INFO-H420
Management of Data Science and Business Workflows

Part II
Explainability

Dimitris SACHARIDIS

2023-2024

Explanations

- recommendations are everywhere – explanations are there as well!

amazon

Related to items you've viewed

Frequently Bought Together

Customers Who Bought This Item Also Bought

Recommended for You Based on

NETFLIX

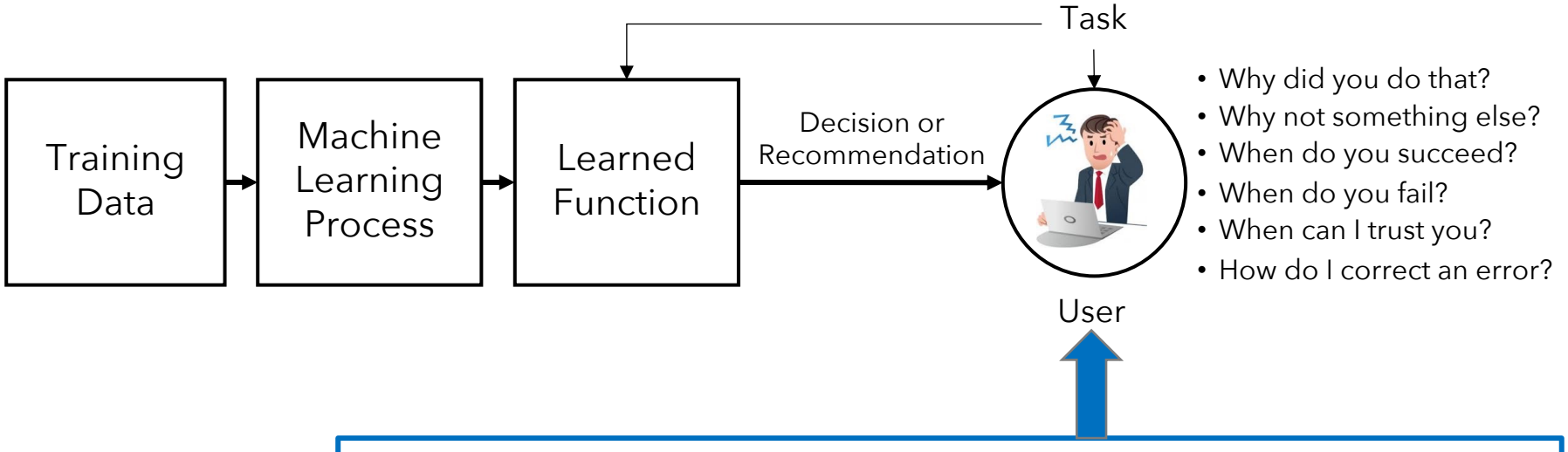
Popular on Netflix

Trending Now

Watch It Again

Because you watched Dark

Explainable AI



- The target of XAI is an end user who:
  - depends on decisions, recommendations, or actions of the system
  - needs to understand the rationale for the system's decisions to understand, appropriately trust, and effectively manage the system

[DARPA] Explainable Artificial Intelligence (XAI)

Explainable AI

Human-understandable explanations of **outcomes** of algorithmic decision-making systems

Explanations that

- enable understanding of overall strengths & weaknesses
- convey an understanding of how the system will behave in the future
- convey how to correct the system's mistakes

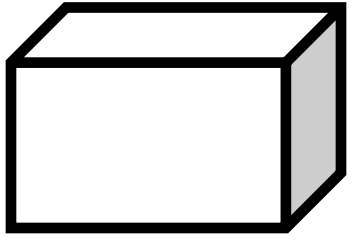
XAI is used to:

- detect biases and prevent discrimination
- detect training data errors, debug the model
- comply with regulations for transparency

[DARPA] Explainable Artificial Intelligence (XAI)

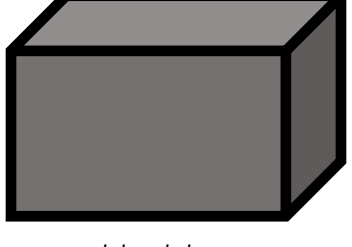
White-Box vs Black-Box

- a **white box** is a transparent AI system
  - complete knowledge of inner workings
  - often interpretable by humans = interpretation



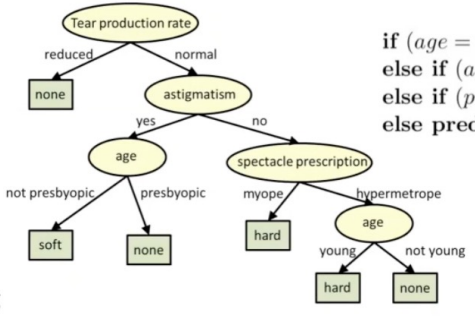
white box

- a **black box** is an opaque, complex AI system
  - no knowledge of inner workings
  - we can only try to make an educated guess about how it reaches a decision = explanation



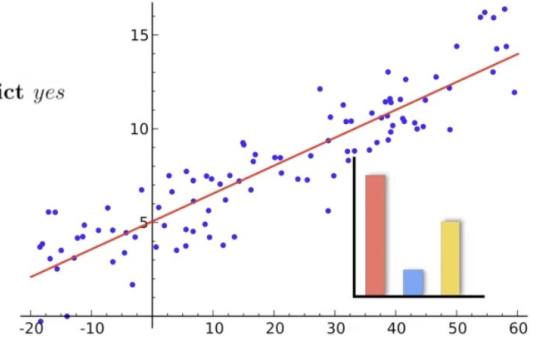
black box

How to achieve Explainability

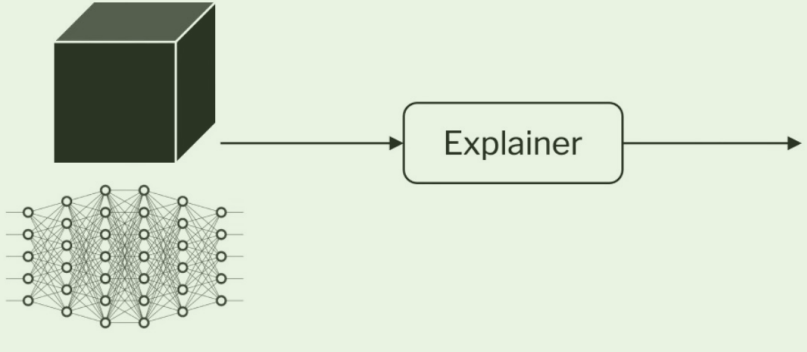


if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

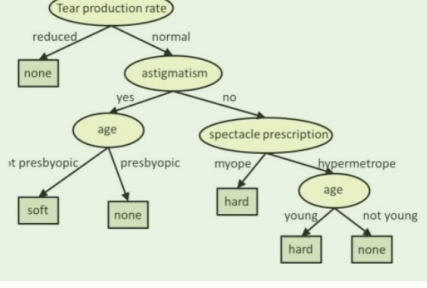
Intrinsic



Post hoc



if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no



[2022 ACM SIGMOD Pradhan et al.] Explainable AI: Foundations, Applications, and Opportunities for Data Management Research https://explainable-ai-tutorial.github.io

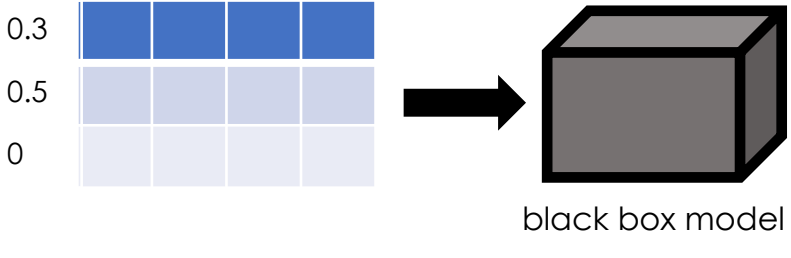
Types of Explanations

Name	Age	Saving	Month	…	Credit History
Maeve	<25	<100 DM	24	…	Paid duly

Why was my loan application rejected?

Feature-based Explanations

Data-based Explanations



Why Explain?

Purpose	Description
Transparency	Explain how the system works
Effectiveness	Help users make good decisions
Trust	Increase users' confidence in the system
Persuasiveness	Convince users to try or buy
Satisfaction	Increase the ease of use or enjoyment
Education	Allow users to learn something from the system
Scrutability	Allow users to tell the system it is wrong
Efficiency	Help users make decisions faster
Debugging	Allows users to identify that there are defects in the system

[1984 B. G. Buchanan et al.] *Explanations as a Topic of AI Research, in Rule-based Systems*

[2017 UMUAI I. Nunes et al.] *A systematic review and taxonomy of explanations in decision support and recommender systems.*

[2007 ICDE_w N. Tintarev et al.] *A survey of explanations in recommender systems.*

What is a good Explanation?

	Co-12 Property	Description
Content	**Correctness**	Describes how faithful the explanation is w.r.t. the black box.
		Key idea: Nothing but the truth
	Completeness	Describes how much of the black box behavior is described in the explanation.
		Key idea: The whole truth
	Consistency	Describes how deterministic and implementation-invariant the explanation method is.
		Key idea: Identical inputs should have identical explanations
	Continuity	Describes how continuous and generalizable the explanation function is.
		Key idea: Similar inputs should have similar explanations
	Contrastivity	Describes how discriminative the explanation is w.r.t. other events or targets.
		Key idea: Answers "why not?" or "what if?" questions
	Covariate complexity	Describes how complex the (interactions of) features in the explanation are.
		Key idea: Human-understandable concepts in the explanation
Presentation	**Compactness**	Describes the size of the explanation.
		Key idea: Less is more
	Composition	Describes the presentation format and organization of the explanation.
		Key idea: *How* something is explained
	Confidence	Describes the presence and accuracy of probability information in the explanation.
		Key idea: Confidence measure of the explanation or model output
User	**Context**	Describes how relevant the explanation is to the user and their needs.
		Key idea: How much does the explanation matter in practice?
	Coherence	Describes how accordant the explanation is with prior knowledge and beliefs.
		Key idea: Plausibility or reasonableness to users
	Controllability	Describes how interactive or controllable an explanation is for a user.
		Key idea: Can the user influence the explanation?

[2023 ACM Comp. Surveys M. Nata et al.] *From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI*

A use case: the Adult Dataset

Adult

Donated on 4/30/1996

Predict whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

Dataset Characteristics
Multivariate

Subject Area
Social Science

Associated Tasks
Classification

Feature Type
Categorical, Integer

 # Instances
48842

 # Features
14

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

https://archive.ics.uci.edu/dataset/2/adult

A use case: the Adult Dataset

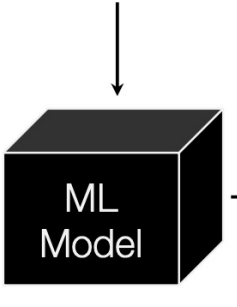
Bank

Rosa

Other customers of the bank

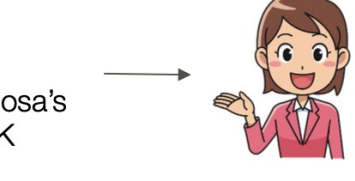
Name	Age	Education	Marital	…	Race	Sex	Hours
Rosa	34	Bachelors	Unmarried	…	White	female	40

