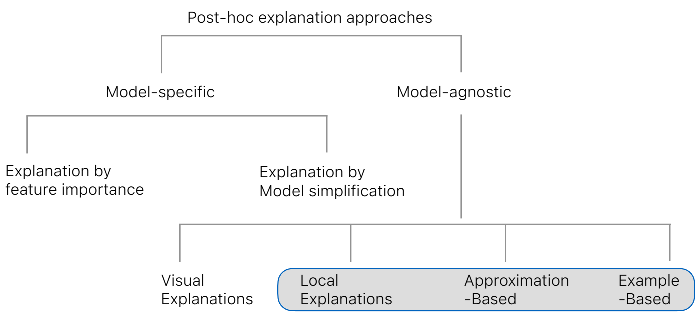
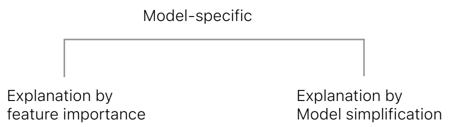
Counterfactual Explanations

Post-hoc explanation approaches

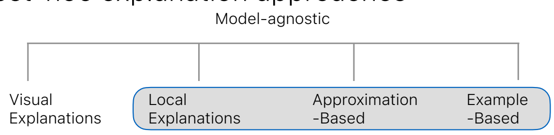


Post-hoc explanation approaches



- **Feature Importance**: finds the most influential features contributing to the model's overall accuracy or for a particular decision.
- **Model simplification**: finds an interpretable model that imitates the opaque model closely.

Post-hoc explanation approaches



- **Visual Explanations**: Plot the change in the model's prediction as one or multiple features are changed.
- **Local Explanations**: only explain a single prediction.
- Approximation-based: Sample new datapoints in the vicinity of the explainee datapoint
  — Then fit a linear model or extract rule sets.
- Example-based: Seek to find data-points in the vicinity of the explainee datapoint. — Try to explain in terms of neighbours that have either the same prediction or a different prediction.

- Explainability has two parts:
  - Why — that is, why the model has produced the desired outcome
  - Suggestions — Make suggestions to achieve the desired outcome.
- The actionable feedback is also known as Algorithmic Recourse
- CFEs do not explicitly answer the "why" the model made a prediction.
- CFEs provide suggestions to achieve the desired outcome.
- CFEs are also applicable to black-box models.
  - Only the predict function of the model is accessible. Do not require model disclosure

Social Implications of Machine

- Ensure equitable social implications of machine learning.
- Establish Fairness
- Make automated tool's decision explainable

recedivism

- Fairness: ensure that the decisions produced by the system are not biased against a particular demographic group of of individuals.
  - i.e. groups defined based on sensitive and protected features such as race, sex, religion.
  - Anti-discrimination laws

Use cases of explainability

- Military classifier to distinguish enemy and friendly tanks
- Classifier to distinguish husky from wolf. — Husky was classified as a wolf because of the presence of snow in the background.

Explainability Problem

- Model Explanation
- Outcome Explanation

- Model explanation searches for interpretable and transparent global explanations of the original model
  - Some approaches are model agnostic
  - Some explain NN using single decision tree and rule sets.

- Outcome explanation:
  - Provide an explanation for a specific prediction from the model.
  - This explanation need not be a global explanation or explain the internal logic of the method.

Class Activation Maps.

  - Model-specific approaches for deep neural networks (CAM) and model agnostic approaches (LIME) have been proposed.
  - CFE

Locally Interpretable model agnostic

Grad-CAM

- Example-based approach is another kind of outcome explanation technique
- CFE is an example-based approach.

- CFEs are applicable to supervised ML — we know the desired prediction and want to explain when the desired prediction is not obtained.
- CFE is mostly applied to classification settings.

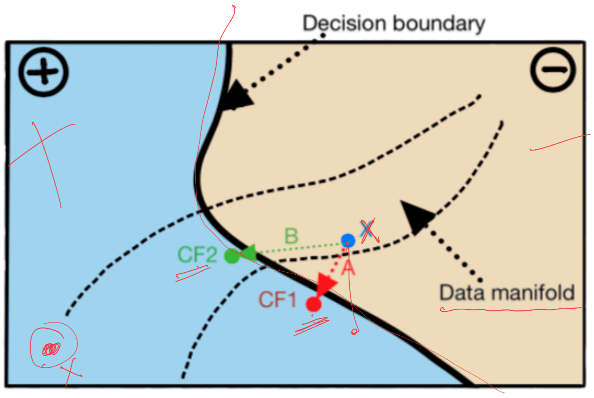
- Why was the loan denied
- What can be done differently so that the loan can be approved in the future?
- What small changes can be made to the feature vector in order to end up on the other side of the decision boundary?
- Increase salary/ increase education.
- Change should be relatively small which leads to the desired outcome.
- Change should be in few things, rather than changing many features.
- Advice should be realistic and actionable

Desiderata and major research themes

- Major themes of research have sought to focus to incorporate increasingly complex constraints on counterfactuals.
  - Aiming to make the resulting explanation truly actionable and helpful.
- How to address the desiderata in a way that is generalisable across algorithms and computationally efficient?

$f(x) \quad y \quad f(x') \quad y$

- Validity
- Actionability
- Sparsity
- Data Manifold Closeness
- Causality
- Amortised inference
- Black-box access
- Model Agnosticity



Validity

- Optimisation problem
  - Minimize the distance between the counterfactual x' and the data point x subject to … output of the classifier on the counterfactual is the desired label $y' \in \mathcal{Y}$.

x, y
x' y'

$$\arg\min_{x'} d(x, x') \text{ subject to } f(x') = y'$$

Differentiable unconstrained form

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

The distance metric d is used to emphasise that the counterfactual must be a small change relative to the starting point.

x

Actionability

- Mutable features — income, age
- Immutable features — race, country
- CF should never change an immutable feature or a legally sensitive feature.
- If changing a legally sensitive feature produces a large change in the prediction, it means there is bias.
- We can use the actionable set of features $\mathcal{A}$ and we can specify a preference order on these features.

$$\arg\min_{x' \in \mathcal{A}} \max_{\lambda} \lambda(f(x') - y')^2 + d(x, x')$$

Sparsity

- CF ideally should change a small number of features
- Easier to understand shorter explanations
- A penalty function $g(x - x')$ encourages sparsity in the difference between x and x'.

$$\arg\min_{x' \in \mathcal{A}} \max_{\lambda} \lambda(f(x') - y')^2 + d(x, x') + g(x - x')$$

- Using L_0 norm minimises the number of features changed
- Using L_1 norm minimises the total change in the features