Counterfactual explanations

CSEP 590B: Explainable AI

Hugh Chen, Ian Covert & Su-In Lee
University of Washington

Course announcements

- For next week's class (5/24), we'll have two guest lectures:
  - 6:30 – 7:30 **James Zou (Stanford)**
  - 7:40 – 8:40 **Dan Weld (UW)**
  - 8:45 – 9:20 Paper discussion

- HW3 will be posted tomorrow
  - Due on 6/1 (two weeks)

Motivation

- **Previously:** feature importance, concept explanations, neuron interpretation

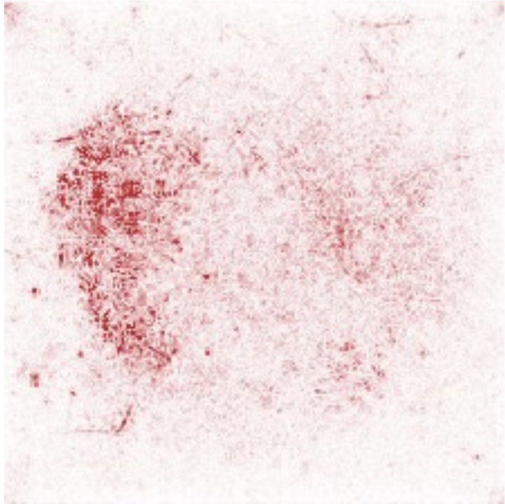
- **Today:** a new type of explanation for individual predictions
  - Not asking what's important to a prediction…
  - Instead asking: "how can we change it?"

Medical image example

Original image

Predicted: benign

Saliency map



Can we go beyond localization?

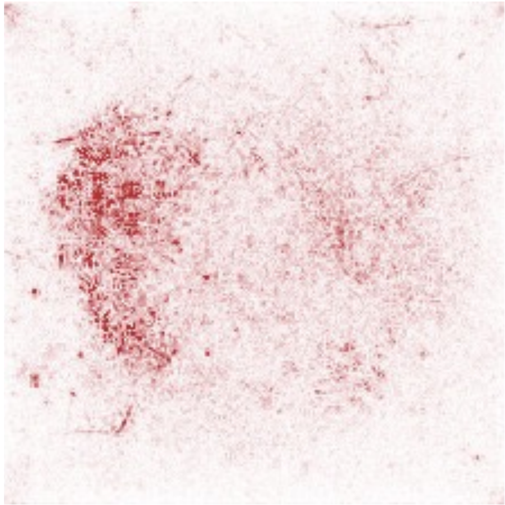
Provided by Alex DeGrave, MD/PhD student in the AIMS lab

Medical image example

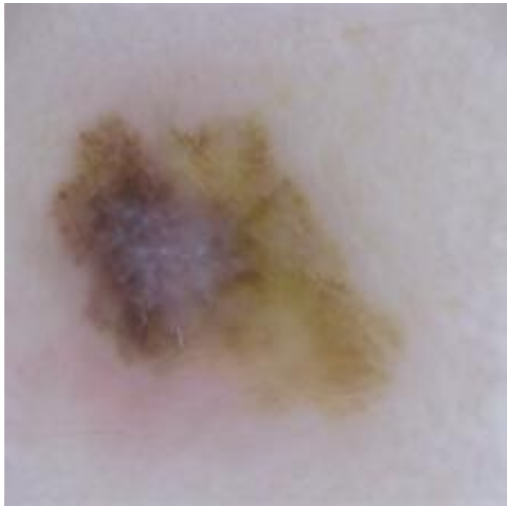
Original image

Predicted: benign

Saliency map



Modified image

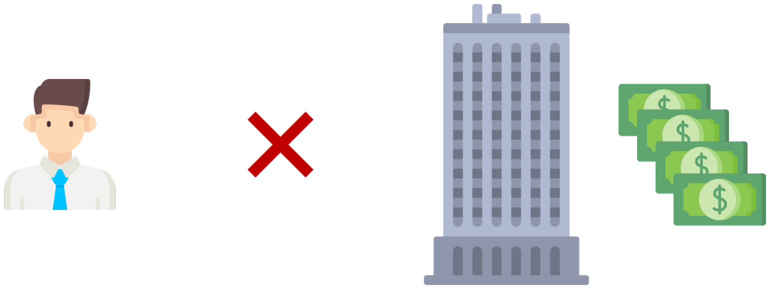


Predicted: malignant

Provided by Alex DeGrave, MD/PhD student in the AIMS lab

Loan approval example

- A bank customer applies for a loan, but his request is denied



- The customer may want to understand why
  - Not just which features are important, but which can be adjusted to change the outcome
  - **Problem:** feature importance methods do not answer this question (at least not exactly)

