Post-hoc XAI Methods: Counterfactuals and XCBR Applications

Betül Bayrak

19.12.2024

NTNU | Norwegian University of Science and Technology

DART DATA and AI Research Unit

EXAIGON

Outline

Motivation

Part-1:

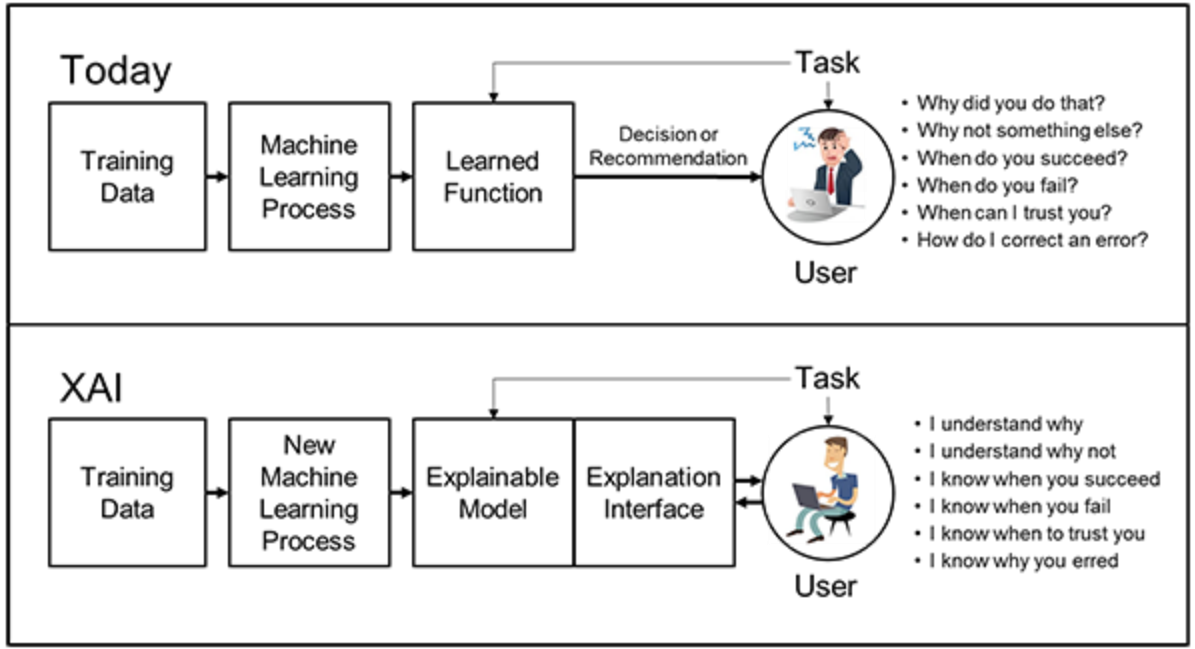
- Terms and concepts
- XAI framework: Factors, Components, Tasks
- Overview of explainers

Part-2:

My Ph.D.:

- RQs
- Papers
- Contributions
- Limitations

Motivation (XAI)



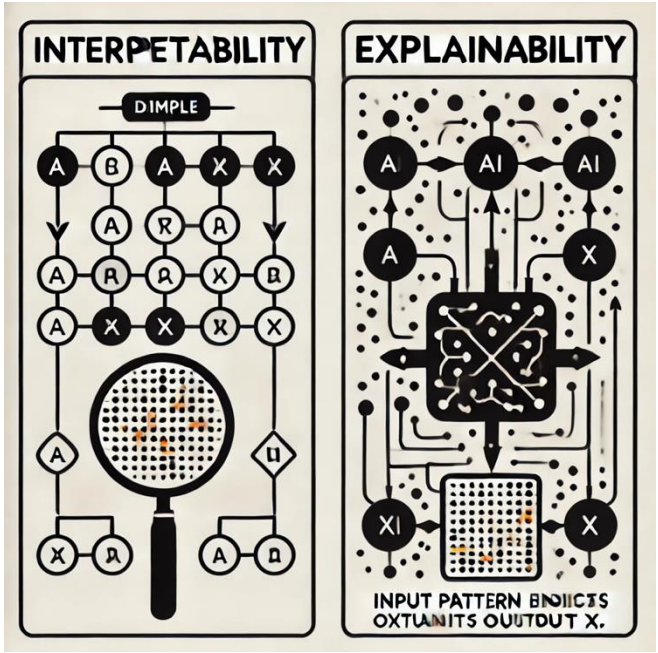
https://tensor-solutions.com/explainable-ai
https://www.geeksforgeeks.org/explainable-artificial-intelligencexai/

PART-1

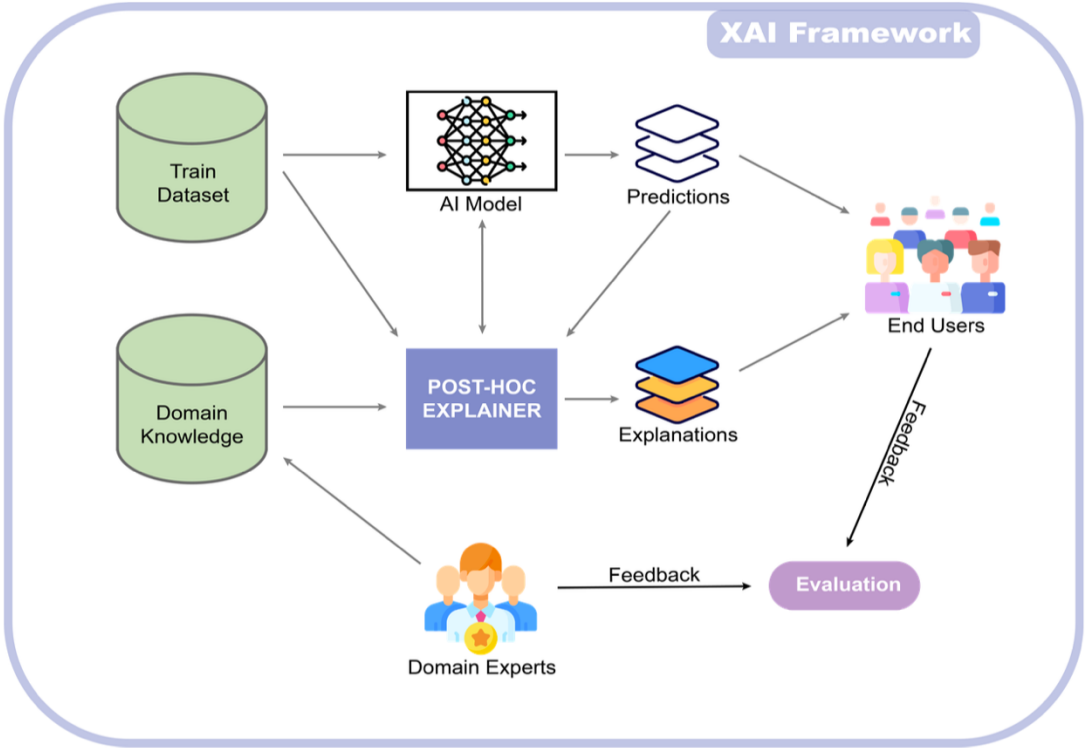
Explainability vs Interpretability

Interpretability refers to the visibility and **understanding of the inner logic** and mechanics of the AI model.

Explainability is the **ability to describe the behavior of a system** in an understandable medium to humans.



XAI Framework Overview



Each component supports the broader goal of explainability, allowing AI systems to generate insights that align with technical demands and user requirements.

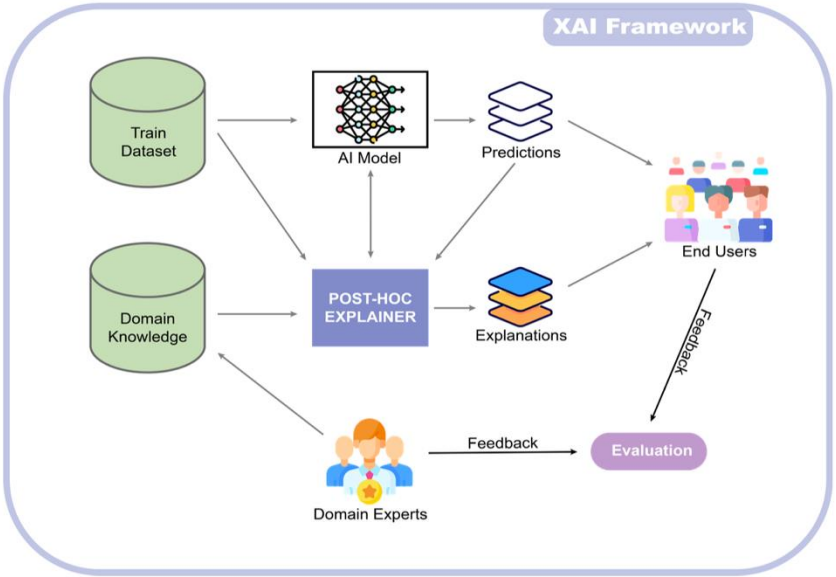
XAI framework:

1. Human-centered design
2. Factors in development
3. Components and tasks

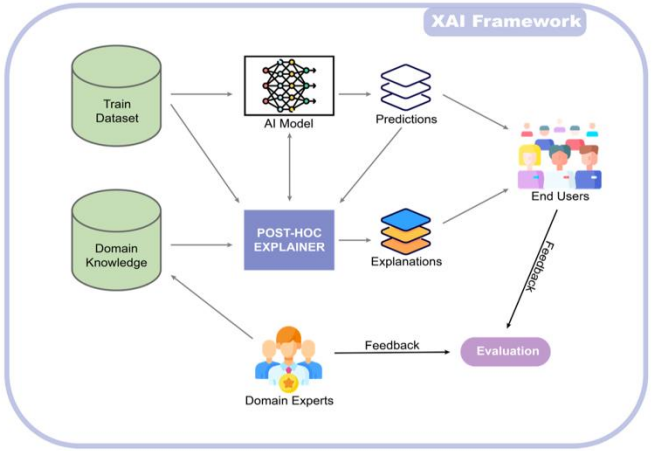
Human-centered design

Users are the **recipients** of the system (in most cases)

The human-centered approach focuses on the **what, when, and how to explain things to human end users** through the explanation process that involves the users in the development process.



