Counterfactual Explanations in Explainable AI: A Tutorial

Distributed Data Lab,

XAI Team

HUAWEI

KDD 2021 Tutorial: https://sites.google.com/view/kdd-2021-counterfactual

Cong WANG

Xiao-Hui LI

Han GAO

Shendi WANG

Luning WANG

Caleb Chen CAO

Lei CHEN

CONTENTS

01

Introduction

Lei CHEN

Artificial Intelligence(AI) in our daily life

Shenzhen Power Supply Bureau

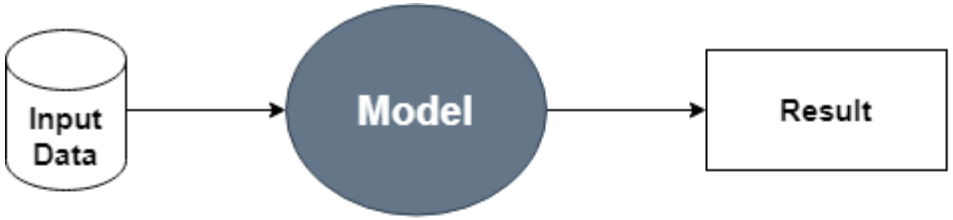
Smart City - Shen Zhen

https://www.huawei.com/minisite/building-an-intelligent-world-together/cn/

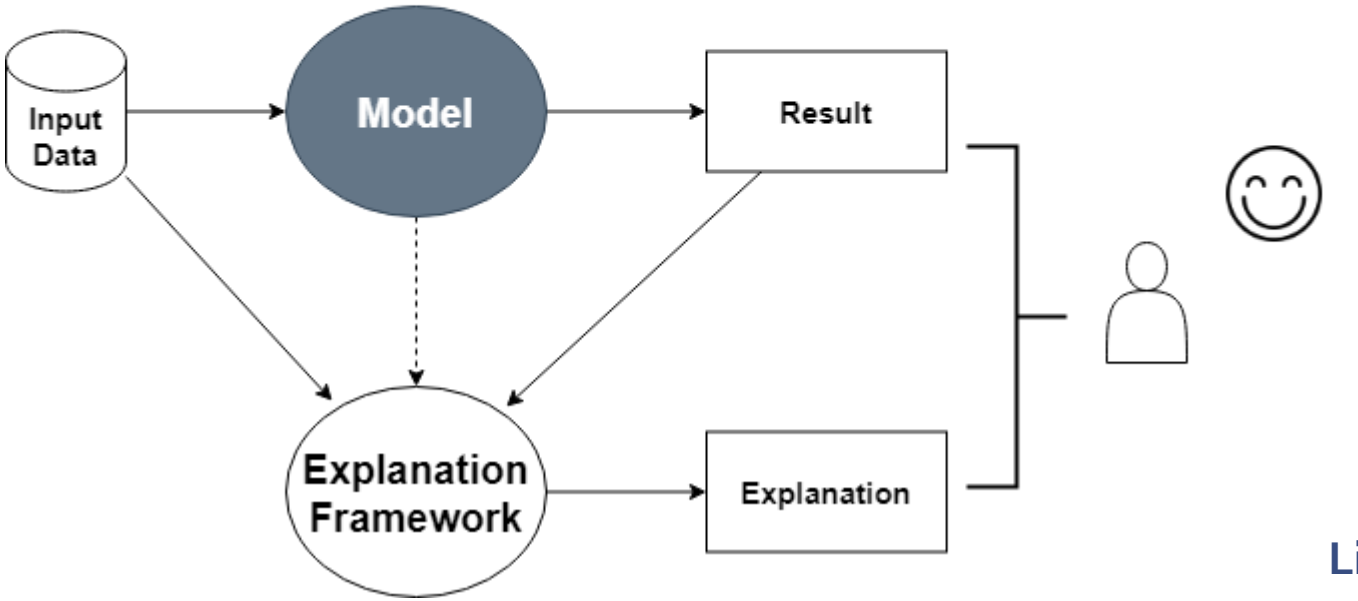
Artificial Intelligence(AI) in our daily life

	Financial Services	Government/ Public sector	Healthcare & Life sciences	Manufacturing	Retail
Predictive Analytics	√	√		√	√
Real-time Operations Management	√	√		√	√
Risk Management and Analytics		√	√		
Customer Services					√
R&D				√	
Fraud detection	√				
Social engagement			√		
Knowledge creation			√		

We need explanation for black-box model…



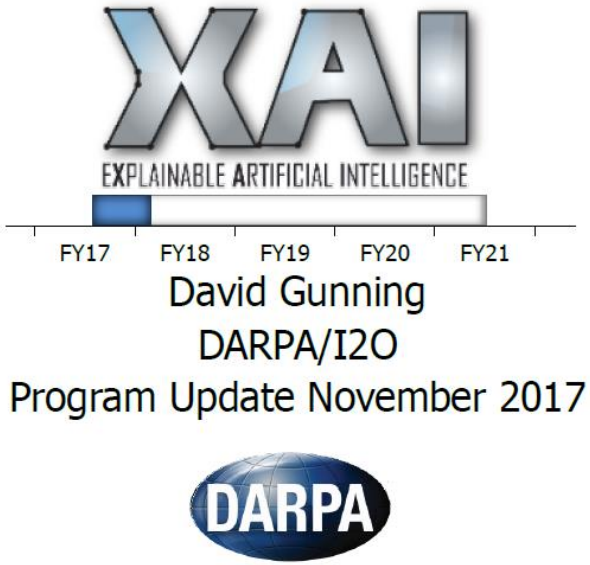
We get result, but we don't know why.



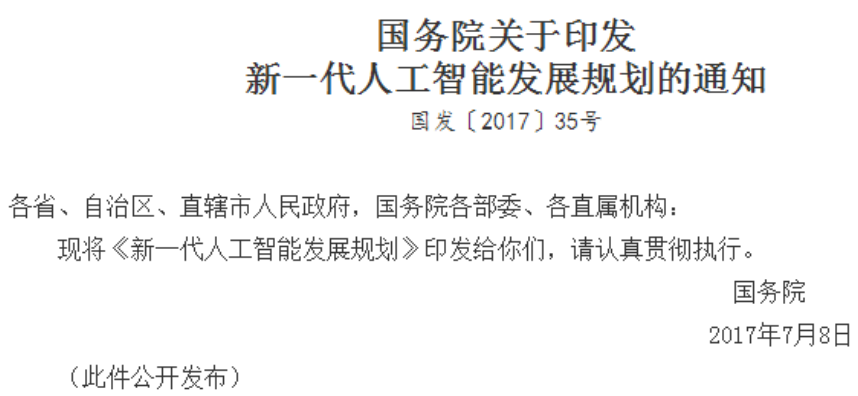
After get explanation, we could more understand the result.

Life is hard, we need XAI.

XAI Booming



[US] Defense Advanced Research Projects Agency(DARPA) initiates "XAI" project in 2017



国务院关于印发
新一代人工智能发展规划的通知

国发〔2017〕35号

各省、自治区、直辖市人民政府，国务院各部委、各直属机构：

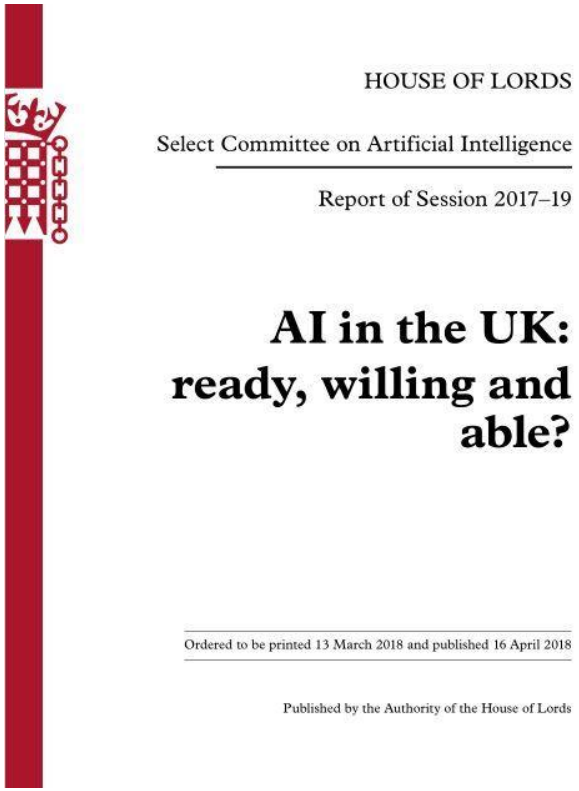
现将《新一代人工智能发展规划》印发给你们，请认真贯彻执行。

国务院

2017年7月8日

（此件公开发布）

[CHN] XAI included in "New Generation Artificial Intelligence Development Plan" issued by state council



HOUSE OF LORDS

Select Committee on Artificial Intelligence

Report of Session 2017–19

AI in the UK: ready, willing and able?

Ordered to be printed 13 March 2018 and published 16 April 2018

Published by the Authority of the House of Lords

[UK] XAI included in AI white paper issued by house of UK

XAI Progress

2017

IJCAI
IJCAI 2017 WORKSHOP ON EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

NIPS
INTERPRETABLE ML SYMPOSIUM

2018

IEEE VIS
1ST WORKSHOP ON VISUALIZATION FOR AI EXPLAINABILITY

IJCAI/ECAI
IJCAI/ECAI 2018 WORKSHOP ON EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

ICML
2018 WORKSHOP ON HUMAN INTERPRETABILITY IN MACHINE LEARNING (WHI)

NIPS & GOOGLE BRAIN
VISUALIZATION FOR MACHINE LEARNING

2019

IEEE VIS
2ND WORKSHOP ON VISUALIZATION FOR AI EXPLAINABILITY

IJCAI
IJCAI 2019 WORKSHOP ON EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

ICAPS
2ND ICAPS WORKSHOP ON EXPLAINABLE PLANNING (XAIP-2019)

HCI
HUMAN-CENTERED MACHINE LEARNING PERSPECTIVES WORKSHOP

AAAI
AAAI 2019 TUTORIAL ON EXPLAINABLE AI

CVPR
CVPR-19 WORKSHOP ON EXPLAINABLE AI

ICCC
2019 ICCV WORKSHOP ON INTERPRETING AND EXPLAINING VISUAL ARTIFICIAL INTELLIGENCE MODELS

KDD
KDD 2019 TUTORIAL EXPLAINABLE AI IN INDUSTRY

2020

CVPR
CVPR 2-2- TUTORIAL TUTORIAL ON INTERPRETABLE MACHINE LEARNING

FAT
FAT 2020 TUTORIAL EXPLAINABLE AI IN INDUSTRY:PRACTICAL CHALLENGES AND LESSONS LEARNED

KDD
KDD 2020 TUTORIAL INTELLIGIBLE AND EXPLAINABLE MACHINE LEARNING-BEST PRACTICES

XAI Importance

- XAI **DECIDES** AI's acceptance in many major industries - **Healthcare**

Healthcare

- Explainability leads to responsibility
  - Hospital, insurance and patients know how to share risk
- Interpretability leads to trust
  - Doctors know when and how to rely on AI

XAI Importance

- XAI **DECIDES** AI's acceptance in many major industries – **Legal Industry**

Opinion

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

Legal Industry (*LegalTech*)

- Legal practitioners needs explanation
  - In forms of *rules*, *decision trees*, *similar cases*, *charts*, *etc.*
- AI gains trust from the public through explainable models and results

XAI Importance

- XAI **DECIDES** AI's acceptance in many major industries – **Financial Industry**


The Big Read Artificial intelligence + Add to myFT

Insurance: Robots learn the business of covering risk

Artificial intelligence could revolutionise the industry but may also allow clients to calculate if they need protection

Save

Oliver Ralph MAY 16, 2017 24

https://www.ft.com/content/e07cee0c-3949-11e7-821a-6027b8a20f23

community.fico.com/s/explainable-machine-learning-challenge

Legal Industry (*FinTech*)

- Explainability leads to trust of AI investment
- Insurance
  - Explainable AI quoting & underwriting
- Loan
  - Explainable credit accessment

XAI Importance

Europe - General Data Protection Regulation (GDPR)

Art. 14(2)(g) GDPR: the controller shall provide the data subject with the following information "the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, **meaningful information about the logic involved**, as well as the significance and the envisaged consequences of such processing for the data subject."

Art. 22 GDPR Automated individual decision-making, including profiling

(1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

https://gdpr-info.eu/art-22-gdpr/

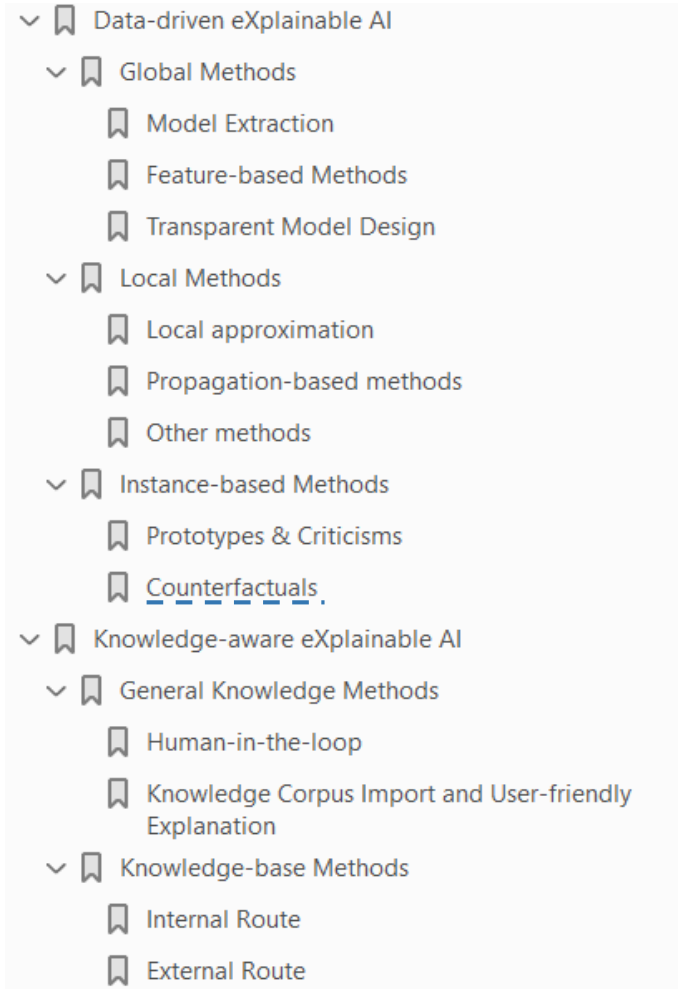
XAI Importance

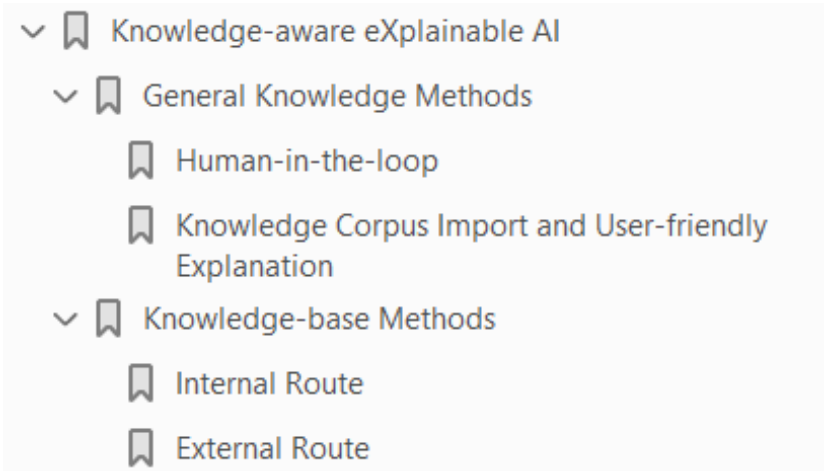
China: Next-Generation Artificial Intelligence Development Plan

3.1 Advanced machine learning theory focuses on making breakthroughs in adaptive learning and autonomous learning to achieve AI with high explanability and strong generalization capabilities.

http://www.gov.cn/zhengce/content/

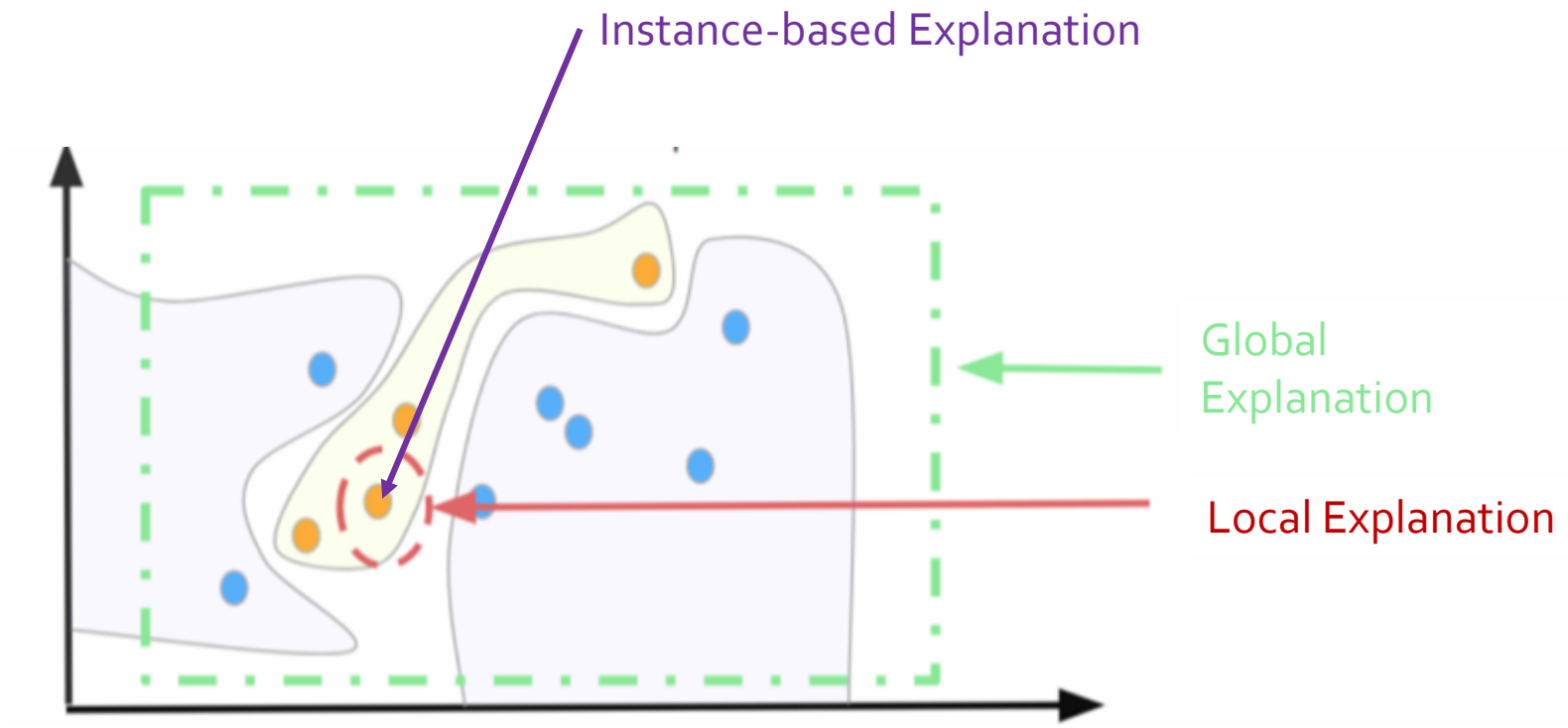
Overview for XAI Methods





Li, X. H., Cao, C. C., Shi, Y., Bai, W., Gao, H., Qiu, L., ... & Chen, L. (2020). A survey of data-driven and knowledge-aware explainable ai. IEEE Transactions on Knowledge and Data Engineering.

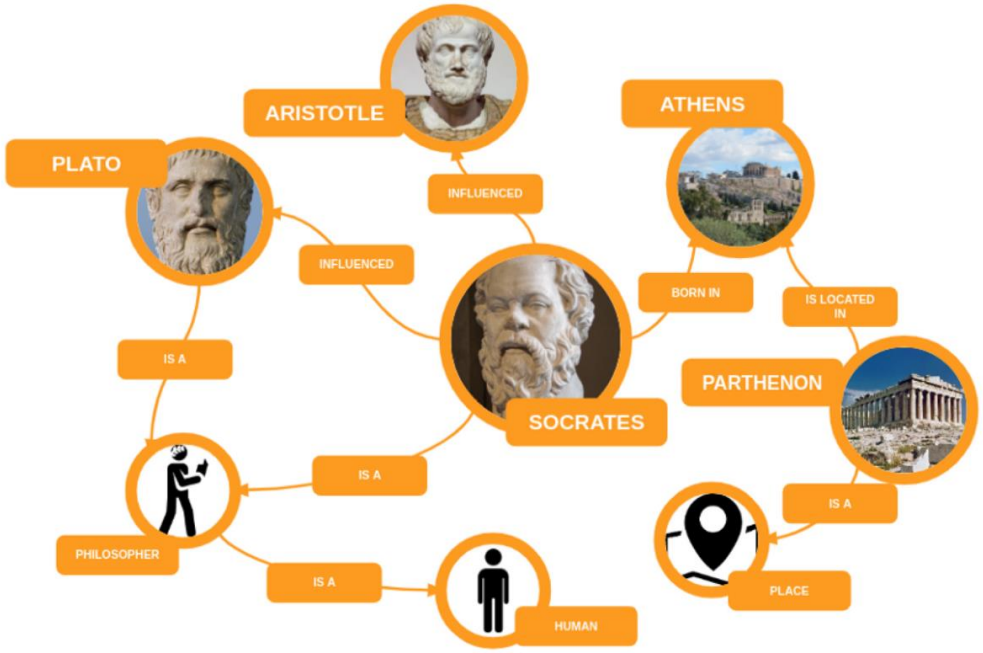
Data-driven XAI Methods



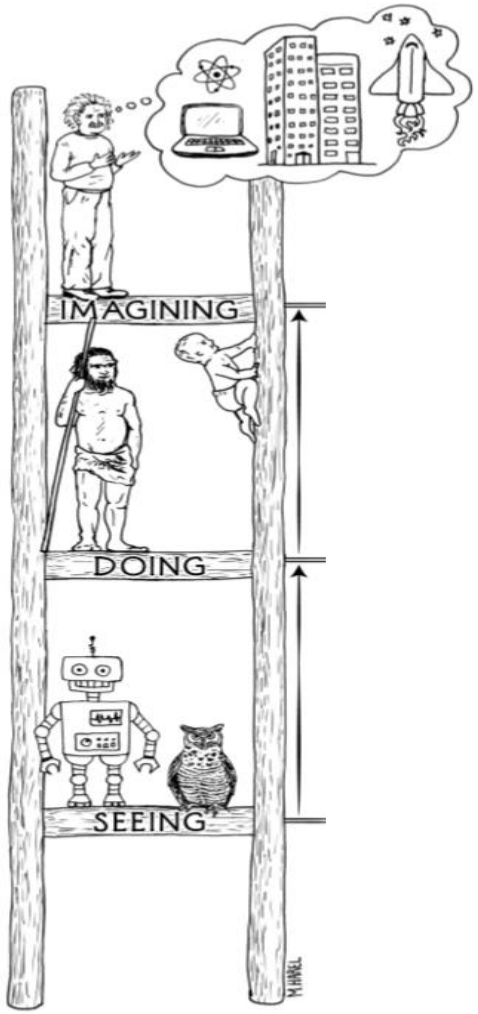
Knowledge-aware XAI Methods

General Knowledge Methods: General knowledge could be diverse forms, natural language, audios or rules with strict logic.

Knowledge-based methods: Knowledge Graph



Causal Hierarchy of Human Thinking



The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

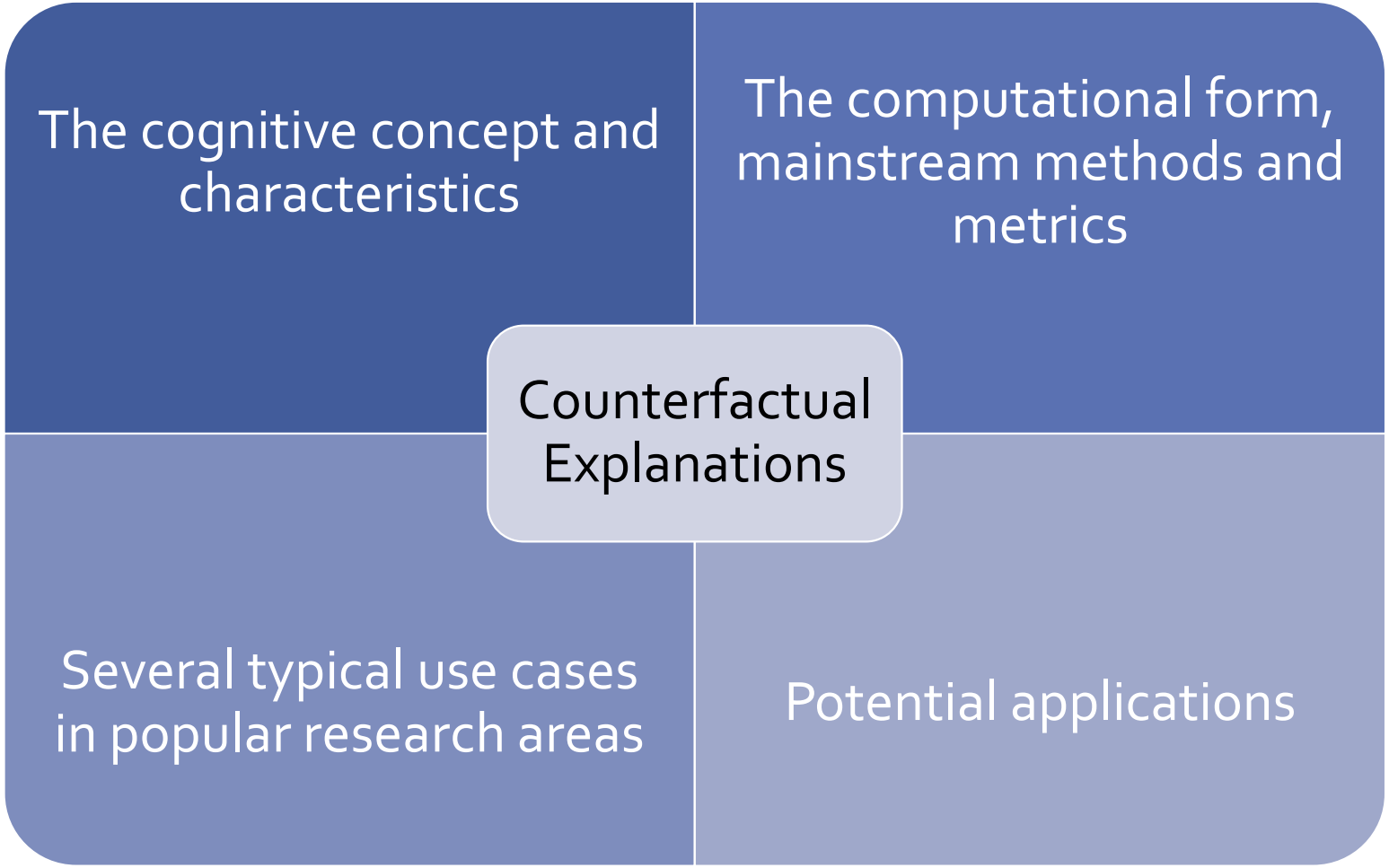
Figure 1: The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Counterfactual Thinking in Daily Life

If I had gone to bed earlier last night, I wouldn't have been late

We hope you could enjoy:



What is Counterfactual Explanation

02

Caleb Chen CAO

Example 1

applied

loan

rejected

Income
number of credit cards
Age
etc

Minimum Change?

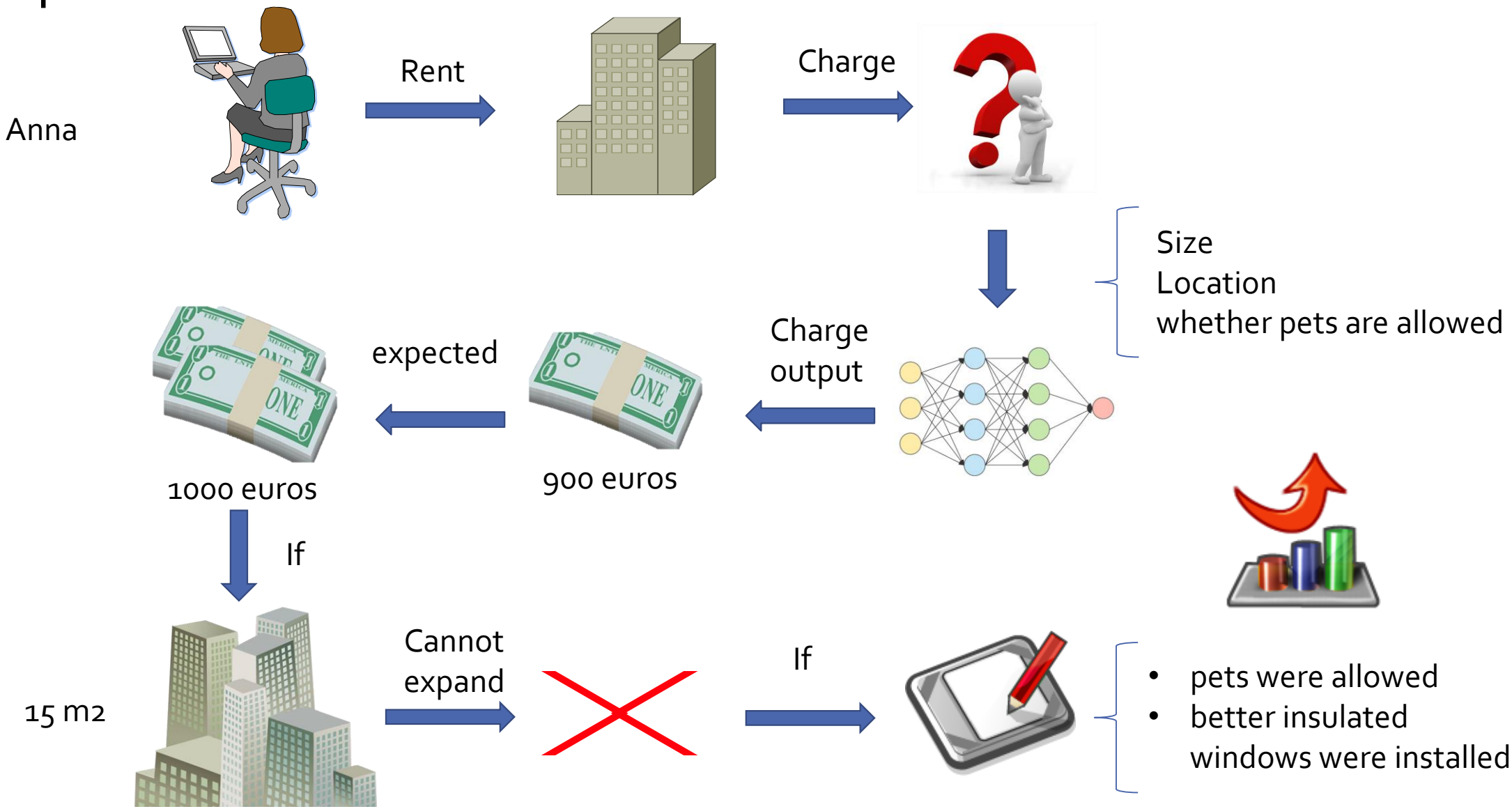
If

Get loan

- earns an extra 10,000 euros a year.
- fewer credit cards
- hadn't defaulted on his loan five years ago

Molnar, C. (2020). Interpretable machine learning. Lulu. com.

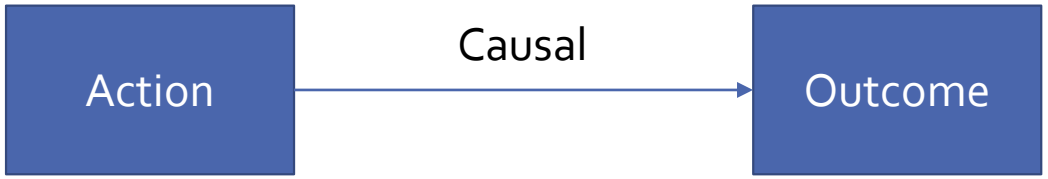
Example 2



Molnar, C. (2020). Interpretable machine learning. Lulu. com.

Counterfactuals and Causes

Counterfactuals amplify the causal link between an action and its outcome



"if the car had swerved into the middle of the road instead, the passenger would not have been injured"

Their judgments of a causal relation between the antecedent, swerving into the wall, and the outcome, the passenger being injured, are amplified.

Byrne, R. M. (2019, August). Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *IJCAI* (pp. 6276-6282).

Counterfactual Content

- Exceptions: People create a counterfactual by changing an exceptional event to be normal.

