Redevabilité pour les systèmes d'aide à la décision : explications interactives

Clément Hénin, Daniel Le Métayer, Inria

Quelques éléments de contexte

- Privatics : équipe-projet Inria consacrée à la protection de la vie privée … et à l'usage éthique du numérique
- Par essence : démarche transversale et interdisciplinaire
- Evolution du domaine :
  - Au-delà de la collecte des données, importance de l'usage
  - Au-delà de la vie privée, autres exigences éthiques (non-discrimination, autonomie, etc.)

Droit, éthique

Exemples de traductions juridiques de ces évolutions :

- France : loi pour une République numérique (2016,2017)
- Europe : RGPD, Directive Police Justice (2016,2018)
- Canada : directive sur la prise de décision automatisée

  (2019,2020)

Sur le plan éthique : multiplication de déclarations, codes, chartes, rapports, comités, etc. notamment sur l'éthique de l'IA

Rapport pour le Parlement Européen

- Understanding algorithmic decision-making : Opportunities and challenges, Claude Castelluccia, Daniel Le Métayer, mars 2019
- Objectif : rendre accessible (même aux parlementaires européens) les enjeux majeurs sans occulter les questions techniques essentielles

http://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2019)624261

Plan

1. Definitions
2. Opportunities and risks related to the use of algorithms
3. Desiderata for algorithms
4. Technical issues and approaches (ADS safety, security, privacy, fairness, explainability)
5. Legal instruments
6. Open questions and remaining challenges
7. Policy options

Notre démarche

- La notion la plus importance et la seule capable de répondre aux enjeux est celle de redevabilité (accountability)
- Cette redevabilité doit être :
    - Envers tous les acteurs concernés (parties prenantes)
    - Interactive et non unilatérale

Notre démarche

L'analyse d'impact algorithmique est un composant essentiel mais non suffisant de la redevabilité et doit être menée de manière rigoureuse de façon à évaluer la légitimité

- Des finalités du SAAD
- Des techniques mise en œuvre par le SAAD
- Des paramètres choisis pour un usage du SAAD

Comment rendre compte ?

- Transparence : montrer, rendre visible
- Explicabilité : donner des raisons (pourquoi le système a-t-il décidé ceci ?), rendre compréhensible
- Justifiabilité : donner des justifications (pourquoi cette décision est-elle bonne ?), des motivations, convaincre, rendre acceptable

NB: le bien-fondé d'une justification doit généralement être établi in fine par des humains

De quoi rendre compte ?

Sur tout le cycle de vie d'un SAAD :

- rendre compte de l'intention : documents de conception du SAAD
- rendre compte de la mise en œuvre du SAAD : code du SAAD
- rendre compte de l'application du SAAD : journaux d'exécution, paires (données d'entrée, résultat)

Nombreux défis

- 3 x 3 possibilités non exclusives mais complémentaires
- Beaucoup de défis
- Nos travaux actuels: explicabilité de l'application des ADS (boîte noire)
- Prochaine étape: justifiabilité

A Generic Framework for Black-box Interactive Explanations

Clément Henin & Daniel Le Métayer

September 11, 2019

Presentation of the black-box setting

F: the black-box function (spam classifier)

$$F : X \rightarrow Y \tag{1}$$

X: Input space (all possible emails)
Y: Output space (boolean spam or non-spam)

D: dataset representing the population (email dataset)

E: scope of the explanation
ex: $E = \{x_e\}$ (local explanation)
$E = D$ (global explanation)

Sampling: creation of emails used to inspect the model

Scope

x_e = "Hello,
I am very happy to be at IJCAI.
Clément Henin, PhD student at Inria"

SAMPLING

Samples

s_1 = "Hello,
I am very happy to be at IJCAI.
Clément Henin, PhD student at Inria"

s_2 = "Hello,
I am very happy to be at IJCAI."