Factors in development



I. Domain Analysis and Requirements

- End Users
- Knowledge Sources
- Actionability
- Ethical Considerations
- Inputs and Outputs
- Assessments
- Reusability

II. Multi-modal Interaction and Human-Centered Design

- Coverage
- Personal Preferences
- Usability Testing

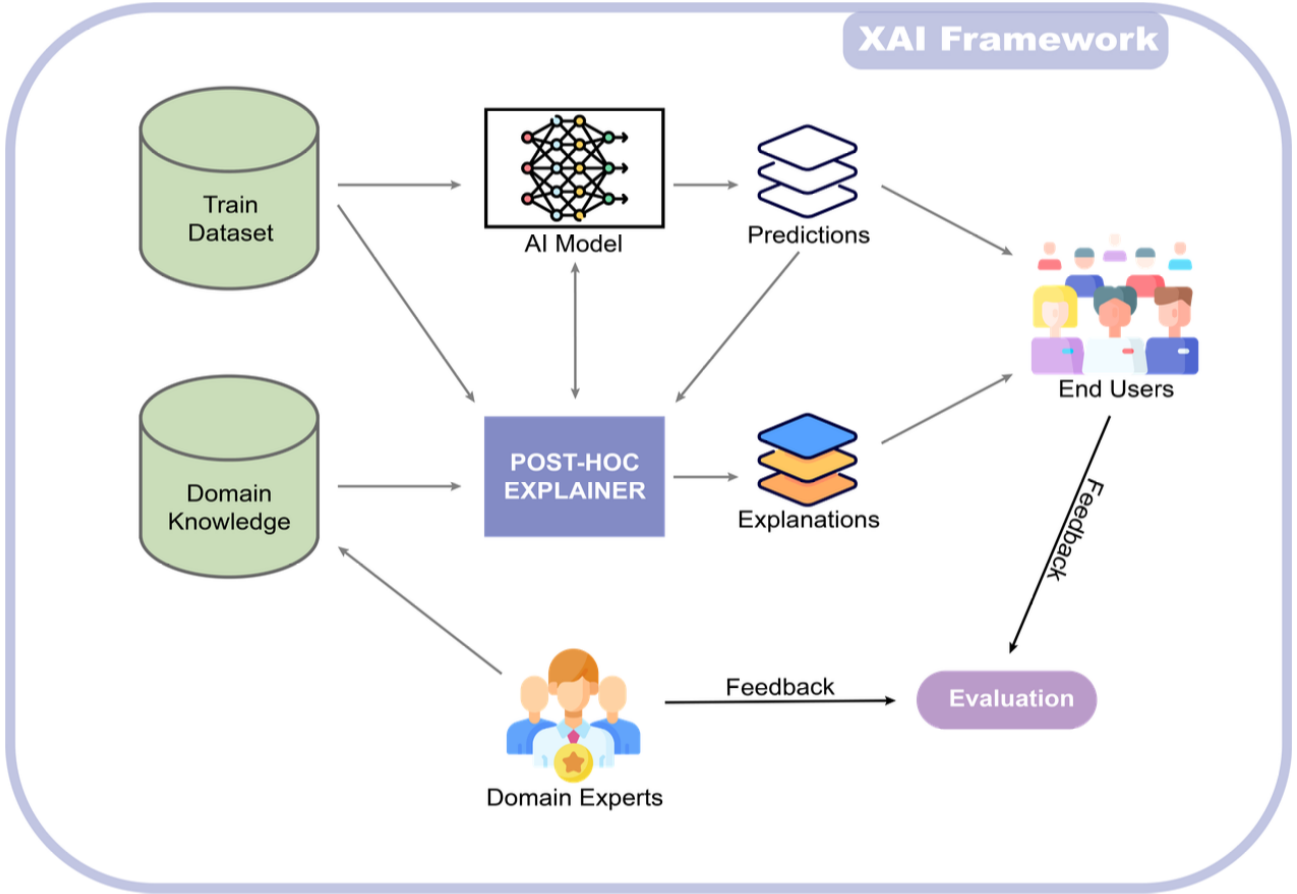
Components and tasks

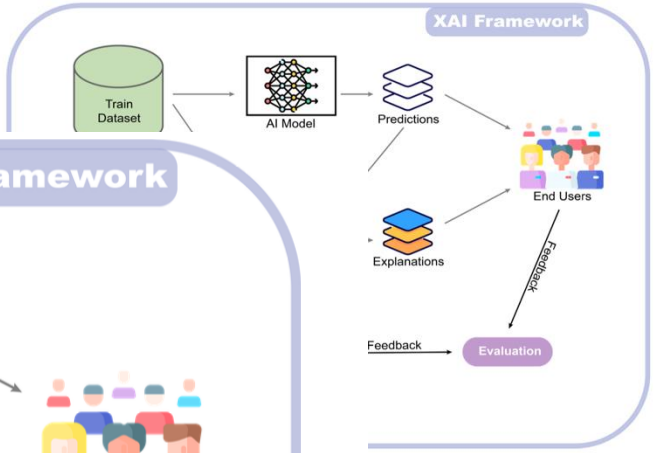
I. C

- A
- D
- S
- F
- E

II. T

- F
- E
- D
- E





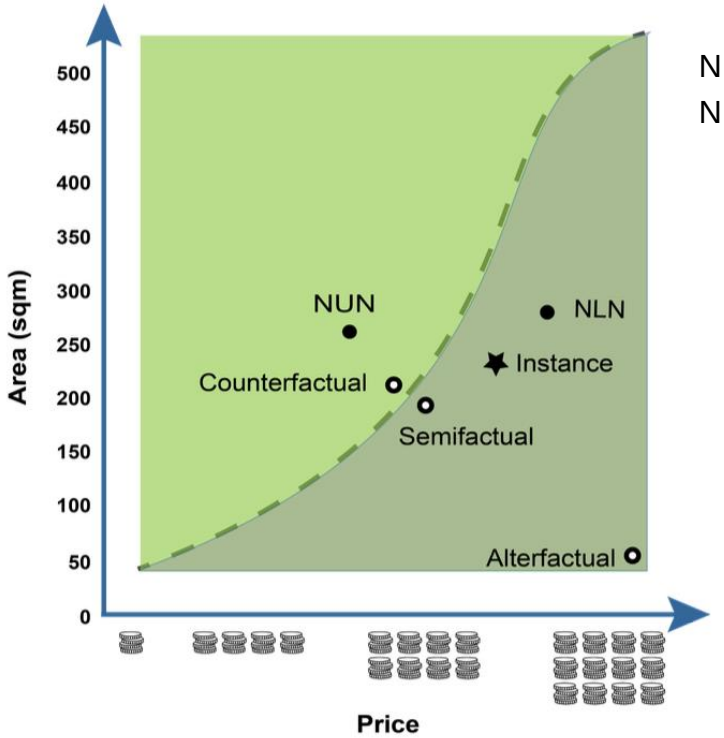
Explainers

Explainers

Ante-hoc Explainers	**Decision Trees**	
	Generalized Additive Models (GAMs	
	Bayesian Rule Lists (BRLs)	
	Linear Models & Logistic Regression	
Post-hoc Explainers	**Instance-Based Explainers**	**Counterfactuals** **Semifactuals** **Alterfactuals** **Prototypes & Criticisms**
	Attribution-Based Explainers	**SHAP** **Integrated Gradients** **Grad-CAM** **Saliency Maps**
	Rule-Based Explainers	**Anchors** **Decision rules**
	Visualization-Based Explainers	**PDPs**: Marginal effect of features **ICE Plots** **ALE Plots**

Instance-based Explainers

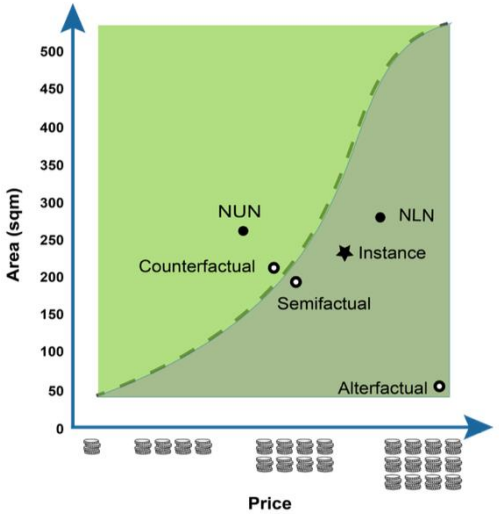
	Price	Area	Decision
Instance		230 sqm	
NLN		270 sqm	
NUN		270 sqm	
Counterfactual		210 sqm	
Semifactual		180 sqm	
Alterfactual		50 sqm	

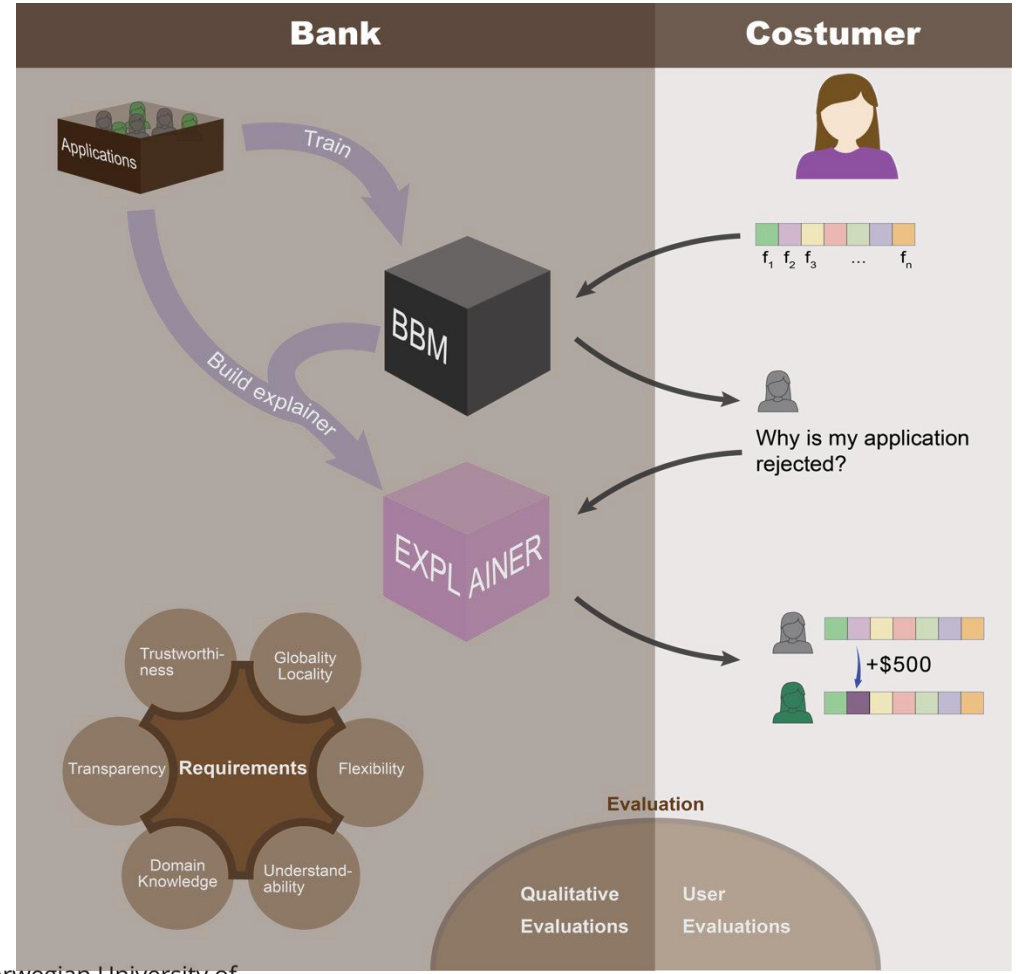


NLN: Nearest Like Neighbour
NUN: Nearest Unlike Neighbour

Instance-based Explainers

- **NLN:** Nearest Like Neighbour
- **NUN:** Nearest Unlike Neighbour
- **Counterfactual Explanations:** Generate hypothetical scenarios to describe how changing the input features can alter the prediction.
- **Semifactual Explanations:** Suggest that even with changes in certain attributes, the model's outcome would remain the same. "even if... still..." scenarios
- **Alterfactual Explanations:** Demonstrate the irrelevance of certain features by proposing changes that do not alter the model