New explanation approach

- **Idea:** find input changes that alter a model predictions in the desired direction
  - Ideally, without changing the original input *too much*

- Two main goals:
  - Understand the model via input modifications
  - Identify options for *algorithmic recourse* (to reverse unfavorable decisions)

What's a counterfactual?

- Modifying a factual event and assessing the consequences of that change
  - Typically, "what if" or "if only I had" thoughts
- Example:
  - A person sips their tea and burns their tongue
  - "If I had waited 10 more minutes, I wouldn't have burned myself"
  - **Insight:** the burn was caused by drinking tea too soon

Counterfactual thinking

- Frequently discussed in the social sciences
  - Philosophers: Aristotle, Plato, Leibniz, Mill
  - Cognitive psychologists: Daniel Kahneman, Amos Tversky

- Key idea: counterfactual thinking is a tool for understanding causality

Downhill rule

- Study on *mental undoing*: how people reverse unwanted outcomes
  - See “Thinking, Fast and Slow” (Kahneman, 2011) or “The Undoing Project” (Lewis, 2017)
- When many changes are possible, people tend to undo/remove surprising occurrences
  - E.g., a car crash that occurred when driving home on an unusual route
  - Counterfactuals are naturally constrained by realism

Kahneman & Tversky, “The simulation heuristic” (1982)

Counterfactual explanations

- Can use counterfactuals to explain ML models
- For a given sample (**explicand**), find a similar sample with different prediction (**counterfactual**)
  - A form of local explanation
  - Alternative to local feature importance
  - Arguably more intuitive due to parallels in human psychology

Today

- Section 1
  - Black-box counterfactual explanations
  - Review of variations
  - Explanation by progressive exaggeration
- Section 2
  - Instance explanations

Setup

- Consider a differentiable black-box model f_θ with parameters θ, input x and label y
- Recall: such models are typically trained by optimizing their parameters:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)$$

- Models are often differentiable with respect to both parameters and inputs

Main idea

- Fix an input x^e with output $f_\theta(x^e)$
  - Choose desired outcome y'
  - Determine an input x' near x^e such that $f_\theta(x') \approx y'$
- Find this input by optimizing w.r.t. the input
  - Optimize via gradient descent
  - Like activation maximization, but with a different objective

Wachter et al., "Counterfactual explanations without opening the black box: Automated decisions and the GDPR" (2017)

Optimization problem

- Solve the following problem:

$$\arg\min_{x'} \max_{\lambda} \lambda(f_\theta(x') - y')^2 + d(x^e, x')$$

- Finds a counterfactual that…
  1. produces the desired output y'
  2. is as close to x^e as possible
- Notation:
  - λ controls the balance between objectives
  - d is a distance function

Optimization problem (cont.)

- The original version is equivalent to:

$$\arg\min_{x'} d(x^e, x') \quad \text{s.t.} \quad f_\theta(x') = y'$$

  - A simpler view, but still difficult to solve

- Relaxed, more practical version:

$$\arg\min_{x'} \lambda(f_\theta(x') - y')^2 + d(x^e, x')$$

  - Fix λ to a large value

Distance metric

- Wachter et al. use a weighted version of L_1 norm, or Manhattan distance:

$$d(x^e, x') = \sum_k \frac{|x_k^e - x_k'|}{w_k}$$

- Weights are inverse median absolute deviation:

$$w_k = \frac{1}{\text{median}_j\left(\left|X_{j,k} - \text{median}_l\left(X_{l,k}\right)\right|\right)}$$

  - $X_{j,k}$ is the jth sample of kth feature

Distance properties

- Encourages small changes
- Captures natural variability of the space
  - Median absolute deviation is like standard deviation, but more robust to outliers
- Encourages sparsity in the counterfactual due to L_1 norm (like lasso linear regression)
  - Many features should remain unchanged

Example

- Three-layer MLP on LSAT dataset (common dataset in fairness literature)
  - Predicting first-year average grade based on:
    - GPA prior to law school
    - Entrance exam scores (LSAT)
    - Race (0 for white, 1 for black)
- Generating counterfactuals such that $f(x') = 0$
  - In their dataset, this represents an average score
  - The question is: "what change would make model predict an average score?"