Rosa's data



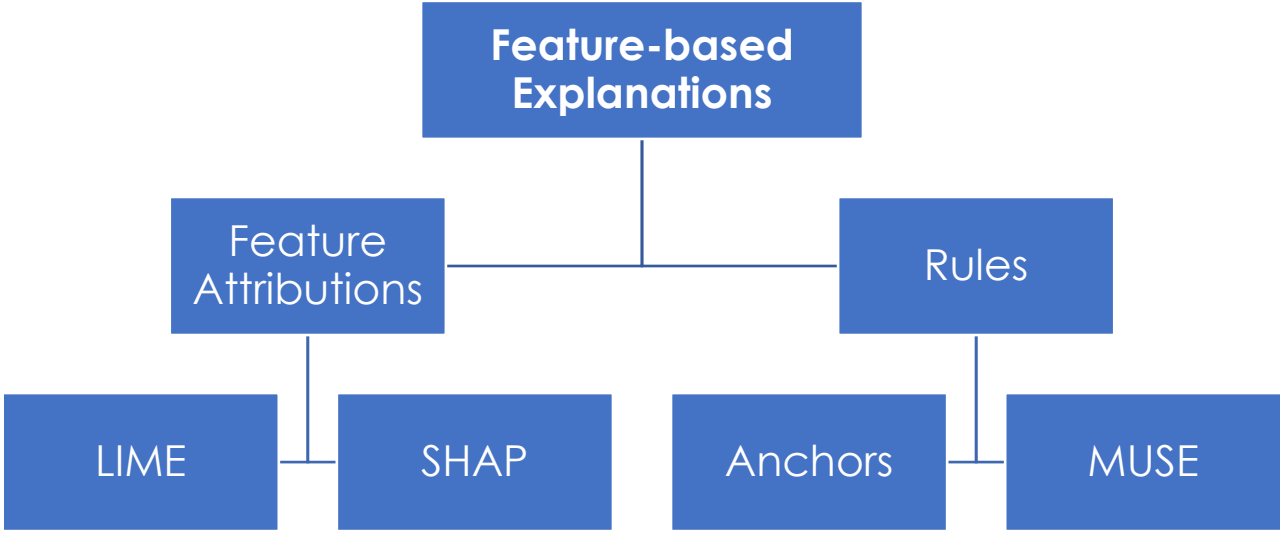
83%

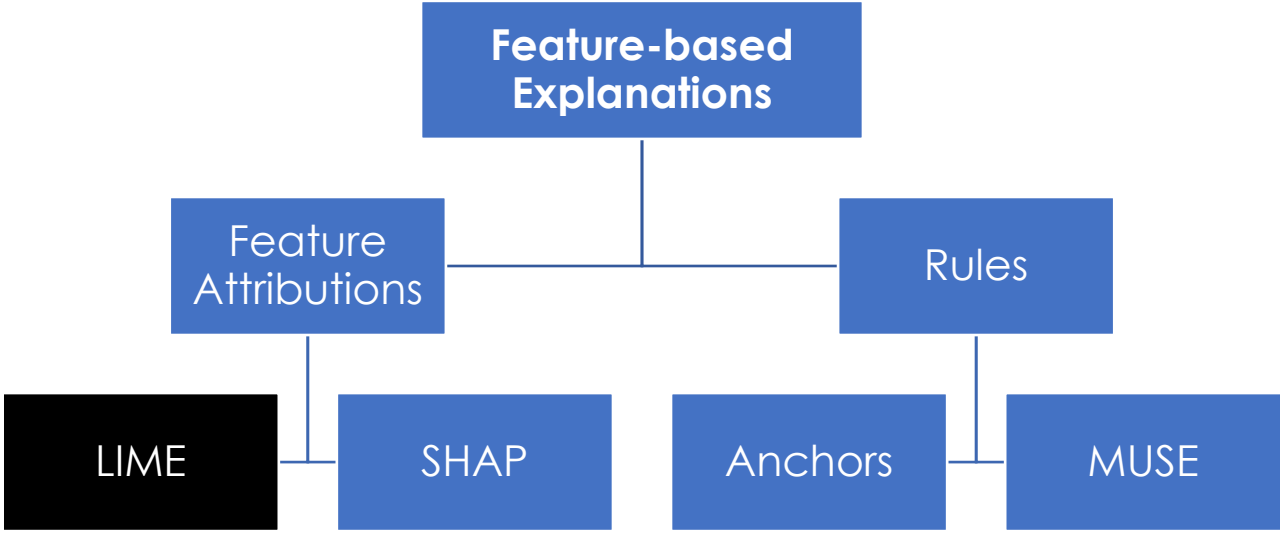
Probability that Rosa's income < 50K



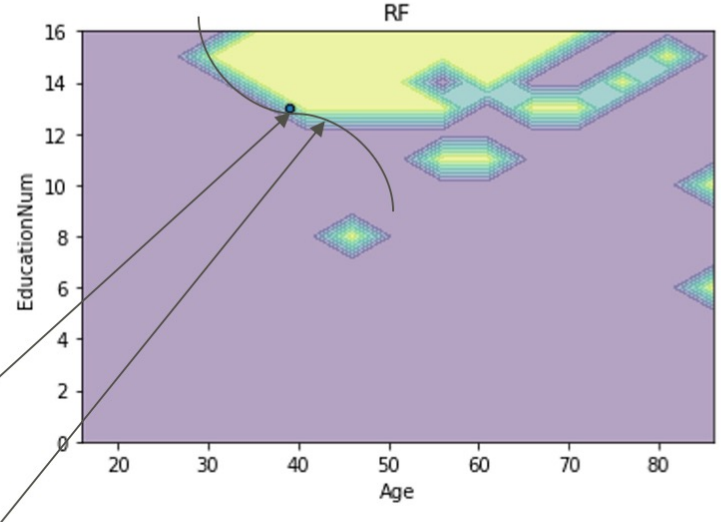
Manager

Feature-based Explanations


Feature-based Explanations
Feature Attributions
Rules
LIME
SHAP
Anchors
MUSE


Feature-based Explanations
Feature Attributions
Rules
LIME
SHAP
Anchors
MUSE

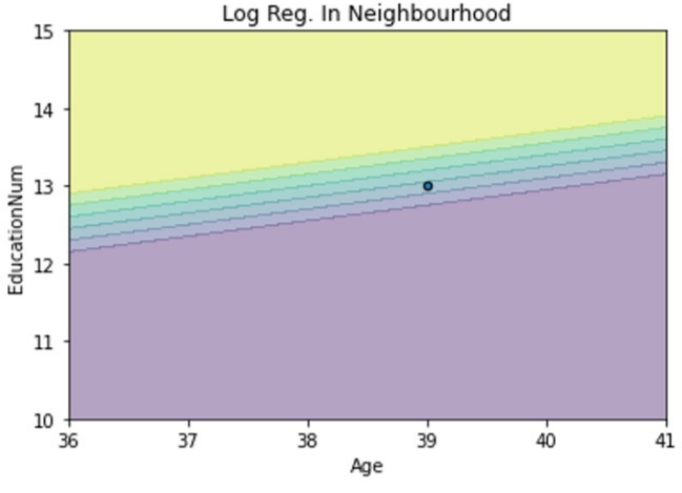
Complex Decision Boundaries



Want explanation for this point

Take neighbouring points and fit linear model

2D Decision Boundary of a Complex Classifier

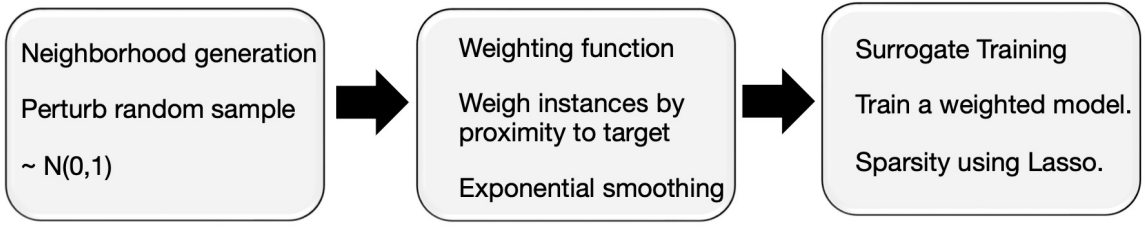


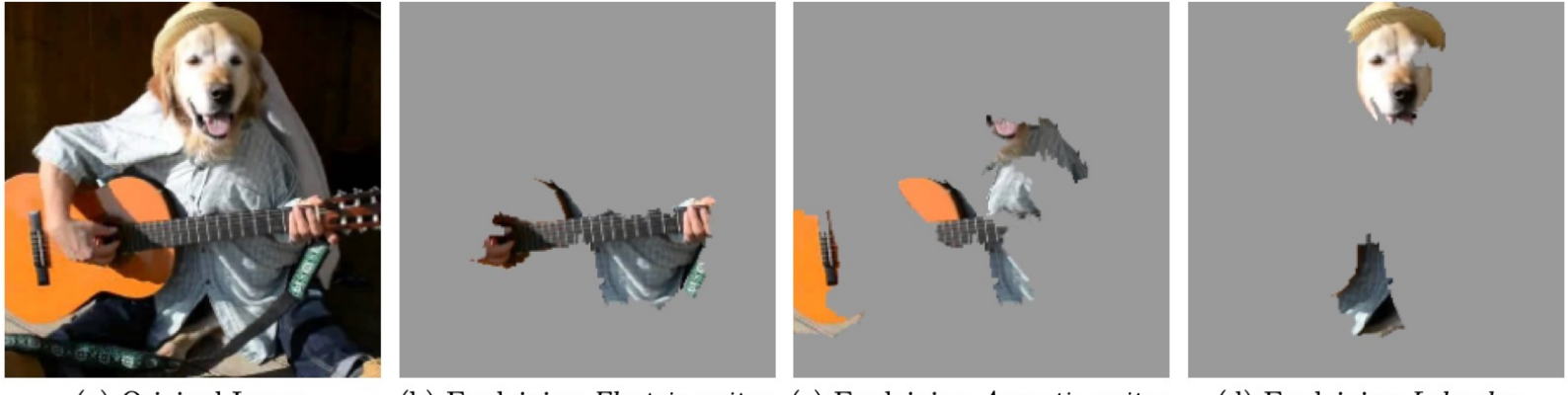
… are simpler locally

LIME – Algorithm

Local Interpretable Model-Agnostic Explanations (LIME)

- fit an interpretable model locally
- get feature weights

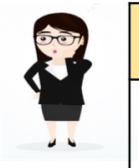




(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

[2016 ACM SIGKDD Ribeiro et al.] "Why should I trust you?" Explaining the predictions of any classifier

LIME – Issues

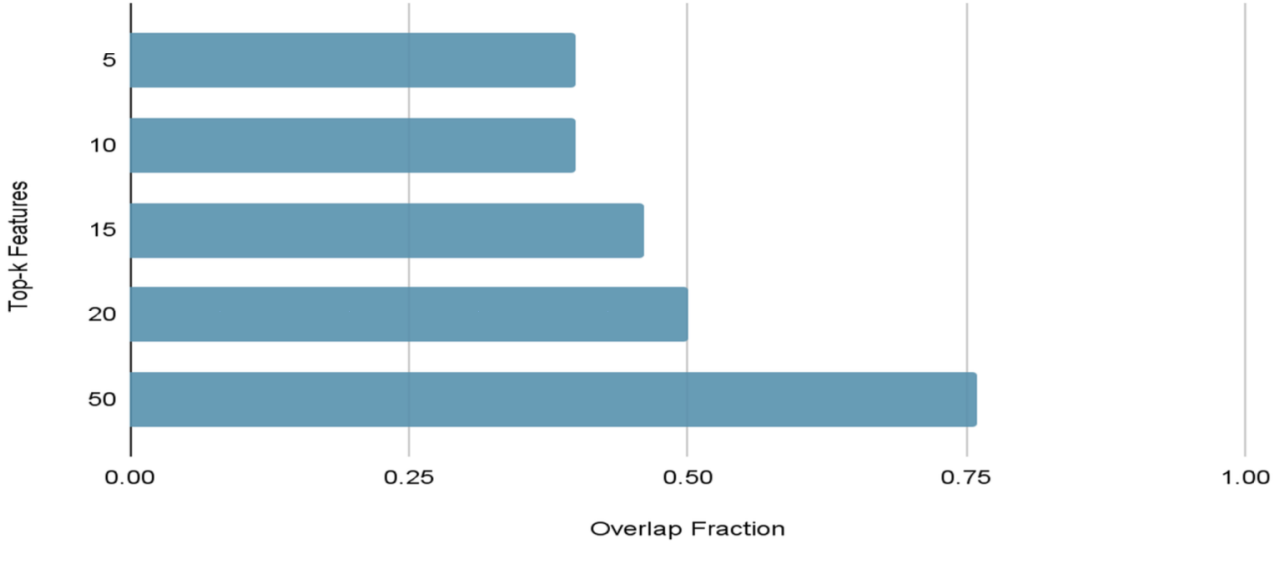


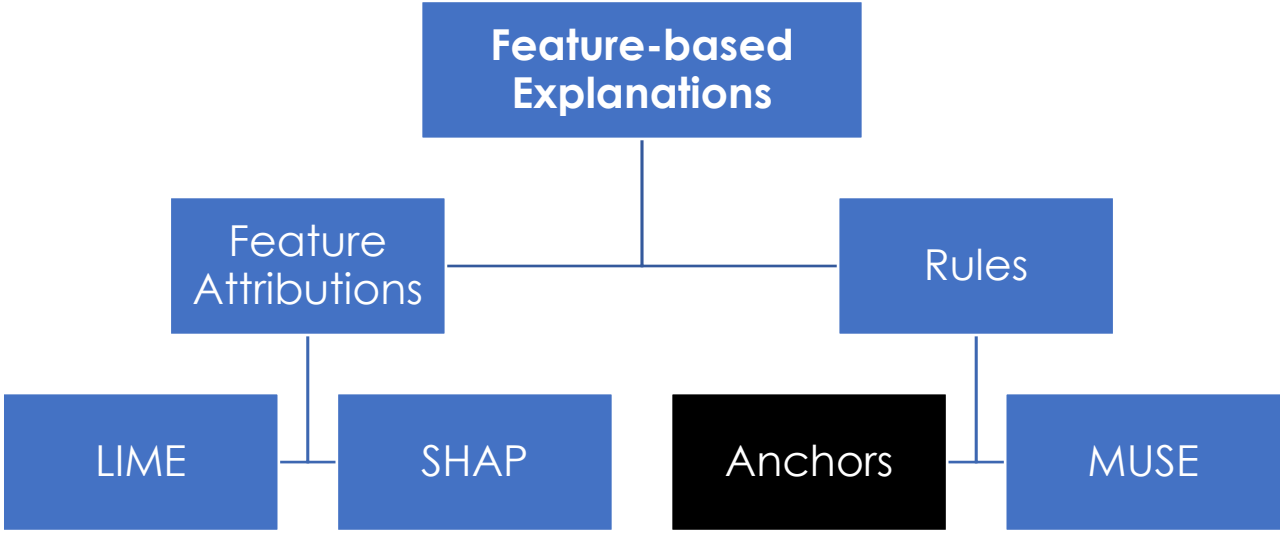
Rosa

Name	Age	Race	…	Hours
Rosa	34	White	…	**40**

Emma

Name	Age	Race	…	Hours
Emma	34	White	…	**30**



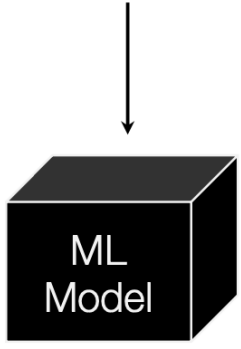

Feature-based Explanations
Feature Attributions
Rules
LIME
SHAP
Anchors
MUSE

Rule-based Explanations

Rosa

Name	Age	Education	Marital	…	Race	Sex	Hours
Rosa	34	Bachelors	Unmarried	…	White	female	40

Rosa's data



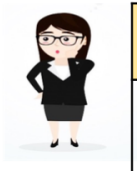
< 50K
Rosa's
income

IF Hours<50 **AND** Marital != "Married", **THEN** <50K

- Local sufficient conditions explaining the output.
- Aims for clear coverage of explanation with precise rules

[2018 AAAI Ribeiro et al.] Anchors: High-precision model-agnostic explanations

LIME vs Anchors



Rosa

Name	Age	Race	...	Hours
Rosa	34	White	...	**40**

Emma

Name	Age	Race	...	Hours
Emma	34	White	...	**30**

Manager

Top 3 features
Age
Education
Hours

Top 3 features

- The scope or validity of LIME is unclear.
- ANCHORS provides clear conditions where explanations hold.

