Explainability for Machine Learning Models: From Data Adaptability to User Perception

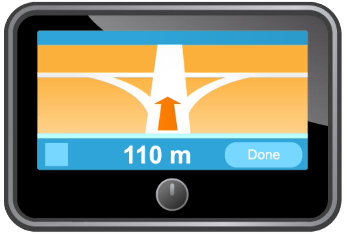
Julien Delaunay

Reviewer:	Marie-Jeanne Lesot, Professor at Sorbonne University
	Andrea Passerini, Associate Professor at Trento University
Jury Member:	Elisa Fromont, Professor at Rennes University
	Pierre Marquis, Professor at Artois University
	Niels van Berkel, Associate Professor at Aalborg University
	Katrien Verbert, Professor at KU Leuven
Director:	Christine Largouët, Associate Professor at Institut Agro
Supervisor:	Luis Galárraga, Researcher at Inria Rennes

December 20, 2023

What Machine Learning Models Do?

What Machine Learning Models Do?



Supervised Machine Learning

Training Dataset

	Age	Tension	Gender	Weight	Level of Insulin
	Features				Class
A	28	150	Female	58	High
B	22	160	Male	65	Low
C	54	155	Female	52	Low
D	72	170	Male	75	High
E	18	170	Male	65	Low

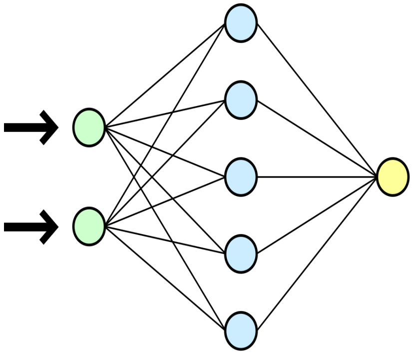
Supervised Machine Learning

Training Dataset

Features | Class

	Age	Tension	Gender	Weight	Level of Insulin
A	28	150	Female	58	High
B	22	160	Male	65	Low
C	54	155	Female	52	Low
D	72	170	Male	75	High
E	18	170	Male	65	Low

Learning Model

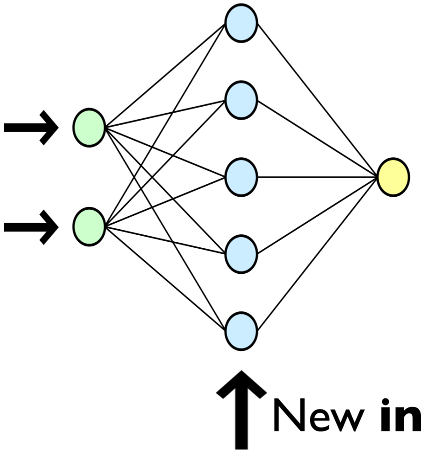


Supervised Machine Learning

Training Dataset

	Features				Class
	Age	**Tension**	**Gender**	**Weight**	**Level of Insulin**
A	28	150	Female	58	**High**
B	22	160	Male	65	**Low**
C	54	155	Female	52	**Low**
D	72	170	Male	75	**High**
E	18	170	Male	65	**Low**

Learning Model



New **instance**

F	45	165	Female	55

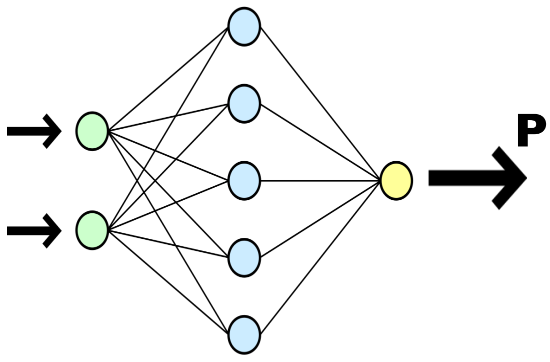
Supervised Machine Learning

Training Dataset

Features | Class

	Age	Tension	Gender	Weight	Level of Insulin
A	28	150	Female	58	High
B	22	160	Male	65	Low
C	54	155	Female	52	Low
D	72	170	Male	75	High
E	18	170	Male	65	Low

Learning Model



Prediction: Low

New **instance**

F	45	165	Female	55

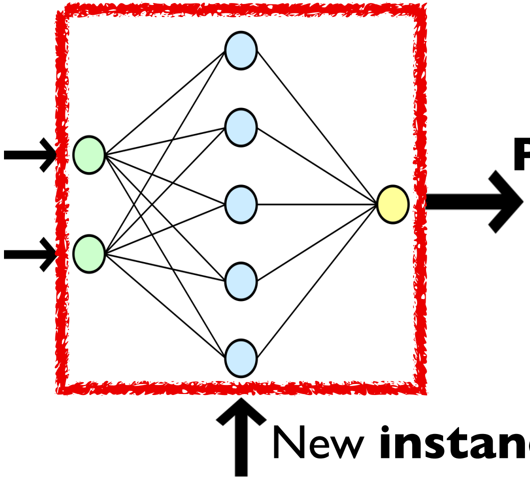
Supervised Machine Learning

Training Dataset

Features / Class

	Age	Tension	Gender	Weight	Level of Insulin
A	28	150	Female	58	High
B	22	160	Male	65	Low
C	54	155	Female	52	Low
D	72	170	Male	75	High
E	18	170	Male	65	Low

Explaining Model



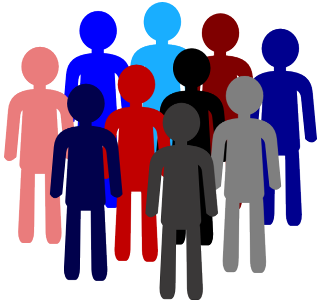
Prediction: Low

New **instance**

F	45	165	Female	55

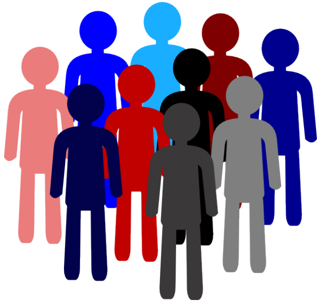
Machine Learning Models Are Used In High-Stakes Tasks

Machine Learning Models Are Used In High-Stakes Tasks

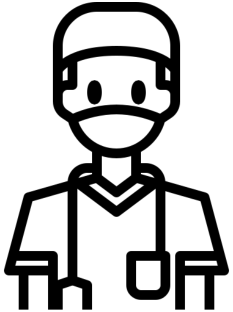


Diabetic patients

Machine Learning Models Are Used In High-Stakes Tasks

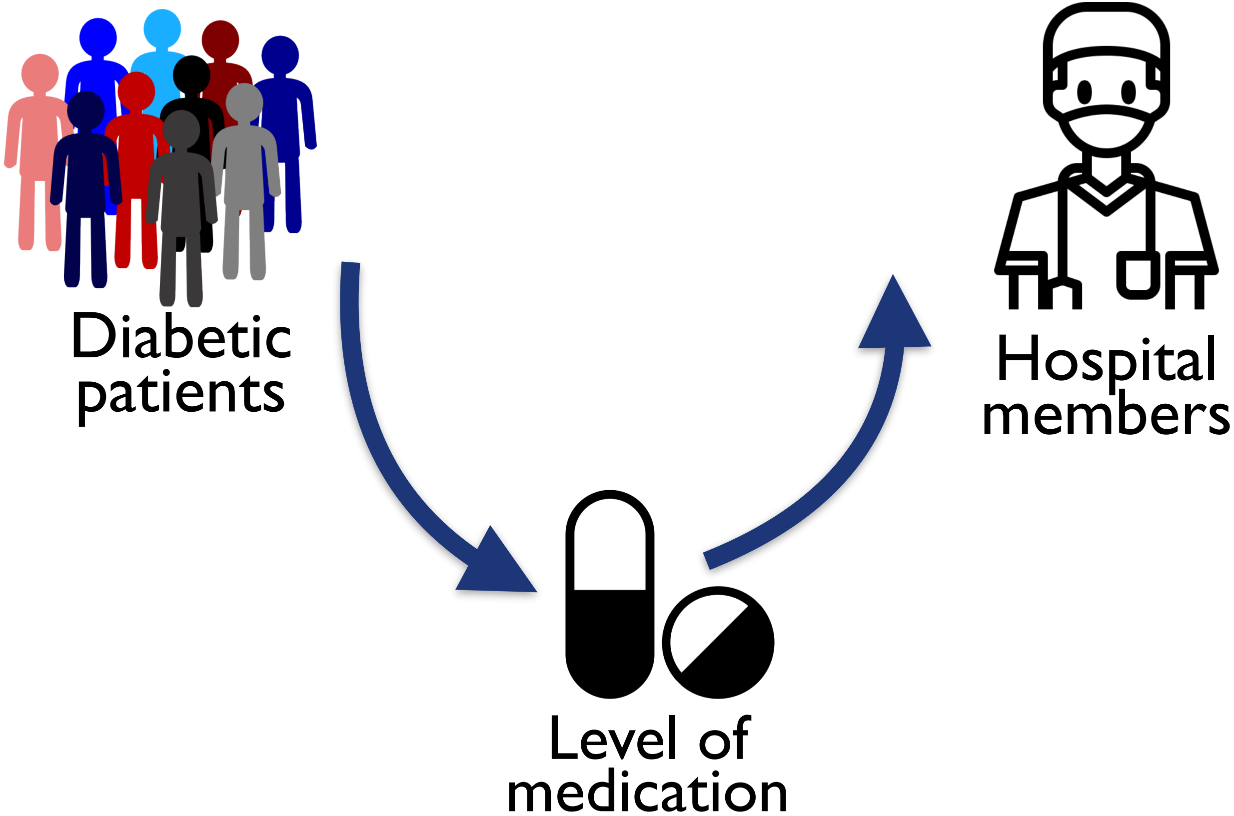


Diabetic patients

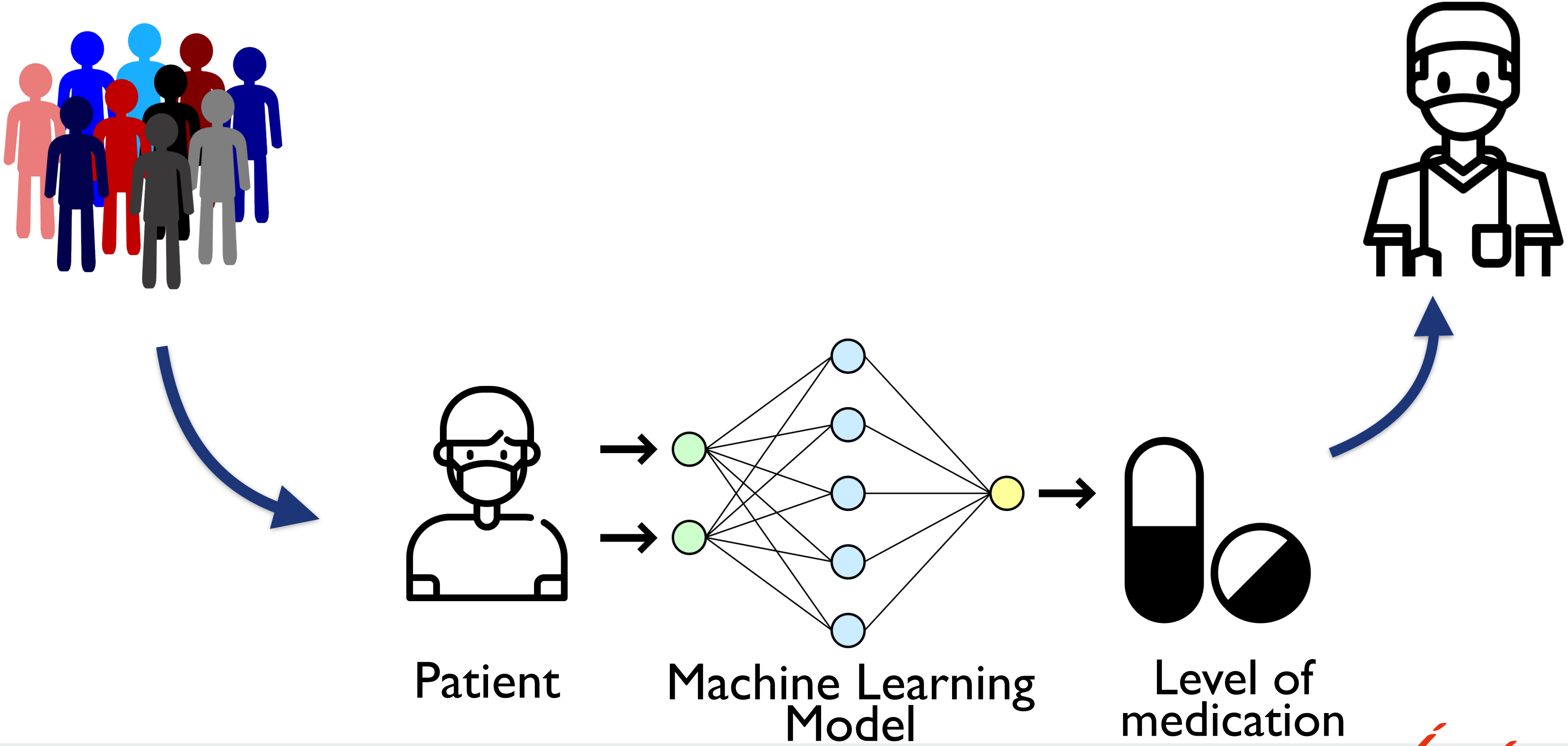


Hospital members

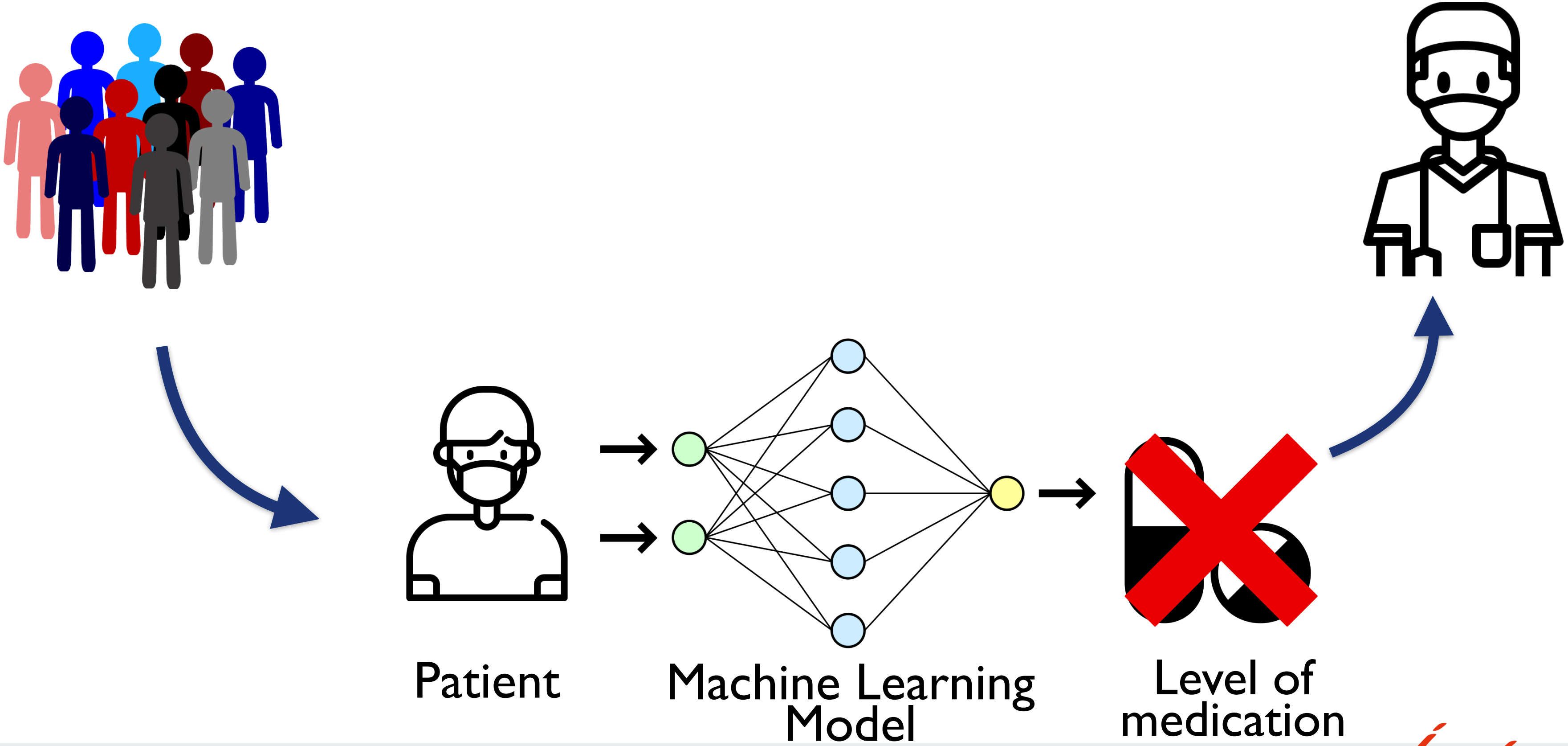
Machine Learning Models Are Used In High-Stakes Tasks



Machine Learning Models Are Used In High-Stakes Tasks

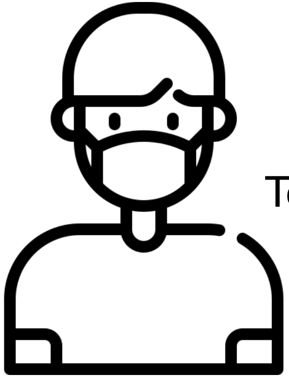


Machine Learning Models Are Used In High-Stakes Tasks



Why Do We Need Explanations?

Why Do We Need Explanations?



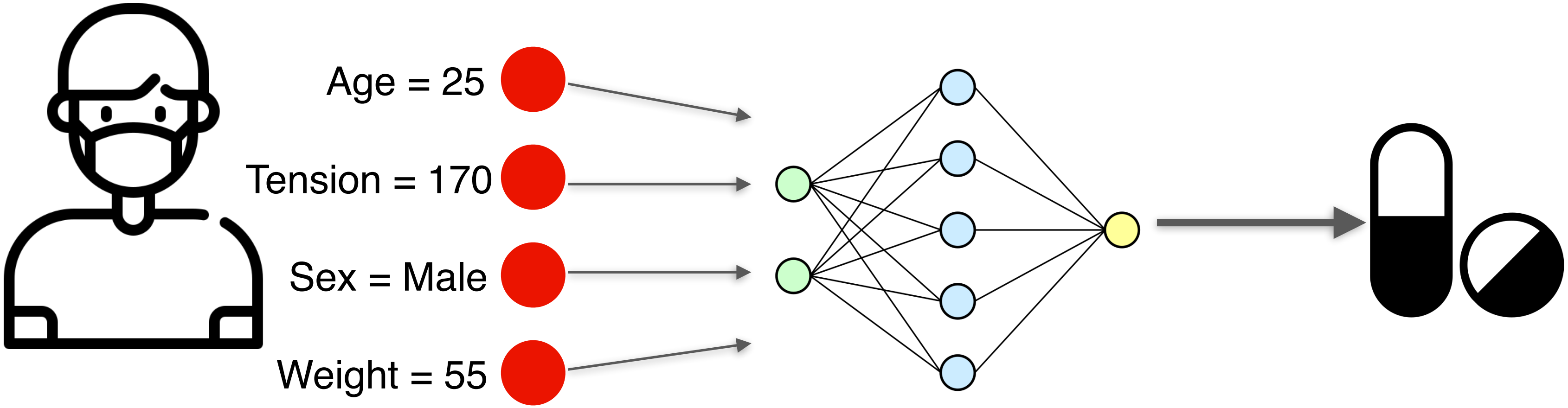
Age = 25

Tension = 170

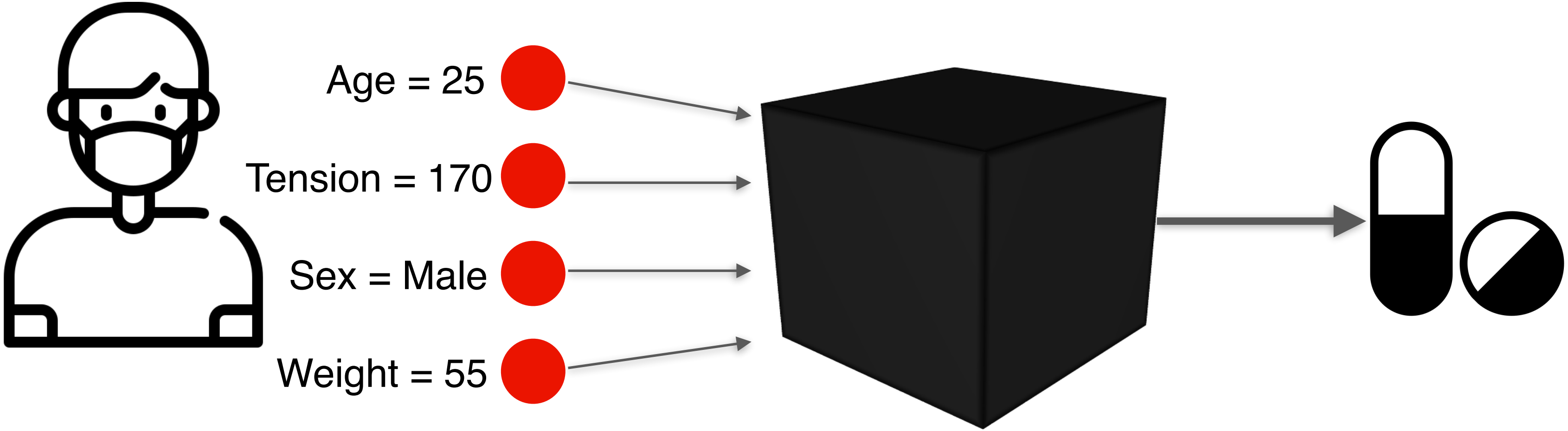
Sex = Male

Weight = 55

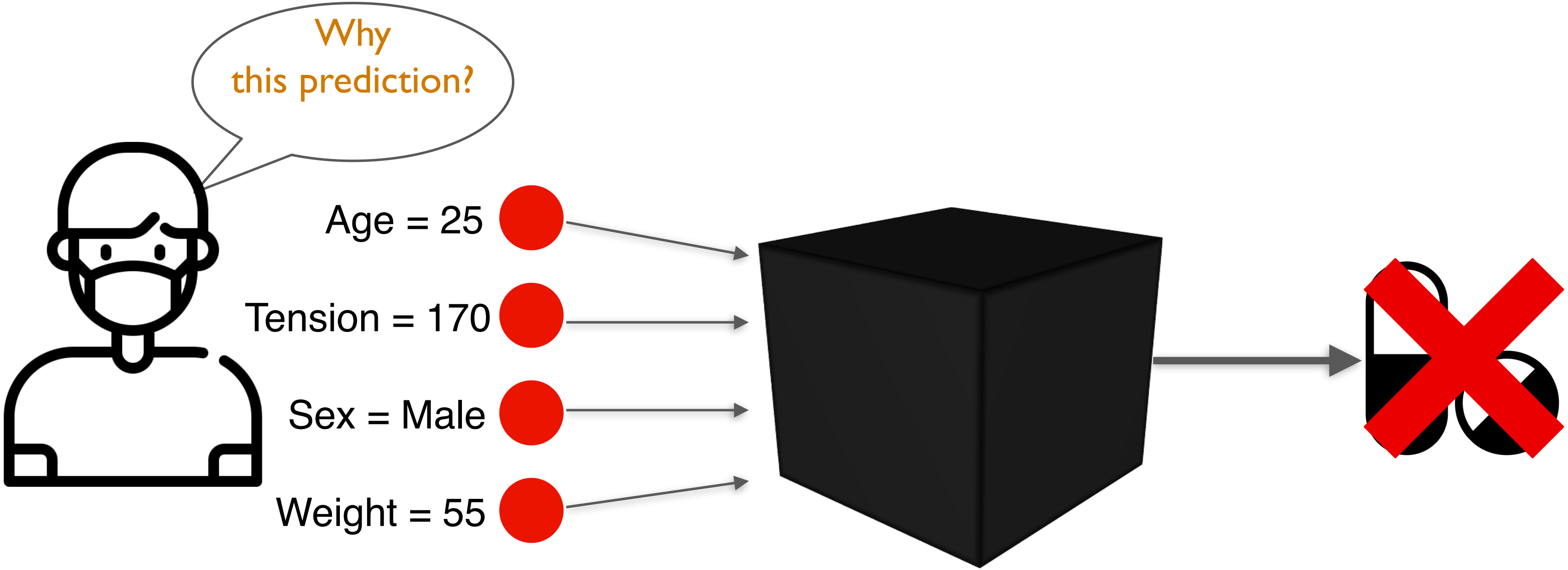
Why Do We Need Explanations?



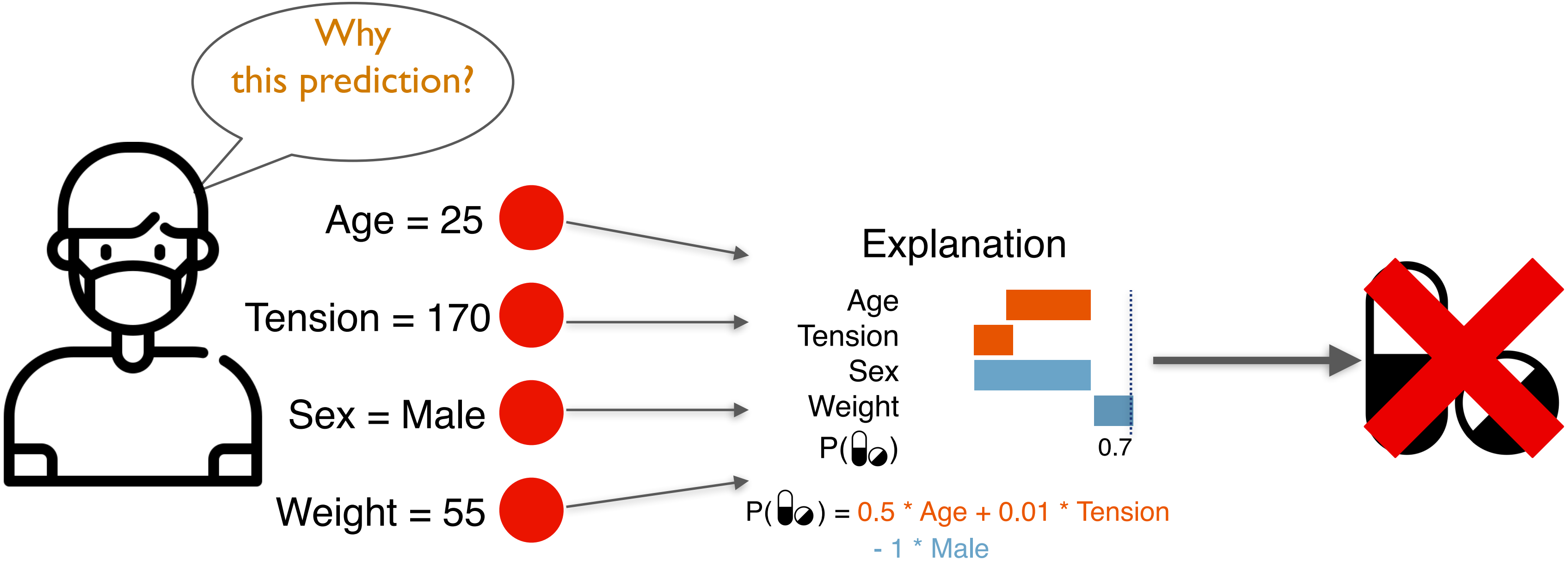
Why Do We Need Explanations?



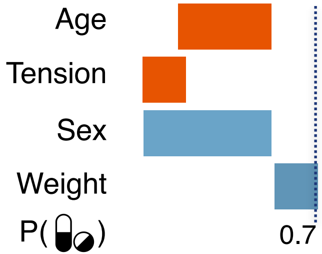
Why Do We Need Explanations?



Why Do We Need Explanations?



