Counterfactual and adversarial explanations

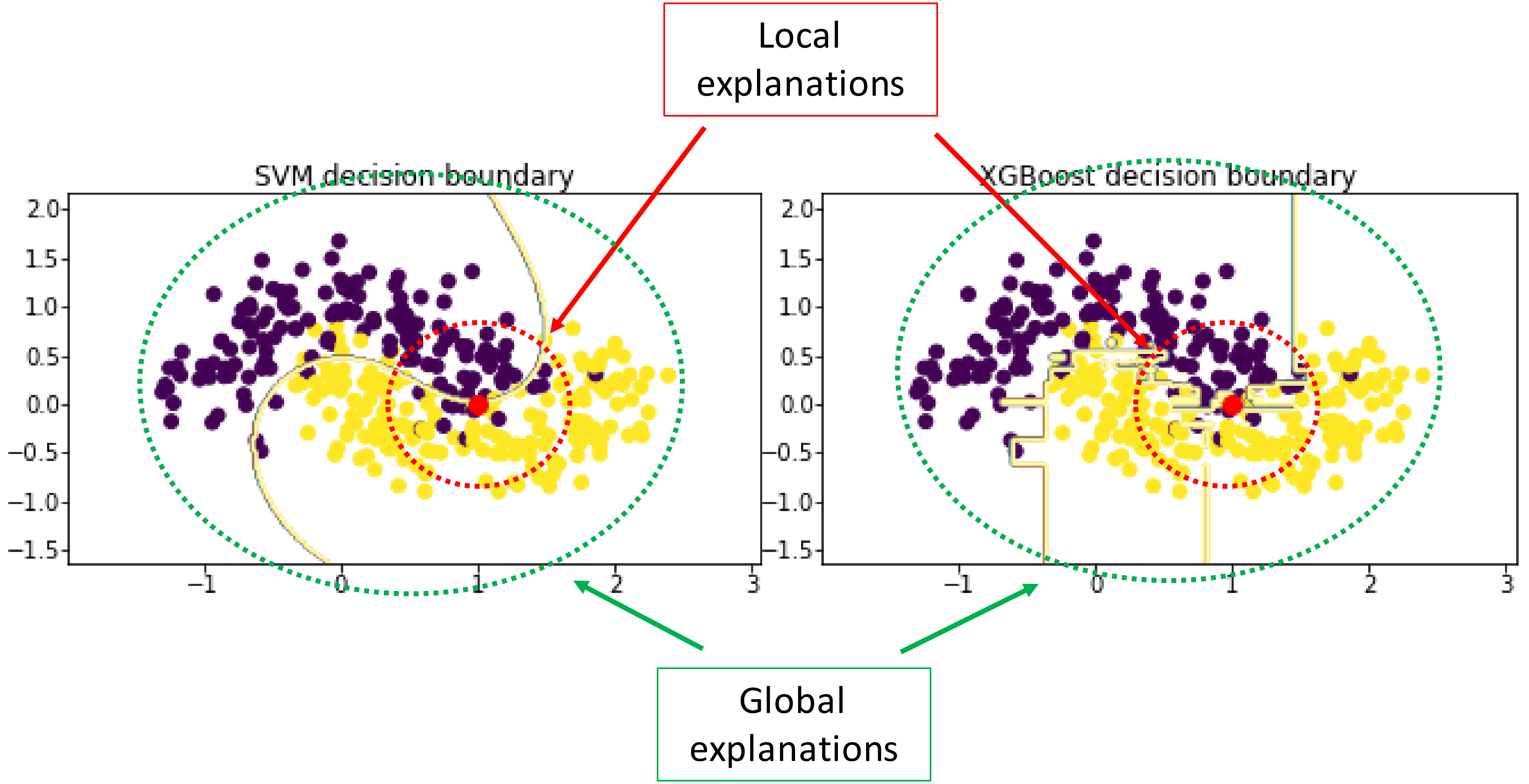
Szymon Bobek

Jagiellonian University
2023

JAGIELLONIAN UNIVERSITY
IN KRAKÓW

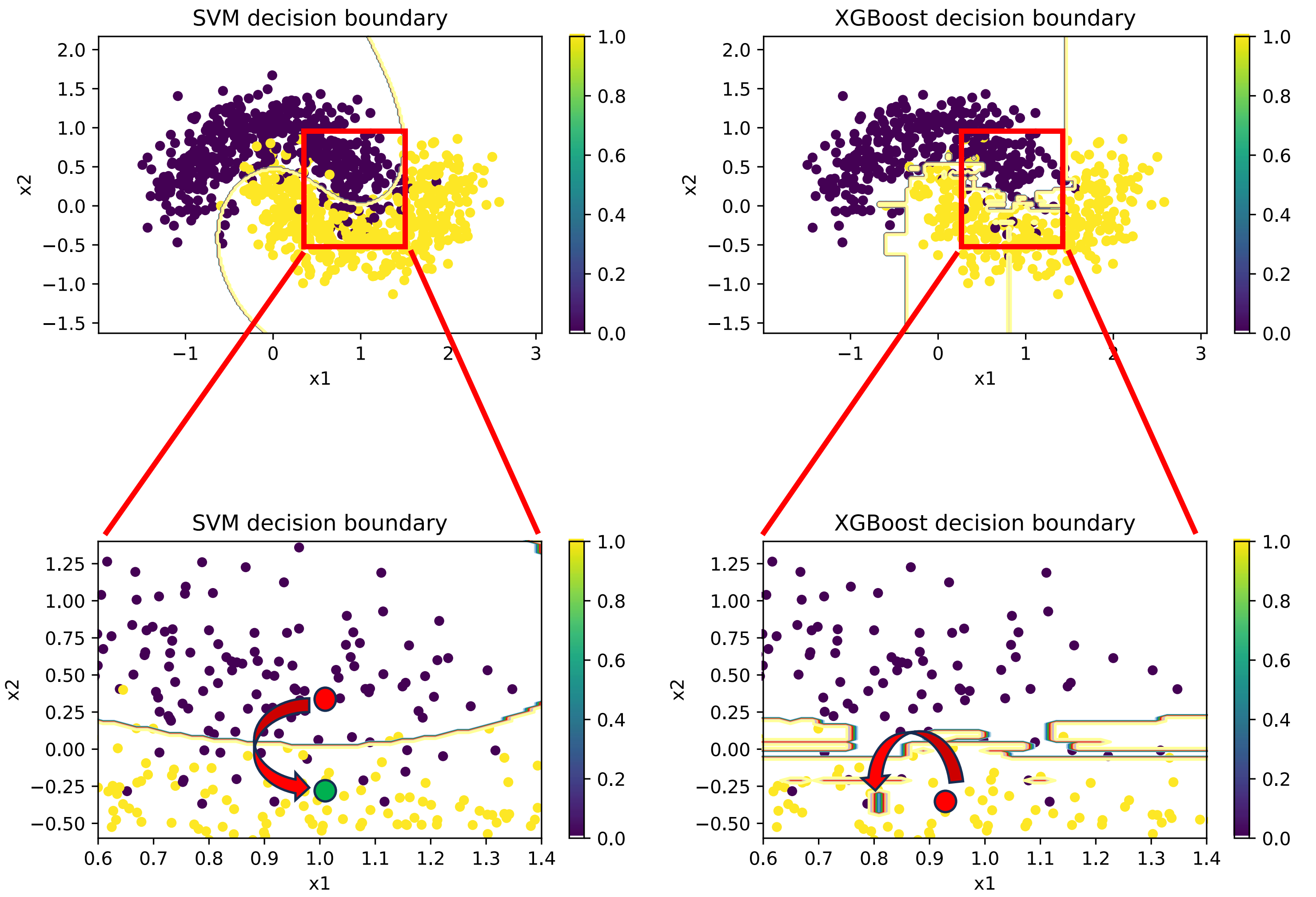
https://geist.re

Local vs Global explanations (or both?)



Counterfactual examples

Counterfactual explanations (CF)



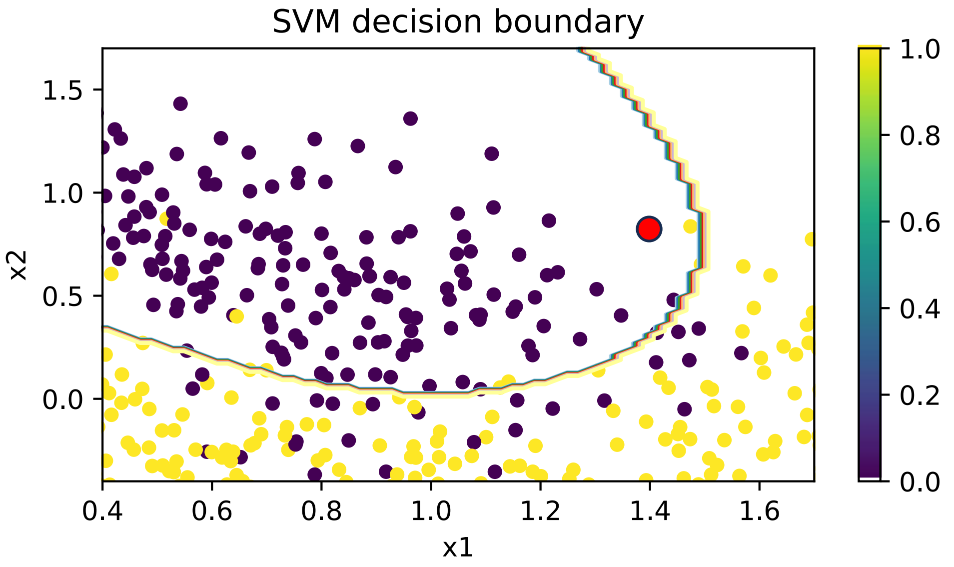
- In this approach we focus on explaining the instance by showing what should be changed in its features to make the model change its decission
- Ususally, the change is required to be *minimal*
- Additional constraints can be added to preserve CF properties
- CFs are example-based explanations, and can be considered counter
- Counterfactuals vs Contrastive

$$y = b(x)$$

$$b(x) = b(x')$$

Properties of CFs

- Validity (Fidelity)
- Minimality (Sparsity)
- Similarity
- Plausability (distribution-aware)
- Discriminative (very subjective)
- Actionability (can not change age)
- Causality (changes of one feature may imply changes in others)