IF Hours<50 **AND** Marital != "Married", **THEN** <50K

Anchors – Quality of Rules

Rosa

Name	Age	Education	Marital	…	Race	Sex	Hours
Rosa	34	Bachelors	Unmarried	…	White	female	40

Rosa's data

PRECISION

COVERAGE

IF Hours<50, **THEN** <50K

IF Hours<50 **AND** Marital != "Married", **THEN** <50K

IF Hours<50 **AND** Marital != "Married" **AND** Race = "White", **THEN** <50K

Precision = # of instances where provided explanation is correct.
Coverage = # of input instances where explanation clause is valid.

Anchors – Algorithm

Multi Armed Bandit Problem

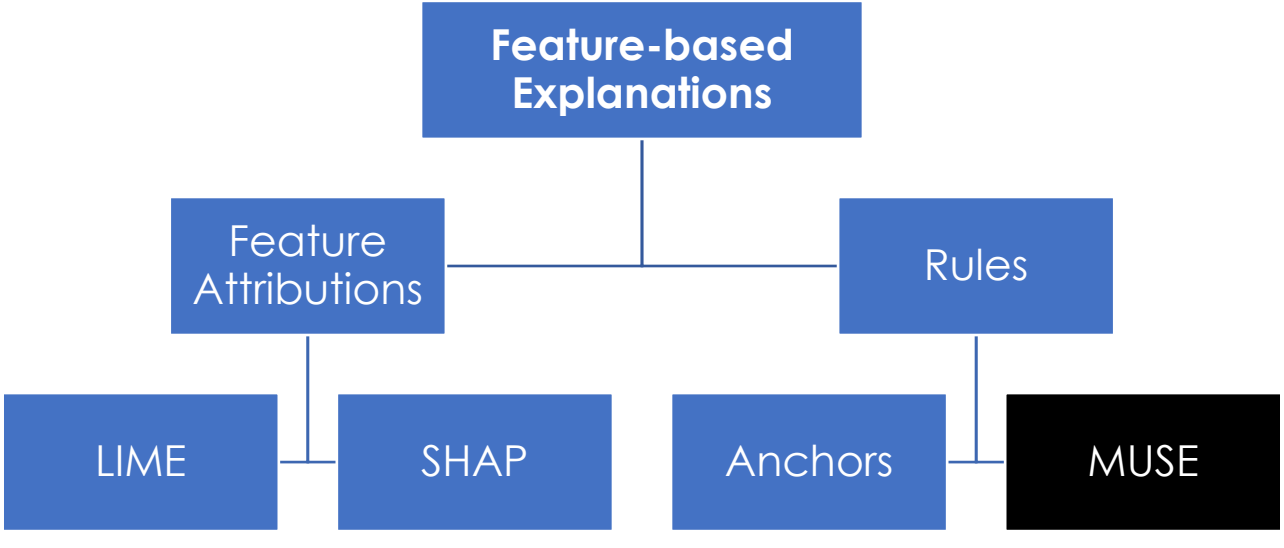
Beam Search

Idea: Create Rules, and Pick the most promising to expand

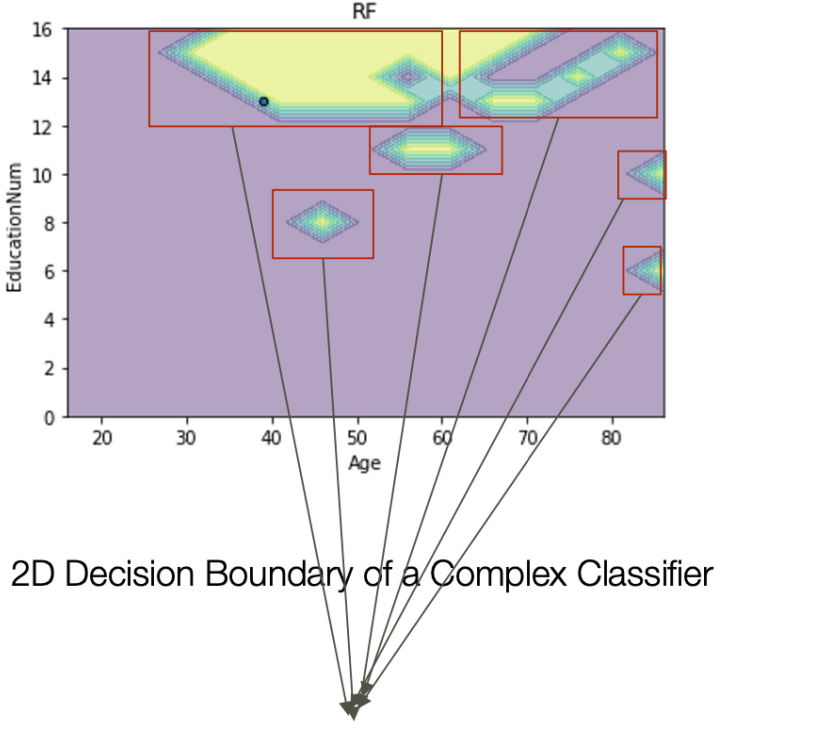
IF Hours<50

IF Hours<50 **AND** Marital != "Married"

IF Hours<50 **AND** Marital != "Married" **AND** Race = "White"


Feature-based Explanations
Feature Attributions
Rules
LIME
SHAP
Anchors
MUSE

Global Explainability



[2019 AAAI/ACM AES Lakkaraju et al.] Faithful and customizable explanations of black box models

MUSE – Global Explanations

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

Entire dataset

If Age < “34”

If Education==“Bachelors” AND Marital!=“Married” , THEN <50K

If Education==“Bachelors“ AND Age > “40” , THEN >50K

If Hours > 40

If Education==”Masters” AND Marital!=“Married” , THEN <50K

Default

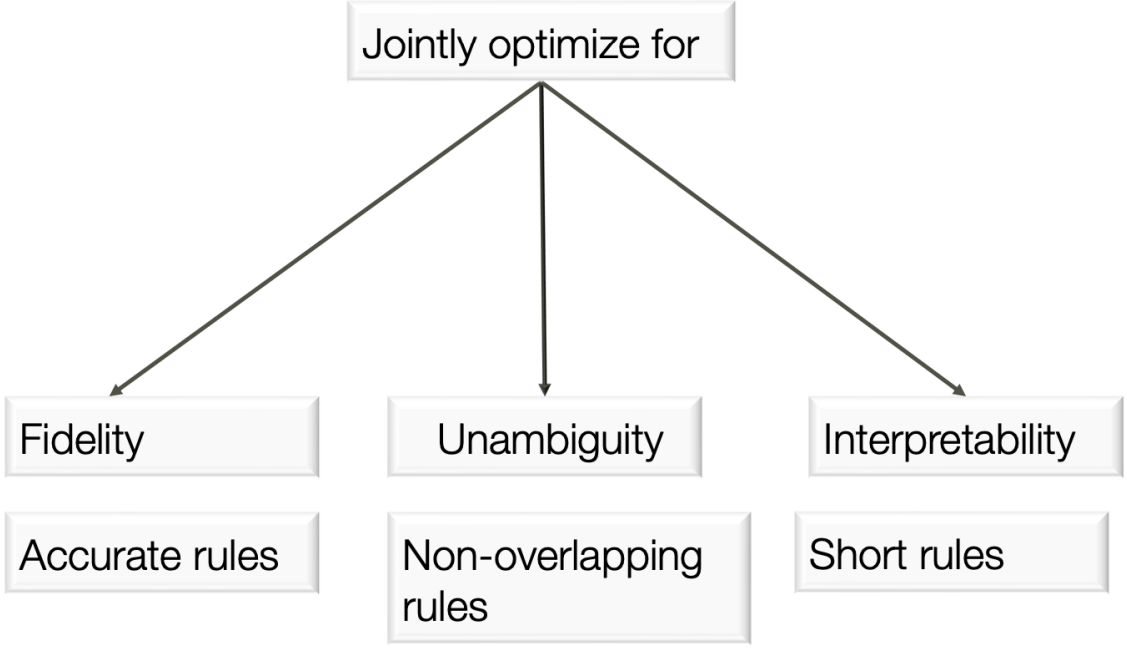
If Education==“Bachelors” AND Marital==“Married” ,
THEN >50K

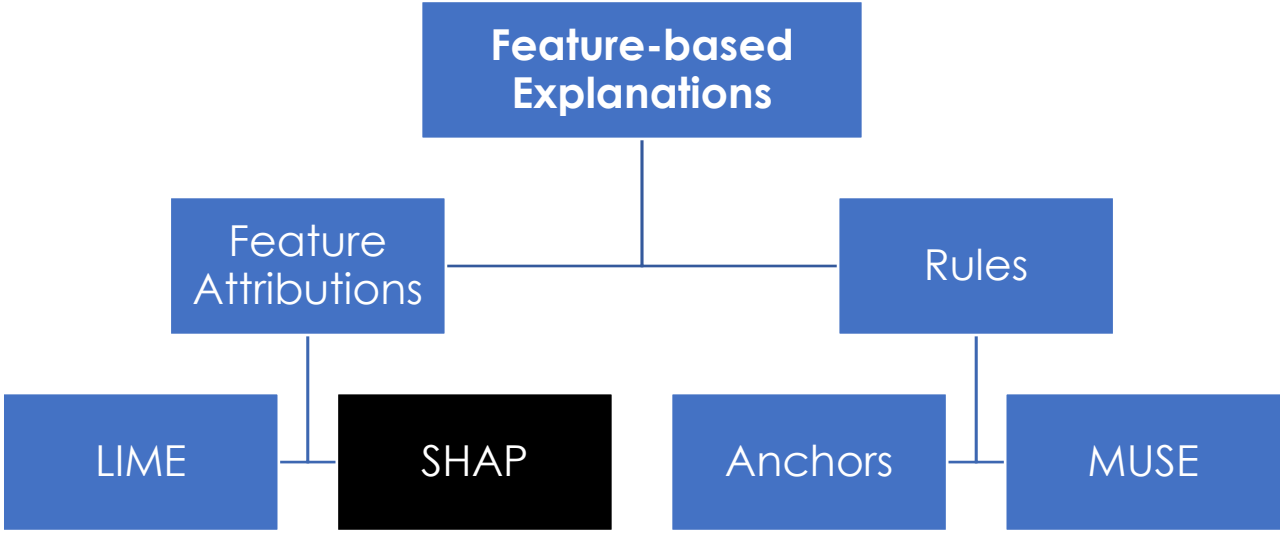
If Race==“Black” AND Hours<“20” , THEN <50K

Global, two-level rule-based explanations.

MUSE – Algorithm

Find a **good set** of rules that explain the model globally



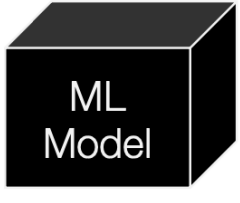

Feature-based Explanations
Feature Attributions
Rules
LIME
SHAP
Anchors
MUSE

Motivation

Rosa

Name	Age	Education	Marital	…	Race	Sex	Hours
Rosa	34	Bachelors	Unmarried	…	White	female	40

Rosa's data



< 50K
Rosa's
income

p=0.83

prediction
probability

find feature importance weights such that:

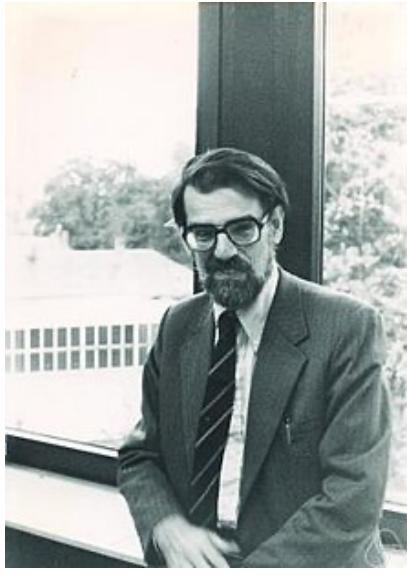
$\phi_{education} + \phi_{age} + \phi_{race} ..... = 0.83$

Properties:

- Feature weights add up to prediction (Efficiency)
- Guarantees for when $\phi_{feature} = 0$ (Null)
- If feature1 ~ feature2 then $\phi_{feature1} = \phi_{feature2}$ (Symmetry)

Shapley Values – Game Theory

Shapley Values is the **unique** solution to the **fair allocation** of payoffs among members in a **coalition** participating in a game characterized by a specific set of axioms.



Lloyd Shapley
2012 Nobel Memorial Prize
in Economic Sciences

[2002 Winter, Eyal.] "The shapley value." Handbook of game theory with economic applications

Shapley Values – Terminology

Coalition : Set of players

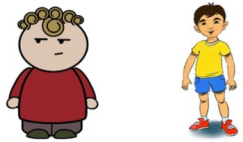
Null Coalition : Empty Set of players

∅

Grand Coalition : Set of All players

S : players in coalition

S' : players out of coalition



V(S) : Value function for S

Score () = 10

Shapley Values – Example

2 players playing table tennis (alone + team)

Q. Who is the better player?

Use Shapley Value to distribute their joint score to them individually, fairly.

p1

p2

Game

Shapley Values – Example

p1 = 6

p2 = 8

p1 p2 = 12

∅ = 0

Shapley Values – Formula

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S))$$

Shapley Value for feature i

Iterating over all possible coalitions without feature i

Weight of coalition

Marginal contribution of feature i to this coalition S

Shapley Values – Example

() — (∅) => 6 - 0 = 6

p1

( ) — () => 12 - 8 = 4

p1 p2 p2

Φ  => ½*(6) + ½*(4) = 5

p1

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S))$$

Shapley Values – Example

() — (∅) => 8 - 0 = 8

p2

( ) — () => 12 - 6 = 6

p2 p1 p1

Φ  => ½*(8) + ½*(6) = 7

p2

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S))$$

Shapley Values – Example

Φ  Φ  ( )

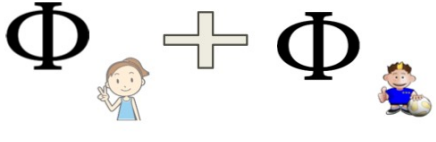
$5 + 7 = 12$

Shapley Values – Feature Weights

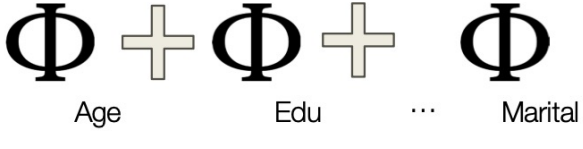
Game

Players

Name	Age	Education	Marital	…	Hours
Rosa	34	Bachelors	Unmarried	…	40



Payoff



Shapley Values – Feature Weights

= 6

Name	Age	Education	Marital	...
?	34	Bachelors	?	...