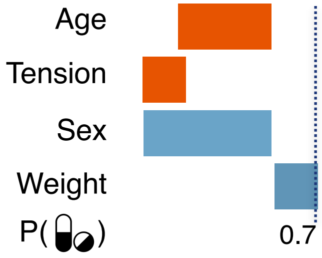
Various Types of Explanation Techniques



P( ) = 0.5 * Age +
0.01 * Tension
- 1 * Male

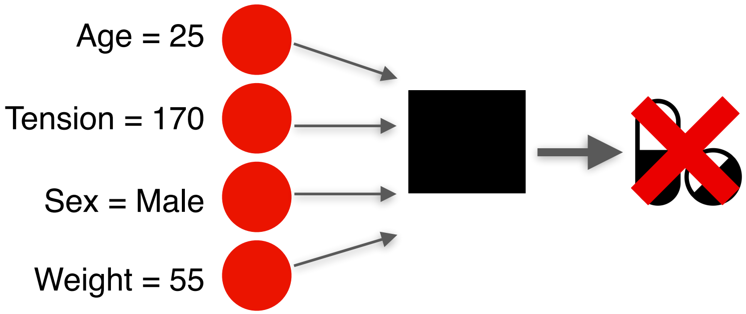
(Feature Attribution)

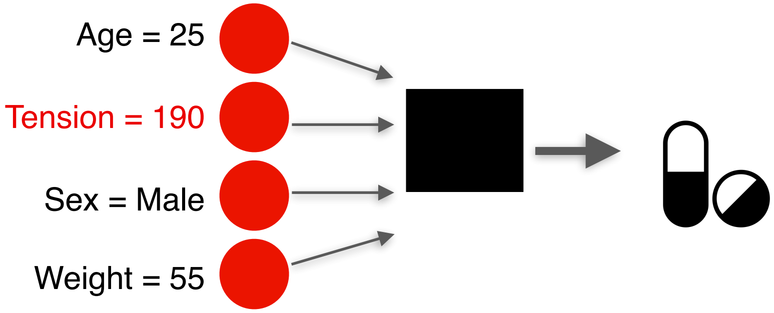
Various Types of Explanation Techniques



P( ) = 0.5 * Age +
0.01 * Tension
- 1 * Male

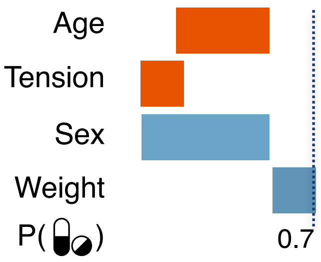
(Feature Attribution)



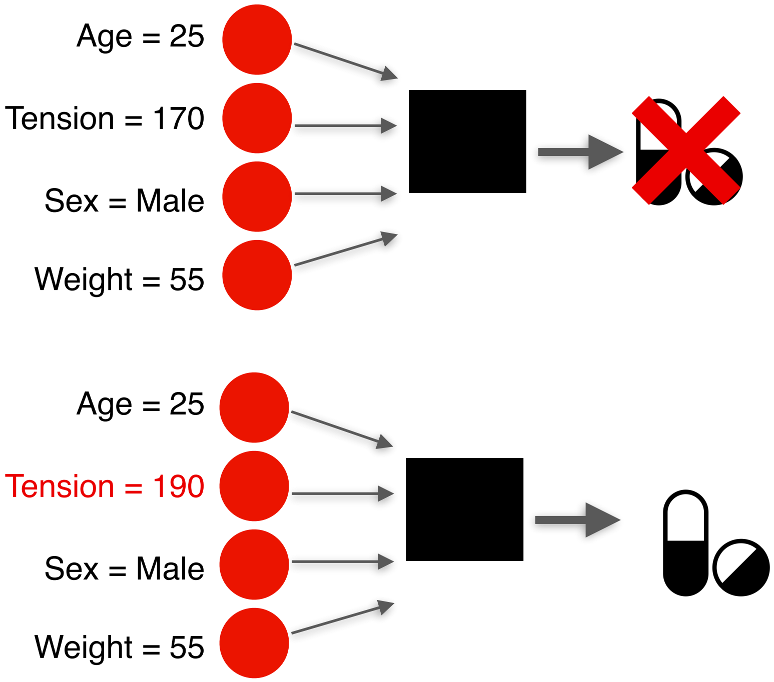


(Example-based)

Various Types of Explanation Techniques



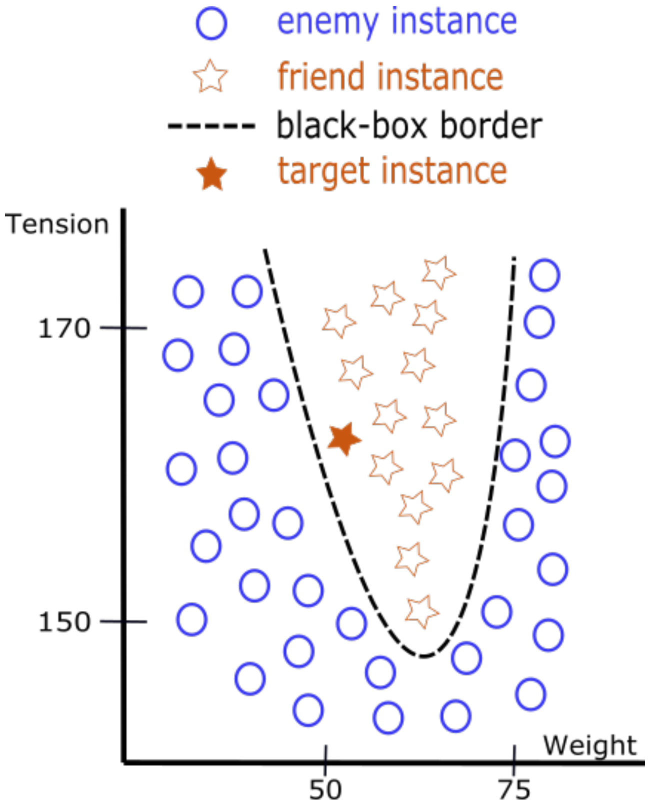
P( ) = 0.5 * Age +
0.01 * Tension
- 1 * Male



If the user has a tension between 150 and 170, while being under 28, then the level of insulin is moderate

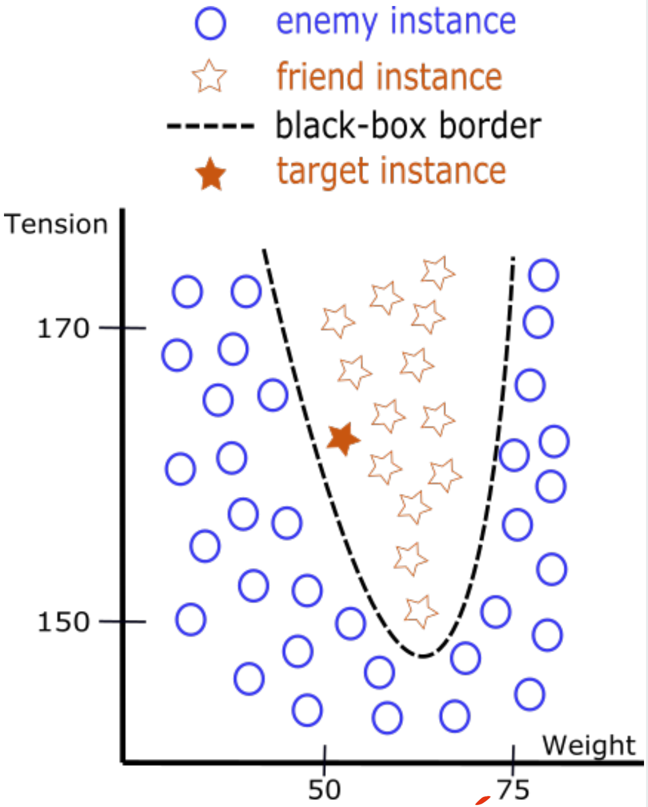
(Feature Attribution) (Example-based) (Rule-based)

Feature Attribution Explanation Techniques



Feature Attribution Explanation Techniques

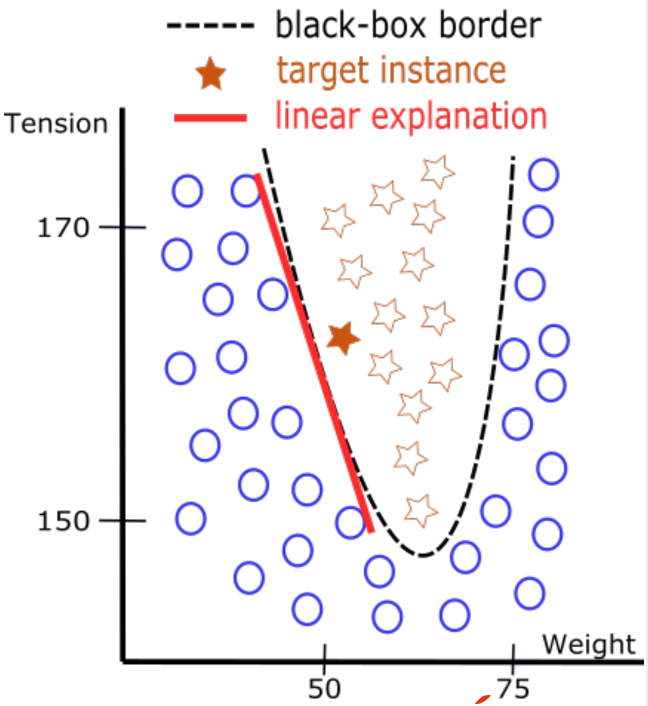
- Methods most widely used (LIME [1], SHAP [2])



(1) Tulio Ribeiro *et al*, ``Why Should I Trust You?”: Explaining the Predictions of Any Classifier, KDD, 2016
(2) Scott Lundberg *et al.*, A Unified Approach to Interpreting Model Predictions, NeurIPS 2017

Feature Attribution Explanation Techniques

- Methods most widely used (LIME [1], SHAP [2])
- LIME and its extensions approximate locally a black box model with a linear function

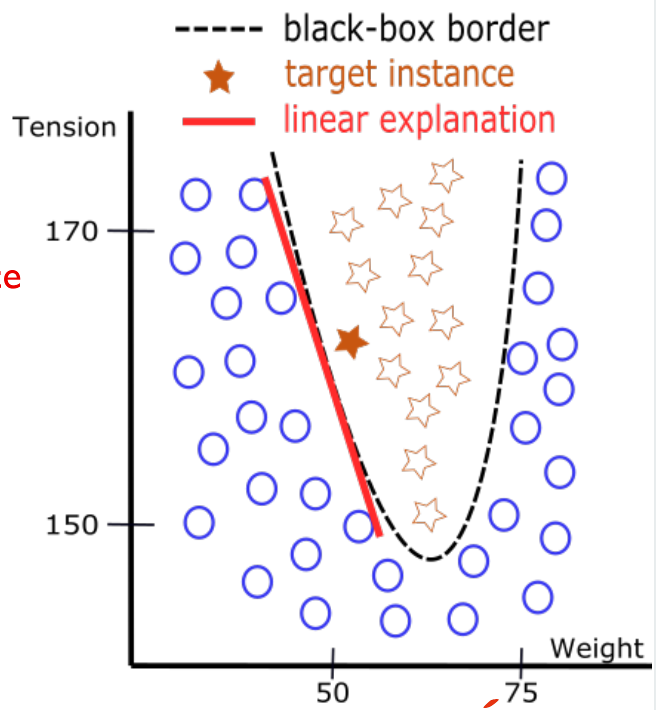


(1) Tulio Ribeiro *et al*, ``Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD, 2016
(2) Scott Lundberg *et al.*, A Unified Approach to Interpreting Model Predictions, NeurIPS 2017

Feature Attribution Explanation Techniques

- Methods most widely used (LIME [1], SHAP [2])
- LIME and its extensions approximate locally a black box model with a linear function
- The coefficients of the linear model represents their importance

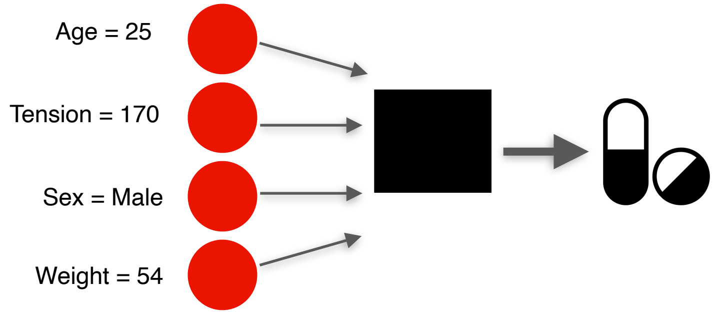
P( ) = 0.5 * Age + 0.01 * Tension - 1 * Male

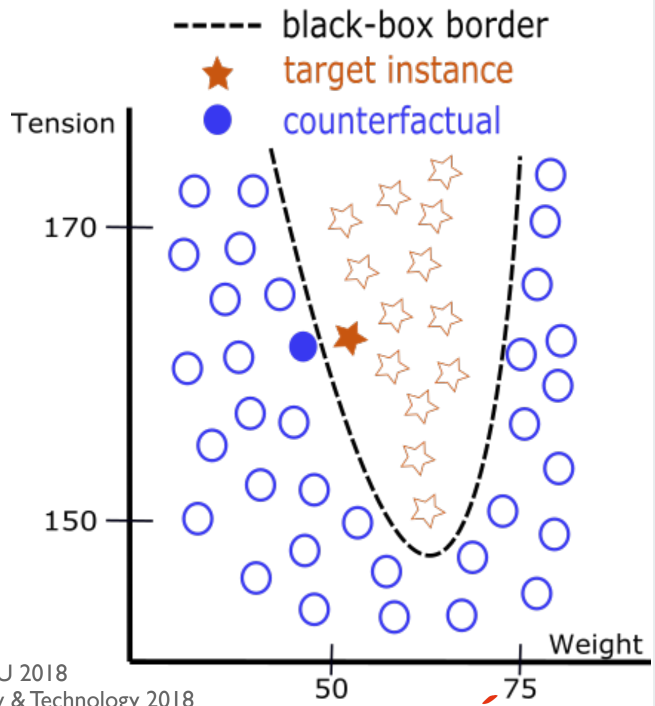


(1) Tulio Ribeiro *et al*, ``Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD, 2016
(2) Scott Lundberg *et al.*, A Unified Approach to Interpreting Model Predictions, NeurIPS 2017

Example-based Explanation Techniques

- Search for the closest instance classified differently
  - Growing Spheres [3], Wachter [4]



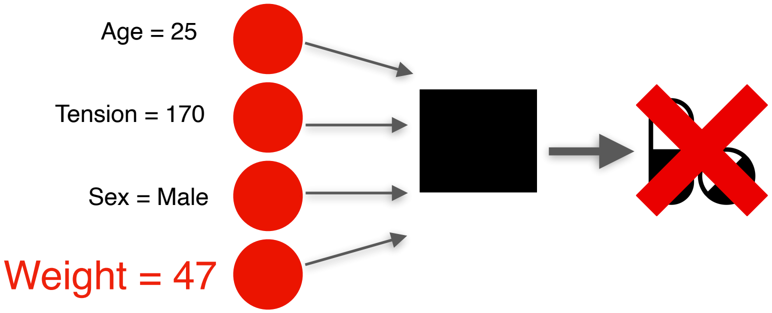


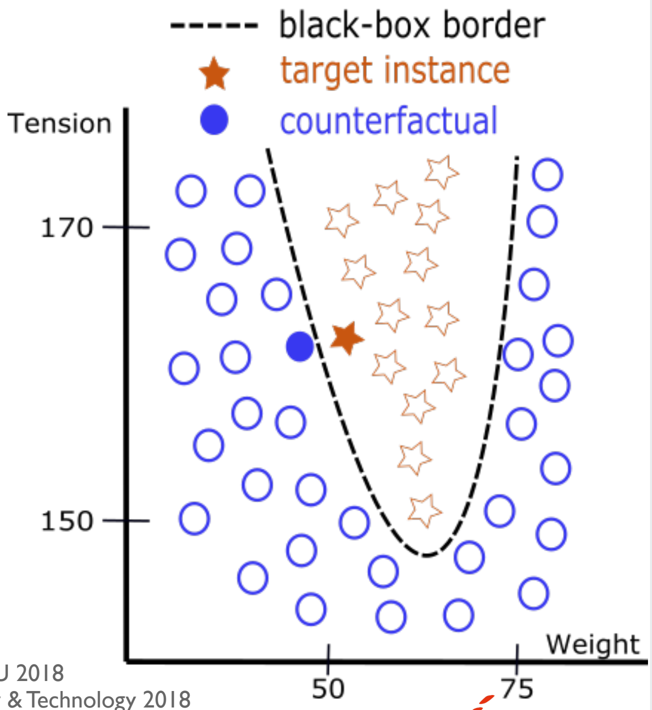
(3) Thibault Laugel *et al.*, Inverse Classification for Comparison-based Interpretability in Machine Learning. IPMU 2018
(4) Sandra Wachter *et al.*, Counterfactual Explanations Without Opening the Black Box. Harvard Journal of Law & Technology 2018

Example-based Explanation Techniques

- Search for the closest instance classified differently
  - Growing Spheres [3], Wachter [4]

- Shows the minimum changes required to modify the prediction

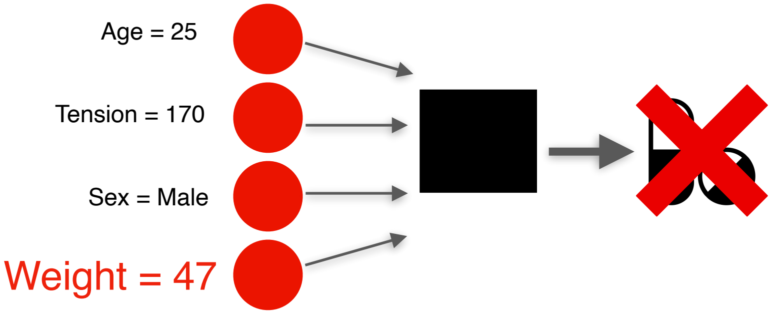


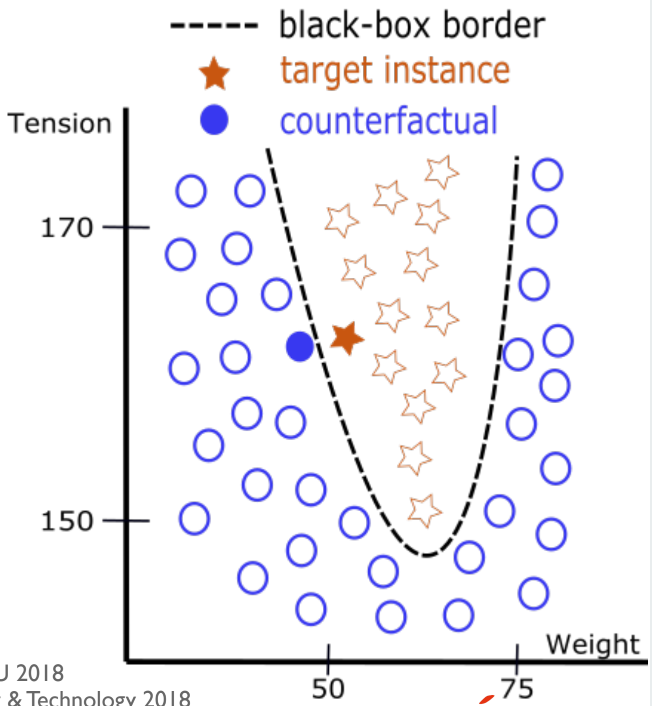


(3) Thibault Laugel *et al.*, Inverse Classification for Comparison-based Interpretability in Machine Learning. IPMU 2018
(4) Sandra Wachter *et al.*, Counterfactual Explanations Without Opening the Black Box. Harvard Journal of Law & Technology 2018

Example-based Explanation Techniques

- Search for the closest instance classified differently
  - Growing Spheres [3], Wachter [4]
- Shows the minimum changes required to modify the prediction
- Close to how human reason and explain

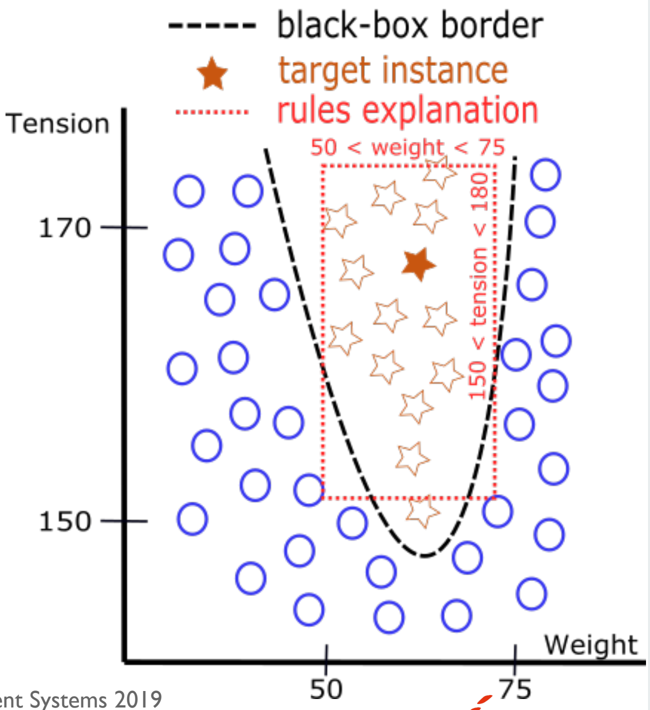




(3) Thibault Laugel *et al.*, Inverse Classification for Comparison-based Interpretability in Machine Learning. IPMU 2018
(4) Sandra Wachter *et al.*, Counterfactual Explanations Without Opening the Black Box. Harvard Journal of Law & Technology 2018

Rule-based Explanation Techniques

- Local approximation of a black box model with decision rules
  - Anchors [5], LORE [6]

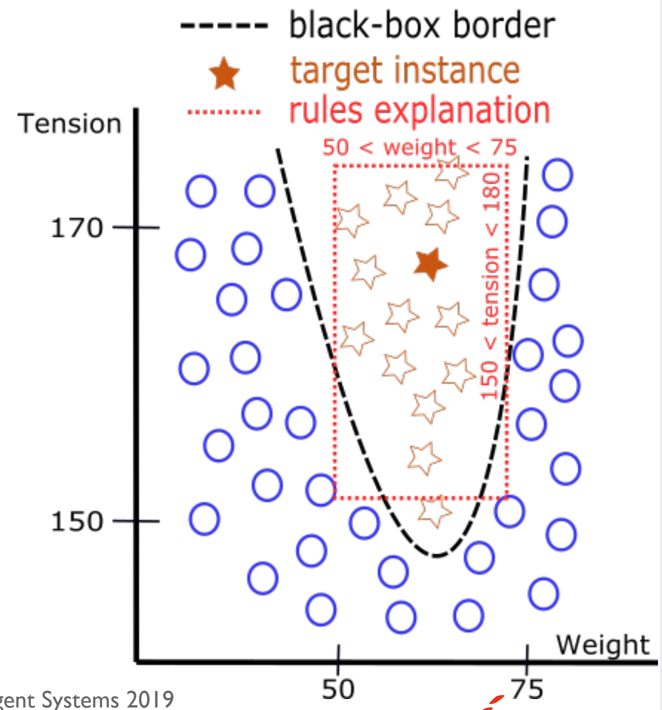


(5) Tulio Ribeiro *et al.*, Anchors: High-Precision Model-Agnostic Explanations. AAAI 2018
(6) Riccardo Guidotti *et al.*, Factual and counterfactual explanations for black box decision making. IEEE Intelligent Systems 2019

Rule-based Explanation Techniques

- Local approximation of a black box model with decision rules
  - Anchors [5], LORE [6]
- Computes the necessary conditions for a particular outcome

If the user has a tension between 150 and 180, while weighing between 50 and 75 kilos, then the level of insulin is moderate

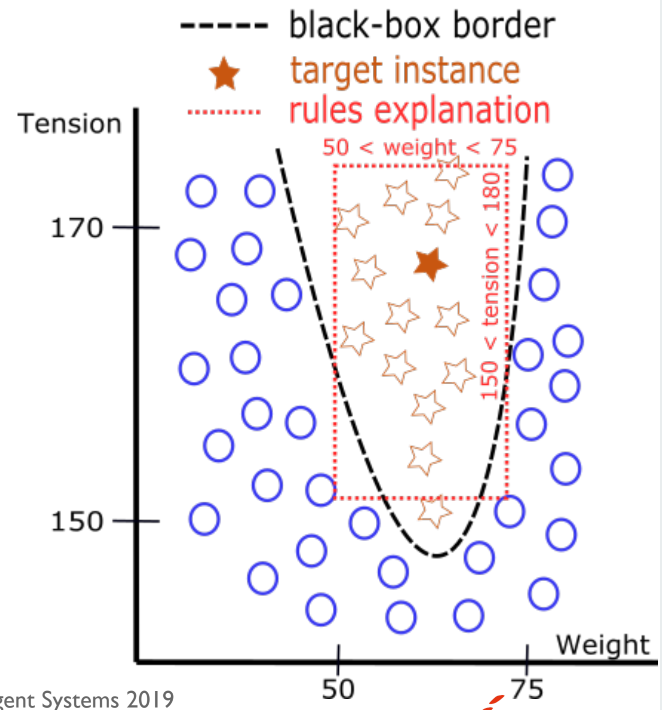


(5) Tulio Ribeiro *et al.*, Anchors: High-Precision Model-Agnostic Explanations. AAAI 2018
(6) Riccardo Guidotti *et al.*, Factual and counterfactual explanations for black box decision making. IEEE Intelligent Systems 2019

Rule-based Explanation Techniques

- Local approximation of a black box model with decision rules
  - Anchors [5], LORE [6]
- Computes the necessary conditions for a particular outcome
- Employed for a long time as proxy for domain expert

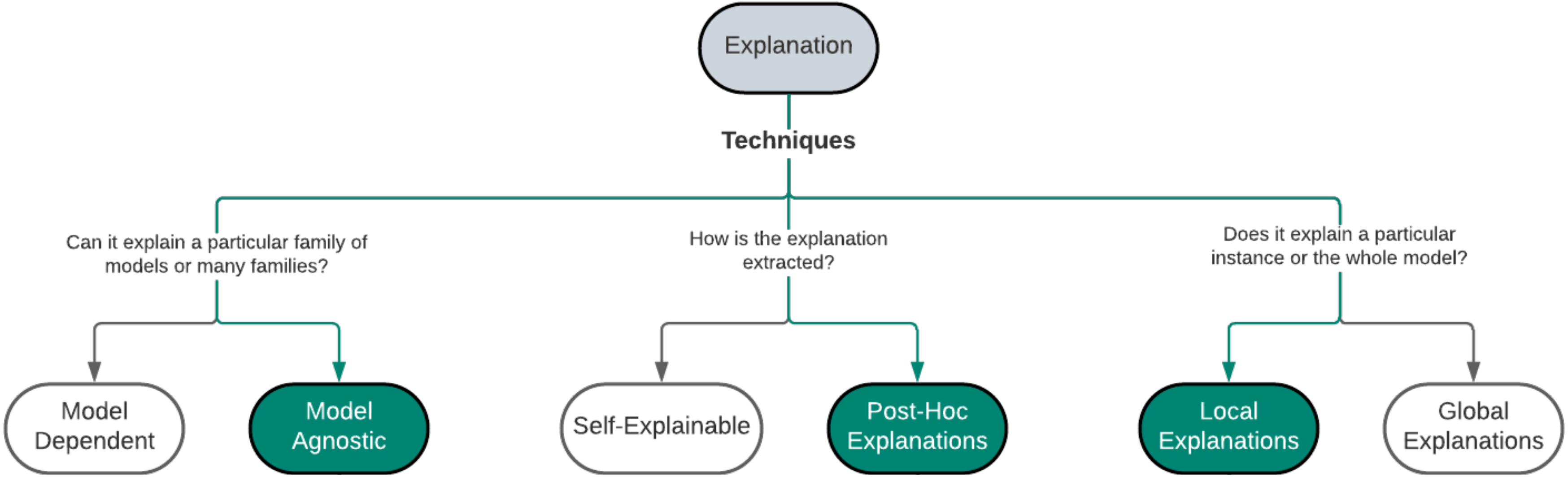
If the user has a tension between 150 and 180, while weighing between 50 and 75 kilos, then the level of insulin is moderate



(5) Tulio Ribeiro *et al.*, Anchors: High-Precision Model-Agnostic Explanations. AAAI 2018
(6) Riccardo Guidotti *et al.*, Factual and counterfactual explanations for black box decision making. IEEE Intelligent Systems 2019

Taxonomy of Methods Generating Explanations

- Various types of explanation techniques:
  - Model dependent / Model Agnostic
  - Self-explainable / Post-Hoc Explanations
  - Local / Global Explanations

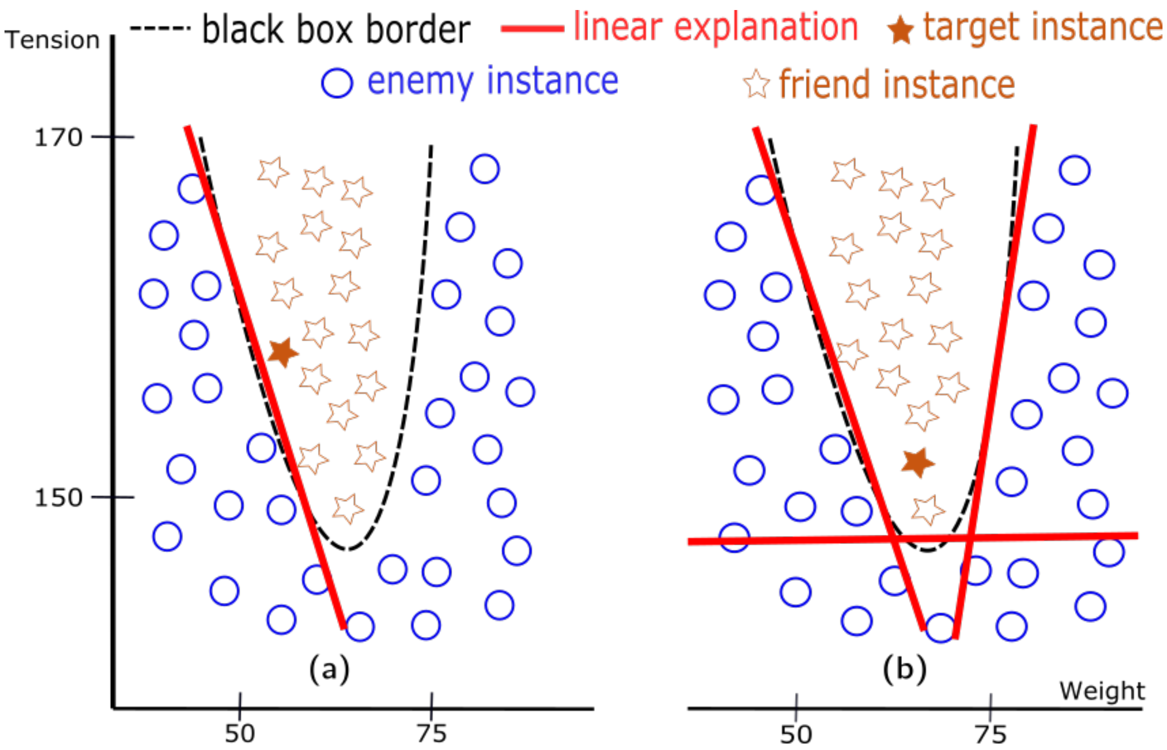


Research Questions — Part I

- How to generate the best explanation from a data perspective?

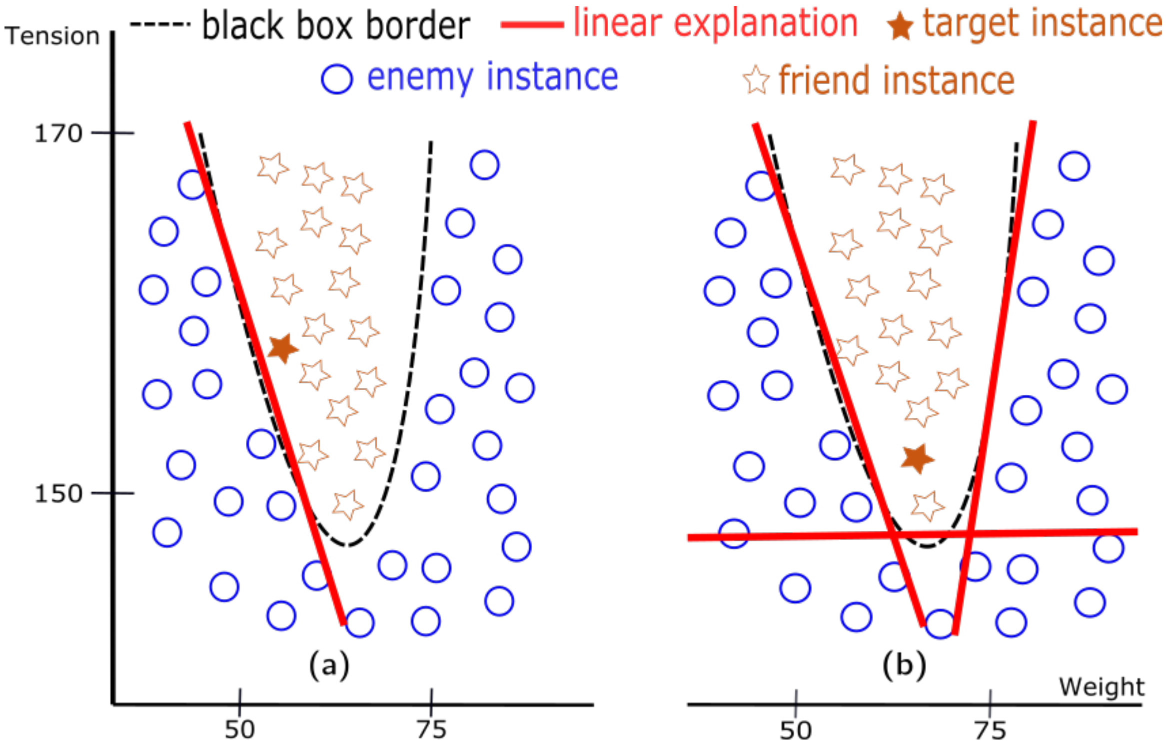
Research Questions — Part I

- How to generate the best explanation from a data perspective?
- Linear explanations are widely employed



Research Questions — Part I

- How to generate the best explanation from a data perspective?
- Linear explanations are widely employed
- But are they adapted to every local situation?
  - When Should We Use Linear Explanations? [7]



(7) **Julien Delaunay**, *et al.*, When Should We Use Linear Explanations?, CIKM, 2022

Research Questions — Part II

- How to generate the best explanation from a user perspective?