Wachter et al., "Counterfactual explanations without opening the black box: Automated decisions and the GDPR" (2017)

Example

$f(x^e)$	x^e			x' (normalized L_2)			x' (normalized L_1)		
Score	**GPA**	**LSAT**	**Race**	**GPA**	**LSAT**	**Race**	**GPA**	**LSAT**	**Race**
0.17	3.1	39.0	0	3.0	37.0	0.2	3.1	35.0	0.1
-0.57	2.7	18.3	0	2.8	28.1	-0.4	2.7	35.8	0.1
-0.77	3.3	28.0	1	3.5	39.8	0.4	3.3	34.4	0.1

Higher LSAT scores raise predicted grade

Evidence of racial bias in model

- Observations:
  - L_2 results are less sparse than L_1
  - Categorical variables (e.g., race) are difficult to optimize
  - None of these variables are modifiable in real life

Wachter et al., "Counterfactual explanations without opening the black box: Automated decisions and the GDPR" (2017)

Today

- Section 1
  - Black-box counterfactual explanations
  - Review of variations
  - Explanation by progressive exaggeration
- Section 2
  - Instance explanations

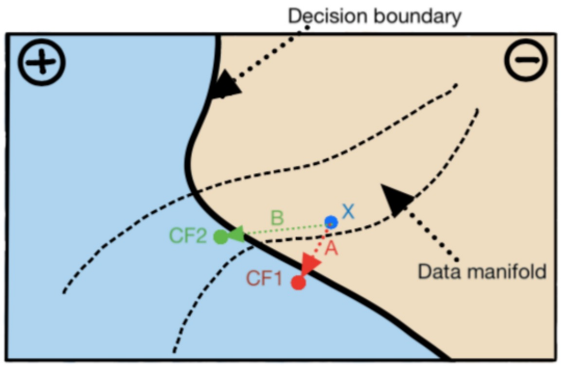
Review paper

- Examines 39 recent papers on counterfactual explanations
  - Explores variations on the original approach (Wachter et al., 2017)
  - Categorizes desiderata satisfied by different implementations
  - Identifies gaps and remaining challenges

Verma et al., "Counterfactual explanations for machine learning: A review" (2020)

Many counterfactuals

- Alice is denied a loan, wants to know what to change to get approved
- Problem: many possible counterfactuals!
  - Increase income and education
  - Increase credit score and decrease age



Verma et al., "Counterfactual explanations for machine learning: A review" (2020)

Desiderata

- What desiderata help prioritize counterfactuals?
- Validity
  - Does the counterfactual correctly change the prediction?
    - Does the counterfactual Alice get a loan?
- Distance
  - Is the counterfactual close to the explicand?
    - May only need to increase income by $10K rather than $50K
- Actionability
  - Does the counterfactual change *mutable* features?
    - Certain features cannot be changed (e.g., race, country of origin are *immutable*)

Desiderata (cont.)

- Sparsity
  - How many features does the counterfactual change?
    - Easier to change few things rather than many
- Data manifold
  - Is the counterfactual realistic?
    - Highly unlikely to be 20 years old and have a PhD
- Causality
  - Does the counterfactual comply with causality?
    - Getting a new educational degree necessitates increasing age by some amount

Implementing desiderata

- **Validity** + **distance** (Wachter et al., 2017)

$$\arg\min_{x'} d(x^e, x') \quad \text{s.t.} \quad f(x') = y'$$

- **Actionability**

$$\arg\min_{x' \in \mathcal{A}} d(x^e, x') \quad \text{s.t.} \quad f(x') = y'$$

  - Only actionable features $\mathcal{A}$ can change
  - Can be implemented softly via distance weighting

Implementing desiderata (cont.)