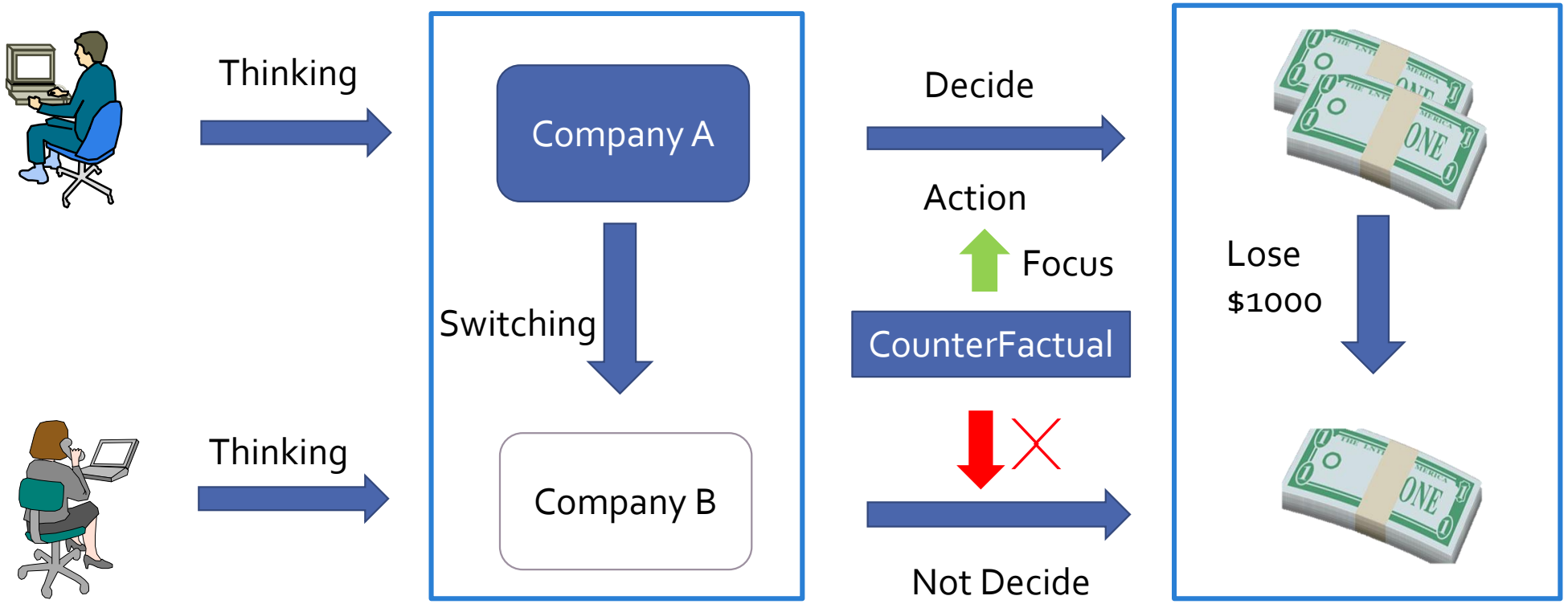
- Controllability: People tend to create a counterfactual in which they change an event within a protagonist's control

  Two envelopes, one is easy sum the other is hard sum. People tended to create counterfactuals that changed the event within her control, "if only she had chosen the other envelope…"

Byrne, R. M. (2019, August). Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *IJCAI* (pp. 6276-6282).

Counterfactual Content

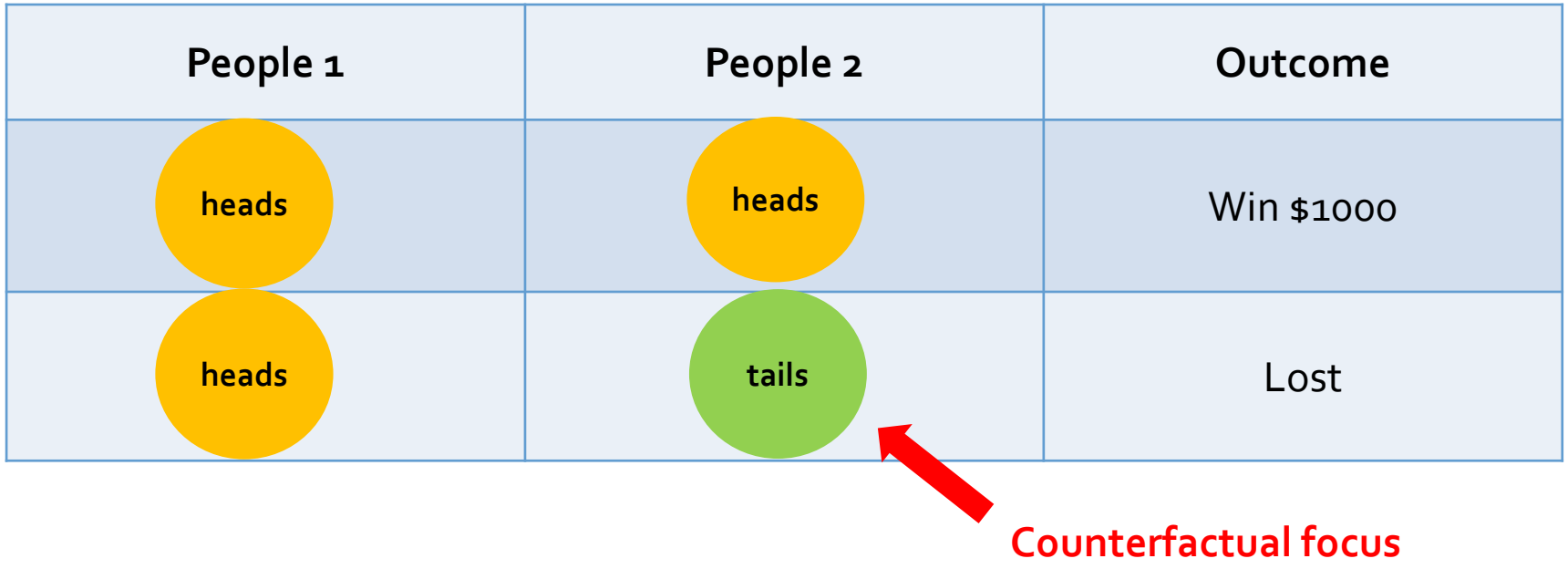
- Actions: People modify actions rather than failures to act when they create a counterfactual



Byrne, R. M. (2019, August). Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *IJCAI* (pp. 6276-6282).

Counterfactual Content

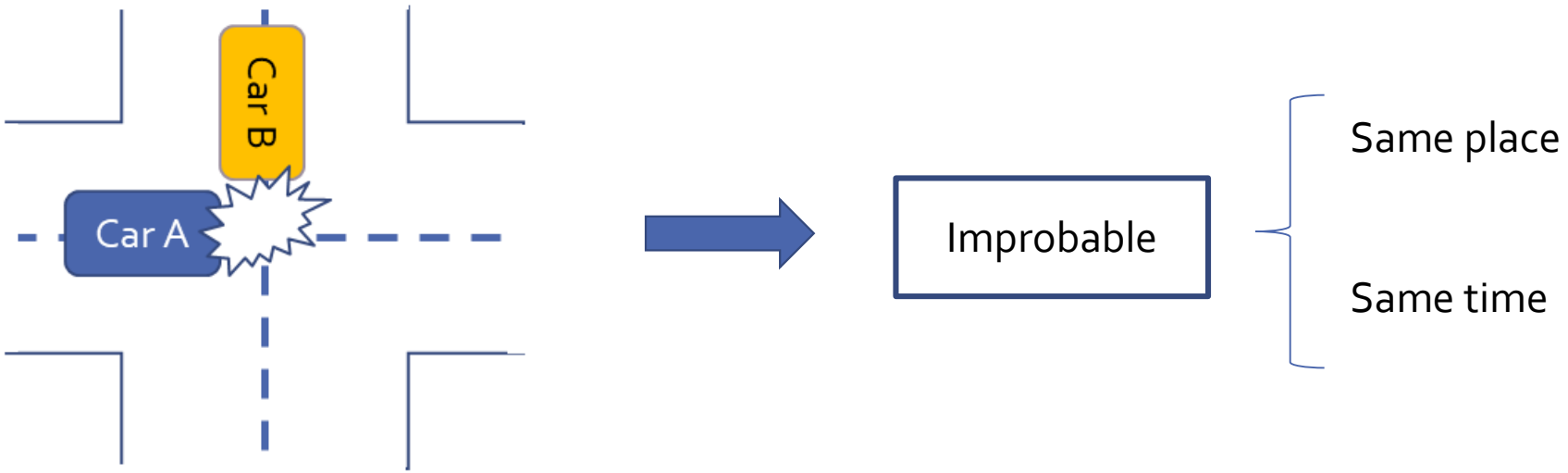
- Recent Events: People create a counterfactual in which they change the most recent event in a temporal sequence of independent events



Byrne, R. M. (2019, August). Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *IJCAI* (pp. 6276-6282).

Counterfactual Content

- Probability: Counterfactual thoughts tend to be rooted in reality – people rarely imagine fantastical alternatives



Byrne, R. M. (2019, August). Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *IJCAI* (pp. 6276-6282).

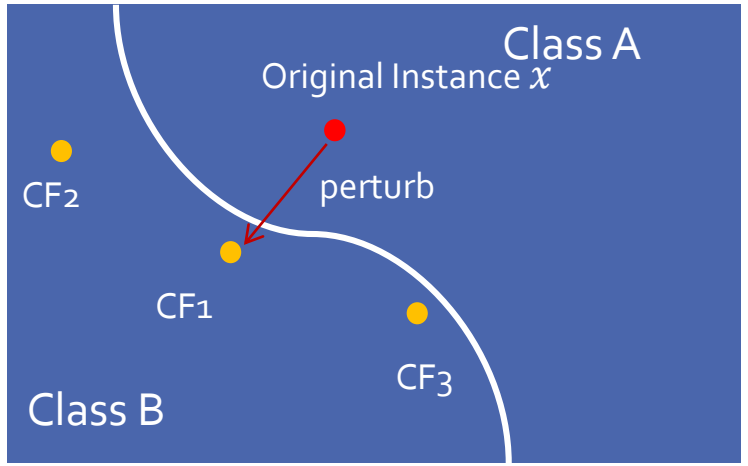
Terminology

- Original instance: The instance of interest that we want to know what action we do could alter the prediction output.

- Perturb: we would like to use the word perturb to illustrate the action of modifying / editing / changing we made on the original instance.

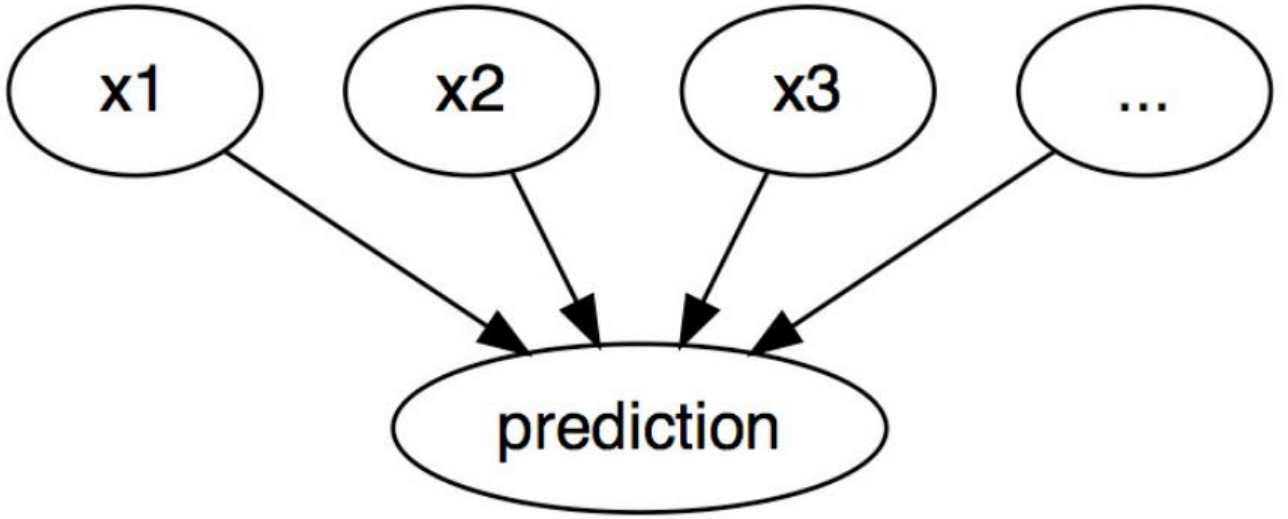
  The counterfactual content human imagined could be considered as the perturbation on original instance.

  - CF: Counterfactual
  - CFX: Counterfactual Explanations



The Relationship Between Causes and Predictions

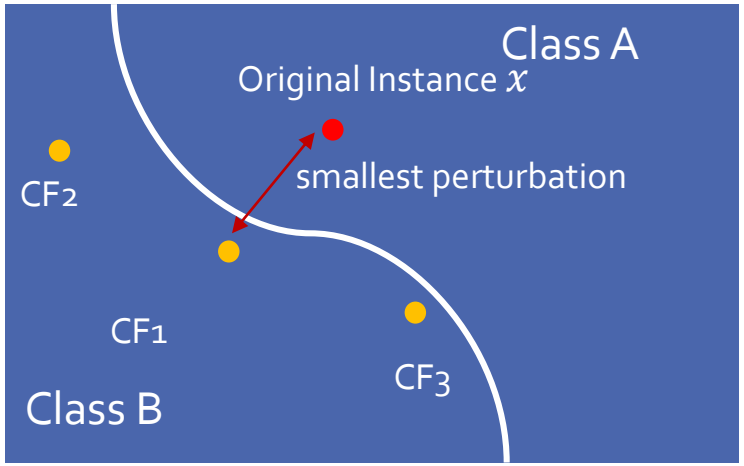
“Event" is the prediction result of an instance, and "cause" is the specific features of that instance that are input into the model and "cause" some prediction.



Molnar, C. (2020). Interpretable machine learning. Lulu. com.

Definitions of Counterfactual Explanation

"A counterfactual explanation of a prediction describes the smallest perturbation to the feature values that changes the prediction to a predefined output."



By creating counterfactual instances, we can learn how the model makes decisions and how individual instances are interpreted.

Molnar, C. (2020). Interpretable machine learning. Lulu. com.

Format of Counterfactual Explanation

"If X had not occurred, Y would not have occurred".

It answered the question "What would have happend if …?" and "Why"

Molnar, C. (2020). Interpretable machine learning. Lulu. com.
Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

03

How to Compute

Caleb Chen CAO

Xiao-hui LI

Han GAO

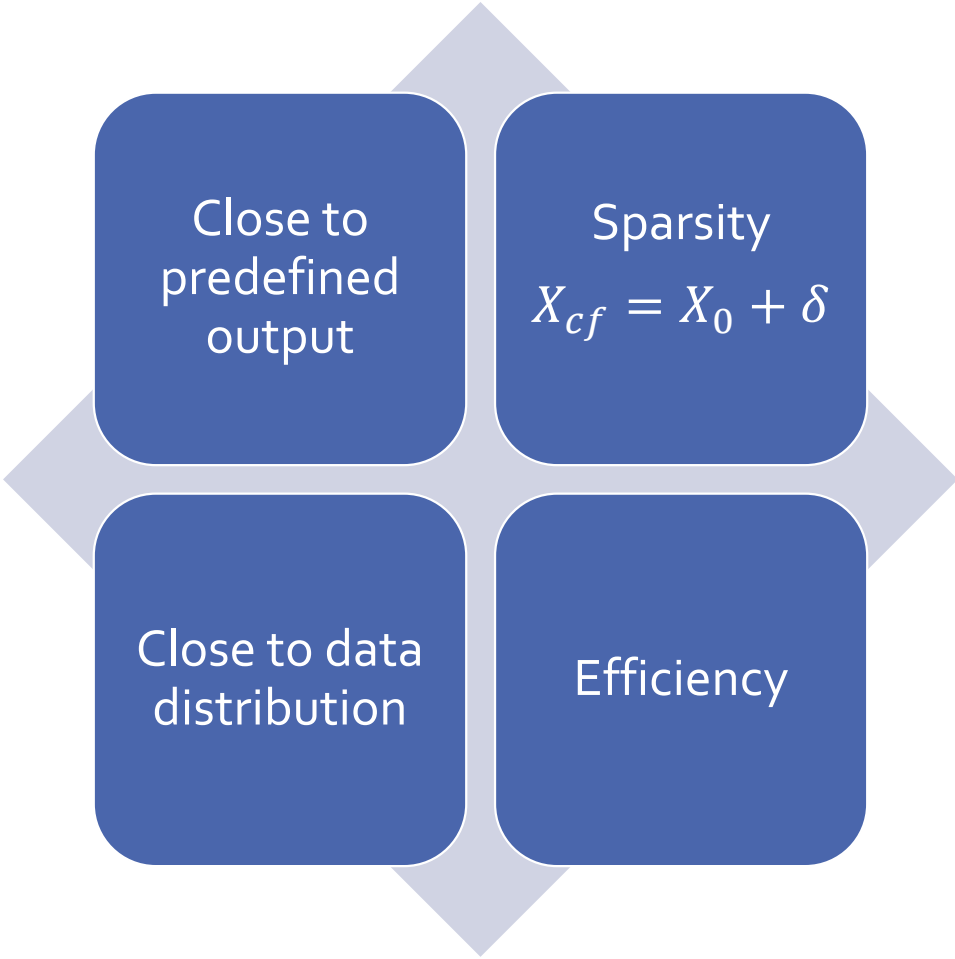
How to Define a "Good" Counterfactual Explanation

Close to predictions	• Counterfactual instances should produce predefined predictions as **close** as possible.
Similarity	• Counterfactuals should be as **similar** as possible to instances. Change as few features as possible
Rationality	• Counterfactual instances should have **possible** features.

Molnar, C. (2020). Interpretable machine learning. Lulu. com.

Desiderata

A counterfactual instance X_{cf} should have the following desirable properties:



Computation Methods

Classic Methods

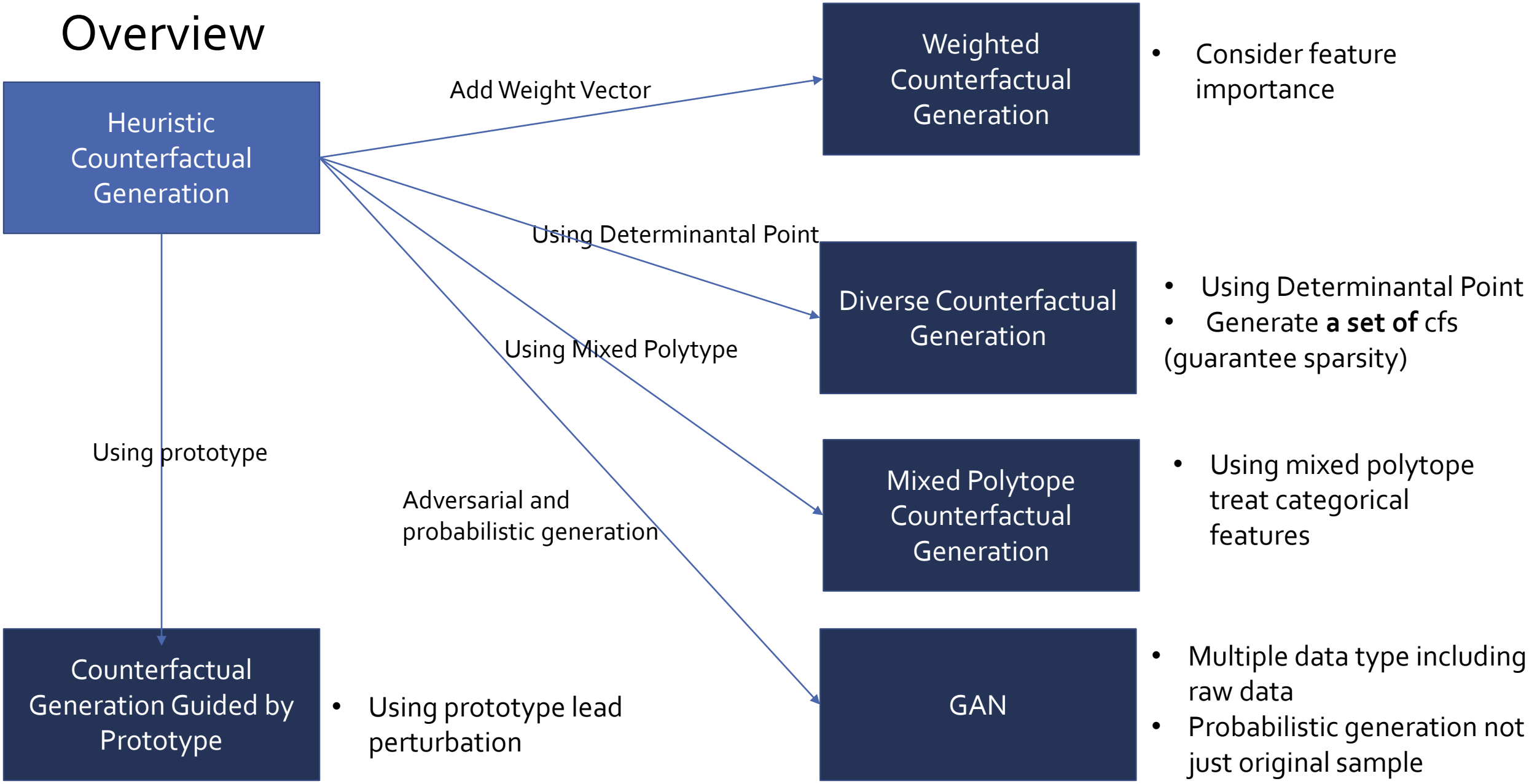
- Heuristic

Keywords for Advanced Methods

- Weighted
- Diverse
- Mixed Polytope
- Prototype

- and GAN …

Overview


Heuristic Counterfactual Generation
Add Weight Vector
Weighted Counterfactual Generation
Consider feature importance
Using Determinantal Point
Diverse Counterfactual Generation
Using Determinantal Point
Generate a set of cfs (guarantee sparsity)
Using Mixed Polytype
Mixed Polytope Counterfactual Generation
Using mixed polytope treat categorical features
Using prototype
Adversarial and probabilistic generation
Counterfactual Generation Guided by Prototype
Using prototype lead perturbation
GAN
Multiple data type including raw data
Probabilistic generation not just original sample

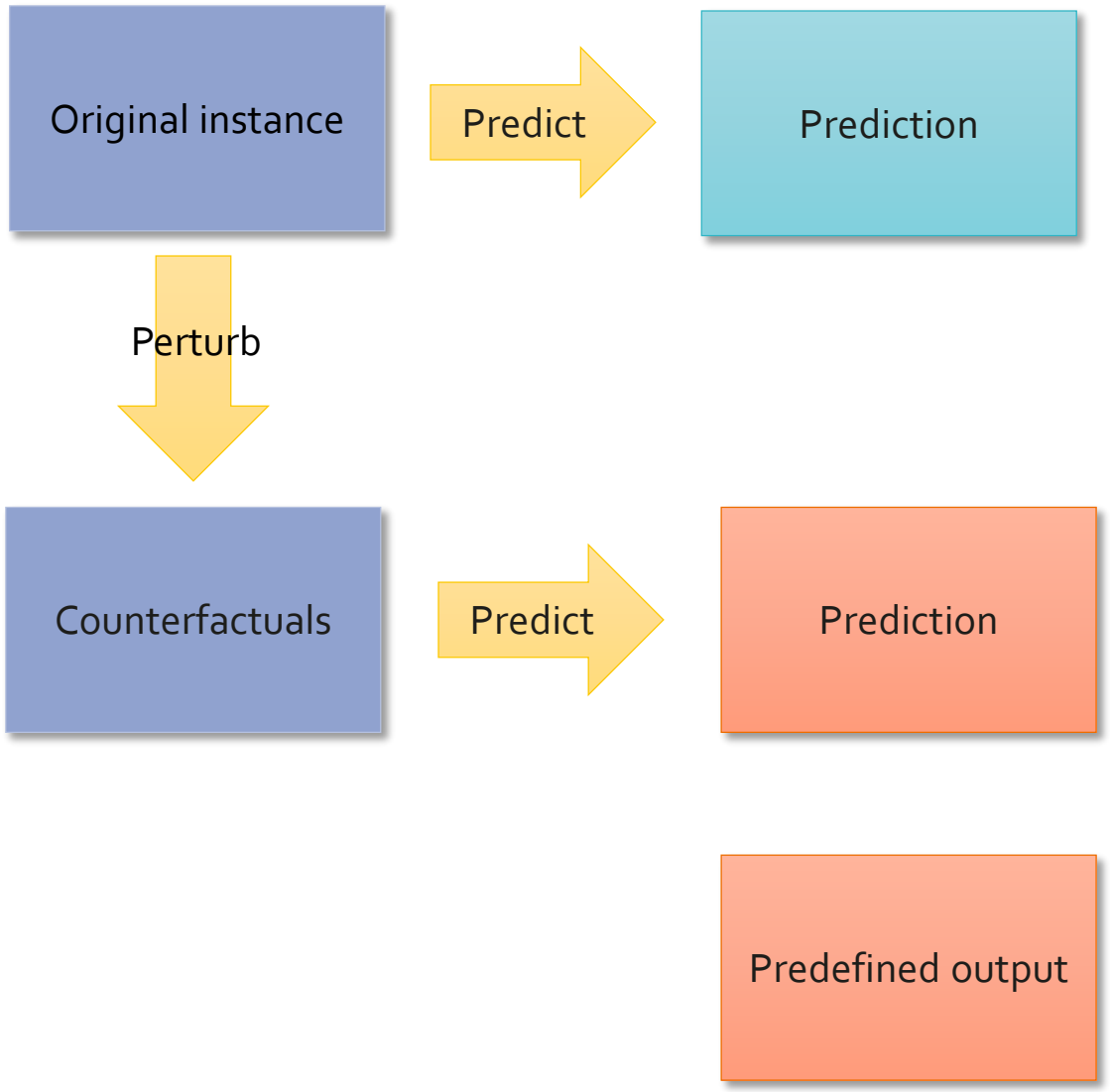
Heuristic Counterfactual Generation

the distance in predictions

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

the distance in instances

Distance metric: $$d(x, x') = \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j},$$



Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. *Harv. JL & Tech.*, *31*, 841.

Heuristic Counterfactual Generation - Algorithm

Algorithm 1: Counterfactual generation heuristic

sample a random instance as the initial x'
optimise $L(x, x', y', \lambda)$ with initial x'
while $|\hat{f}(x') - y'| > \varepsilon$ **do**
    increase λ by step-size α
    optimise $L(x, x', y', \lambda)$ 1 with new x'
return x'

x: interested instance
x′ : counterfactual
y′: desired outcome
λ : balances two terms
ϵ : tolerance for the distance

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. *Harv. JL & Tech.*, *31*, 841.

Heuristic Counterfactual Generation – LSAT Example

Score	GPA	LSAT	Race	GPA x'	LSAT x'	Race x'
0.17	3.1	39.0	0	3.1	34.0	0
0.54	3.7	48.0	0	3.7	32.4	0
-0.77	3.3	28.0	1	3.3	33.5	0
-0.83	2.4	28.5	1	2.4	35.8	0
-0.57	2.7	18.3	0	2.7	34.9	0

Predictions

Instances

Counterfactual

Desired prediction : Score = 0

LSAT: law school admission test.

Counterfactual Explanation for Person 1:
If your LSAT was 34.0, you would have an average predicted score (0).

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. *Harv. JL & Tech.*, *31*, 841.

Heuristic Counterfactual Generation – LSAT Example Problem

Score	GPA	LSAT	Race	GPA x'	LSAT x'	Race x'
0.17	3.1	39.0	0	3.1	34.0	0
0.54	3.7	48.0	0	3.7	32.4	0
-0.77	3.3	28.0	1	3.3	33.5	0
-0.83	2.4	28.5	1	2.4	35.8	0
-0.57	2.7	18.3	0	2.7	34.9	0

Predictions

Instances

Counterfactual

Desired prediction : Score = 0

Counterfactual Explanation for Person 3:

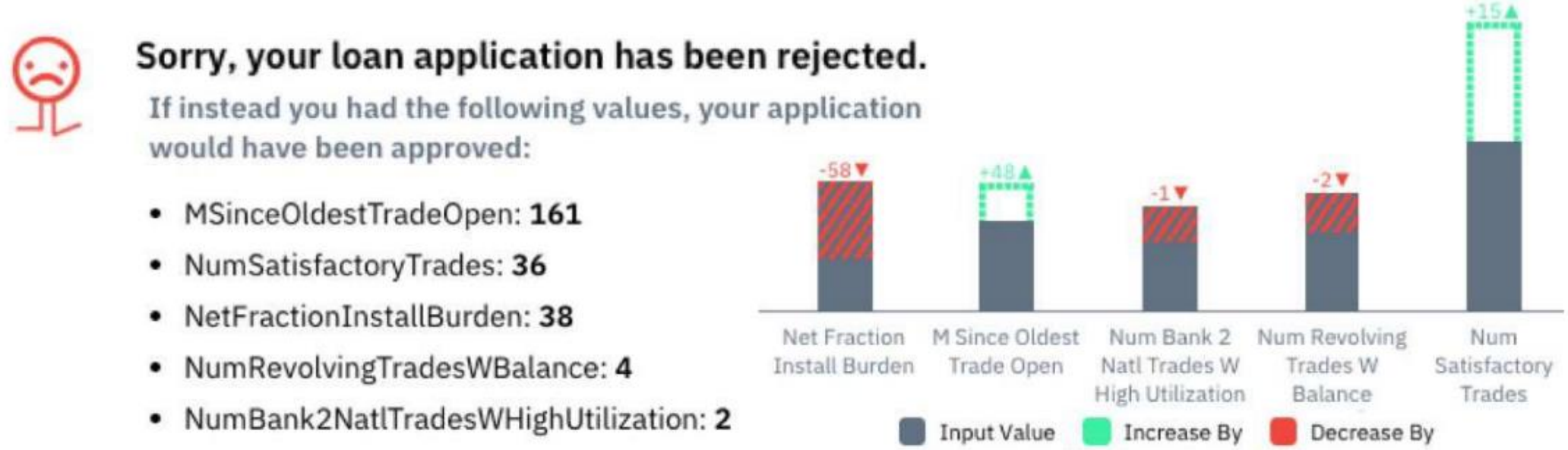
If your LSAT was 33.5, and you were 'white', you would have an average predicted score (0).

But the race should be considered as important as any other features.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. *Harv. JL & Tech.*, *31*, 841.

Weighted Counterfactual Generation - Motivation

- **Heuristic Counterfactual Generation** assumes all features are equally important



- However, each feature's ability to change and the magnitude of the change may vary

Mc Grath, R., Costabello, L., Le Van, C., Sweeney, P., Kamiab, F., Shen, Z., & Lecue, F. (2018, December). Interpretable Credit Application Predictions With Counterfactual Explanations. In *NIPS 2018-Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy*.

Weighted Counterfactual Generation – Methodology

- Promoting highly discriminative features during the generation of counterfactuals

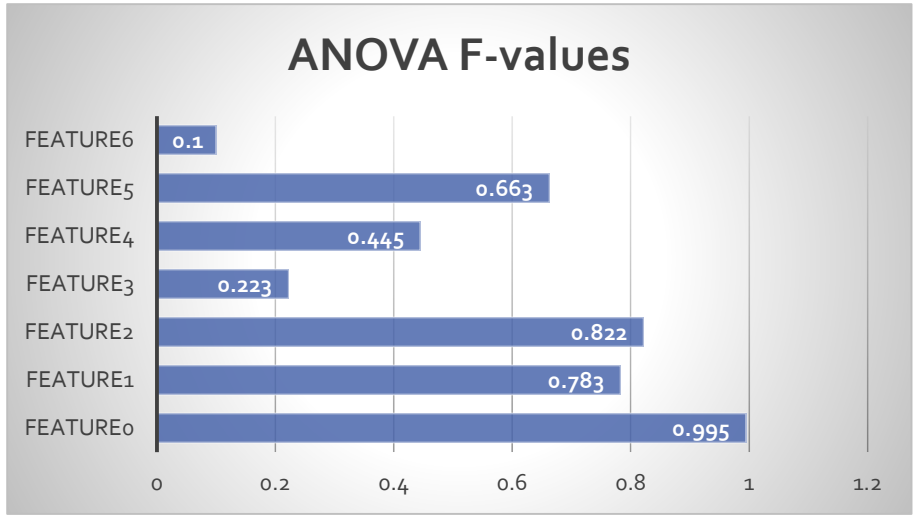
Distance metric:

$$d_2(x, x') = \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j} \theta_j,$$

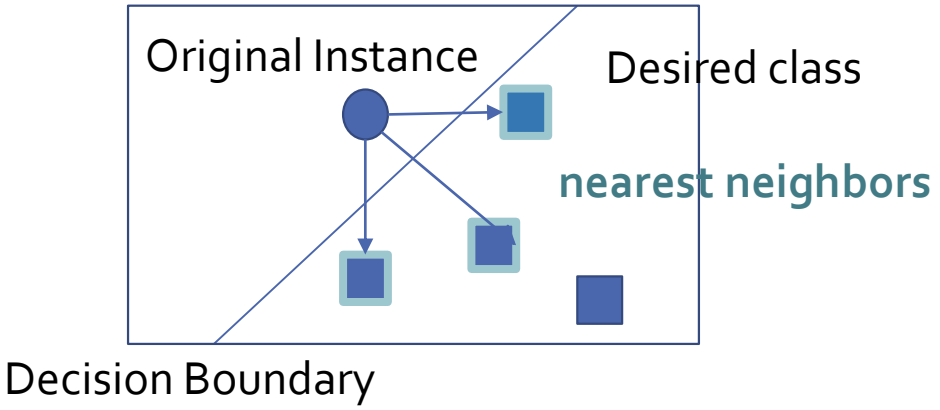
weight vector

- Two weighting strategies:

Global feature importance



K-Nearest Neighbors



Mc Grath, R., Costabello, L., Le Van, C., Sweeney, P., Kamiab, F., Shen, Z., & Lecue, F. (2018, December). Interpretable Credit Application Predictions With Counterfactual Explanations. In *NIPS 2018-Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy*.

Weighted Counterfactual Generation – Experiment Results

- Best results in bold

Model	HELOC		
	Baseline	Importance	KNN
LogReg	4.86±1.84	**3.95±1.69**	4.71±1.72
MLP	8.88±2.54	**8.34±2.58**	8.45±2.53
GradBoost	1.5±0.6	**1.49±0.58**	1.5±0.58
SVC	2.5±1.32	**2.01±1.14**	2.44±1.27

Importance=global feature importance strategy, KNN=k-nearest neighbours. KNN uses k = 20,

Value: Average size (i.e. average number of features) of generated counterfactual explanations, smaller means better

Mc Grath, R., Costabello, L., Le Van, C., Sweeney, P., Kamiab, F., Shen, Z., & Lecue, F. (2018, December). Interpretable Credit Application Predictions With Counterfactual Explanations. In *NIPS 2018-Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy*.