Sampling: creation of emails used to inspect the model

Scope

x_e = "Hello,
I am very happy to be at IJCAI.
Clément Henin, PhD student at Inria"

Population

x_1 = "Hello,
If a machine is expected to be infallible,
it cannot also be intelligent.
Alan Turing"

x_2 = "Hello,
Information is the resolution of uncertainty.
Claude Shannon"

x_3 = "Hello,
The theory has been developed for a hypothetical
nervous system, or machine, called a perceptron.
frank rosenblatt"

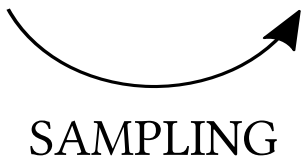
Samples

s_1 = "Hello,
I am very happy to be at IJCAI.
Clément Henin, PhD student at Inria"

s_2 = "Hello,
I am very happy to be at IJCAI.
Alan Turing"

s_3 = "Hello,
I am very happy to be at IJCAI.
Claude Shannon"

s_4 = "Hello,
I am very happy to be at IJCAI.
frank rosenblatt"



Sampling: creation of emails used to inspect the model

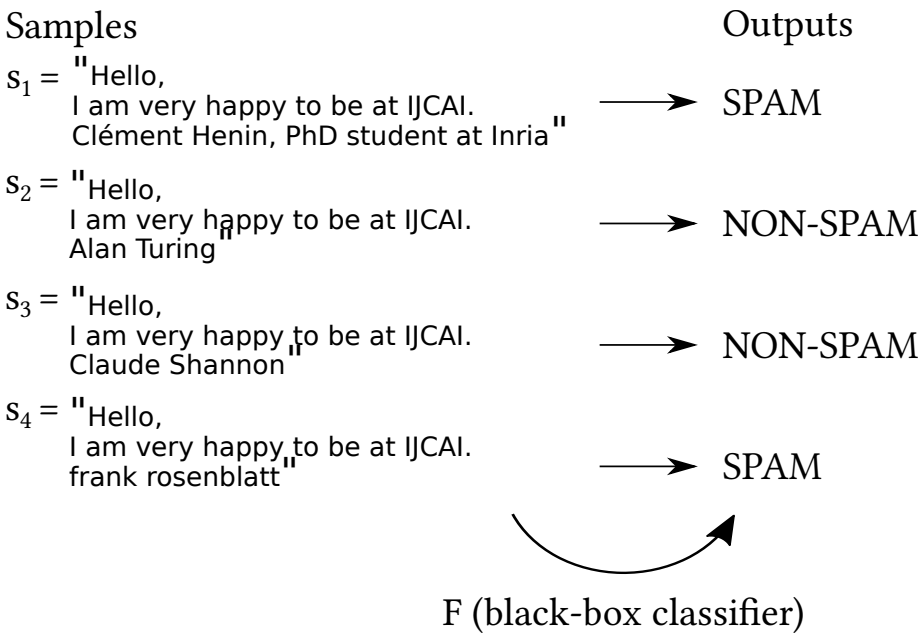
$$S = \{h_\theta(x_e, x_p) \mid (\theta, x_e, x_p) \in \Theta \times E \times D, Z(\theta, x_e, x_p) = 1\} \quad (2)$$

with

$$h_\theta : E \times D \to X \quad (3)$$

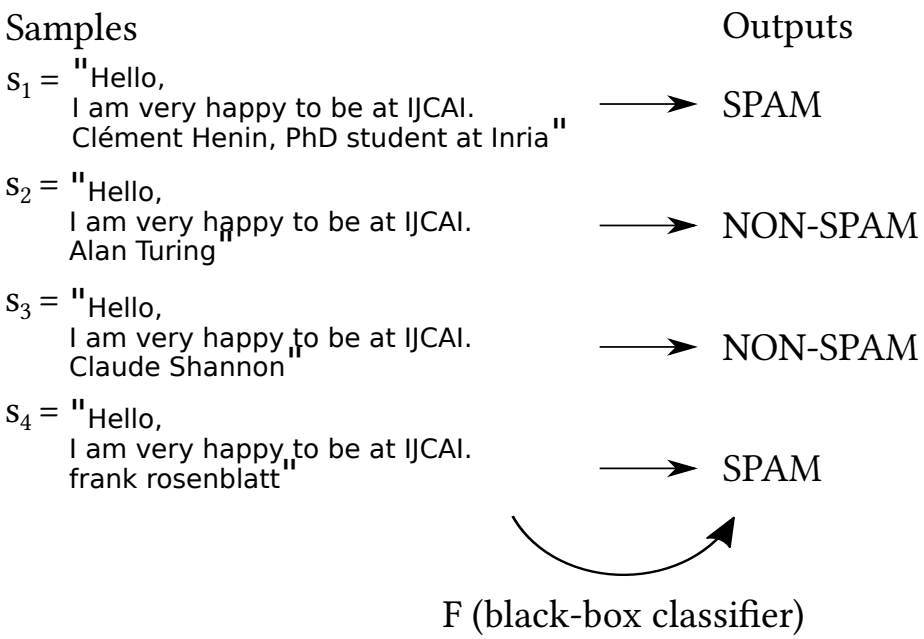
Name	Description	Example
X	Input space of F	Space of emails
E	Scope of the explanation	Email x_e
Θ	Parameters of the sampling	Part of the email
D	Dataset describing the overall population	Training set of F