=

= 8

Name	Age	Education	Marital	...
?	?	Bachelors	?	...

=

Finding prediction for partial instances.
This is now a missing value problem.

Shapley Values – Feature Weights – Marginalise

Marginalise on all members in S' to find the value function.

$v(\mathbf{X}_S=\mathbf{x}_S) = E_D[f(\mathbf{x}_S,\mathbf{X}_{S'})]$

Name	Education	Marital
Rosa	Bachelors	Unmarried
James	Bachelors	Unmarried
Imran	Masters	Married
Ilya	PhD	Married

Rosa's explanation.

S=(Marital),S'=(Education)

$v(\mathbf{X}_S$=(Marital=Unmarried))

Marginalise over Education

Name	Education	Marital
Rosa	Bachelors	Unmarried
P1	Masters	Unmarried
P2	PhD	Unmarried

Issue : Out of distribution instances

Used in Quantitative Input Influence

[2016 (SP). IEEE Datta, et al.] Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems

Shapley Values – Feature Weights – Condition

Condition on all members in S to find the value function.

$v(\mathbf{X}_S=\mathbf{x}_S) = E[\, f(\, \mathbf{X} \mid \mathbf{X}_S= \mathbf{x}_S \,) \,]$

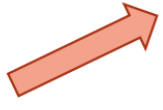
Name	Education	Marital
Rosa	Bachelors	Unmarried
James	Bachelors	Unmarried
Imran	Masters	Married
Ilya	PhD	Married

Rosa's explanation.

S=(Marital),S'=(Education)

$v(\mathbf{X}_S=_{\text{(Marital=Unmarried)}})$

Name	Education	Marital
Rosa	Bachelors	Unmarried
James	Bachelors	Unmarried



Condition on Marital

Used in SHAP

[2017 NeurIPS Lundberg, et al.] A unified approach to interpreting model predictions

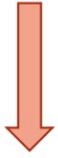
Shapley Values – Feature Weights – Issues

Choosing the players is non-trivial

Two variables Education(A) and Marital(B).

Add third redundant variable Higher Degree(C), such that,

$P(X_c=X_B)=1$



Name	Education (A)	Marital (B)	Higher Degree (C)
Rosa	Bachelors	Unmarried	No
James	Bachelors	Unmarried	No
Imran	Masters	Married	Yes
Tian	Masters	Unmarried	Yes

G_1 : Players=(A,B,C) : $f(X_a,X_b,X_c)$

$\Phi_{G1}(B) = \Phi_{G1}(C)$

G_2 : Players=(A,B): $f(X_a,X_b,X_b)$

$\Phi_{G2}(B) != \Phi_{G1}(B)$ and

$\Phi_{G2}(B) != \Phi_{G1}(B) + \Phi_{G1}(C)$

[2020 ICML Kumar, et al.] Problems with Shapley-value-based explanations as feature importance measures

Shapley Values – Feature Weights – Issues

Actionability

Manager

Φ Age Φ Edu ⋯ Φ Marital

Contrastiveness

Rosa

Emma

[2020 ICML Kumar, et al.] Problems with Shapley-value-based explanations as feature importance measures

Shapley Values – Feature Weights – Issues

Computing exact shapley values is intractable.

SHAP values are approximations of shapley values.

Computing SHAP	Linear Regression, Decision Trees	Logistic Regression, NN
With Feature Independence assumption	Tractable	#P-hard
Without Feature Independence assumption	#P-hard	#P-hard

[2021 AAAI Van den Broeck et al.] On the tractability of SHAP explanations

Data-Based Explanations

Data-Based Explanations

Use **training data points** to explain ML model behavior

Rosa

Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data

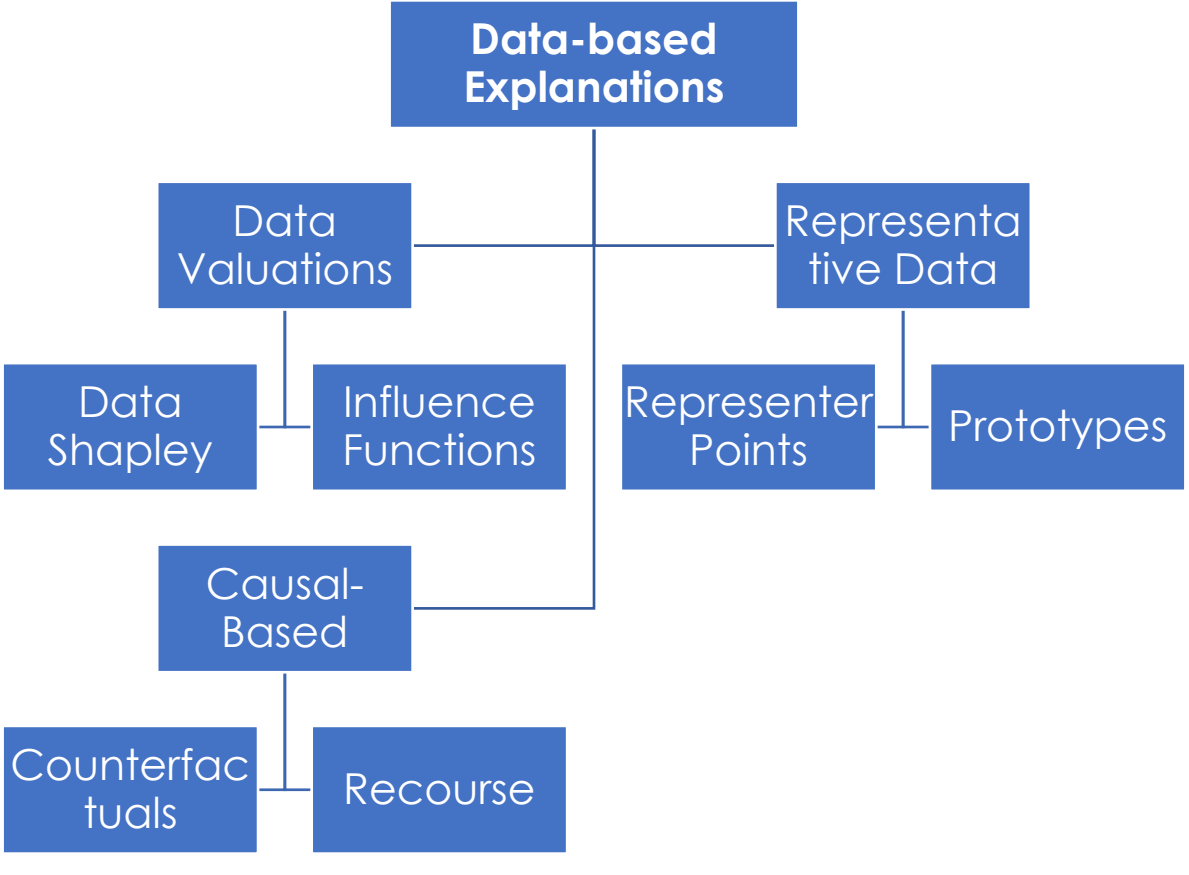
ML Model

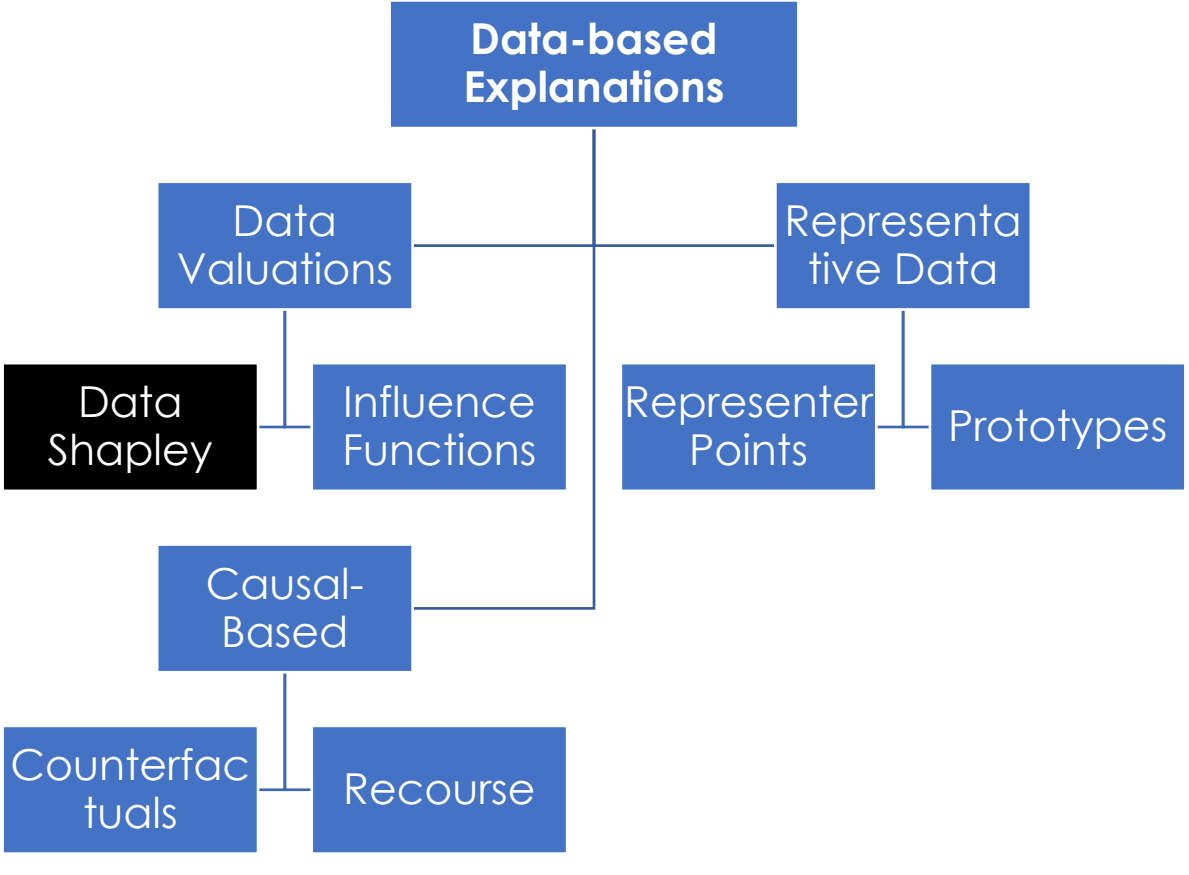
83%

Probability that Rosa's income < 50K

These instances are the most responsible for Rosa's negative outcome

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…


Data-based Explanations
Data Valuations
Representative Data
Data Shapley
Influence Functions
Representer Points
Prototypes
Causal-Based
Counterfactuals
Recourse


Data-based Explanations
Data Valuations
Representative Data
Data Shapley
Influence Functions
Representer Points
Prototypes
Causal-Based
Counterfactuals
Recourse

Shapley Values for Data

- Given training data D and performance metric P, the value of a data point t_1 is given by Data Shapley Value

$$\text{Value}\ (t_1) = \underbrace{\sum_{\substack{\text{Subsets S}\subset D \\ \text{not containing } t_1}}}_{\text{All possible subsets}} \frac{\overbrace{P(S \cup t_1) - P(S)}^{\text{Marginal contribution of } t_1}}{\underbrace{\binom{|D|-1}{|S|}}_{\substack{\text{\# of subsets of} \\ \text{size |S| in D}}}}$$

Expected contribution to all possible sizes of training data subsets

[2019 ICML Ghorbani et al.] Data Shapley: Equitable Valuation of Data for Machine Learning

Shapley Values for Data – Benefits

- The difference between the prediction and the average prediction is fairly distributed among the feature values of the instance

- Allows contrastive explanations; can compare value of a datum to a subset or even to a single data point

- Backed by solid theory (efficiency, symmetry, additivity axioms)

Shapley Values for Data – Challenges

- Is computationally expensive
  - For one data point: train and evaluate a model (expensive) for every possible subset of the training data (exponentially many)
- Can be misinterpreted
  - NOT the difference in the predicted values before and after removing the data point from model training
- Need access to the data to calculate the Shapley value for a new data point
- Does not account for distributional shift

Shapley Values for Data – Efficiency

- **Monte-Carlo method** (Ghorbani et al.)