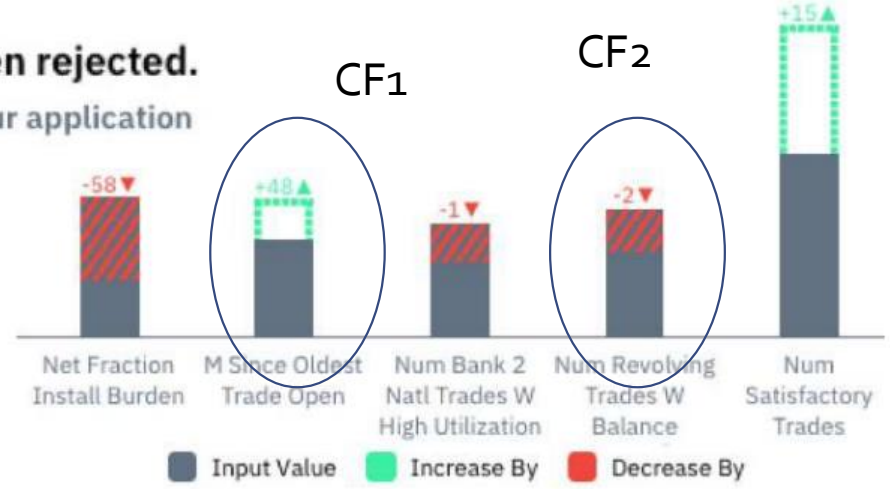
Diverse Counterfactual Generation - Motivation

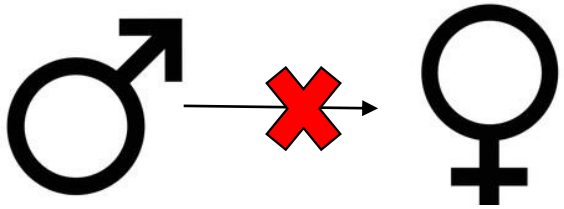
It's more user-friendly to generate multiple counterfactuals that are **actionable**:

Sorry, your loan application has been rejected.

If instead you had the following values, your application would have been approved:

- MSinceOldestTradeOpen: **161**
- NumSatisfactoryTrades: **36**
- NetFractionInstallBurden: **38**
- NumRevolvingTradesWBalance: **4**
- NumBank2NatlTradesWHighUtilization: **2**





Not actionable and gender discriminative

Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607-617).*

Diverse Counterfactual Generation - Methodology

- Loss function:

$$C(\boldsymbol{x}) = \underset{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_k}{\arg\min} \frac{1}{k}\sum_{i=1}^{k} \text{yloss}(f(\boldsymbol{c}_i), y) + \frac{\lambda_1}{k}\sum_{i=1}^{k} dist(\boldsymbol{c}_i, \boldsymbol{x}) - \lambda_2 \,\text{dpp\_diversity}(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k)$$

Diversity via Determinantal Point Processes, $dist(c_i, c_j)$ denotes the distance between two counterfactual example

$$dpp\_diversity = det(\mathbf{K})$$

$$\mathbf{K}_{i,j} = \frac{1}{1+dist(\boldsymbol{c}_i, \boldsymbol{c}_j)}$$

Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607-617).*

Diverse Counterfactual Generation - Practical Implementation

- Choice of $yloss$

$$hinge\_yloss = max(0, 1 - z * logit(f(\boldsymbol{c}))),$$

- Choice of distance function

For continuous features

$$\text{dist\_cont}(\boldsymbol{c}, \boldsymbol{x}) = \frac{1}{d_{cont}} \sum_{p=1}^{d_{cont}} \frac{|\boldsymbol{c}^p - \boldsymbol{x}^p|}{MAD_p}$$

d_{cont} : the number of continuous variables
MAD_p: the median absolute deviation for the p-th continuous variable.

For categorical features

$$\text{dist\_cat}(\boldsymbol{c}, \boldsymbol{x}) = \frac{1}{d_{cat}} \sum_{p=1}^{d_{cat}} I(\boldsymbol{c}^p \neq \boldsymbol{x}^p),$$

d_{cat} : the number of categorical variables
$I(\cdot)$: Indicator function

Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607-617).*

Diverse Counterfactual Generation – Experiment Results

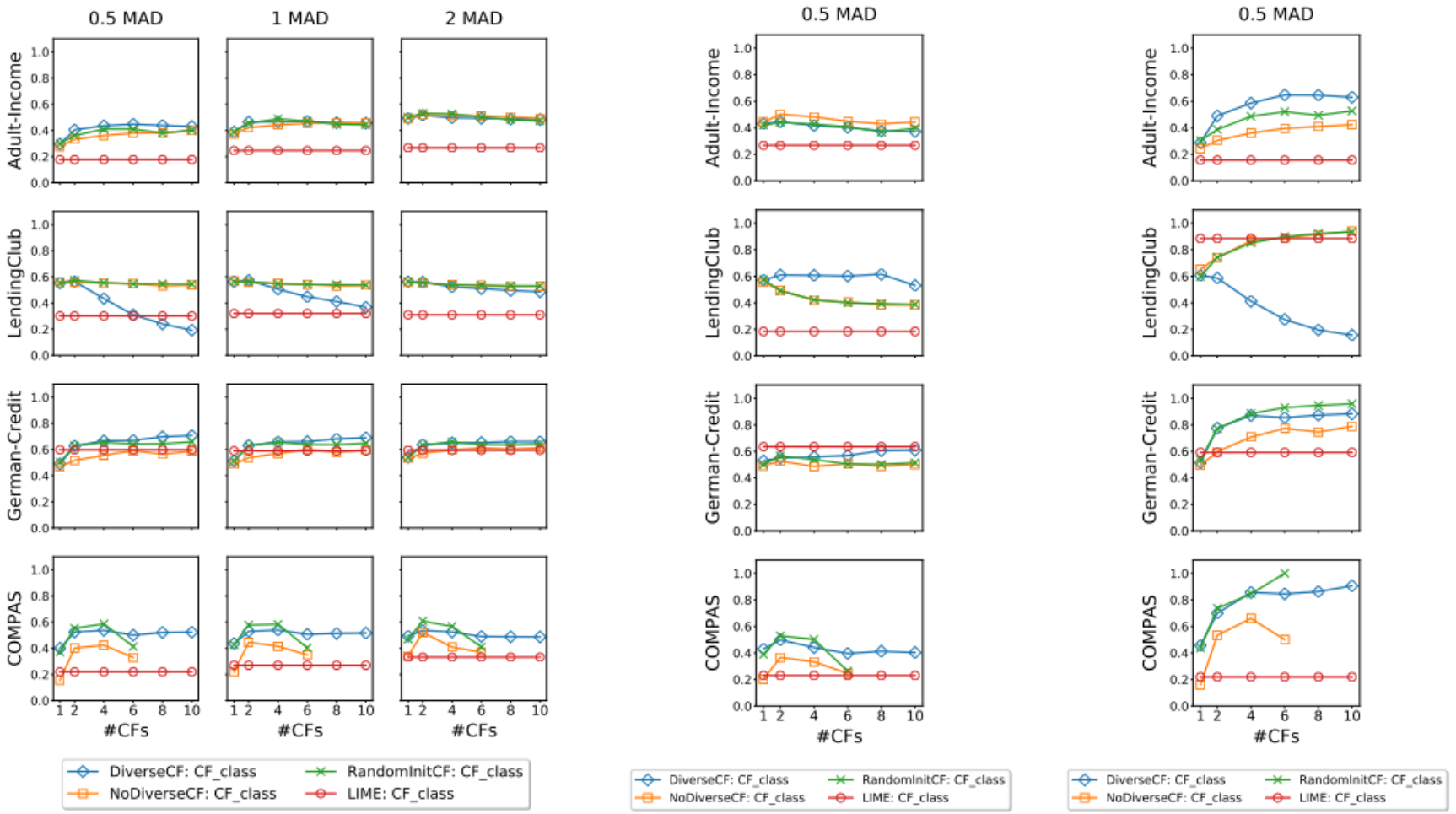
Adult	HrsWk	Education	Occupation	WorkClass	Race	AgeYrs	MaritalStat	Sex
Original input (outcome: <=50K)	45.0	HS-grad	Service	Private	White	22.0	Single	Female
Counterfactuals (outcome: >50K)	–	Masters	–	–	–	65.0	Married	Male
	–	Doctorate	–	Self-Employed	–	34.0	–	–
	33.0	–	White-Collar	–	–	47.0	Married	–
	57.0	Prof-school	–	–	–	–	Married	–

LendingClub	EmpYrs	Inc$	#Ac	CrYrs	LoanGrade	HomeOwner	Purpose	State
Original input (outcome: Default)	7.0	69996.0	4.0	26.0	D	Mortgage	Debt	NY
Counterfactuals (outcome: Paid)	–	61477.0	–	–	B	–	Purchase	–
	10.0	83280.0	1.0	23.0	A	–	–	TX
	10.0	69798.0	–	40.0	A	–	–	–
	10.0	130572.0	–	–	A	Rent	–	–

COMPAS	PriorsCount	CrimeDegree	Race	Age	Sex
Original input (outcome: Will Recidivate)	10.0	Felony	African-American	>45	Female
Counterfactuals (outcome: Won't Recidivate)	–	–	Caucasian	–	–
	0.0	–	–	–	Male
	0.0	–	Hispanic	–	–
	9.0	Misdemeanor	–	–	–

Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607-617).*

Diverse Counterfactual Generation – Experiment Results



(a) F1 score of the counterfactual class. (b) Precision. (c) Recall.

Figure 3: Performance of 1-NN classifiers learned from counterfactuals at different distances from the original input. `DiverseCF` outperforms LIME and baseline CF methods in F1 score on correctly predicting the counterfactual class, except in `LendingClub` dataset. For `Adult-Income` and `COMPAS` datasets, both precision and recall is higher for `DiverseCF` compared to `LIME`.

Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607-617).*

Mixed Polytope Counterfactual Generation - Motivation

- **Heuristic Counterfactual Generation** had difficulty with categorical features:

Score	GPA	LSAT	Race	GPA x'	LSAT x'	Race x'
0.17	3.1	39.0	0	3.1	34.0	0
0.54	3.7	48.0	0	3.7	32.4	0
-0.77	3.3	28.0	1	3.3	33.5	-0.7
-0.83	2.4	28.5	1	2.4	35.8	0
-0.57	2.7	18.3	0	2.7	34.9	0

Nonsense value for categorical features could be generated

Russell, C. (2019, January). Efficient search for diverse coherent explanations. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 20-28).

Mixed Polytope Counterfactual Generation - Methodology

Mixed polytope of variable i is described by the linear constraints:

$$\sum_j d_{i,j} + d_{i,c} = 1$$

$$F_{i,} - l_i + u_i = c_i$$

$$0 \leq l_i \leq (L_i - F_i) d_{i,c}$$

$$0 \leq r_i \leq (R_i - F_i) d_{i,c}$$

$$d_{i,j} \in [0,1] \qquad \forall j$$

$d_{i,j}$ for the jth component of the ith set of indicator variables
c_i is the i-th contiguous variable takes range [Li ,Ui]
$d_{i,c}$ is an additional indicator value that shows that variable v is takes a contiguous value

Russell, C. (2019, January). Efficient search for diverse coherent explanations. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 20-28).

Mixed Polytope Counterfactual Generation - Methodology

- Using novel integer program based upon a "mixed polytope" that is guaranteed to generate coherent counterfactuals that map back into the same form as the original data.

$$\arg\min_{x'} d(x, x')$$
$$\text{such that: } f(x') = c$$

Original problem formulation

$$\arg\min_{x'} ||\hat{x} - x'||_{1,w}$$
$$\text{such that: } f(x') \leq 0$$
$$x' \text{ lies on the mixed polytope}$$
$$d_{i,j} \in \{0, 1\} \qquad \forall i, j$$

Mixed polytope problem formulation

Russell, C. (2019, January). Efficient search for diverse coherent explanations. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 20-28).

Mixed Polytope Counterfactual Generation - Explanation

Consider five individuals:

Person	1	2	3	4	5
Race	0	0	1	1	0
LSAT	39.0	48.0	28.0	28.5	18.3
GPA	3.1	3.7	3.3	2.4	2.7

Explanation by Heuristic Counterfactual Generation:

Person 1: If your LSAT was 34.0, you would have an average predicted score (0).
Person 2: If your LSAT was 32.4, you would have an average predicted score (0).

Explanation by Mixed Polytope Counterfactual Generation :

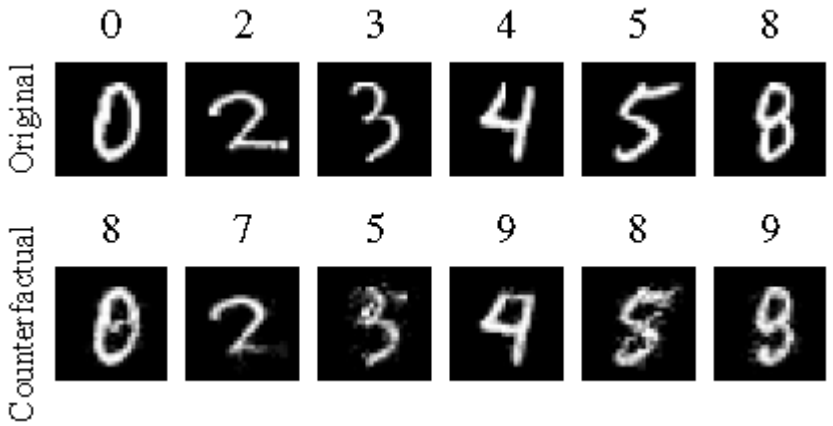
```
You got score 'above average'.
 One way you could have got score 'below average' is if:
  lsat took value 33.9 rather than 39.0
_____
Another way you could have got score 'below average'is if :
  gpa had taken value 2.5 rather than 3.1
______
Another way you could have got score 'below average' is if :
  isblack had taken value 1 rather than 0
```

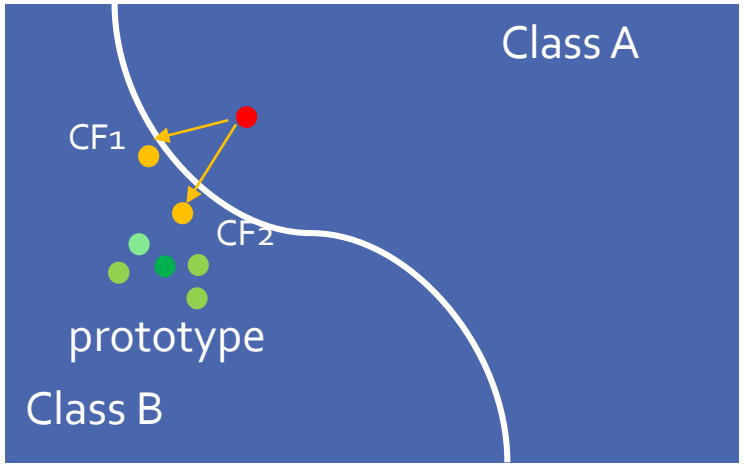
```
You got score 'above average'.
 One way you could have got score 'below average' is if:
  lsat took value 32.3 rather than 48.0
_____
Another way you could have got score 'below average'is if :
  isblack took value 1 rather than 0.
```

Russell, C. (2019, January). Efficient search for diverse coherent explanations. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 20-28).

Counterfactual Generation Guided by Prototype - Motivation

1. Heuristic generation for counterfactual explanation takes long to converge
2. The converging counterfactual may not be interpretable





Using prototype as a guide, there can be a good direction for the perturbation generation

Van Looveren, A., & Klaise, J. Interpretable Counterfactual Explanations Guided by Prototypes. *Age*, *46*, 46.

Counterfactual Generation Guided by Prototype - Methodology

Objective function:

$$L = c \cdot L_{\text{pred}} + \beta \cdot L_1 + L_2 + L_{\text{AE}} + L_{\text{proto}}$$

$$\max([f_{\text{pred}}(x_0 + \delta)]_{t_0} - \max_{i \neq t_0}[f_{\text{pred}}(x_0 + \delta)]_i, -\kappa)$$

To change prediction label

$$\beta \cdot \|\delta\|_1 + \|\delta\|_2^2$$

To generate proximate CF

$$\gamma \cdot \|x_0 + \delta - \text{AE}(x_0 + \delta)\|_2^2$$

CFs within data manifold

$$\theta \cdot \|\text{ENC}(x_0 + \delta) - \text{proto}_j\|_2^2$$

To make CF closer to prototype

Van Looveren, A., & Klaise, J. Interpretable Counterfactual Explanations Guided by Prototypes. *Age*, *46*, 46.

Counterfactual Generation Guided by Prototype - Results

Ablation study on different losses combination:

$$A = c \cdot L_{\text{pred}} + \beta \cdot L_1 + L_2$$
$$B = c \cdot L_{\text{pred}} + \beta \cdot L_1 + L_2 + L_{\text{AE}}$$
$$C = c \cdot L_{\text{pred}} + \beta \cdot L_1 + L_2 + L_{\text{proto}}$$
$$D = c \cdot L_{\text{pred}} + \beta \cdot L_1 + L_2 + L_{\text{AE}} + L_{\text{proto}}$$
$$E = \beta \cdot L_1 + L_2 + L_{\text{proto}}$$
$$F = \beta \cdot L_1 + L_2 + L_{\text{AE}} + L_{\text{proto}}$$

Results

Method	Time (s)	Gradient steps	IM1	IM2 ($\times 10$)
A	13.06 ± 0.23	5158 ± 82	1.56 ± 0.03	1.65 ± 0.04
B	8.40 ± 0.38	2380 ± 113	1.36 ± 0.02	1.60 ± 0.03
C	2.37 ± 0.09	751 ± 31	1.23 ± 0.02	1.46 ± 0.03
D	2.05 ± 0.08	498 ± 27	1.26 ± 0.02	1.29 ± 0.03
E	4.39 ± 0.04	1794 ± 12	1.20 ± 0.02	1.52 ± 0.03
F	2.86 ± 0.06	773 ± 16	1.22 ± 0.02	1.29 ± 0.03

Table 1: Summary statistics with 95% confidence bounds for each loss function for the MNIST experiment.

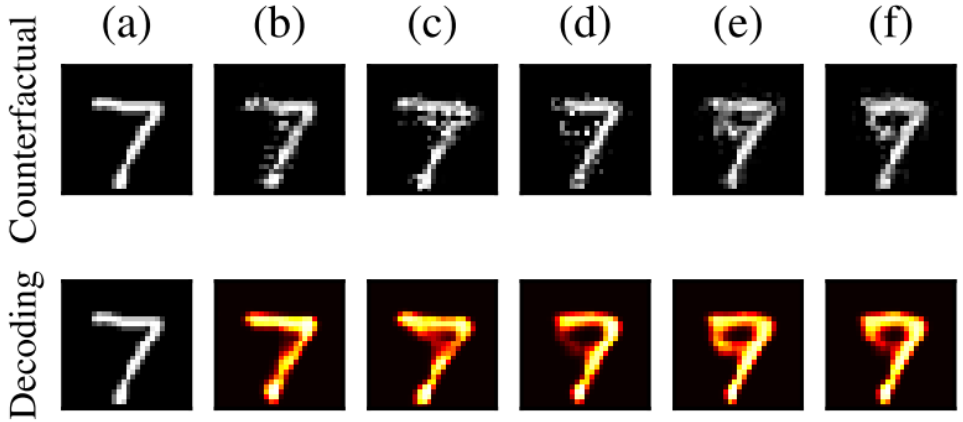


Figure 4: (a) Shows the original instance, (b) to (f) on the first row illustrate counterfactuals generated by using loss functions A, B, C, D and F. (b) to (f) on the second row show the reconstructed counterfactuals using AE.

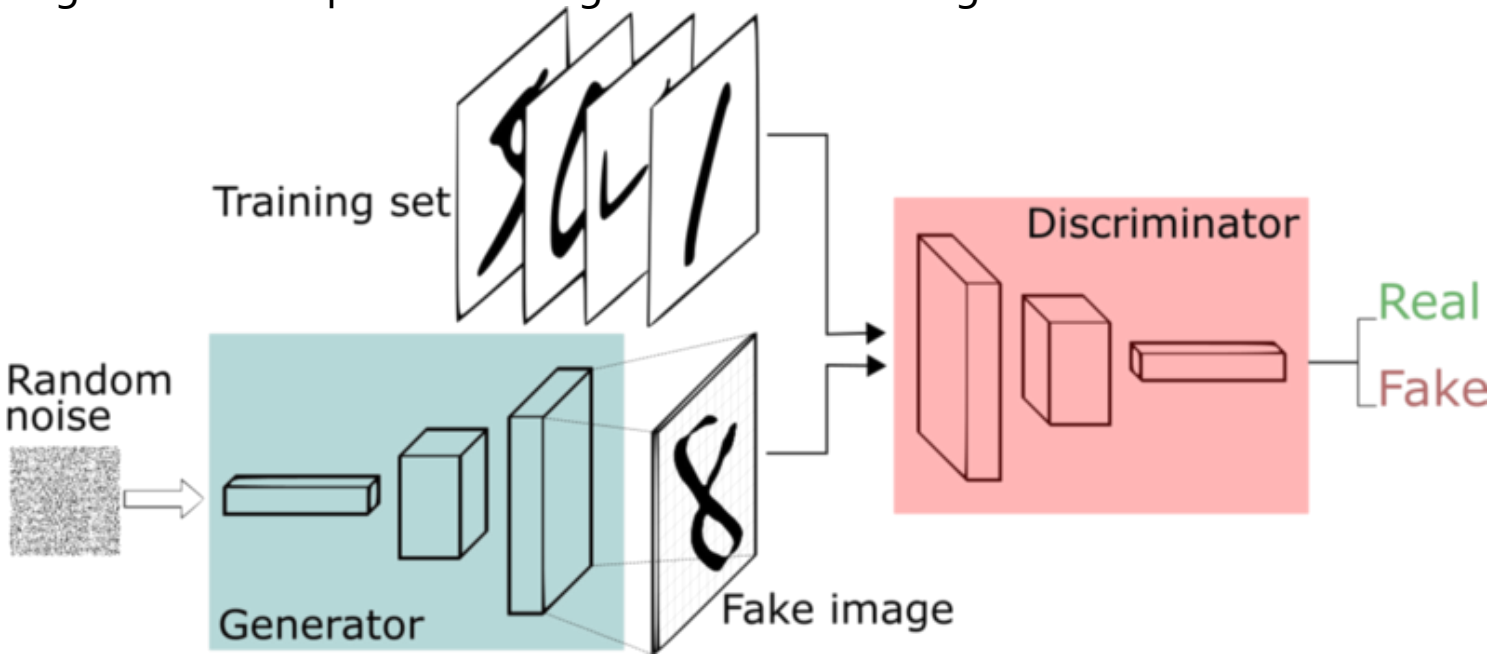
Van Looveren, A., & Klaise, J. Interpretable Counterfactual Explanations Guided by Prototypes. *Age*, *46*, 46.

GAN Introduction

- **GAN (Generative Adversarial Network) has 2 functions locked in a game. They are generator and discriminator**

1) The generator trying to maximize the probability of making the discriminator mistakes its inputs as real.
2) The discriminator guiding the generator to produce more realistic images.

Generator is able to generate samples matching distribution of original dataset.



https://www.freecodecamp.org/news/an-intuitive-introduction-to-generative-adversarial-networks-gans-7a2264a81394/

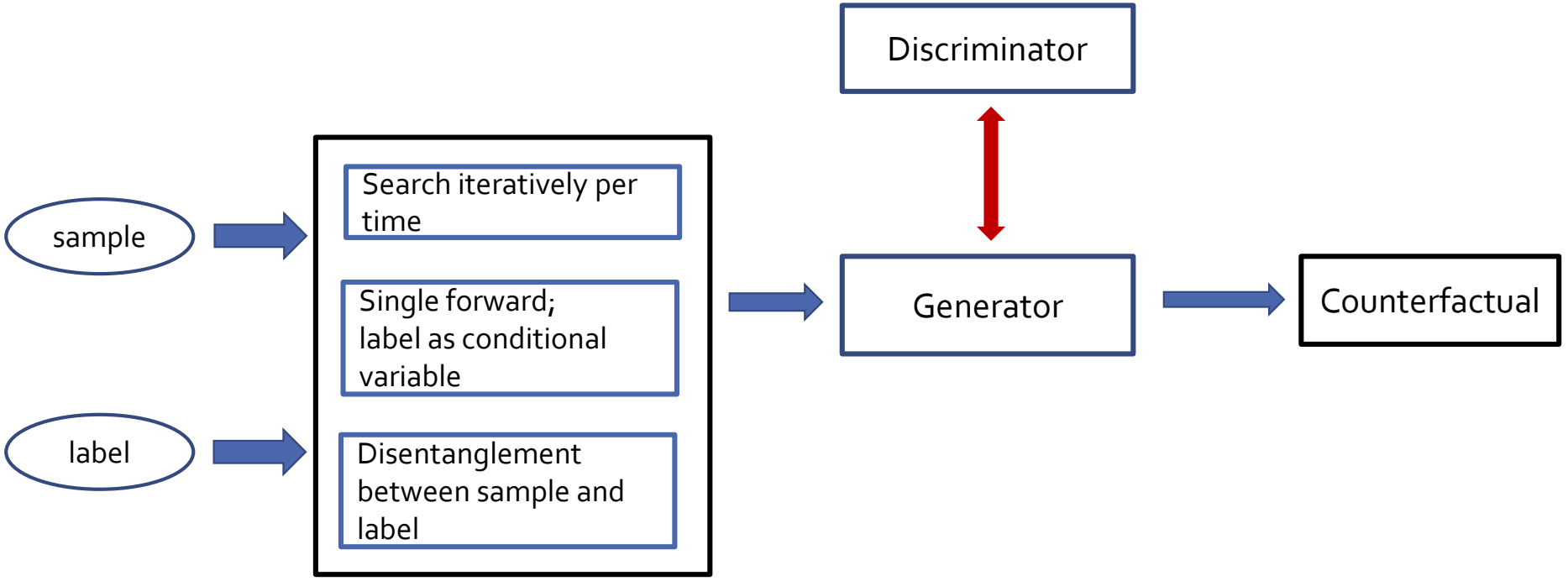
Use GAN to Compute and Generate Counterfactuals

- General GAN Framework to generate counterfactuals

a) Search iteratively per time with an existing GAN

b) Forward computing per time with label embedding as conditional variable into GAN

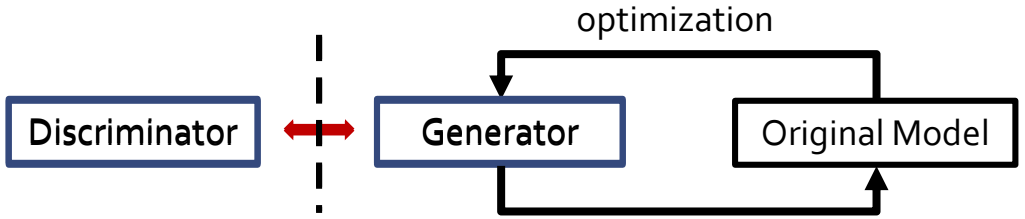
- Pay attention to disentanglement between sample's embedding and label's embedding



https://www.freecodecamp.org/news/an-intuitive-introduction-to-generative-adversarial-networks-gans-7a2264a81394/

General GAN Framework to Generate Counterfactuals

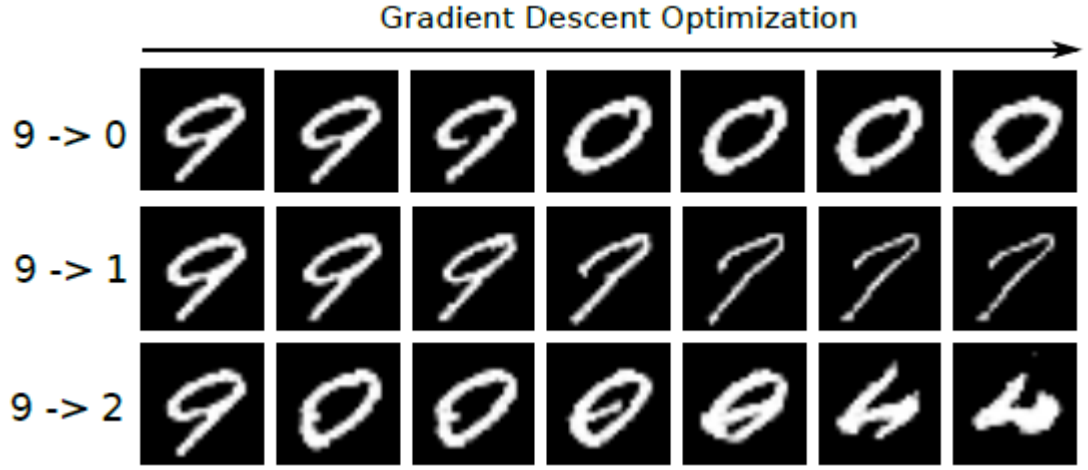
- **Based on an existing GAN, optimizing the classical Counterfactual loss function to generate Counterfactual given each sample**



Here I is an image, $A = \{a_1, \dots, a_N\}$ is the set of attributes. C is the original model and is a classifier as an example here. A' is the target attribute vector. G is the generator from an existing GAN. c' is the target class.

$$\min_{A'} \quad \|I - I(A')\|_p \qquad \text{s.t.} \quad c' = C(I(A')) \qquad I(A') = G(I; A')$$

Below, the input image is number 9. Target classes are 0, 1, ...etc.



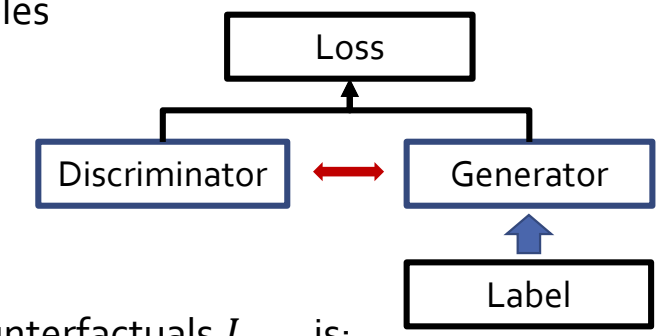
Liu, S., Kailkhura, B., Loveland, D., & Han, Y. (2019, November). Generative counterfactual introspection for explainable deep learning. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (pp. 1-5). IEEE.

General GAN Framework to Generate Counterfactuals

- **Fine tuning GAN to generate Counterfactual with a single forward pass without optimization**

Compared with previous method, it can generate counterfactual with input and a target class directly.

Here, regard labels as conditional variables



The loss of generator for generating counterfactuals $L_{G_{CF}}$ is:

Conditional variable concatenated in Batchnorm layer

minimize $L_{G_{\text{CF}}} = L_{\text{M}} + L_{l_p} + L_{\chi}$

$$L_{\text{M}} = w_{\text{M}} d_{\text{M}}(y_{\text{CF}}, y_{\text{T}})$$

$$L_{l_p} = w_{l_p} d_p(x, x_{\text{CF}}),$$

$$L_{\text{G}} = \mathbb{E}_{x\sim\chi, z\sim p(z)}[\log(1 - D(x_{\text{CF}}, y_{\text{T}}))]$$

$$L_{\text{CC}} = \mathbb{E}_{x\sim\chi, z\sim p(z)}[\|x - G_{\text{CF}}(x_{\text{CF}}, y_{\text{CF}}, y_{\text{M}}, z)\|_1]$$

$$L_{\chi} = w_G L_{\text{G}} + w_{\text{CC}} L_{\text{CC}}$$

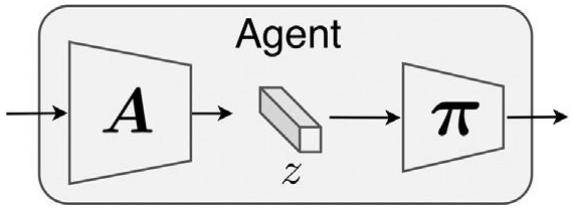
$$L_{\text{D}} = -\mathbb{E}_{x\sim\chi}[\log(D(x, y_{\text{M}}))]$$

minimize $-\mathbb{E}_{x\sim\chi, z\sim p(z)}[\log(1 - D(x_{\text{CF}}, y_{\text{T}}))],$

y_T and y_{CF} are embedding of original and counterfactual label. d_M represents a divergence metric. L_{l_p} induces sparsity of counterfactual. L_{χ} penalizes out-of-distribution counterfactuals.

Van Looveren, A., Klaise, J., Vacanti, G., & Cobb, O. (2021). Conditional Generative Models for Counterfactual Explanations. *arXiv preprint arXiv:2101.10123*.

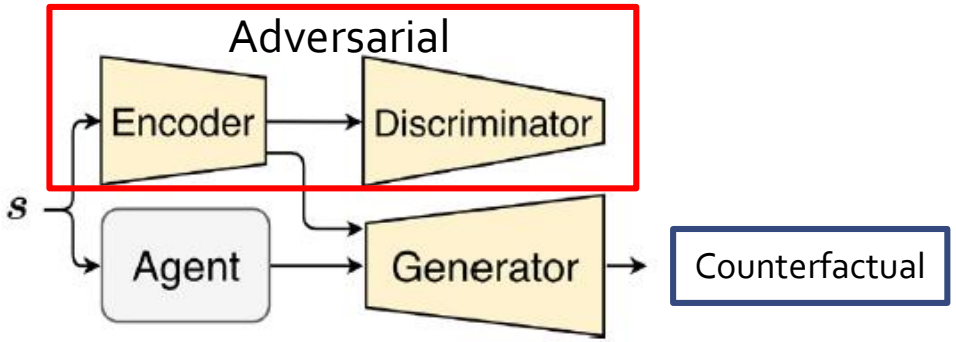
Pay Attention to Disentanglement Between Sample's Embedding and Label's Embedding



- **Use GAN framework to train an encoder which can only extract information of sample**

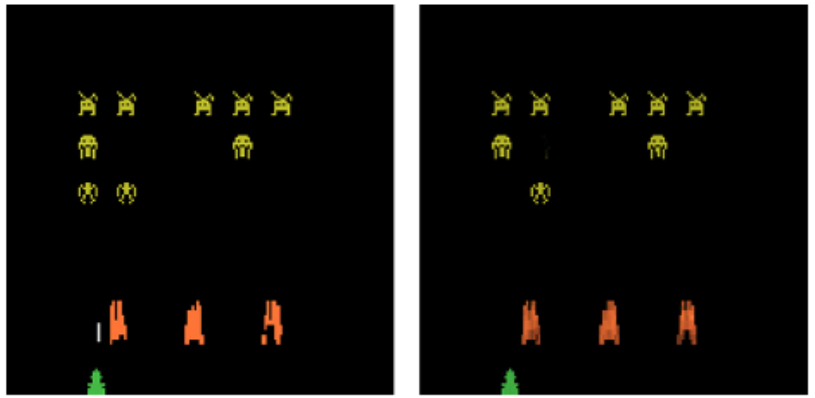
Task: Given an original state s and a target action a', what is the counterfactual state s' ?