- **Sparsity**

$$\arg\min_{x'} d(x^e, x') \quad \text{s.t.} \quad f(x') = y'$$

  - Can set distance d to encourage sparsity (L_0 or L_1 norm)

- **Data manifold**

$$\arg\min_{x' \in \mathcal{A}} d(x^e, x') + l(x'; X) \quad \text{s.t.} \quad f(x') = y'$$

  - l penalizes counterfactuals that are far from the data manifold defined by the training set X
  - Not straightforward in practice: we rarely have l

Implementation properties

- Model access
  - Complete access, gradients only, predictions only
- Model class
  - Model-agnostic, differentiable models, linear models
- Amortization
  - We can train a model to generate counterfactuals (faster than optimizing for each explicand)
- Counterfactual attributes
  - Sparsity, data manifold, causality
- Optimization attributes
  - Actionable features, distance for categorical features

Verma et al., "Counterfactual explanations for machine learning: A review" (2020)

Comparing methods

	Assumptions		Optimization amortization		CF attributes			CF opt. problem attributes	
Paper	Model access	Model domain	Amortized Inference	Multiple CF	Sparsity	Data manifold	Causal relation	Feature preference	Categorical dist. func
[72]	Black-box	Agnostic	No	No	Changes iteratively	No	No	Yes	-
[111]	Gradients	Differentiable	No	No	L1	No	No	No	-
[104]	Complete	Tree ensemble	No	No	No	No	No	No	-
[74]	Black-box	Agnostic	No	No	L0 and post-hoc	No	No	No	-
[57]	Black-box	Agnostic	No	Yes	Flips min. split nodes	No	No	No	Indicator
[29]	Gradients	Differentiable	No	No	L1	Yes	No	No	-
[56]	Black-box	Agnostic	No	No	No	No	No	No[2]	-
[95]	Complete	Linear	No	Yes	L1	No	No	No	N.A.[3]
[107]	Complete	Linear	No	No	Hard constraint	No	No	Yes	-
[98]	Black-box	Agnostic	No	Yes	No	No	No	Yes	Indicator
[30]	Black-box or gradient	Differentiable	No	No	L1	Yes	No	No	-
[91]	Black-box	Agnostic	No	No	No	No	No	No	-
[61]	Gradients	Differentiable	No	No	No	Yes	No	No	-
[90]	Gradients	Differentiable	No	No	No	No	No	No	-

Verma et al., "Counterfactual explanations for machine learning: A review" (2020)

Comparing methods (cont.)

Reviewed a lot of methods!

	Assumptions		Optimization amortization		CF attributes			CF opt. problem attributes	
Paper	Model access	Model domain	Amortized Inference	Multiple CF	Sparsity	Data manifold	Causal relation	Feature preference	Categorical dist. func
[72]	Black-box	Agnostic	No	No	Changes iteratively	No	No	Yes	-
[111]	Gradients	Differentiable	No	No	L1	No	No	No	-
[104]	Complete	Tree ensemble	No	No	No	No	No	No	-
[74]	Black-box	Agnostic	No	No	L0 and post-hoc	No	No	No	-
[57]	Black-box	Agnostic	No	Yes	Flips min. split nodes	No	No	No	Indicator
[29]	Gradients	Differentiable	No	No	L1	Yes	No	No	-
[56]	Black-box	Agnostic	No	No	No	No	No	No[2]	-
[95]	Complete	Linear	No	Yes	L1	No	No	No	N.A.[3]
[107]	Complete	Linear	No	No	Hard constraint	No	No	Yes	-
[98]	Black-box	Agnostic	No	Yes	No	No	No	Yes	Indicator
[30]	Black-box or gradient	Differentiable	No	No	L1	Yes	No	No	-
[91]	Black-box	Agnostic	No	No	No	No	No	No	-
[61]	Gradients	Differentiable	No	No	No	Yes	No	No	-
[90]	Gradients	Differentiable	No	No	No	No	No	No	-
[113]	Black-box	Agnostic	No	No	Changes one feature	No	No	No	-
[85]	Gradients	Differentiable	No	Yes	L1 and post-hoc	No	No	No	Indicator
[89]	Black-box	Agnostic	No	No	No	Yes[4]	No	No	-
[108]	Black-box or gradient	Differentiable	No	No	L1	Yes	No	No	Embedding
[82]	Gradients	Differentiable	Yes	Yes	No	Yes	Yes	Yes	-
[64]	Complete	Linear	No	Yes	Hard constraint	No	No	Yes	Indicator
[87]	Gradients	Differentiable	No	No	No	Yes	No	Yes	N.A.[5]
[67]	Black-box	Agnostic	No	No	Yes	Yes	No	No	-
[65]	Complete	Linear and causal graph	No	No	L1	No	Yes	Yes	-
[66]	Gradients	Differentiable	No	No	No	No	Yes	Yes	-
[76]	Gradients	Differentiable	No	No	Changes iteratively	Yes	No	No[6]	-
[26]	Black-box	Agnostic	No	Yes	L0	Yes	No	Yes	Indicator
[63]	Complete	Linear and tree ensemble	No	No	No	Yes	No	Yes	-
[47]	Complete	Random Forest	No	Yes	L1	No	No	No	-
[79]	Complete	Tree ensemble	No	No	L1	No	No	No	-