- Diversity (representing different valid options)

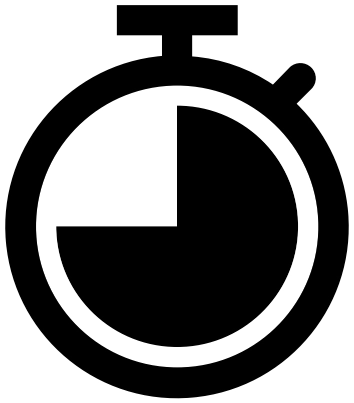


$$y = b(x)$$

$$b(x) = b(x')$$

Properties and types of CF explainer

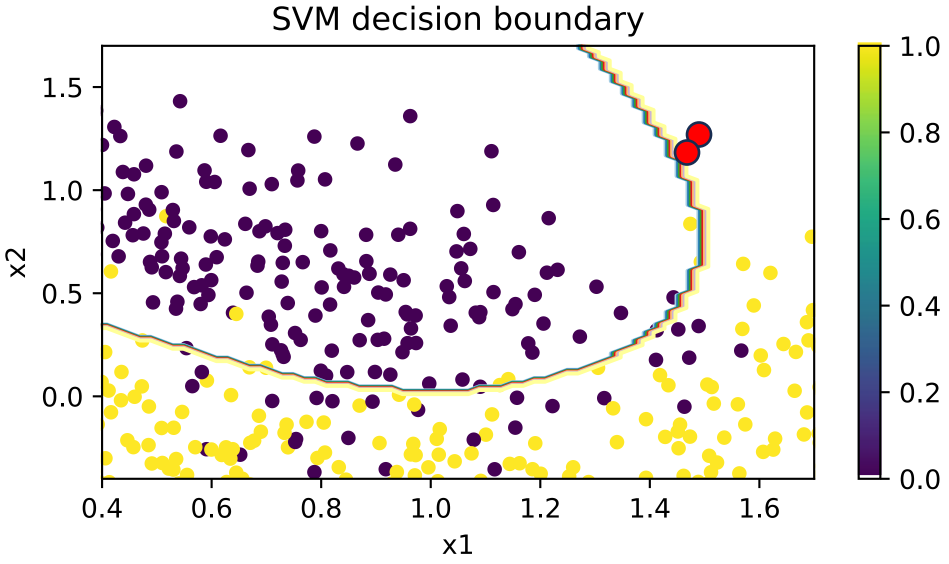
- Properties of Explainers
  - **Efficiency**
  - Stability (similar instances, similar CF)
  - Fairness (related to causality and plausability)
- Types of search
  - Optimization
  - Heuristic Search
  - Instance-Based
  - Decision Tree
- Types of CF generaiton
  - Endogenous
  - Exogenous (majority)



Properties and types of CF explainer

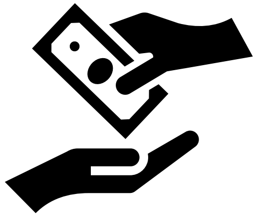
- Properties of Explainers
  - Efficiency
  - **Stability** (similar instances, similar CF)
  - Fairness (related to causality and plausability)
- Types of search
  - Optimization
  - Heuristic Search
  - Instance-Based
  - Decision Tree
- Types of CF generaiton
  - Endogenous
  - Exogenous (majority)

$$\hat{L}(\Phi^{e\rightarrow m}, X) = \max_{x_j \in N_\epsilon(x_i)} \frac{||x_i - x_j||_2}{||\Phi_i^{e\rightarrow m} - \Phi_j^{e\rightarrow m}||_2 + 1}$$