An overview of the architecture is shown on the right.



$$L_{AE} = \frac{1}{|\mathcal{X}|} \sum_{(\boldsymbol{s},\boldsymbol{a})\in\mathcal{X}} ||G(E(\boldsymbol{s}), \boldsymbol{\pi}(A(\boldsymbol{s}))) - \boldsymbol{s}||_2^2$$

A single VAE loss will stress more on state while ignoring information from agent. **We train the encoder and discriminator adversarially to disentangle information from state and agent.**



Left: The game state in which an agent takes action "move left and shoot".

Right: The counterfactual state where the agent will take the action "move right". An enemy is removed.

Olson, Matthew L., et al. "Counterfactual state explanations for reinforcement learning agents via generative deep learning." *Artificial Intelligence* 295 (2021): 103455.

Pay Attention to Disentanglement Between Sample's Embedding and Label's Embedding

- **Combined with VAE and contrastive loss to achieve it.**

Z is embedding of a sample x. Y is embedding of class attribute

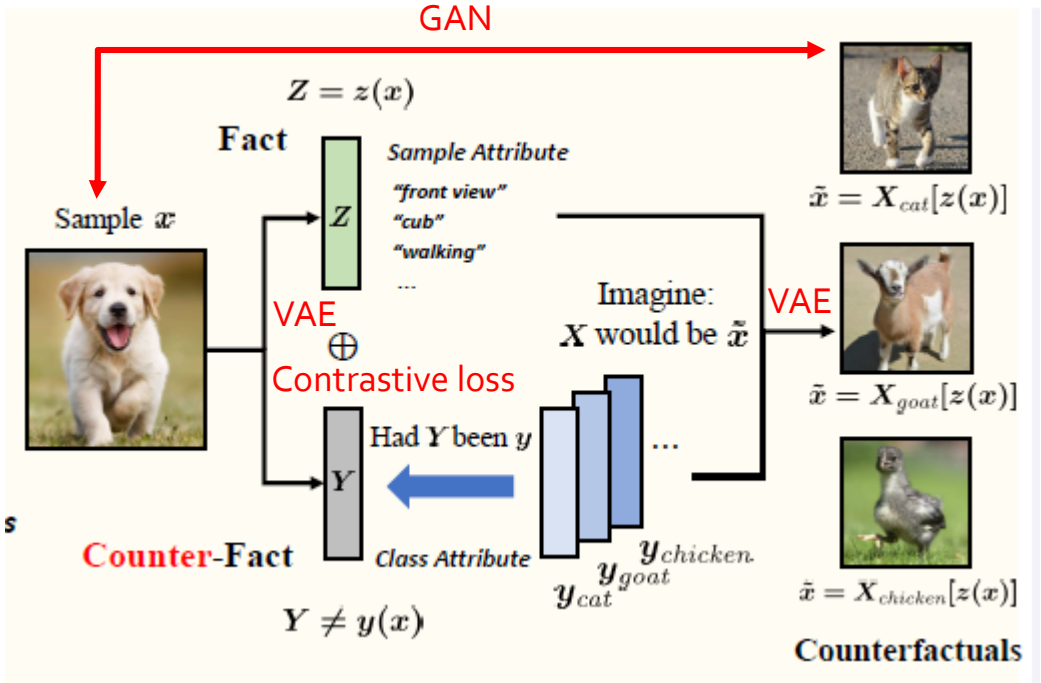
Claim: A counterfactual is faithful is faithful if and only if the sample attribute Z and class attribute Y are group disentangled

Disentangle Z from Y

$$\mathcal{L}_Z = -\mathbb{E}_{Q_\phi(Z|X)}\left[P_\theta(X \mid Z, Y)\right] + \beta D_{KL}\left(Q_\phi(Z \mid X) \parallel P(Z)\right)$$

Disentangle Y from Z

$$\mathcal{L}_Y = -log\frac{exp\left(-dist(\mathbf{x}, \mathbf{x_y})\right)}{\sum_{\mathbf{x}' \in \tilde{X} \cup \{\mathbf{x_y}\}} exp\left(-dist(\mathbf{x}, \mathbf{x}')\right)}$$



x is original sample. x' is counterfactual. x_y is the sample with label y. $\tilde{X}$ is the set including counterfactuals with label except y. The overall loss is shown below, where $\mathcal{L}_F$ is the loss of a discriminator which identifies counterfactuals from real samples.

$$\min_{\theta,\phi} \mathcal{L}_Z + \nu\mathcal{L}_Y + \max_{\omega} \rho\mathcal{L}_F$$

Yue, Z., Wang, T., Sun, Q., Hua, X. S., & Zhang, H. (2021). Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15404-15414).

Selected References

1. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.
2. Mc Grath, R., Costabello, L., Le Van, C., Sweeney, P., Kamiab, F., Shen, Z., & Lecue, F. (2018, December). Interpretable Credit Application Predictions With Counterfactual Explanations. In NIPS 2018-Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy.
3. Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607-617).
4. Russell, C. (2019, January). Efficient search for diverse coherent explanations. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 20-28).
5. Van Looveren, A., & Klaise, J. Interpretable Counterfactual Explanations Guided by Prototypes. Age, 46, 46.
6. Liu, S., Kailkhura, B., Loveland, D., & Han, Y. (2019, November). Generative counterfactual introspection for explainable deep learning. In 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP) (pp. 1-5). IEEE.
7. Van Looveren, A., Klaise, J., Vacanti, G., & Cobb, O. (2021). Conditional Generative Models for Counterfactual Explanations. arXiv preprint arXiv:2101.10123.
8. Olson, Matthew L., et al. "Counterfactual state explanations for reinforcement learning agents via generative deep learning." Artificial Intelligence 295 (2021): 103455.
9. Yue, Z., Wang, T., Sun, Q., Hua, X. S., & Zhang, H. (2021). Counterfactual zero-shot and open-set visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15404-15414).

Evaluation & Metric

Evaluation for Explanations

- Computational Metrics
- Cognitive Metrics

Metrics for Counterfactual

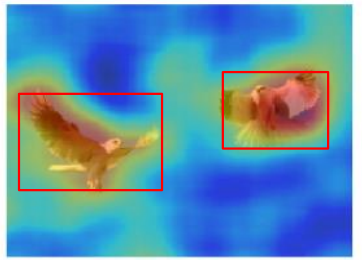
- Computational Metrics
  - Validity
  - Proximity
  - Sparsity
  - Diversity
- Cognitive Metrics
  - Intuitiveness, friendliness & comprehensibility
  - Understandablility

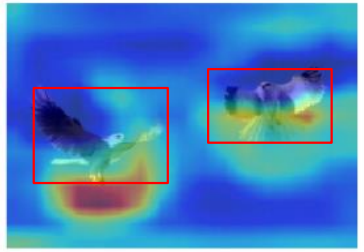
Evaluation for Explanations: Computational Metrics

Evaluations through Computational Metrics

Computational Metrics for Saliency Methods

Property (Desideratum)	Description	Metrics
Faithfulness	whether the salient pixels indicate true importance for the model	Insertion AUC
Localization	whether the salient pixels locate the object for explanation	Pointing Game
Sensitivity Check	check whether the saliency method has the desired sensitivity	Class Sensitivity
Stability	whether the saliency method is stable with insignificant variation	Sens_max



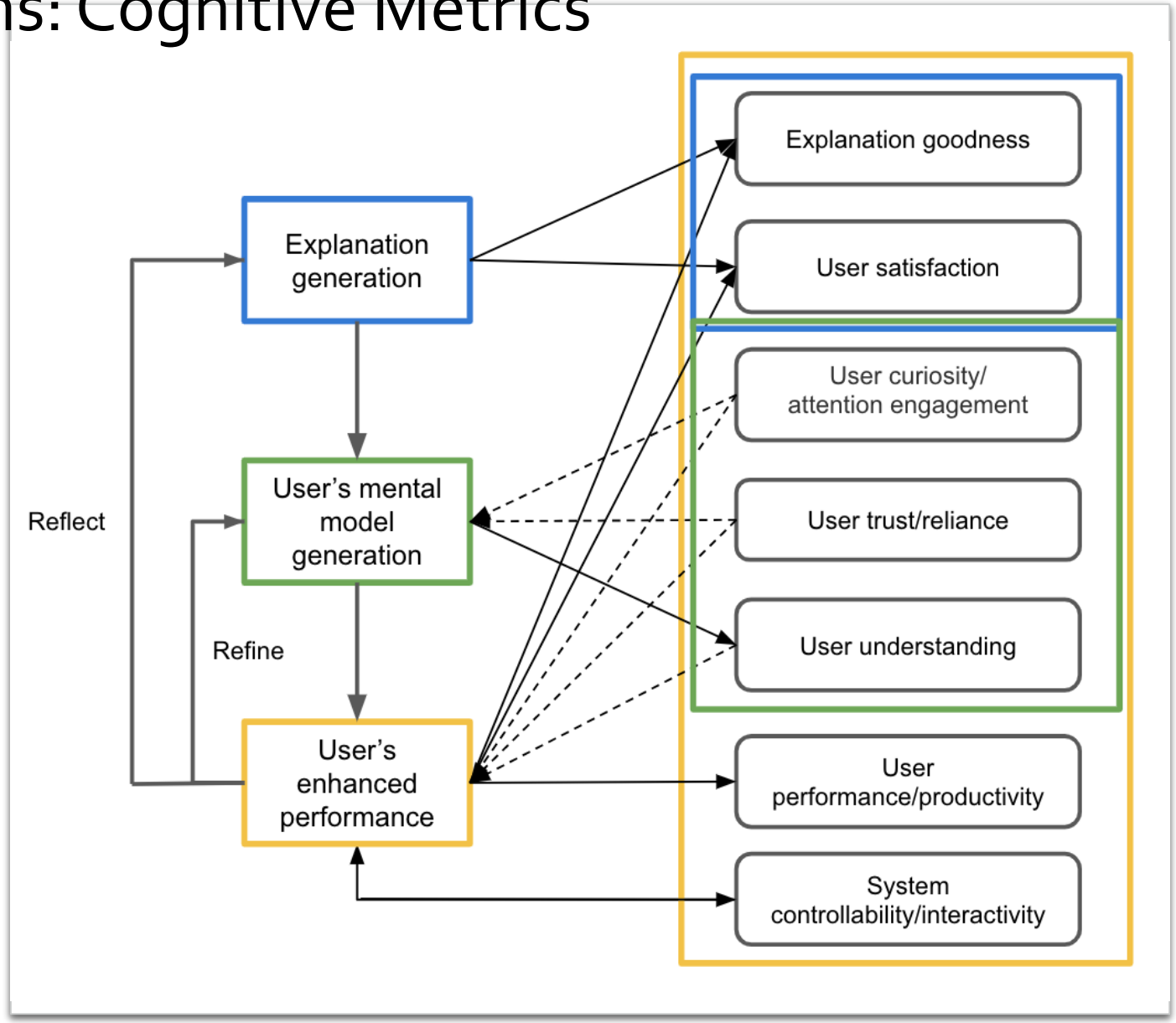


Metrics design varies according to the forms of explanations, the tasks

X. H. Li, Y. Shi, et.al An Experimental Study of Evaluations on Saliency Methods, accepted by SIGKDD 2021 (Presentation at 1:00-2:30pm 8.17)

Evaluation for Explanations: Cognitive Metrics

Hypothesized evaluative model for XAI, including three functional stages and seven associated cognitive metrics. The solid lines represent metrics that can be assessed at the stage, and the dashed lines represent metrics that moderate the stage.



Hsiao, J., Ngai, H. H. T., Qiu, L., Yang, Y., & Cao, C. C. (2021) "Roadmap for designing cognitive metrics for explainable artificial intelligence (xai)." Artificial Intelligence.

Evaluation for Explanations: Cognitive Metrics

The recommended measure types for each cognitive metric

	Subjective: external stimuli	Subjective: internal states	Objective: Cognitive states/processes	Subjective/ Objective: Cognitive models	Subjective/ Objective: Temporal dynamics
Explanation Goodness	xx		xxx	xxx	
User Satisfaction		xx		xxx	xxx
User Curiosity/ Attention Engagement		xx	xxx	xxx	xxx
User Trust/Reliance		xx	xx	xxx	xxx
User Understanding		x	xx	xxx	xxx
User Performance/Productivity			xxx	xxx	xxx
System Controllability/ Interactivity	xx		xxx		

xxx: Best options
xx: Good options
x: Acceptable options

Hsiao, J., Ngai, H. H. T., Qiu, L., Yang, Y., & Cao, C. C. (2021) "Roadmap for designing cognitive metrics for explainable artificial intelligence (xai)." Artificial Intelligence.

Evaluation for Explanations: Cognitive Metrics

Table: Sample questions in comparing two XAIs from the perspectives of different cognitive metrics

Question	Cognitive metric evaluated
Which XAI highlights more accurate details?	Explanation goodness
Which XAI highlights a more appropriate amount of details?	Explanation goodness
Which XAI highlights fewer irrelevant details?	Explanation goodness
Which XAI has a more reasonable size of highlighted region?	Explanation goodness
Which XAI seems more complete?	Explanation goodness
Which XAI are you more satisfied with?	User satisfaction
Which XAI is more reliable and you can count on more?	User trust/reliance
Which XAI provides explanations that make you want to know more about how the AI system works?	User curiosity/attention engagement
Which XAI gives you more confidence in the AI system that it works well?	User trust/reliance
Which XAI facilitates your understanding of AI system more?	User understanding

Some related questionnaires:

- Explanation Goodness Checklist
- Explanation Satisfaction Scale
- Curiosity Checklist
- Scale of Trust
- System Causability Scale

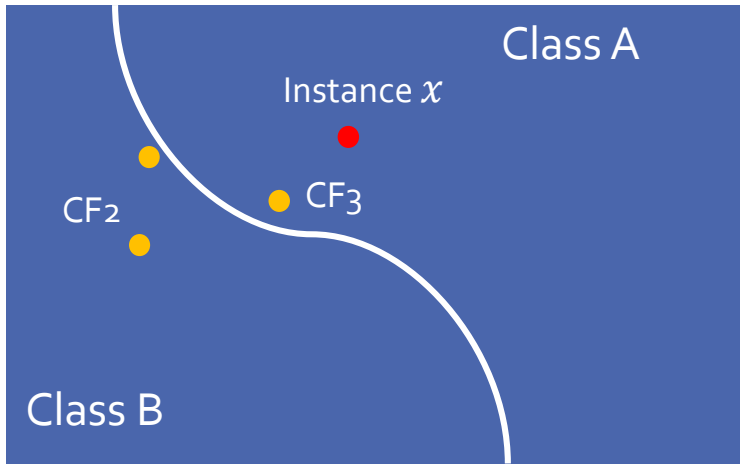
Hsiao, J., Ngai, H. H. T., Qiu, L., Yang, Y., & Cao, C. C. (2021) "Roadmap for designing cognitive metrics for explainable artificial intelligence (xai)." Artificial Intelligence.

Computational Metrics for CF: Validity

- Validity: Validity measures whether the counterfactuals that actually have the desired class label

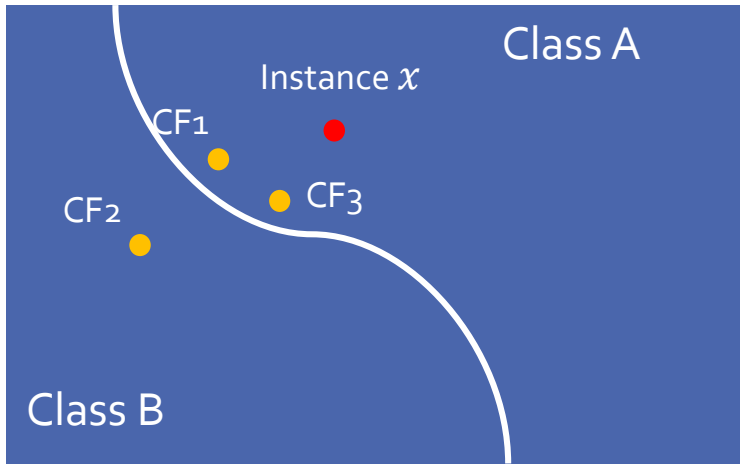
$$Validity: \frac{\sum_i I(f(CF_i') == y_i')}{K}$$

Method 1:



Validity = 2/3

Method 2:



Validity = 1/3

R.K. Mothilal, A. Sharma and C. Tan, Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations, FAT 2020

Computational Metrics for CF: Proximity

- Proximity: Proximity measures the distance of a counterfactual from the input datapoint

CF1 more proximate than CF2

Continuous-Proximity: $\frac{1}{K}\sum_i dist\_cont(CF_i, x)$

Categorical-Proximity: $\frac{1}{K}\sum_i dist\_cat(CF_i, x)$

Common continuous distance metrics:
L1 norm
L2 norm
Mahalanobis distance

R.K. Mothilal, A. Sharma and C. Tan, Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations, FAT 2020

Computational Metrics for CF: Sparsity

- Sparsity: Number of features difference between original input and a CF example, common metric L_0

$$Sparsity\colon\ L_0(x - CF)$$

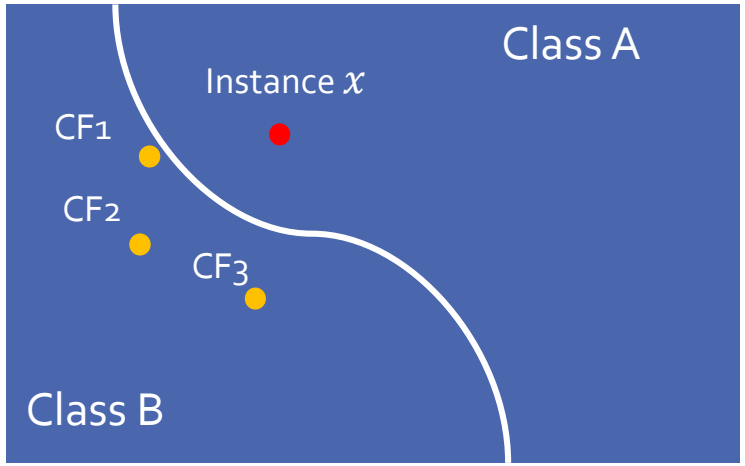
Features	Test Case	"Good" Counterfactual	"Bad" Counterfactual
Weight	80 kg	80 kg	80 kg
Duration	1 hr	***1.5 hrs***	***3 hrs***
Gender	Male	Male	***Female***
Meal	Empty	Empty	***Full***
Units	6	6	***6.5***
Bac Level	Over	Under	Under
		Sparsity= 1	Sparsity= 4

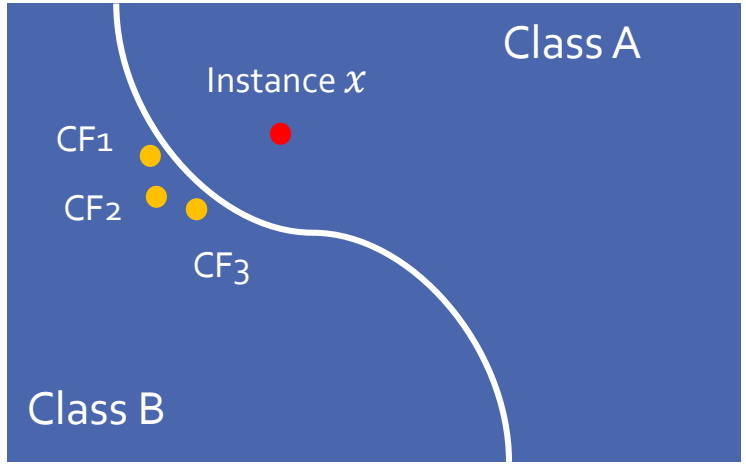
R.K. Mothilal, A. Sharma and C. Tan, Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations, FAT 2020

Computational Metrics for CF: Diversity

- Diversity: Diversity measures feature-wise distances between different generated counterfactuals for one data sample

$$\text{Diversity: } \frac{1}{C_k^2} \sum_i \sum_j dist(CF_i, CF_j)$$





R.K. Mothilal, A. Sharma and C. Tan, Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations, FAT 2020

Cognitive Metrics for CF: Intuitiveness, friendliness & comprehensibility

- Directly design questionnaires on Intuitiveness, friendliness & comprehensibility

Feature	Bare_nuclei	MarAdh	CluThic	mitoses	CelSizUni	CelShaUni	NorNuc	SinEpCeSi	**Model Prediction**
Value	10.0	5.0	8.0	1.0	8.0	5.0	3.0	2.0	**Malignant**

Explanation

*"Had **bare_nuclei** been 3.0 point lower and **CluThic** been 7.0 point lower, the patient would have been diagnosed as Benign rather than Malignant"*

Q1:Given a scale from 1 to 10, "how intuitive and friendly is the explanation to you?" (1 is least preferable, 10 is most preferable)

0

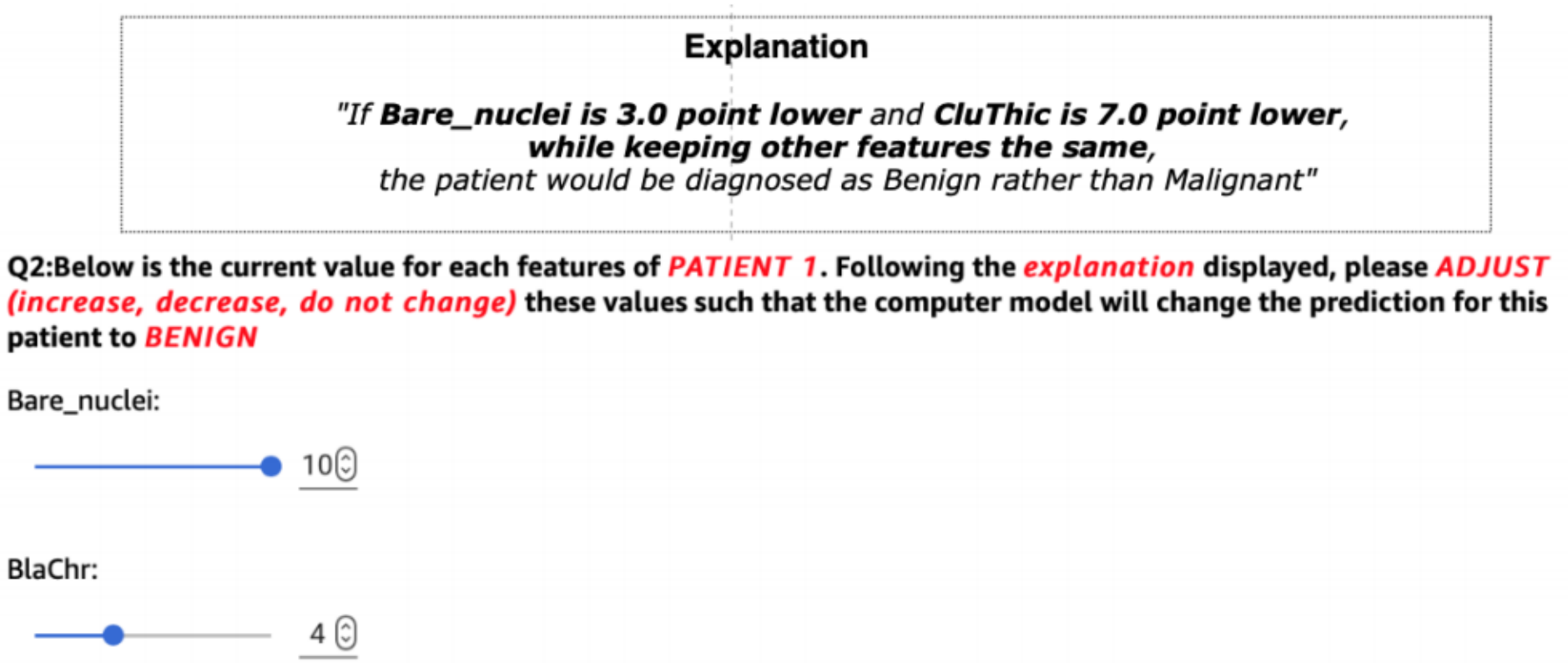
Q2:Given a scale from 1 to 10, "how understandable is the explanation to you?" (1 is least preferable, 10 is most preferable)

0

T. Le, S. Wang and D. Lee, GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction, SIGKDD, 2020

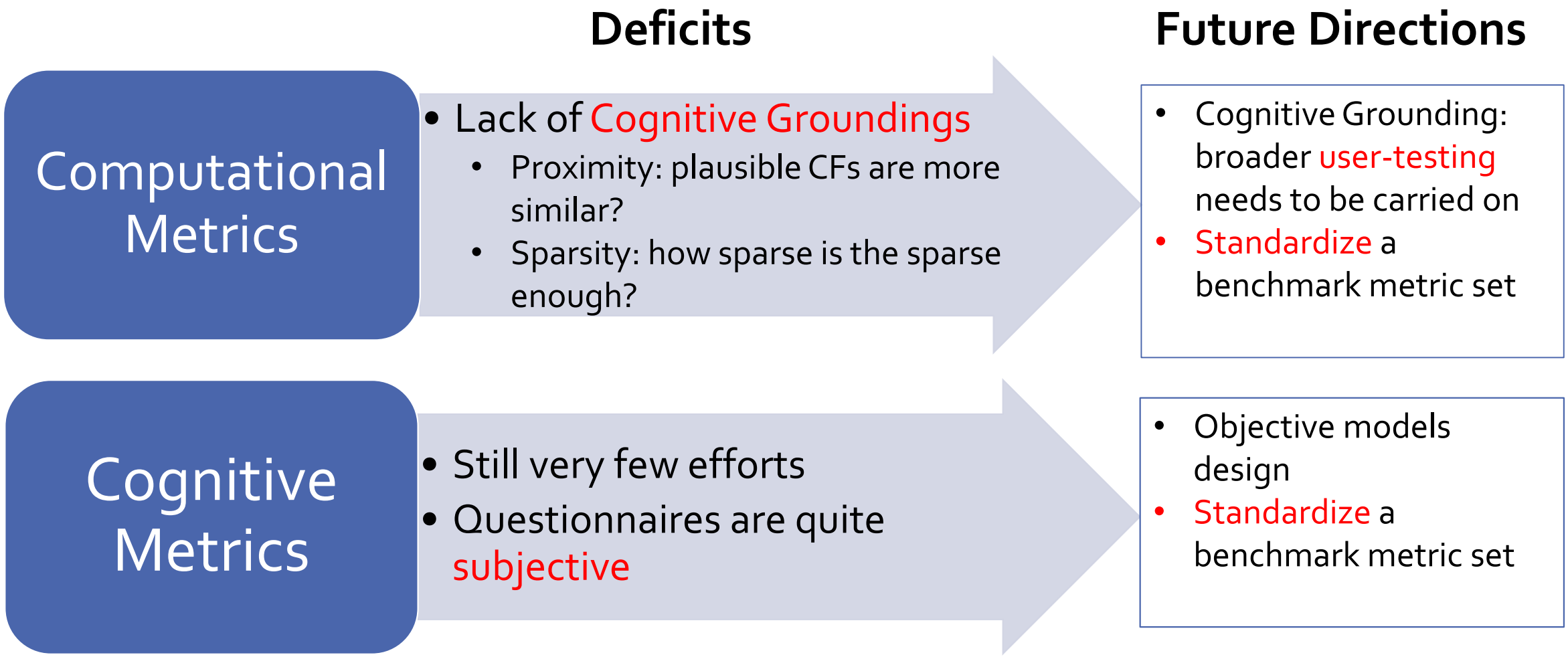
Cognitive Metrics for CF: Understandability

- Test on the understandability of users on the explanation



T. Le, S. Wang and D. Lee, GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction, SIGKDD, 2020

Deficits and Future Research Directions



M. T. Keane, E. M. Kenny, E. Delaney and B. Smyth, If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques, IJCAI 2021

Selected References

1. R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for Explainable AI: Challenges and Prospects, arXiv: 1912.04608
2. R.K. Mothilal, A. Sharma and C. Tan, Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations, FAT 2020
3. T. Le, S. Wang and D. Lee, GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction, SIGKDD, 2020
4. M. T. Keane, E. M. Kenny, E. Delaney and B. Smyth, If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques, IJCAI 2021
5. Hsiao, J., Ngai, H.H., Qiu, L., Yang, Y., & Cao, C.C. (2021). Roadmap of Designing Cognitive Metrics for Explainable Artificial Intelligence (XAI). ArXiv, abs/2108.01737.

04

Counterfactual Explanations in Different Areas

Shendi WANG

Cong WANG

Han GAO

Counterfactual Explanations in Different Areas

- **Counterfactual in NLP**
- **Counterfactual in RS**
- **Counterfactual in CV**
- **Counterfactual in GNN**

Counterfactual in NLP

- Arbitrary
- Specified

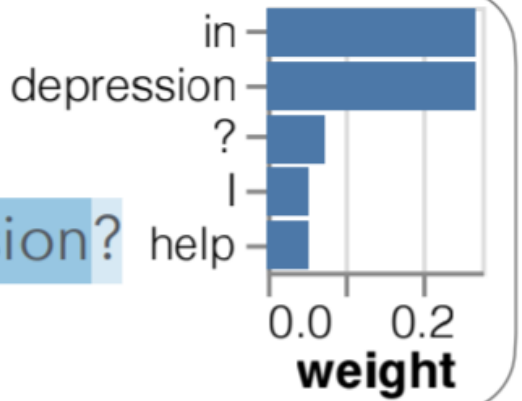
Counterfactual Explanations VS. Feature Importance for NLP

Q1: How can I help a friend experiencing serious depression?

A

Q2: How do I help a friend who is in depression?

Predict $f(x)$: = *Duplicate* (98.2% confident)



	$\hat{x}$, perturbed Q2	$f(\hat{x})$
B	Q2: How do I help a •woman who is in depression?	≠
C	Q2: How do I help a friend who is •suicidal?	=
D	Q2: How do I •find a friend who is in depression?	=

Wu, T., Ribeiro, M. T., Heer, J., & Weld, D. S. (2021). Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.

Counterfactual Generation for NLP

Original x

Prediction

It is not great for kids.

 y negative

Removing/Inserting

It is ~~not~~ great for kids.

 $y' \neq y$ positive

Replacing

It is not **great** → **bad** for kids.

 $y' \neq y$ positive

Arbitrary Removing Words

How removing certain words from a document can change its predicted class.

Number of words in document = m

1 word	---->	m	candidates
2 words	---->	m . (m-1)	candidates
3 words	---->	m . (m-1) . (m-2)	candidates
:		:	
k words	---->	m . (m-1) ... (m-k+1)	candidates

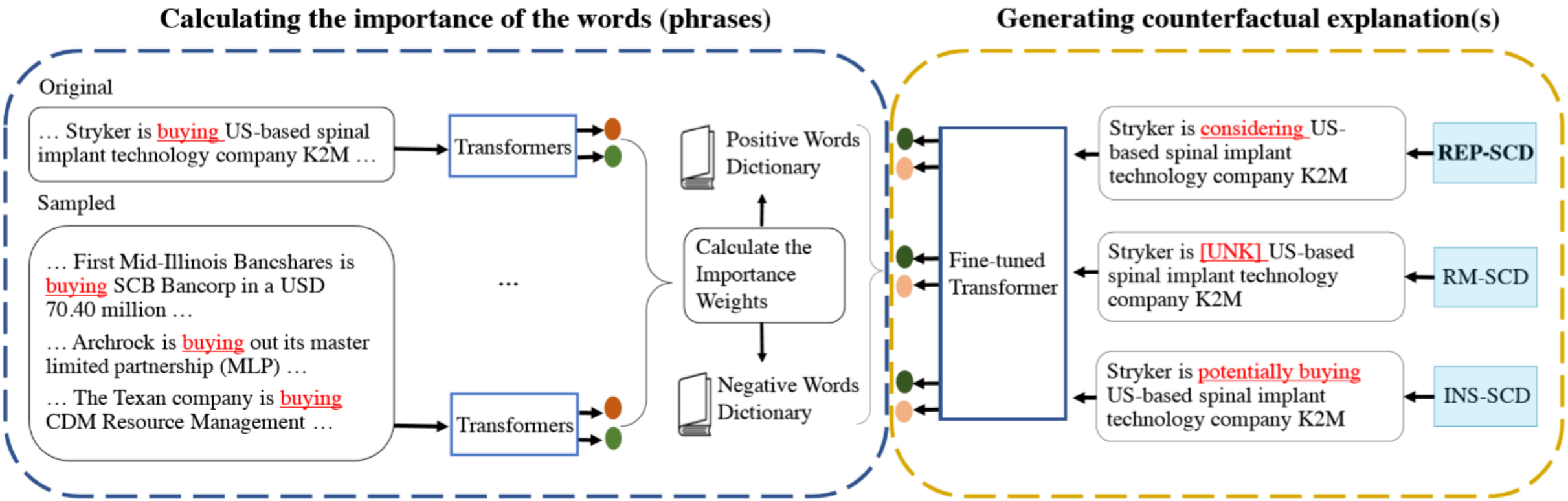
$$= \frac{m!}{(m-k)!}$$

$$= O(m^k)$$

Best-First Search with Pruning

Martens D, Provost F. Explaining data-driven document classifications[J]. 2013.

