression (LR), Quadratic Discriminant Analysis (QDA)

Dataset	% Action of The Original	
	80% Counterfactual Effect	100% Counterfactual Effect
German Credits (Features = 9)	24.3%	44.9%
Hotel Bookings (Features=29)	14.6%	26.0%
COMPAS (Features=15)	14.8%	30.0%
HELOC (Features=23)	13.4%	44.7%

No assumptions on CE or ML models

E.g. no assumptions on:

- ML Model architecture like tree-based etc.
- ML Model's differentiability
- CE algorithms

Physical Meaning of P-SHAP

$$\mathcal{W}_1(f(\mathbf{x}), \mathbf{y}^*) \leq L \sqrt{\sum_{i=1}^{j=m} p_{ij}^{\mathrm{OT}} \|\mathbf{x}_i - \mathbf{r}_j\|_2^2}$$

Guaranteed proximity

$$\|\mathbf{z} - \mathbf{x}\|_F \leq \|\mathbf{r} - \mathbf{x}\|_F$$

Low computational complexity

```python
from xai_cola import data_interface
from xai_cola import ml_model_interface
from counterfactual_explainer import DiCE
from xai_cola.counterfactual_limited_actions import COLA

# Initialize the COLA
refiner = COLA(
        data=data,
        ml_model=ml_model,
        x_factual=factual,
        x_counterfactual=counterfactual,
        )

# Choose the policy
refiner.set_policy(
        matcher="ect", # We prefer "ect_matcher" with DiCE, you can also
        attributor="pshap",
        Avalues_method="max"
        )

# Choose the number of actions
factual, ce, ace = refiner.get_refined_counterfactual(limited_actions=10)
```

arXiv 2410.05419 License MIT

GitHub

Counterfactual explanations with Limited Actions (COLA)

factual

	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	27	1	2	1	1	1	3552	24	4	1
1	31	1	2	2	1	1	3161	24	0	1
2	34	0	3	1	1	2	2064	24	4	1
3	20	0	2	2	1	1	2039	18	4	1
4	29	1	3	1	1	2	11328	24	7	1
5	22	0	2	1	2	1	741	12	2	1
6	24	0	2	2	1	1	1207	24	1	1
7	53	1	2	0	1	1	7119	48	4	1

factaul -> corresponding counterfactual

	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	27	1	2	1	1	1	3552 -> 1886	24	4	1 -> 0
1	31	1	2	2	1	1	3161	24 -> 20	0	1 -> 0
2	34	0	3	1	1 -> 2	2	2064 -> 3077	24	4	1 -> 0
3	20	0	2	2	1	1	2039 -> 9594	18	4	1 -> 0
4	29	1	3 -> 2	1	1	2	11328 -> 4852	24	7	1 -> 0
5	22	0	2	1	2	1 -> 2	741 -> 10076	12	2	1 -> 0
6	24	0	2	2	1	1	1207 -> 4342	24 -> 19	1	1 -> 0
7	53	1	2 -> 3	0	1	1	7119	48 -> 32	4	1 -> 0

factual -> action-limited counterfactual

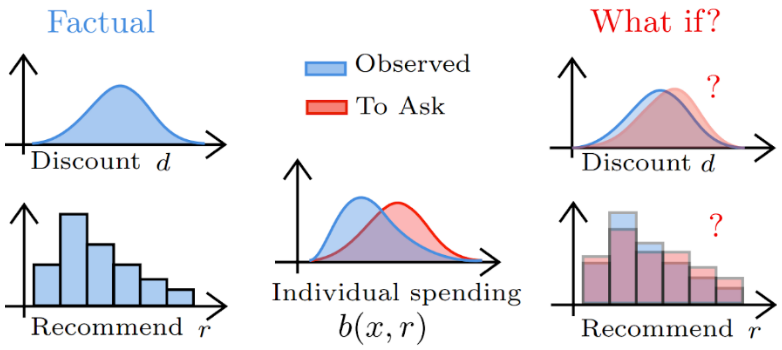
	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	27	1	2	1	1	1	3552 -> 1886	24	4	1 -> 0
1	31	1	2	2	1	1	3161	24 -> 20	0	1 -> 0
2	34	0	3	1	1	2	2064 -> 3077	24	4	1 -> 0
3	20	0	2	2	1	1	2039 -> 9594	18	4	1 -> 0
4	29	1	3	1	1	2	11328 -> 4852	24	7	1 -> 0
5	22	0	2	1	2	1	741 -> 10076	12	2	1 -> 0
6	24	0	2	2	1	1	1207 -> 4342	24 -> 19	1	1 -> 0
7	53	1	2 -> 3	0	1	1	7119	48 -> 32	4	1 -> 0