Research Questions — Part II

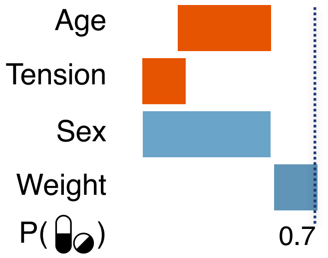
- How to generate the best explanation from a user perspective?

- Few user studies has been conducted to measure [8][9] impact of explanation:

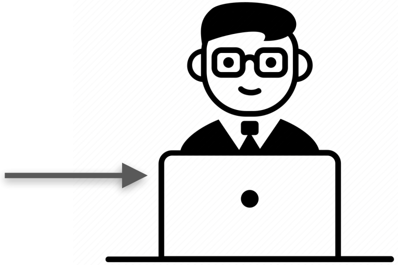
(8) Doshi-Velez and Kim., Towards A Rigorous Science of Interpretable Machine Learning. Machine Learning 2018
(9) Adadi *et al.*, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 2018

Research Questions — Part II

- How to generate the best explanation from a user perspective?
- Few user studies has been conducted to measure [8][9] impact of explanation:



If the user has a tension between 150 and 170, while being under 28, then the level of insulin is moderate

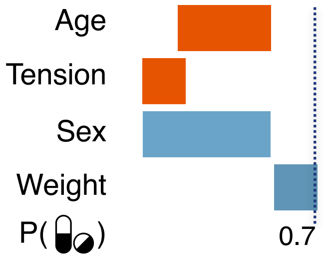


(8) Doshi-Velez and Kim., Towards A Rigorous Science of Interpretable Machine Learning. Machine Learning 2018
(9) Adadi *et al.*, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 2018

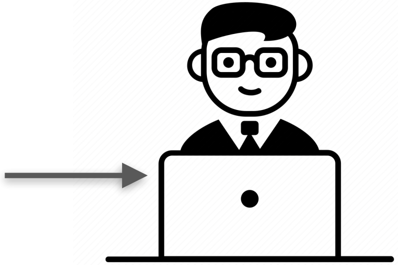
Research Questions — Part II

- How to generate the best explanation from a <u>user</u> perspective?

- Few user studies has been conducted to measure [8][9] impact of explanation:
    - Impact of Explanation Techniques and Representations on Users' Trust and Understanding [10]



If the user has a tension between 150 and 170, while being under 28, then the level of insulin is moderate



(8) Doshi-Velez and Kim., Towards A Rigorous Science of Interpretable Machine Learning. Machine Learning 2018
(9) Adadi *et al.*, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 2018
(10) **Julien Delaunay**, *et al.*, Impact of Explanation Techniques and Representations on Users' Trust and Understanding. Under Review CSCW 2024

Part I: How to generate the best explanation from a <u>data</u> perspective?

When Should We Use Linear Explanations? [CIKM '22]

Julien Delaunay

When Should We Use Linear Explanations? — Contributions

When Should We Use Linear Explanations? — Contributions

- A novel technique to detect the closest decision boundary

When Should We Use Linear Explanations? — Contributions

- A novel technique to detect the closest decision boundary

- An oracle to answer the question: "When are linear explanations adapted?"

When Should We Use Linear Explanations? — Contributions

- A novel technique to detect the closest decision boundary

- An oracle to answer the question: "When are linear explanations adapted?"

- Two methods that generate:
  - Linear explanations if adapted
  - Rule-based explanations otherwise

Input Assumptions

Input Assumptions

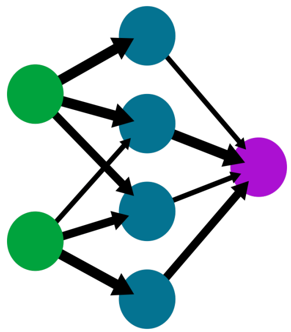
Age	Tension	Gender	Weight
28	150	Female	58
22	160	Male	65
54	155	Female	52
72	170	Male	75
18	170	Male	65

A dataset

Input Assumptions

Age	Tension	Gender	Weight
28	150	Female	58
22	160	Male	65
54	155	Female	52
72	170	Male	75
18	170	Male	65

A dataset

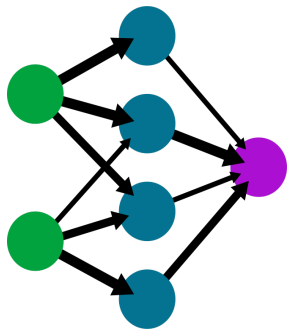


A black box

Input Assumptions

Age	Tension	Gender	Weight
28	150	Female	58
22	160	Male	65
54	155	Female	52
72	170	Male	75
18	170	Male	65

A dataset

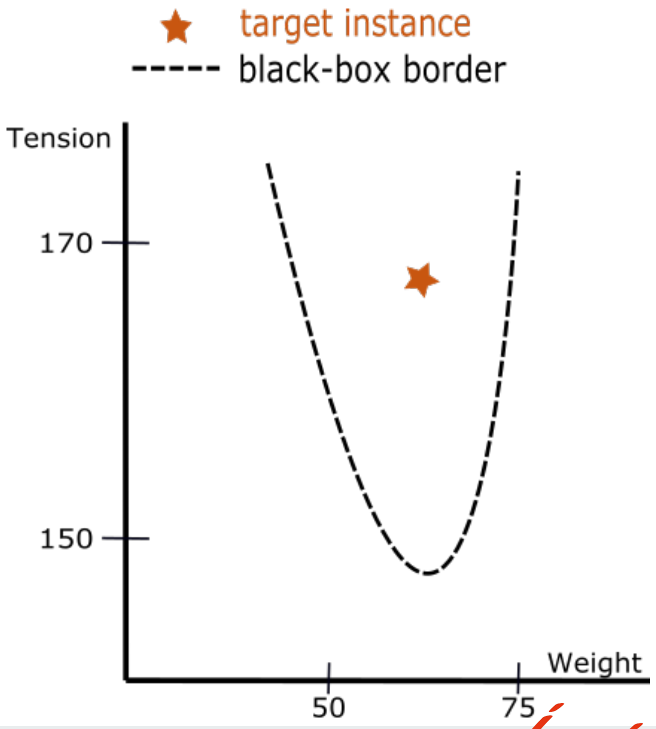


A black box

F	45	165	Female	55

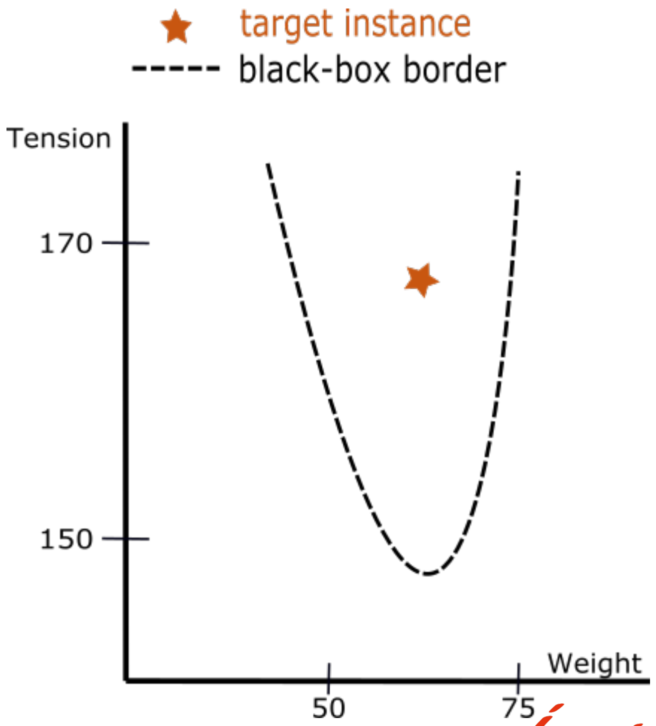
Target Instance

Where is the closest decision boundary?



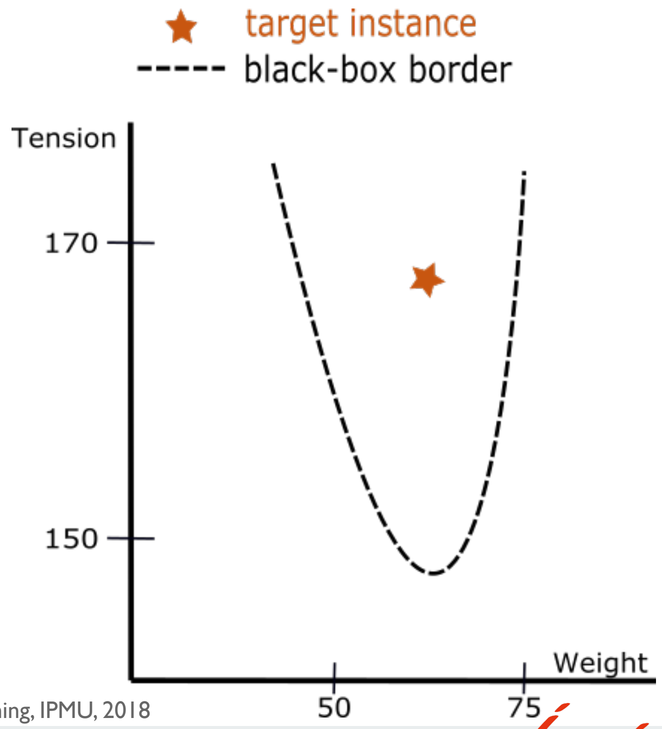
Where is the closest decision boundary?

- The closest counterfactual indicates the decision boundary



Where is the closest decision boundary?

- The closest counterfactual indicates the decision boundary
- Growing Spheres[3]:

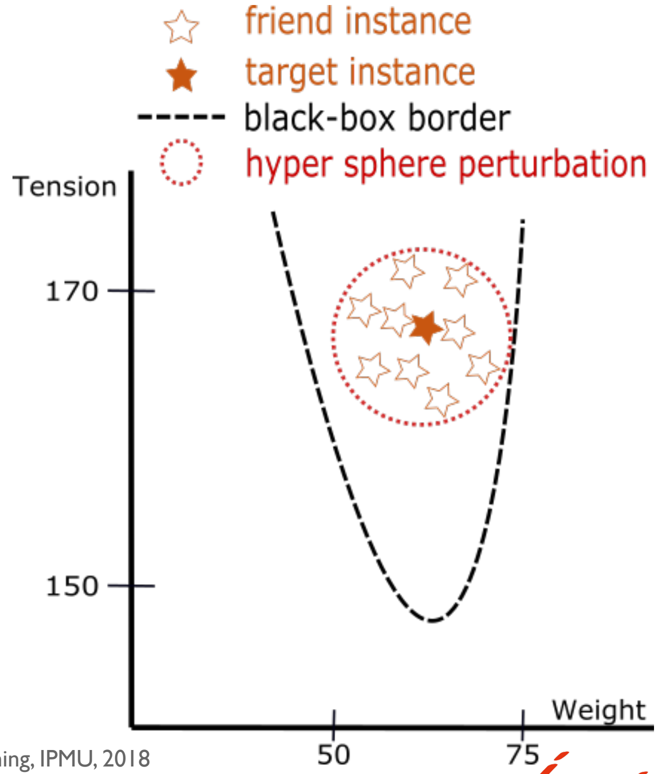


(3) Thibault Laugel *et al*, Inverse Classification for Comparison-based Interpretability in Machine Learning, IPMU, 2018

Where is the closest decision boundary?

- The closest counterfactual indicates the decision boundary

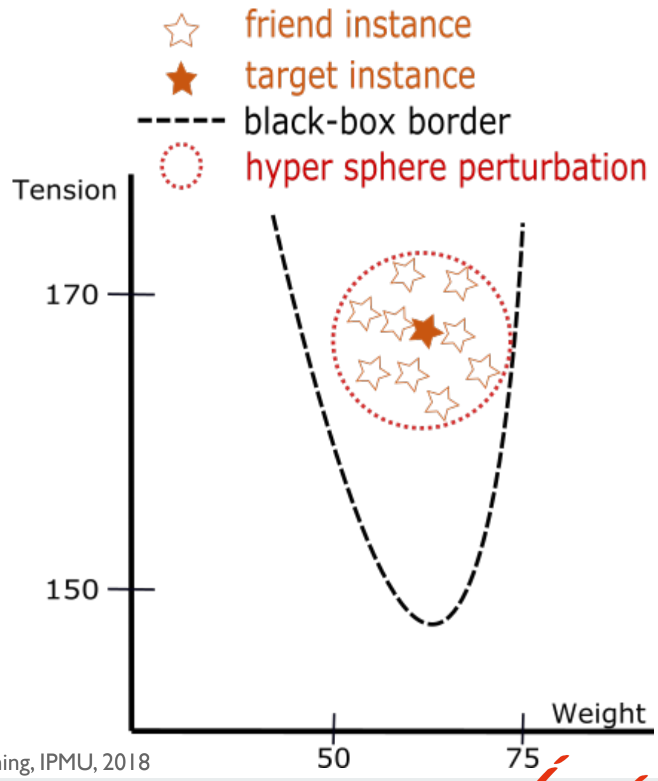
- Growing Spheres[3]:
    - Generates instances inside an hypersphere



(3) Thibault Laugel *et al*, Inverse Classification for Comparison-based Interpretability in Machine Learning, IPMU, 2018

Where is the closest decision boundary?

- The closest counterfactual indicates the decision boundary
- Growing Spheres[3]:
  - Generates instances inside an hypersphere
  - While there is no instance from the other class:

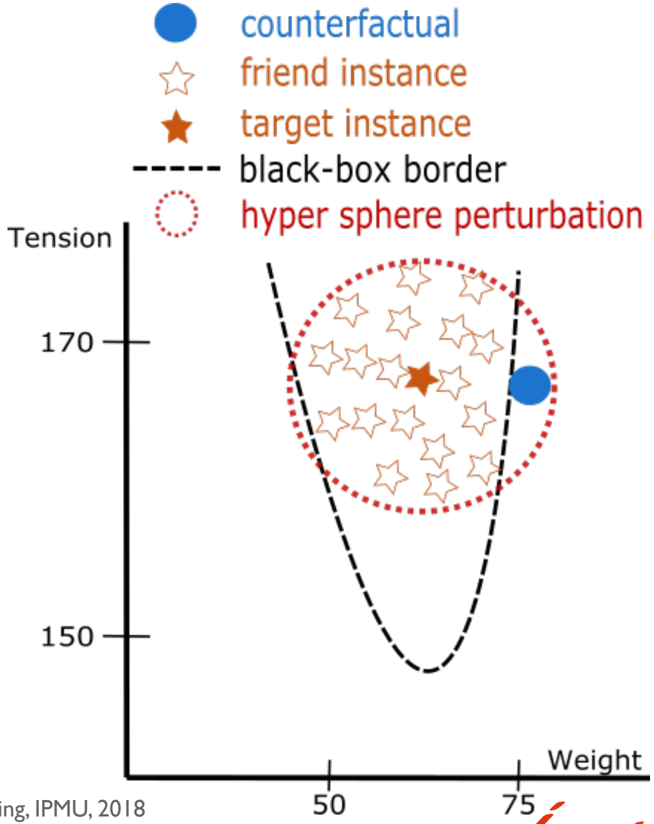


(3) Thibault Laugel *et al*, Inverse Classification for Comparison-based Interpretability in Machine Learning, IPMU, 2018

Where is the closest decision boundary?

- The closest counterfactual indicates the decision boundary

- Growing Spheres[3]:
  - Generates instances inside an hypersphere
  - While there is no instance from the other class:
    I. Increases the perturbation
    II. Until the first counterfactual is met



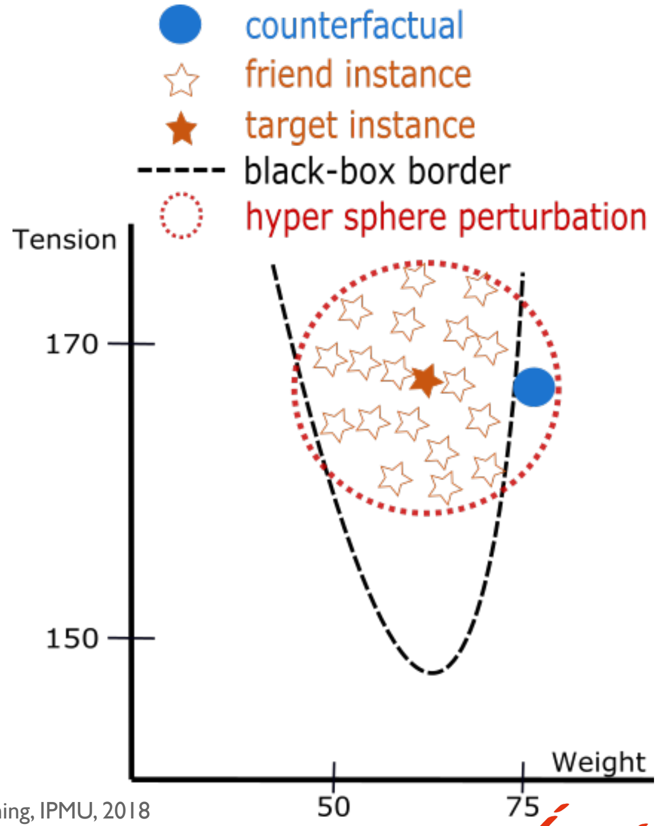
(3) Thibault Laugel *et al*, Inverse Classification for Comparison-based Interpretability in Machine Learning, IPMU, 2018

Where is the closest decision boundary?

- The closest counterfactual indicates the decision boundary

- Growing Spheres[3]:
  - Generates instances inside an hypersphere
  - While there is no instance from the other class:
    - I. Increases the perturbation
    - II. Until the first counterfactual is met

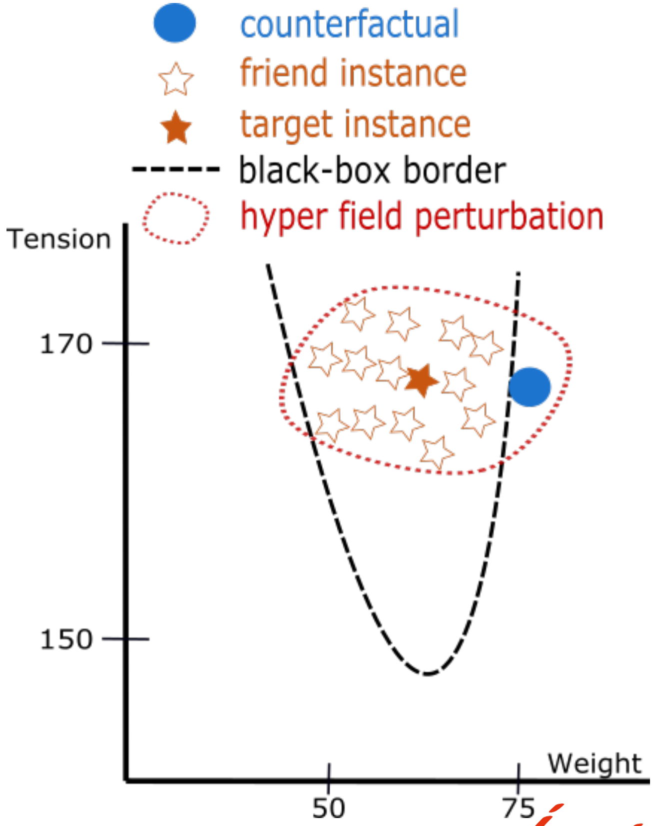
- Drawback of Growing Spheres:
  - Perturbs in all direction at the same rate
  - Does not deal with categorical features



(3) Thibault Laugel *et al*, Inverse Classification for Comparison-based Interpretability in Machine Learning, IPMU, 2018

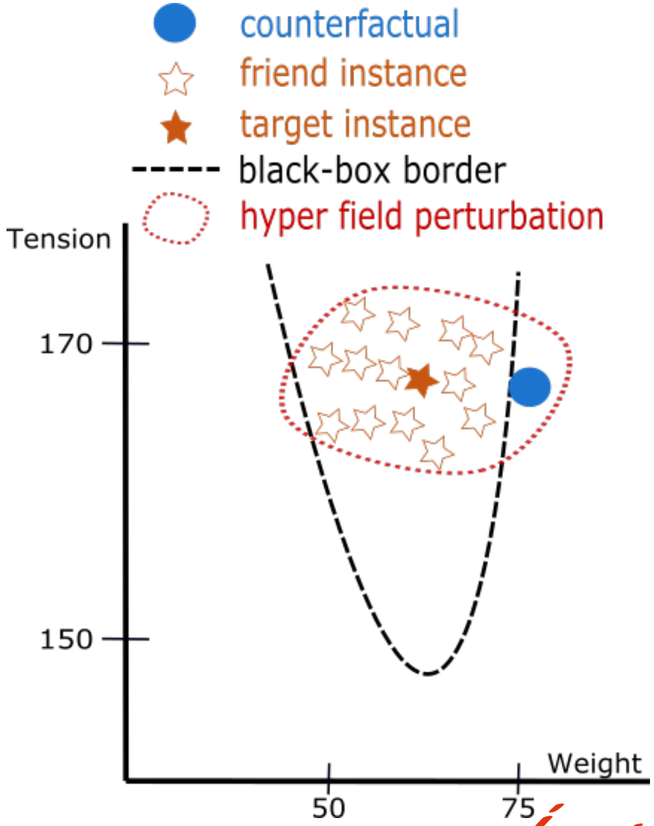
Growing Fields — 1st Contribution

- Generates instances inside an hyper field



Growing Fields — 1st Contribution

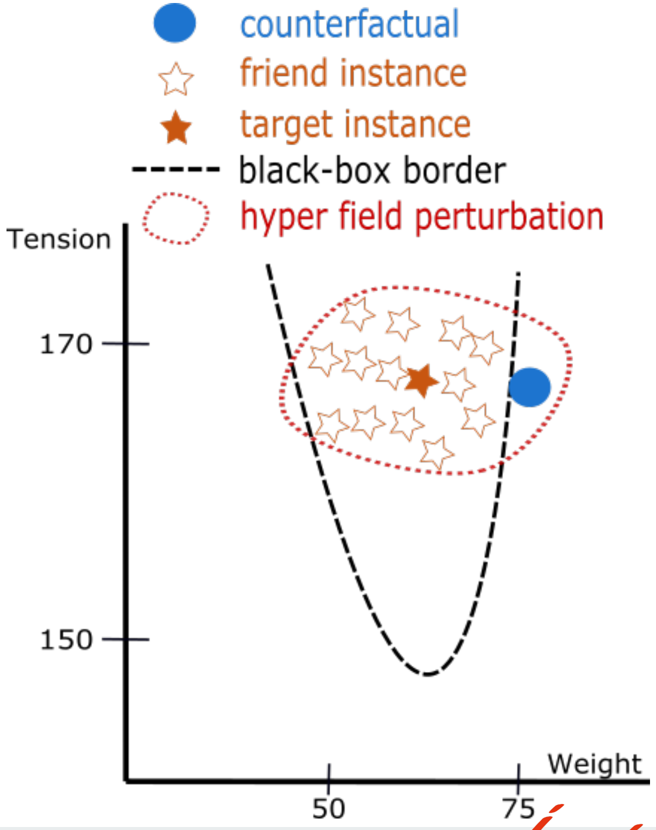
- Generates instances inside an hyper field
  - Employs the mean and standard deviation of each features to:
    - Control the rate of perturbation
    - Perturb more accurately



Growing Fields — 1st Contribution

- Generates instances inside an hyper field
  - Employs the mean and standard deviation of each features to:
    - Control the rate of perturbation
    - Perturb more accurately

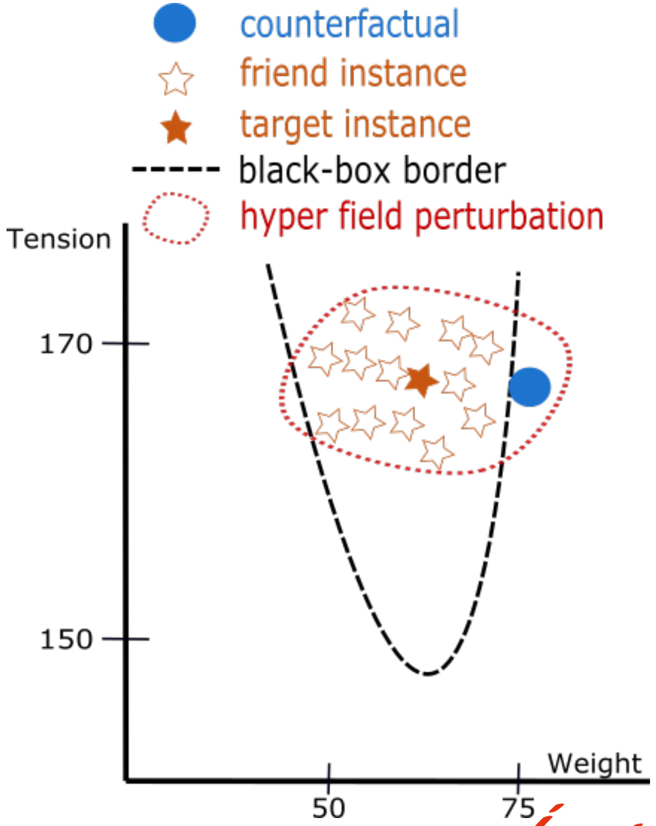
- Employs the normalized standardized Euclidean distance:



Growing Fields — 1st Contribution

- Generates instances inside an hyper field
  - Employs the mean and standard deviation of each features to:
    - Control the rate of perturbation
    - Perturb more accurately

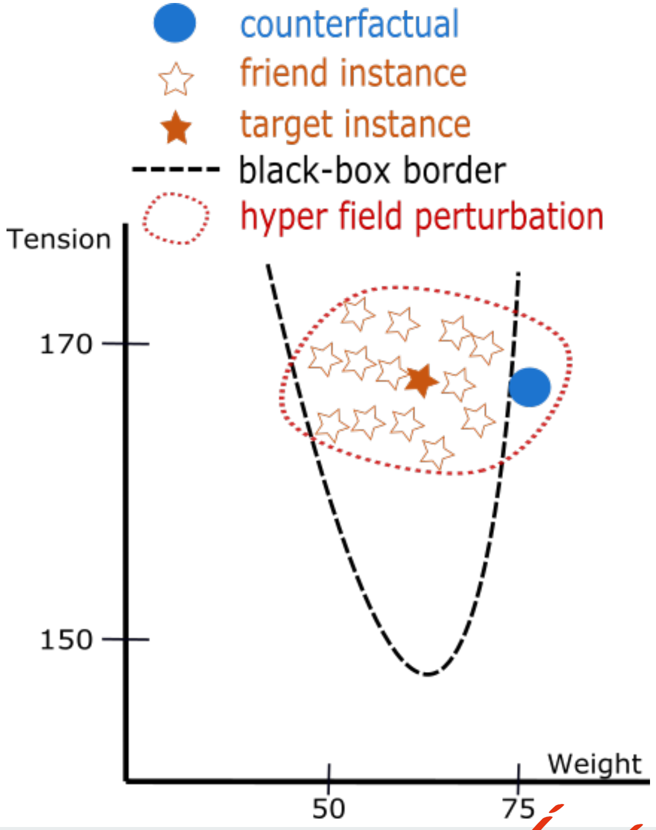
- Employs the normalized standardized Euclidean distance:
  - Perturbation rate is comprised between 0 and 1



Growing Fields — 1st Contribution

- Generates instances inside an hyper field
  - Employs the mean and standard deviation of each features to:
    - Control the rate of perturbation
    - Perturb more accurately

- Employs the normalized standardized Euclidean distance:
  - Perturbation rate is comprised between 0 and 1
  - Convert the perturbation rate into a probability of changing a categorical value



Experiments — Realism Comparison

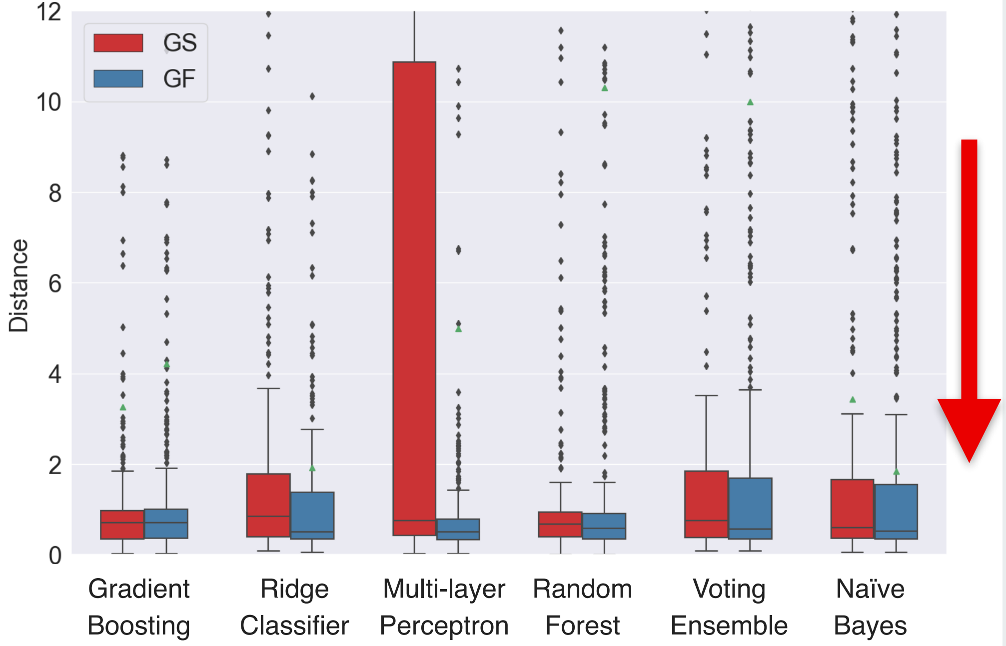
Experiments — Realism Comparison

- Realism is measured through the distance between:
  - The counterfactual generated by:
    A. Growing Spheres (GS)
    B. Growing Fields (GF)
  - The closest instance from dataset