Specified Perturbations



Yang, L., Kenny, E., Ng, T. L. J., Yang, Y., Smyth, B., & Dong, R. (2020, December). Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6150-6160).

Removing/Inserting Word

INS-SCD: Recasting *fact* as *hoped for*	Ori: Stryker **is buying** US-based spinal implant technology company K2M Group Holdings for USD 1.40 billion in cash Rev: Stryker **is potentially buying** US-based spinal implant technology company K2M Group Holdings for USD 1.40 billion in cash
INS-SCD: Inserting the negative word	Ori: WPP has **confirmed** the recent speculation that it has entered into exclusive negotiations with private equity firm Bain Capital... Rev: WPP has **not confirmed** the recent speculation that it has entered into exclusive negotiations with private equity firm Bain Capital...
RM-SCD: Removing the negative limitation(s)	Ori: This suitor is the Namdar and Washington Prime consortium, the insiders noted, adding that there can be **no certainty** a deal will complete... Rev: This suitor is the Namdar and Washington Prime consortium, the insiders noted, adding that there can be **certainty** a deal will complete...

Yang, L., Kenny, E., Ng, T. L. J., Yang, Y., Smyth, B., & Dong, R. (2020, December). Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6150-6160).

Replacing Word

REP-SCD: Replacing with the certainty word	Ori: Professional vacation services provider ILG is **consider-ing** a merger with Diamond Resorts International... Rev: Professional vacation services provider ILG is **announc-ing** a merger with Diamond Resorts International...
REP-SCD: Changing the deal value	Ori: Vivendi is in early discussions to sell a 10.0 per cent stake in Universal Music Group (UMG) to Tencent for roughly EUR 3.00 **billion**... Rev: Vivendi is in early discussions to sell a 10.0 per cent stake in Universal Music Group (UMG) to Tencent for roughly EUR 3.00 **million**

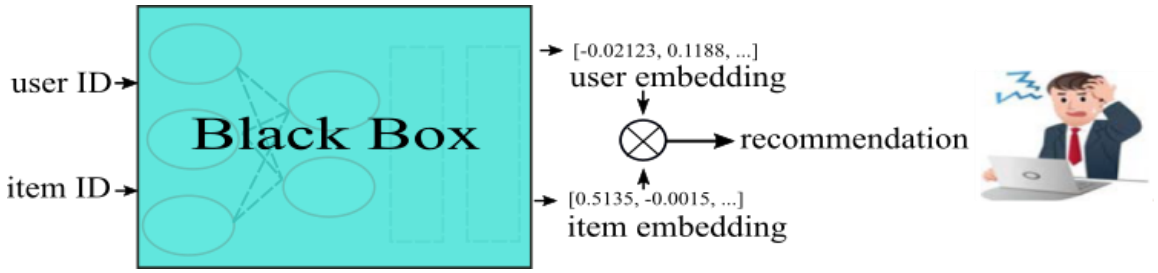
Yang, L., Kenny, E., Ng, T. L. J., Yang, Y., Smyth, B., & Dong, R. (2020, December). Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6150-6160).

Selected References

1. Martens D, Provost F. Explaining data-driven document classifications[J]. 2013.
2. Yang, L., Kenny, E., Ng, T. L. J., Yang, Y., Smyth, B., & Dong, R. (2020, December). Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 6150-6160).
3. Wu, T., Ribeiro, M. T., Heer, J., & Weld, D. S. (2021). Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.

Counterfactual in RS

Background



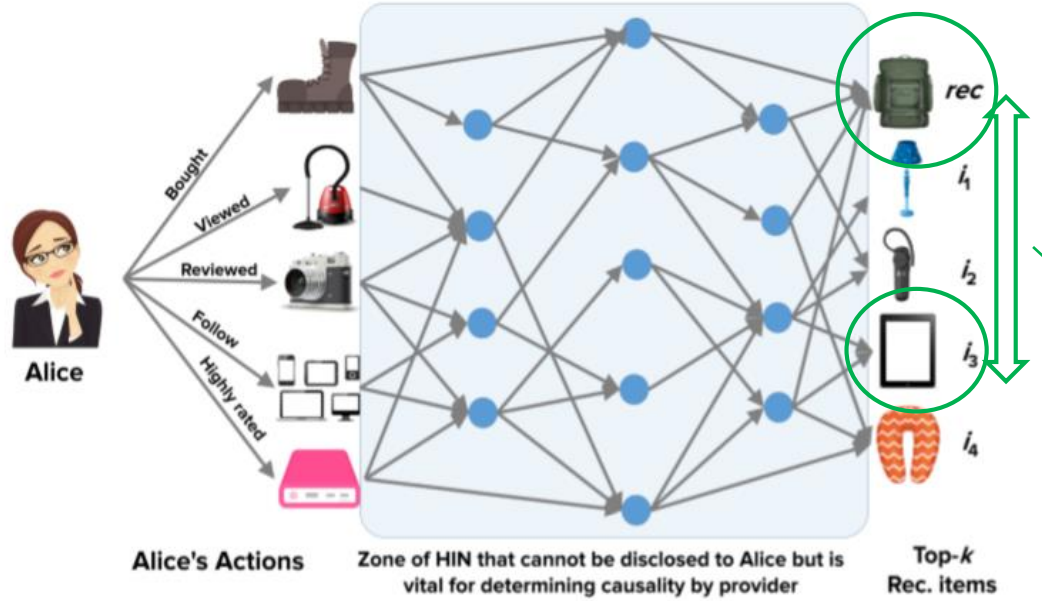
Unexplainable

- Lack of trust
- Lack of recognition
- Reduced Satisfaction

Explanations for recommenders forms:

- How systems compute
- Post-hoc rationalizations
- **End-users: likes/dislikes/ratings or demographic factors**

PRINCE



Alice: Why did I receive this recommendation "Jack Wolfskin backpack"?

PRINCE: You **bought** "Adidas Hiking Shoes";
You **reviewed** "Nikon Coolpix Camera" with "Sleek! Handy on hikes!";
You **rated** "Intenso Travel Power Bank" highly.

If you **had not** done these actions:
"iPad Air" **would have replaced** "Jack Wolfskin backpack".

Figure 1: PRINCE generates explanations as a minimal set of actions using counterfactual evidence on user-specific HINs.

identifies a small (and actually minimal) set

↓

removing these actions

↓

replacing the recommended item

replaced

Ghazimatin, A., Balalau, O., Saha Roy, R., & Weikum, G. (2020, January). PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 196-204)

Personalized PageRank (PPR) for Recommenders

PPR score of node v personalized for s.

$$PPR(s, \cdot) = \alpha e_s + (1 - \alpha) PPR(s, \cdot) W$$

a single seed s

teleportation probability α

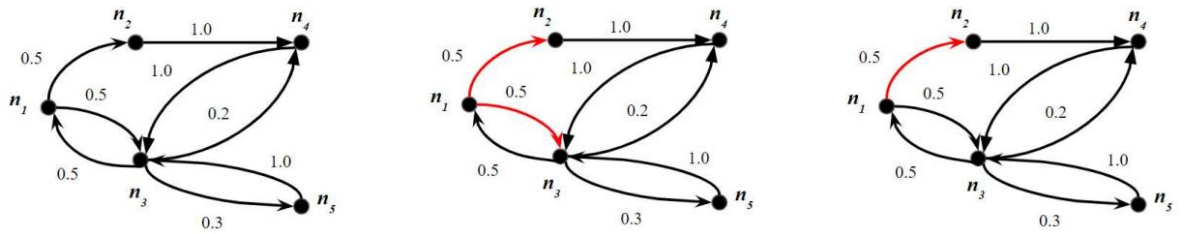
one-hot vector e_s

transition matrix W

Top-1 recommendation

$$rec = \underset{i \in I \setminus N_{out}(u)}{\arg\max} \; PPR(u, i)$$

swap orders of rec and i



(a) $PPR(n_1, n_4) = 0.160$
$PPR(n_1, n_5) = 0.085$
$PPR(n_1, n_4) > PPR(n_1, n_5)$

(b) $A = \{(n_1, n_2), (n_1, n_3)\}$
$W(n_1, n_2)[PPR(n_2, n_4|A) - PPR(n_2, n_5|A)] = 0.095$
$W(n_1, n_3)[PPR(n_3, n_4|A) - PPR(n_3, n_5|A)] = -0.022$

(c) $A^* = \{(n_1, n_2)\}$
$PPR(n_1, n_4|A^*) = 0.078$
$PPR(n_1, n_5|A^*) = 0.110$
$PPR(n_1, n_5|A^*) > PPR(n_1, n_4|A^*)$

Figure 2: Toy Example. (a) A weighted and directed graph where the *PPR* scores are personalized for node n_1. Node n_4 has higher *PPR* than n_5. (b) Scores in a graph configuration where outgoing edges (n_1, n_2), and (n_1, n_3) are removed (marked in red). (c) Removing (n_1, n_2) causes n_5 to outrank n_4.

Ghazimatin, A., Balalau, O., Saha Roy, R., & Weikum, G. (2020, January). PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 196-204)

User Study:
PRINCE VS. Highest Contributions (HC) & Shortest Paths (SP)

Method	Explanation for "Baby stroller" with category "Baby" [Amazon]
PRINCE	**Action 1:** You rated highly "Badger Basket Storage Cubby" with category "Baby" **Replacement Item:** "Google Chromecast HDMI Streaming Media Player" with categories "Home Entertainment"
HC	**Action 1:** You rated highly "Men's hair paste" with category "Beauty" **Action 2:** You reviewed "Men's hair paste" with category "Beauty" with text "Good product. Great price." **Action 3:** You rated highly "Badger Basket Storage Cubby" with category "Baby" **Action 4:** You rated highly "Straw bottle" with category "Baby" **Action 5:** You rated highly "3 Sprouts Storage Caddy" with category "Baby" **Replacement Item:** "Bathtub Waste And Overflow Plate" with categories "Home Improvement"
SP	**Action 1:** You rated highly "Men's hair paste" with category "Beauty" **Action 2:** You rated highly "Badger Basket Storage Cubby" with category "Baby" **Action 3:** You rated highly "Straw bottle" with category "Baby" **Action 4:** You rated highly "3 Sprouts Storage Caddy" with category "Baby" **Replacement Item:** "Google Chromecast HDMI Streaming Media Player" with categories "Home Entertainment"

Method	Explanation for "The Multiversity" with categories "Comics, Historical-fiction, Biography, Mystery" [Goodreads]
PRINCE	**Action 1:** You rated highly "Blackest Night" with categories "Comics, Fantasy, Mystery, Thriller" **Action 2:** You rated highly "Green Lantern" with categories "Comics, Fantasy, Children" **Replacement item:** "True Patriot: Heroes of the Great White North" with categories "Comics, Fiction"
HC	**Action 1:** You follow User ID *x* **Action 2:** You rated highly "Blackest Night" with categories "Comics, Fantasy, Mystery, Thriller" **Action 3:** You rated highly "Green Lantern" with categories "Comics, Fantasy, Children" **Replacement item:** "The Lovecraft Anthology: Volume 2" with categories "Comics, Crime, Fiction"
SP	**Action 1:** You follow User ID *x* **Action 2:** You rated highly "Fahrenheit 451" with categories "Fantasy, Young-adult, Fiction" **Action 3:** You rated highly "Darkly Dreaming Dexter (Dexter, #1)" with categories "Mystery, Crime, Fantasy" **And 6 more actions** **Replacement item:** "The Lovecraft Anthology: Volume 2" with categories "Comics, Crime, Fiction"

Table 4: Anecdotal examples of explanations by PRINCE and the counterfactual baselines.

Ghazimatin, A., Balalau, O., Saha Roy, R., & Weikum, G. (2020, January). PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 196-204)

CF Explanations to Improve Recommender Models

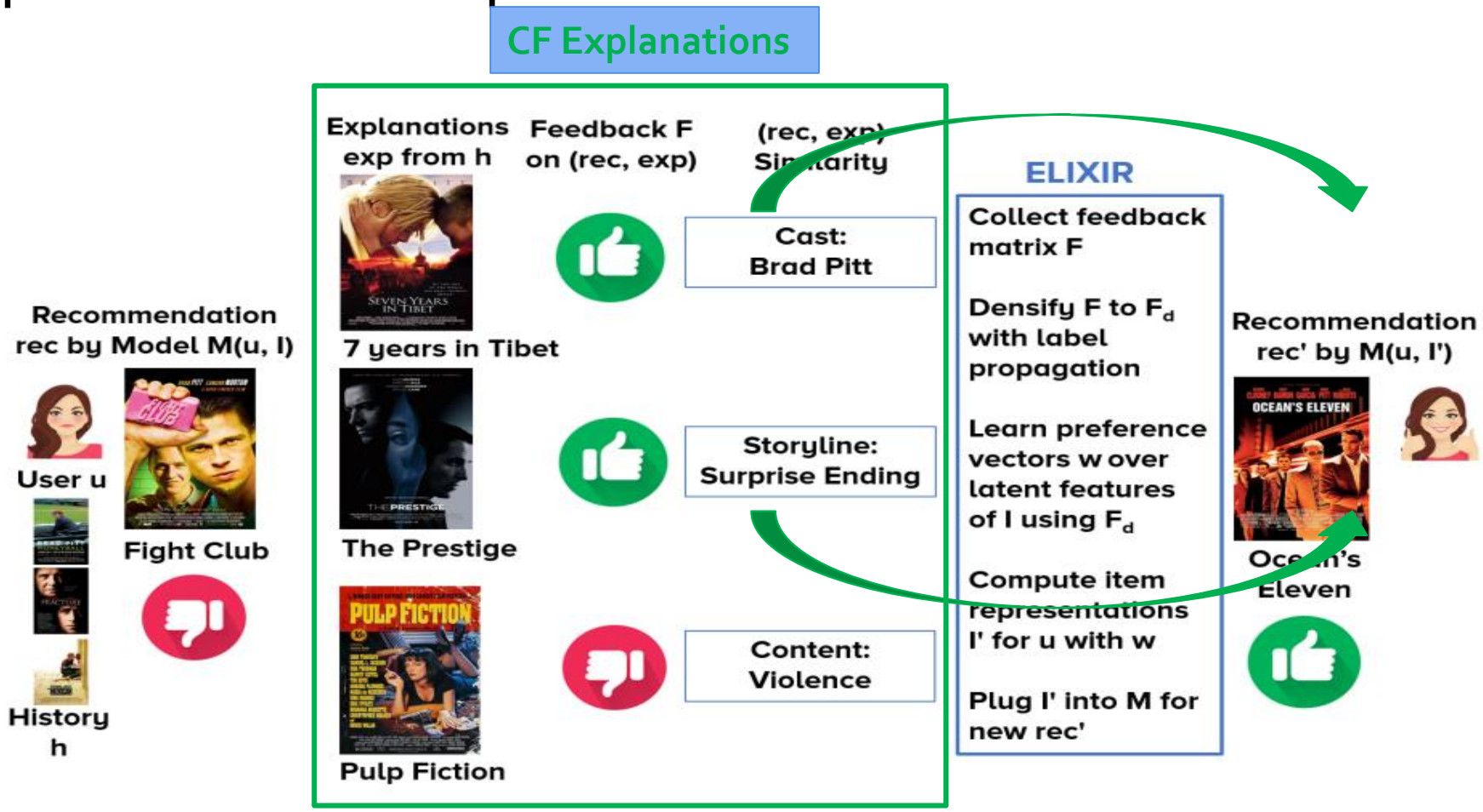


Figure 1: Example illustrating the intuition for ELIXIR.

Ghazimatin, A., Pramanik, S., Saha Roy, R., & Weikum, G. (2021, April). ELIXIR: Learning from User Feedback on Explanations to Improve Recommender Models. In *Proceedings of the Web Conference 2021* (pp. 3850-3860).

Strategies for Candidate Counterfactuals

Random Search (Rnd): randomly considers cardinality and the selected items.

Exhaustive Search (Exh): considers candidates in increasing cardinality.
e.g. from cardinality 1, 2, and so forth.

Breadth First Search (BFS):

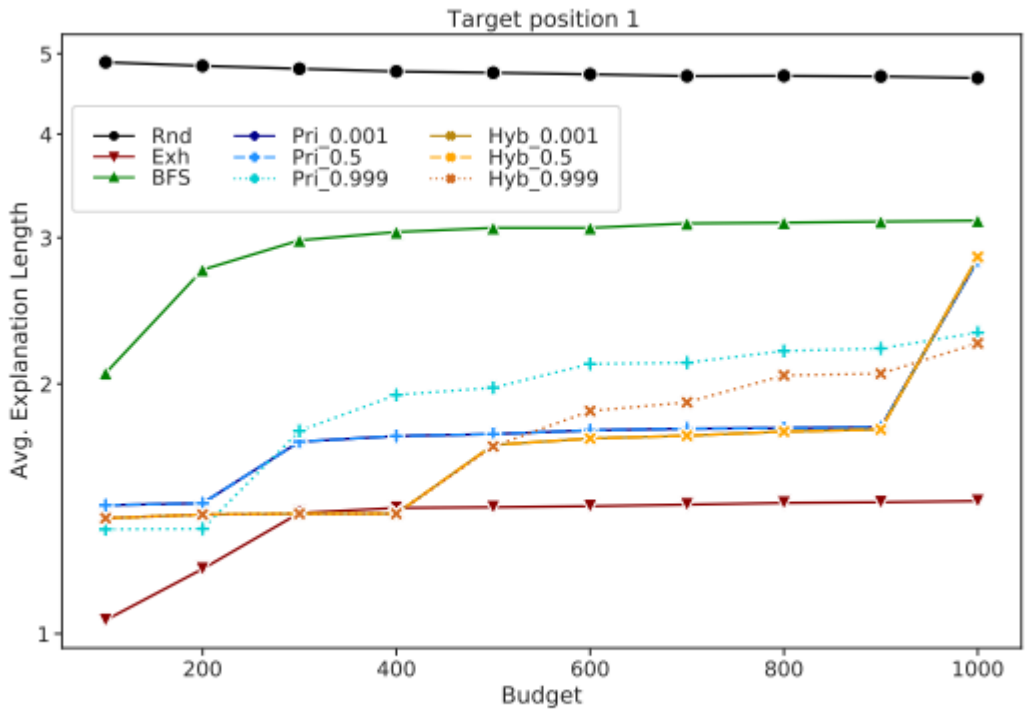
1. tries to quickly identify a candidate
2. refine the explanation

Priority Search (Pri):
Pri invokes the recommender, computes its priority score, and add it.

Hybrid Search (Hyb): hybrid of the exhaustive and the priority search.

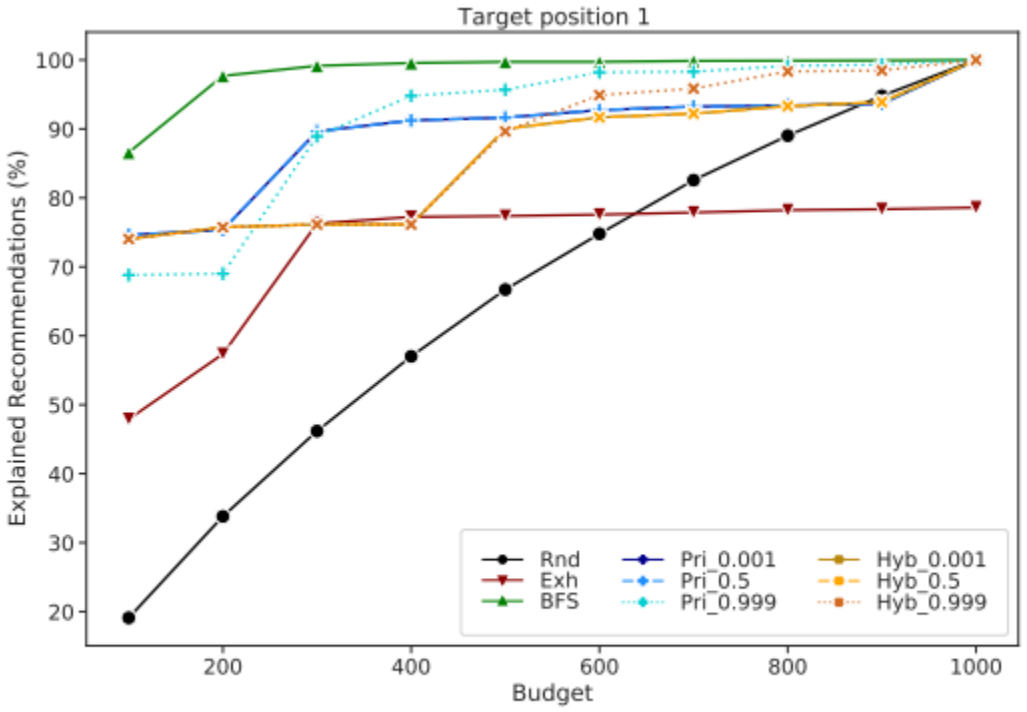
Kaffes, V., Sacharidis, D., & Giannopoulos, G. (2021, June). Model-Agnostic Counterfactual Explanations of Recommendations. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 280-285).

The average **length** of the returned counterfactual explanation



(a) Average computed length of counterfactuals per budget.

The **percentage** of recommendations that were explained



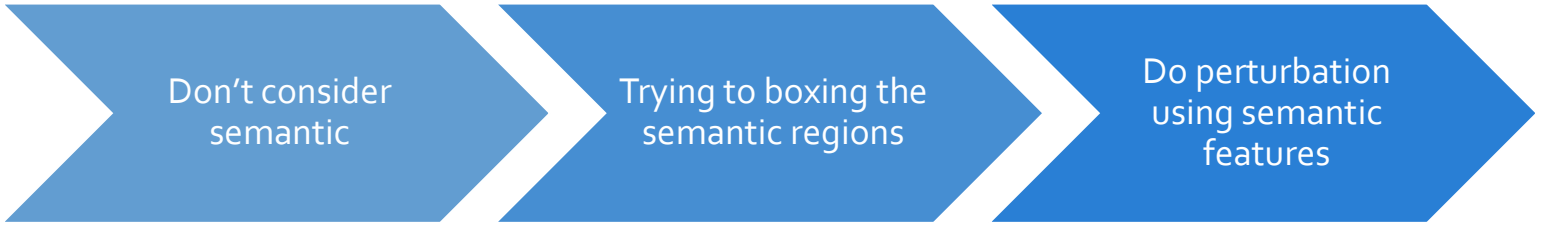
(b) Recommendations successfully explained per budget.

Kaffes, V., Sacharidis, D., & Giannopoulos, G. (2021, June). Model-Agnostic Counterfactual Explanations of Recommendations. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 280-285).

Selected References

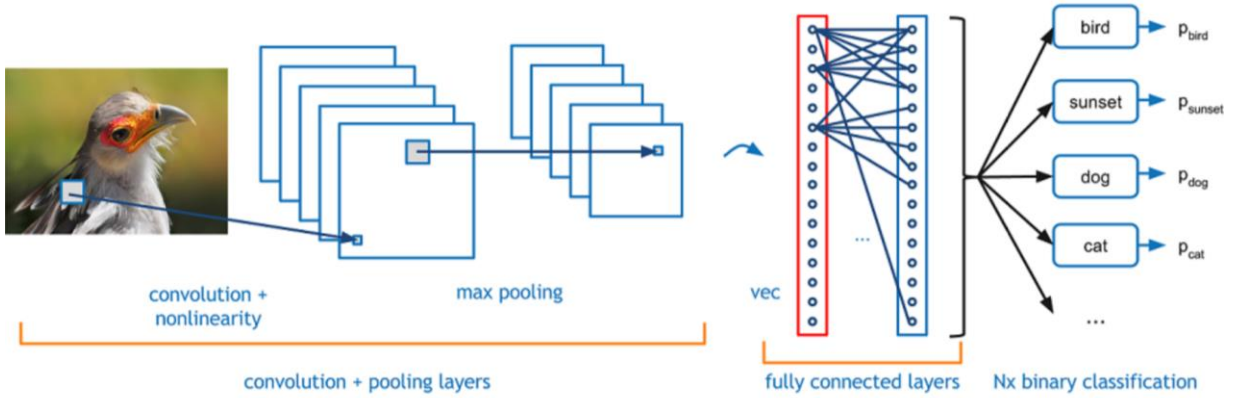
1. Ghazimatin, A., Balalau, O., Saha Roy, R., & Weikum, G. (2020, January). PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 196-204)
2. Ghazimatin, A., Pramanik, S., Saha Roy, R., & Weikum, G. (2021, April). ELIXIR: Learning from User Feedback on Explanations to Improve Recommender Models. In *Proceedings of the Web Conference 2021* (pp. 3850-3860).
3. Kaffes, V., Sacharidis, D., & Giannopoulos, G. (2021, June). Model-Agnostic Counterfactual Explanations of Recommendations. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 280-285).

Counterfactual in CV



Apply Counterfactual to Computer Vision

- Mostly used in fine-grained cases, to avoid the perturbation be the whole object.
- Perturbation in this area could be consider as modify features extracted by network.



A general CNN working flow

- Depending on the extent to which they consider semantic features

Don't consider semantic → Trying to boxing the semantic regions → Do perturbation using semantic features

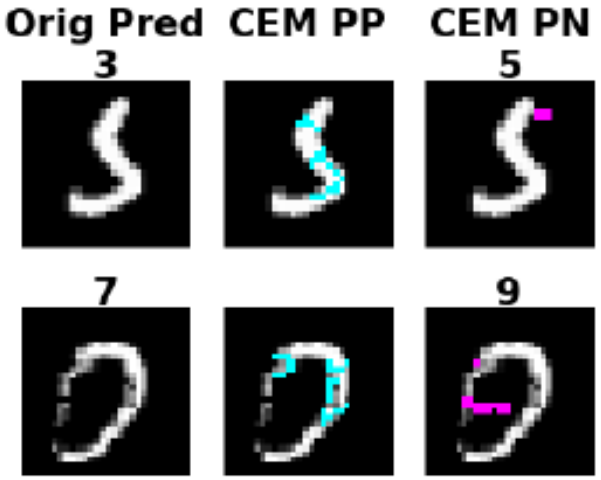
Picture from https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529

Contrastive Explanation Method (CEM)

Contribution

Find what should be minimally and sufficiently present (eg. important object pixels in an image) for a given input to justify its classification and analogously **what should be minimally and necessarily absent** (eg. certain background pixels).

Pertinent Negatives, could be seen as counterfactuals



Dhurandhar, A., Chen, P. Y., Luss, R., Tu, C. C., Ting, P., Shanmugam, K., & Das, P. (2018, December). Explanations based on the missing: towards contrastive explanations with pertinent negatives. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 590-601).

CEM

Solution process: Finding Pertinent Negatives (PN)

Solve:

elastic net regularizer

L2 reconstruction error of x evaluated by the autoencoder.

$$\min_{\delta \in \mathcal{X}/\mathbf{x}_0} \; c \cdot f_\kappa^{\text{neg}}(\mathbf{x}_0, \boldsymbol{\delta}) + \beta\|\boldsymbol{\delta}\|_1 + \|\boldsymbol{\delta}\|_2^2 + \gamma\|\mathbf{x}_0 + \boldsymbol{\delta} - \text{AE}(\mathbf{x}_0 + \boldsymbol{\delta})\|_2^2. \quad (1)$$

First term is:

control separation

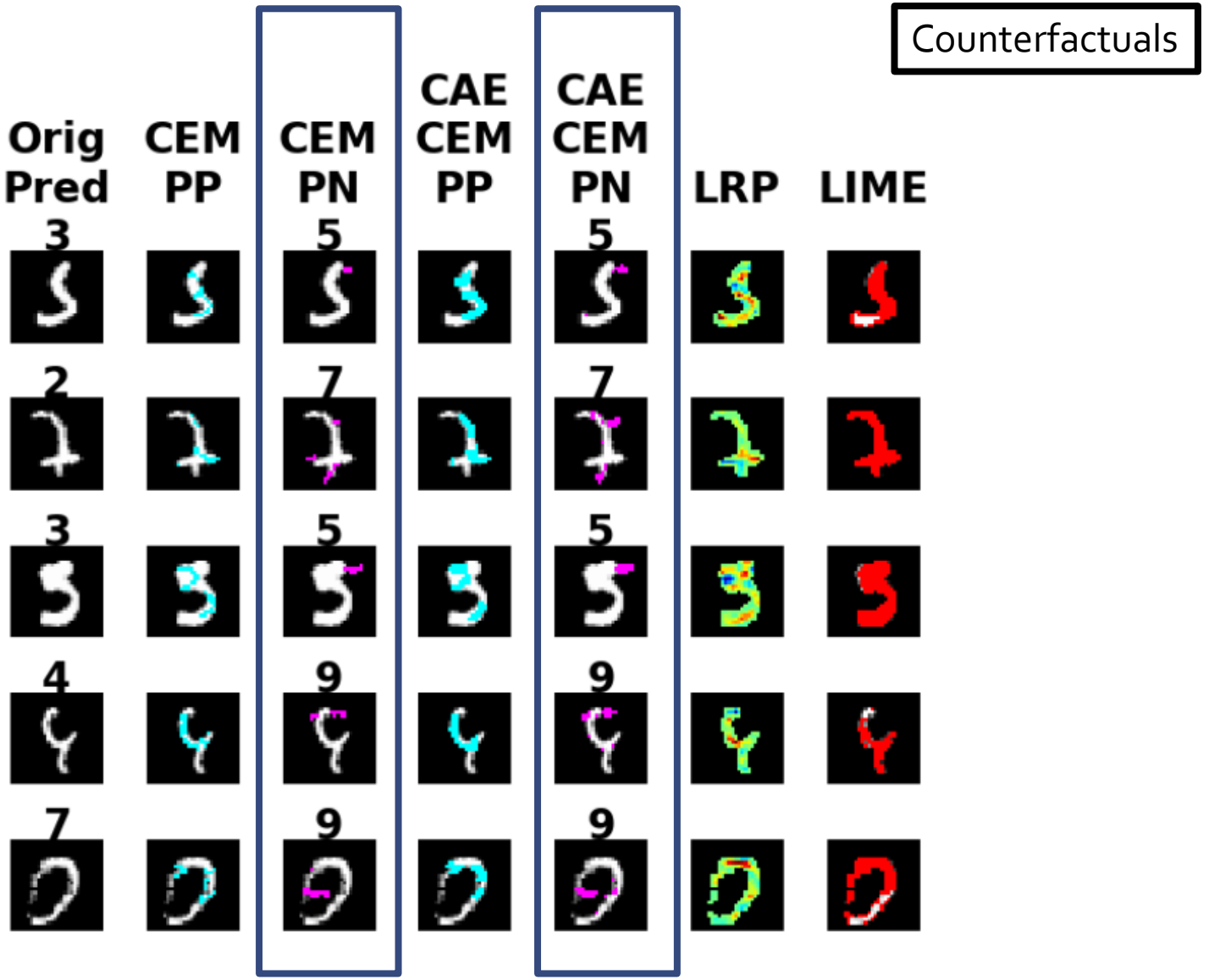
$$f_\kappa^{\text{neg}}(\mathbf{x}_0, \boldsymbol{\delta}) = \max\{[\text{Pred}(\mathbf{x}_0 + \boldsymbol{\delta})]_{t_0} - \max_{i \neq t_0}[\text{Pred}(\mathbf{x}_0 + \boldsymbol{\delta})]_i, -\kappa\} \quad (2)$$

prediction score on original class t_0

aim to find an perturbation δ

Dhurandhar, A., Chen, P. Y., Luss, R., Tu, C. C., Ting, P., Shanmugam, K., & Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *arXiv preprint arXiv:1802.07623*.

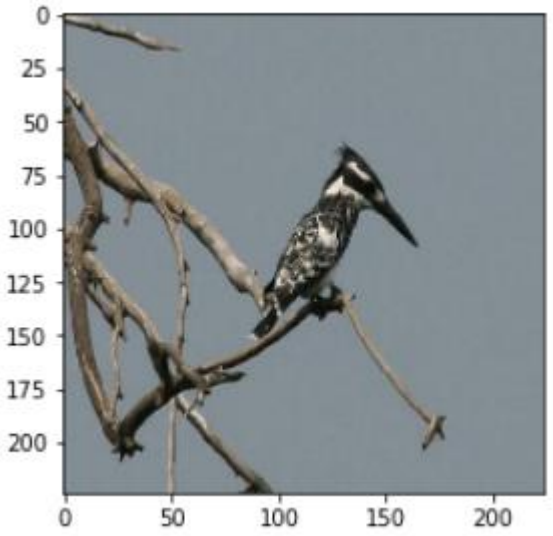
CEM



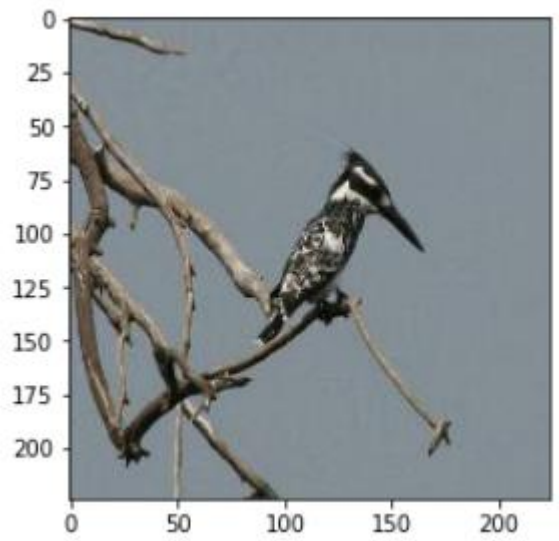
Dhurandhar, A., Chen, P. Y., Luss, R., Tu, C. C., Ting, P., Shanmugam, K., & Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *arXiv preprint arXiv:1802.07623*.