Generation: analyse samples to create explanations



Generation process

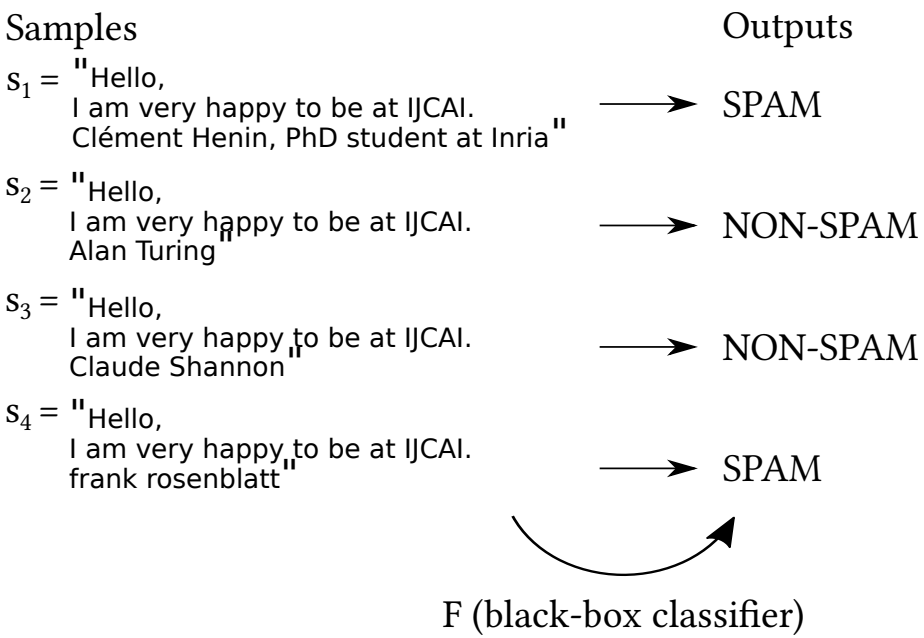
Generation: analyse samples to create explanations



Generation process

- Rule based model (RBM)

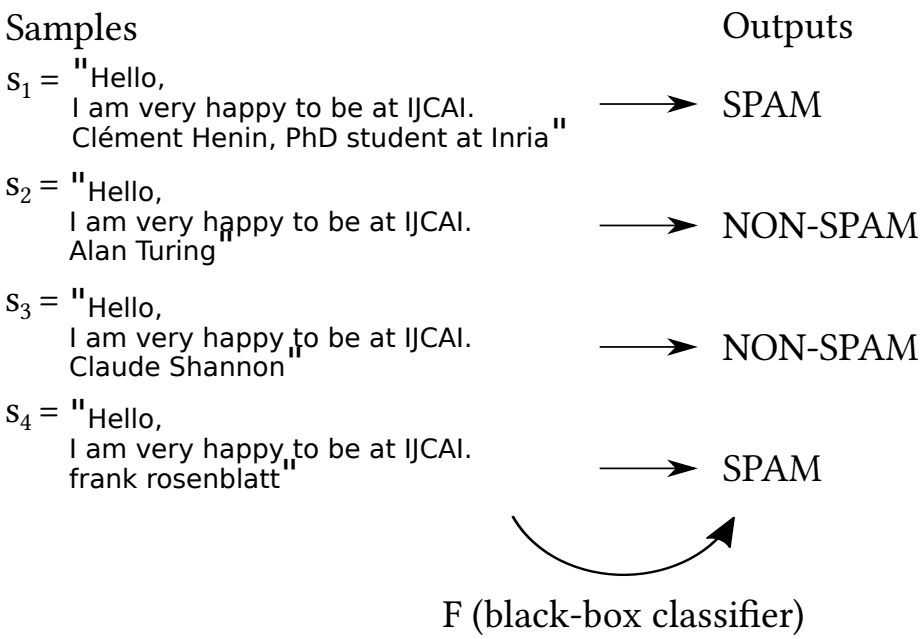
Generation: analyse samples to create explanations