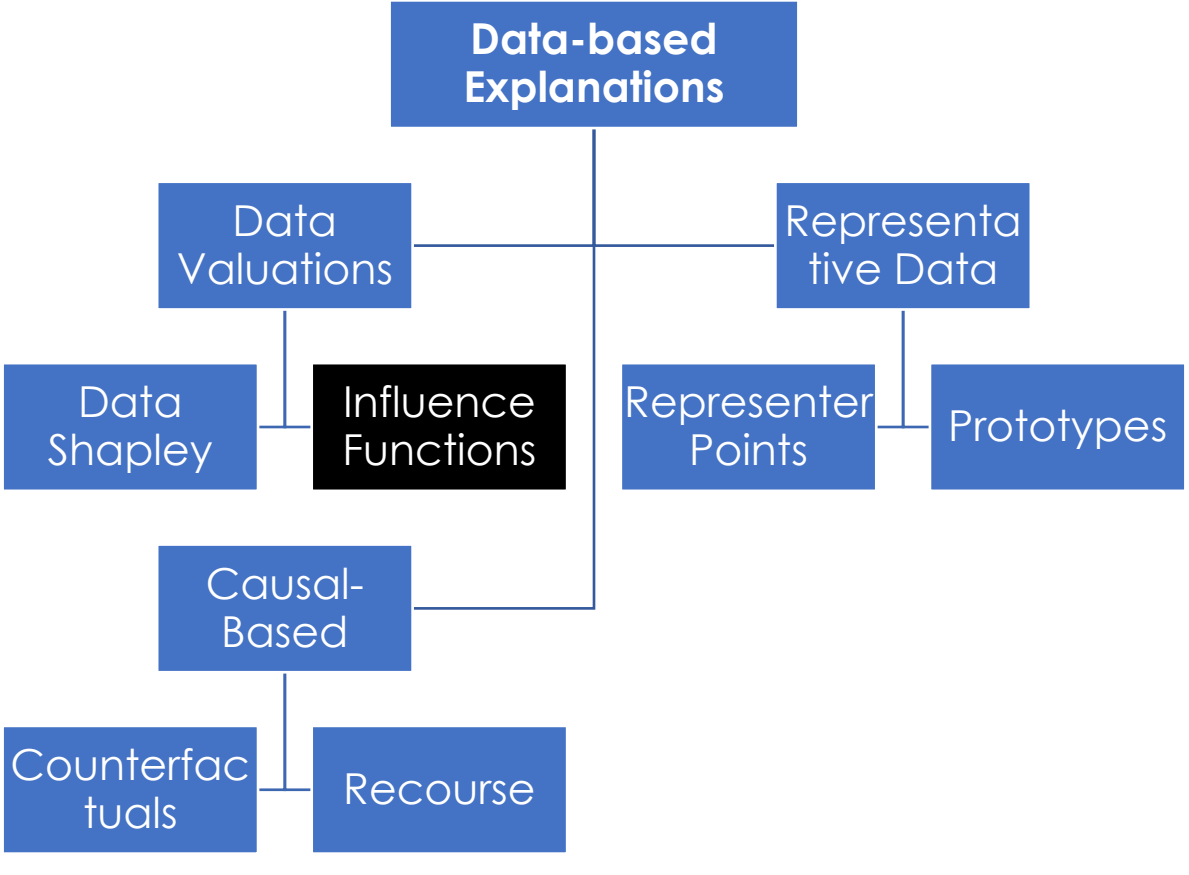
$$\text{Value}\ (t_i) = \mathbb{E}_{\pi \sim \Pi}[V(S_\pi^i \cup i) - V(S_\pi^i)]$$

Random permutation sampled from all permutations of data points

Data appearing before i in a permutation

- **Truncated Monte Carlo Shapley:** Use early stopping when calculating data Shapley for a sampled permutation

[2019 ICML Ghorbani et al.] Data Shapley: Equitable Valuation of Data for Machine Learning


Data-based Explanations
Data Valuations
Representative Data
Data Shapley
Influence Functions
Representer Points
Prototypes
Causal-Based
Counterfactuals
Recourse

Influence of a Training Data Point

- How much does each training data point contribute to the learned model?

Rosa

Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data

Original Model

83%

Probability that Rosa's income < 50K

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
Nazia	36	Bachelors	Unmarried	…	Asian	female	40	<50K
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

Influence of a Training Data Point

- Leave-one-out approach

Name	Age	Education	Marital	…	Race	Gender	Hours
Rosa	34	Bachelors	Unmarried	…	Black	female	40

Rosa's data

Updated Model

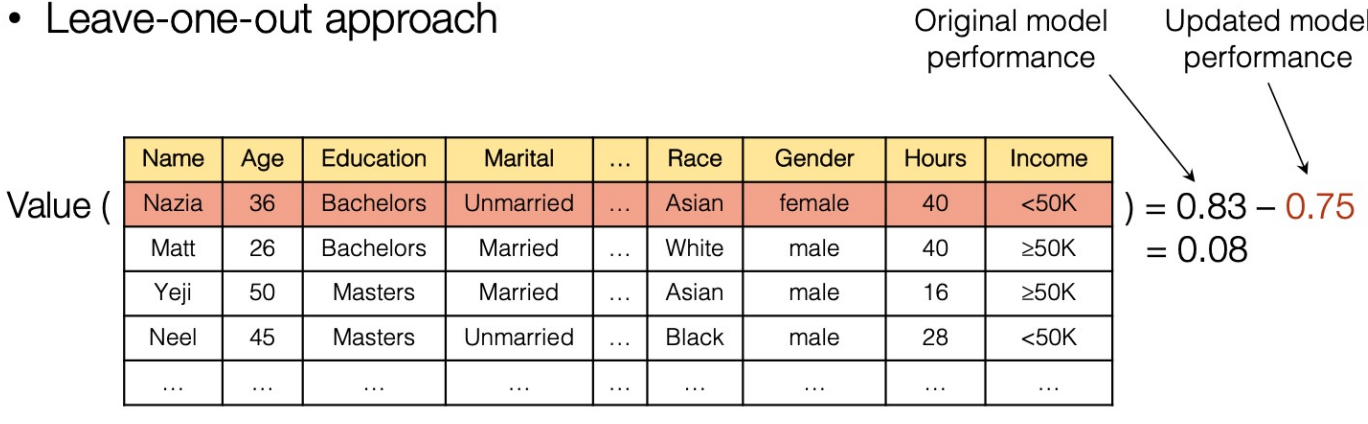
75%

Probability that Rosa's income < 50K

Name	Age	Education	Marital	…	Race	Gender	Hours	Income
~~Nazia~~	~~36~~	~~Bachelors~~	~~Unmarried~~	…	~~Asian~~	~~female~~	~~40~~	~~<50K~~
Matt	26	Bachelors	Married	…	White	male	40	≥50K
Yeji	50	Masters	Married	…	Asian	male	16	≥50K
Neel	45	Masters	Unmarried	…	Black	male	28	<50K
…	…	…	…	…	…	…	…	…

Influence of a Training Data Point

- Leave-one-out approach



- Might be expensive to compute for each data point
- Not reasonable when training data has duplicates

Influence Function

- Given data points $z \in Z$ and loss function L, the learning algorithm searches for a model with parameters $\theta \in \Theta$ such that

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(z_i, \theta)$$

- Influence of training data point z_{train} (from robust statistics)
  - How do the model parameters change as we upweight z_{train} by an infinitesimal amount ϵ?

$$\theta_\epsilon = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(z_i, \theta) + \epsilon L(z_{\text{train}}, \theta)$$

[2017 ICML Koh et al.] Understanding Black-box Predictions via Influence Functions

Influence Function

- Original model parameters: θ, updated model parameters: θ_ϵ

- Influence of z_{train} on test loss is then measured as

$$\mathcal{I}_{\text{up,loss}}(z_{\text{train}}, z_{\text{test}}) = L(z_{test}, \theta_\epsilon) - L(z_{test}, \theta^*)$$

$$\stackrel{\text{def}}{=} \frac{dL(z_{\text{test}}, \hat{\theta}_{\epsilon, z_{\text{train}}})}{d\epsilon}\Big|_{\epsilon=0}$$

$$= -\underbrace{\nabla_\theta L(z_{\text{test}}, \hat{\theta})^\top}_{\text{Gradient of test loss}} H_{\hat{\theta}}^{-1} \underbrace{\nabla_\theta L(z_{\text{train}}, \hat{\theta})}_{\text{Gradient of training data loss}},$$

where $\underbrace{H_{\hat{\theta}} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 L(z_i, \hat{\theta})}_{\text{Hessian matrix}}.$

Influence Function – Challenges

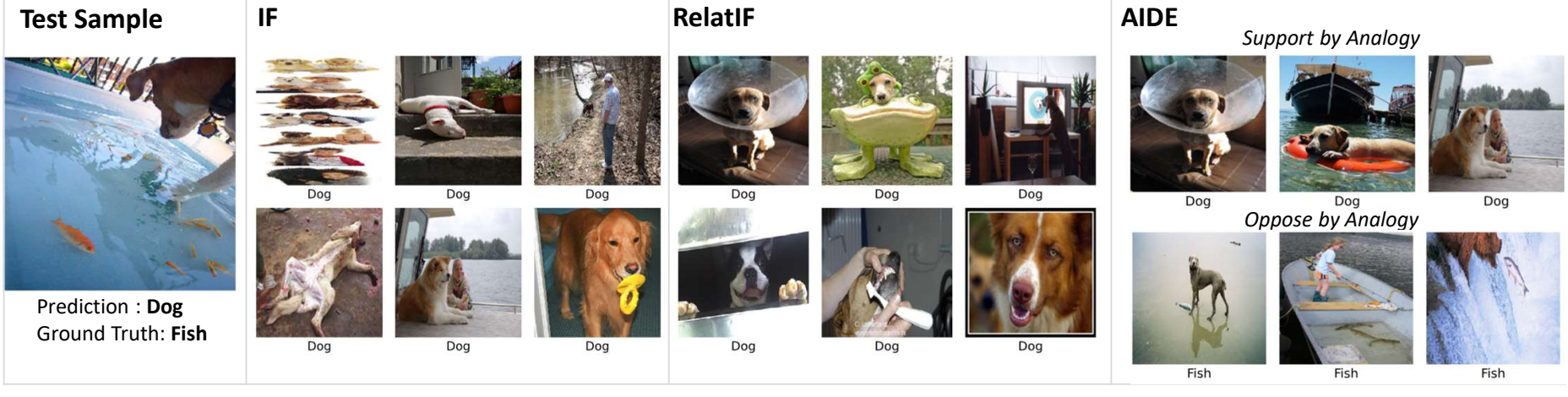
- Computationally **expensive**
  - Computing and inverting the Hessian matrix
    - $(O(np^2 + p^3))$, n: #data points, p: #model parameters
  - Computing the influence over test loss for all training points

- Applicable to models with **twice-differentiable loss function**
  - Alternatives for non-parametric gradient-boosted decision trees

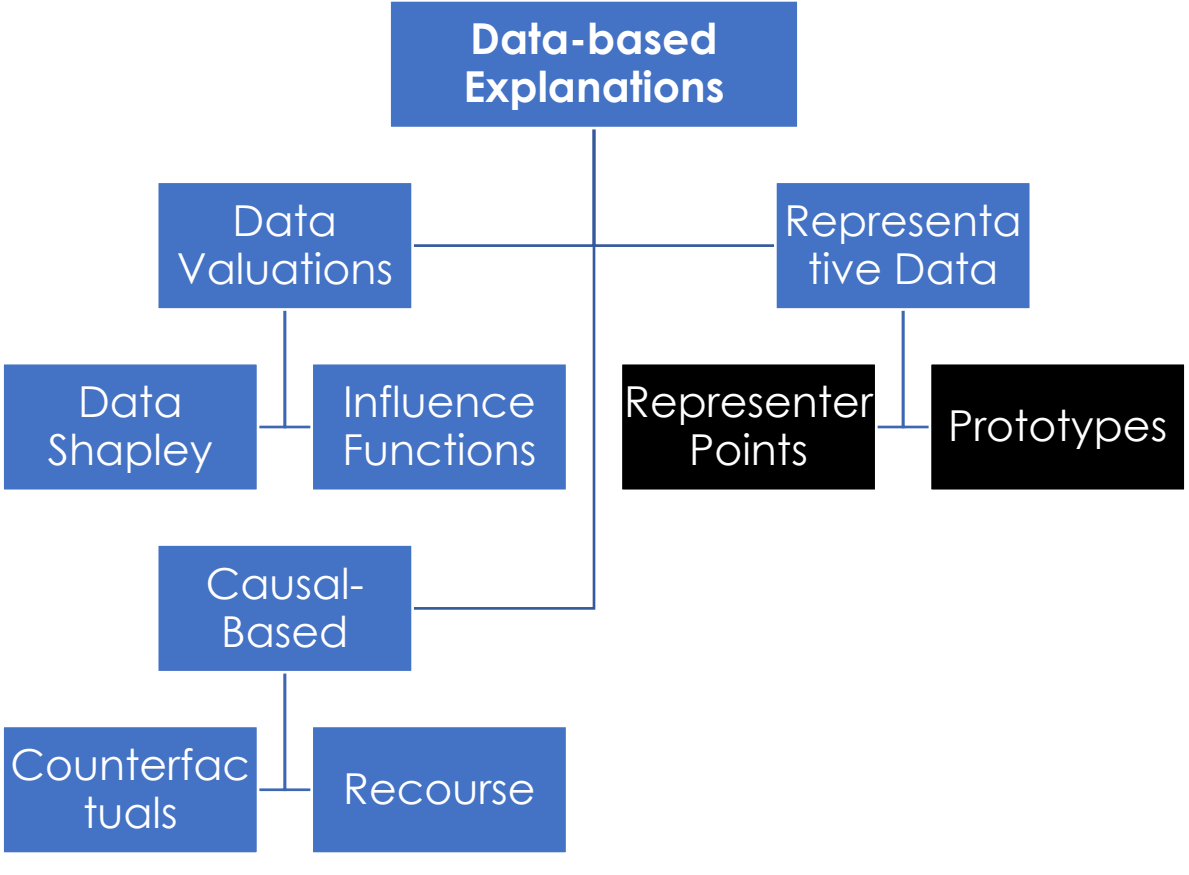
- **Influence functions are approximate:** first-order approximations; the approach forms a quadratic expansion around the parameters

- No clear threshold to decide when a training data point is influential or non-influential

Influence Function – Example



[Nematov et al.] AIDE: Antithetical, Intent-based, and Diverse Explanations


Data-based Explanations
Data Valuations
Representative Data
Data Shapley
Influence Functions
Representer Points
Prototypes
Causal-Based
Counterfactuals
Recourse