One Problem

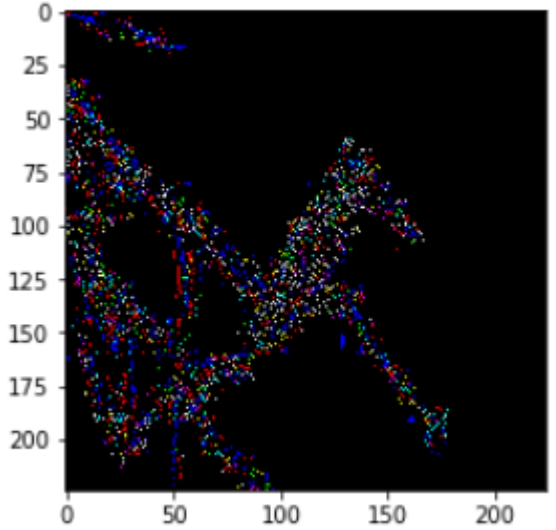
For don’t consider semantic features: **Hard to recognize the difference**



Original Instances
Class: Pied Kingfisher



Counterfactual
Class: Green Kingfisher



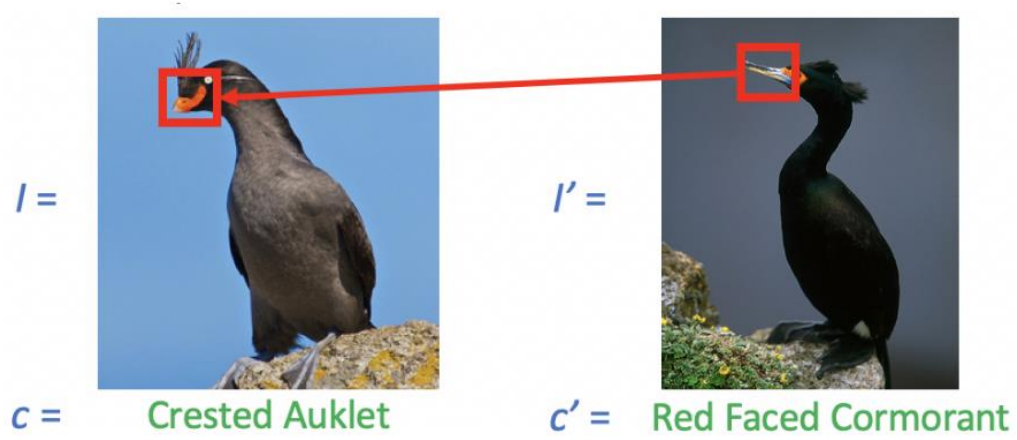
Perturbation (threshold 10)

Using Alibi Code: https://github.com/SeldonIO/alibi

Counterfactual visual explanations (CVE)

Contribution

propose an approach to generate **counterfactual visual explanations,** i.e. what region in the image made the model predict class c instead of class c'

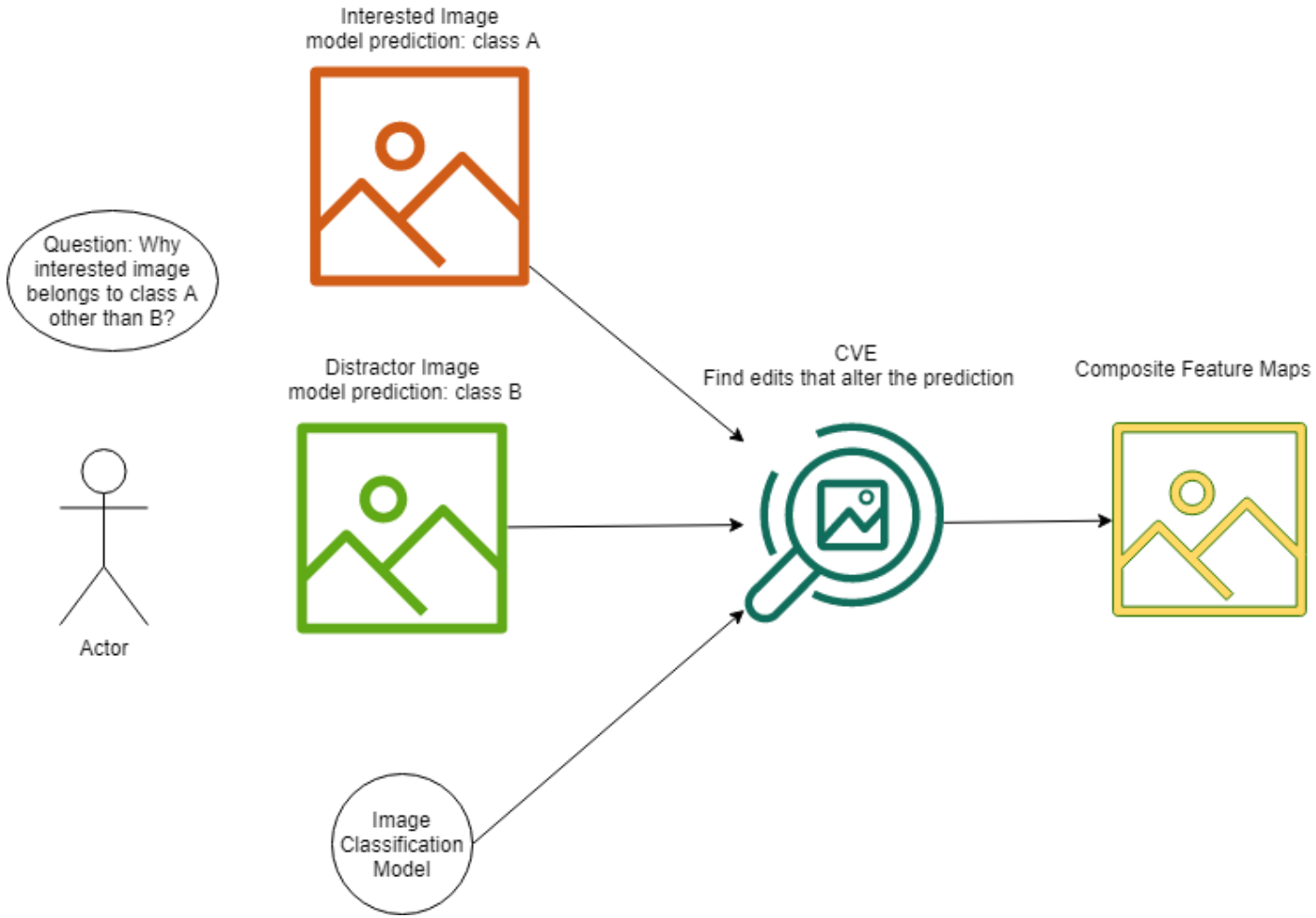


if the highlighted region in the left image looked like the highlighted region in the right image, the resulting image I∗ would be classified more confidently as c'.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019, May). Counterfactual visual explanations. In *International Conference on Machine Learning* (pp. 2376-2384). PMLR.

CVE

Solution process



Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019, May). Counterfactual visual explanations. In *International Conference on Machine Learning* (pp. 2376-2384). PMLR.

CVE

Solution process

Decompose the model to **feature extractor** and **decision network**

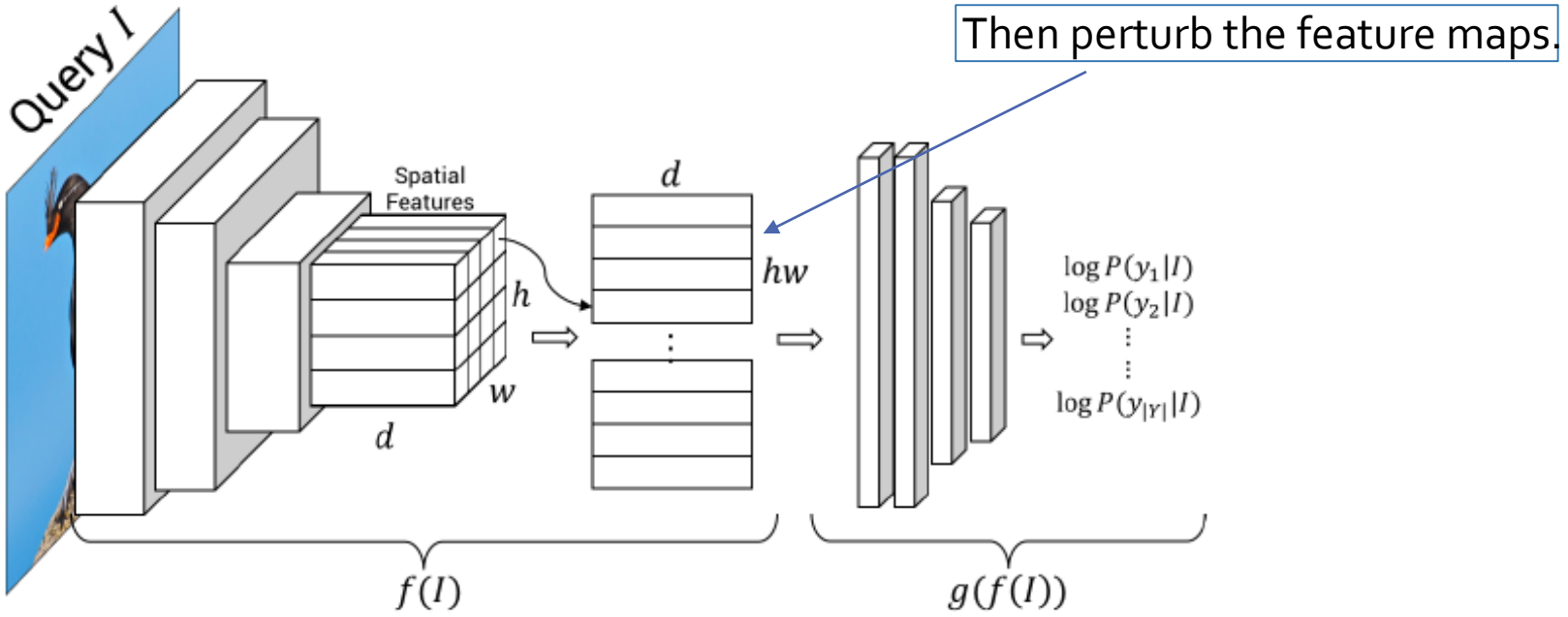


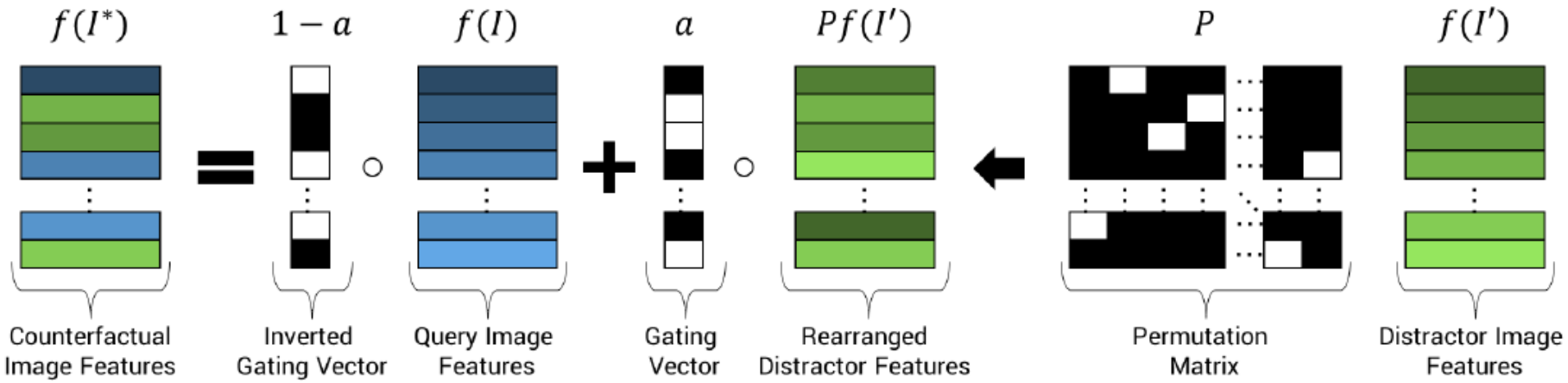
Figure 3. We decompose a CNN as a spatial feature extractor $f(I)$ and a decision network $g(f(I))$ as shown above.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019, May). Counterfactual visual explanations. In *International Conference on Machine Learning* (pp. 2376-2384). PMLR.

CVE

Solution process

Then perturb the feature maps



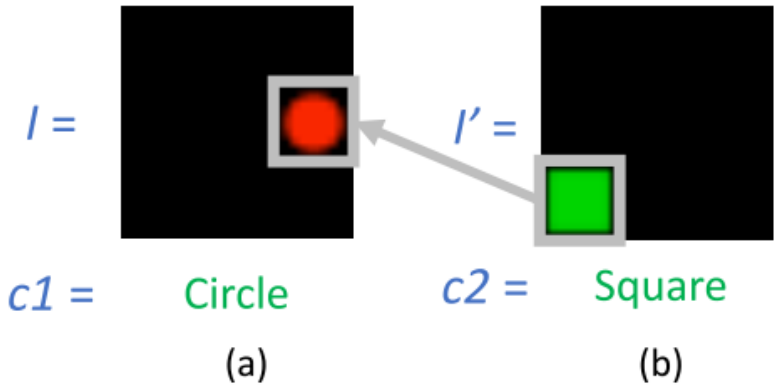
$f(I^*)$ would be the composite feature map.

a: gating vector

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019, May). Counterfactual visual explanations. In *International Conference on Machine Learning* (pp. 2376-2384). PMLR.

CVE

SHAPES



Counterfactual Explanation:
If the middle right cell in image I looked like the bottom left cell in image I?,
the models prediction would have been Square.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019, May). Counterfactual visual explanations. In *International Conference on Machine Learning* (pp. 2376-2384). PMLR.

CVE

For the results, this work present composite images.

MNIST
Qualitative Results

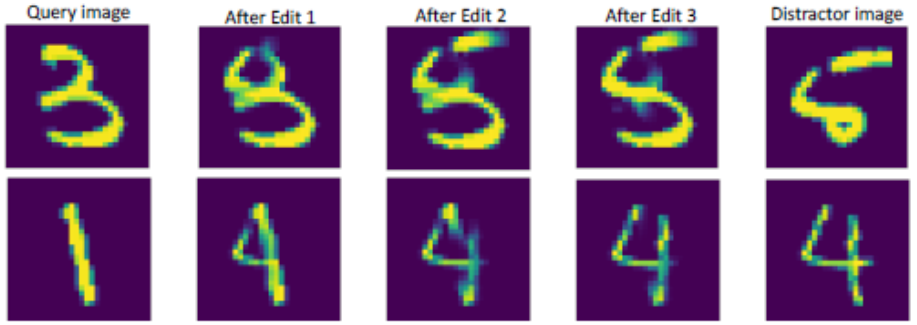


Figure 6. Examples of multiple edits on MNIST digits.

Quantitative Analysis
It takes our approach 2.67 edits to change the model's prediction from c to c'.
It takes 15 s per image on a Titan XP GPU

Caltech-UCSD Birds (CUB)

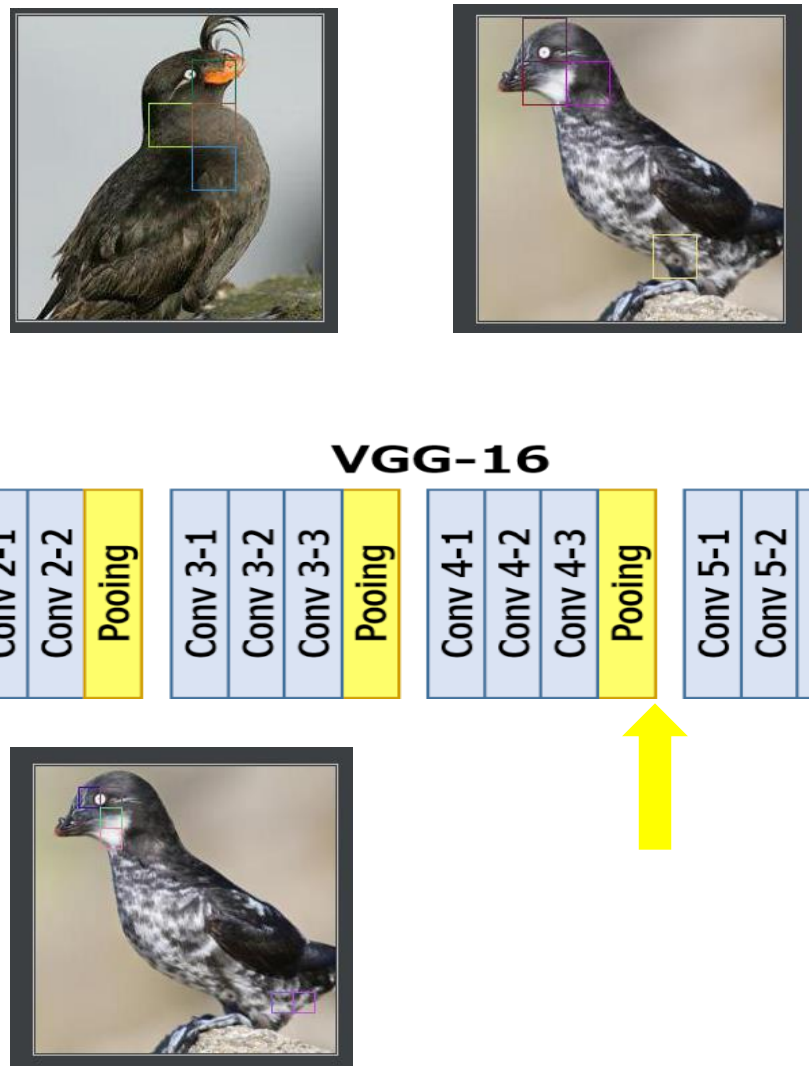
Qualitative Results.

Quantitative Analysis
it takes our approach 7.4 edits to change the model's prediction from c to c'
Runtimes are 1.85 and 1.34 sec/image for random and NN distractor classes respectively on a Titan XP GPU.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019, May). Counterfactual visual explanations. In *International Conference on Machine Learning* (pp. 2376-2384). PMLR.

TDLS - CVE

Boxing more smaller regions..

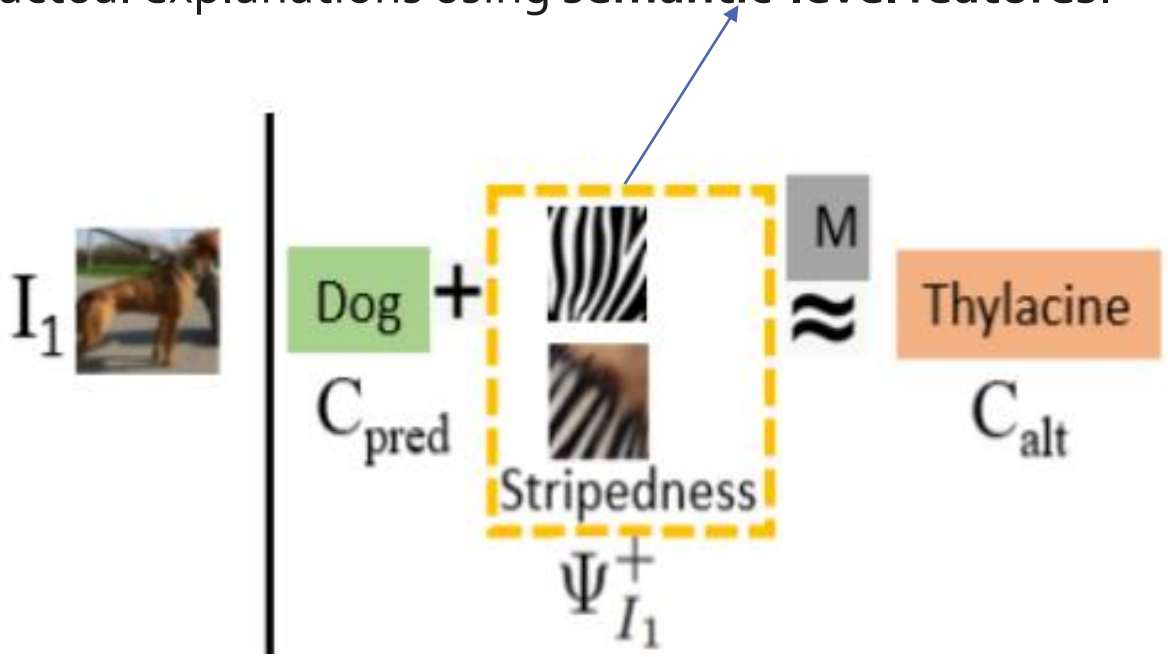


Generating counterfactual explanation using Top-Down Layer Searching algorithm.

Wang, C., Han, H., & Cao, C.C. (2021). TDLS: A Top-Down Layer Searching Algorithm for Generating Counterfactual Visual Explanation.

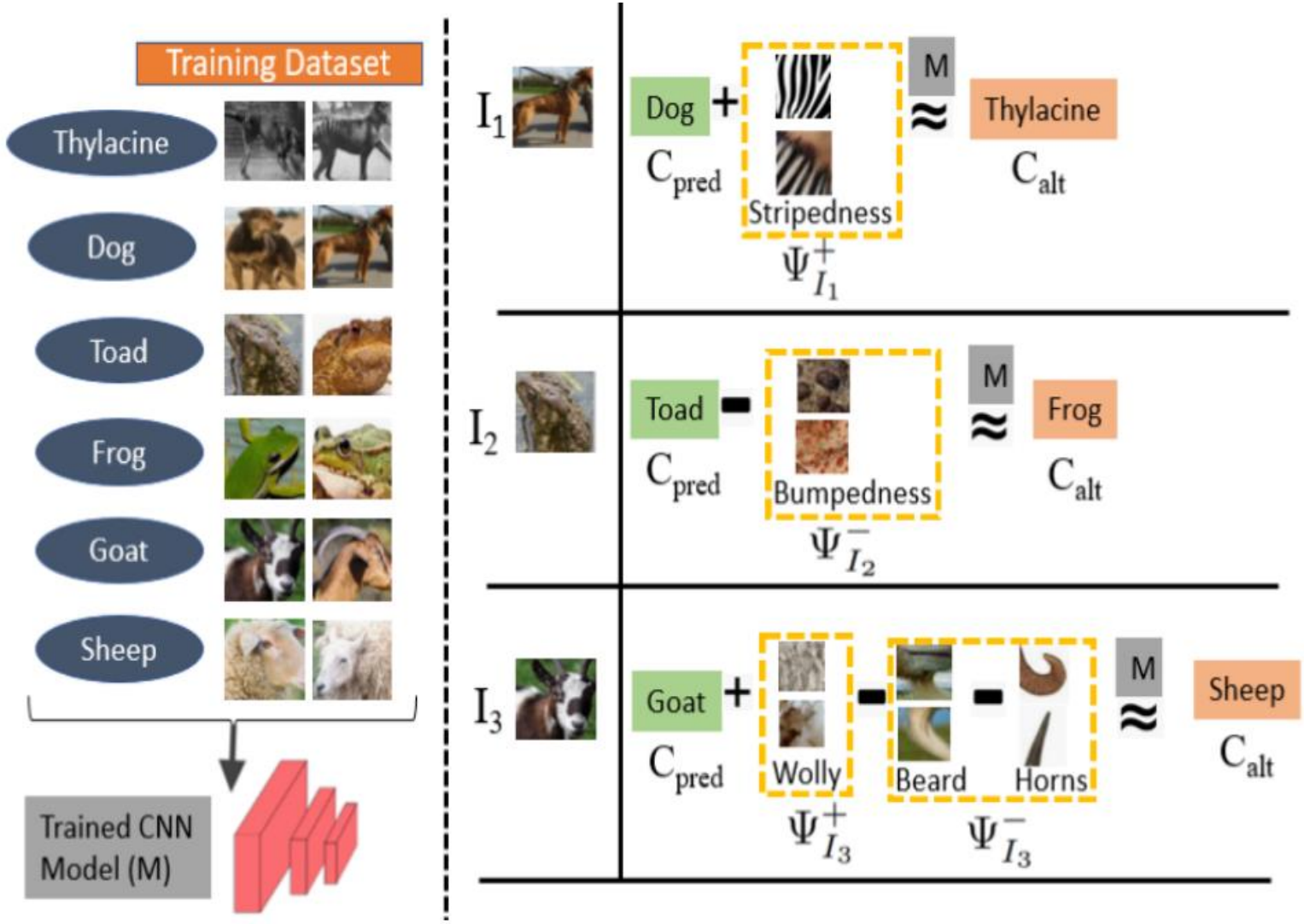
Conceptual and Counterfactual Explanations (CoCoX)

Contribution:
Generate counterfactual explanations using **semantic-level features**.



Akula, A., Wang, S., & Zhu, S. C. (2020, April). Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 03, pp. 2594-2601).

CoCoX



Question:
Why does the machine classify the image I3 as Goat instead of Sheep?

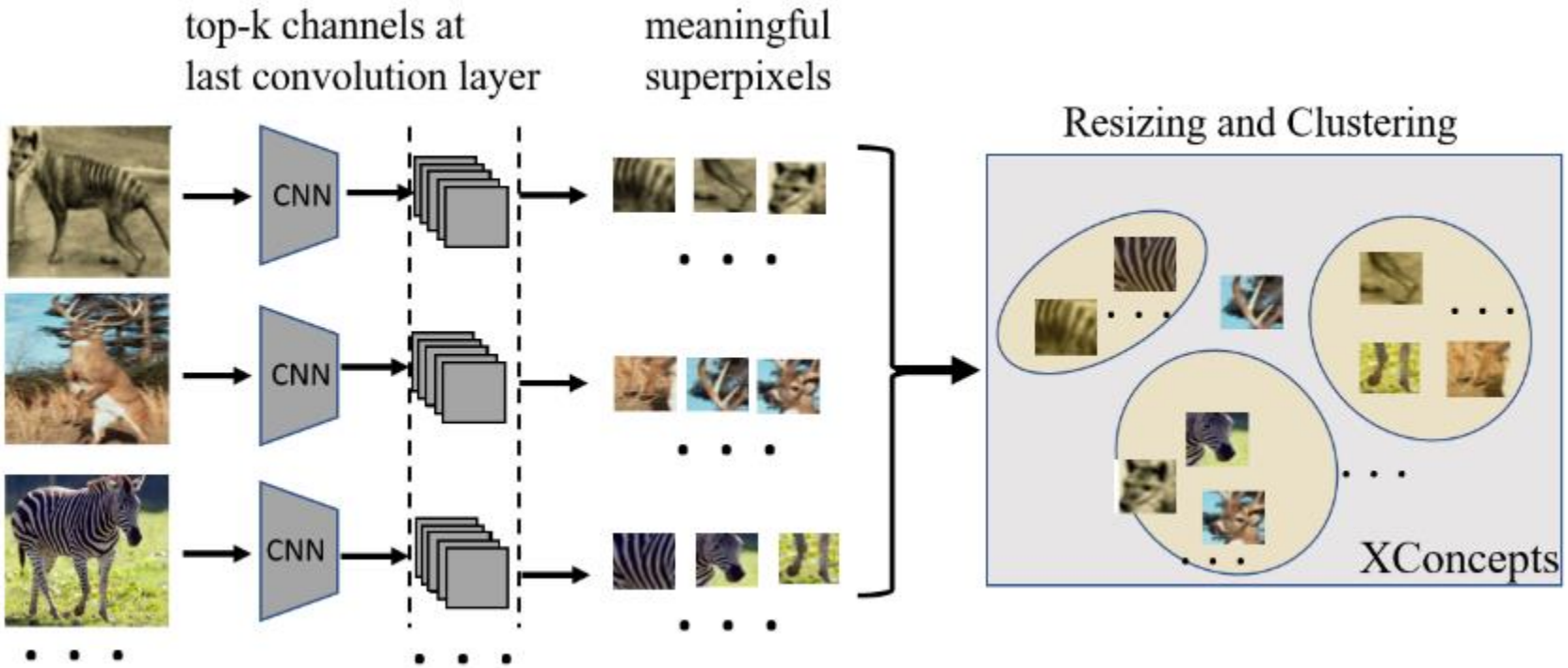
Answer:
Machine thinks the input image is Goat and not Sheep mainly because Sheep's feature woolly is absent in I3 and Goat's features beard and horns are present in I3.

Akula, A., Wang, S., & Zhu, S. C. (2020, April). Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 03, pp. 2594-2601).

CoCoX

Solution Process:

Mining Xconcepts: Generating XConcepts and identifying Class-Specific Xconcepts



Akula, A., Wang, S., & Zhu, S. C. (2020, April). Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 03, pp. 2594-2601).

CoCoX

Using TCAV during identifying class-specific Xconcepts

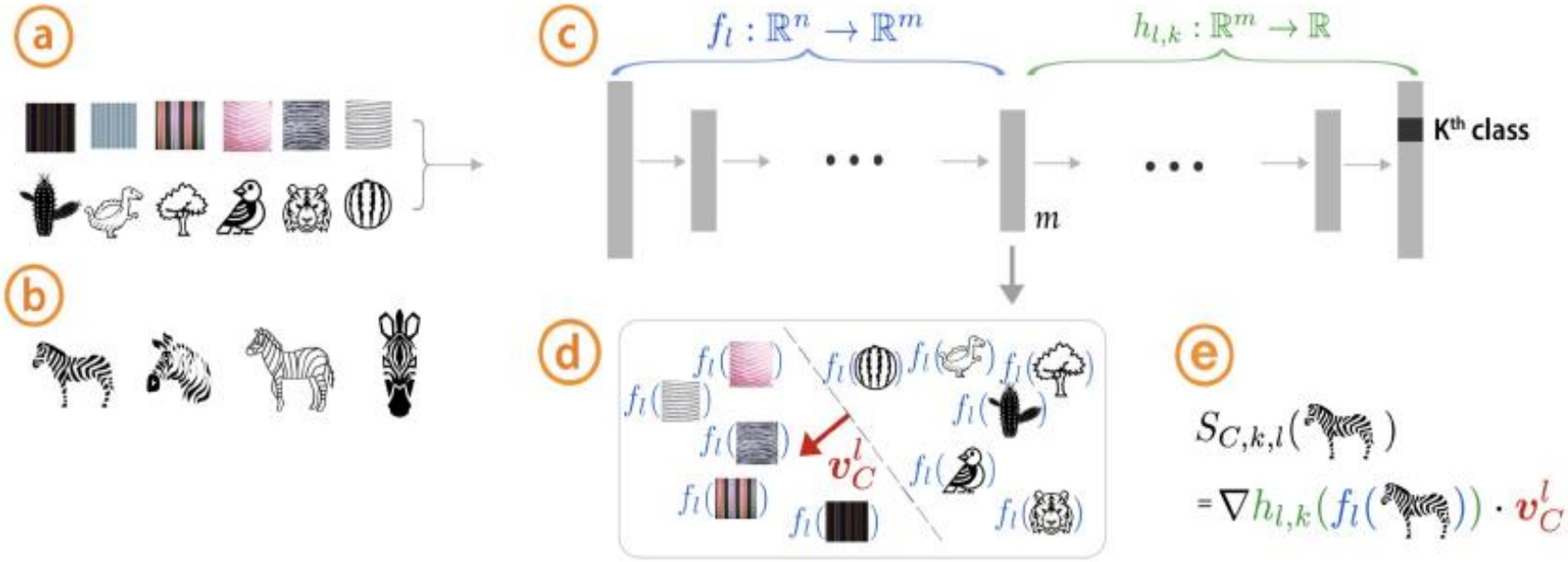


Figure 1. **Testing with Concept Activation Vectors:** Given a user-defined set of examples for a concept (e.g., 'striped'), and random examples ⓐ, labeled training-data examples for the studied class (zebras) ⓑ, and a trained network ⓒ, TCAV can quantify the model's sensitivity to the concept for that class. CAVs are learned by training a linear classifier to distinguish between the activations produced by a concept's examples and examples in any layer ⓓ. The CAV is the vector orthogonal to the classification boundary (v_C^l, red arrow). For the class of interest (zebras), TCAV uses the directional derivative $S_{C,k,l}(x)$ to quantify conceptual sensitivity ⓔ.

Akula, A., Wang, S., & Zhu, S. C. (2020, April). Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 03, pp. 2594-2601).

CoCoX

Finding Pertubation δ

$$\underset{\delta_{pred},\delta_{alt}}{\text{minimize}} \quad \alpha D(\delta_{pred},\delta_{alt}) + \beta \left\|\delta_{pred}\right\|_1 + \lambda \left\|\delta_{alt}\right\|_1 ;$$

get minimal set of xconcepts that alter the prediction.

$$D(\delta_{pred},\delta_{alt}) = max\{g^{pred}(I^{'}) - g^{alt}(I^{'}), -\tau\};$$

τ : Control the difference between the predictions.

$$I^{'} = A^{m,L} \circ v_{pred}^{\top}\delta_{pred} \circ v_{alt}^{\top}\delta_{alt};$$

$$\delta_{pred}^{i} \in \{-1,0\},\ \delta_{alt}^{i} \in \{0,1\}\ \forall i \text{ and } \alpha,\beta,\lambda,\tau \geq 0. \qquad (4)$$

For generate counterfactuals I', don't directly perturb the original image but using Hadamard product between activations.