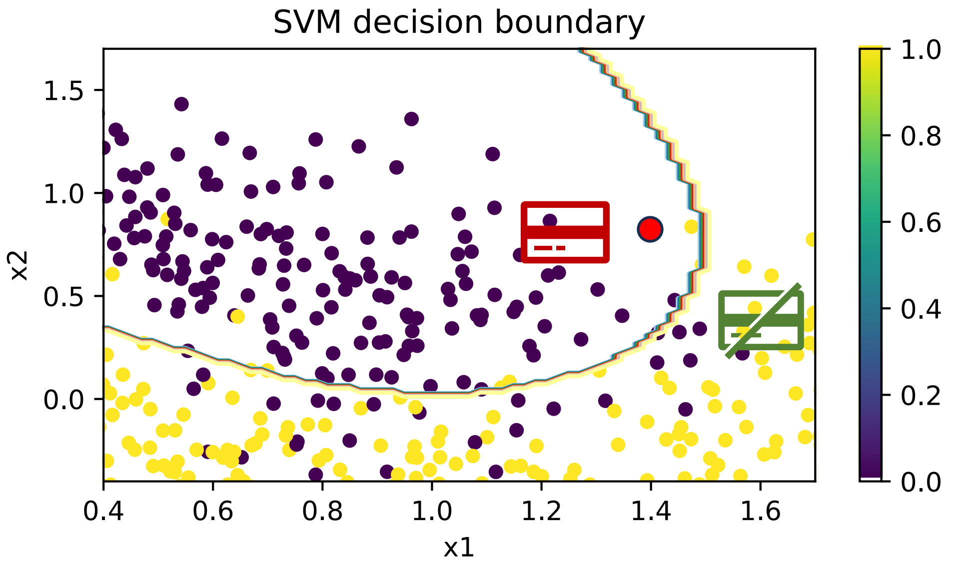


Properties and types of CF explainer

- Properties of Explainers
  - Efficiency
  - Stability (similar instances, similar CF)
  - **Fairness** (related to causality and plausability)
- Types of search
  - Optimization
  - Heuristic Search
  - Instance-Based
  - Decision Tree
- Types of CF generaiton
  - Endogenous
  - Exogenous (majority)