Experiments — Realism Comparison

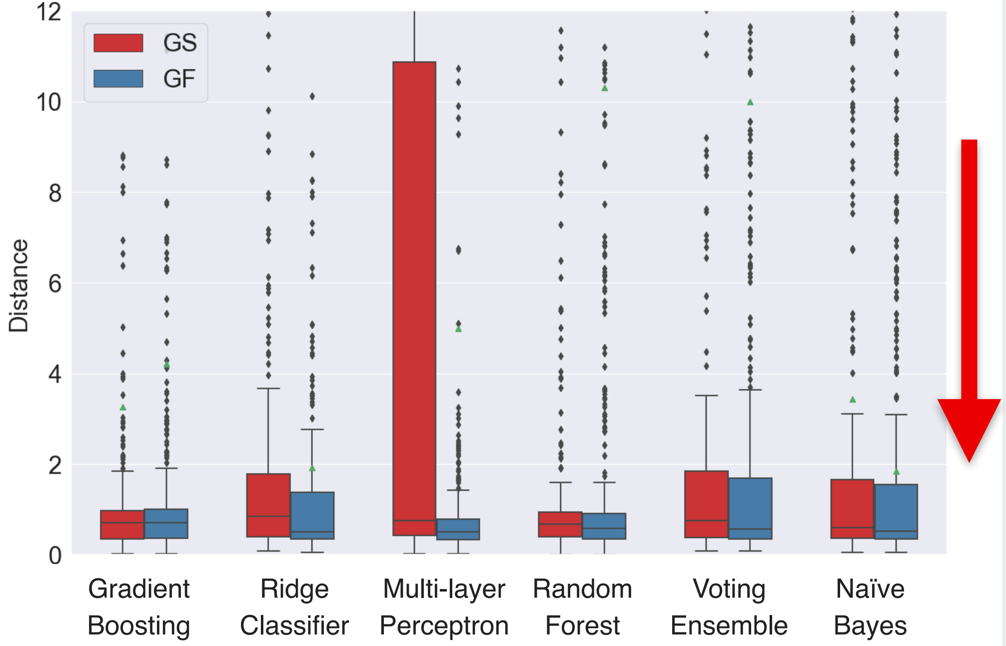
- Realism is measured through the distance between:
  - The counterfactual generated by:
    - A. Growing Spheres (GS)
    - B. Growing Fields (GF)
  - The closest instance from dataset

- Averaged over 7 continuous datasets



Experiments — Realism Comparison

- Realism is measured through the distance between:
  - The counterfactual generated by:
    A. Growing Spheres (GS)
    B. Growing Fields (GF)
  - The closest instance from dataset



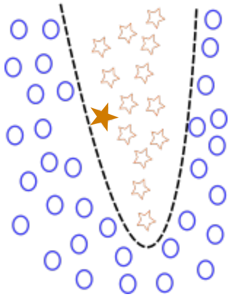
- Averaged over 7 continuous datasets

- GF generates more realistic instances than GS

When Are Linear Explanations Adapted? — Oracle

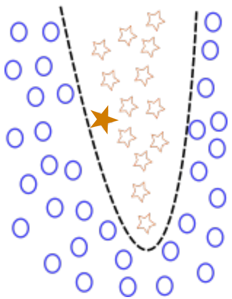
When Are Linear Explanations Adapted? — Oracle

Input Dataset
& Black-box

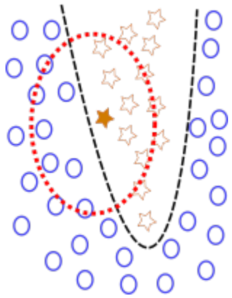


When Are Linear Explanations Adapted? — Oracle

Input Dataset
& Black-box

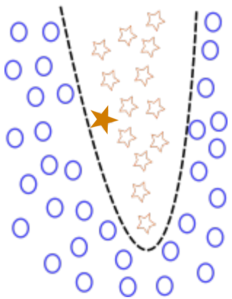


Growing Fields

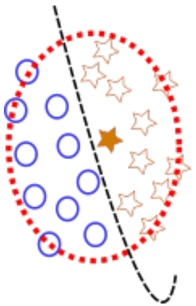


When Are Linear Explanations Adapted? — Oracle

Input Dataset
& Black-box

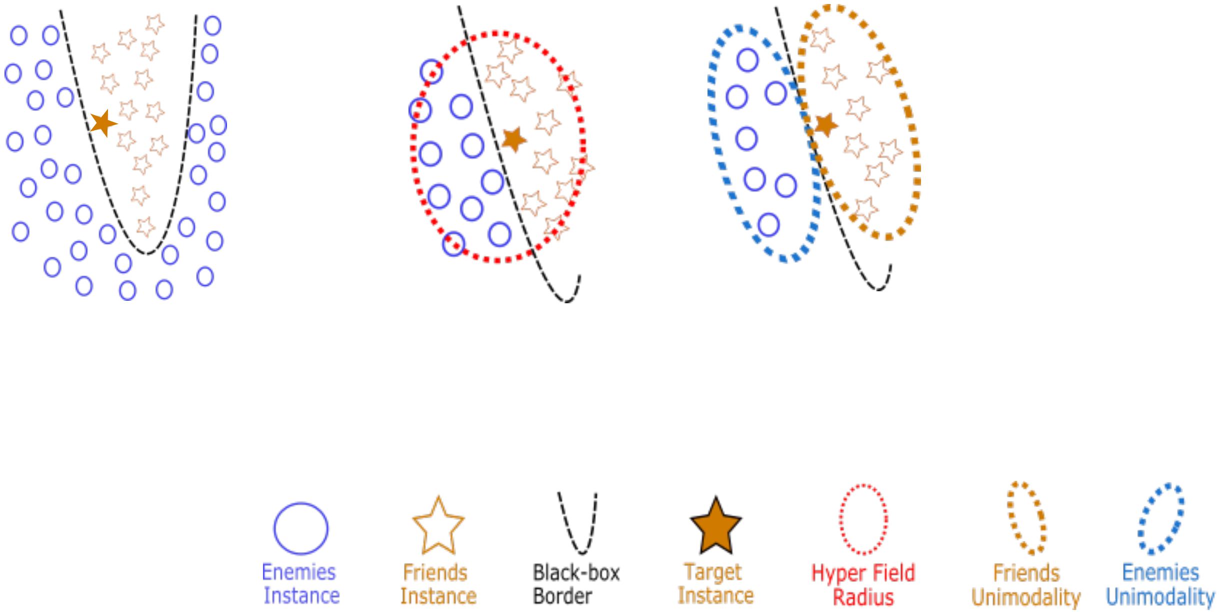


Growing Fields

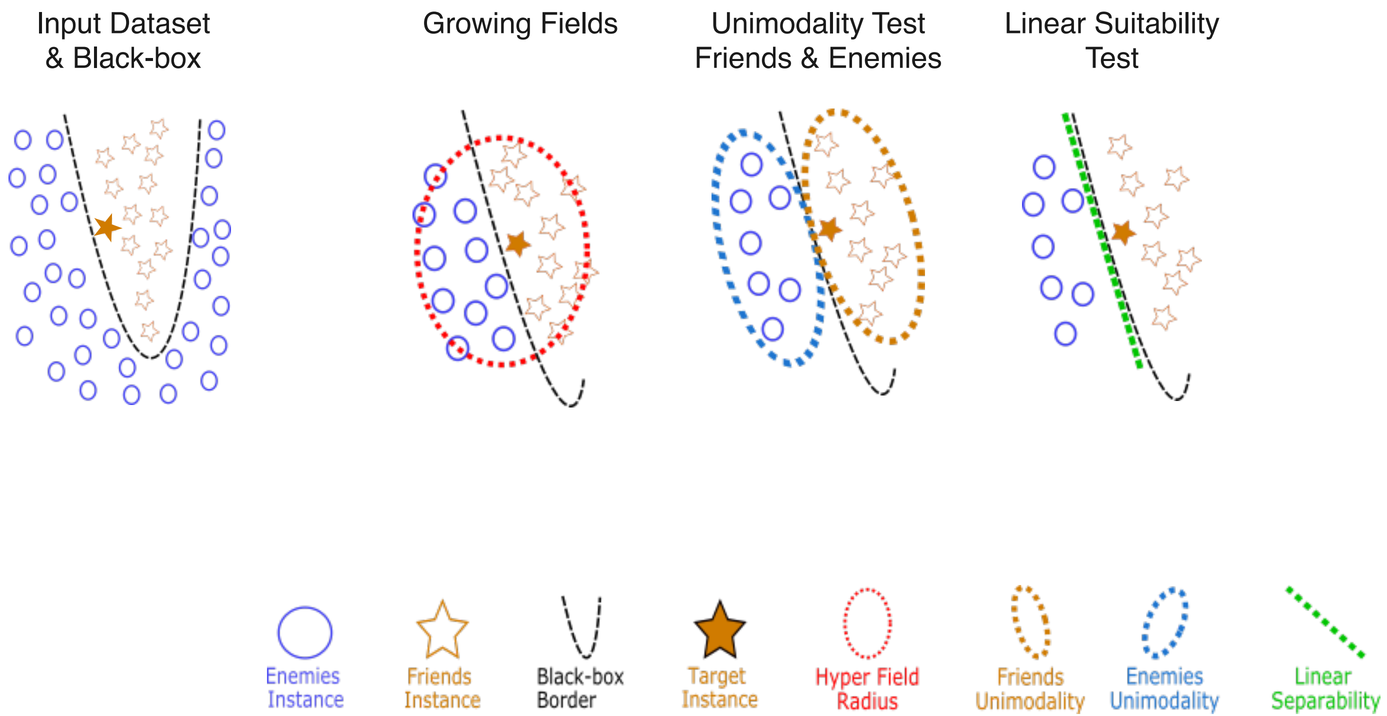


When Are Linear Explanations Adapted? — Oracle

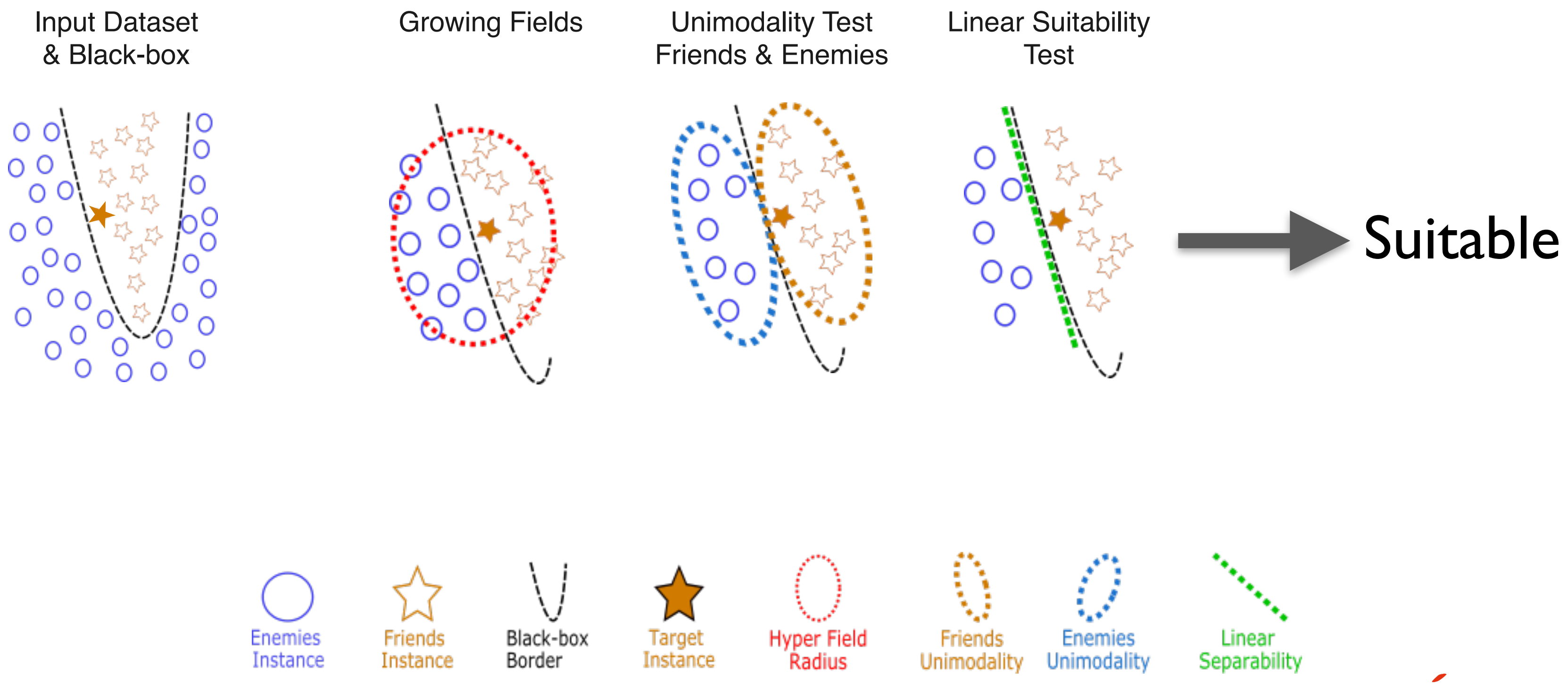
Input Dataset & Black-box | Growing Fields | Unimodality Test Friends & Enemies



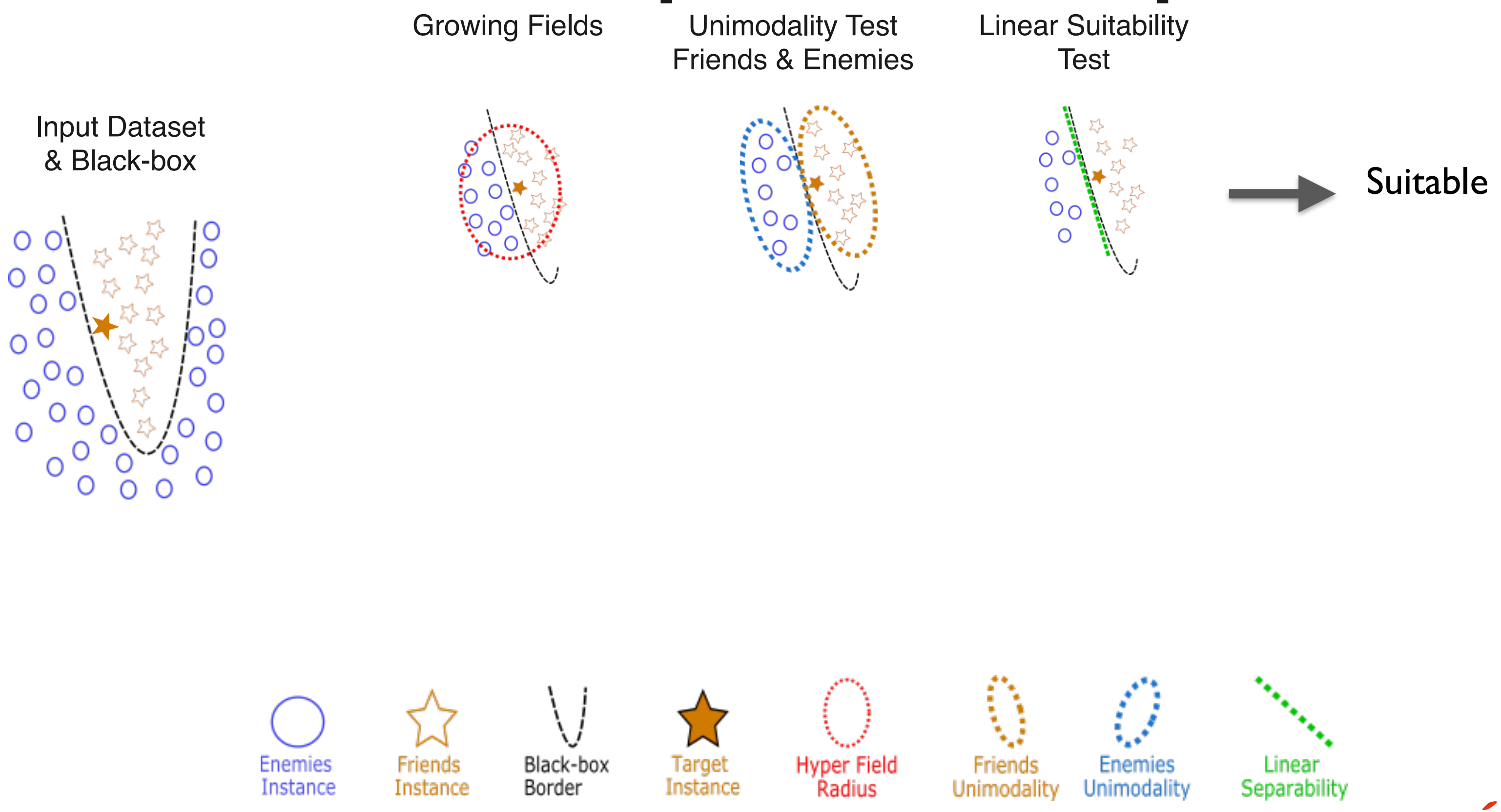
When Are Linear Explanations Adapted? — Oracle



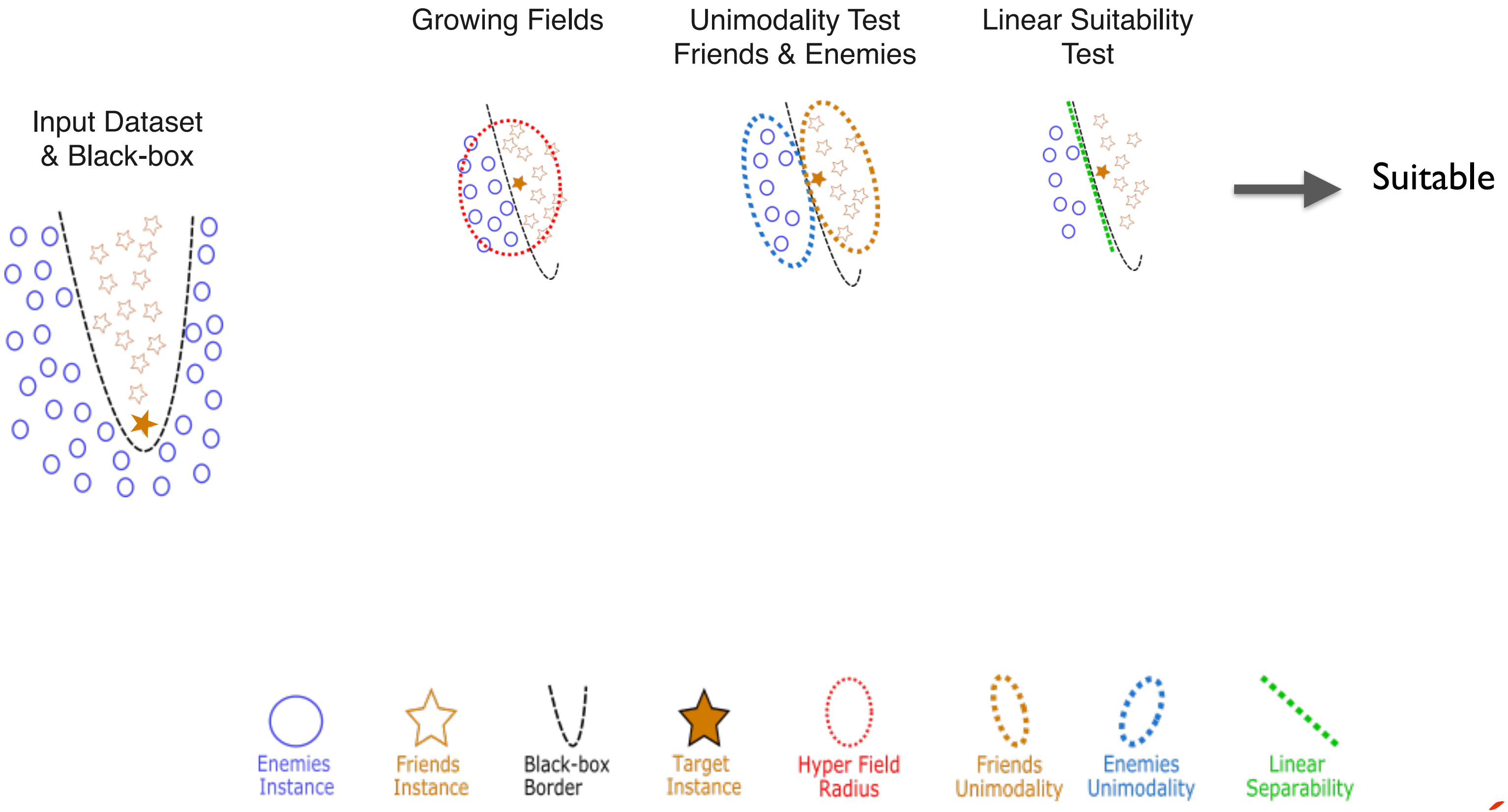
When Are Linear Explanations Adapted? — Oracle



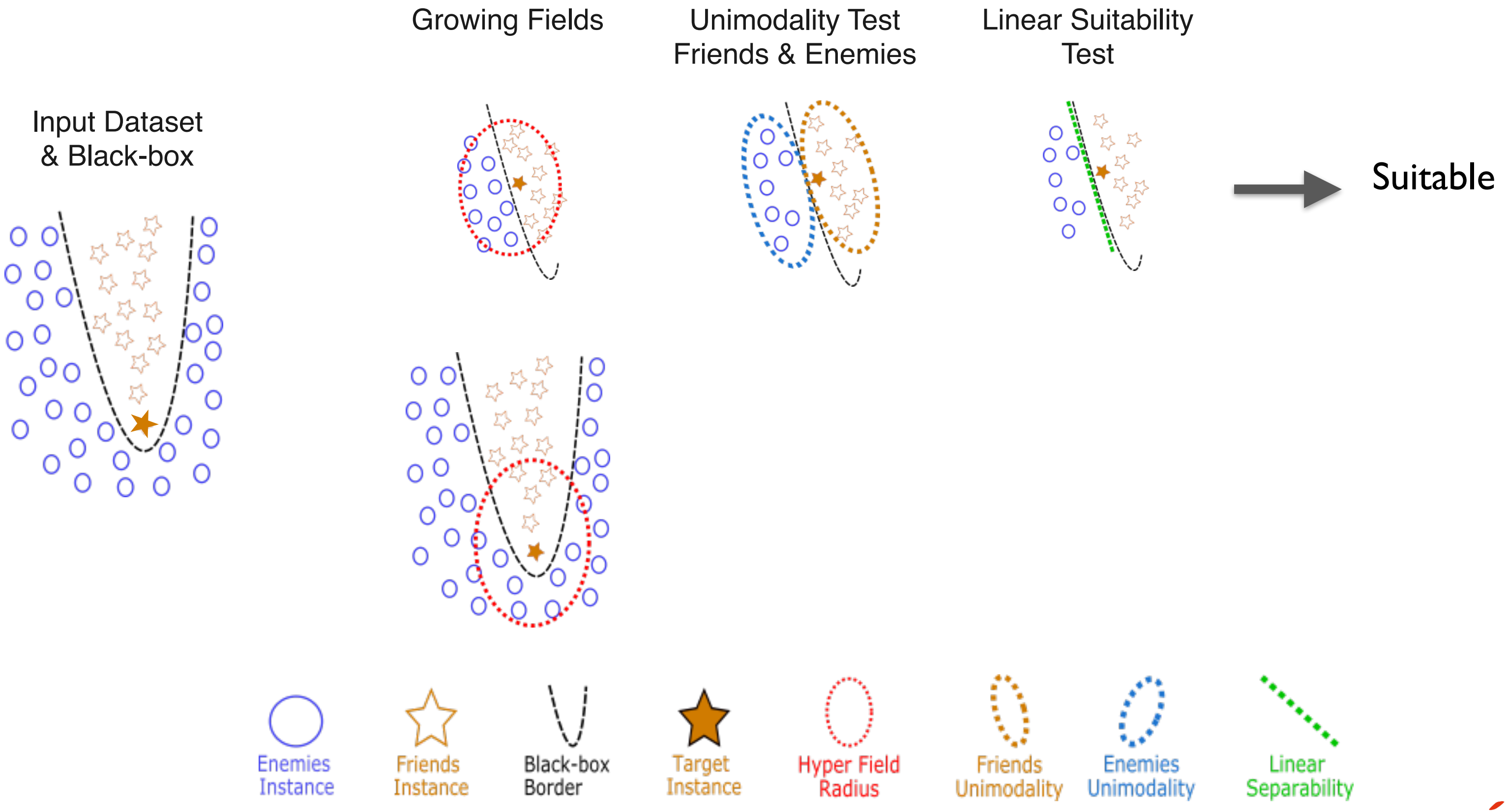
When Are Linear Explanations Adapted? — Oracle



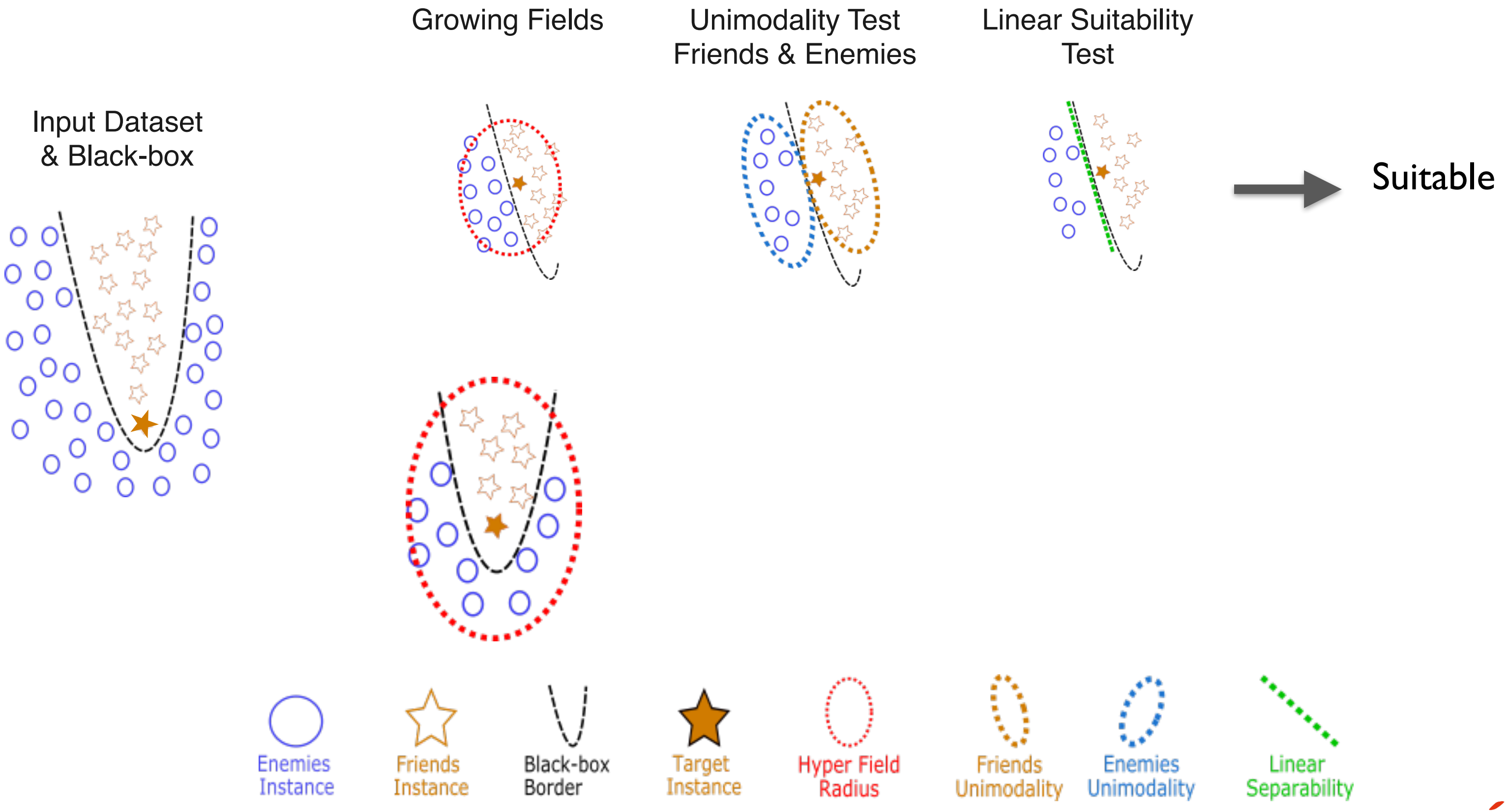
When Are Linear Explanations Adapted? — Oracle