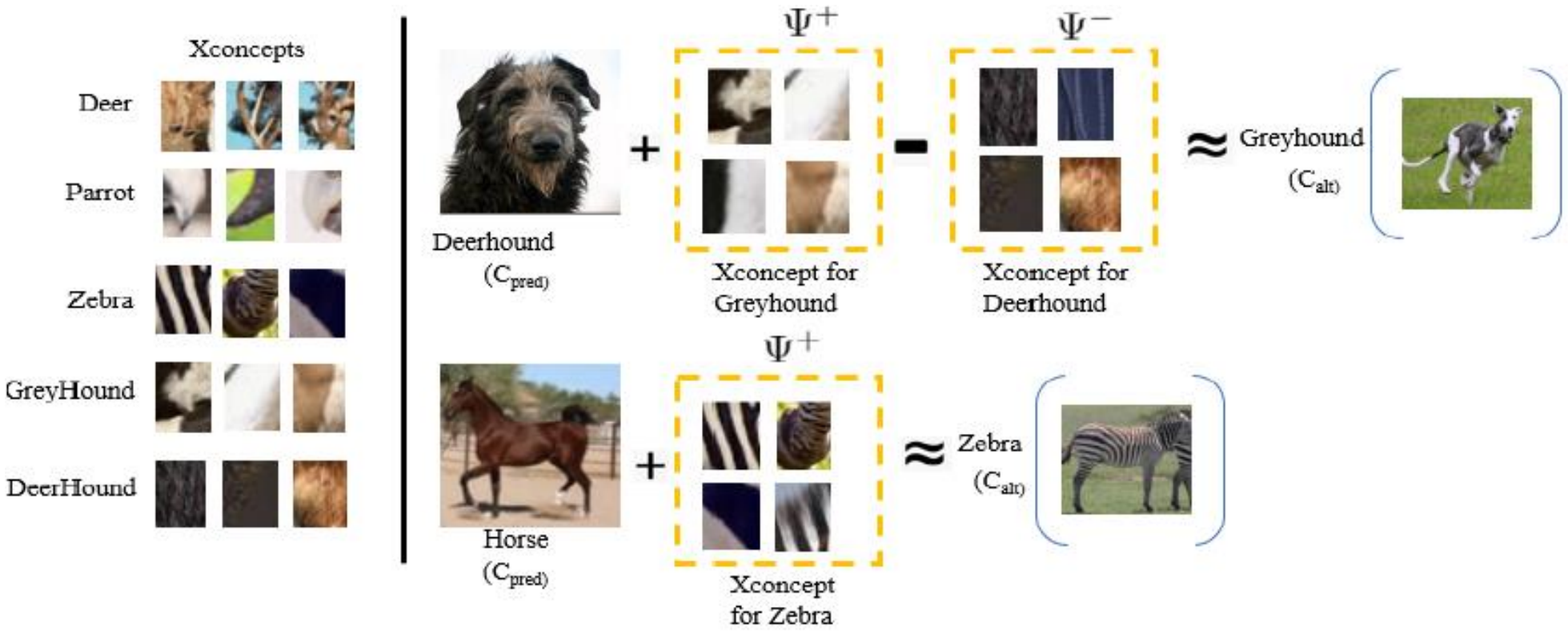
UsingFISTA

$pred$: the prediction class of original instance
alt: the desired class we try to alter to

Akula, A., Wang, S., & Zhu, S. C. (2020, April). Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 03, pp. 2594-2601).

CoCoX

CASE STUDY



Examples of xconcepts (Left) and counterfactual explanations (Right) identified by CoCoX

Akula, A., Wang, S., & Zhu, S. C. (2020, April). Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 03, pp. 2594-2601).

CoCoX

Experiments

Justified Trust (Quantitative Metric): given an image, it evaluates whether the users could reliably predict the model's output decision

Explanation Satisfaction (ES) (Qualitative Metric): measure human subjects' feeling of satisfaction at having achieved an understanding of the machine in terms of usefulness, sufficiency, appropriated detail, confidence, and accuracy

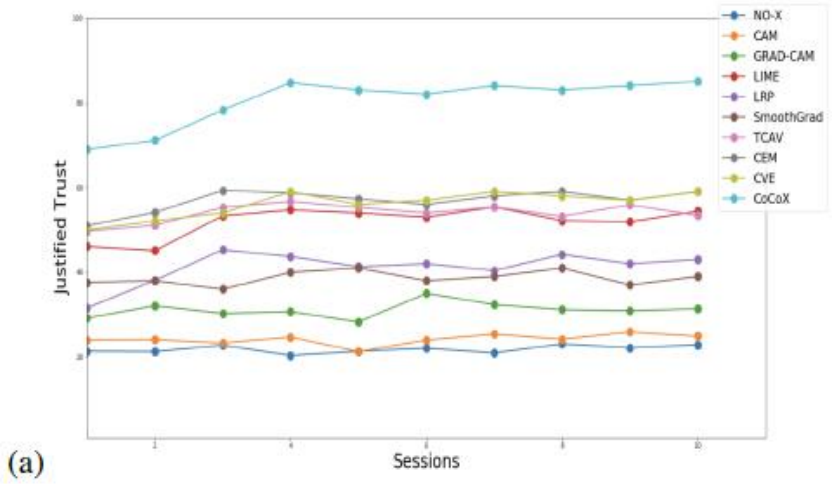
	XAI Framework	Justified Trust (±std)	Explanation Satisfaction (±std)				
			Confidence	Usefulness	Appropriate Detail	Understandability	Sufficiency
Non-Expert Subject Pool	Random Guessing	6.6 %	N/A	N/A	N/A	N/A	N/A
	NO-X	21.4 % ± 2.7 %	N/A	N/A	N/A	N/A	N/A
	CAM (Zhou et al. 2016)	24.0 % ± 1.9 %	4.2 ± 1.8	3.6 ± 0.8	2.2 ± 1.9	3.2 ± 0.9	2.6 ± 1.3
	Grad-CAM (Selvaraju et al. 2017)	29.2 % ± 3.1 %	4.1 ± 1.1	3.2 ± 1.9	3.0 ± 1.6	4.2 ± 1.1	3.2 ± 1.0
	LIME (Ribeiro, Singh, and Guestrin 2016)	46.1 % ± 1.2 %	5.1 ± 1.8	4.2 ± 1.6	3.9 ± 1.1	4.1 ± 2.0	4.3 ± 1.6
	LRP (Bach et al. 2015)	31.1 % ± 2.5 %	1.1 ± 2.2	2.8 ± 1.0	1.6 ± 1.7	2.8 ± 1.0	2.1 ± 1.8
	SmoothGrad (Smilkov et al. 2017)	37.6 % ± 2.9 %	1.4 ± 1.0	2.2 ± 1.8	2.8 ± 1.0	3.1 ± 0.8	2.9 ± 0.8
	TCAV (Kim et al. 2018)	49.7 % ± 3.3 %	3.6 ± 2.1	3.2 ± 1.8	3.3 ± 1.6	3.6 ± 2.1	3.9 ± 1.1
	CEM (Dhurandhar et al. 2018)	51.0 % ± 2.1 %	4.1 ± 1.4	3.4 ± 1.4	3.1 ± 2.1	2.9 ± 0.9	3.3 ± 1.6
	CVE (Goyal et al. 2019)	50.9 % ± 3.0 %	3.8 ± 1.9	3.1 ± 0.9	3.6 ± 2.1	4.1 ± 1.2	4.2 ± 1.2
	CoCoX (Fault-lines)	**69.1 % ± 2.1 %**	**6.2 ± 1.2**	**6.6 ± 0.7**	**7.2 ± 0.9**	**7.1 ± 0.6**	**6.2 ± 0.8**
Expert Subject Pool	NO-X	28.1 % ± 4.1 %	N/A	N/A	N/A	N/A	N/A
	CAM (Zhou et al. 2016)	37.1 % ± 3.9 %	3.2 ± 1.8	3.3 ± 1.4	3.1 ± 2.1	3.1 ± 1.8	2.9 ± 1.9
	Grad-CAM (Selvaraju et al. 2017)	39.1 % ± 2.1 %	3.7 ± 1.2	3.1 ± 2.2	2.7 ± 1.9	3.7 ± 1.1	3.4 ± 1.6
	LIME (Ribeiro, Singh, and Guestrin 2016)	42.1 % ± 3.1 %	3.1 ± 2.2	3.0 ± 1.2	2.8 ± 1.9	3.1 ± 2.2	2.8 ± 1.7
	LRP (Bach et al. 2015)	51.1 % ± 3.1 %	3.2 ± 4.1	3.5 ± 1.6	4.2 ± 1.5	4.3 ± 1.0	3.9 ± 0.9
	SmoothGrad (Smilkov et al. 2017)	40.7 % ± 2.1 %	3.1 ± 1.0	2.9 ± 1.2	3.8 ± 1.5	3.3 ± 1.1	3.1 ± 1.0
	TCAV (Kim et al. 2018)	55.1 % ± 3.3 %	3.9 ± 2.8	3.6 ± 1.6	4.1 ± 1.3	4.9 ± 1.2	3.9 ± 0.8
	CEM (Dhurandhar et al. 2018)	61.1 % ± 2.2 %	4.8 ± 1.6	3.7 ± 1.6	4.0 ± 1.2	3.7 ± 1.0	4.0 ± 1.1
	CVE (Goyal et al. 2019)	64.5 % ± 3.7 %	4.1 ± 2.3	3.9 ± 1.5	4.6 ± 1.5	4.5 ± 1.4	3.9 ± 1.2
	CoCoX (Fault-lines)	**70.5 % ± 1.3 %**	**5.7 ± 1.1**	**4.9 ± 0.8**	**5.8 ± 1.2**	**6.9 ± 1.1**	**6.4 ± 1.0**

Table 1: Quantitative (Justified Trust) and Qualitative (Explanation Satisfaction) comparison of CoCoX with random guessing baseline, no explanation (NO-X) baseline, and other state-of-the-art XAI frameworks such as CAM, Grad-CAM, LIME, LRP, SmoothGrad, TCAV, CEM, and CVE.

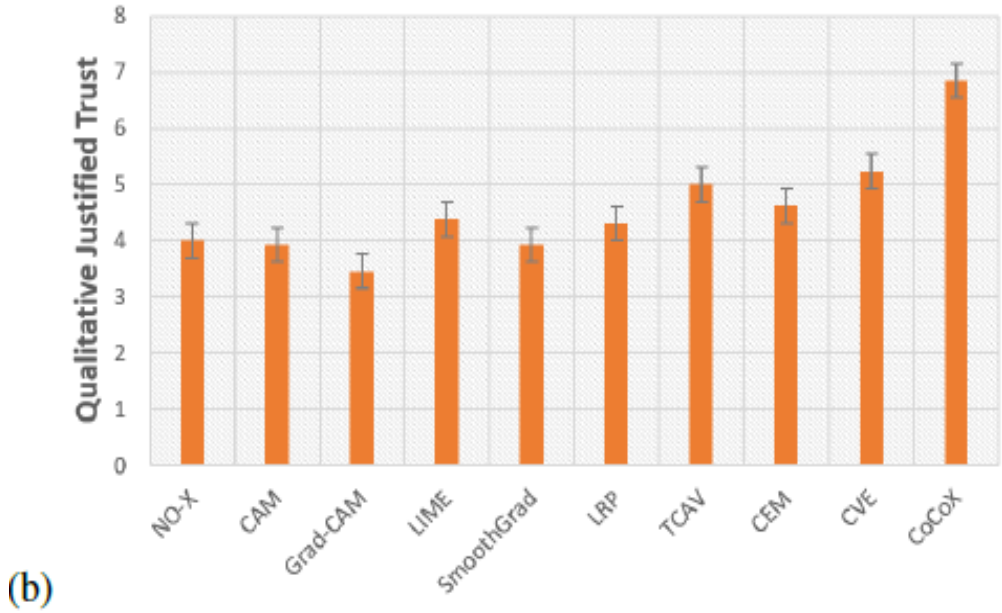
Akula, A., Wang, S., & Zhu, S. C. (2020, April). Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 03, pp. 2594-2601).

CoCoX

Experiment



(a) Gain in Justified Trust over time

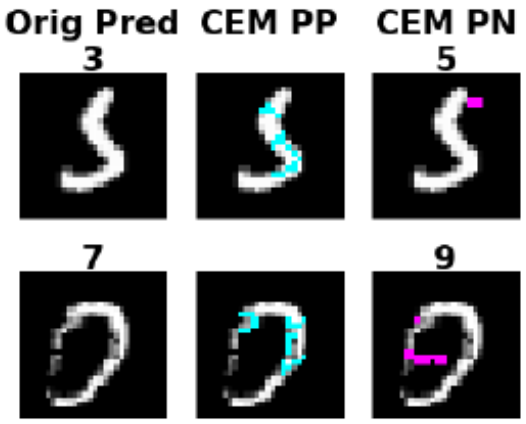


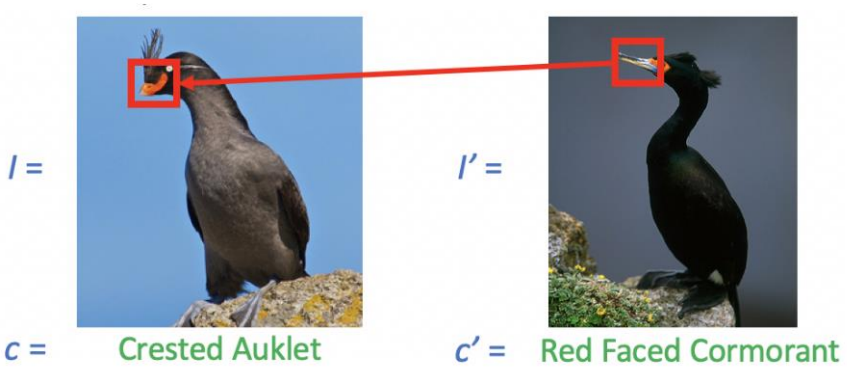
(b) Average Qualitative Justified Trust (on a Likert scale of 0 to 9). Error bars denote standard errors of the means.

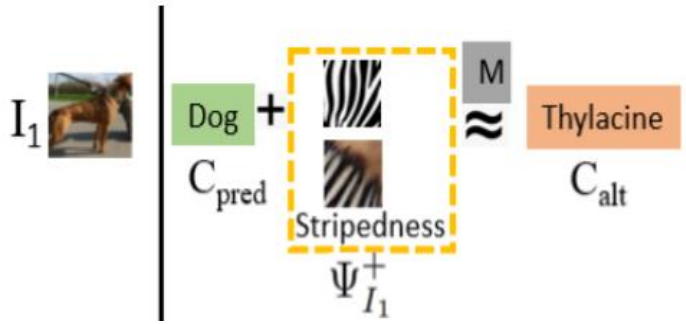
Akula, A., Wang, S., & Zhu, S. C. (2020, April). Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 03, pp. 2594-2601).

Selected References

1. Dhurandhar, A., Chen, P. Y., Luss, R., Tu, C. C., Ting, P., Shanmugam, K., & Das, P. (2018, December). Explanations based on the missing: towards contrastive explanations with pertinent negatives. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 590-601).
2. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019, May). Counterfactual visual explanations. In International Conference on Machine Learning (pp. 2376-2384). PMLR.
3. Wang, C., Han, H., & Cao, C.C. (2021). TDLS: A Top-Down Layer Searching Algorithm for Generating Counterfactual Visual Explanation.
4. Akula, A., Wang, S., & Zhu, S. C. (2020, April). Cocox: Generating conceptual and counterfactual explanations via fault-lines. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 03, pp. 2594-2601).





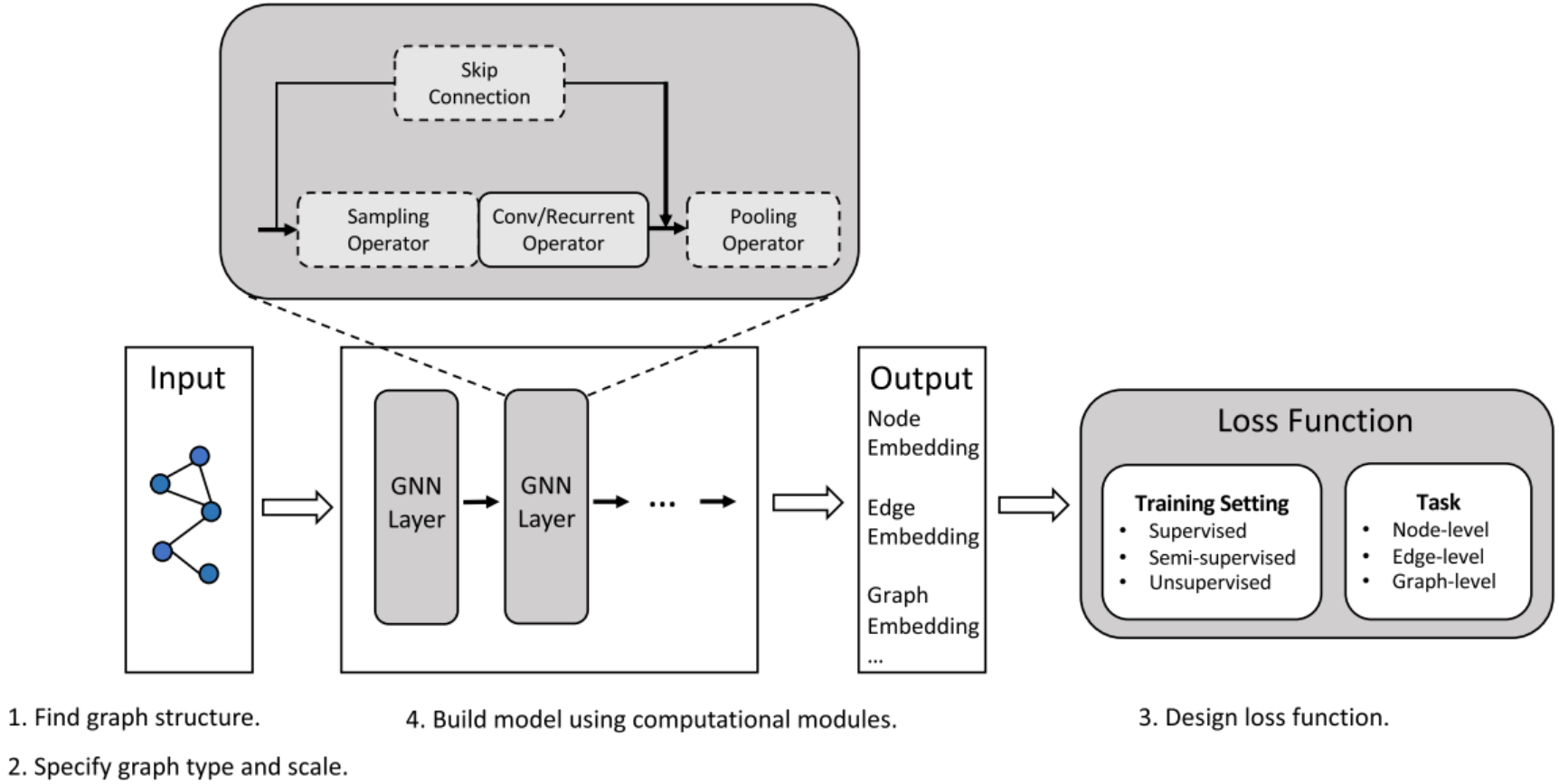


Counterfactual in GNN

- Counterfactuals for explaining GNN
  - Change both graph structures and node features
  - Only change graph structures
- GNN + Counterfactuals to explain other models

GNN Introduction

Graph neural networks (GNNs) are neural models that capture the dependence of graphs via message passing between the nodes of graphs.



The general design pipeline for a GNN model

Zhou, Jie, et al. "Graph neural networks: A review of methods and applications." AI Open 1 (2020): 57-81.

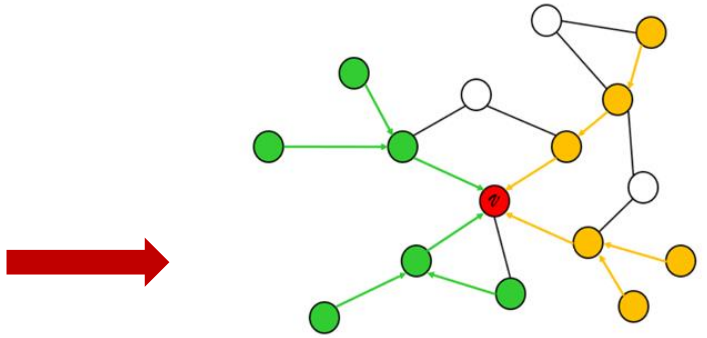
Counterfactuals for Explaining GNN

In practical, at a certain layer l of a GNN model, it aggregates neural message from adjacent nodes via their node features and edge connections.

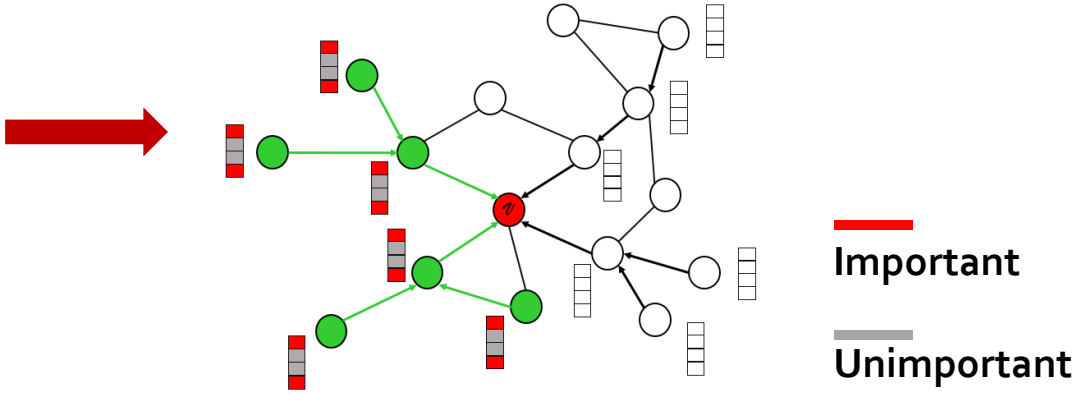
Counterfactuals for explaining GNN refers to perturb node features or graph structures to find minimal change that can convert the node class.

Most existing work can be divided into 2 classes:

a) Only perturb graph structures to generate CounterFactuals



b) Perturb both graph structures and node features



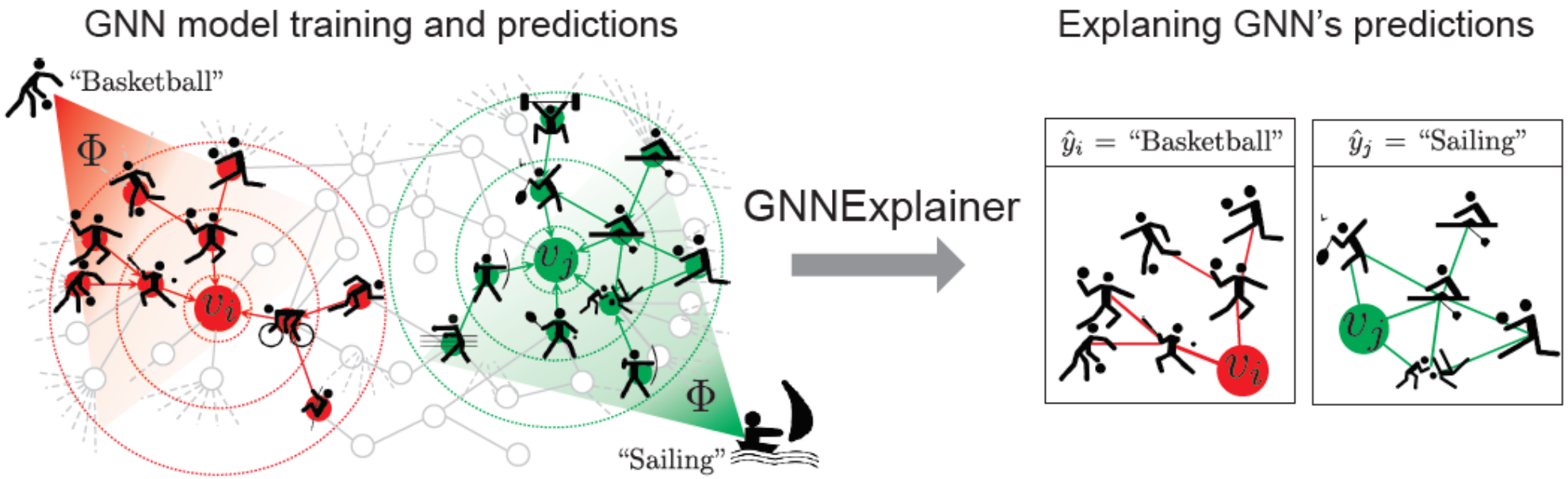
Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, *32*, 9240.

Change Both Graph Structures and Node Features

model-agnostic approach

any GNN-based model

any graph-based machine learning task



$$\max_{G_S} MI\,(Y, (G_S, X_S)) = H(Y) - H(Y|G = G_S, X = X_S)$$

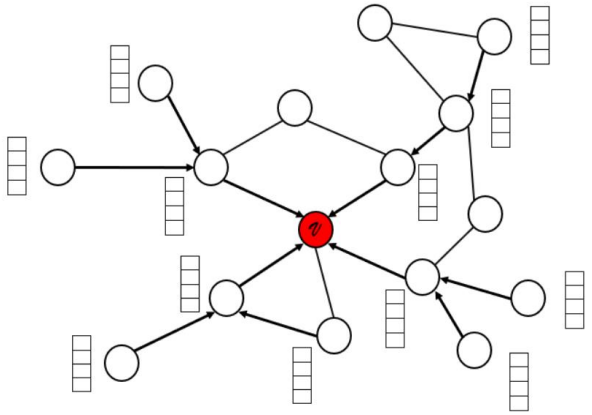
Given an instance, GNNEXPLAINER identifies a compact subgraph structure and a small subset of node features that have a crucial role in GNN's prediction

Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, *32*, 9240.

Change Both Graph Structures and Node Features

GOAL: To identify a set of important pathways and features for prediction

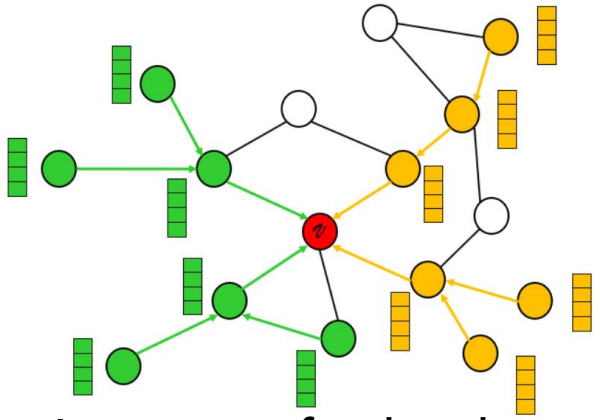
- G_c : Computation graph for making prediction at node *v*
- : Neural message-passing pathways from neighbors of node *v*
- : Features of each node



Computation Graph G_c

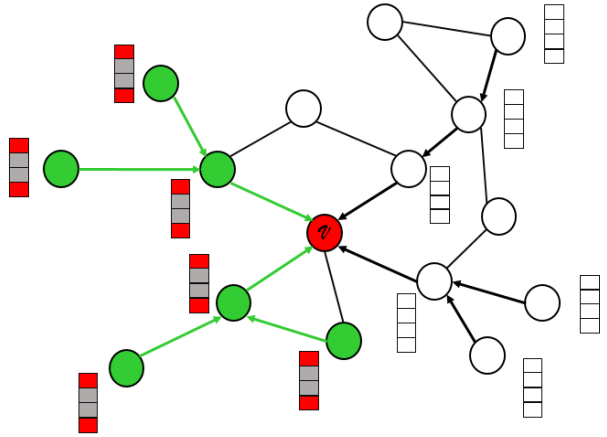
- : **Important** for prediction at node *v*
- : **Unimportant** for prediction at node *v*
- Message aggregation:

AGG(, ,)



Importance of each pathway

- To identify what feature dimensions of neighbors are important for prediction
- : **Important** feature
- : **Unimportant** feature



Importance of features

Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, *32*, 9240.

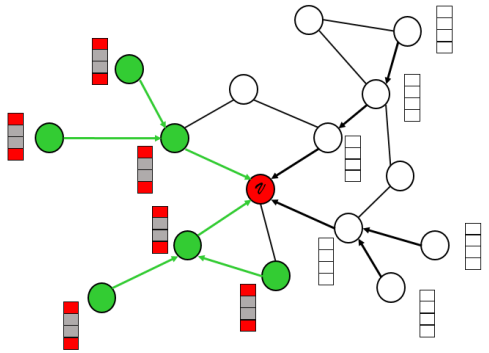
Change Both Graph Structures and Node Features

CounterFactuals in GNNExplainer: Maximizing mutual information MI between prediction distribution and that conditioned on the target subgraph as explanation.

Target: $G_S \subseteq G_c$

$X_S = \{x_j \in G_S\}$

Y is predicted label distribution



$$\max_{G_S} MI\,(Y, (G_S, X_S)) = H(Y) - H(Y|G = G_S, X = X_S).$$

constant — minimize

$$H(Y|G=G_S, X=X_S) = -\mathbb{E}_{Y|G_S,X_S}\left[\log P_\Phi(Y|G=G_S, X=X_S)\right]$$

Maximize

Minimize H →maximize probability of $\hat{y}$

Example: $v_j \in G_c(v_i), v_j \neq v_i$ if removed, probability of prediction $\hat{y}_i$ will strongly decreases, v_j can be regarded as a good counterfactual explanation

Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, *32*, 9240.

Only Change Graph Structures

CF-GNNEXPLAINER: iteratively removes edges from the original adjacency matrix based on matrix sparsification techniques.

Define a node $v = (A_v, x)$, where A_v is the subgraph adjacency matrix, x is the feature vector for v

A general GCN is $f(A, X; W) = \text{softmax}\left[\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}XW\right]$, where $\tilde{A} = A + I$. $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ are entries in the degree matrix.

Pipeline:

Train the model **f** → Init a binary perturbation matrix **P** → Init a CF generating function **g** → Optimize loss function to update **P** → $CF = P \odot A_v$

GNN model: $f(A_v, X_v; W) = \text{softmax}\left[(D_v + I)^{-1/2}(A_v + I)(D_v + I)^{-1/2}X_vW\right]$ (update)

CF generation function: $g(A_v, X_v. W; P) = \text{softmax}\left[\bar{D_v}^{-1/2}(P \odot A_v + I)\bar{D_v}^{-1/2}X_vW\right]$ (update, fix)

Lucic, A., ter Hoeve, M., Tolomei, G., de Rijke, M., & Silvestri, F. (2021). CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. *arXiv preprint arXiv:2102.03322*.

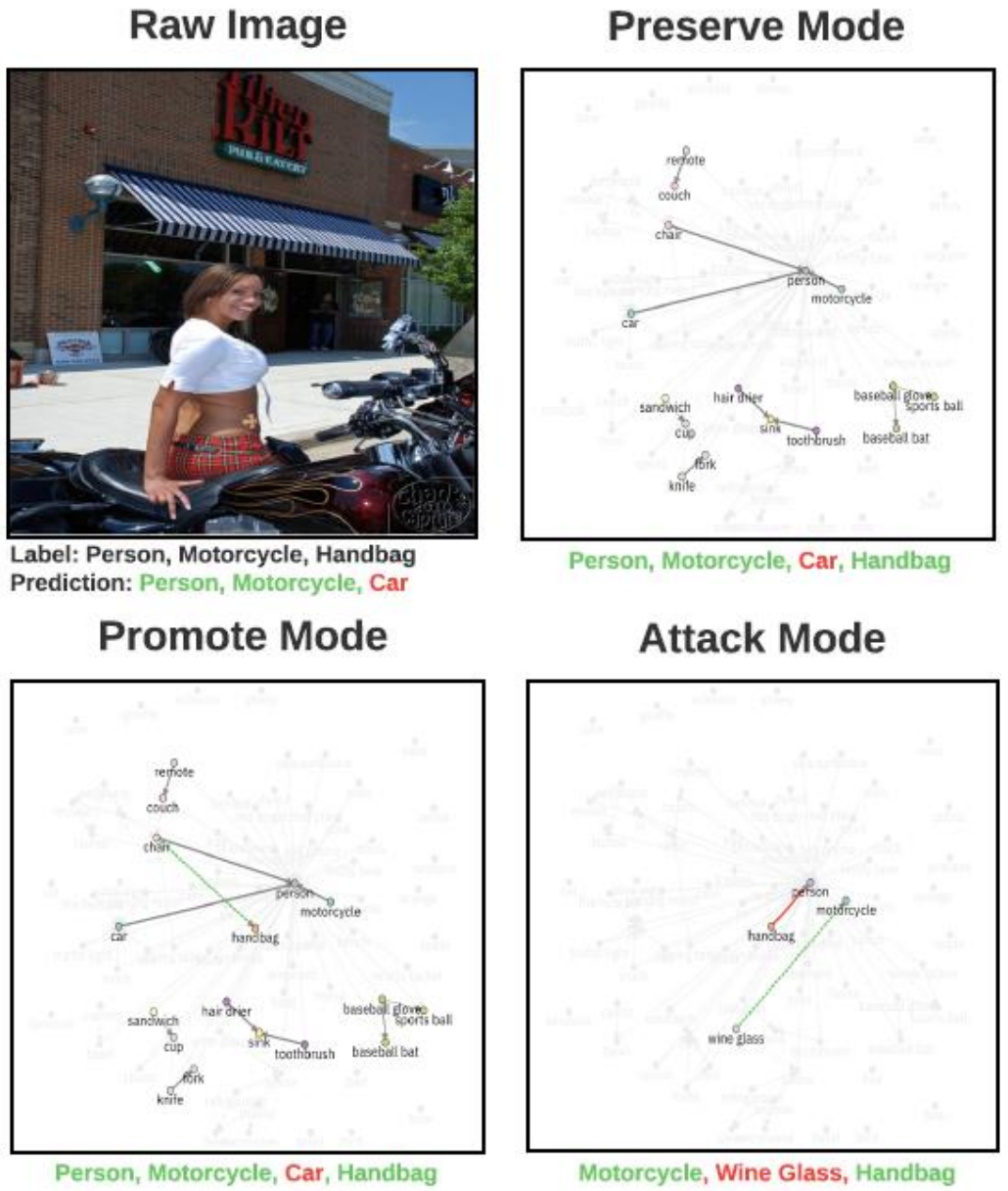
Only Change Graph Structures

Preserve, Promote or Attack GNN's predictions: Build an unified framework to measure the influence of graph topology perturbation into model predictions.

Preserve mode seeks the sparse graph pattern which provides factual GNN explanation.

Promote mode aims to fix GNN's wrong predictions.

Attack mode aims for counterfactual explanations.



Sun, Y., Valente, A., Liu, S., & Wang, D. (2021). Preserve, Promote, or Attack? GNN Explanation via Topology Perturbation. *arXiv preprint arXiv:2103.13944*.