Verma et al., "Counterfactual explanations for machine learning: A review" (2020)

Open questions

- Scalability
  - Solving per-explicand optimization problem is slow
- Adversarial examples
  - Counterfactuals are susceptible to adversarial examples
  - How to mitigate, or prove solutions aren't adversarial?
- Local preferences
  - Actionable, mutable, and immutable features may change per explicand (user preferences)
- Categorical features
  - More difficult to optimize via gradient descent
- And more

Verma et al., "Counterfactual explanations for machine learning: A review" (2020)

Today

- Section 1
  - Black-box counterfactual explanations
  - Review of variations
  - Explanation by progressive exaggeration
- Section 2
  - Instance explanations

Motivation

- Images are more challenging than tabular data
  - Prone to adversarial examples
  - Want meaningful visual changes, realistic images

- This work creates a series of realistic, visually meaningful counterfactual images
  - Requires a deep learning classifier
  - Involves training other deep learning modules

Singla et al. "Explanation by progressive exaggeration" (2019)

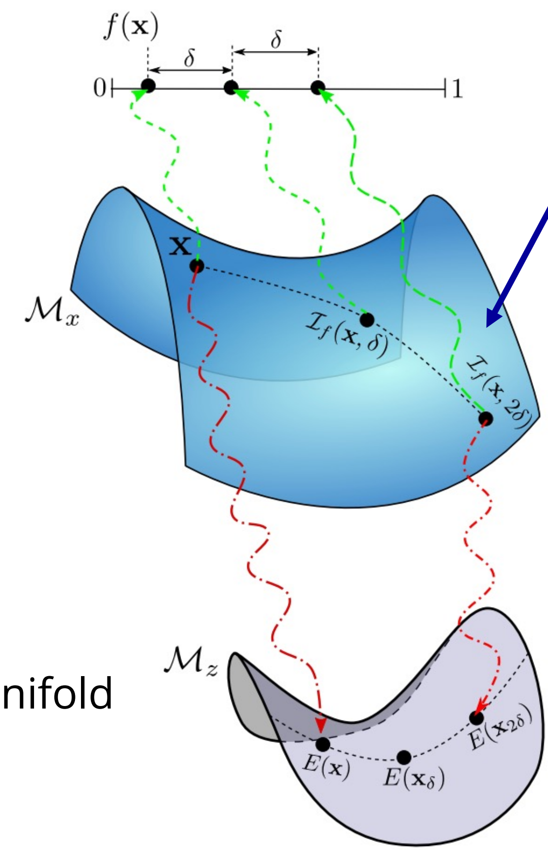
Premise

Size of prediction change δ

Series of counterfactuals

Counterfactual function $I(x, \delta)$

Data manifold



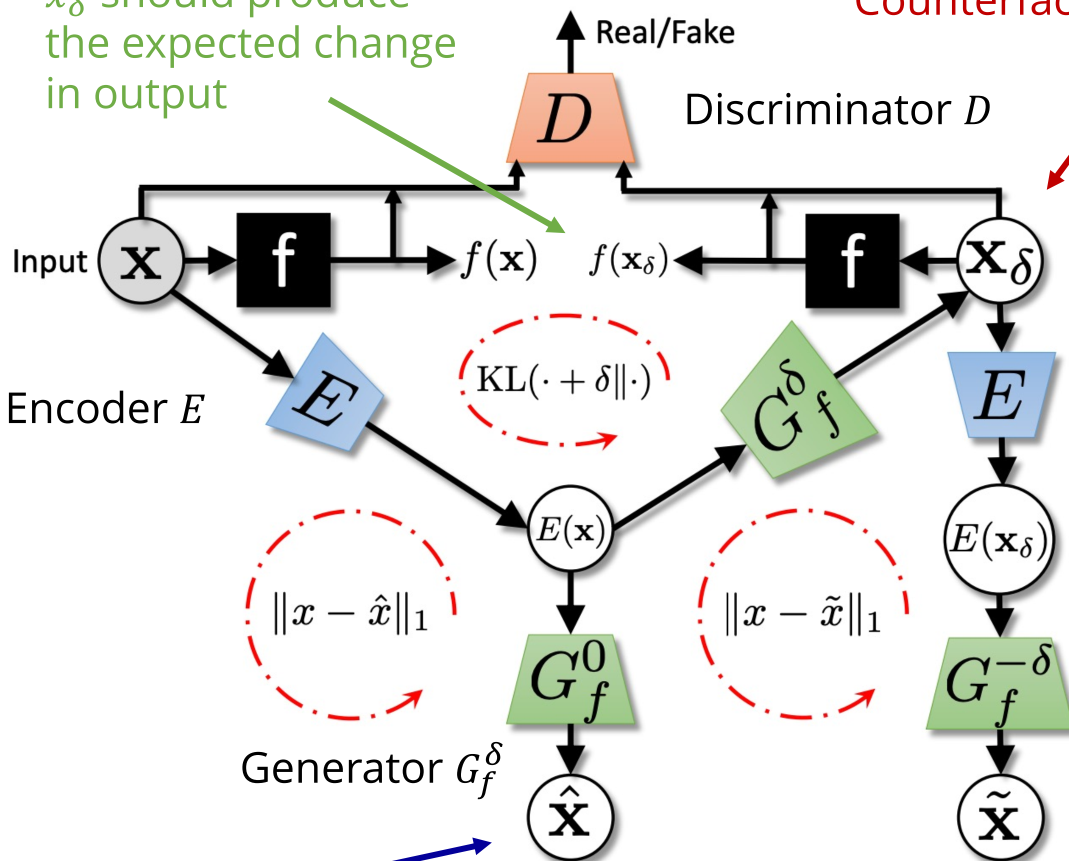
Encoder E (data → embedding)