Representer Points

- Representer points
  - Decompose a predictor into a linear combination of functions of training data points
  - Weighted similarities between test data point and each training data point

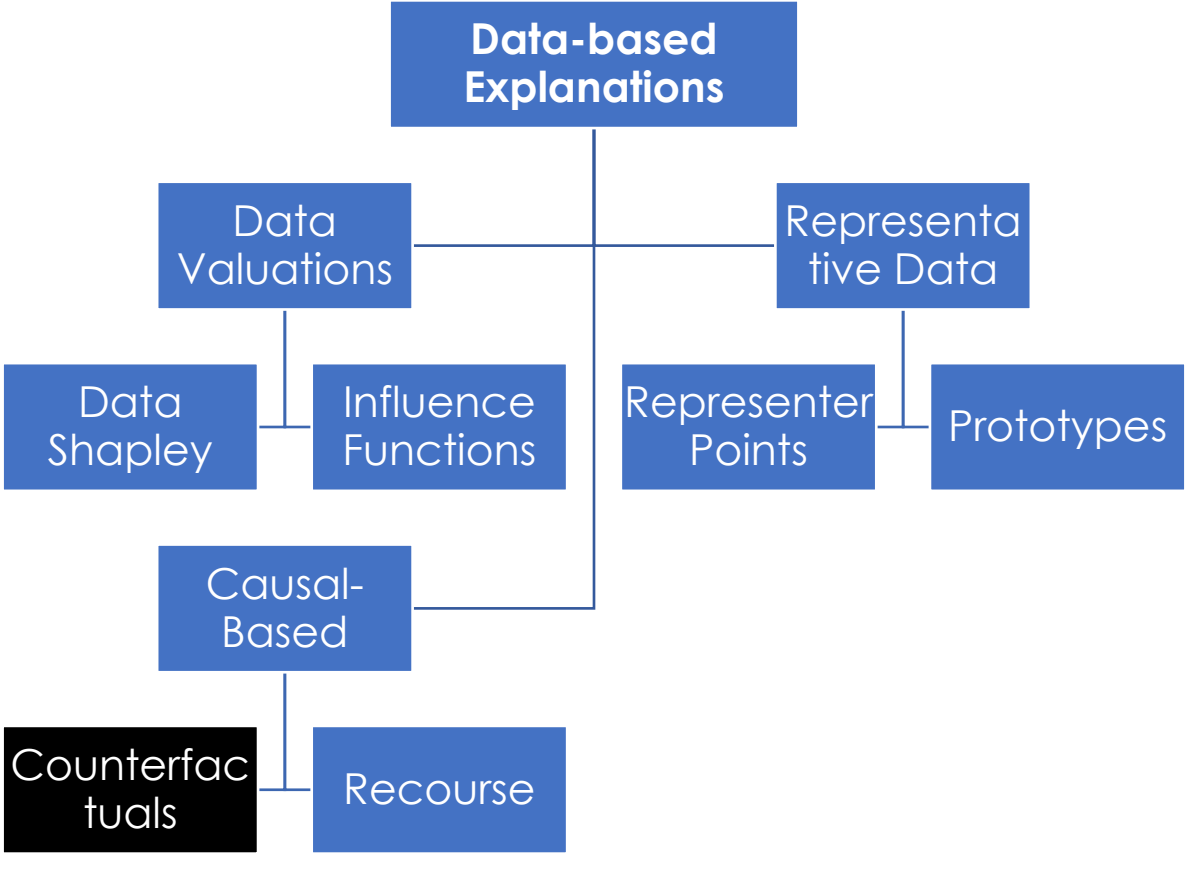
[2018 NeurIPS Yeh et al.] Representer Point Selection for Explaining Deep Neural Networks

Data Prototypes

- **Prototypes:** data points that are representative of all the data

- **Criticisms:** data points not represented by prototypes

[2016 NeurIPS Kim et al.] Examples are not enough, learn to criticize! Criticism for interpretability


Data-based Explanations
Data Valuations
Representative Data
Data Shapley
Influence Functions
Representer Points
Prototypes
Causal-Based
Counterfactuals
Recourse

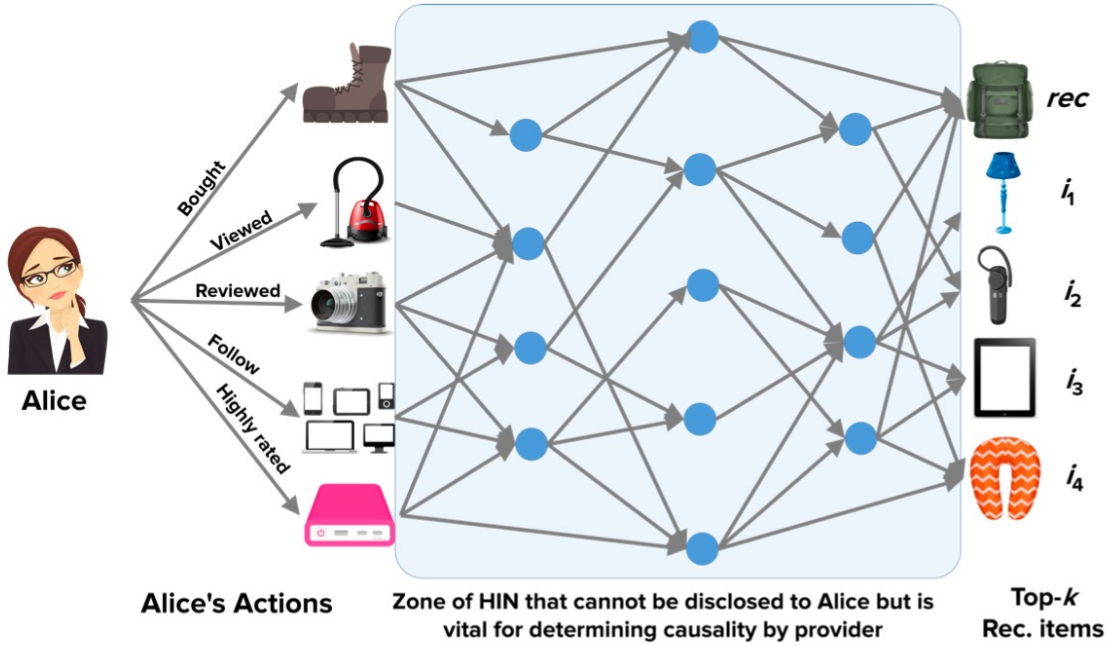
Motivation – Causality

- consider a **causal relationship**: "*If X had not occurred, Y would not have occurred*"
- it explains why Y occurred: it's because X occurs

- a **counterfactual explanation** of a specific **outcome** describes the smallest **change** to **preferences** that results in **not seeing** that outcome
  - Example: to explain "Why was I recommended item Y?" look for smallest changes in preferences so that item Y no longer appears in the recommendations

- Original input = **factual**
- Fictitious, changed input = **counterfactual**

Motivation – Causality



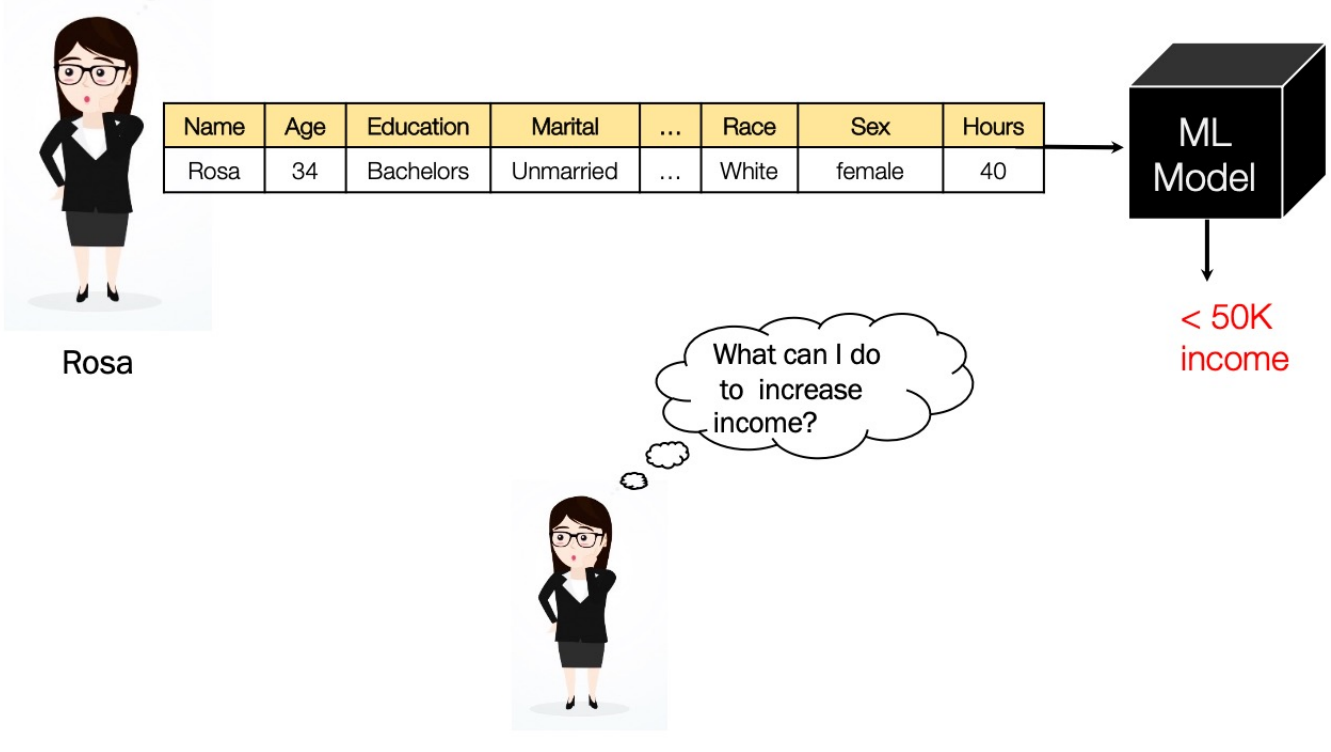
Alice: Why did I receive this recommendation "Jack Wolfskin backpack"?

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".

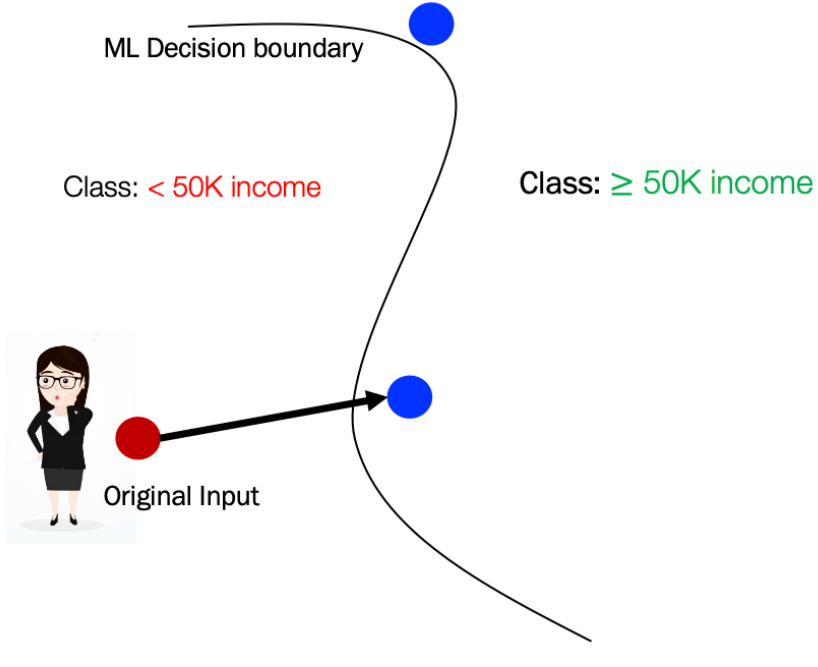
[2020 WSDM A. Ghazimatin et al.] *PRINCE: Provider-side Interpretability with Counterfactual Explanations in Recommender Systems*

Counterfactual Explanations



Counterfactual Explanations

- What features need to be changed and by how much to flip a model's prediction?