Only Change Graph Structures

Preserve, Promote or Attack GNN's predictions

A is a binary adjacency matrix

$\bar{\mathbf{A}} = 11^T - \mathbf{I} - \mathbf{A}$, the supplement of A

$\mathbf{C} = \bar{\mathbf{A}} - \mathbf{A}$, All possible perturbations of A

$\mathbf{C} = \mathbf{C}^- + \mathbf{C}^+,\ \mathbf{C}^- \leq 0,\ \mathbf{C}^+ \geq 0,$

$\mathbf{A}'(\mathbf{S}) = \mathbf{A} + \mathbf{C}^- \circ \mathbf{S}^- + \mathbf{C}^+ \circ \mathbf{S}^+$, the topology attribution map

$$\mathrm{A} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \qquad \bar{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

$$\mathrm{C} = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}$$

Sparsity-promoting optimization

$$\begin{aligned} \underset{\mathbf{S}}{\text{maximize}} \quad & \mathcal{R}(\mathbf{A}'(\mathbf{S}); \mathbf{X}, \mathbf{W}^*) + \lambda_1 \|\mathbf{S}^-\|_1 - \lambda_2 \|\mathbf{S}^+\|_1 \\ \text{subject to} \quad & S_{ij} \in [0,1], \forall i, j, \end{aligned}$$

$$C^- = \begin{pmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix} \qquad C^+ = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

$$S^- = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad S^+ = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

Attack mode (Counterfactual explanations)

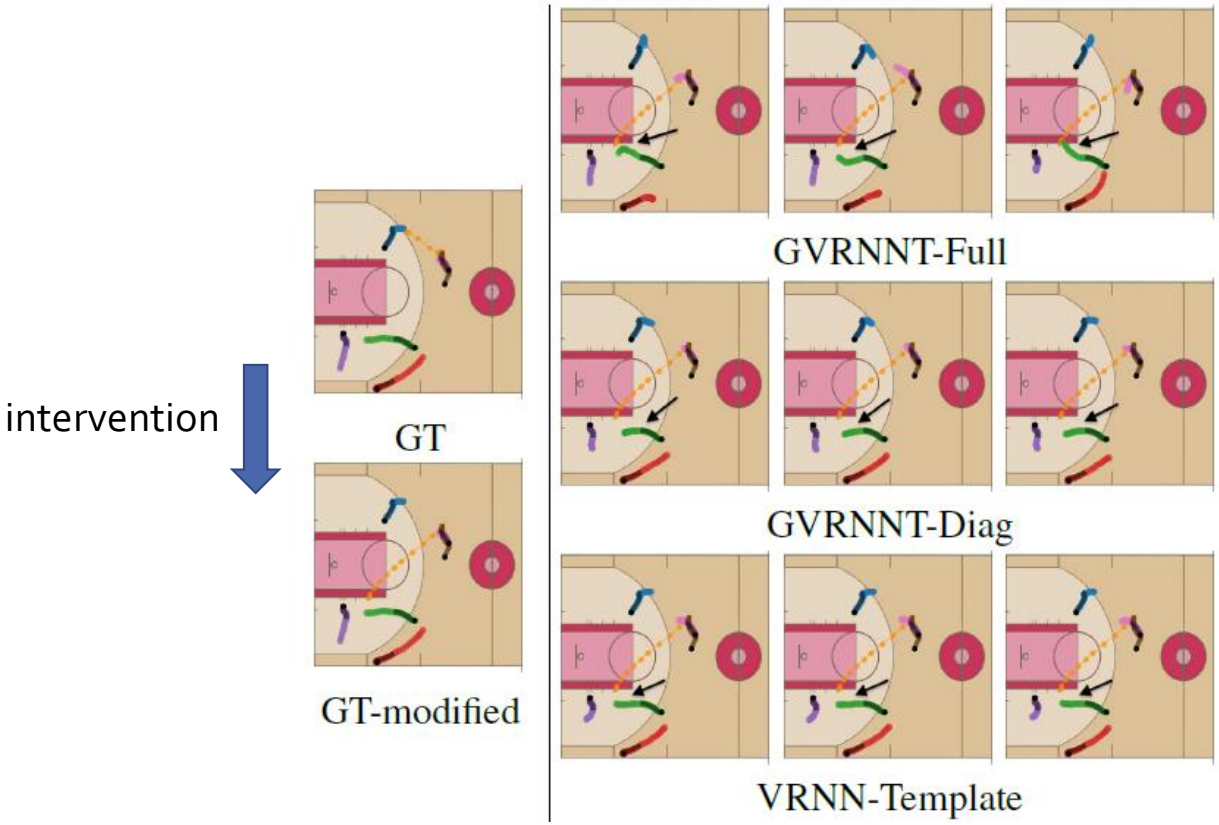
$$\mathcal{R}(\mathrm{A}'(\mathrm{S}); \mathrm{X}, \mathrm{W}^*) = \min \left\{ \max_{\mathrm{t} \notin \Omega} \mathrm{p_t}(\mathrm{A}'(\mathrm{S}); \mathrm{X}, \mathrm{W}^*) - \max_{\mathrm{c} \in \Omega} \mathrm{p_c}(\mathrm{A}'(\mathrm{S}); \mathrm{X}, \mathrm{W}^*), \kappa \right\}$$

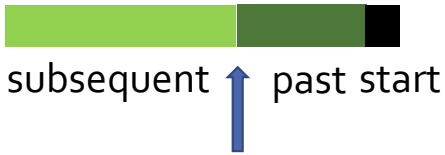
Sun, Y., Valente, A., Liu, S., & Wang, D. (2021). Preserve, Promote, or Attack? GNN Explanation via Topology Perturbation. *arXiv preprint arXiv:2103.13944*.

GNN + Counterfactuals to explain other models

Graph Variational RNN (GVRNN) can be used to answer counterfactual questions like:

"What would happen if player A passed the ball to player B instead of player C?"



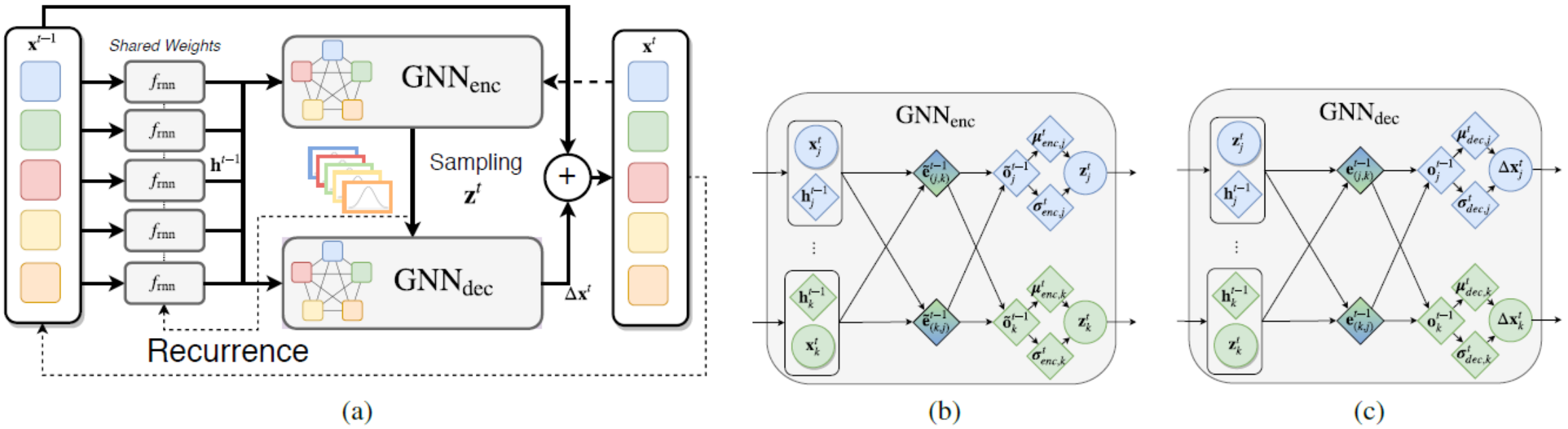


- At a high level, a VRNN is a VAE at every time step. Therefore we can modify conditions to have a look at how interactions among multi agents change via GNN.

Yeh, R. A., Schwing, A. G., Huang, J., & Murphy, K. (2019). Diverse generation for multi-agent sports games. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4610-4619).

GNN + Counterfactuals to explain other models

Graph Variational RNN (GVRNN) for modeling multi-agent sports



(a) The overall architecture. (b)(c) Details of the GNN between two agents as examples

Yeh, R. A., Schwing, A. G., Huang, J., & Murphy, K. (2019). Diverse generation for multi-agent sports games. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4610-4619).

Selected References

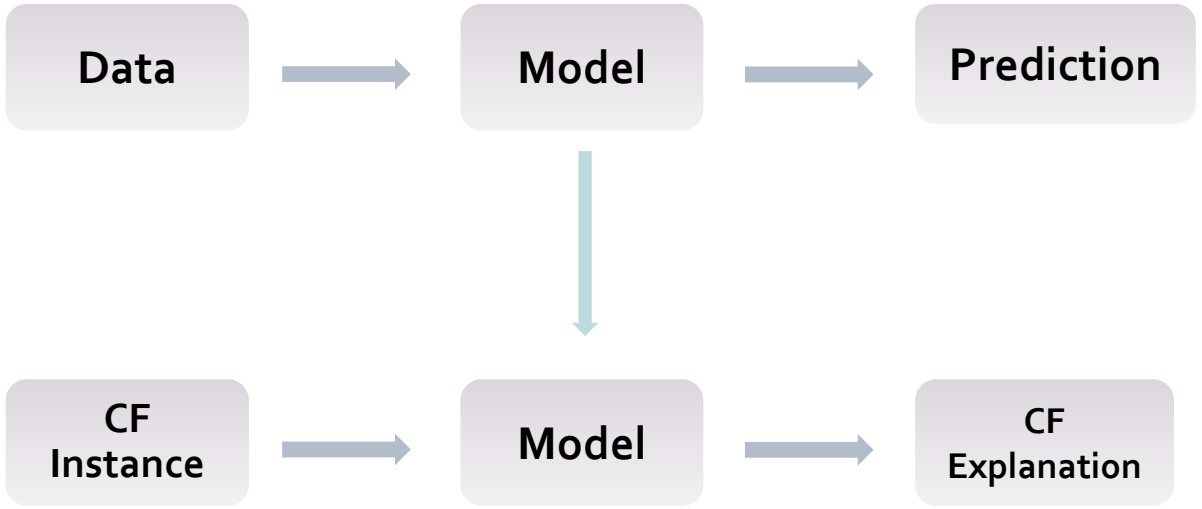
1. Zhou, Jie, et al. "Graph neural networks: A review of methods and applications." AI Open 1 (2020): 57-81.
2. Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. Advances in neural information processing systems, 32, 9240.
3. Lucic, A., ter Hoeve, M., Tolomei, G., de Rijke, M., & Silvestri, F. (2021). CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. arXiv preprint arXiv:2102.03322.
4. Sun, Y., Valente, A., Liu, S., & Wang, D. (2021). Preserve, Promote, or Attack? GNN Explanation via Topology Perturbation. arXiv preprint arXiv:2103.13944.
5. Yeh, R. A., Schwing, A. G., Huang, J., & Murphy, K. (2019). Diverse generation for multi-agent sports games. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4610-4619).

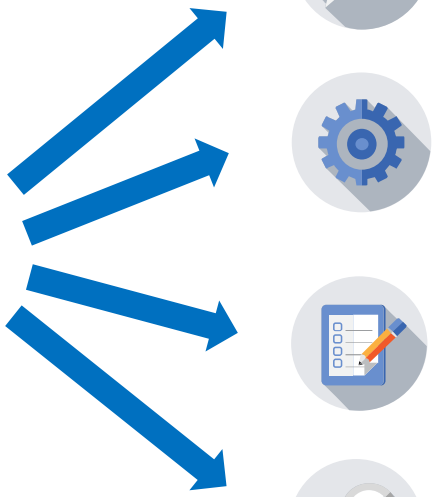
05

Applications of counterfactual

Luning WANG

Application Scenarios of Counterfactual Explanation





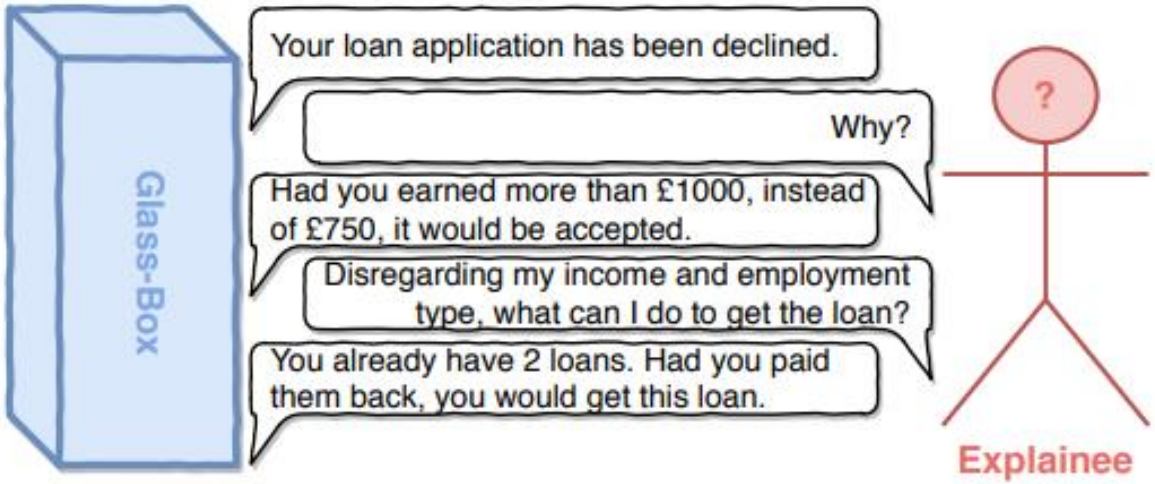
Human – Model Understanding

Business – Model Debug

Data – Service Enhancement

Standard – Reliability Proof

Interact with Human – A More Human-Centric Explanation



Interact with the system in a natural way – Dialogue[1]

"Your loan application has been *declined*. If you were a *skilled employee* instead of an *unskilled – resident*, your loan application would be *accepted*."

A user-centric interpretation, not a precooked template[2]

Sokol, Kacper, and Peter A. Flach. "Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant." IJCAI. 2018.
Sokol, Kacper, and Peter A. Flach. "Conversational Explanations of Machine Learning Predictions Through Class-contrastive Counterfactual Statements." IJCAI. 2018.

Interact with Business - A More Explicit Debug Approach

CF explanation of a Financial Model

Your loan application has been **declined**. If your **savings account** had had *more than* **100** pounds, you had **not** had a savings account or its status had been **unknown**, your loan application would be **accepted**.

Your loan application has been **declined**. Assuming that you had **asked for** *less or equal* to **£663** or *between* **883** and **1285** pounds, instead of **836** pounds, your loan application would be **accepted**.

More than 100 pounds	or	**Not have a savings account**	**Business logic conflict**
$x \leq 663$	or	**$883 \leq x \leq 1285$**	**Non-monotonic**

Locate the boundary of the anomaly prediction[1]

Sokol, Kacper, and Peter A. Flach. "Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety." SafeAI@ AAAI. 2019.

Interact with Data - A More Effective Data Augmentation Approach

Input Sentence	**Token-based Substitution** (Ribeiro et al. 2020), (Devlin et al. 2018)	**Adversarial Attack** (Michel et al. 2019)	**GYC** (Ours)
I am very *disappointed* with the service	I am very **pleased** with the service. I am very **happy** with the service. I am very **impressed** with the service.	I am very **witty** with the service.	I am very **pleased to get a good** service. I am very **happy** with **this** service. I am very **pleased** with the service.
1st time burnt *pizza* was *horrible*.	1st time burnt **house** was **destroyed**. 1st time burnt **place** was **burned**. 1st time burnt **house** was **burnt**.	**personable** time burnt pizza was horrible.	1st time burnt pizza **is delicious**. 1st time burnt **coffee** was **delicious**. 1st time burned pizza was **delicious**.

Model : Sentiment Classifier
Source: Negative Class Label, Target: Positive Class Label

Input Sentence: *I am very disappointed with the service.*
Counterfactual Text Samples :
[1] I am very pleased with the service.
[2] I am very happy with this service
[3] I am very pleased to get a good service.

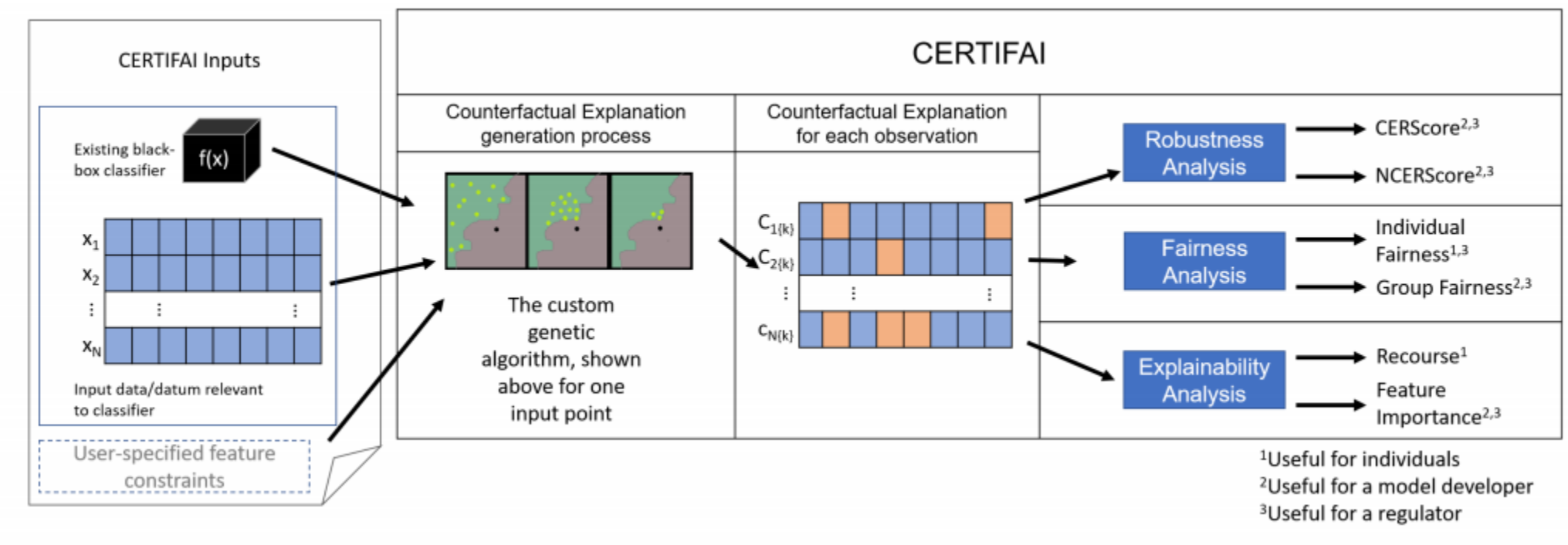
Model : Topic Classifier
Source Topic: World, Target Topic: Sci-Fi

Input Sentence: *The country is at war with terrorism.*
Counterfactual Text Samples :
[1] The country is at war with piracy at international waters.
[2] The country is at war with its own beurocracy.
[3] The country is at war with piracy offenses.

Counterfactual Instance – examples of GYC framework

Madaan, Nishtha, et al. "Generate your counterfactuals: Towards controlled counterfactual generation for text." arXiv preprint arXiv:2012.04698 (2020).

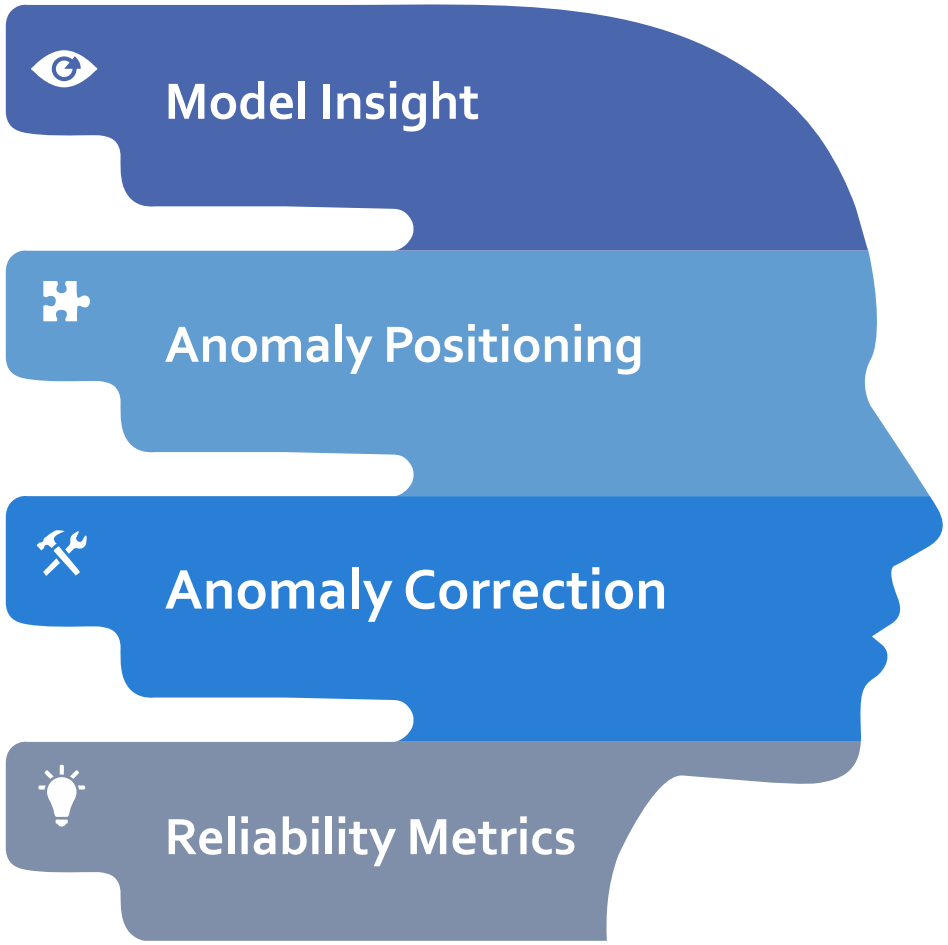
Interact with Standard - A More Reliable Service



Robustness, fairness, transparency metrics designed based on counterfactual interpretation

Sharma, Shubham, Jette Henderson, and Joydeep Ghosh. "Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models." arXiv preprint arXiv:1905.07857 (2019).

Inspiration of Future Application



Improve Model Effectiveness

- Flexible Interaction
- User-Centric Presentation
- …

- Decision Boundaries
- Business Logic
- …

- Data Augmentation
- Model Enhancement
- …

- Fairness
- Robustness
- …

Selected References

1. Sokol, K., & Flach, P. A. (2018, July). Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant. In *IJCAI (pp. 5868-5870).*
2. Sokol, K., & Flach, P. A. (2018, July). Conversational Explanations of Machine Learning Predictions Through Class-contrastive Counterfactual Statements. In *IJCAI (pp. 5785-5786).*
3. Sokol, K., & Flach, P. A. (2019, January). Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. In *SafeAI@ AAAI.*
4. *Madaan, N., Padhi, I., Panwar, N., & Saha, D. (2020). Generate your counterfactuals: Towards controlled counterfactual generation for text. arXiv preprint arXiv:2012.04698.*
5. *Madaan, N., Padhi, I., Panwar, N., & Saha, D. (2020). Generate your counterfactuals: Towards controlled counterfactual generation for text. arXiv preprint arXiv:2012.04698.*

05

Conclusion

Lei CHEN

Conclusion: Computation of Counterfactual

	Method Keyword	Description
Classic Counterfactual Generation	Heuristic	Define two distance term, one between predictions, one between instances
Advanced Counterfactual Generation	Weighted	Consider features' importance
	Diverse	Using determinantal point
	Mixed Polytope	Using mixed polytope deal with categorical features
	Prototype	Using prototype guide the perturbation
	GAN	Probabilistic generation from a generator trained with discriminator adversarially

Conclusion: Metrics for Counterfactual

Category	Metric	Description
Computational Metrics	Validity	Whether the counterfactuals that actually have the desired class label
	Proximity	Distance of a counterfactual from the input datapoint
	Sparsity	Number of features difference between original input and a CF example
	Diversity	Measures feature-wise distances between different generated counterfactuals
Cognitive Metrics	Intuitiveness, friendliness & comprehensibility	How intuitive, user-friendly and comprehensible the counterfactuals are
	Understandability	How easy it is for the users to understand the explanations

Conclusion: Counterfactual Explanations in Different Areas

NLP: Perturb on words

Original x

It is not great for kids.

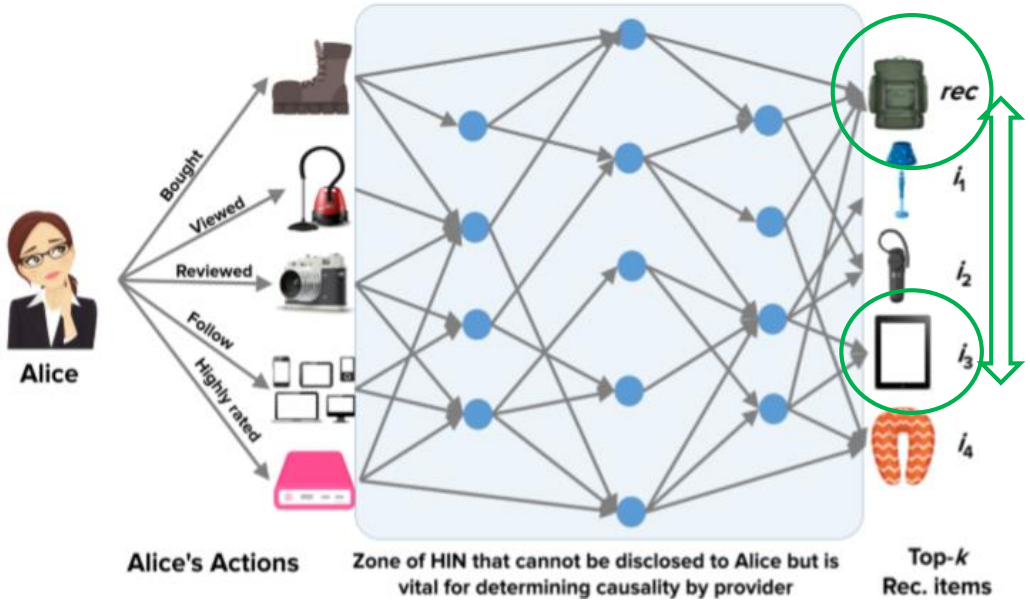
Removing/Inserting

It is ~~not~~ great for kids.

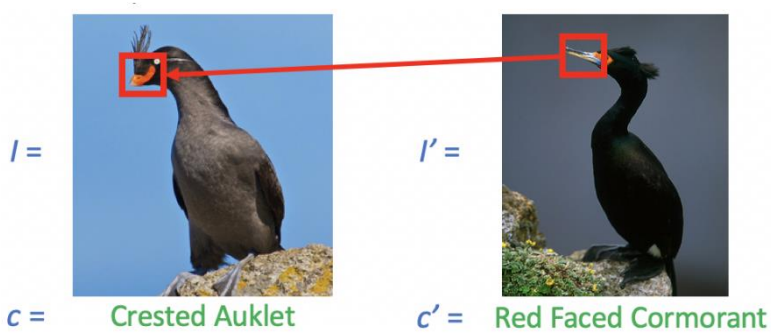
Replacing

It is not great → bad for kids.

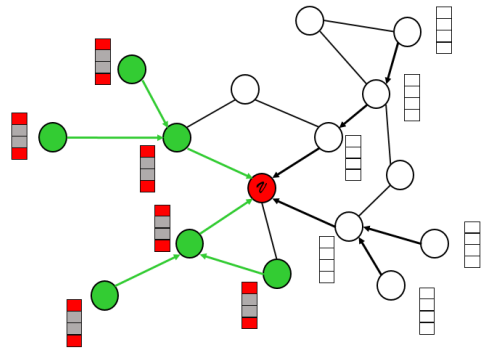
RS: Perturb on actions



CV: Perturb on feature maps

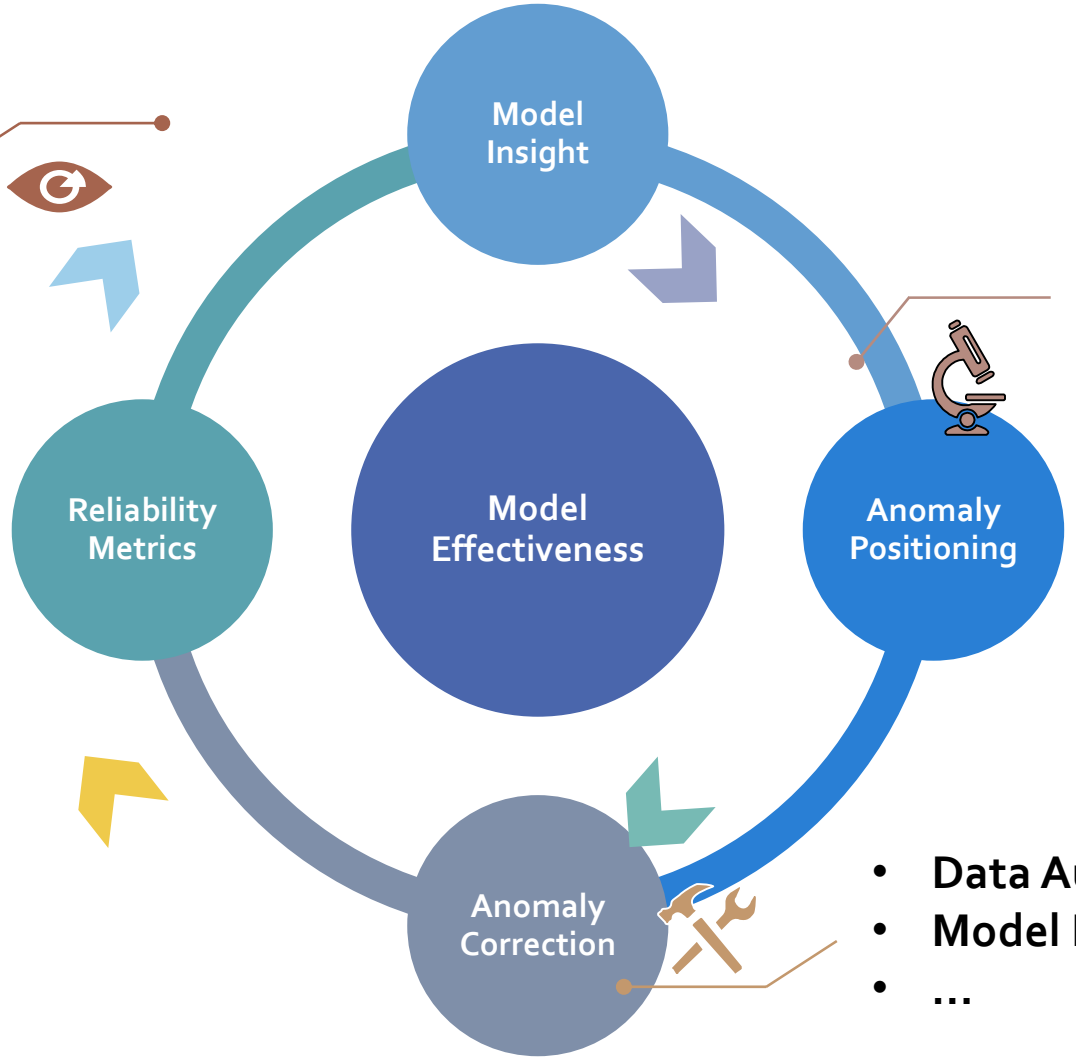


GNN: Perturb on edges/nodes



Conclusion: Applications of Counterfactual Explanations

- **Flexible Interaction**
- **User-Centric Presentation**
- **…**

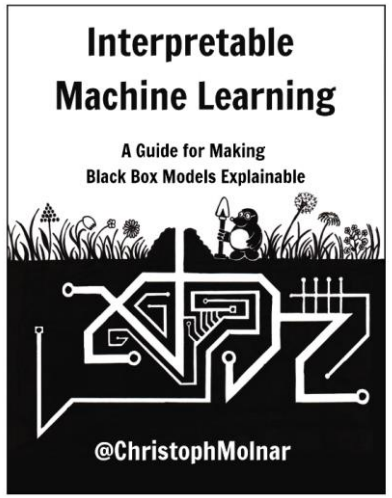


- **Decision Boundaries**
- **Business Logic**
- **…**

- **Fairness**
- **Robustness**
- **…**

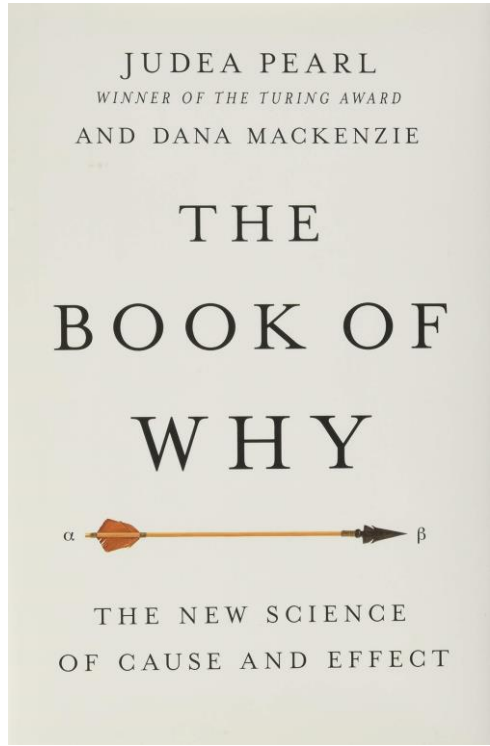
- **Data Augmentation**
- **Model Enhancement**
- **…**

There are many brilliant researchers/teams



Christoph Molnal

XAI project from Italy



Judea Pearl

IBM Research

MindInsight built on MindSpore

and more…

Q & A

Thank You.

XAI Team, DDL

14/AUG/2021