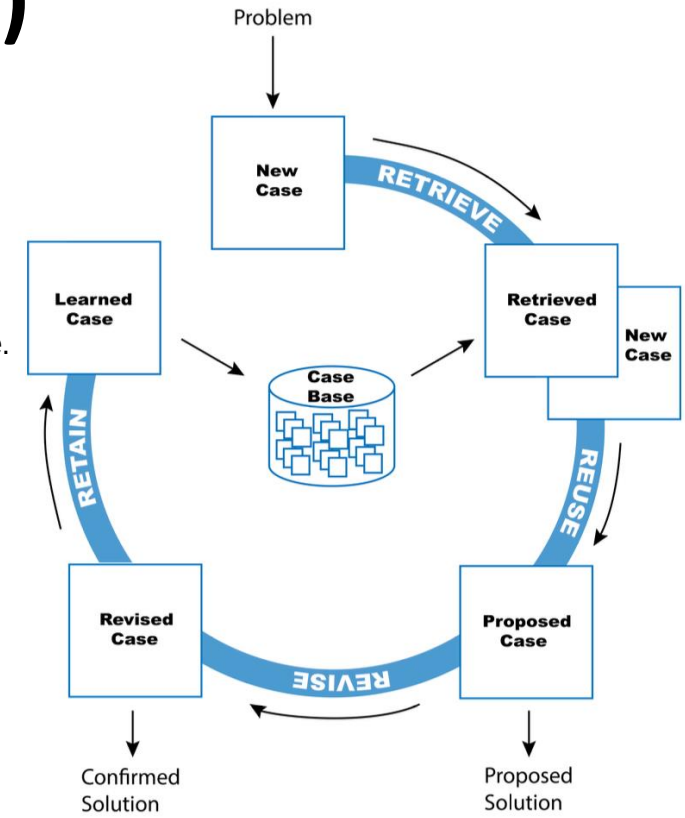

Bank
Costumer
Applications
Train
BBM
Build explainer
EXPL AINER
f_1 f_2 f_3 ... f_n
Why is my application rejected?
+$500
Trustworthi-ness
Globality Locality
Transparency
Requirements
Flexibility
Domain Knowledge
Understand-ability
Evaluation
Qualitative Evaluations
User Evaluations

Case-Based Reasoning (CBR)

CBR is a problem-solving methodology that uses past experiences to address new problems.

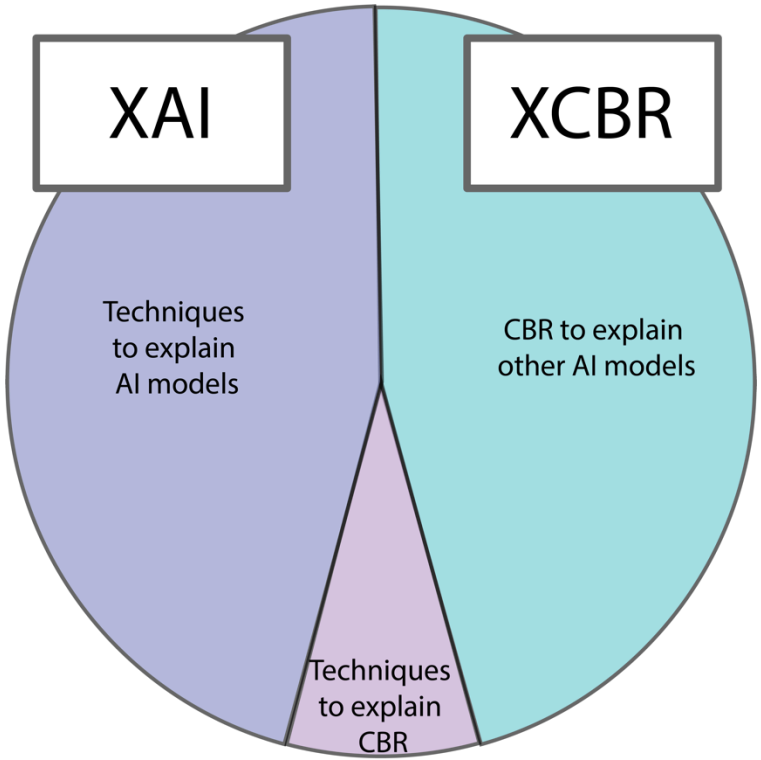
CBR operates through a four-step cycle:

1. **Retrieve**: Identify and retrieve the most relevant past cases from the case base.
2. **Reuse**: Adapt the retrieved case(s) to solve the new problem.
3. **Revise**: Test the proposed solution and refine it if necessary.
4. **Retain**: Store the new case and solution in the case base for future use.



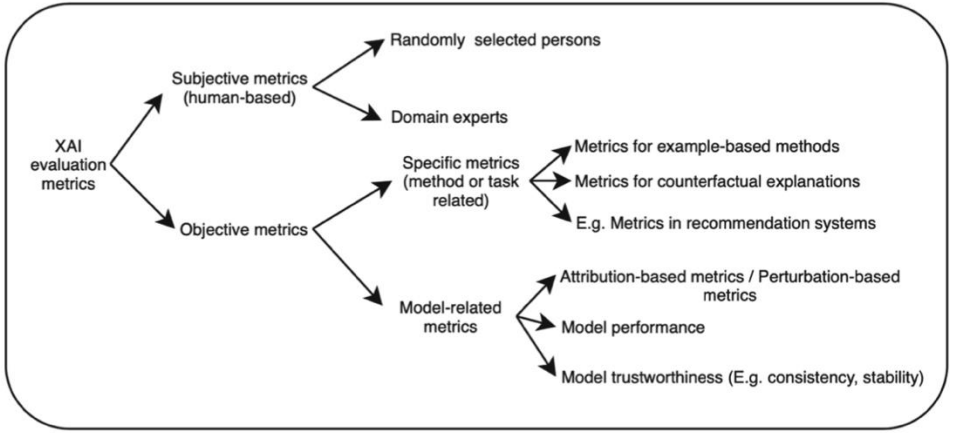
XCBR

- CBR is a flexible and interpretable methodology.
- It is used to explain AI models



https://gaia.fdi.ucm.es/research/excbr/

Evaluation



Subjective Evaluation

- User Satisfaction
- Comprehensibility
- Trustworthiness
- Actionability

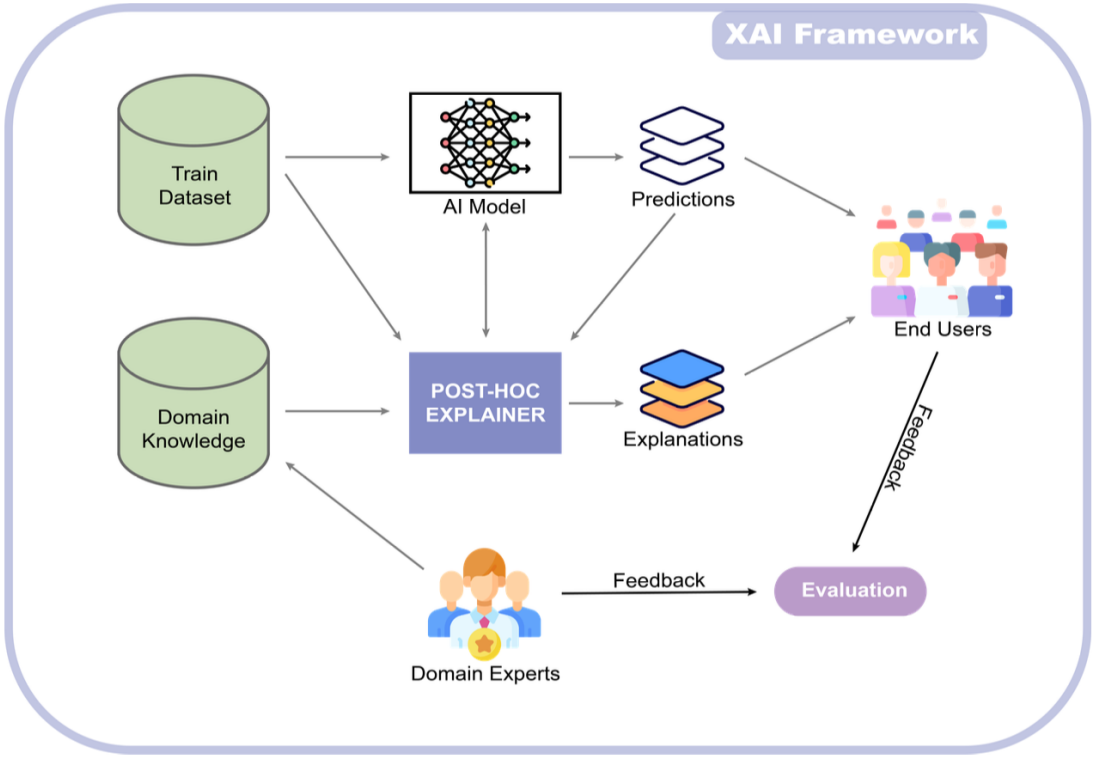
Objective Evaluation

- Fidelity
- Stability
- Proximity
- Comprehensiveness
- Sufficiency
- Sparsity

Coroama, Loredana, and Adrian Groza. "Evaluation metrics in explainable artificial intelligence (XAI)." *International conference on advanced research in technologies, information, innovation and sustainability.* Cham: Springer Nature Switzerland, 2022.

PART-2

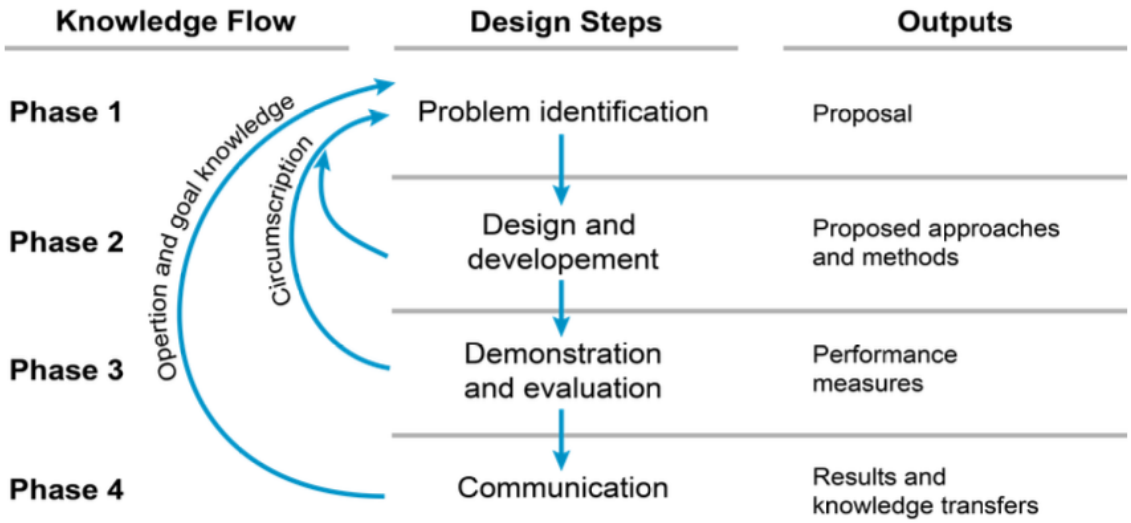
Focus


XAI Framework
Train Dataset
AI Model
Predictions
End Users
Domain Knowledge
POST-HOC EXPLAINER
Explanations
Feedback
Feedback
Evaluation
Domain Experts

Focus

Explainers			
	Ante-hoc Explainers	Decision Trees	
		Generalized Additive Models (GAMs	
		Bayesian Rule Lists (BRLs)	
		Linear Models & Logistic Regression	
	Post-hoc Explainers	Instance-Based Explainers	Counterfactuals Semifactuals Alterfactuals Prototypes & Criticisms
		Attribution-Based Explainers	SHAP Integrated Gradients Grad-CAM Saliency Maps
		Rule-Based Explainers	Anchors Decision rules
		Visualization-Based Explainers	PDPs: Marginal effect of features ICE Plots ALE Plots

My Ph.D.