Counterfactual Explanations

Goal: **Minimum change in input attributes to flip a model's prediction**

Popular Techniques

- Feature-based explanations: Feature score captures its likelihood to change outcome
- Instance-level explanations: Identify counterfactual scenarios to change outcome
  - Inverse Classification
  - Nearest-counterfactual explanation

Inverse Classification

- Identify a close neighbor of the point which has the opposite outcome

Name	Age	Education	Marital	…	Race	Sex	Hours
Rosa	34	Bachelors	Unmarried	…	White	female	40

ML Model → < 50K income

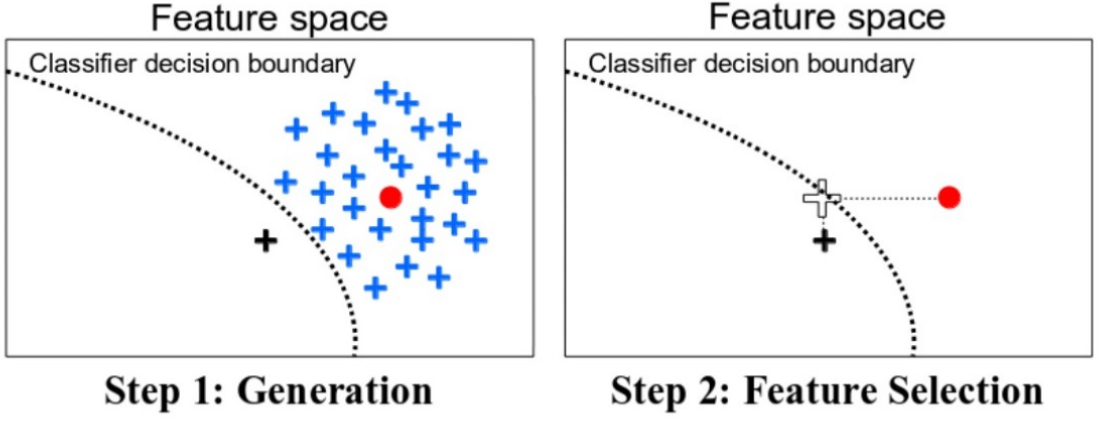
Name	Age	Education	Marital	…	Race	Sex	Hours
Counterfactual Rosa	34	Bachelors	Unmarried	…	White	female	65

ML Model → ≥ 50K income

[IPMU 2018 Laugel et al.] Inverse Classification for Comparison-based Interpretability in Machine Learning

Inverse Classification

- Identify a close neighbor of the point which has the opposite outcome
- Growing sphere algorithm

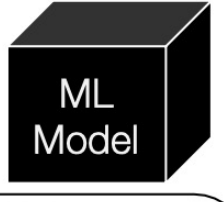


[IPMU 2018 Laugel et al.] Inverse Classification for Comparison-based Interpretability in Machine Learning

DICE: Diverse Counterfactual Explanations

- Post-hoc, "what-if" style of explanations
- Generate multiple diverse counterfactuals that receive a favorable outcome

Name	Age	Education	Marital	…	Race	Sex	Hours
Rosa	34	Bachelors	Unmarried	…	White	female	40



Counterfactual Examples by DiCE

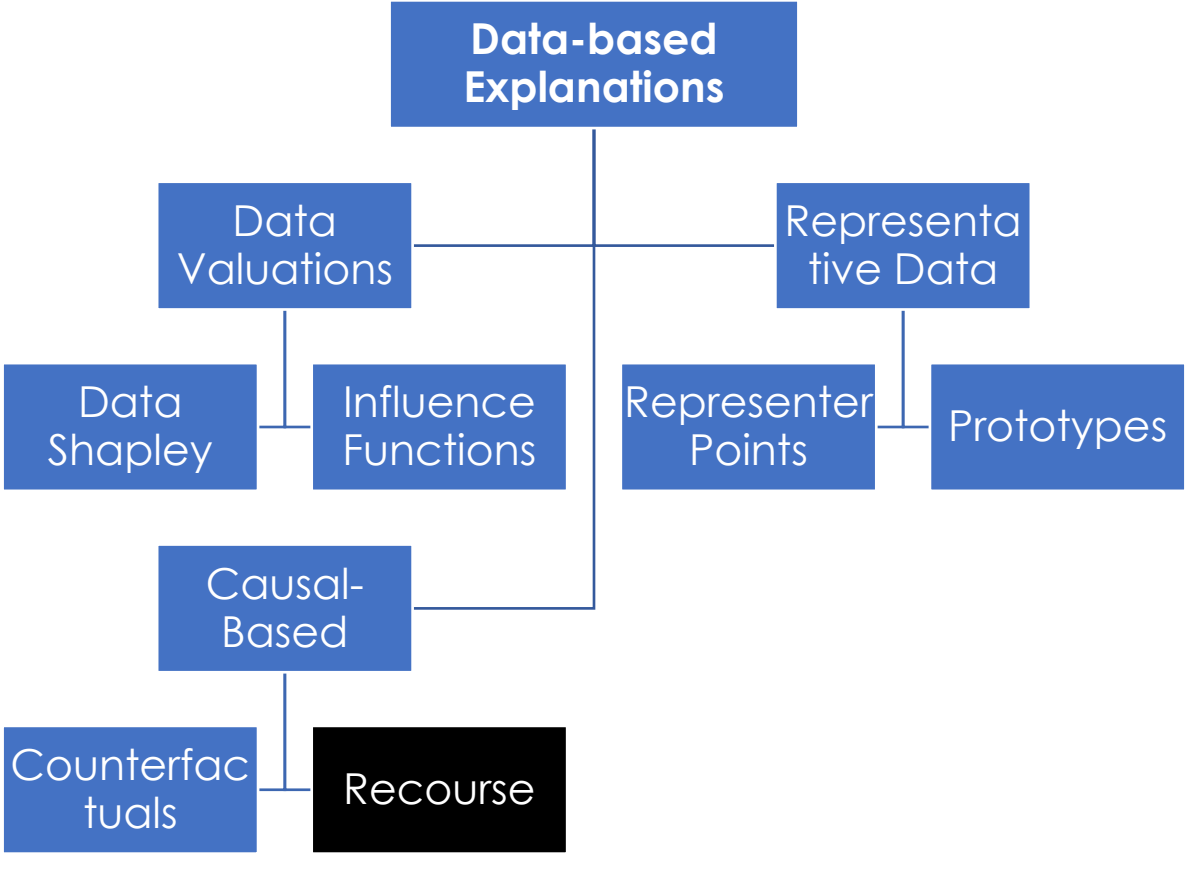
1. Increase hours to at least 65
2. Change education to Masters and increase work experience by 1 year

[2020 ACM FAT* Mothilal et al.] Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations

DICE – Algorithm

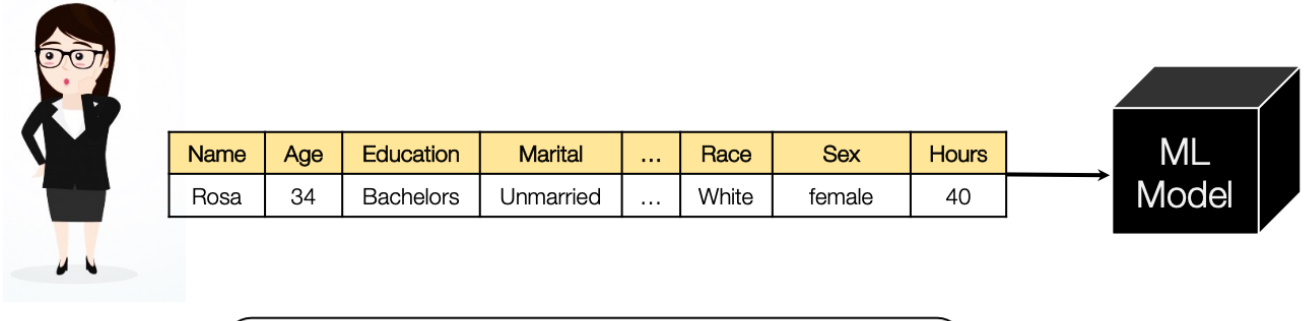
- Model the search space of counterfactuals as an optimization problem

- Three different objectives
  - Proximity to original input
  - Sparsity of change
  - Diversity of explanation examples


Data-based Explanations
Data Valuations
Representative Data
Data Shapley
Influence Functions
Representer Points
Prototypes
Causal-Based
Counterfactuals
Recourse

Algorithmic Recourse

- **Input**: Individual with negative outcome
- **Goal**: Identify smallest cost intervention that can flip the outcome in the future



Algorithmic Recourse

- Model recourse as an optimization problem
- Linear Program:
  - Objective:
    Minimize cost(

Age	Education	Marital	...	Hours
a1	a2	a3	...	ak

)

  - Constraint

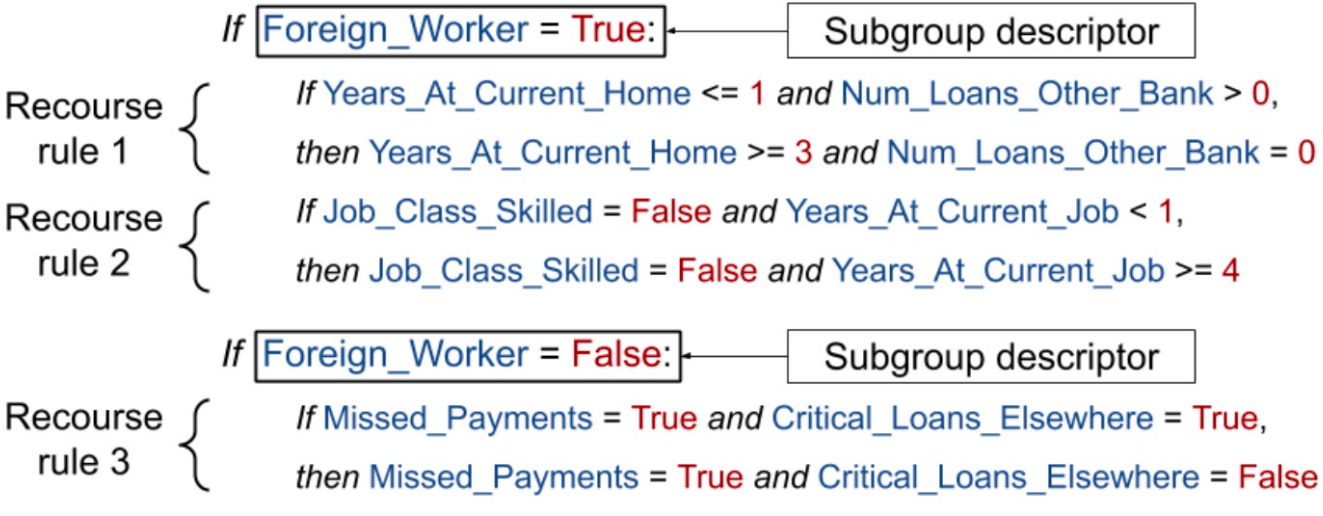
$$f\left(\begin{array}{|c|c|c|c|c|c|c|c|}\hline \text{Name} & \text{Age} & \text{Education} & \text{Marital} & \ldots & \text{Race} & \text{Sex} & \text{Hours} \\ \hline \text{Rosa} & 34 & \text{Bachelors} & \text{Unmarried} & \ldots & \text{White} & \text{female} & 40 \\ \hline \end{array} + \begin{array}{|c|c|c|c|c|}\hline \text{Age} & \text{Education} & \text{Marital} & \ldots & \text{Hours} \\ \hline a1 & a2 & a3 & \ldots & ak \\ \hline \end{array}\right) = 1$$

- Assumes f is a linear function

[2019 FAT Ustun et al.] Actionable Recourse in Linear Classification

ARES – Global Recourse Summary

- explain not a specific instance but the whole model (global explainability)



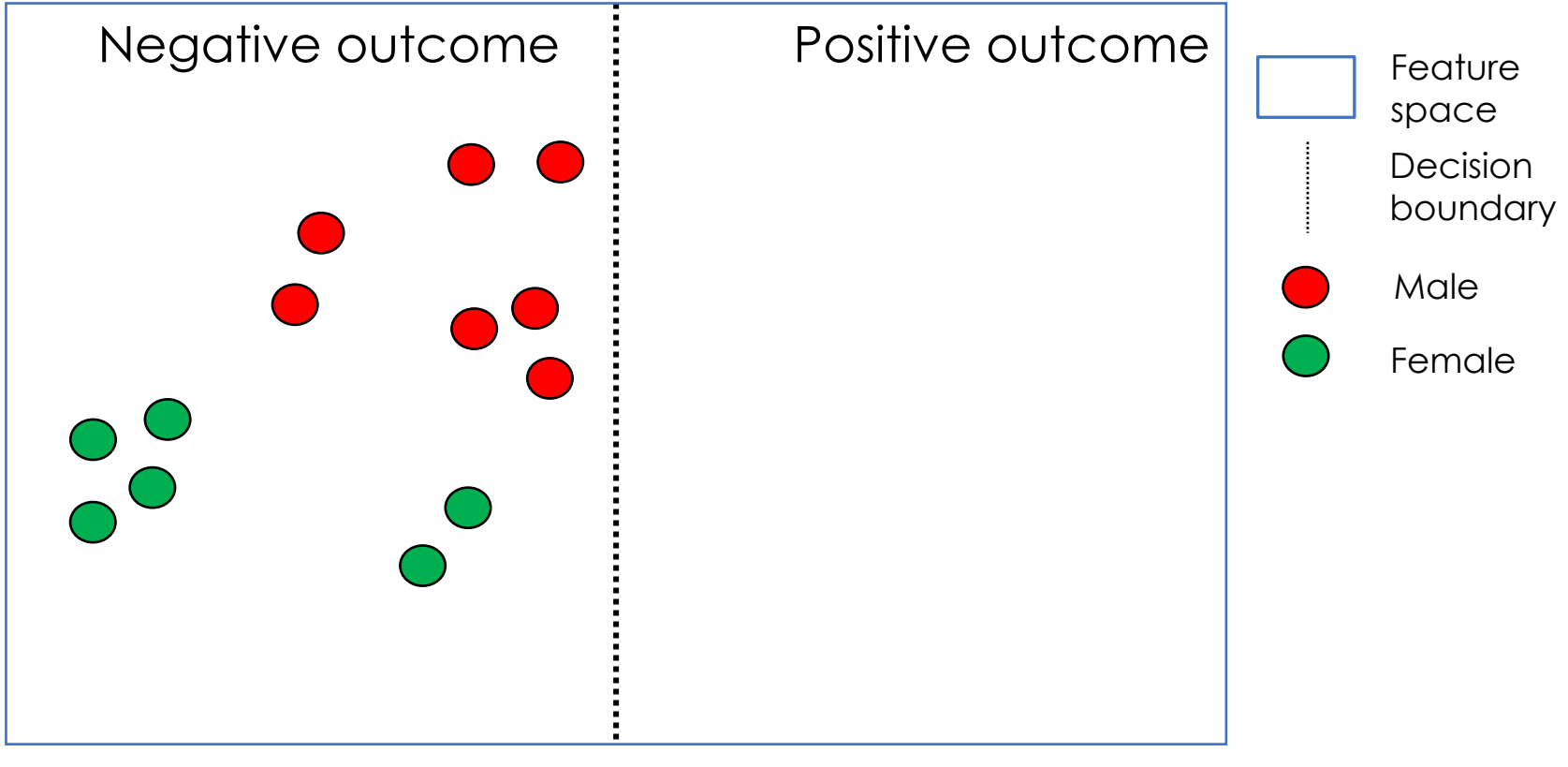
[2020 NeurIPS Rawal et al.] Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses