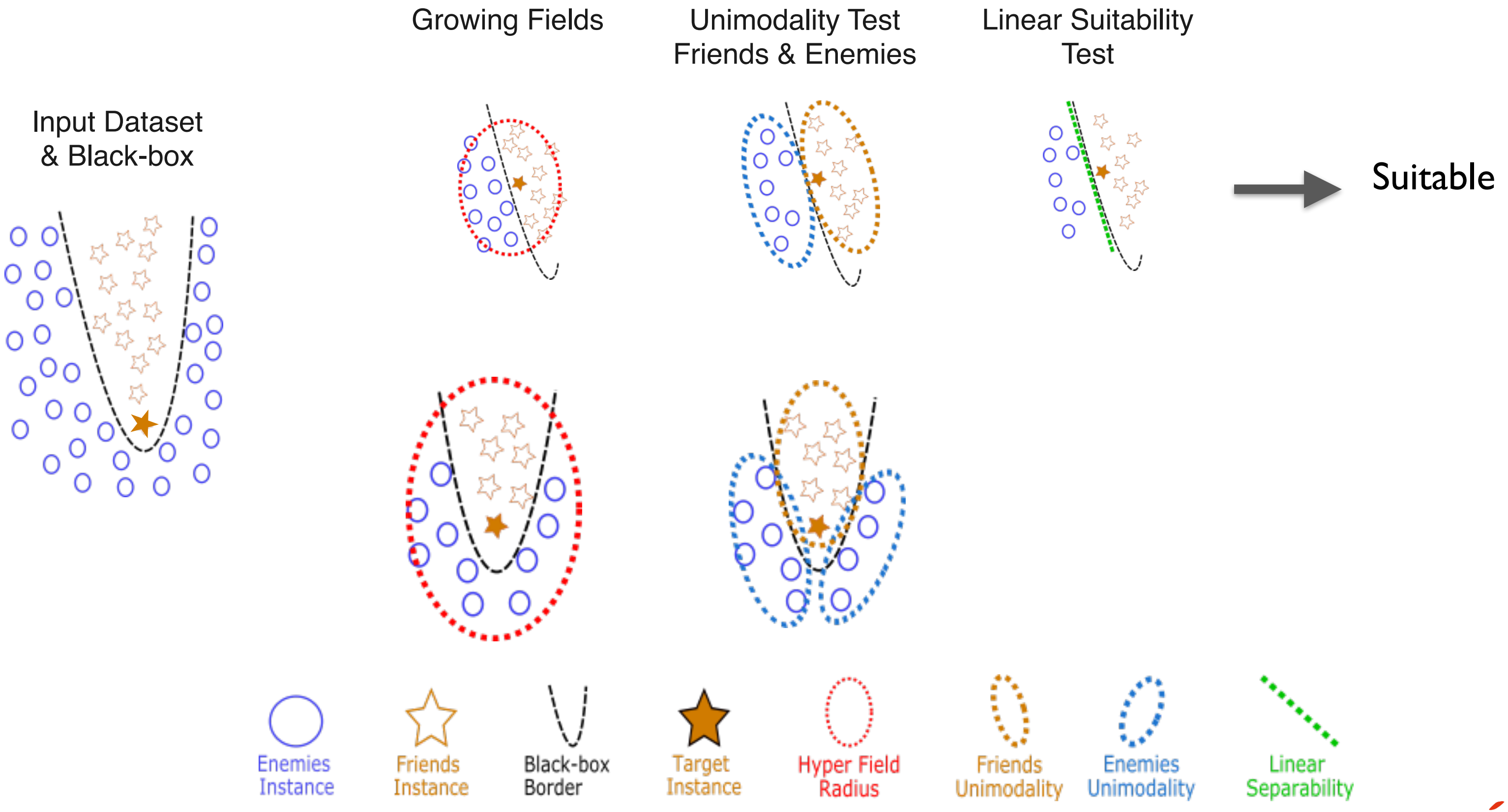
When Are Linear Explanations Adapted? — Oracle



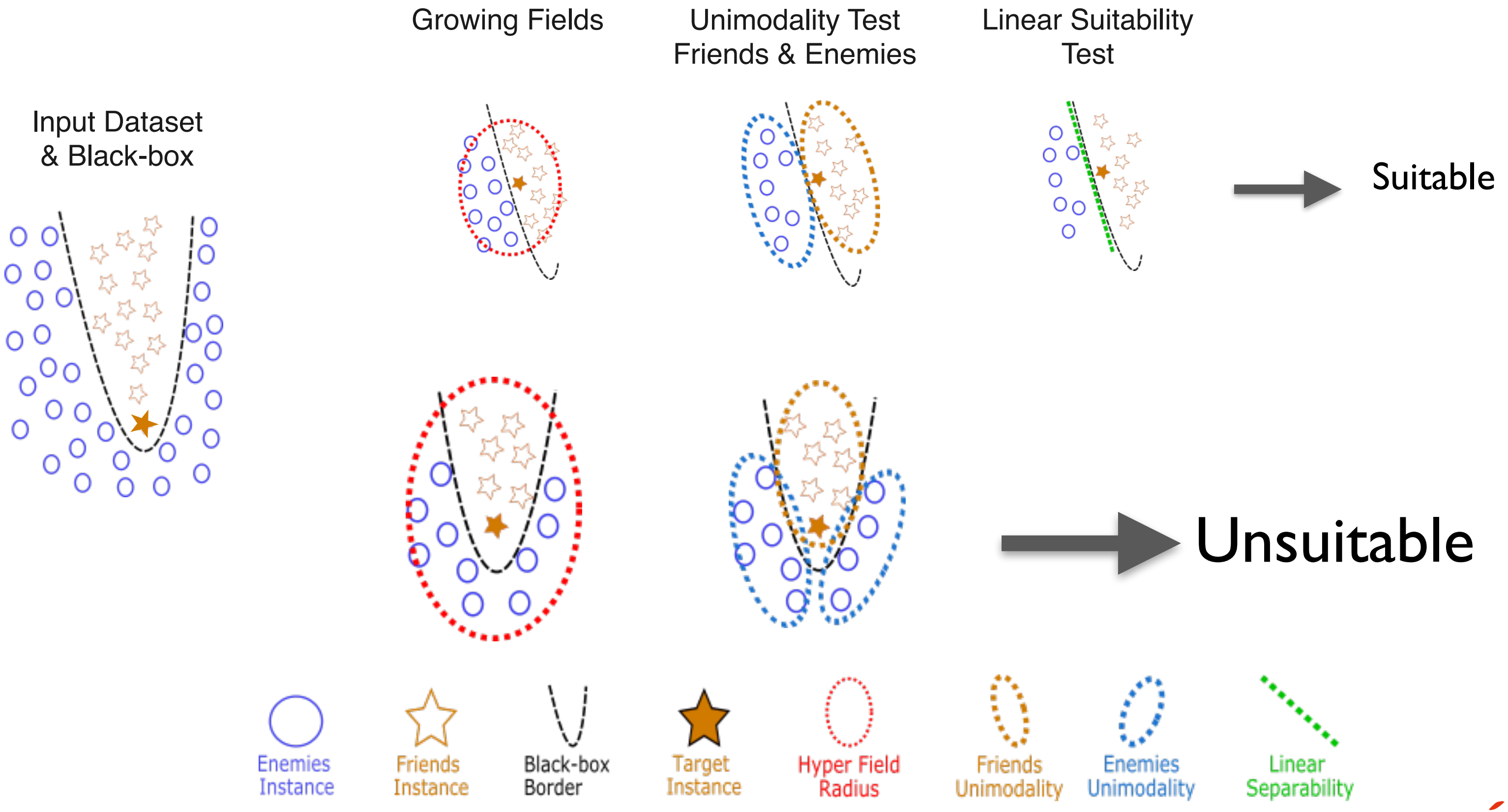
When Are Linear Explanations Adapted? — Oracle



When Are Linear Explanations Adapted? — Oracle



When Are Linear Explanations Adapted? — Oracle



Adherence Experiments — Oracle

- Adherence:
    - Agreement between linear explanation and black box model predictions

Adherence Experiments — Oracle

- Adherence:
  - Agreement between linear explanation and black box model predictions

- Per instance:
  - Growing Fields generates artificial instances
  - We compute the average accuracy over:
    1. The artificial instances
    2. Linear explanation and black box outcome

Adherence Experiments — Oracle

- Adherence:
  - Agreement between linear explanation and black box model predictions

- Per instance:
  - Growing Fields generates artificial instances
  - We compute the average accuracy over:
    1. The artificial instances
    2. Linear explanation and black box outcome

- Comparison of Linear Explanation (LE) average accuracy when
  - APE Oracle indicates suitable LE_{uni}
  - APE Oracle indicates not suitable LE_{mul}

$$\Delta acc = acc(LE_{uni}) - acc(LE_{mul})$$

Adherence Experiments — Oracle

- Adherence:
  - Agreement between linear explanation and black box model predictions

- Per instance:
  - Growing Fields generates artificial instances
  - We compute the average accuracy over:
    1. The artificial instances
    2. Linear explanation and black box outcome

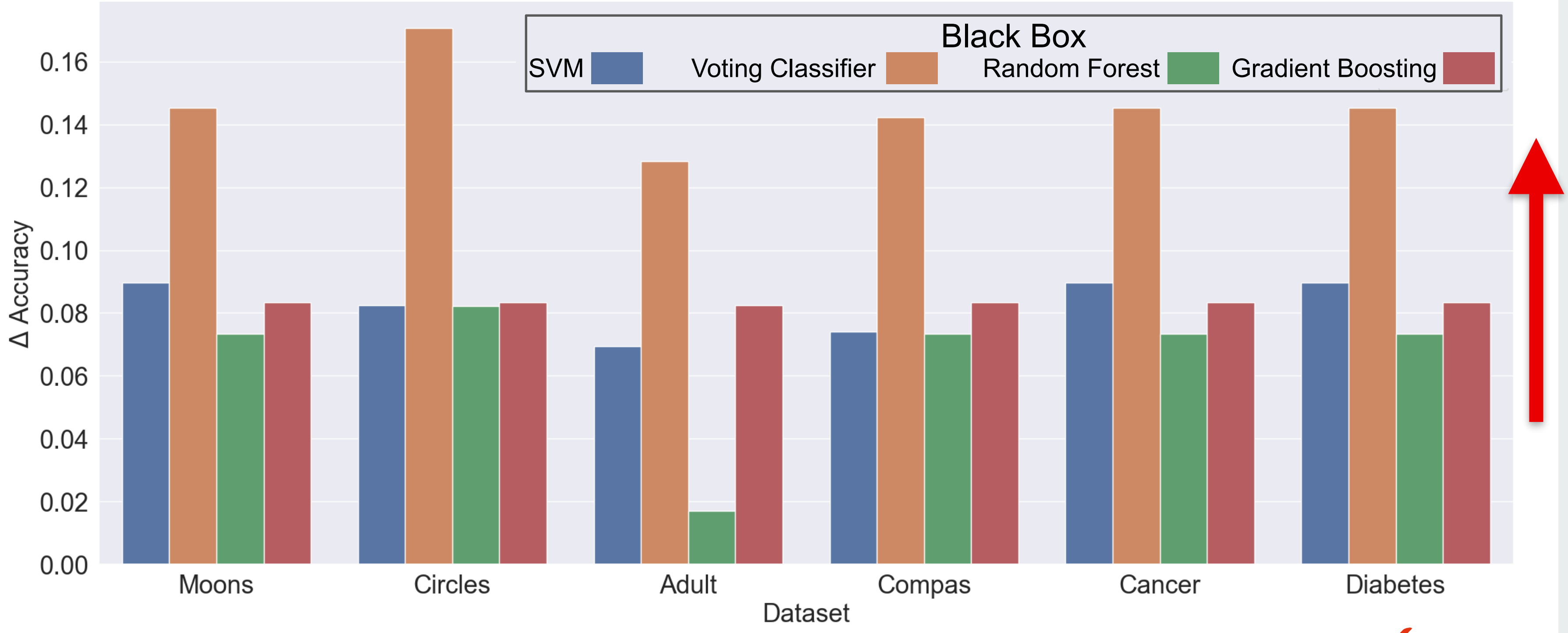
- Comparison of Linear Explanation (LE) average accuracy when
  - APE Oracle indicates suitable LE_{uni}
  - APE Oracle indicates not suitable LE_{mul}

$$\Delta acc = acc(LE_{uni}) - acc(LE_{mul})$$

- On 12 datasets & 6 black boxes

Adherence Results — Oracle

- Oracle' abilities to determine in which situations a single linear explanations is adapted



Fidelity Results — Oracle

- Fidelity: Features returned by the linear explanation are features actually used by the black box

Fidelity Results — Oracle

- Fidelity: Features returned by the linear explanation are features actually used by the black box

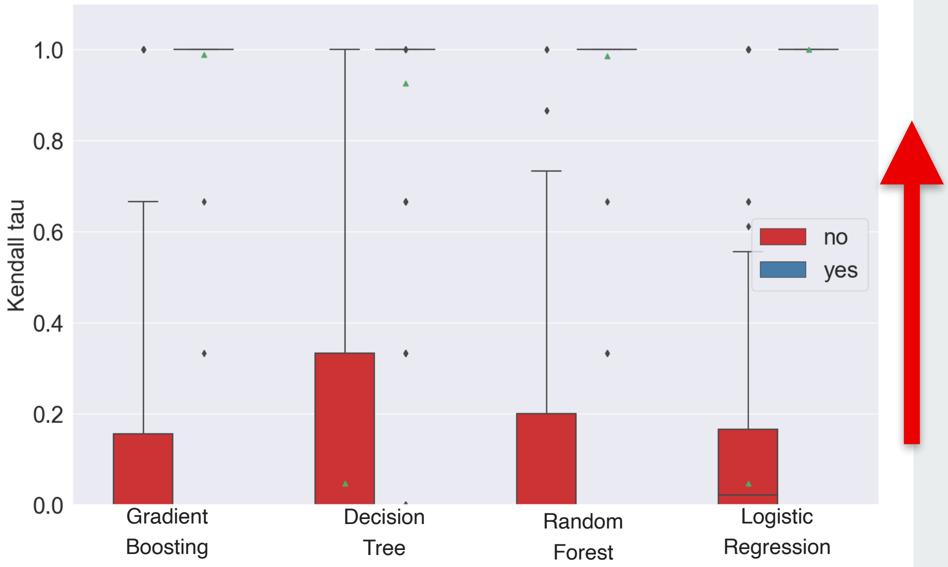
- “Glass-box” classifiers:
    - Not all features are employed to classify
    - Features employed are known

Fidelity Results — Oracle

- Fidelity: Features returned by the linear explanation are features actually used by the black box

- "Glass-box" classifiers:
  - Not all features are employed to classify
  - Features employed are known

- Comparison of average kendall tau when
  - APE Oracle indicates **suitable** "yes"
  - APE Oracle indicates **not suitable** "no"



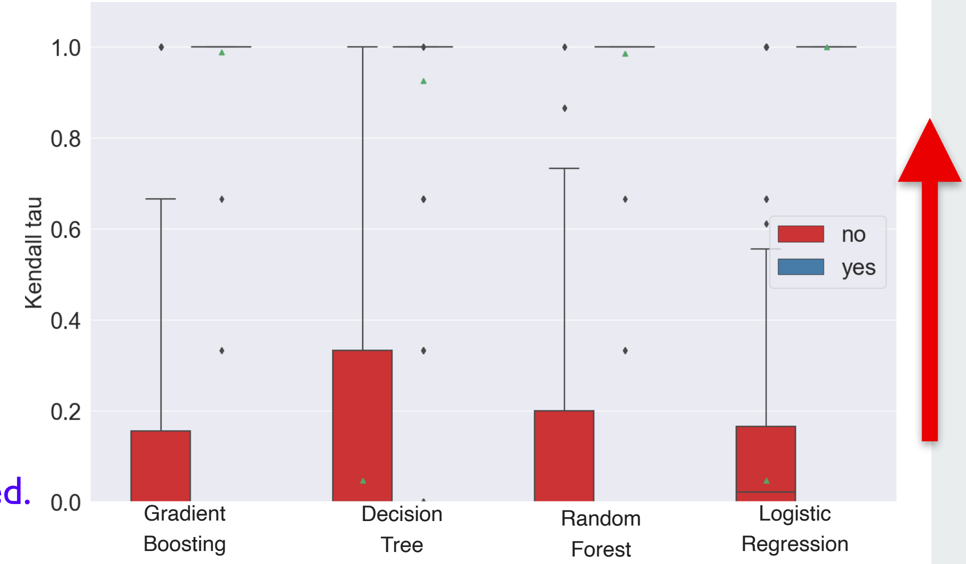
Fidelity Results — Oracle

- Fidelity: Features returned by the linear explanation are features actually used by the black box

- "Glass-box" classifiers:
  - Not all features are employed to classify
  - Features employed are known

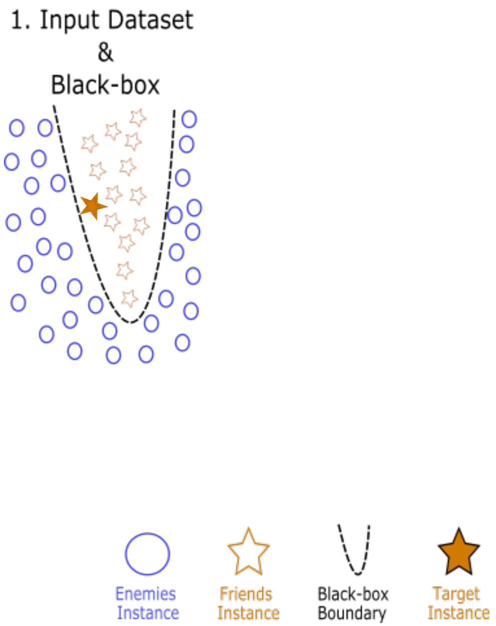
- Comparison of average kendall tau when
  - APE Oracle indicates **suitable** "yes"
  - APE Oracle indicates **not suitable** "no"

- Linear Explanation finds the features employed when the Oracle indicates adapted.

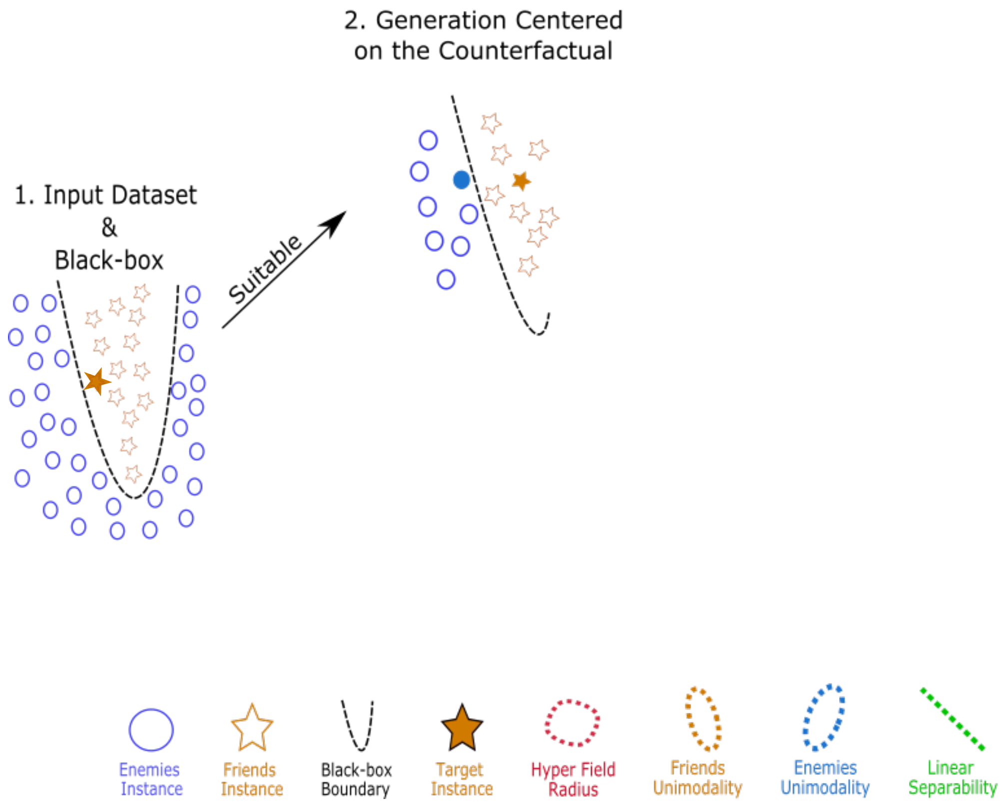


APE: Adapted Post-hoc Explanations — Framework

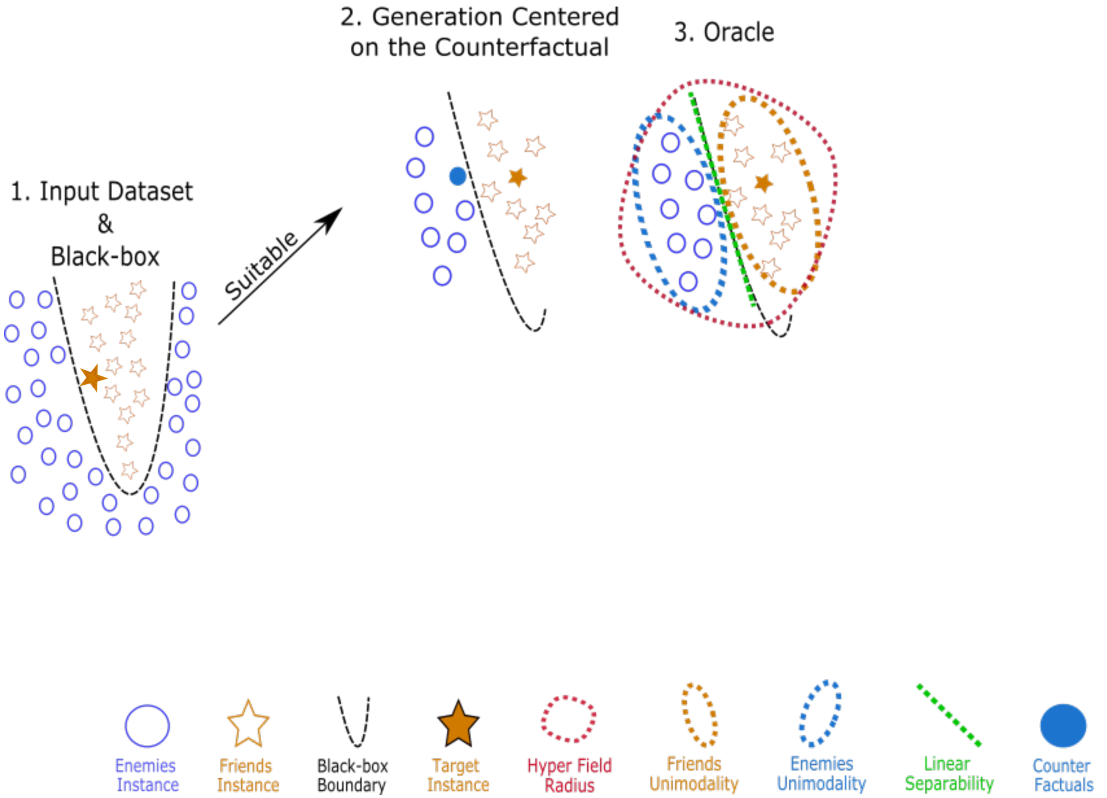
APE: Adapted Post-hoc Explanations — Framework



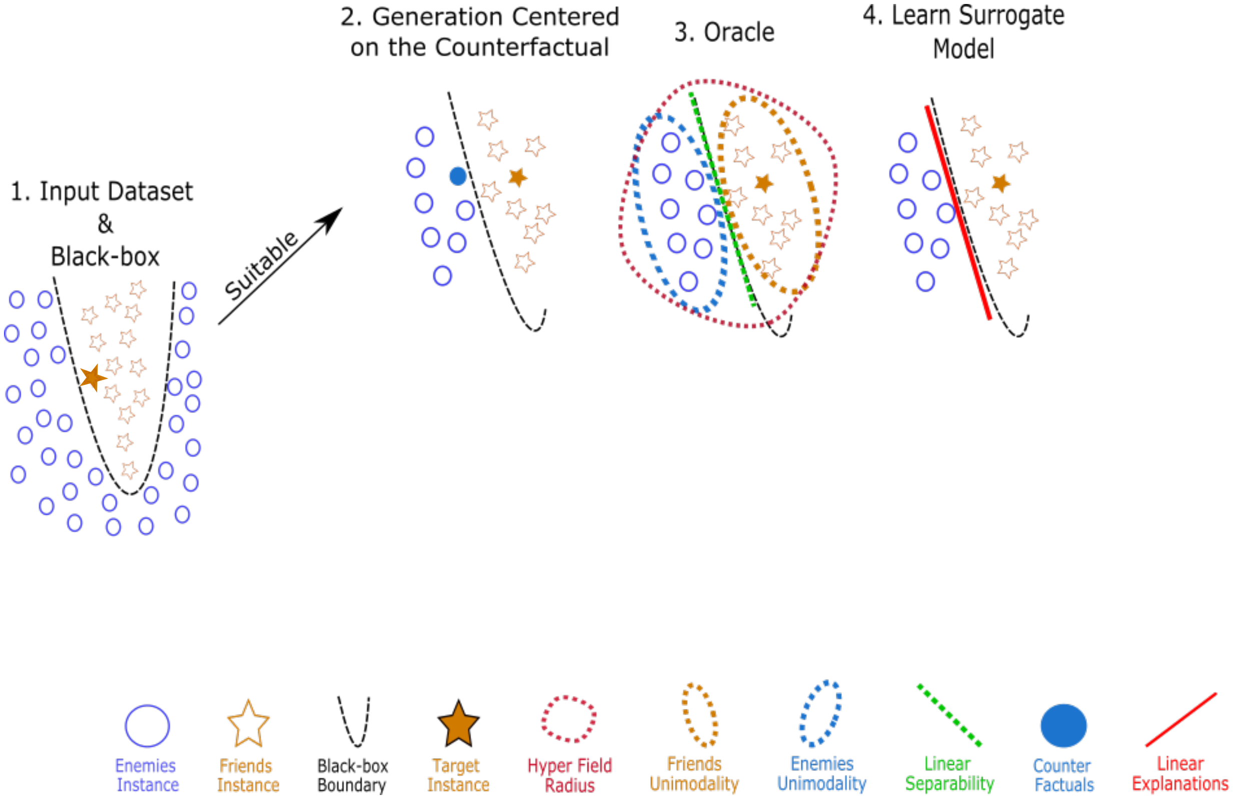
APE: Adapted Post-hoc Explanations — Framework



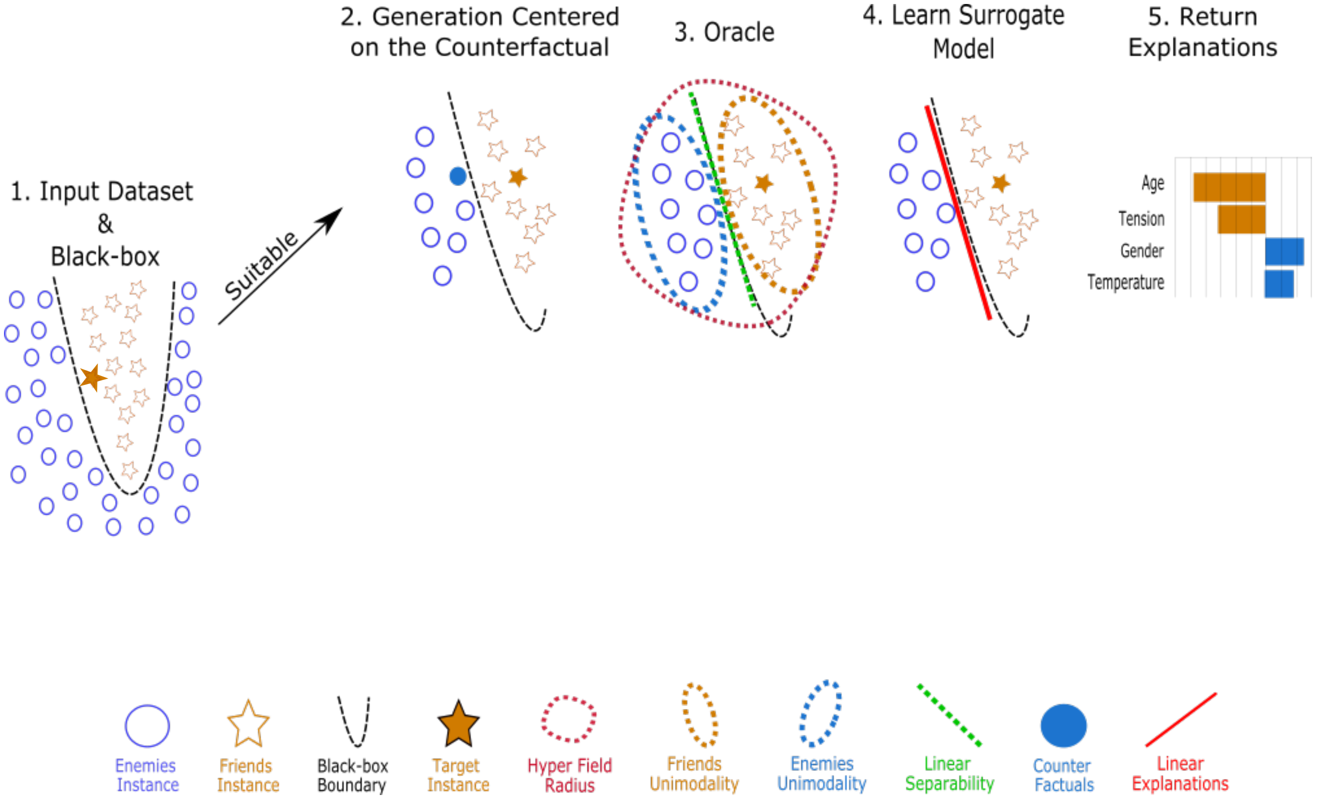
APE: Adapted Post-hoc Explanations — Framework



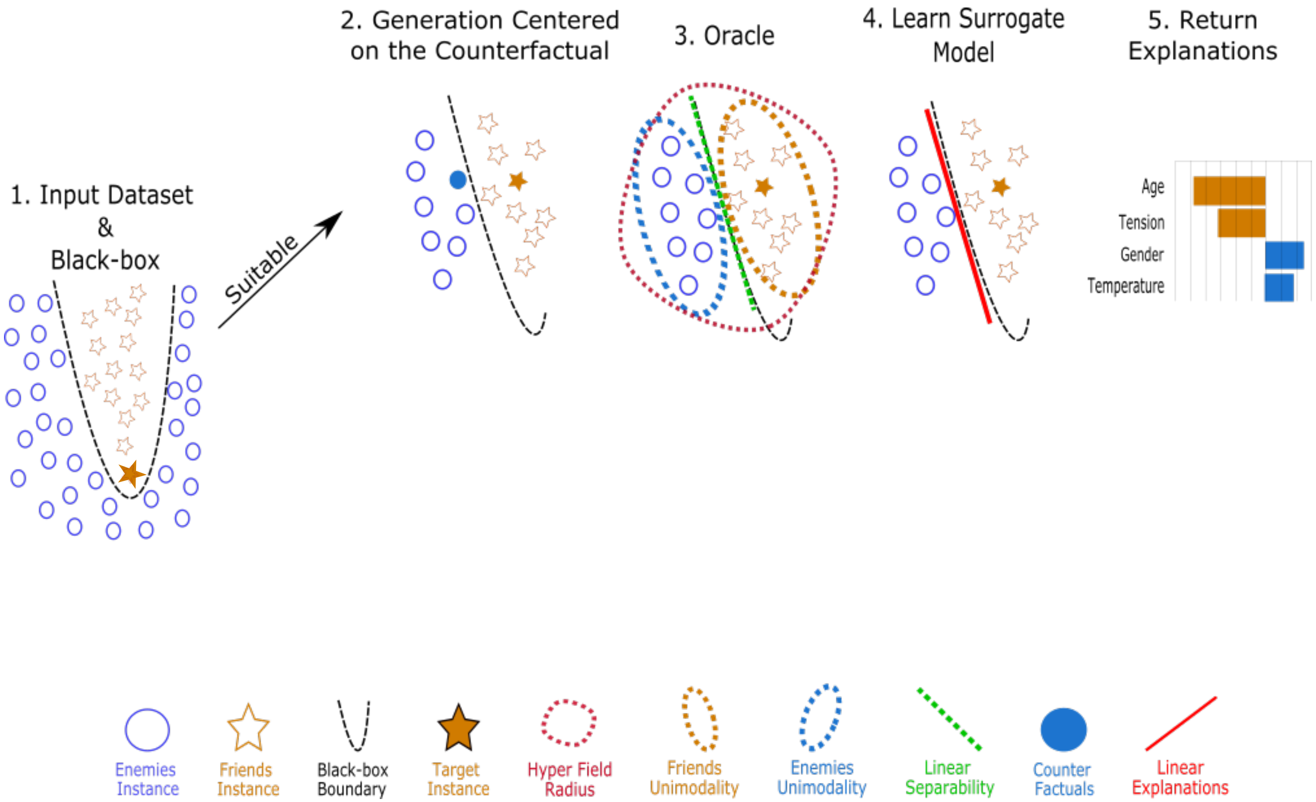
APE: Adapted Post-hoc Explanations — Framework