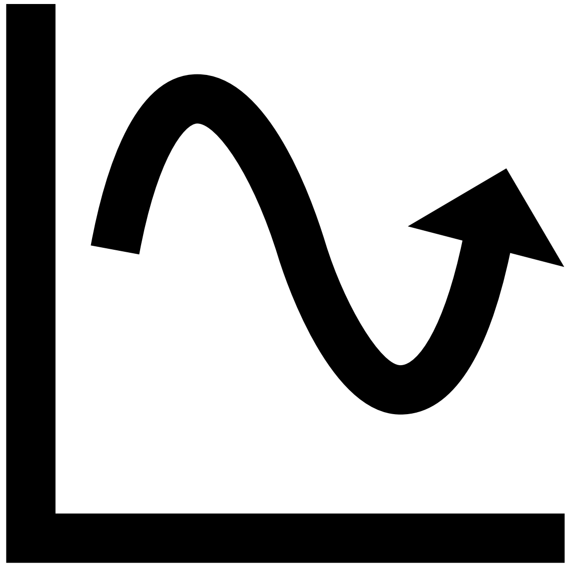


Changes in the model have reflection in real world



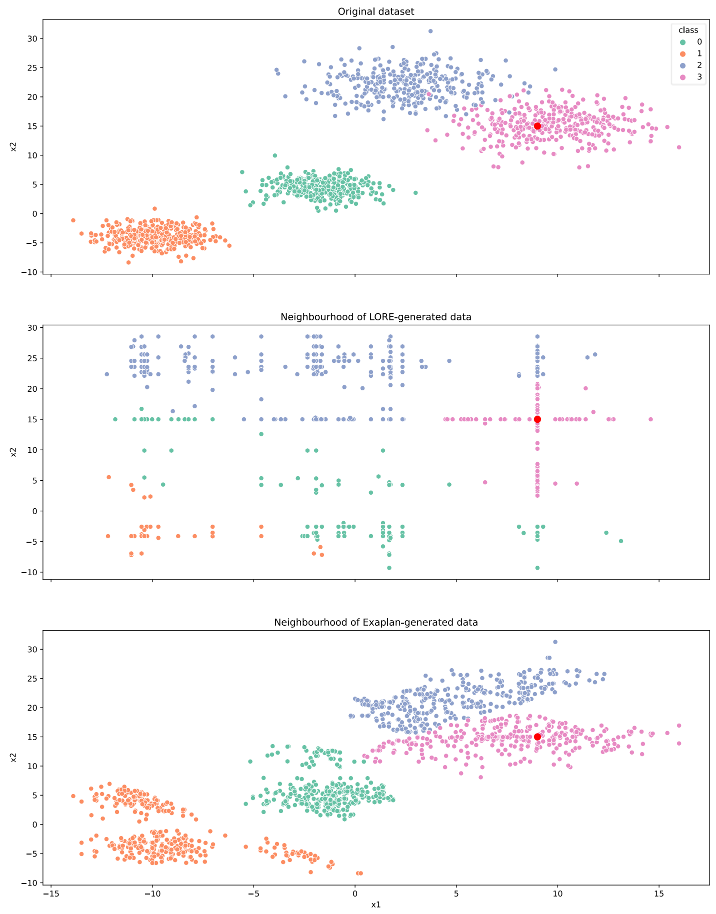
Properties and types of CF explainer

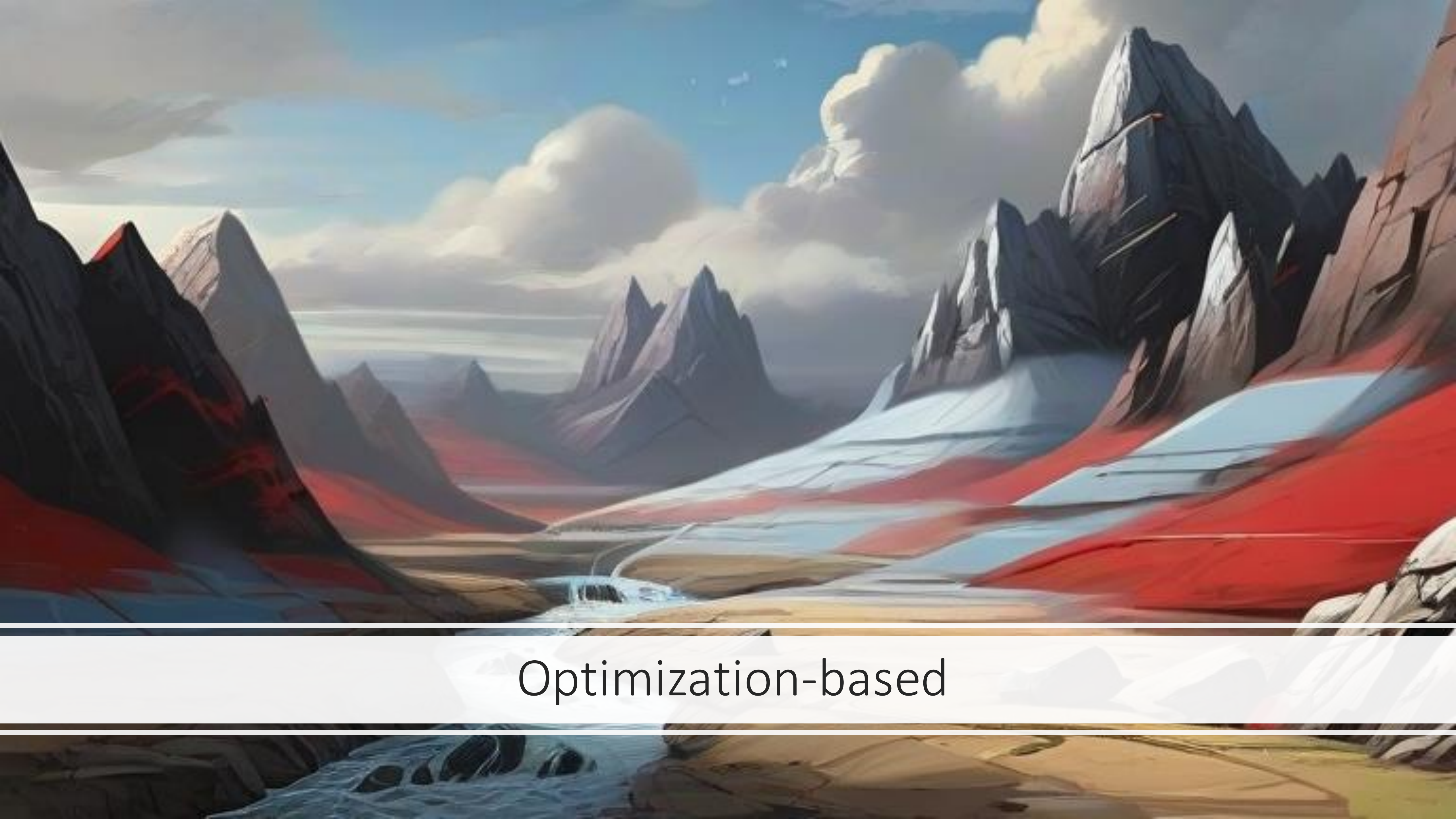
- Properties of Explainers
  - Efficiency
  - Stability (similar instances, similar CF)
  - Fairness (related to causality and plausability)
- **Types of search**
  - Optimization
  - Heuristic Search
  - Instance-Based
  - Decision Tree
- Types of CF generaiton
  - Endogenous
  - Exogenous (majority)



Properties and types of CF explainer

- Properties of Explainers
  - Efficiency
  - Stability (similar instances, similar CF)
  - Fairness (related to causality and plausability)
- Types of search
  - Optimization
  - Heuristic Search
  - Instance-Based
  - Decision Tree
- **Types of CF generaiton**
  - Endogenous
  - Exogenous (majority)





Optimization-based

Optimization-based techniques (BF)

- Brute force optimization is the simplest way of finding CF of an instance x over features F
- Generates all the possible variations of x with respect to any of the subsets in F (possibly limited to m')
- It replaces a feature value in x with any representative value from r for different subsets m'

$$m = |F|$$

$$O(\binom{|F|}{m'} \cdot m \cdot r)$$

F can be replaced with features that are actionable to reduce computations cost

Optimization-based techniques (WACH)

- WACH – the one of the first famous CF model
- It searches for CF by minimizing balance between difference between instances and their predictions
- The balance is modified by the lambda parameter
- In the original paper, the authors minimize the loss and maximize lambda at the same time

$$\lambda(b(x') - y')^2 + d(x, x')$$

$$|b(x') - y'| \leq \epsilon$$

$$\arg\min_{x'}\max_{\lambda} L(x, x', y', \lambda).$$

$$d(x, x') = \sum_{i}^{m} \frac{|x_i - x'_i|}{MAD_i}$$

$$\boxed{d(x, x') = \sum_{i}^{m} \frac{|x_i - x'_i|}{MAD_i}\theta_j}$$

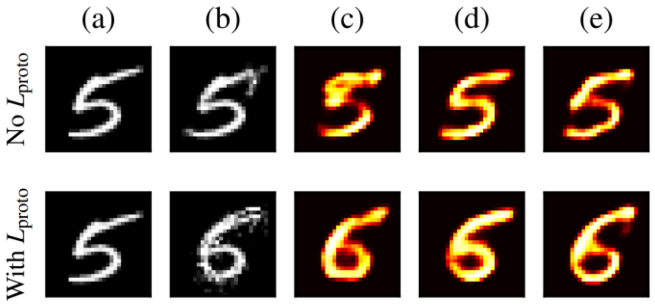
$$\mathrm{MAD} = \mathrm{median}(|X_i - \tilde{X}|)$$

Rory McGrath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. CoRR abs/1811.05245 (2018)

Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J. L. & TECH. 841 (2018).

Optimization-based techniques (CEM)

- CF is defined as $x' = x + \delta$
- Such that $b(x+\delta) \neq b(x)$
- The AE is used to assure plausability of CF
- CEM is using FISTA algorithm as an optimization backend
- CEGP is an Prototypical extension of CEM



$$\overbrace{||proto_j - ENC(x+\delta)||}^{\text{Distance of an encoded CF to latent prototype of opposite class}}$$