Knowledge Flow
Design Steps
Outputs
Phase 1
Phase 2
Phase 3
Phase 4
Opertion and goal knowledge
Circumscription
Problem identification
Design and developement
Demonstration and evaluation
Communication
Proposal
Proposed approaches and methods
Performance measures
Results and knowledge transfers

Research questions

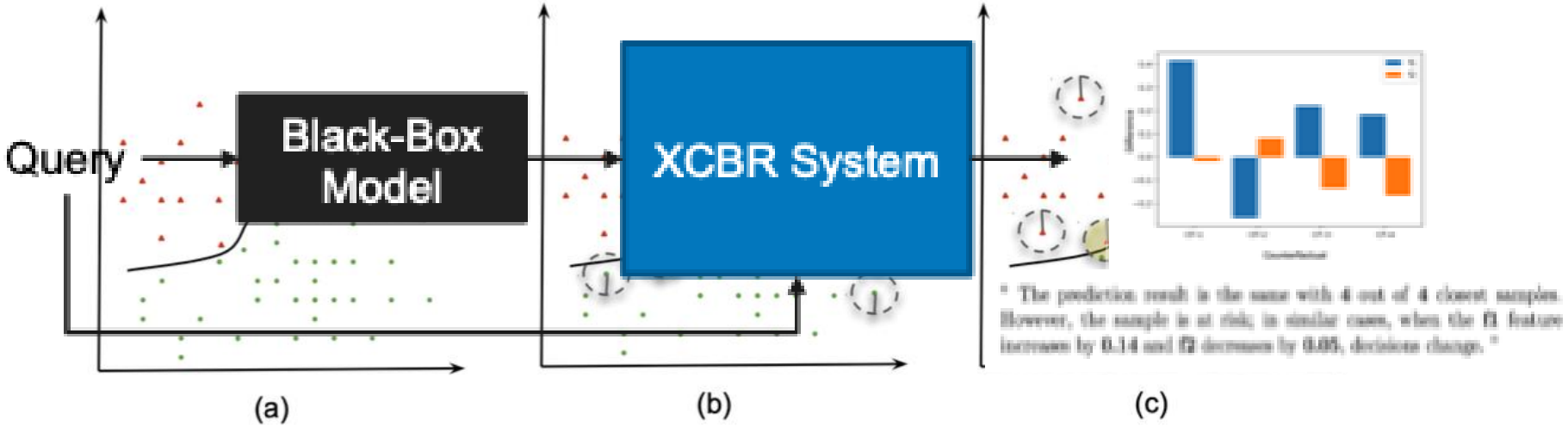
Research Questions		
	RQ1: Domain Knowledge and Design	
	RQ2: Generating plausible explanations	RQ2.1: High-quality counterfactuals
		RQ2.2: Multimodel counterfactuals
	RQ3: Improving comprehensibility	RQ3.1: Instance-based explanations
		RQ3.2: Evaluating explanations

Research questions - contributions

Research Questions		P-I	P-II	P-III	P-IV	P-V
RQ 1: Domain Knowledge and Design		*	*			*
RQ 2: Generating Plausible Explanations	RQ 2.1: High-quality Counterfactuals			*	*	
	RQ 2.2: Multimodal Counterfactuals					*
RQ 3: Improving Comprehensibility	RQ 3.1: Instance-Based Explanations	*	*		*	*
	RQ 3.2: Evaluating Explanations				*	*

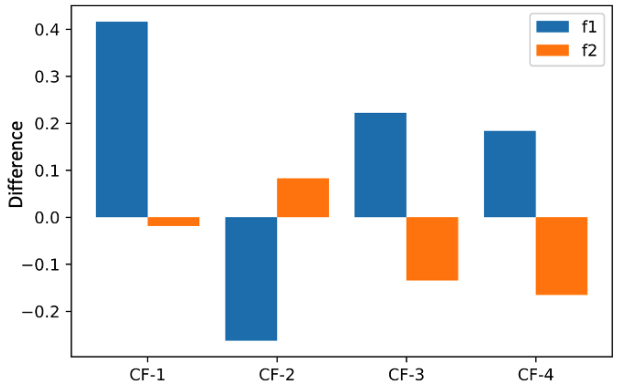
Paper 1: When to Explain?

- A model-agnostic XCBR framework that **selectively** triggers explanations



Paper 1: When to Explain?

Explanation Case #324
s: <f_1, f_2, . . . , f_n, label>
cf: <f_1, f_2, . . . , f_n, ¬label>



" The prediction result is the same with **4** out of **4** closest samples. However, the sample is at risk; in similar cases, when the **f1** feature increases by **0.14** and **f2** decreases by **0.05**, decisions change. "

1. **Store Data**: Maintain a case base with pairs of samples and their counterfactuals.
2. **Retrieve Pair**: Identify the most similar sample-counterfactual pair.
3. **Show Differences**: Highlight key differences between the sample and counterfactual to provide actionable insights.
4. **Visualization**: Use a combination of a **bi-directional bar graph** and **text annotations** to present the results effectively.

Paper 1: Contributions

Research Questions		
	RQ1: Domain Knowledge and Design	
	RQ2: Generating plausible explanations	RQ2.1: High-quality counterfactuals
		RQ2.2: Multimodel counterfactuals
	RQ3: Improving comprehensibility	RQ3.1: Instance-based explanations
		RQ3.2: Evaluating explanations

Paper2 - A Twin XCBR System Using Supportive and Contrastive Explanations

Integrate domain knowledge

Create case bases
Import cases

Model local similarity metrics

Model global similarity metrics

Expert Knowledge

Data

Black Box Model

Data-driven similarity modeling

Shap values per feature/class

Expert knowledge + Data

+ local sim

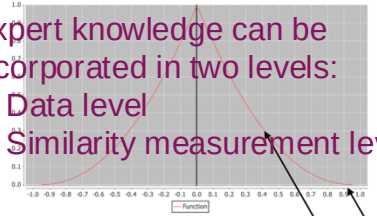
cases

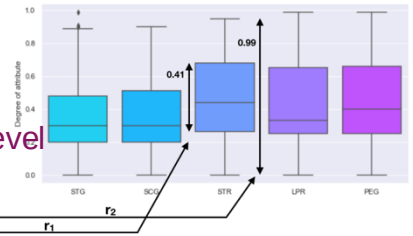
+ global sim

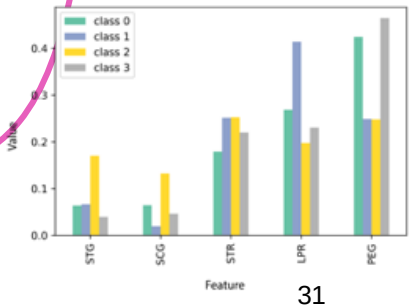
MA CBR System

Expert knowledge can be incorporated in two levels:

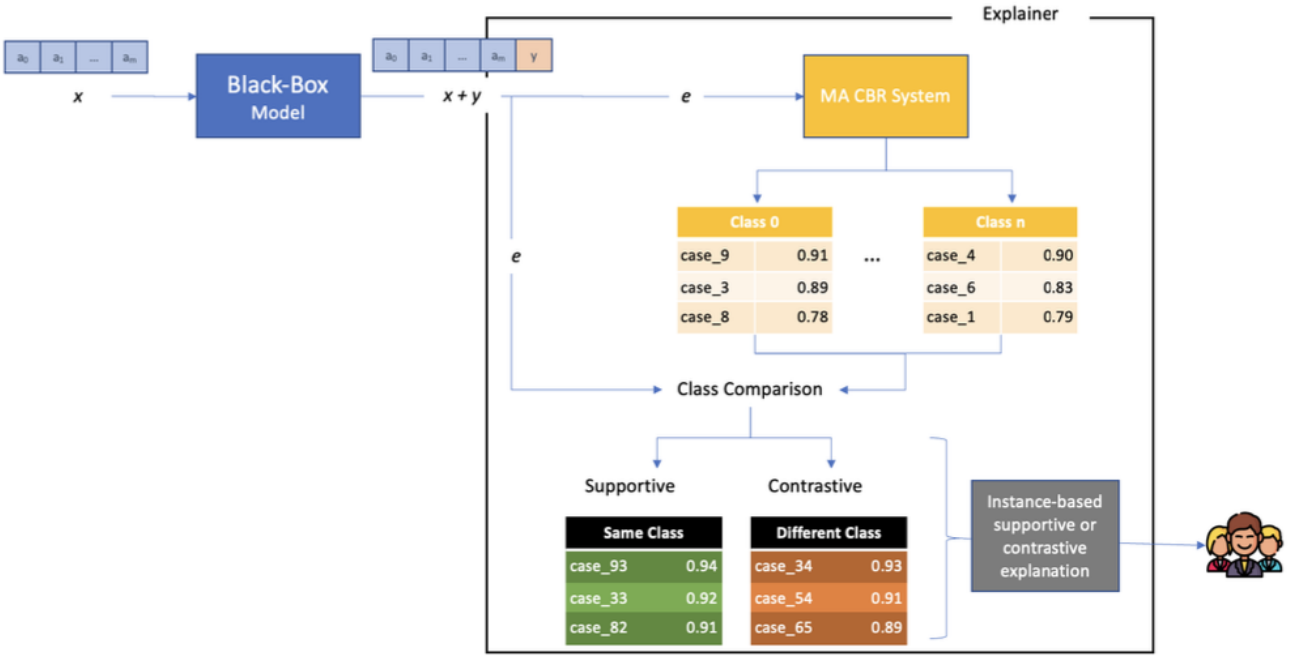
- Data level
- Similarity measurement level







Paper2 - A Twin XCBR System Using Supportive and Contrastive Explanations



Paper2 - A Twin XCBR System Using Supportive and Contrastive Explanations

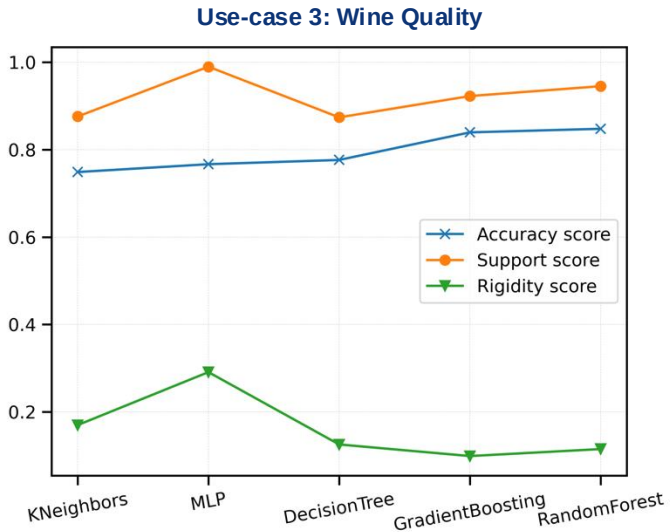
Use-case 1: Depression Screening	Use-case 2: App Usage	Use-case 3: Wine Quality
Real-world data	Real-world data	Open data
3 classes	3 classes	3 classes
105 instances	230 instances	4898 instances
102 features	26 features	12 features
Unbalanced	Unbalanced	Unbalanced
Data level DK	Similarity level DK	No DK

	accuracy	support	rigidity
Use-case 1	0.24	0.476	0.984
Use-case 2	0.58	0.696	0.199
Use-case 3	0.77	0.9897	0.2197

$$support = \frac{|S|}{n}$$

$$rigidity = |1 - \frac{support}{acc}|$$

n : Number of test instances
|*S*| : Number of supported predictions
acc : Accuracy score