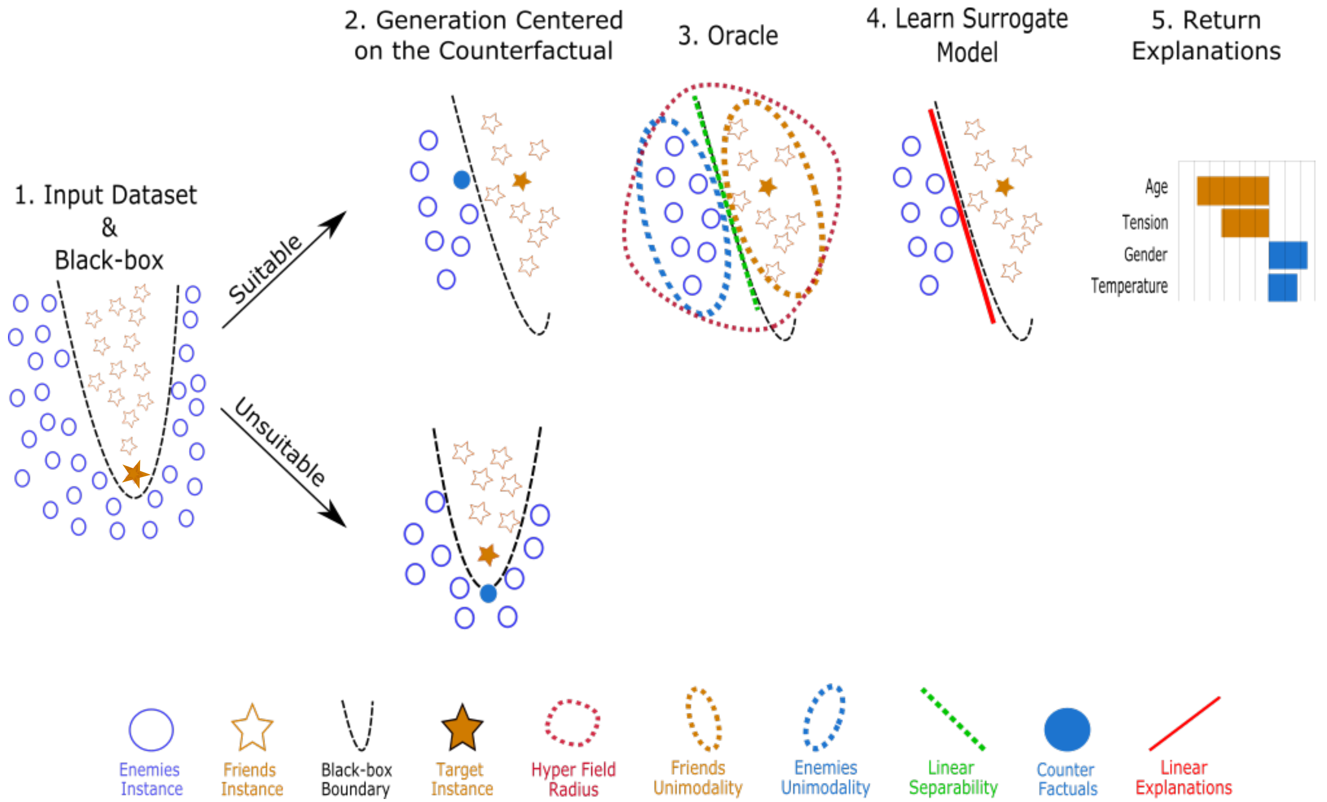
APE: Adapted Post-hoc Explanations — Framework



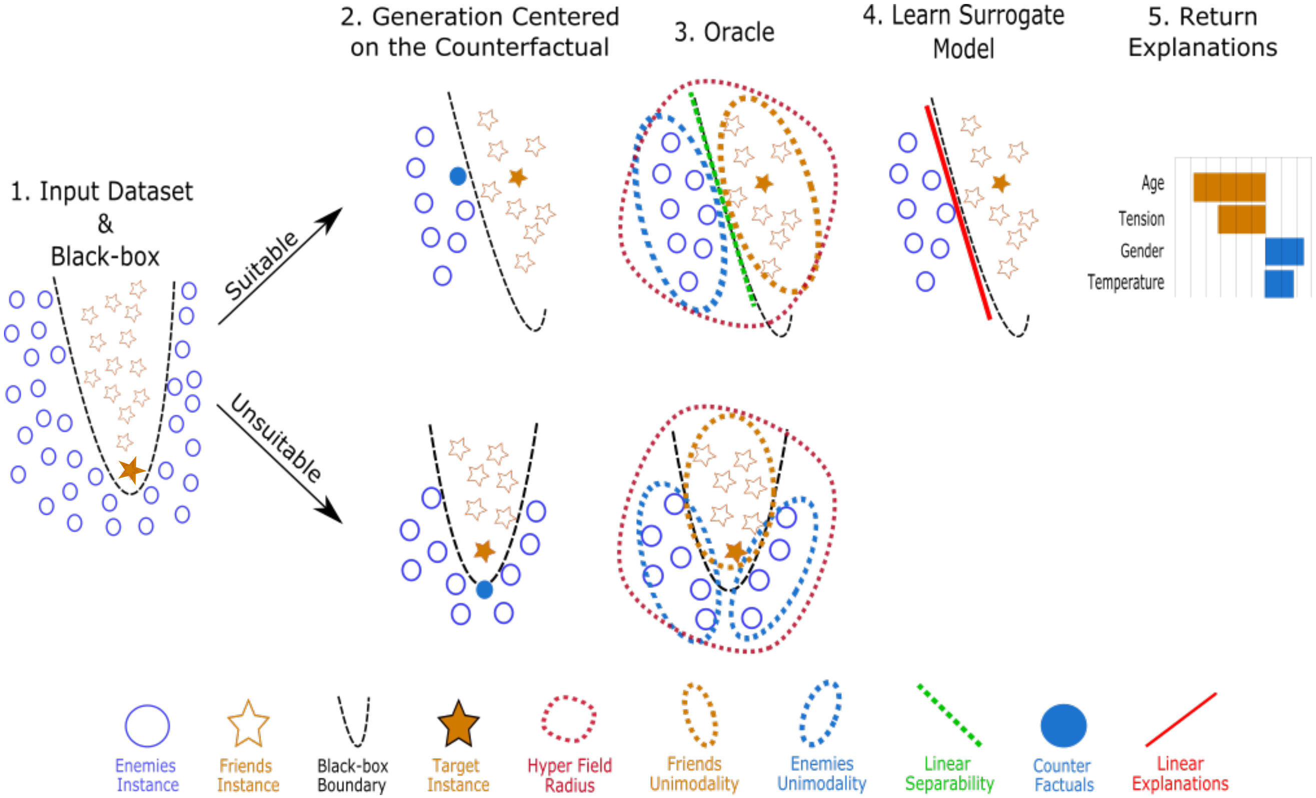
APE: Adapted Post-hoc Explanations — Framework



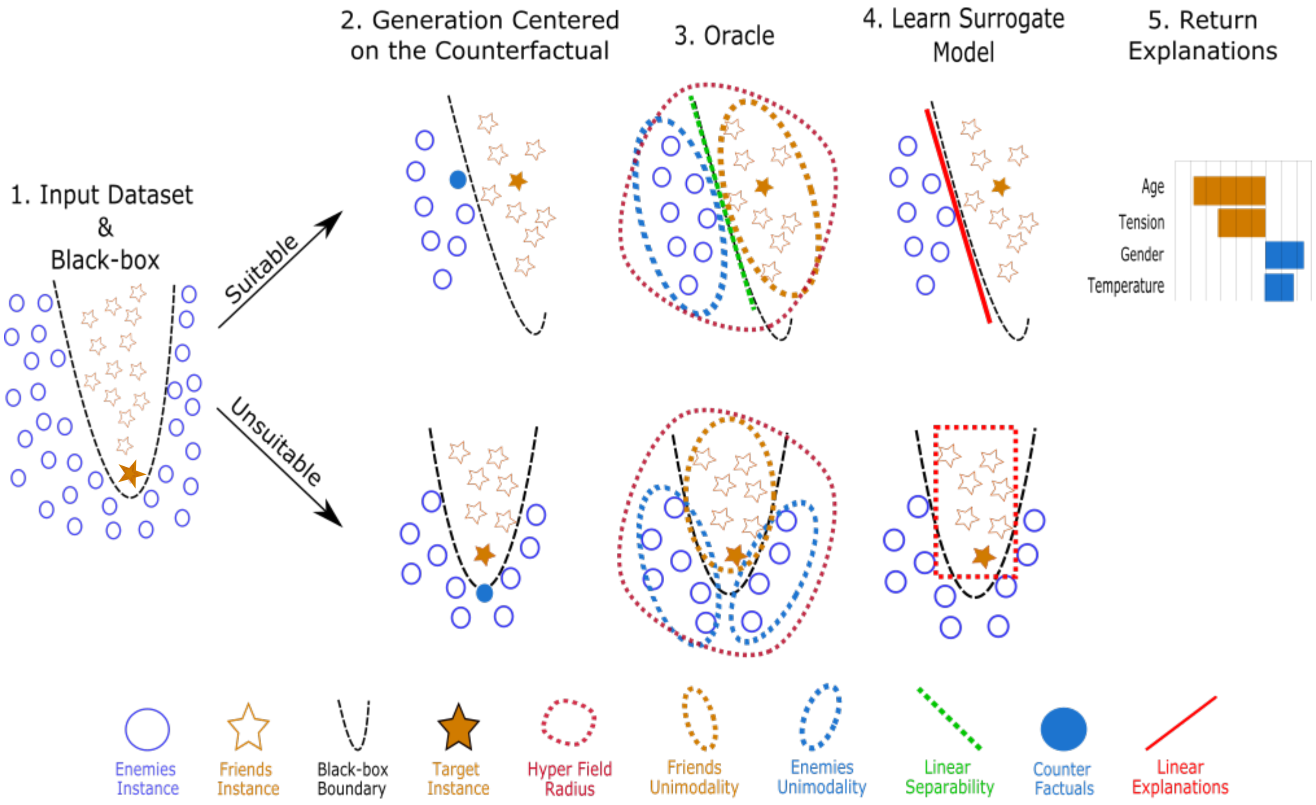
APE: Adapted Post-hoc Explanations — Framework



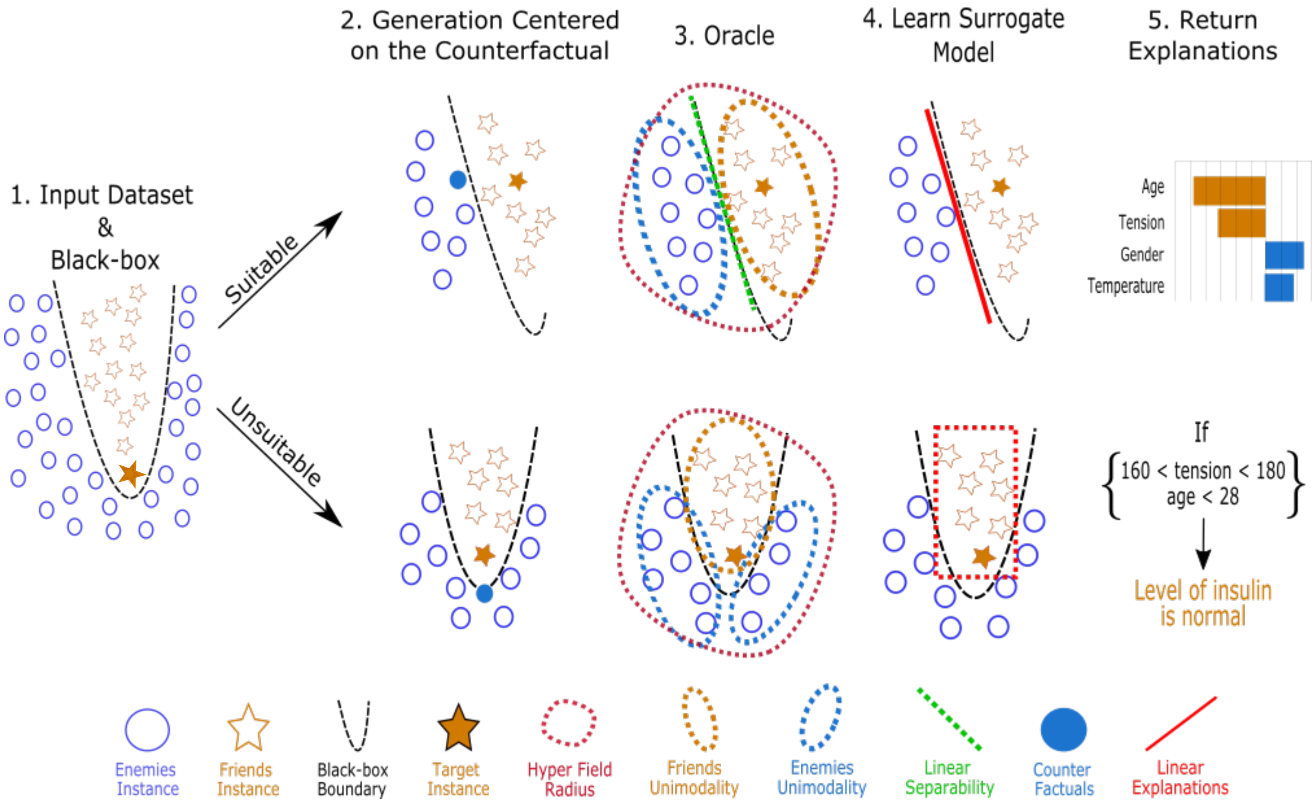
APE: Adapted Post-hoc Explanations — Framework



APE: Adapted Post-hoc Explanations — Framework



APE: Adapted Post-hoc Explanations — Framework



APE: Adapted Post-hoc Explanations

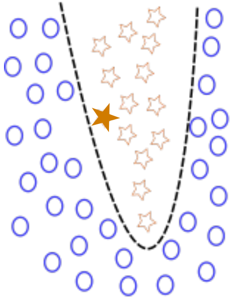
- We propose 2 novels explanation methods:
  A. APEa: Linear if suitable and Anchors *(5)* otherwise
  B. APEt: Linear if suitable and a shallow decision tree otherwise

(5) Tulio Ribeiro *et al*,. Anchors: High Precision Model-Agnostic Explanations. AAAI 2018

APE: Adapted Post-hoc Explanations

- We propose 2 novels explanation methods:
  A. APEa: Linear if suitable and Anchors (5) otherwise
  B. APEt: Linear if suitable and a shallow decision tree otherwise

Input Dataset & Black-box

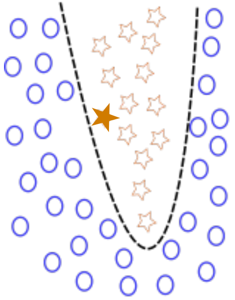


(5) Tulio Ribeiro *et al*,. Anchors: High Precision Model-Agnostic Explanations. AAAI 2018

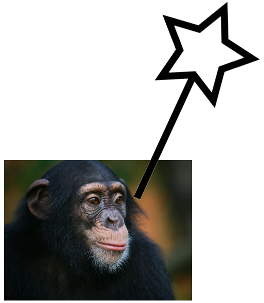
APE: Adapted Post-hoc Explanations

- We propose 2 novels explanation methods:
  A. APEa: Linear if suitable and Anchors (5) otherwise
  B. APEt: Linear if suitable and a shallow decision tree otherwise

Input Dataset & Black-box



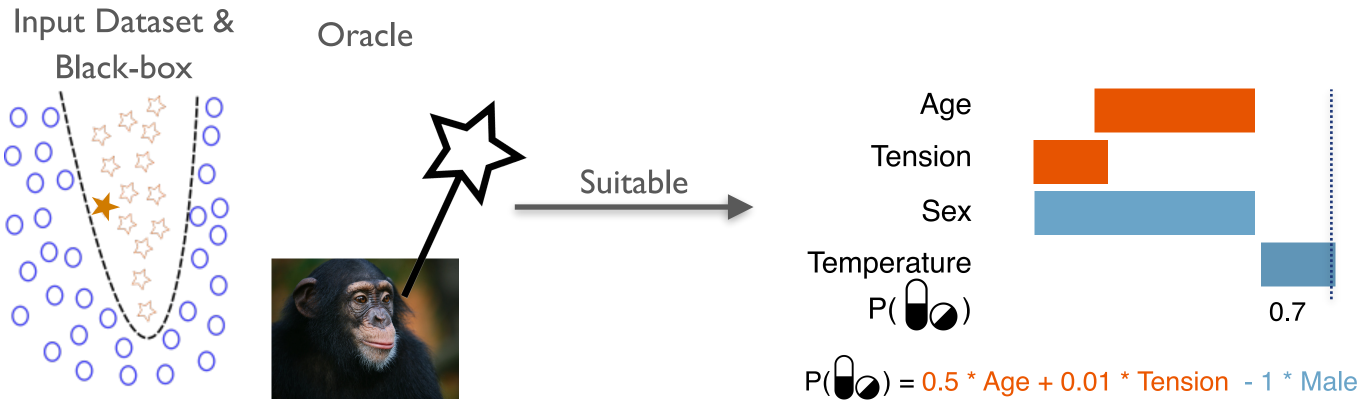
Oracle



(5) Tulio Ribeiro *et al*,. Anchors: High Precision Model-Agnostic Explanations. AAAI 2018

APE: Adapted Post-hoc Explanations

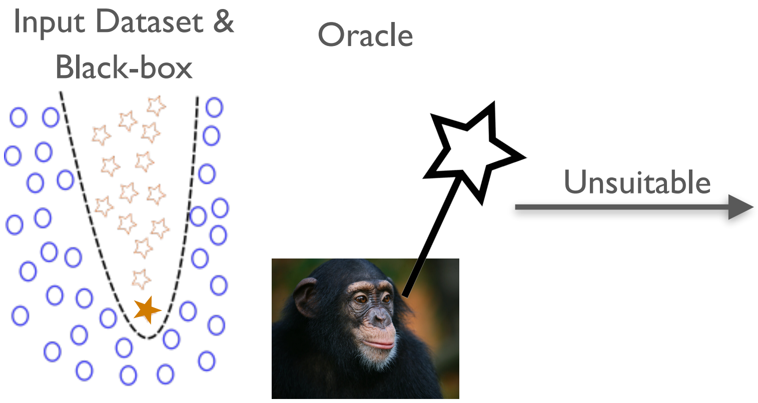
- We propose 2 novels explanation methods:
  A. APEa: Linear if suitable and Anchors *(5)* otherwise
  B. APEt: Linear if suitable and a shallow decision tree otherwise

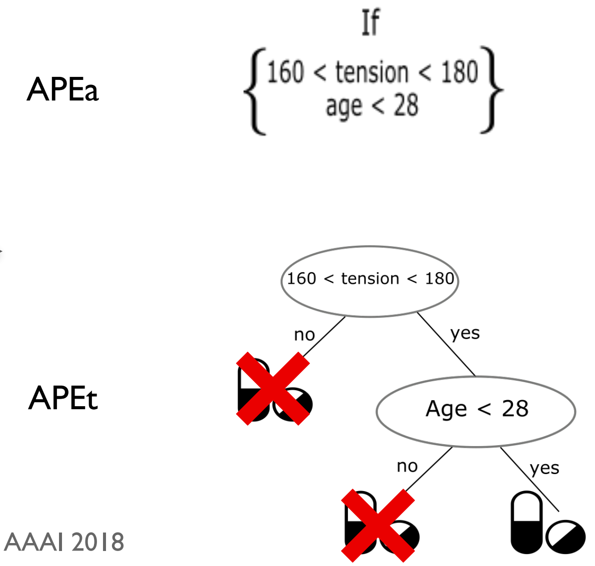


(5) Tulio Ribeiro *et al*,. Anchors: High Precision Model-Agnostic Explanations. AAAI 2018

APE: Adapted Post-hoc Explanations

- We propose 2 novels explanation methods:
  A. APEa: Linear if suitable and Anchors *(5)* otherwise
  B. APEt: Linear if suitable and a shallow decision tree otherwise





(5) Tulio Ribeiro *et al*,. Anchors: High Precision Model-Agnostic Explanations. AAAI 2018

Experiments — Framework

Experiments — Framework

- We compute the average adherence of 4 explanation methods:
  - LIME *(1)*
  - Local Surrogate (LS) *(11)*
  - APEa: LS if suitable and Anchors otherwise
  - APEt: LS if suitable and a shallow decision tree otherwise

(1) Tulio Ribeiro *et al.*, ``Why Should I Trust You?'': Explaining the Predictions of Any Classifier. KDD 2016
(11) Thibault Laugel *et al.*, Defining Locality for Surrogates in Post-hoc Interpretablity. ICML 2018

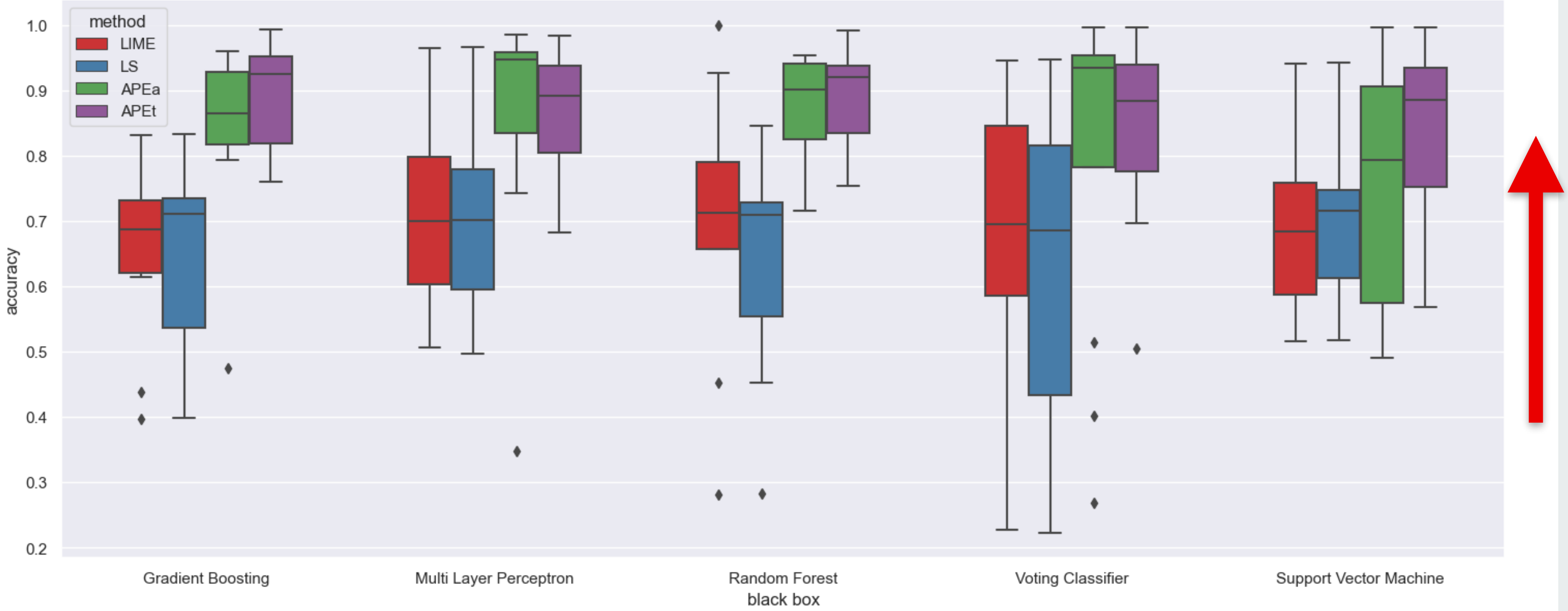
Experiments — Framework

- We compute the average adherence of 4 explanation methods:
  - LIME *(I)*
  - Local Surrogate (LS) *(II)*
  - APEa: LS if suitable and Anchors otherwise
  - APEt: LS if suitable and a shallow decision tree otherwise
- Based on the prediction of 5 black box models:
  - Gradient Boosting
  - Multi Layer Perceptron
  - Random Forest
  - Voting Classifier
  - Support Vector Machines

(I) Tulio Ribeiro *et al.*, ``Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD 2016
(II) Thibault Laugel *et al.*, Defining Locality for Surrogates in Post-hoc Interpretablity. ICML 2018

Results — Comparison With Linear

- Adherence gain of our methods compare to linear explanations alone



Summary

Summary

- We introduce Growing Fields, a method to:
    - Detect the closest decision boundary
    - Generate artificial instances based on the data distribution

Summary

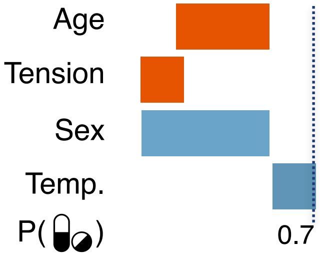
- We introduce Growing Fields, a method to:
  - Detect the closest decision boundary
  - Generate artificial instances based on the data distribution

- We present an Oracle to determine *a priori*:
  - The suitability of a linear explanation to approximate locally a black box model

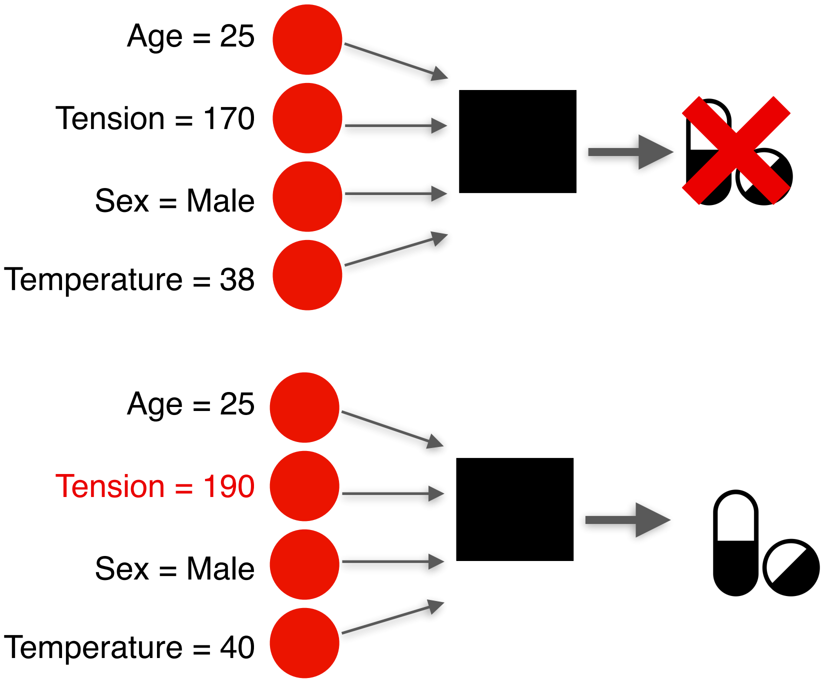
Summary

- We introduce Growing Fields, a method to:
  - Detect the closest decision boundary
  - Generate artificial instances based on the data distribution
- We present an Oracle to determine *a priori*:
  - The suitability of a linear explanation to approximate locally a black box model
- We develop APE a novel method that:
  - Returns linear explanation if adapted
  - Returns rule-based explanation otherwise

What about the user?



P( ) = 0.5 * Age +
0.01 * Tension
- 1 * Male



If the user has a tension between 160 and 180, while being under 28, then the level of insulin is moderate

(Feature Attribution) (Example-based) (Rule-based)

Part II: How to generate the best explanation from a user perspective?

Impact of Explanation Techniques and Representations on Users Trust and Comprehension [Under Review CSCW '24]

Julien Delaunay

Second Contribution of my Thesis

Second Contribution of my Thesis

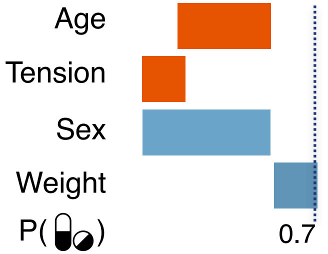
- Methodological framework for conducting user studies:
    - Investigate the impact of explanation on users
    - Metrics to measure users' trust and understanding

Second Contribution of my Thesis

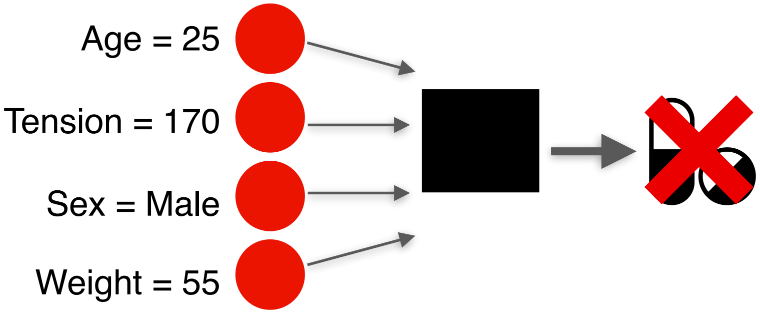
- Methodological framework for conducting user studies:
  - Investigate the impact of explanation on users
  - Metrics to measure users' trust and understanding

- A user study:
  - 280 crowdworkers
  - Two domains (healthcare and law)

Problem Statement — Users Perception

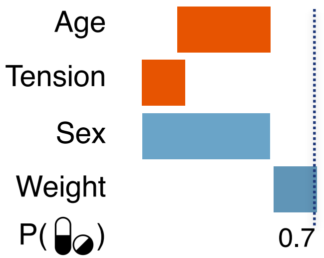


P( ) = 0.5 * Age +
0.01 * Tension
- 1 * Male

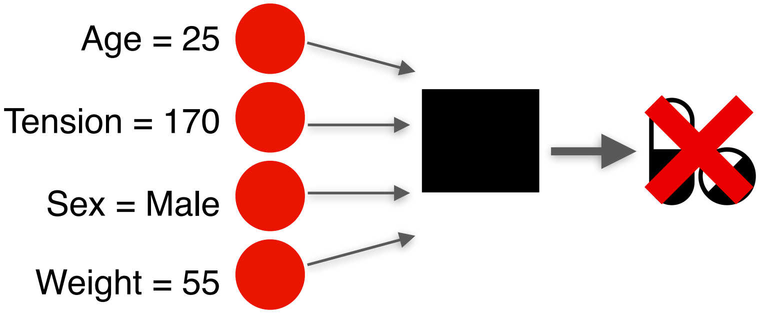


If the user has a tension between 150 and 170, while being under 28, then the level of insulin is moderate

Problem Statement — Users Perception



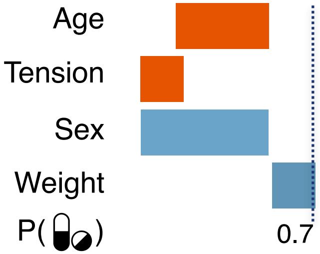
P( ) = 0.5 * Age +
0.01 * Tension
- 1 * Male



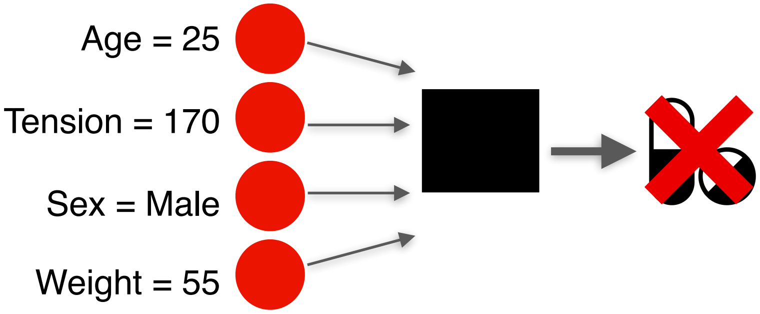
If the user has a tension between 150 and 170, while being under 28, then the level of insulin is moderate

RQ1: Which explanation technique provides the best explanations in terms of users' trust and comprehension of the AI model?