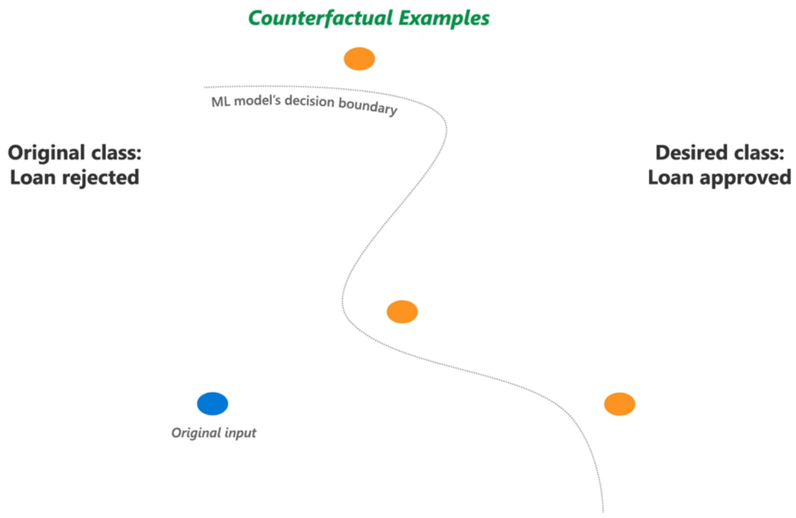
$$\underbrace{\alpha f(b(x), b(x+\delta))}_{\text{Difference in desired and acutal outcome}} + \overbrace{\beta||\delta||_1 + ||\delta||_2^2}^{\text{Elastic net regularizers}} + \gamma\underbrace{||x+\delta - AE(x+\delta)||_2^2}_{\text{Reconstruciton loss of perturbed sample}}$$

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: towards contrastive explanations with pertinent negatives. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Curran Associates Inc., Red Hook, NY, USA, 590–601.

Optimization-based techniques (DICE)

- It approaches CF generation under constraints of feasibility and diversity
- User can define mutable and immutable features
- It searches simultaneously for k CF, penalizing similar CFs

$$\arg\min_{x'_1,\ldots,x'_k} \frac{1}{k}\sum_{i=1}^{k} \max(0, 1 - y' logit(b(x'_i))) + \frac{\lambda_1}{k}\sum_{i=1}^{k} d(x'_i, x) - \lambda_2 div(x'_1, \ldots, x'_k)$$



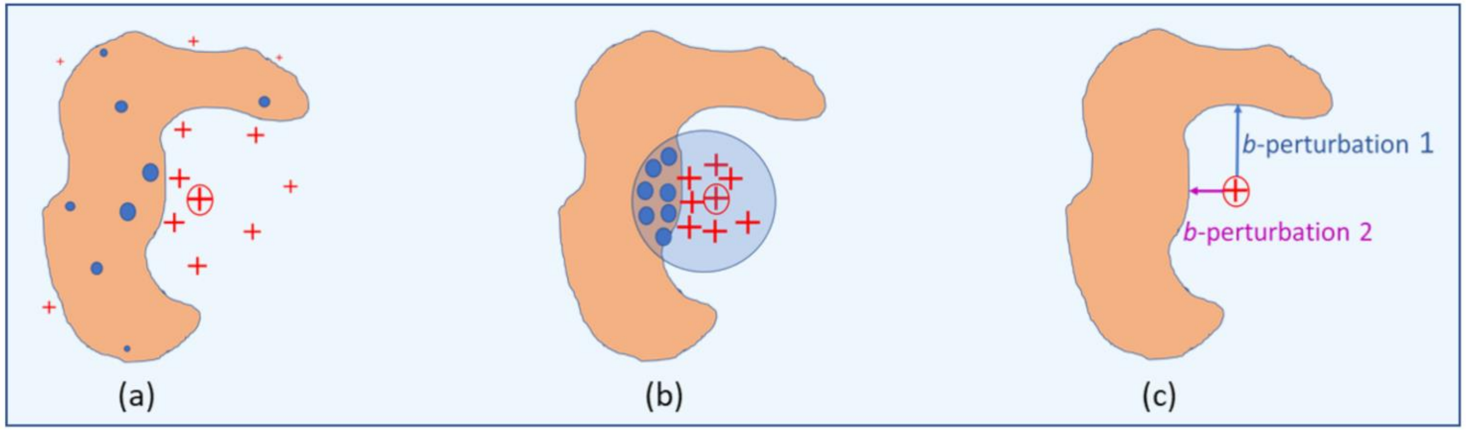
https://github.com/interpretml/DiCE

Mothilal RK, Sharma A, Tan C (2020) Explaining machine learning classifiers through diverse counterfactual explanations. In: FAT*'20: conference on fairness, accountability, and transparency, Barcelona, Spain, January 27–30, 2020, ACM, pp 607–617

Heuristic-based

Heuristic-based methods (CLEAR)

- CLEAR generates a random synthetic neighborhood around x
- It selects a small balanced sub-sample composed of instances at diversified level of probability from b(x) according to predefined parameters modeling the margins around the decision boundary.
- Then for each feature, finds a counterfactual instance varying only a feature with a brute force approach and extends the balanced neighborhood with them.
- Finally, it trains a local surrogate linear regressor r on the balanced neighborhood and estimates the counterfactual instances retrieved at the previous step.
- CLEAR returns as explanations the actual and estimated counterfactuals as well as the regressors unveiling the feature coefficients and the approximation error between b and r.



Standalone LIME cannot generate and evaluat realiably CFs

Create balanced neighbourhood

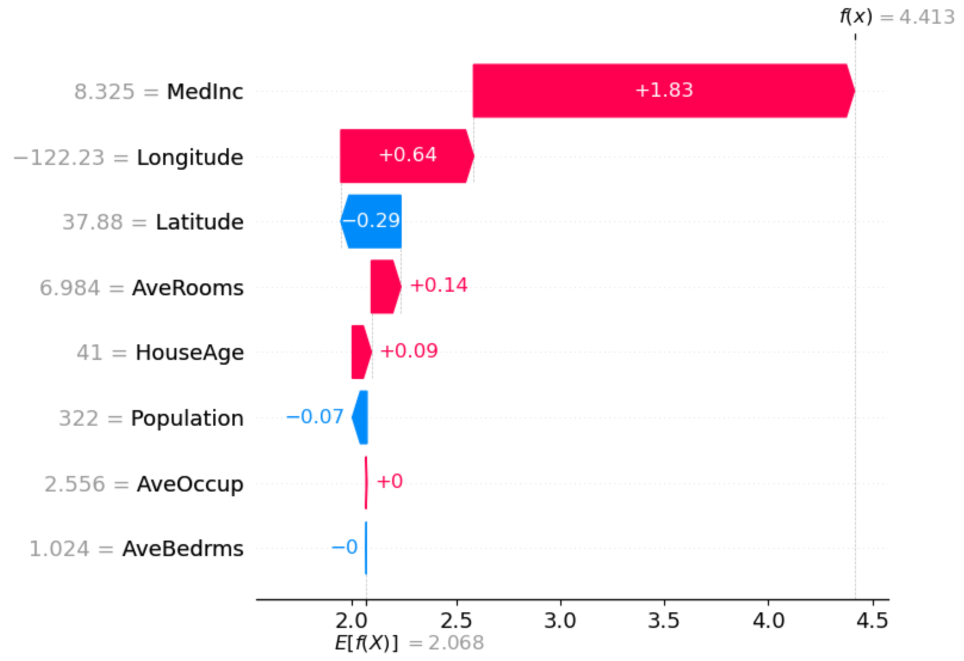
Extend the neighbourhood with brute-force found CFs