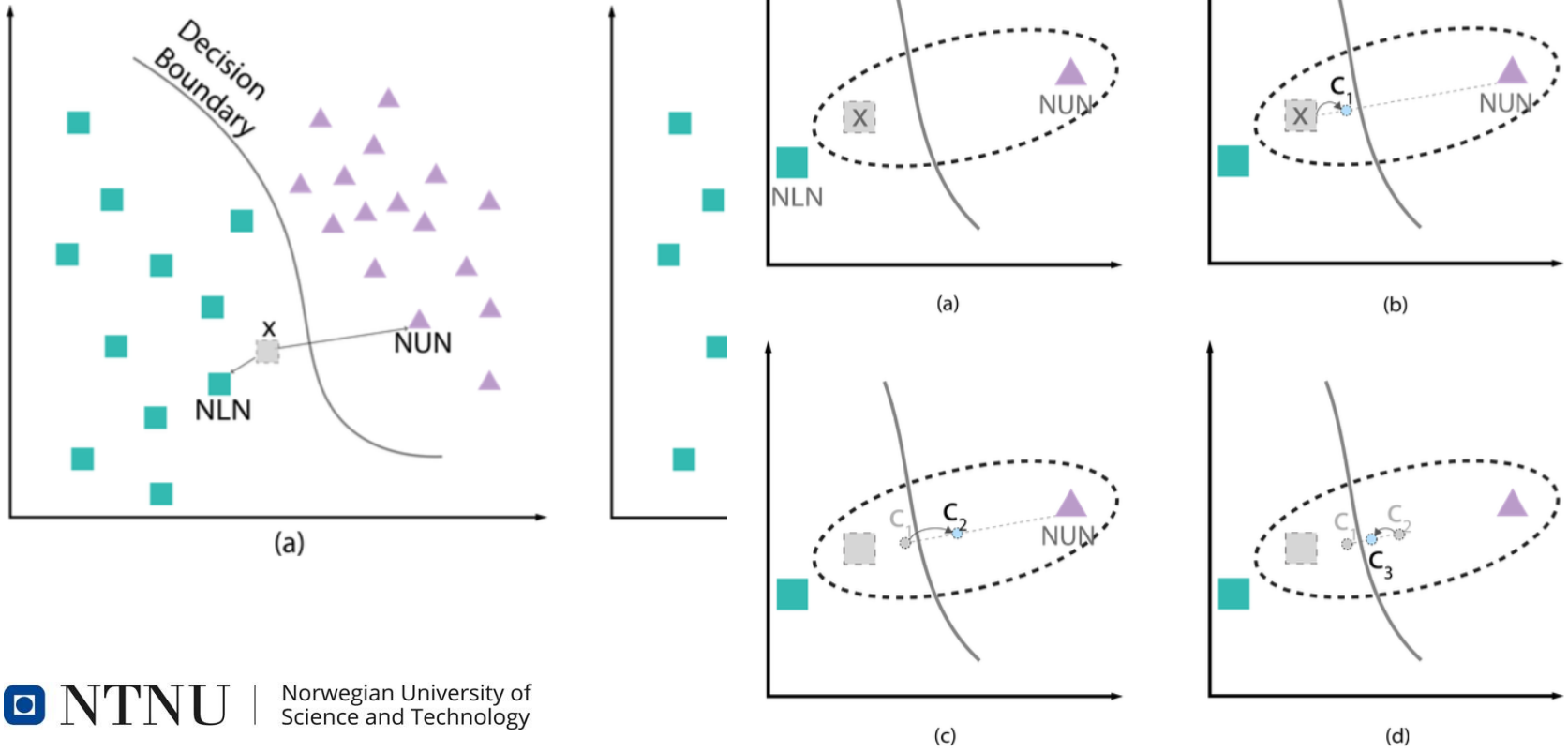


NTNU | Norwegian University of Science and Technology

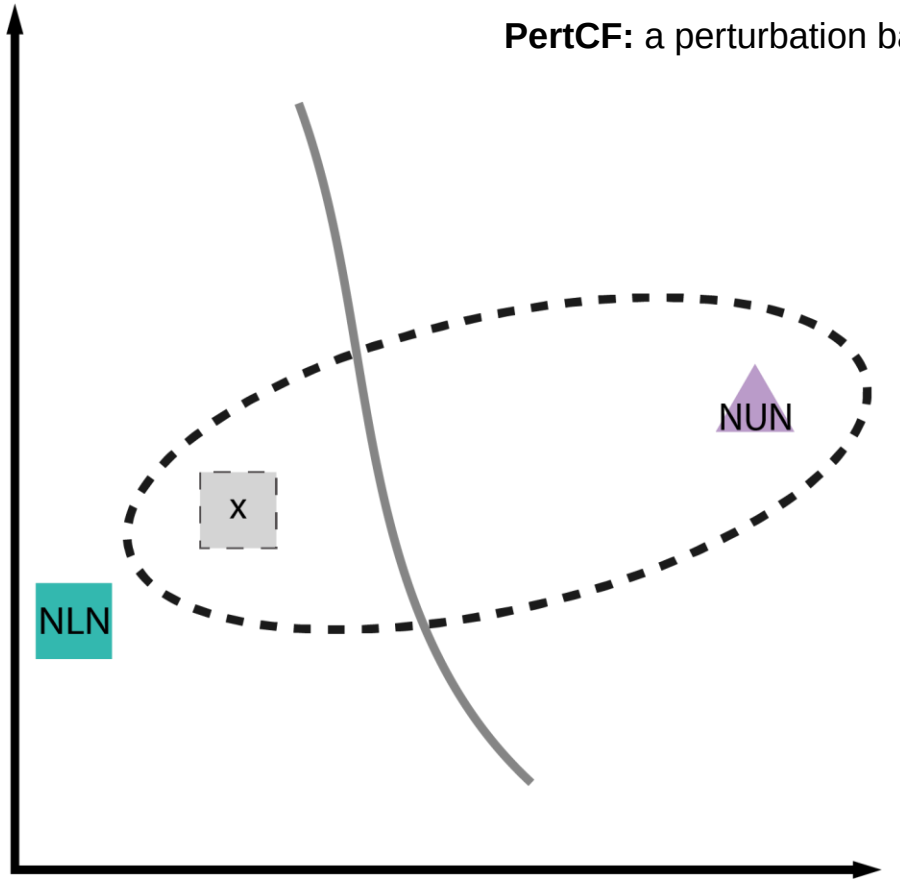
Paper 2: Contributions

Research Questions		
	RQ1: Domain Knowledge and Design	
	RQ2: Generating plausible explanations	RQ2.1: High-quality counterfactuals
		RQ2.2: Multimodel counterfactuals
	RQ3: Improving comprehensibility	RQ3.1: Instance-based explanations
		RQ3.2: Evaluating explanations

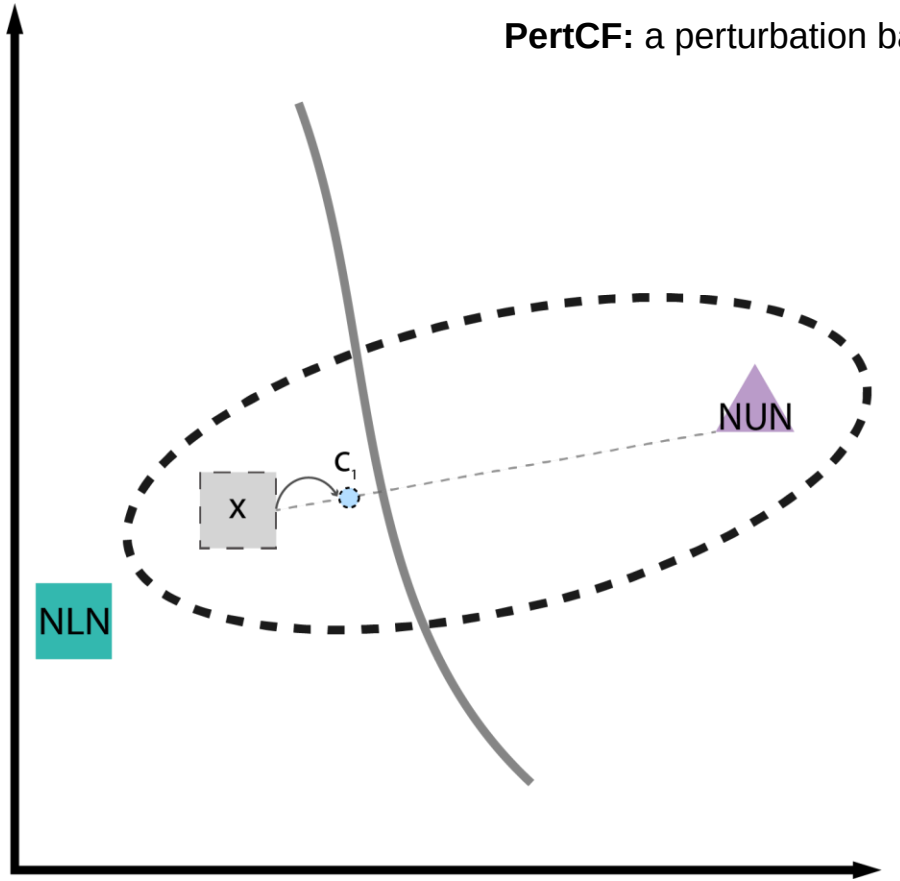
Paper3 - PertCF: A Perturbation-Based Counterfactual Generation Approach



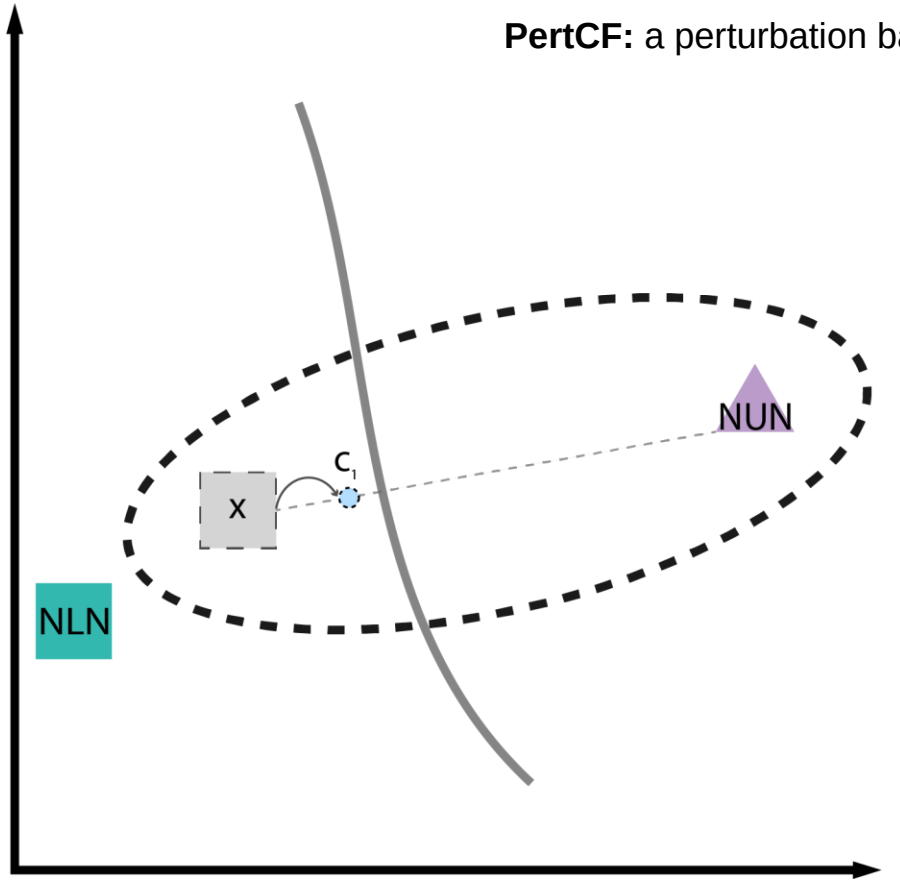
PertCF: a perturbation based CF generation approach