Problem Statement — Users Perception



P( ) = 0.5 * Age +
0.01 * Tension
- 1 * Male



If the user has a tension between 150 and 170, while being under 28, then the level of insulin is moderate

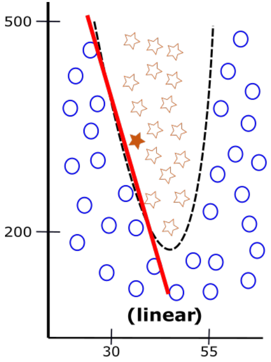
RQ1: Which explanation technique provides the best explanations in terms of users' trust and comprehension of the AI model?

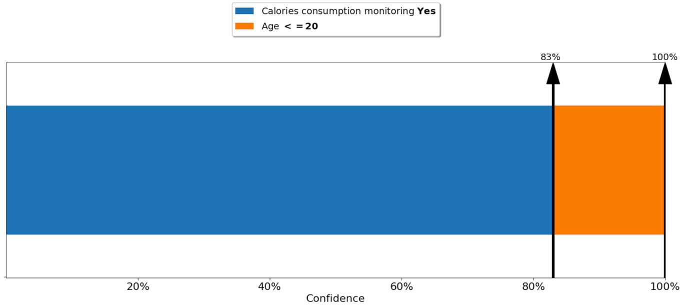
RQ2: Does the explanation's representation impact the users' trust and understanding?

Challenges We Faced When Designing The Study

Challenges We Faced When Designing The Study

1. How to represent these three different explanations techniques under one common representation?





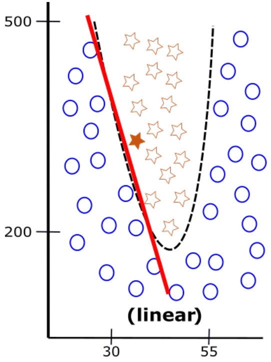
Based on the above data, the artificial intelligence (AI) tool has predicted **obesity**.

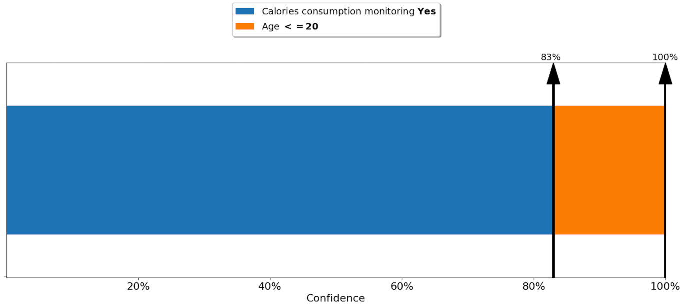
- First, because a family member **suffers** from overweight.
- Second, she is **aged between 23 and 26 years old**.
- Third, she **doesn't practice** physical activity weekly.

All together, it brings an AI's confidence of 95% for this **obesity** prediction

Challenges We Faced When Designing The Study

1. How to represent these three different explanations techniques under one common representation?





Based on the above data, the artificial intelligence (AI) tool has predicted **obesity**.

- First, because a family member **suffers** from overweight.
- Second, she is **aged between 23 and 26 years old**.
- Third, she **doesn't practice** physical activity weekly.

All together, it brings an AI's confidence of 95% for this **obesity** prediction

2. Which use case?
   - Domain understandable for a layperson / complex enough to require an AI model
     1. Risk of obesity
     2. Risk of recidivism

Participants' Initial Prediction

Participants' Initial Prediction

Information About
an Individual

Gender	Female
Age	23
Height	166
Family member has overweight	No
Frequent consumption of high caloric food	No
Frequency of consumption of vegetables	Sometimes
Number of daily meals	More than 3
Consumption of food between meals	Sometimes
Smoke	No
Consumption of water daily	More than 2L
Calories consumption monitoring	Yes
Physical activity frequency per week	2 or 4 days
Time using technology devices daily	0-2 hours
Consumption of alcohol	Sometimes
Transportation used	Public transportation

Participants' Initial Prediction

Information About an Individual

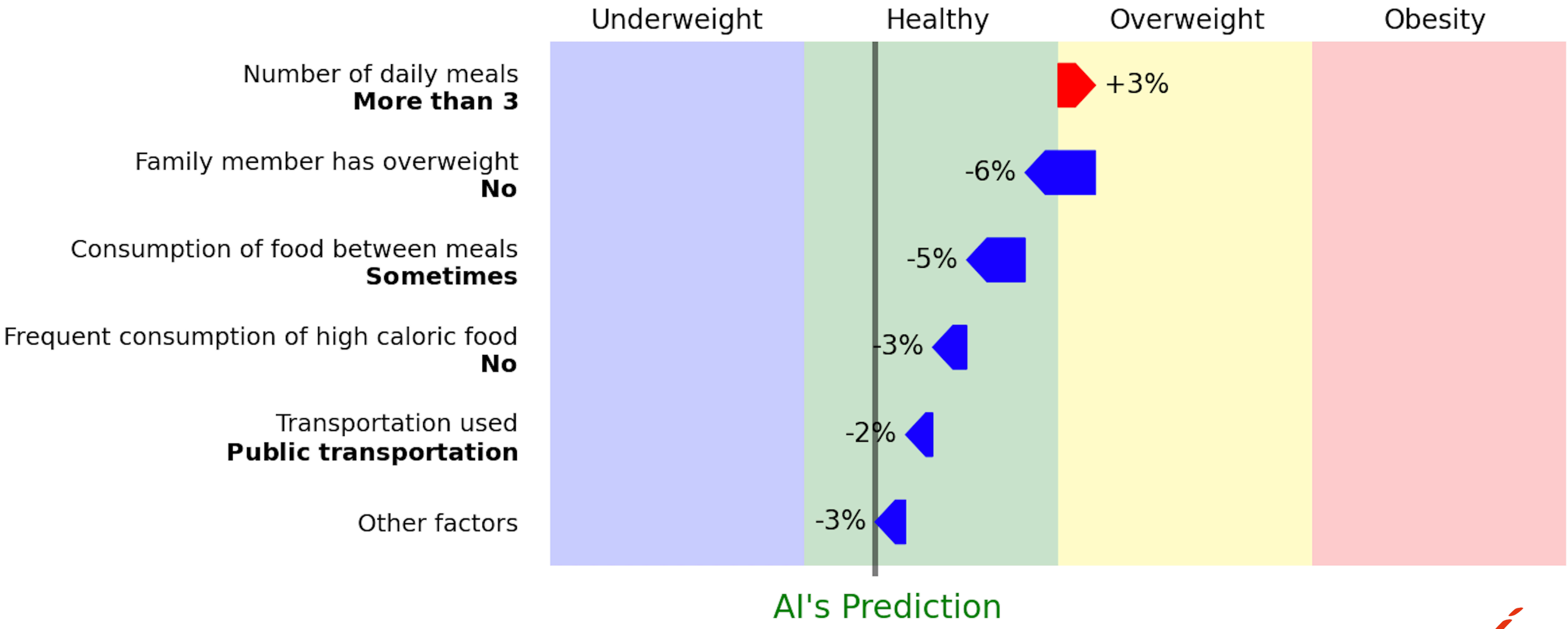
Gender	Female
Age	23
Height	166
Family member has overweight	No
Frequent consumption of high caloric food	No
Frequency of consumption of vegetables	Sometimes
Number of daily meals	More than 3
Consumption of food between meals	Sometimes
Smoke	No
Consumption of water daily	More than 2L
Calories consumption monitoring	Yes
Physical activity frequency per week	2 or 4 days
Time using technology devices daily	0-2 hours
Consumption of alcohol	Sometimes
Transportation used	Public transportation

Prediction Task

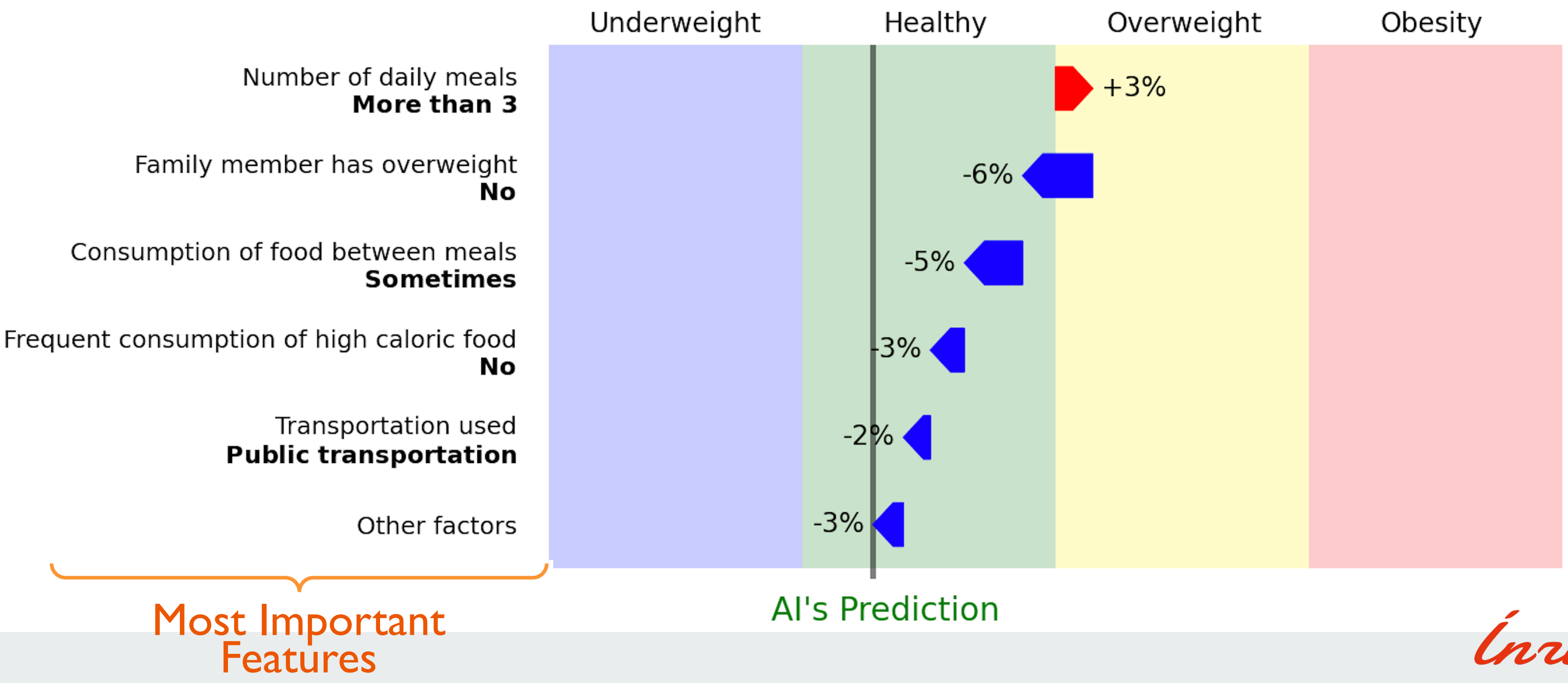
Based on the above information, to which of these four categories do you think this individual belongs?

- Underweight
- Healthy
- Overweight
- Obesity

Graphical Representation — Feature Attribution

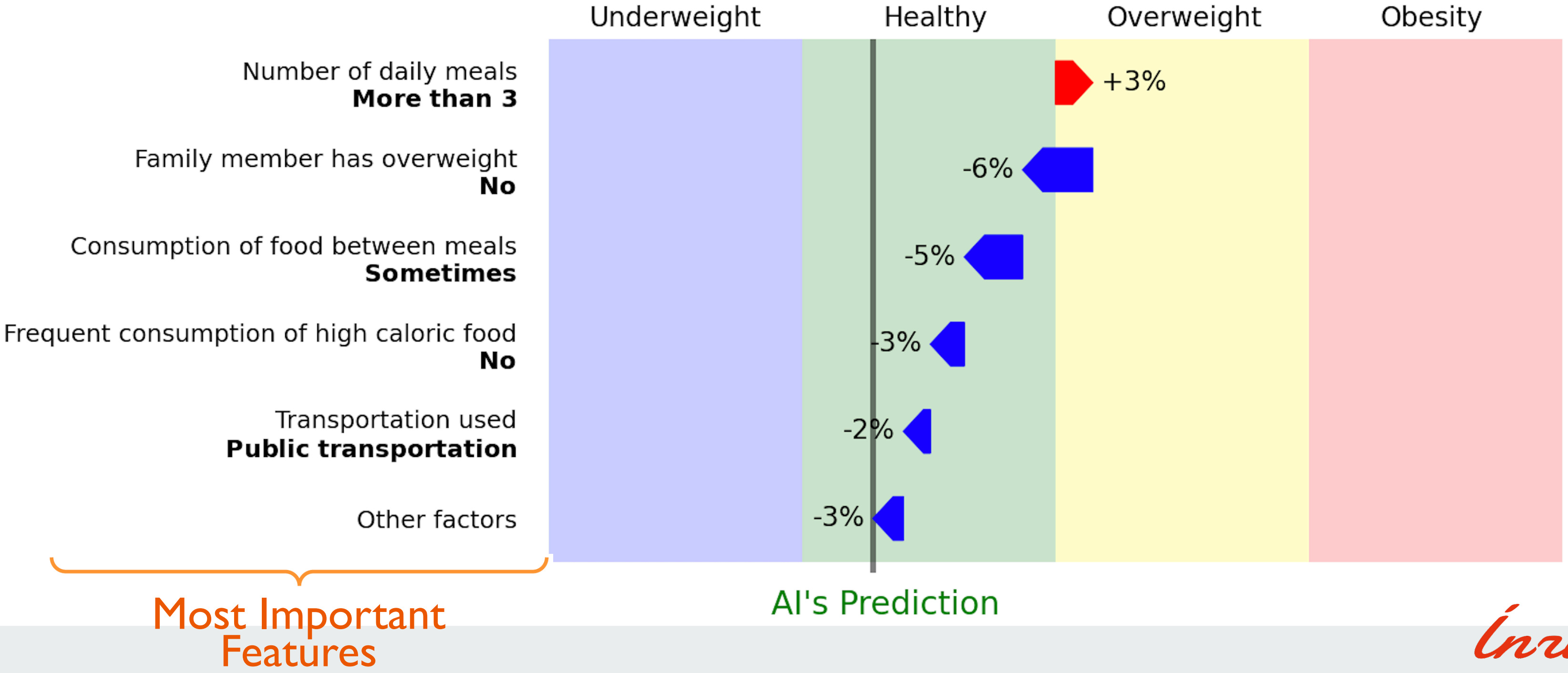


Graphical Representation — Feature Attribution

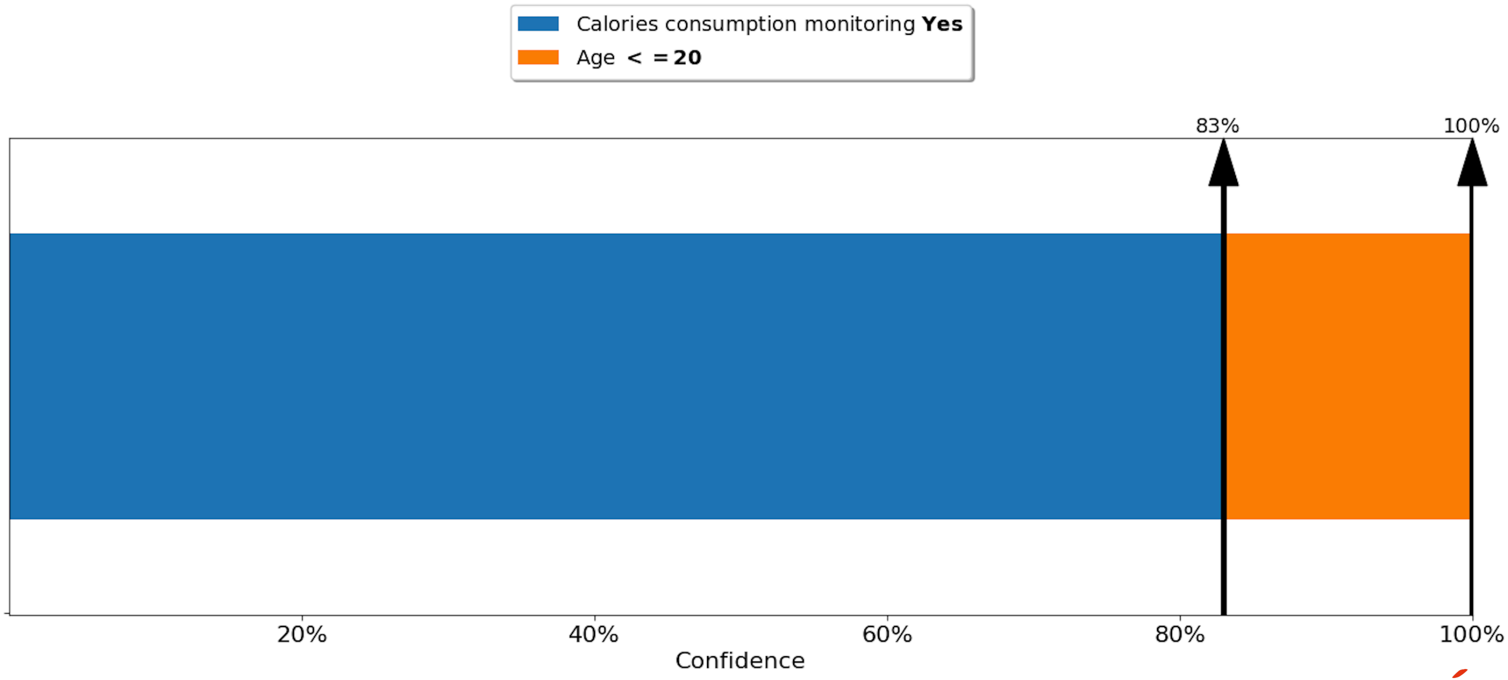


Graphical Representation — Feature Attribution

- Features that impacted the prediction:
  - Red (Blue) bars indicate an increased chance of being overweight or obese (underweight or healthy)
  - The values on the side correspond to the impact of the specific features on the prediction

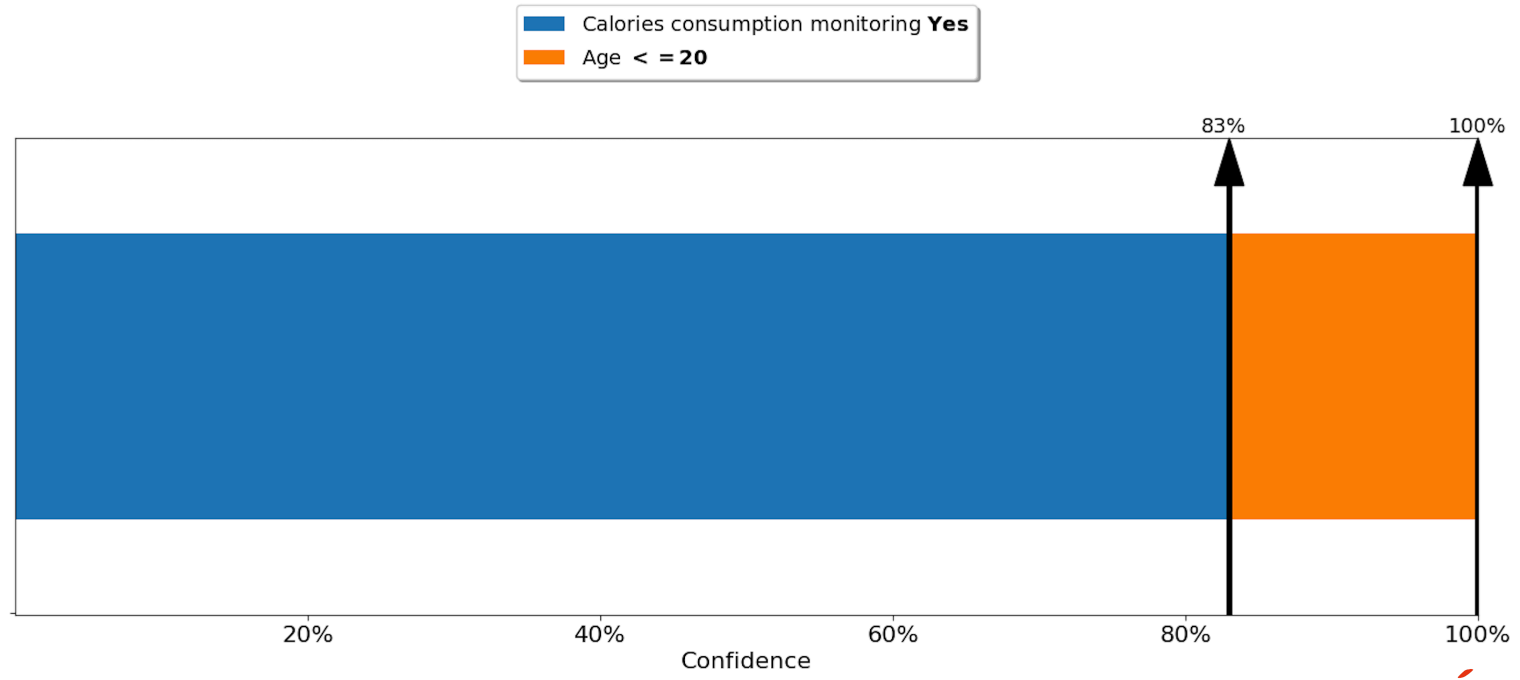


Graphical Representation — Rule-based

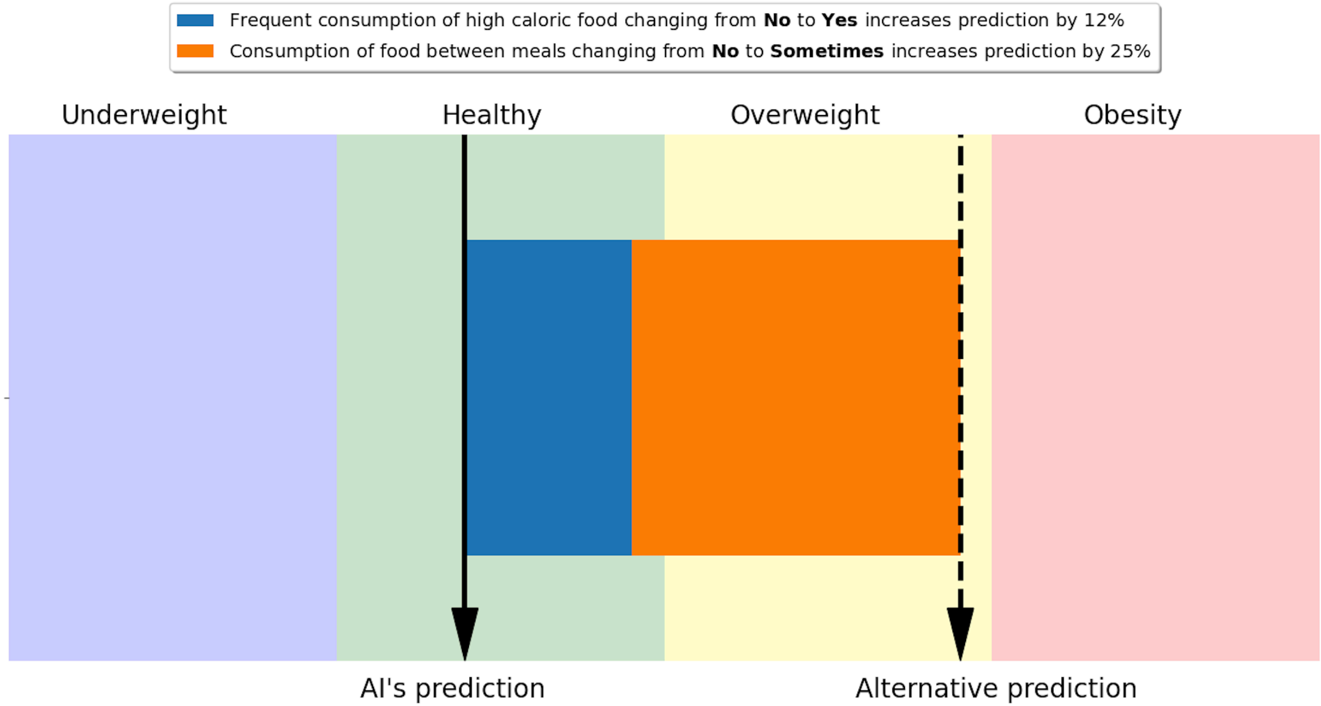


Graphical Representation — Rule-based

- Colored bars represent the importance of one user's answer to the prediction:
  - Numerical values correspond to the proportion of users for which the AI tool predicts **healthy**

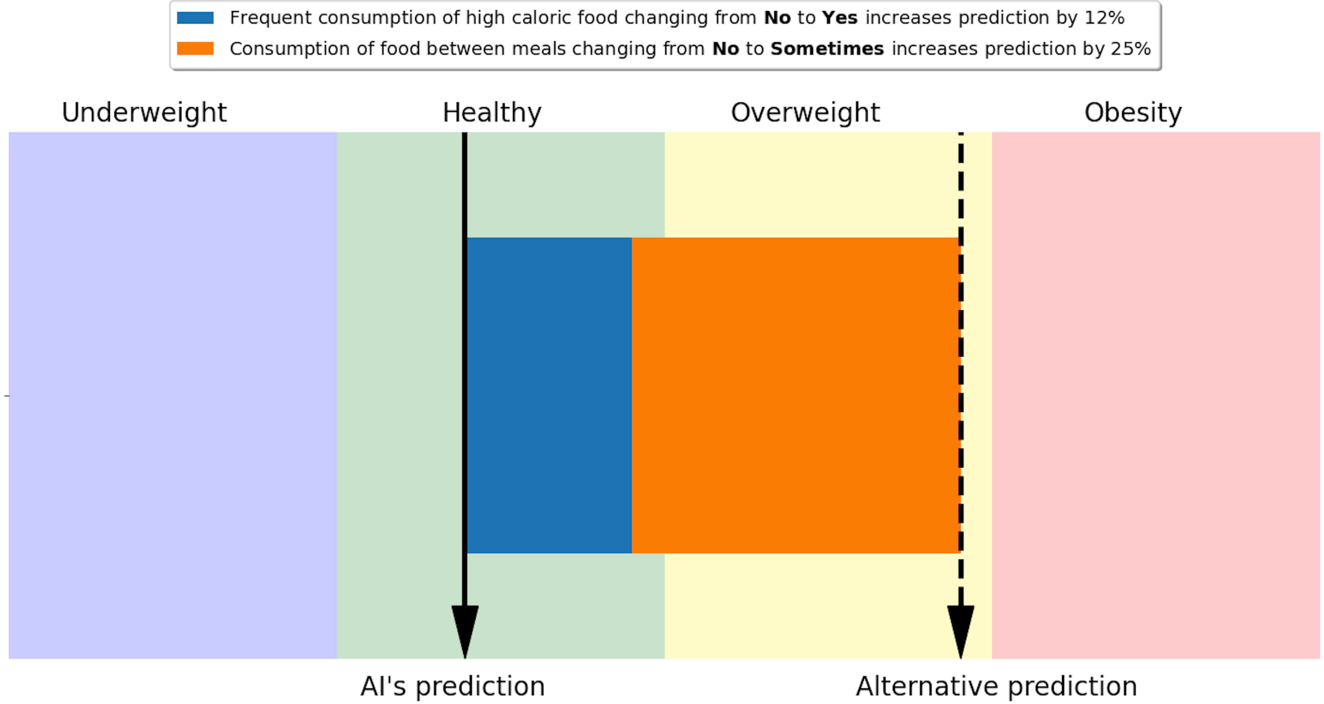


Graphical Representation — Counterfactual



Graphical Representation — Counterfactual

- Colored bars indicate most effective features to modify the prediction:
  - Length of the bars correspond to the importance of changing one answer's value to another

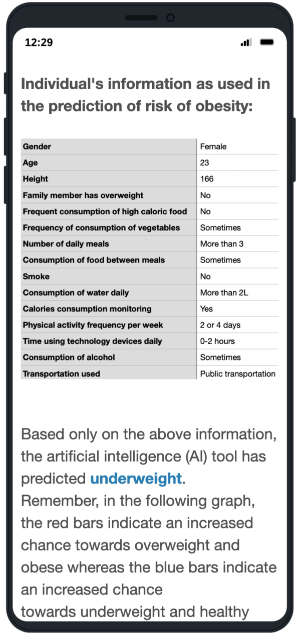


What Does a Survey Looks Like

Individual's information as used in the prediction of risk of obesity:

Gender	Female
Age	23
Height	166
Family member has overweight	No
Frequent consumption of high caloric food	No
Frequency of consumption of vegetables	Sometimes
Number of daily meals	More than 3
Consumption of food between meals	Sometimes
Smoke	No
Consumption of water daily	More than 2L
Calories consumption monitoring	Yes
Physical activity frequency per week	2 or 4 days
Time using technology devices daily	0-2 hours
Consumption of alcohol	Sometimes
Transportation used	Public transportation

Based only on the above information, the artificial intelligence (AI) tool has predicted **underweight**.
Remember, in the following graph, the red bars indicate an increased chance towards