Embedding manifold

Architecture + training

x_δ should produce the expected change in output

Counterfactual $x_\delta = G_f^\delta(E(x))$

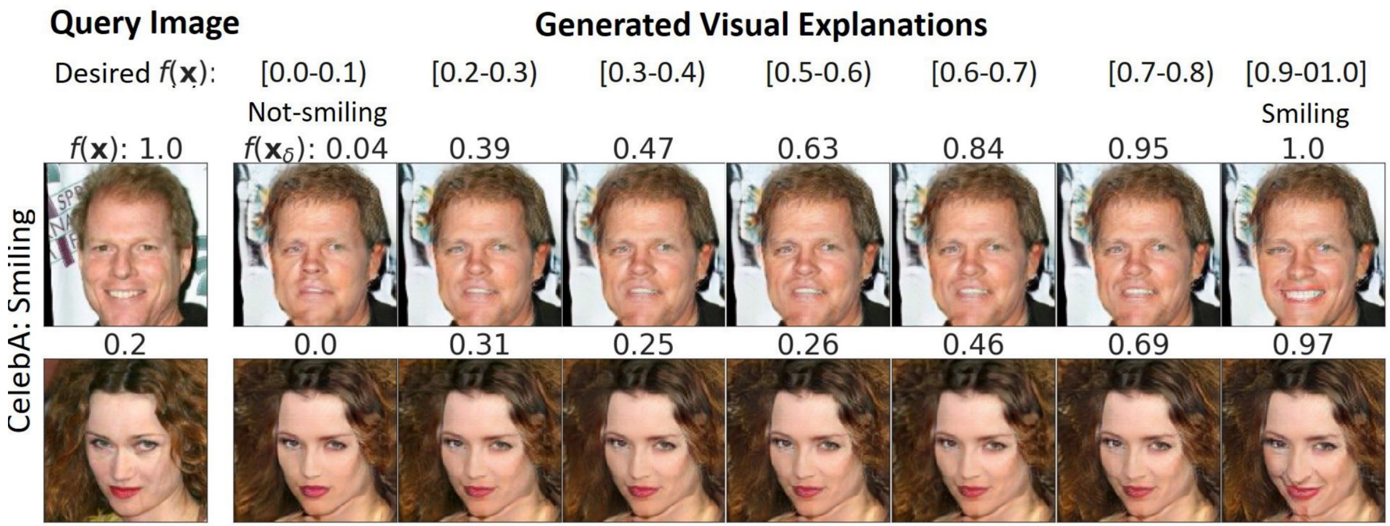


Should have $x_\delta \approx x$ when $\delta = 0$

Counterfactual function should be invertible

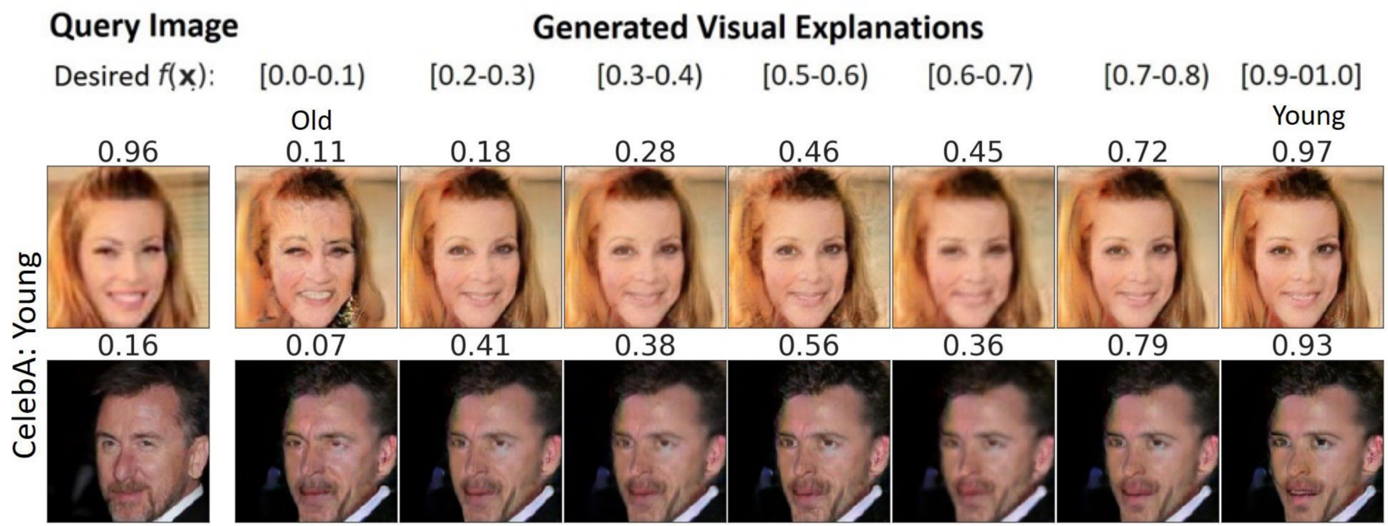
Example result

Not smiling → smiling



Example result

Not young → young



Conclusions

- Several ways to find counterfactual explanations
  - Easiest for differentiable models with tabular data and continuous features
  - We can handle categorical features and non-differentiable models (did not discuss), plus other data types

- Limitation: counterfactuals change model outputs, but not necessarily reality
  - E.g., in medical risk assessment, no treatment and short stay may be correlated with positive outcomes; but these are counterproductive interventions
  - Should rely on causal inference methods instead

Counterfactuals in ML

- Counterfactual reasoning is not unique to these methods
  - Feature importance also uses counterfactuals
    - Gradients: change from small input perturbation
    - Removal-based methods: observe outcomes with held-out feature values
  - A fundamental tool in causal inference
    - See “Causality” textbook by Judea Pearl (2009)

- As a result, *counterfactual explanations* are sometimes known as **recourse explanations**

Today

- Section 1
  - Black-box counterfactual explanations
  - Review of variations
  - Explanation by progressive exaggeration
  - **10 min break**
- Section 2
  - Instance explanations