PertCF: a perturbation based CF generation approach

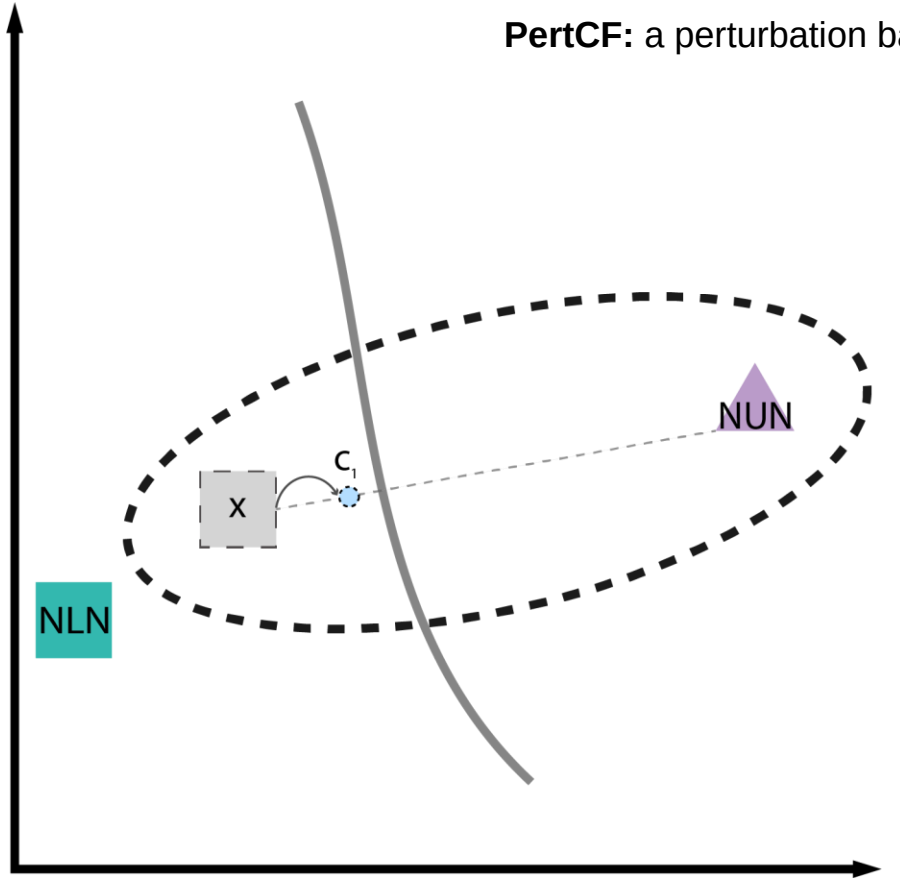


PertCF: a perturbation based CF generation approach



- Perturb '*x*' to generate candidate c_1
- SHAP (SHapley Additive exPlanations)

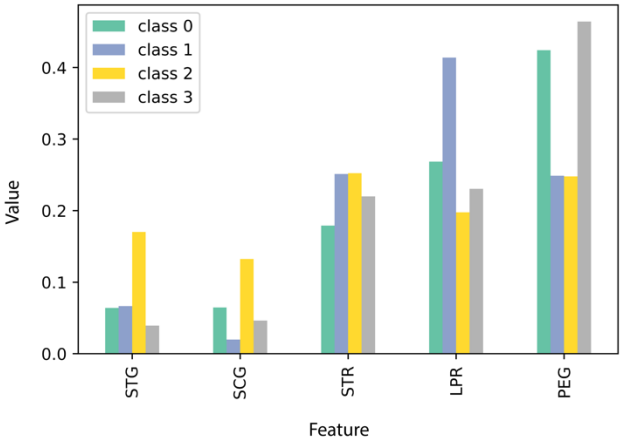
PertCF: a perturbation based CF generation approach



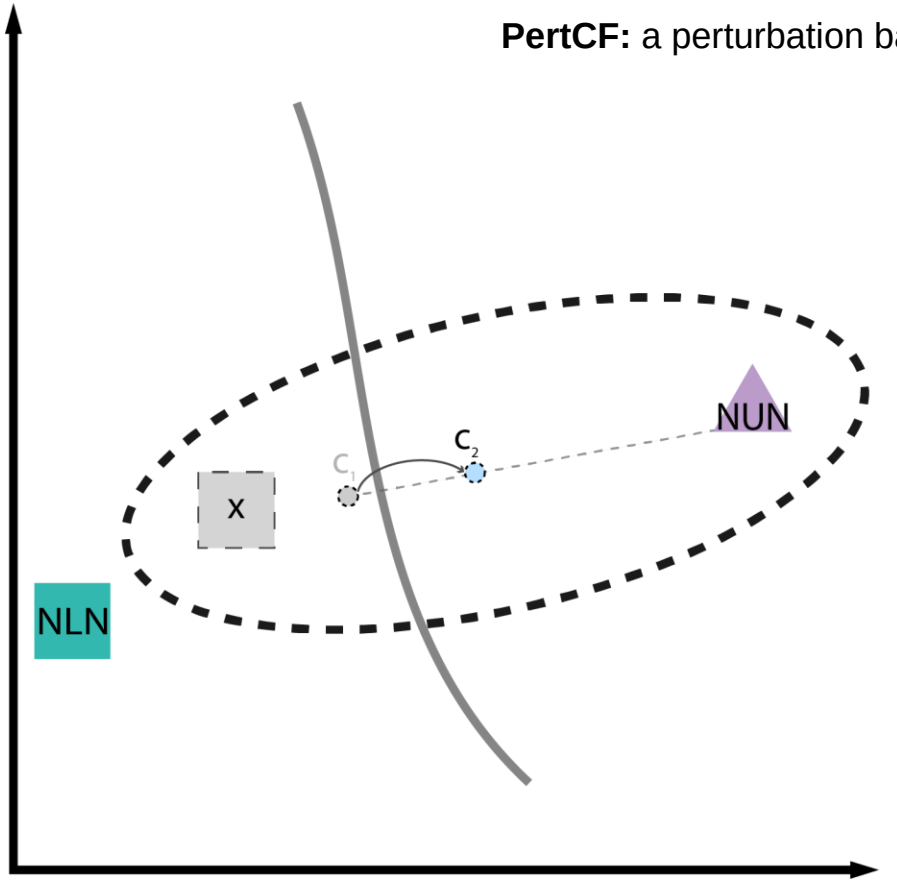
- Perturb '*x*' to generate candidate c_1
- SHAP (SHapley Additive exPlanations)

$c_i = <c_{i1}, c_{i2}, \cdots, c_{im}>$
$c_{ia} = s_a + shap\_target_a * dist(t_a, s_a)$

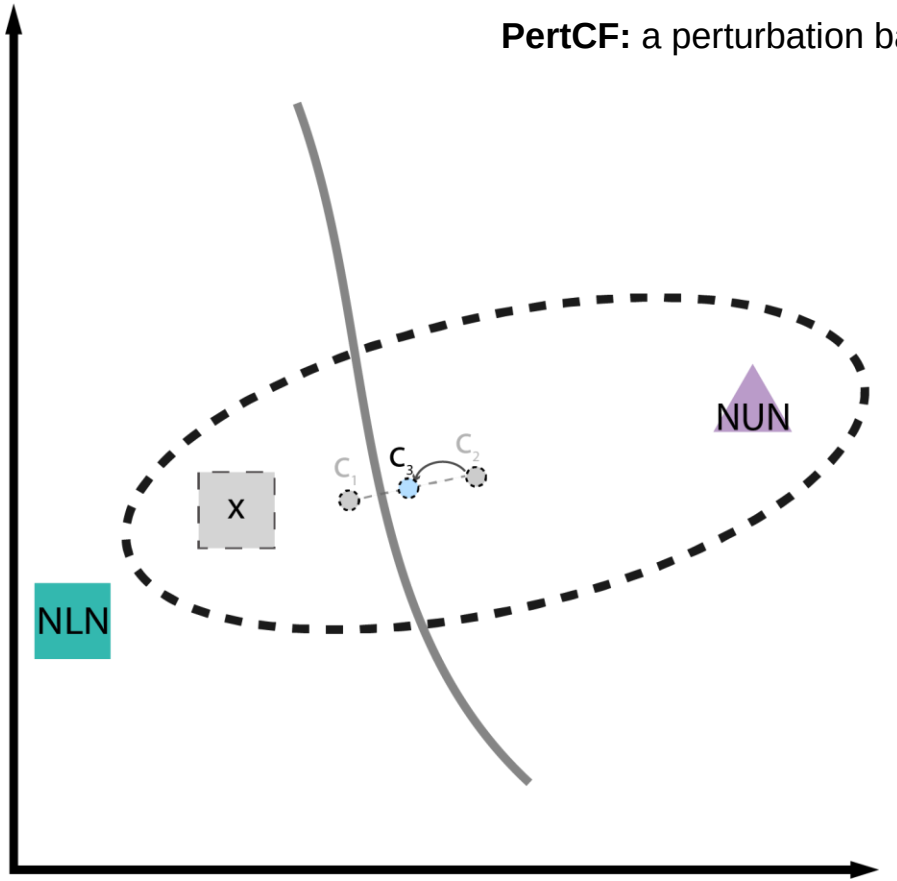
s: source
t: target



PertCF: a perturbation based CF generation approach

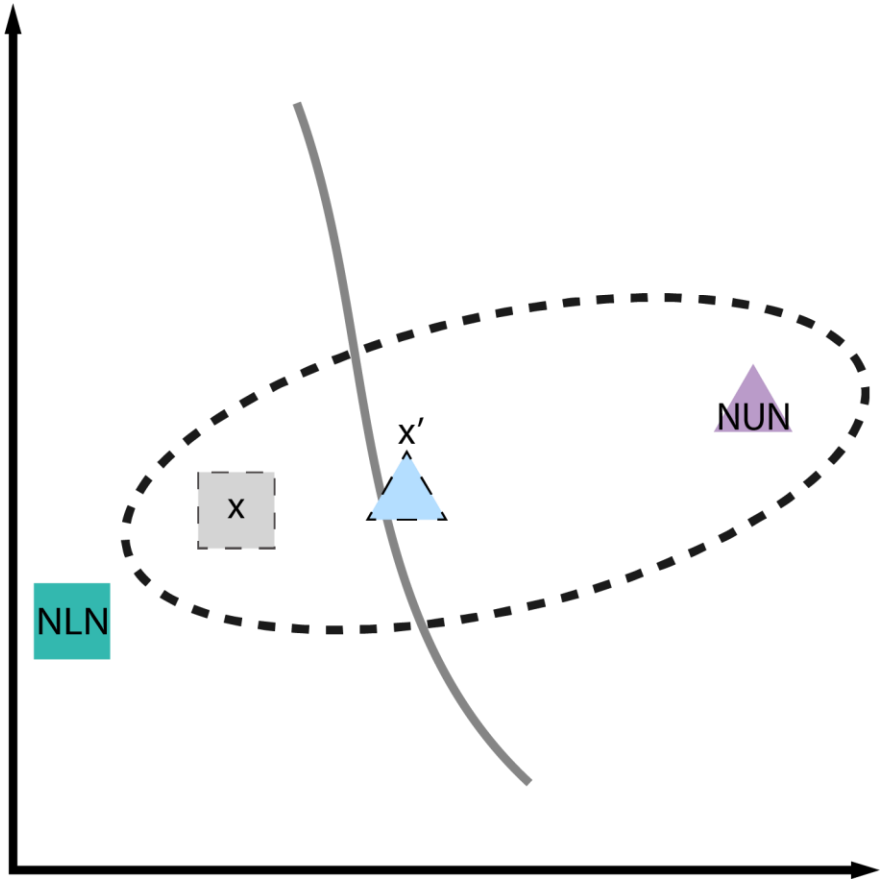


PertCF: a perturbation based CF generation approach



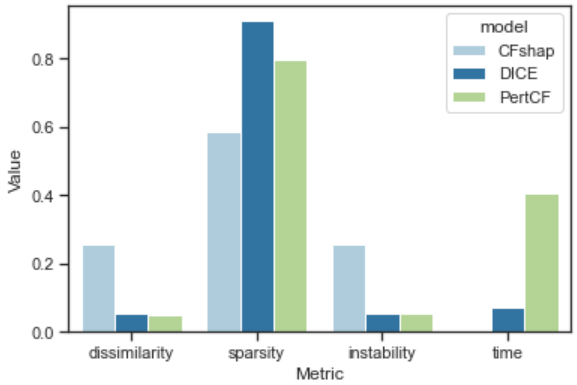
Termination criteria:

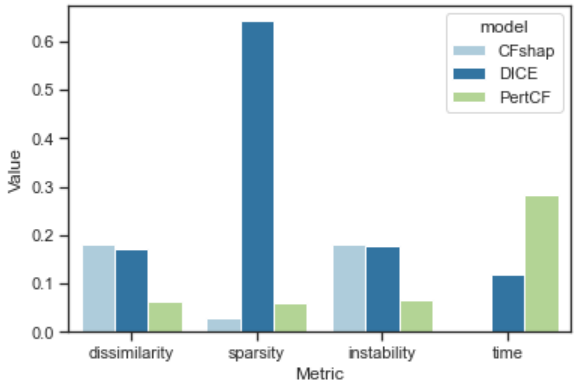
1. Number of iterations
2. Distance between the last two candidate


x′
NUN
x
NLN

Paper3 - PertCF

	Dissim.	Sparsity	Instability	Time
		South German Credit		
DICE	0.055744	**0.911053**	0.056013	0.073553
CFshap	0.255488	0.584211	0.255525	**0.005750**
PertCF	**0.051702**	0.798246	**0.051817**	0.406937
		User Knowledge Modeling		
DICE	0.172721	**0.642276**	0.176896	0.117947
CFshap	0.179172	0.029268	0.180618	**0.001123**
PertCF	**0.063630**	0.058537	**0.066436**	0.282734





Paper 3: Contributions

Research Questions		
	RQ1: Domain Knowledge and Design	
	RQ2: Generating plausible explanations	RQ2.1: High-quality counterfactuals
		RQ2.2: Multimodel counterfactuals
	RQ3: Improving comprehensibility	RQ3.1: Instance-based explanations
		RQ3.2: Evaluating explanations

Paper4 - Evaluation of Instance-based Explanations: An In-depth Analysis of Counterfactual Evaluation Metrics, Challenges, and the CEval Toolkit

- **Motivation:**
  - Confusion in Literature
  - Benchmarking Needs
  - Lack of Open-Source Tools

- **Contributions:**
  - **CEval Toolkit:** An open-source platform for evaluating and optimizing instance-based explanations.
  - **Customizable Framework:** Adaptable for different datasets and user needs.
  - **Focus on Optimization**: Integrates methods to enhance explanation quality.

- **Impact:**
  - Provides clarity in evaluating instance-based explanations.
  - Bridges technical evaluation with user-centric goals.
  - Serves as a benchmark for future XAI systems.

Metrics

$$validity = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{z_i \neq y} \mathbb{1}_{z_i = f(x_i')}$$

$$proximity = \frac{1}{m}\sum_{i=1}^{m} dist(x, x_i')$$

$$sparsity = \frac{1}{m\rho}\sum_{i=1}^{m}\sum_{j=1}^{\rho} \mathbb{1}_{a_j'^i \neq a_j}$$

$$number\_of\_explanations = |e|$$

$$diversity\_dpp = det(K)$$

$$lcc = \frac{u_z}{l} \qquad diversity_{lcc} = lcc * diversity\_dpp$$

$$yNN = \frac{1}{mk}\sum_{i=1}^{m}\sum_{\psi_j \in kNN(x_i')} \mathbb{1}_{z_i = f(\psi_j)}$$

$$feasibility = \frac{1}{mk}\sum_{i=1}^{m}\sum_{\psi_j \in kNN(x_i')}^{k} dist(x_i', \psi_j)$$

$$kNLN\_distance = \frac{1}{mk}\sum_{i=1}^{m}\sum_{\vartheta_j \in kNLN(x_i')}^{k} dist(x_i', \vartheta_j)$$

$$realtive\_distance = \frac{1}{m}\sum_{i=1}^{m} \frac{dist(x_i', x)}{dist(NUN(x), x)}$$

$$redundancy = \frac{1}{m|S|}\sum_{i=1}^{m}\sum_{s \in S} \mathbb{1}_{f(\tau'(x_i', x, s)) = z_i}$$

$$robustness = \frac{1}{m}\sum_{i=1}^{m} dist(e_i, explain(P(x)))$$

$$plausibility = \frac{1}{m}\sum_{i=1}^{m} \frac{dist(x_i', NLN(x_i'))}{dist(NLN(x_i'), NUN(NLN(x_i')))}$$

$$coverage = \frac{1}{|X|}\sum_{x \in X} \mathbb{1}_{|e|>0}, \; e = explain(x)$$

$$rigidity = \left|1 - \frac{supp}{acc}\right|$$

discriminative power

vulnerability

computational complexity

constraint violation

TABLE 1. Unified Notation for the Rest of the Article

Symbol	Description
x	An instance to explain, $x =< a_1, ..., a_\rho >$.
ρ	The number of attributes.
X	Set of instances to explain, $x \in X$.
$f()$	Prediction function.
y	Prediction result for x, $y = f(x)$
$explain()$	Explanation function, returns a set of counterfactuals for an instance.
e	Set of counterfactuals for x, $e = explain(x)$ and $e = \{e_1, ..., e_m\}$.
m	Number of counterfactuals that are provided for x.
e_i	$e_i = \{x_i', z_i\}$, $e_i \in e$.
x_i'	$x_i' =< a_1'^i, ..., a_\rho'^i >$
z_i	$z_i = f(x_i')$
u_z	The number of unique class labels for the counterfactuals generated for x.
l	The number of unique class labels.
R_j	Range of j^{th} attribute.
k	Number of neighbours.
$kNN()$	$kNN(x)$ finds k Nearest Neighbours to x.
$kNLN()$	$kNLN(x)$ finds k Nearest Like Neighbours to x.
$NUN()$	Nearest Unlike Neighbour, $NUN(x)$ returns the nearest neighbour which is labelled different than x.
$dist()^*$	Distance function, quantifies the distance between two instances.
$P()^*$	Perturbation function, $P(x)$ returns perturbed/corrupted version of x, x^p.
$\tau()$	$\tau(x_i', x, j)$ copies j^{th} attribute of x to j^{th} attribute of x_i' and returns x_i'.
A	$A = \{1, 2, ..., \rho\}$
$[A]^l$	All subsets of A that have l elements. $[A]^2 = \{\{a, b\} : a, b \in A, a \neq b\}$
S	All possible subsets (power set) of A without $\emptyset$ and A (super set). $S = \bigcup_{l=1}^{\rho-1} [A]^l$
$\|S\|$	$\|S\| = 2^\rho - 2$
$\tau'()$	$\tau'(x_i', x, s)$ copies s attributes of x to s attribute of x_i' and returns x_i'.

* These functions can be implemented in different ways.

TABLE 2. Evaluation metrics for instance-based explanations

	Applicable				Requires		Short Description
	Generated	Existed	Single	Multi	Data	Model	
Validity [1]	x	x	x	x	-	x	Whether the decision was altered.
Proximity [1]	x	x	x	x	-	-	Mean of feature-wise distance between the instance and its counterfactuals.
Sparsity [1]	x	x	x	x	-	-	Mean number of altered features between the instance and its counterfactuals.
Number of counterfactuals [1]	x	x	-	x	-	-	Number of counterfactuals generated for an instance.
Diversity [1]	x	x	-	x	-	-	Mean proximity between counterfactuals.
Diversity_lcc [1]	x	x	-	x	x	-	Diversity with class coverage coefficient.
yNN [1]	x	x	x	x	x	x	Amount of support counterfactuals receive from positively classified background data.
Feasibility [1]	x	x	x	x	x	-	Mean proximity of the counterfactuals to their nearest observations in the background data.
kNLN Distance [1]	x	x	x	x	x	-	Mean distance of counterfactuals to their k-NLN.
Relative Distance [1]	x	-	x	x	x	-	The ratio of the mean distance between the instance and counterfactuals to the mean distance between the instance and its NUN.
Redundancy	x	x	x	x	-	x	Mean count of unnecessary feature changes.
Robustness	x	-	x	x	x	x	Mean proximity between the explanation of the instance and the explanation of a slightly corrupted version of the instance.
Plausibility	x	-	x	x	x	-	The degree of credibility in the context.
Discriminative Power	x	x	-	x	x	-	The ability to differentiate two distinct classes through a naive approach.
Vulnerability	x	-	x	x	x	-	The extent of susceptibility to manipulations.
Complexity	x	x	x	x	x	x	Cost for the explainer to generate a single explanation.
Constraints	x	x	x	x	-	-	Mean count of violated pre-defined constraints.
Coverage	x	x	x	x	-	-	The ability to generate valid counterfactuals across various types of instances.

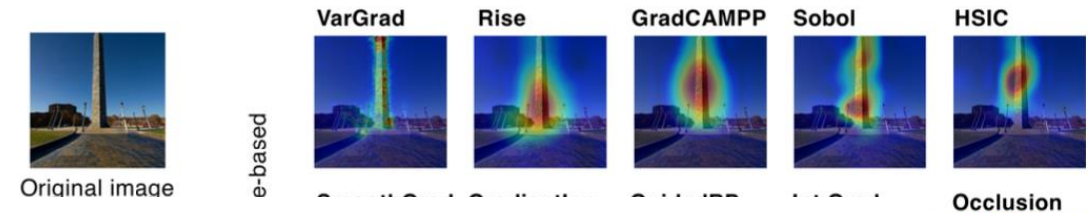
[1]This metric is implemented in the package.