Use local regressor to evaluate CF

White A, d'Avila Garcez AS (2020) Measurable counterfactual local explanations for any classifier. In: ECAI 2020—24th European conference on artificial intelligence, 29 August–8 September 2020, Santiago de Compostela, Spain, August 29–September 8, 2020 - Including 10th conference on prestigious applications of artificial intelligence (PAIS 2020), IOS Press, Frontiers in Artificial Intelligence and Applications, vol 325, pp 2529–2535

Heuristic-based methods (CFSHAP)

- CFSHAP first estimates the Shapely values for each possible target class different from b(x).
- Then, it randomly generates synthetic neighbors of x by permuting x only on the features for which the Shapely values are negative with respect to the desired counterfactual class
- The counterfactuals are slected from these points

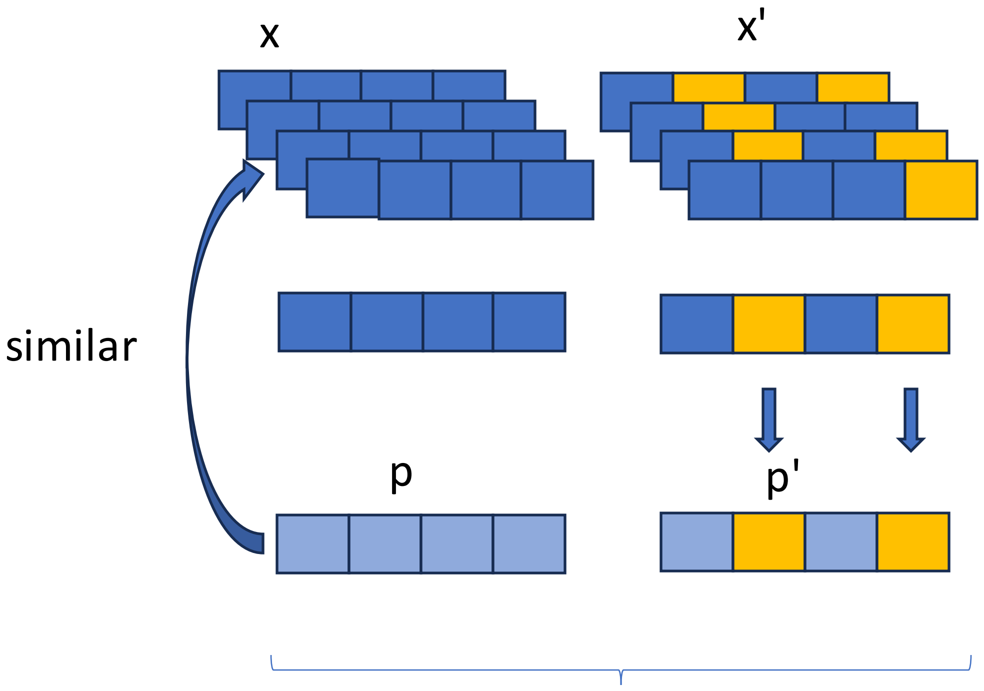


Rathi S (2019) Generating counterfactual and contrastive explanations using SHAP. CoRR arXiv:1906.09293

Instance based

Instance-based methods (CBCE)

- Create a dataset X=(x,x') containing pairs of simlar instances such that b(x) != b(x')
- When instance p is given, first a pair (x,x') from X is found such that p is most similar to x
- Then, p' is constructed by replacing values in p with values form x' that are different from x

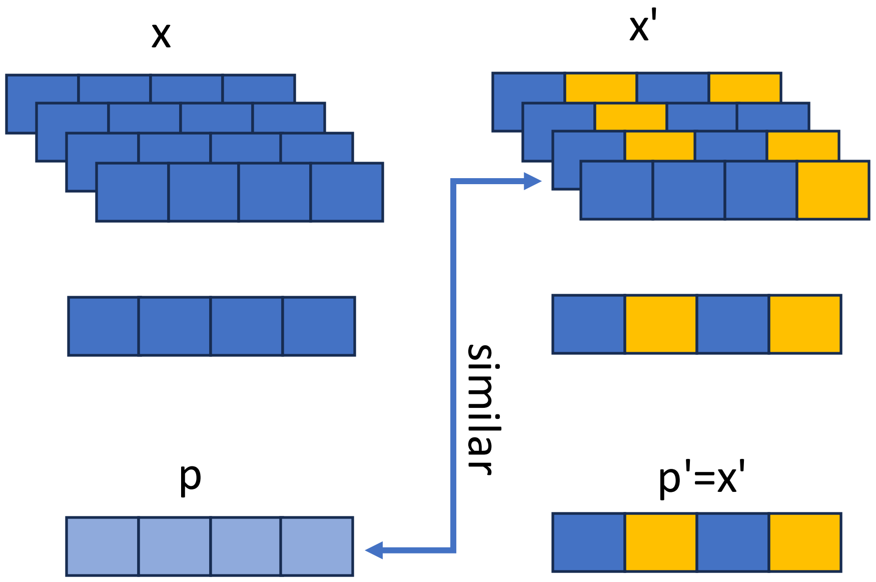


Create CF p' that is different form p in a same way as x' is different form x

Keane MT, Smyth B (2020) Good counterfactuals and where to find them: a case-based technique for generating counterfactuals for explainable AI (XAI). In: Case-based reasoning research and development—28th international conference, ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings, Springer, Lecture notes in computer science, vol 12311, pp 163–178

Instance-based methods (NNCE)

- Endogenous explainer
- It selects neares neighbour from a set of examples that have opposite class then p
- Computationally intensive
- Rather not diverse