Lei You, PhD
Assistant Professor in Applied Mathematics
Technical University of Denmark (DTU)

Summary

In this talk, we explore advanced techniques in Explainable AI (XAI) by integrating concepts from **optimal transport theory**, a mathematical framework for comparing and aligning distributions. Two themes are covered:

Distribution Pattern as Explanations



Traditional counterfactual explanations focus on changing individual inputs to see how they affect outcomes, but they often miss the bigger picture of how groups of data points relate to one another. We extend traditional counterfactual explanations by introducing Distributional Counterfactual Explanation (DCE), which shifts from focusing solely on individual input changes to considering broader patterns within the entire data distribution. As a result, our approach provides stakeholders with valid counterfactual distributions supported by statistical confidence.

Explanations With Actionable Minimality

Given a (group of) factual instance(s), how can we devise an action plan that requires the least feature modifications to achieve a desired counterfactual outcome?

$\mathbf{x}$

200	5	No
150	3	No
100	2	No
150	6	No

$\mathbf{z}'$

250	8	**Yes**
150	3	No
350	9	**Yes**
150	6	No

$\mathbf{z}''$

200	5	No
150	7	**Yes**
100	2	No
350	6	**Yes**

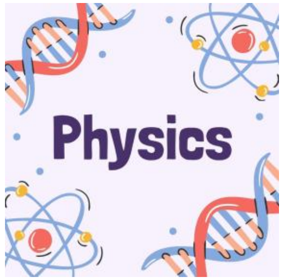
$\mathbf{y}^*$

Yes
Yes
Yes
Yes

We refine counterfactual explanations to enhance actionable efficiency by minimizing unnecessary feature changes, ensuring the proposed interventions are both valid and practical. Using optimal transport, we derive a joint distribution between observed and counterfactual data, which informs Shapley values for more precise feature attributions. This approach ensures minimal, realistic changes that make explanations more feasible and impactful for stakeholders.

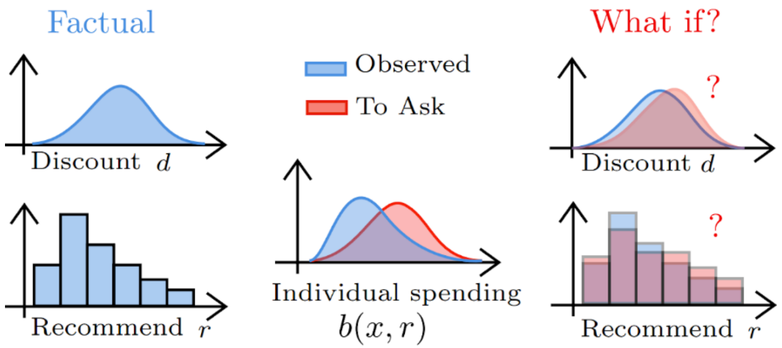
Lei You, PhD
Assistant Professor in Applied Mathematics
Technical University of Denmark (DTU)

Summary



Optimal transport has a physical interpretation of generating data (i.e. explanations)

Distribution Pattern as Explanations



Explanations With Actionable Minimality

Given a (group of) factual instance(s), how can we devise an action plan that requires the least feature modifications to achieve a desired counterfactual outcome?

x

💵	Click	®
200	5	No
150	3	No
100	2	No
150	6	No

z'

💵	Click	®
250	8	**Yes**
150	3	No
350	9	**Yes**
150	6	No

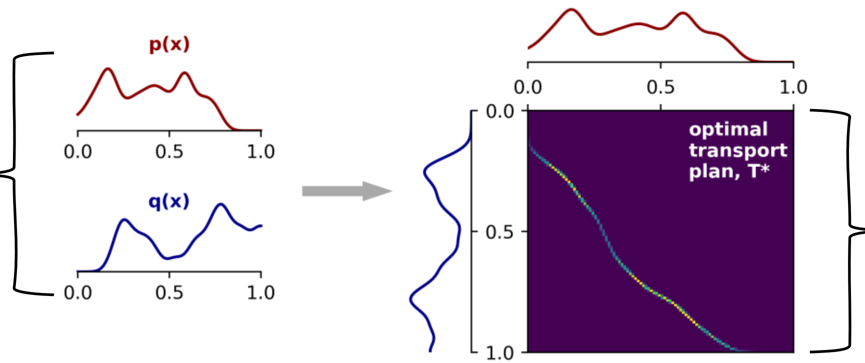
z''

💵	Click	®
200	5	No
150	7	**Yes**
100	2	No
350	6	**Yes**

y^*

®
Yes
Yes
Yes
Yes

A Metric for Dissimilarity



A Plan for Data Transformation