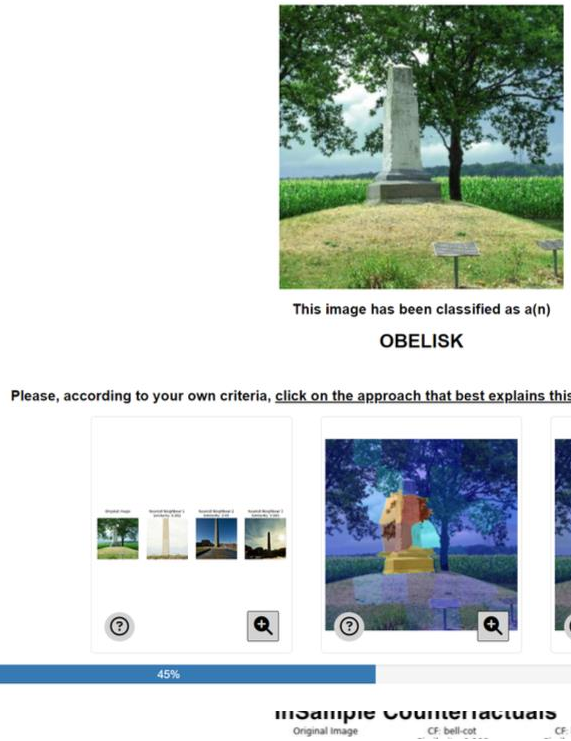
Paper 4: Contributions

Research Questions		
	RQ1: Domain Knowledge and Design	
	RQ2: Generating plausible explanations	RQ2.1: High-quality counterfactuals
		RQ2.2: Multimodel counterfactuals
	RQ3: Improving comprehensibility	RQ3.1: Instance-based explanations
		RQ3.2: Evaluating explanations

Paper5 - An Empirical Analysis of User Preferences Regarding XAI metrics

- **Motivation:**
  - Understanding user preferences is crucial for designing effective XAI systems.
  - A need to bridge the gap between technical metrics and user-centric needs.
- **Aim:**
  - Conduct an empirical study to evaluate which XAI metrics align with user priorities.
  - Provide actionable insights for improving XAI frameworks.

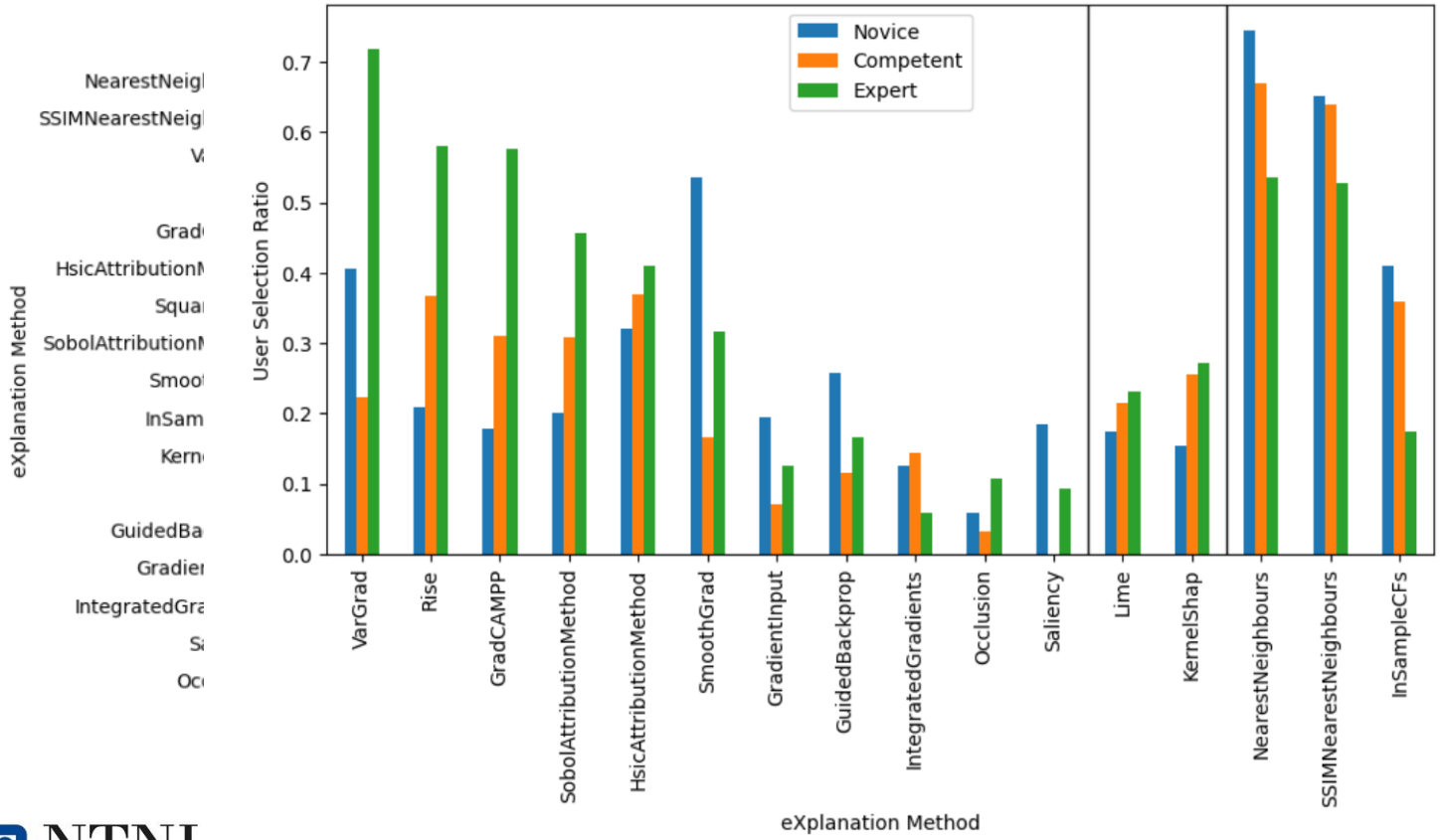




Key Contributions:

1. **Empirical Analysis**: Quantifies user preferences for common XAI metrics (e.g., interpretability, fidelity, robustness).
2. **Insights on Trade-offs**: Highlights trade-offs between technical metrics and user satisfaction.
3. **Guidelines for XAI Design**: Provides recommendations to align technical and user-centric goals.

Paper5


Novice
Competent
Expert
User Selection Ratio
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
VarGrad
Rise
GradCAMPP
SobolAttributionMethod
HsicAttributionMethod
SmoothGrad
GradientInput
GuidedBackprop
IntegratedGradients
Occlusion
Saliency
Lime
KernelShap
NearestNeighbours
SSIMNearestNeighbours
InSampleCFs
eXplanation Method

Paper5

- Some of the objective evaluation metrics align better with user preferences.
- Metrics such as **insertion** and **sparsity** are good predictors of user satisfaction (for this specific task).
- Novice users prefer instance-based explanations.
- Expert users prefer feature-based explanations.
- Explanation **representation MATTERS**.
- Objective + Subjective Evaluation

Paper 5: Contributions

Research Questions		
	RQ1: Domain Knowledge and Design	
	RQ2: Generating plausible explanations	RQ2.1: High-quality counterfactuals
		RQ2.2: Multimodel counterfactuals
	RQ3: Improving comprehensibility	RQ3.1: Instance-based explanations
		RQ3.2: Evaluating explanations

Contributions - overall

Research Questions		P-I	P-II	P-III	P-IV	P-V
RQ 1: Domain Knowledge and Design		*	*			*
RQ 2: Generating Plausible Explanations	RQ 2.1: High-quality Counterfactuals			*	*	
	RQ 2.2: Multimodal Counterfactuals					*
RQ 3: Improving Comprehensibility	RQ 3.1: Instance-Based Explanations	*	*		*	*
	RQ 3.2: Evaluating Explanations				*	*

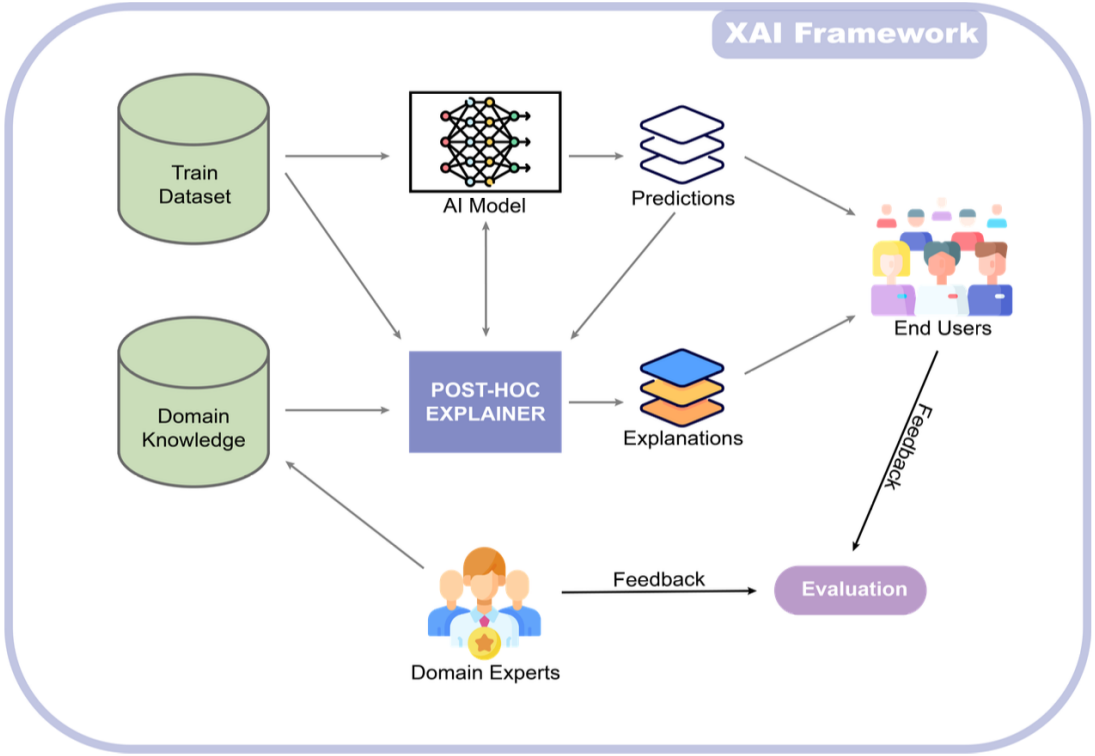
Limitations

- **Human-in-the-Loop Challenges:**
  - Reliance on human evaluation introduces biases and subjectivity.
  - Balancing objective metrics with user feedback remains complex.
- **Multi-Modality Constraints:**
  - Current methods focus on single or limited number of data types (e.g., tabular, image).
  - Limited application to real-world multimodal data (e.g., text, audio, video).
- **Scalability Issues:** High computational demands restrict application to large datasets and real-time systems.
- **Objective Mismatch:** Misalignment between stakeholder needs (users, developers, regulators) and explanation goals.
- **Customization Gaps:** Current metrics do not align with user preferences for tailored, flexible explanations.
- **Empirical Validation:** Limited generalizability to diverse domains and broader user groups.

Future Directions

Scalability and Efficiency	Develop algorithms for scalable, real-time explanations with minimal computational overhead.
Multimodal Explainability	Extend methods to handle diverse data types and cross-modal relationships.
Automated and Adaptive Explanations	Create systems that adapt explanations to user expertise and preferences in real time.
Human-Centered XAI	Incorporate participatory design approaches and longitudinal studies to refine explanations.
Integration with Emerging Trends	Adapt XAI methods for transformers, LLMs, and conversational AI applications.

Conclusion


XAI Framework
Train Dataset
AI Model
Predictions
End Users
Domain Knowledge
POST-HOC EXPLAINER
Explanations
Feedback
Feedback
Evaluation
Domain Experts

Conclusion

Advancing Instance-Based Explanations:	• Developed innovative methods for generating and evaluating instance-based explanations, particularly counterfactuals.
Bridging Human and Machine Understanding:	• Focused on human-centered design to align AI explanations with user needs, fostering trust and usability.
Comprehensive Evaluation Frameworks:	• Proposed and implemented metrics and toolkits like CEval to assess and optimize XAI techniques effectively.
Contribution to XAI Literature:	• Enhanced understanding of trade-offs between technical metrics and user preferences, providing actionable insights for future research.

Thank you for your attention ☺

Betül Bayrak

betul.bayrak@ntnu.no

NTNU | Norwegian University of Science and Technology