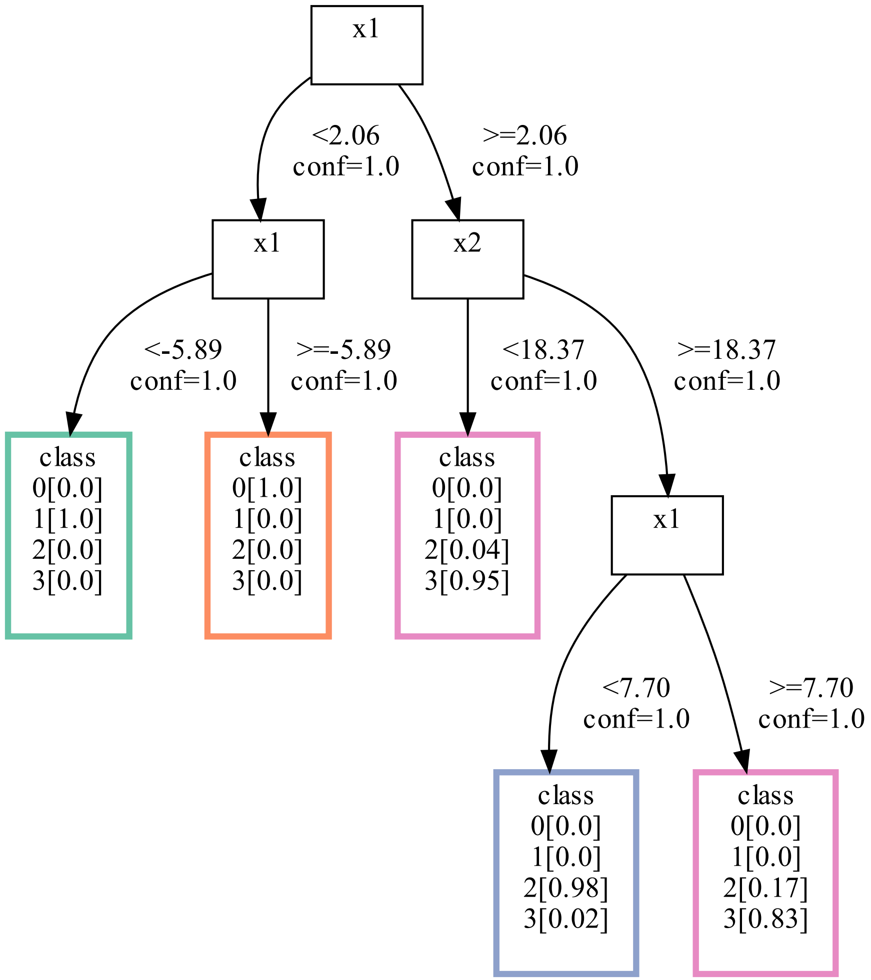


Shakhnarovich G, Darrell T, Indyk P (2008) Nearest-neighbor methods in learning and vision. IEEE Trans Neural Netw 19(2):377

Decision trees

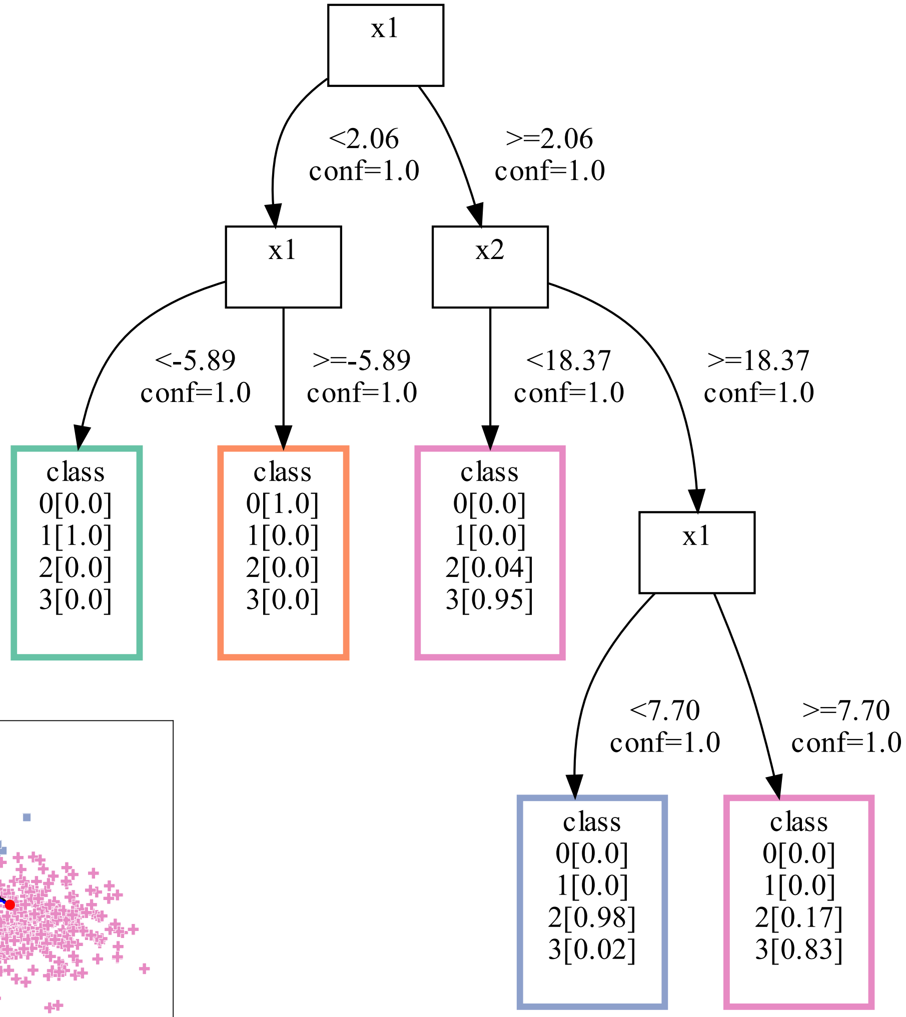
Decision trees (LORE)

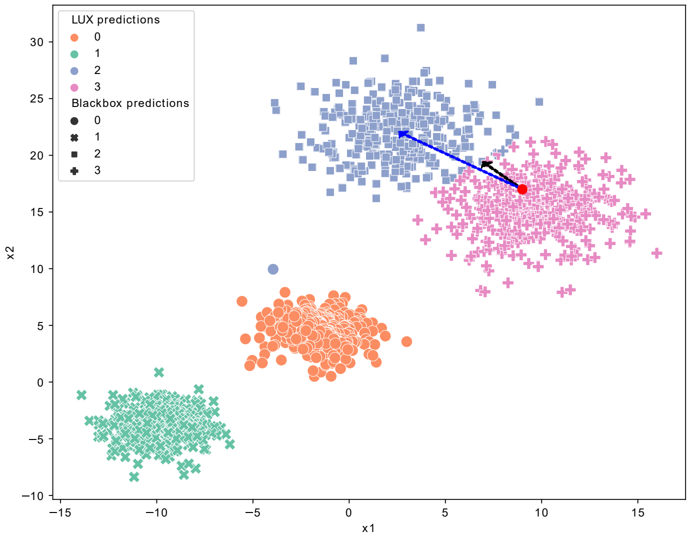
- Lore is a rule based explainer that uses decisiont ree as backend
- It focuses on generating explanations, but in paralel it allows for creating CFs
- The minimality is assured by the minimal split conditions in a treee that x is not satisfied