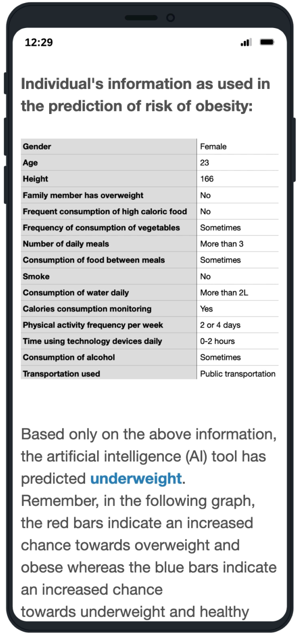


What Does a Survey Looks Like

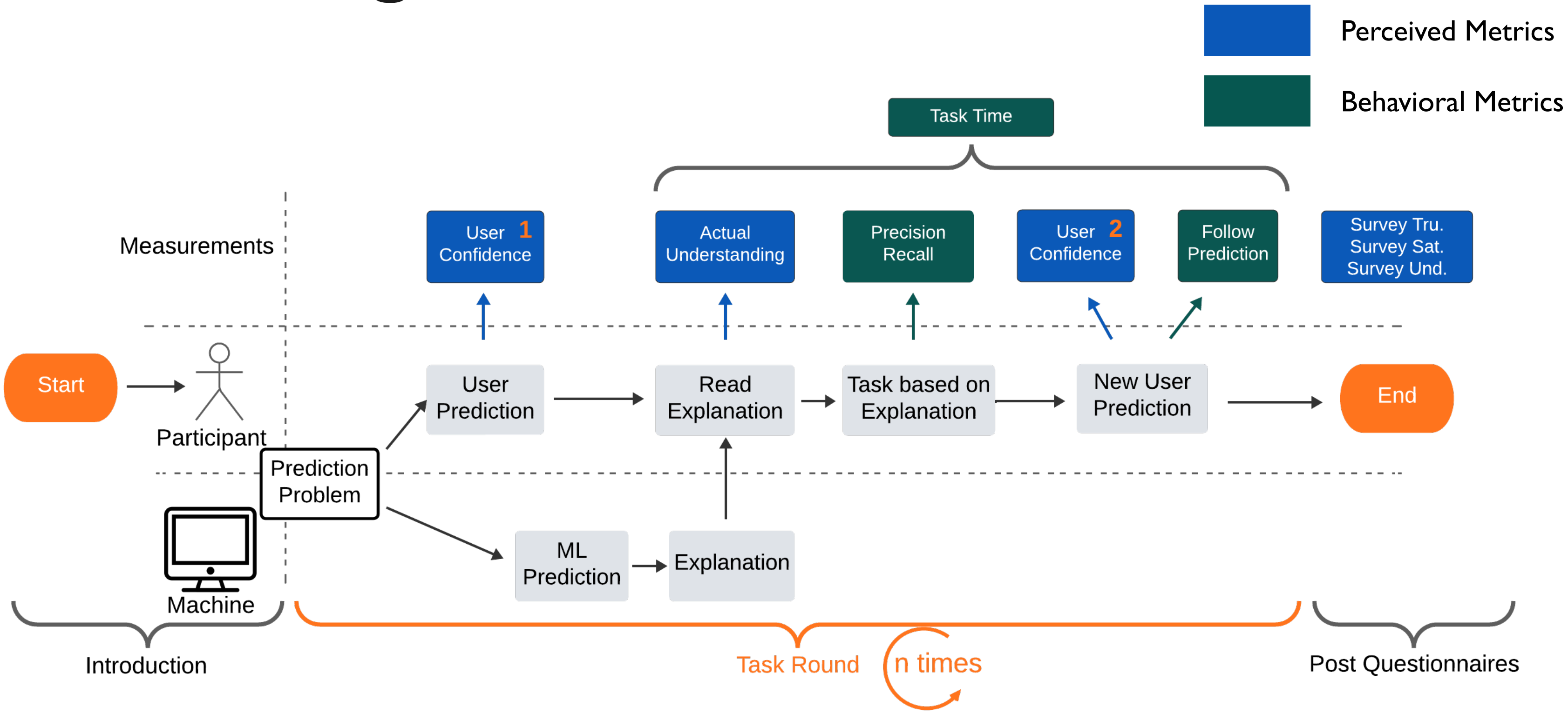
Individual's information as used in the prediction of risk of obesity:

Gender	Female
Age	23
Height	166
Family member has overweight	No
Frequent consumption of high caloric food	No
Frequency of consumption of vegetables	Sometimes
Number of daily meals	More than 3
Consumption of food between meals	Sometimes
Smoke	No
Consumption of water daily	More than 2L
Calories consumption monitoring	Yes
Physical activity frequency per week	2 or 4 days
Time using technology devices daily	0-2 hours
Consumption of alcohol	Sometimes
Transportation used	Public transportation

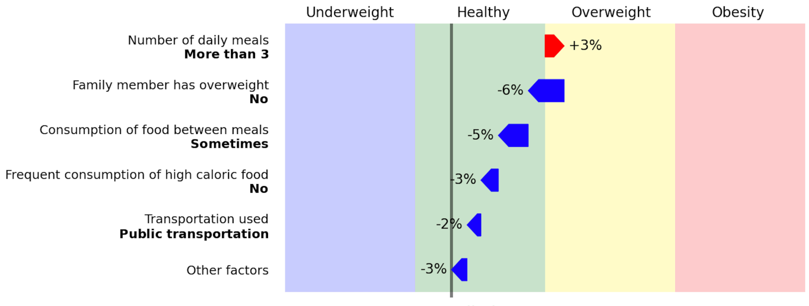
Based only on the above information, the artificial intelligence (AI) tool has predicted **underweight**.
Remember, in the following graph, the red bars indicate an increased chance towards



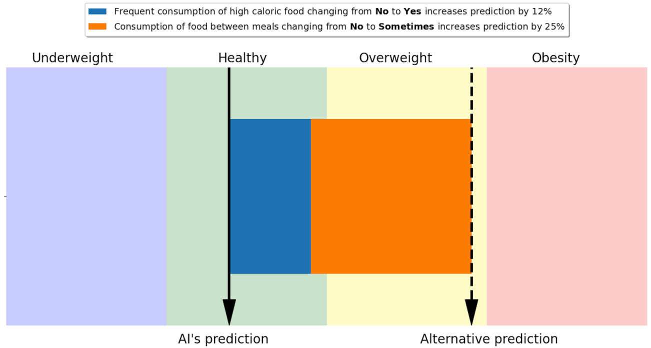
Methodological Framework



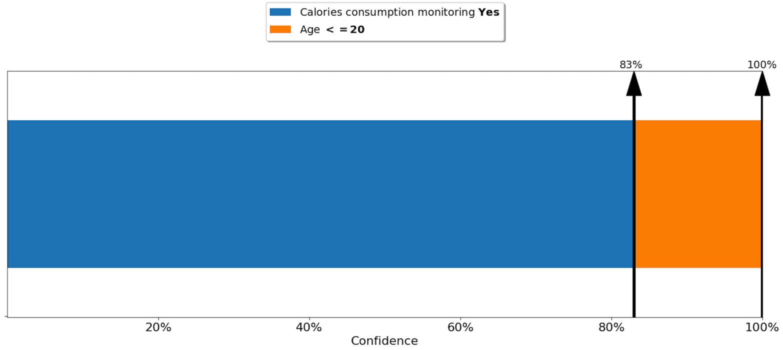
Explanation Representations



Feature Attribution



Example



Rules

Explanation Representations

Based only on the above information, the AI tool has predicted **underweight.**

Remember, the AI associates a score to each response. We obtain a value between 0% and 100% by summing these scores. This value falls into one of four categories: **underweight** (below 25%), **healthy** (between 25% and 50%), **overweight** (between 50% and 75%), and **obesity** (above 75%).

- First, since **no** family member of this individual **suffers** from overweight, the score decreases by 12%.
- Second, since the individual **sometimes** consumes food between meals, the score decreases by 10%.
- Third, **no consuming frequently** high caloric food decreases score by 6%.
- Fourth, using **public transport** decreases the score by 4%.
- Fifth, **monitoring** her calories consumption decreases the score by 2%.

Combining all the **other answers** increases the score by 1% and the final value is 17% implying an **underweight** prediction.

Feature Attribution

Based on the above data, the AI tool has predicted **underweight.**
To turn the AI prediction into an **overweight** prediction, the individual should **have** (at least) a family member **suffering** from overweight and practice physical activity **1 or 2 days** instead of **2 or 4 days** per week.

Example

Based on the above data, the artificial intelligence (AI) tool has predicted **obesity.**

- First, because a family member **suffers** from overweight.
- Second, she is **aged between 23 and 26 years old**.
- Third, she **doesn't practice** physical activity weekly.

All together, it brings an AI's confidence of 95% for this **obesity** prediction

Rules

Explanation Representations

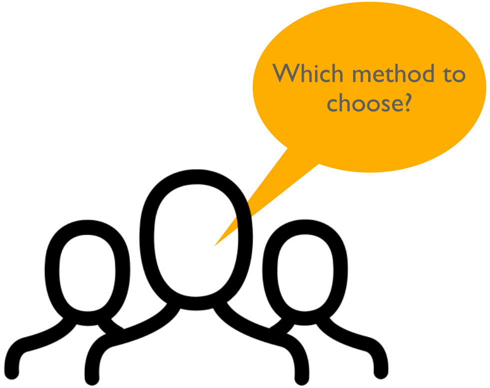
Based only on the above information, the AI tool has predicted **underweight.**

Remember, the AI associates a score to each response. We obtain a value between 0% and 100% by summing these scores. This value falls into one of four categories: **underweight** (below 25%), **healthy** (between 25% and 50%), **overweight** (between 50% and 75%), and **obesity** (above 75%).

- First, since **no** family member of this individual **suffers** from overweight, the score decreases by 12%.
- Second, since the individual **sometimes** consumes food between meals, the score decreases by 10%.
- Third, **no consuming frequently** high caloric food decreases score by 6%.
- Fourth, using **public transport** decreases the score by 4%.
- Fifth, **monitoring** her calories consumption decreases the score by 2%.

Combining all the **other answers** increases the score by 1% and the final value is 17% implying an **underweight** prediction.

Feature Attribution



Based on the above data, the AI tool has predicted **underweight.**
To turn the AI prediction into an **overweight** prediction, the individual should **have** (at least) a family member **suffering** from overweight and practice physical activity **1 or 2 days** instead of **2 or 4 days** per week.

Example

Based on the above data, the artificial intelligence (AI) tool has predicted **obesity.**

- First, because a family member **suffers** from overweight.
- Second, she is **aged between 23 and 26 years old**.
- Third, she **doesn't practice** physical activity weekly.

All together, it brings an AI's confidence of 95% for this **obesity** prediction

Rules

Experimental Design

- 7 groups
  - 2 feature attribution (graphic + text)
  - 2 counterfactual (graphic + text)
  - 2 rule-based (graphic + text)
  - Control group (no explanation)
  - 20 participants per group
- Average completion time ~ 15 min
- Qualtrics
  - Platform to design the 14 surveys (7 per dataset)
- Prolifics
  - Platform to find crowdworkers

Domain	**Healthcare**		**Law**	
Factor	*N*	% sample	*N*	% sample
Gender				
Female	66	47.14	66	47.14
Male	62	44.29	74	52.86
Prefer not to say	1	0.71	0	0.0
Age				
< 20	10	7.14	11	7.86
20 < 30	81	57.86	88	62.86
30 < 40	24	17.14	27	19.29
40 >	14	10.0	14	10.0
Nationality				
Africa	45	32.14	37	26.43
Asia	2	1.43	2	1.43
Australia	0	0.0	1	0.71
Europe	77	55.0	82	58.57
North America	5	3.57	15	10.71
South America	0	0.0	3	2.14

Methodology

- Independent Variable:
  - Explanation Techniques (feature-attribution, rule-based, and counterfactual)
  - Explanation Representation (graphical and text)
  - Demographic Information

Methodology

- Independent Variable:
  - Explanation Techniques (feature-attribution, rule-based, and counterfactual)
  - Explanation Representation (graphical and text)
  - Demographic Information
- Dependent Variable:
  - Users' perception of:
    - Understanding,
    - Trust
  - Users' behavior:
    - Understanding,
    - Trust

Results — Understanding

	Recidivism				Obesity			
	Self Report		Behavioural		Self Report		Behavioural	
	Post Und.	SR Und.	Prec.	Rec.	Post Und.	SR Und.	Prec.	Rec.
Expl. Technique	1.20	0.87	16.24***	1.58	1.35	3.75*	31.42***	6.37***
Represent.	0.36	0.96	0.13	3.00⁻	0.55	0.14	0.05	2.85⁻
Age	0.01	1.07	1.88	0.10	0.06	0.16	6.41*	0.02
Education	0.93	1.63	0.94	0.43	0.34	0.50	0.25	1.31
Gender	1.07	0.54	0.35	0.30	0.03	0.14	0.18	0.36
Surr.:Repr.	0.87	0.28	1.12	0.74	0.16	0.48	0.35	4.99**

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{-}p < 0.1$

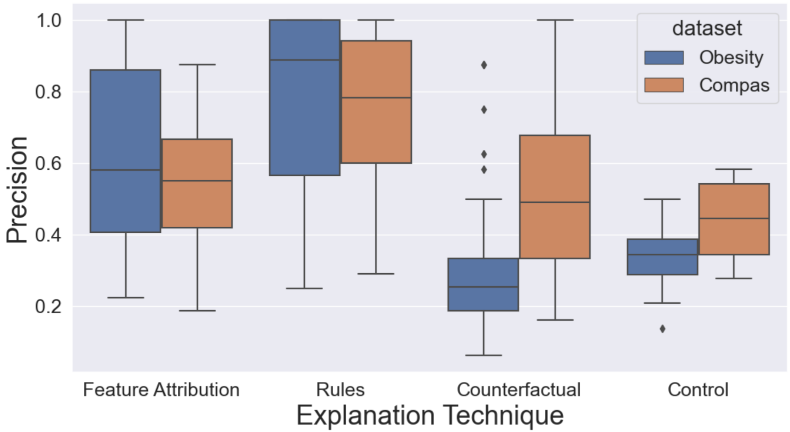
Results — Understanding

- Precision:
  - Alignment between features identified by users and features reported in explanations

	Recidivism				Obesity			
	Self Report		Behavioural		Self Report		Behavioural	
	Post Und.	SR Und.	Prec.	Rec.	Post Und.	SR Und.	Prec.	Rec.
Expl. Technique	1.20	0.87	16.24***	1.58	1.35	3.75*	31.42***	6.37***
Represent.	0.36	0.96	0.13	3.00⁻	0.55	0.14	0.05	2.85⁻
Age	0.01	1.07	1.88	0.10	0.06	0.16	6.41*	0.02
Education	0.93	1.63	0.94	0.43	0.34	0.50	0.25	1.31
Gender	1.07	0.54	0.35	0.30	0.03	0.14	0.18	0.36
Surr.:Repr.	0.87	0.28	1.12	0.74	0.16	0.48	0.35	4.99**

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05, ^{-}p < 0.1$

Does the participants find important features?



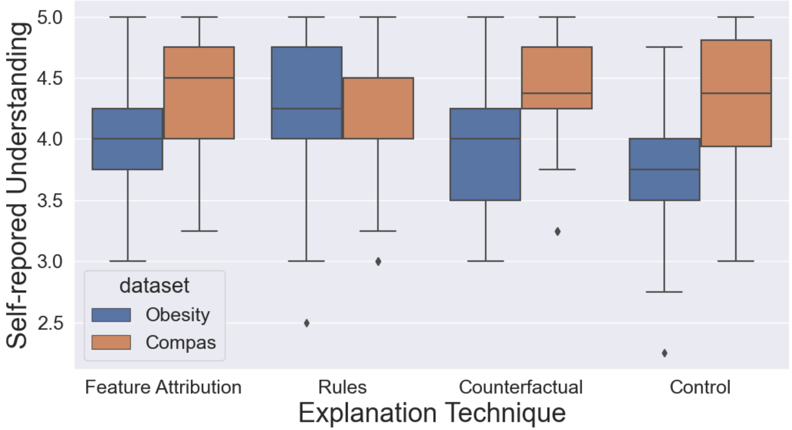
Results — Understanding

- SR Und.:
  - Perceived comprehension of the system's prediction while looking at the explanation

	Recidivism				Obesity			
	Self Report		Behavioural		Self Report		Behavioural	
	Post Und.	SR Und.	Prec.	Rec.	Post Und.	SR Und.	Prec.	Rec.
Expl. Technique	1.20	0.87	16.24***	1.58	1.35	3.75*	31.42***	6.37***
Represent.	0.36	0.96	0.13	3.00⁻	0.55	0.14⁻	0.05	2.85⁻
Age	0.01	1.07	1.88	0.10	0.06	0.16	6.41*	0.02
Education	0.93	1.63	0.94	0.43	0.34	0.50	0.25	1.31
Gender	1.07	0.54	0.35	0.30	0.03	0.14	0.18	0.36
Surr.:Repr.	0.87	0.28	1.12	0.74	0.16	0.48	0.35	4.99**

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05, ^{-}p < 0.1$

Does the participants think they understand?



Results — Trust

	Recidivism			Obesity		
	Self Report		Behav.	Self Report		Behav.
	Post	SR Tru.	Fol.	Post	SR Tru.	Fol.
Expl. Technique	0.03	1.40	0.78	0.42	0.12	0.38
Represent.	0.32	0.04	0.00	0.55	8.22^{**}	0.12
Age	0.18	0.46	2.76^{-}	0.70	0.06	0.00
Education	1.82	0.13	0.34	0.69	2.14^{-}	0.63
Gender	1.35	2.16	0.31	2.32	0.12	1.11
Surr.:Repr.	1.23	0.35	0.75	0.23	0.26	3.55^{*}

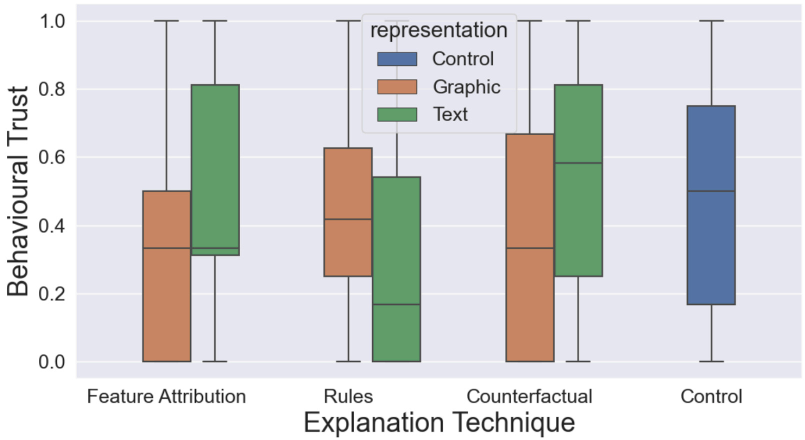
$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05, ^{-}p < 0.1$

Results — Trust

- Behavioural Trust:
  - Proportion of times users modify their initial prediction in favor of the AI's prediction

	Recidivism			Obesity		
	Self Report		Behav.	Self Report		Behav.
	Post	SR Tru.	Fol.	Post	SR Tru.	Fol.
Expl. Technique	0.03	1.40	0.78	0.42	0.12	0.38
Represent.	0.32	0.04	0.00	0.55	8.22**	0.12
Age	0.18	0.46	2.76⁻	0.70	0.06	0.00
Education	1.82	0.13	0.34	0.69	2.14⁻	0.63
Gender	1.35	2.16	0.31	2.32	0.12	1.11
Surr.:Repr.	1.23	0.35	0.75	0.23	0.26	3.55*

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05, ^{-}p < 0.1$



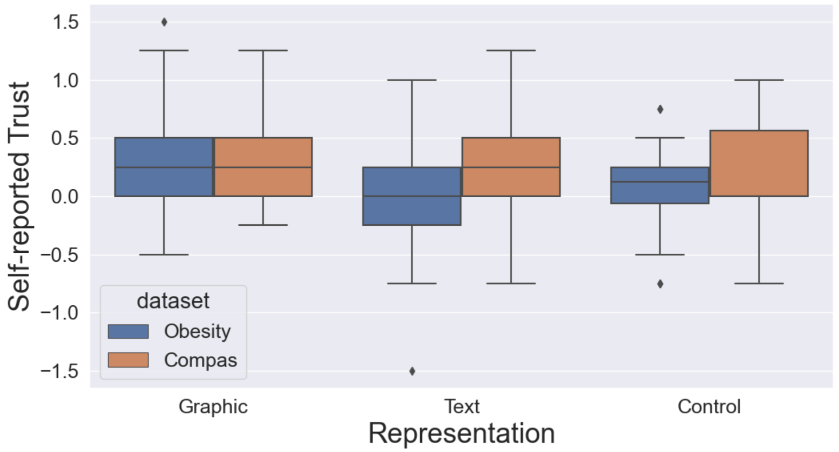
Does the users follow the prediction?

Results — Trust

- Perceived Trust:
  - Changes in self-reported trust before and after accessing AI predictions and explanations

	Recidivism			Obesity		
	Self Report		Behav.	Self Report		Behav.
	Post	SR Tru.	Fol.	Post	SR Tru.	Fol.
Expl. Technique	0.03	1.40	0.78	0.42	0.12	0.38
Represent.	0.32	0.04	0.00	0.55	8.22**	0.12
Age	0.18	0.46	2.76^{-}	0.70	0.06	0.00
Education	1.82	0.13	0.34	0.69	2.14^{-}	0.63
Gender	1.35	2.16	0.31	2.32	0.12	1.11
Surr.:Repr.	1.23	0.35	0.75	0.23	0.26	3.55*

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05, ^{-}p < 0.1$



Does the users feel they can trust the model?

Discussion

Discussion

- Explanations help users:
  - Identify which factors led to a prediction
  - Gain trust in the model's prediction

Discussion

- Explanations help users:
  - Identify which factors led to a prediction
  - Gain trust in the model's prediction

- **Rule-based** explanations are the most effective way
  - It aligns with common educational reasoning principles
  - Clarity of when it is applicable i.e., simplicity

Discussion

- Explanations help users:
  - Identify which factors led to a prediction
  - Gain trust in the model's prediction

- **Rule-based** explanations are the most effective way
  - It aligns with common educational reasoning principles
  - Clarity of when it is applicable i.e., simplicity

- **Counterfactual** explanations yield low users' understanding but high trust
  - Due to the nature of the explanation
  - How we measure the understanding

Discussion

- Explanations help users:
  - Identify which factors led to a prediction
  - Gain trust in the model's prediction

- **Rule-based** explanations are the most effective way
  - It aligns with common educational reasoning principles
  - Clarity of when it is applicable i.e., simplicity

- **Counterfactual** explanations yield low users' understanding but high trust
  - Due to the nature of the explanation
  - How we measure the understanding

- Presentation of explanations shapes users' trust in the model

- Graphical representation increases more user acceptance than textual
  - Cognitive bias related to the apparent complexity of a graphical presentation

Conclusion

Julien Delaunay

Part I — Takeaway Message

- The key to characterize a decision boundary:
  - Conduct a thorough search for counterfactuals
  - Linear separability alone is insufficient to determine linear suitability

Part I — Takeaway Message

- The key to characterize a decision boundary:
  - Conduct a thorough search for counterfactuals
  - Linear separability alone is insufficient to determine linear suitability

- Previous research has focused on:
  - Adapting the explanation to the model

Part I — Takeaway Message

- The key to characterize a decision boundary:
  - Conduct a thorough search for counterfactuals
  - Linear separability alone is insufficient to determine linear suitability
- Previous research has focused on:
  - Adapting the explanation to the model
- We propose to:
  - Adapt the explanation to the specific situation (target, black box)

Part I — Data Perspective in Explainable AI

- There is no one-size-fits-all explanation technique solution:
  - Explanation should be tailored to the data and application

Part I — Data Perspective in Explainable AI

- There is no one-size-fits-all explanation technique solution:
  - Explanation should be tailored to the data and application

- Exploring strategies for impactful explanations:
  - Investigate the influence of the generation strategy on explanation effectiveness
  - Extend the adaptability of our oracle to diverse data types

Part I — Data Perspective in Explainable AI

- There is no one-size-fits-all explanation technique solution:
  - Explanation should be tailored to the data and application

- Exploring strategies for impactful explanations:
  - Investigate the influence of the generation strategy on explanation effectiveness
  - Extend the adaptability of our oracle to diverse data types

- Develop oracles to assess the suitability of various explanation techniques
  - When should we use rule-based explanations?
  - When should we use example-based explanations?

Part I — Data Perspective in Explainable AI

- There is no one-size-fits-all explanation technique solution:
  - Explanation should be tailored to the data and application

- Exploring strategies for impactful explanations:
  - Investigate the influence of the generation strategy on explanation effectiveness
  - Extend the adaptability of our oracle to diverse data types

- Develop oracles to assess the suitability of various explanation techniques
  - When should we use rule-based explanations?
  - When should we use example-based explanations?

- Measure the user-centric impact of adapting the explanation
  - User study combining explanation techniques for a single instance
  - User study with explanation techniques adapted to the target instance

Part II — Takeaway Message

- Factors influencing explanations:
  - Consider the domain specificity when applying explanations (e.g., obesity, recidivism)
  - The chosen technique employed to generate the explanation

Part II — Takeaway Message

- Factors influencing explanations:
  - Consider the domain specificity when applying explanations (e.g., obesity, recidivism)
  - The chosen technique employed to generate the explanation

Part II — Takeaway Message

- Factors influencing explanations:
  - Consider the domain specificity when applying explanations (e.g., obesity, recidivism)
  - The chosen technique employed to generate the explanation

- Optimal representation for explanation depends on the technique:
  - Decision rules are well-suited for textual representation
  - Counterfactuals align effectively with textual representation
  - Feature-attribution find clarity when presented graphically

Part II — User Perspective in Explainable AI

- Investigate if users' preferences are influenced by the data type
  - Explanations' representation differ for text, image, and time series

Part II — User Perspective in Explainable AI

- Investigate if users' preferences are influenced by the data type
  - Explanations' representation differ for text, image, and time series
- Long-term user interaction measurement:
  - How does initial perception of an AI system change over time?
  - Collect user feedback at regular intervals to gauge changes

Part II — User Perspective in Explainable AI

- Investigate if users' preferences are influenced by the data type
  - Explanations' representation differ for text, image, and time series
- Long-term user interaction measurement:
  - How does initial perception of an AI system change over time?
  - Collect user feedback at regular intervals to gauge changes
- Adapting task evaluation to techniques:
  - Utilize "what-if" scenario for counterfactual
  - Identify important features for rule-based
  - Generalize feature-attribution to similar instances

Part II — User Perspective in Explainable AI

- Investigate if users' preferences are influenced by the data type
  - Explanations' representation differ for text, image, and time series
- Long-term user interaction measurement:
  - How does initial perception of an AI system change over time?
  - Collect user feedback at regular intervals to gauge changes
- Adapting task evaluation to techniques:
  - Utilize "what-if" scenario for counterfactual
  - Identify important features for rule-based
  - Generalize feature-attribution to similar instances
- Adapting explanations to users' roles:
  - Assess if computer scientists and domain experts seek similar techniques and representations
  - Adapted explanations based on users' trust in AI and their specific objectives

Envisioning the Future of Explainable AI

- Current explanations may not align with users' requests:
  - Users know "what" is important but lack "why"
  - We should employ large language model to generate explanations

Envisioning the Future of Explainable AI

- Current explanations may not align with users' requests:
  - Users know "what" is important but lack "why"
  - We should employ large language model to generate explanations
- Explore interactive explanations:
  - Identify the sub-population affected by the model prediction
  - Assess the model's performance on a subset of the input data

Envisioning the Future of Explainable AI

- Current explanations may not align with users' requests:
  - Users know "what" is important but lack "why"
  - We should employ large language model to generate explanations
- Explore interactive explanations:
  - Identify the sub-population affected by the model prediction
  - Assess the model's performance on a subset of the input data
- Effectively translate explanation techniques to the user

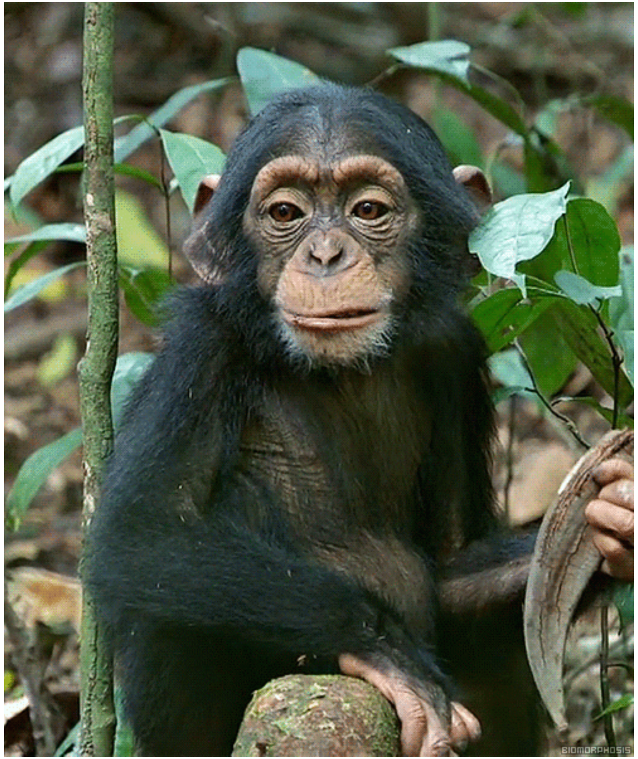
Envisioning the Future of Explainable AI

- Current explanations may not align with users' requests:
  - Users know "what" is important but lack "why"
  - We should employ large language model to generate explanations
- Explore interactive explanations:
  - Identify the sub-population affected by the model prediction
  - Assess the model's performance on a subset of the input data
- Effectively translate explanation techniques to the user
- Leverage the common knowledge embedded in large language models

List of Contributions

- Contribution in the Thesis:
  - How to generate the best explanation from a <u>data</u> perspective?
    - When Should We Use Linear Explanations? [CIKM '22]
    - Improving Anchor-Based Explanations [CIKM '20]
    - Does it make sense to explain a Black Box With a Black Box? [Under Review: NAACL '24]
  - How to generate the best explanation from a <u>user</u> perspective?
    - Methodological Framework [Under Review: CSCW '24]
    - Impact of Explanation Techniques and Representations on Users [Under Review: CSCW '24]
    - Adaptation of AI Explanations to Users' Roles [HCXAI '23]
- Collaboration during the thesis:
  - s-LIME: Reconciling Locality and Fidelity in Linear Explanations [IDA '22]
  - *On Moral Manifestations in Large Language Models* [Moral Agent '23]
  - Global Explanations of NLP Models through Cooperative Generation [BlackboxNLP '23]

Thanks for your attention



Thanks for your attention