FACTS – Recourse and Fairness

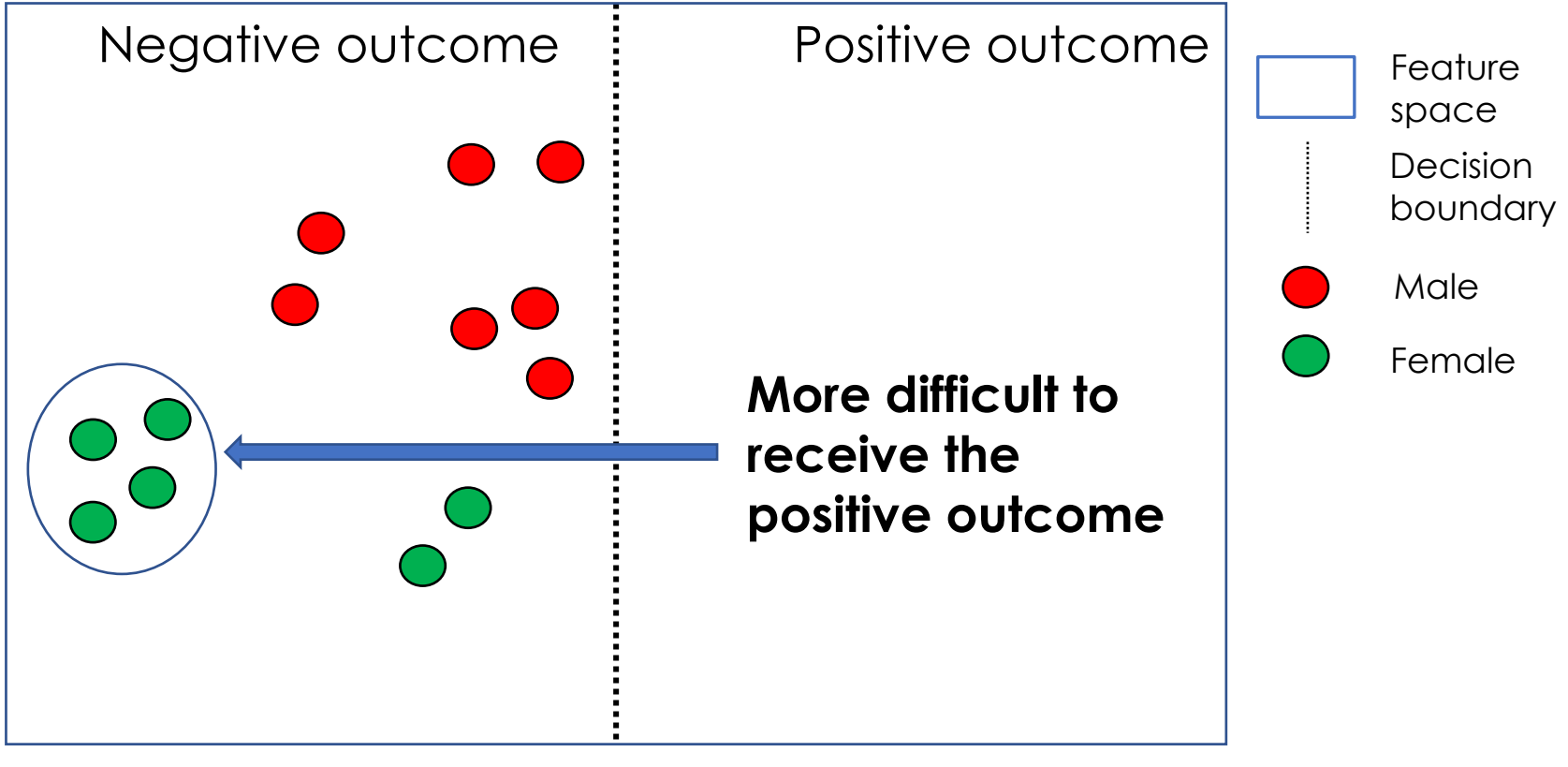
- recourse = the way to reverse an unfavorable outcome
- fairness of recourse = does the recourse cost the same?

[2023 NeurIPS Kavouras et al.] Fairness-Aware Counterfactuals for Subgroups

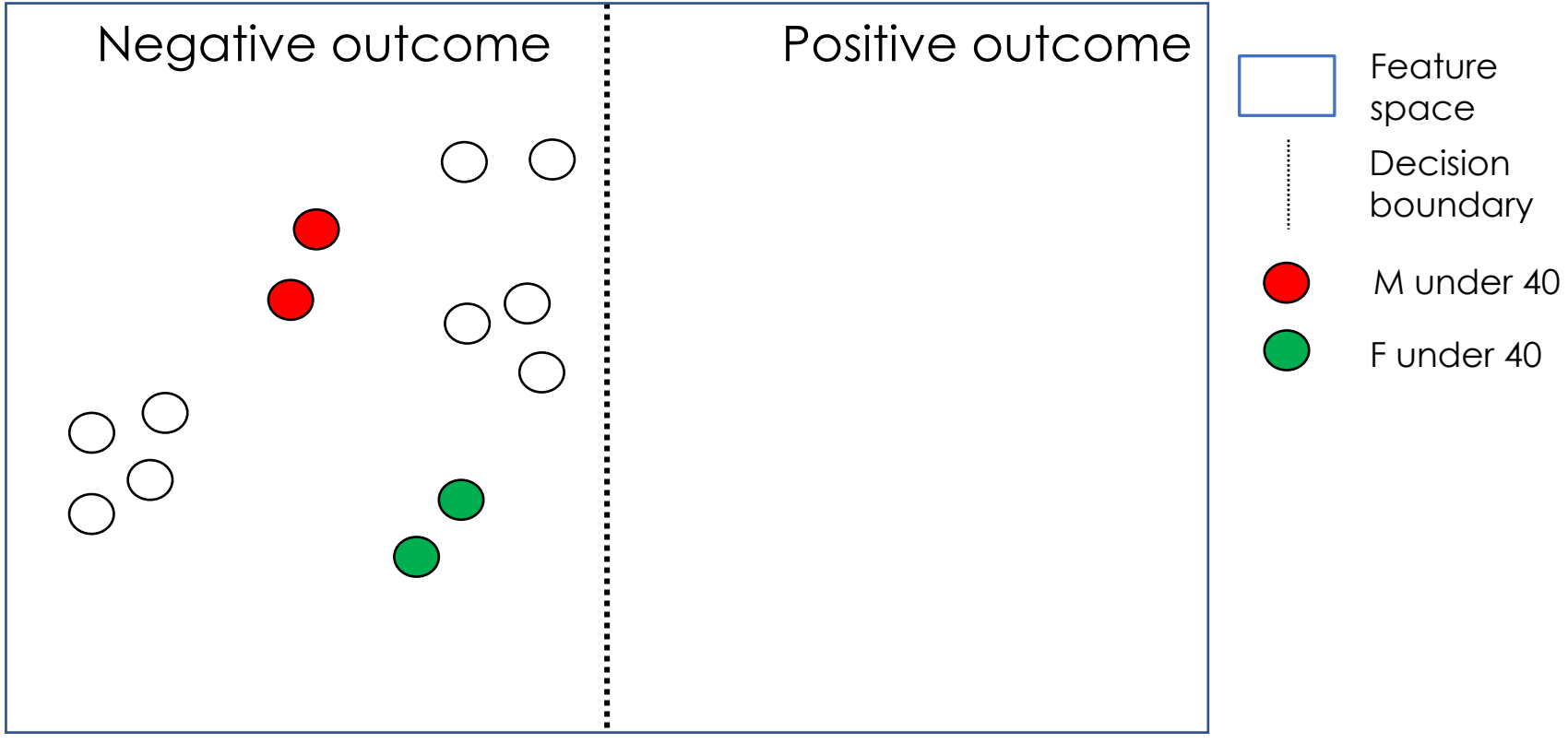
FACTS – Recourse



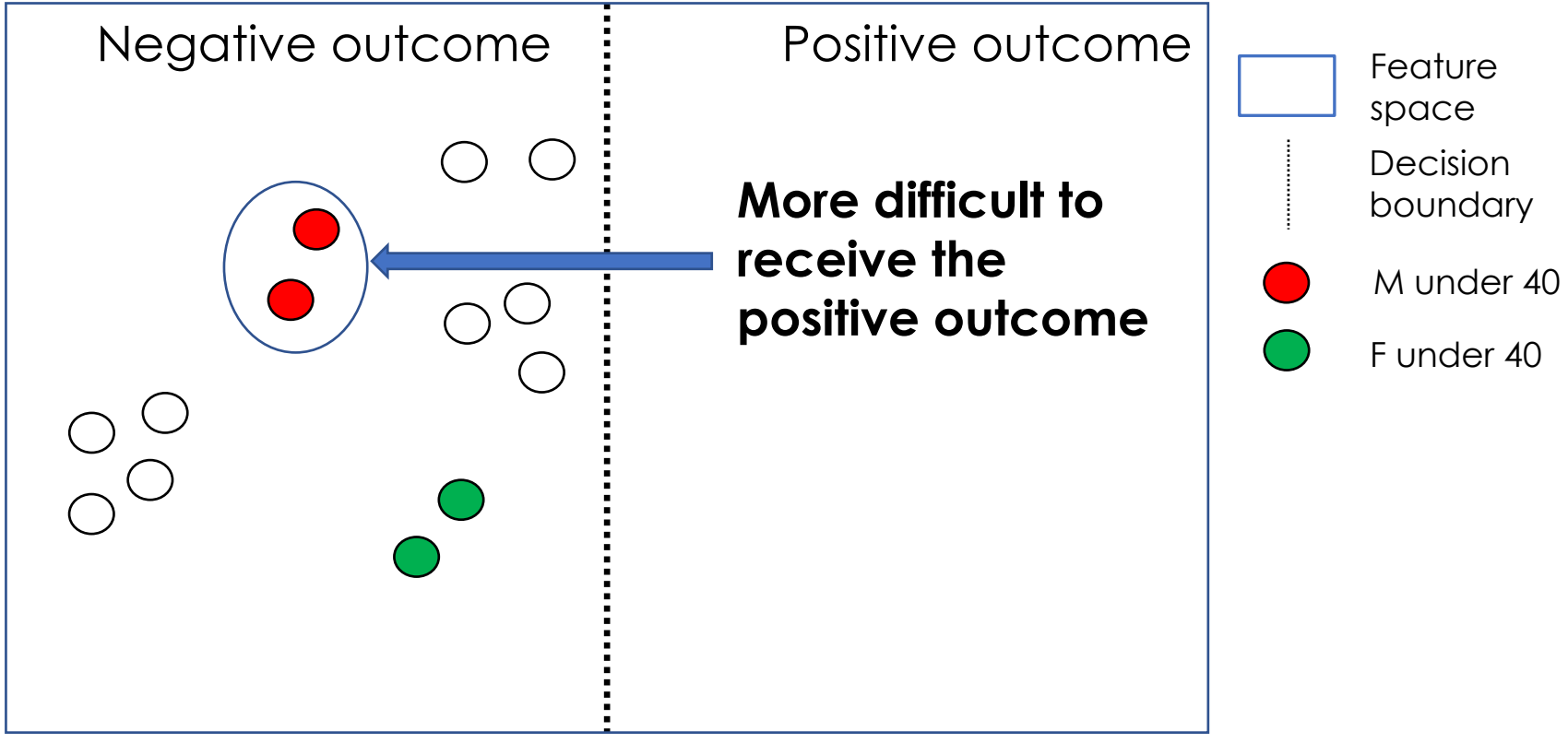
FACTS – Fairness of recourse



FACTS – Fairness of recourse in subgroups



FACTS – Fairness of recourse in subgroups



FACTS – Definitions for Fair Recourse

- Equal Effectiveness
- Equal Choice for Recourse
- Equal Effectiveness within Budget
- Equal Cost of Effectiveness
- Fair Effectiveness-Cost Trade-off
- Equal Mean Recourse

FACTS – Example

```
If capital-gain = 0, education-num = 10, hours-per-week = FullTime, race =  White:
        Protected Subgroup ' Male', 8.40% covered
                Make hours-per-week = OverTime with effectiveness 8.92% and counterfactual cost = 2.0.
                Make education-num = 11 with effectiveness 8.92% and counterfactual cost = 3.0.
                Make hours-per-week = BrainDrain with effectiveness 16.98% and counterfactual cost = 4.0.
                Make education-num = 11, hours-per-week = OverTime with effectiveness 23.16% and counterfactual cost = 5.0.
                Make education-num = 12 with effectiveness 23.16% and counterfactual cost = 6.0.
                Make education-num = 12, hours-per-week = OverTime with effectiveness 31.56% and counterfactual cost = 8.0.
                Aggregate cost of the above recourses = 8.00
        Protected Subgroup ' Female', 8.32% covered
                Make hours-per-week = OverTime with effectiveness 1.79% and counterfactual cost = 2.0.
                Make education-num = 11 with effectiveness 1.79% and counterfactual cost = 3.0.
                Make hours-per-week = BrainDrain with effectiveness 3.27% and counterfactual cost = 4.0.
                Make education-num = 11, hours-per-week = OverTime with effectiveness 4.46% and counterfactual cost = 5.0.
                Make education-num = 12 with effectiveness 4.46% and counterfactual cost = 6.0.
                Make education-num = 12, hours-per-week = OverTime with effectiveness 5.95% and counterfactual cost = 8.0.
                Aggregate cost of the above recourses = inf
        Bias against  Female. Unfairness score = inf.
```

[2023 NeurIPS Kavouras et al.] Fairness-Aware Counterfactuals for Subgroups

Acknowledgements

- Some material from [2022 ACM SIGMOD Pradhan et al.] Explainable AI: Foundations, Applications, and Opportunities for Data Management Research https://explainable-ai-tutorial.github.io