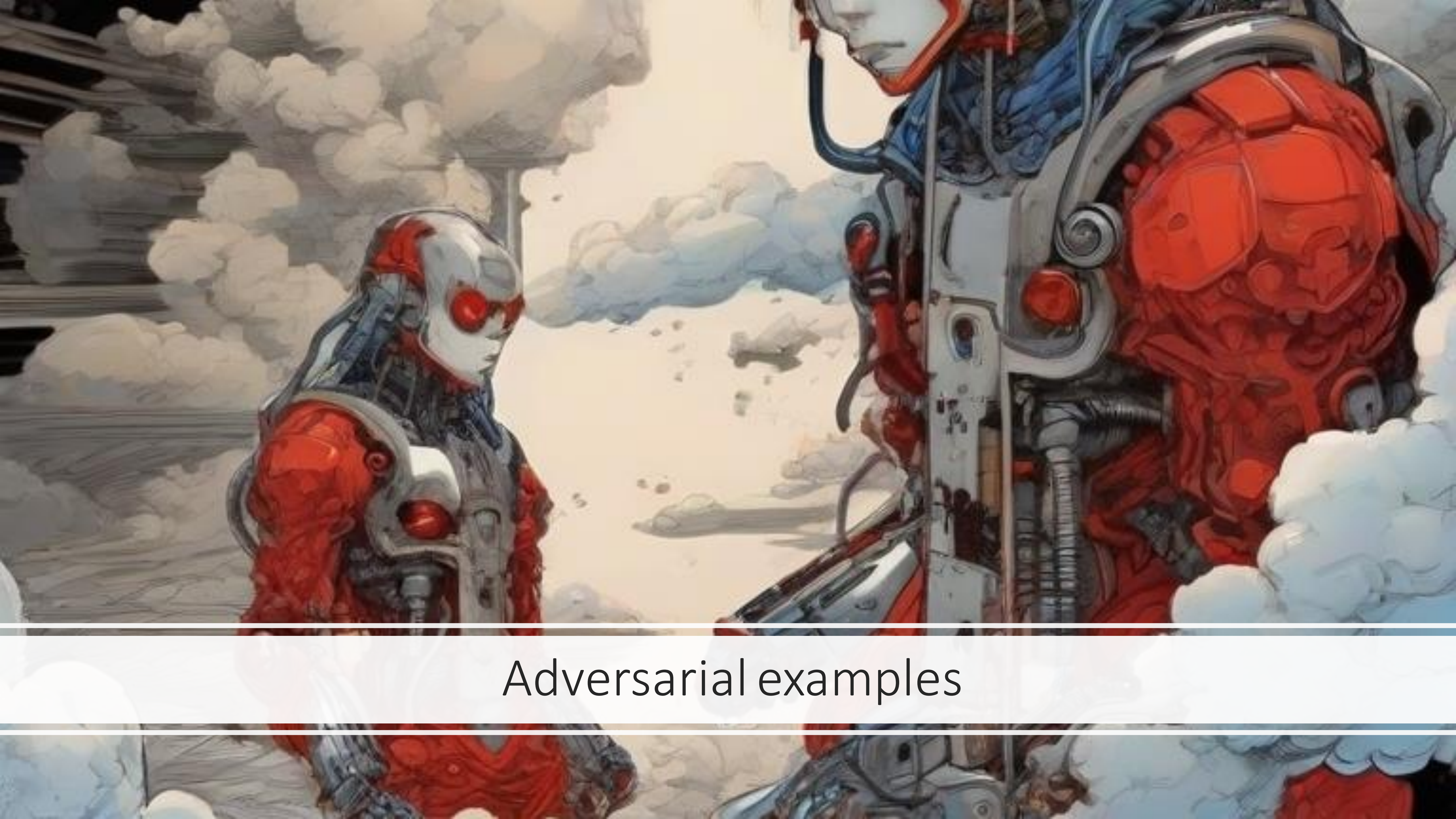


Decision trees (LUX)

- LUX, similarly like LORE generates CF as rules
- Additinally it allows to use background dataset to generate endogenous CFs
- The sparcity is defined in terms of distance to neareast or to medoid of cluster of opposite class





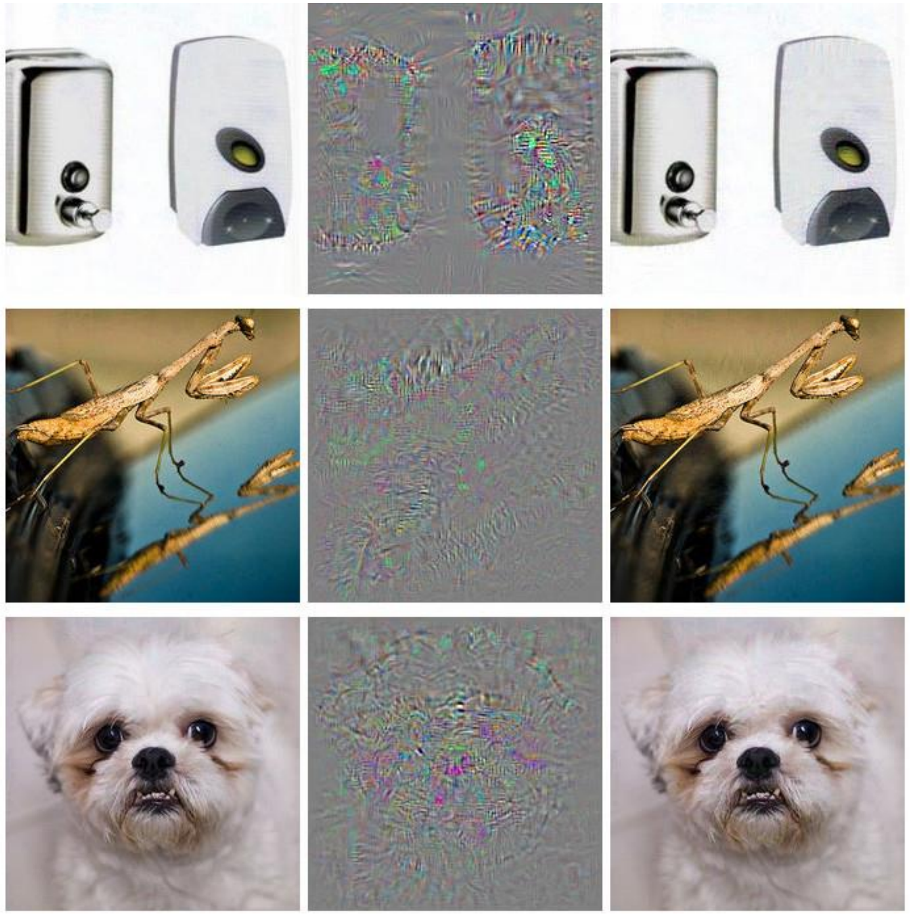


Adversarial examples

Adversarial attacks

- Adversarial attacks can be considered a malicious usage of CF
- It aims at finding the unseen or irrelevant (by human) modificaiton of intput to change deicison of the output

$$loss(\hat{f}(\underbrace{x+r}_{\text{Image and changes}}), l) + \underbrace{c \cdot |r|}_{\text{Changes balanced by c}}$$



Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

Thank you for your attention!

JAGIELLONIAN UNIVERSITY
IN KRAKÓW

https://geist.re