Generation process

- Rule based model (RBM)
- Criteria 1: complexity of the RBM
  Number of rules
- Criteria 2: fidelity of the RBM
  # samples s.t. RBM(s) = F(s)

Generation: analyse samples to create explanations



Generation process

- Rule based model (RBM)
- Criteria 1: complexity of the RBM
  Number of rules
- Criteria 2: fidelity of the RBM
  # samples s.t. RBM(s) = F(s)

RBM 1

If lenght(signature) > 20:
    SPAM
Else:
    NON-SPAM

# rules = 1
precision = 75%

RBM 2

If lenght(signature) > 20 OR no capital letters in signature:
    SPAM
Else:
    NON-SPAM

# rules = 2
precision = 100%

Generation: analyse samples to create explanations

$$f_w : X \to Y \tag{4}$$

$$w^* = \underset{w}{\text{argmin}} \quad \sum_i \lambda_i c_i(w, S) \tag{5}$$

$$\text{subject to} \quad o_i(w, S)$$

Name	Description	Example
X	Input space of F	Space of emails
Y	Output space of F	$[0, 1]$
S	Samples (product of the sampling step)	Emails with changed signature
f_w	Proxy model	Rule-based model

Name	Sampling			Generation		Output Type	Steps[b]	I/O[c]
	E	D	Type[a]	Model	Objectives			
Trepan	Pop.	Pop.	P&R	Decision tree	Complexity, Fidelity	Decision tree	S G_M	I
BETA	Pop.	Pop.	S&D	Rule-based model	Interpretability, Fidelity Unambiguity	Rule-based global model	S G_M	I
GoldenEye	Pop.	Pop.	P&R	Permutation	Fidelity, Interaction size	Important features interactions	S G_M G_E	I
VIN	Pop.	Pop.	P&D	Permutation	ANOVA projection	Important features interactions	S G_M G_E	I
PDP ICE	Pop.	Pop.	P&D	NA	NA	Dependence plot for one feature	S G_E	O
QII	Indiv. or or group	Pop.	P&D	NA	NA	Var. importance	S G_E	O
Anchors	$\{x_e\}$	Pop.	P&R	Rule-based model	Fidelity, Complexity, Generality	Rule-based local model	S G_M	I
LIME	$\{x_e\}$	$\emptyset$	P&R	Linear model	Fidelity, Complexity	Var. importance	S G_M G_E	O
Shapley	$\{x_e\}$	Pop.	P&D	NA	NA	Var. importance	S G_E	O
LEMNA	$\{x_e\}$	$\emptyset$	P&R	Mixture of linear models	Fidelity, Complexity	Var. importance	S G_M G_E	O
Local Gradient	$\{x_e\}$	Pop.	NA	Parzen window	Fidelity	Directions of highest slope	G_M G_E	O
Counter-factuals	$\{x_e\}$	Pop.	NA	Small deviation	Target output, Distance input	Example-based	S G_M G_E	I

Table: Comparative table of the different black-box explanation methods. The columns correspond to the parameters of our framework with the following notation (a) S: Selection, P: Perturbation, D: Deterministic, R: Random ; (b) S: Sampling, G_M: model generation phase, G_E: explanation generation phase; (c) I: Indirect, O: One-shot.

Interactive Black-box EXplanations (IBEX)

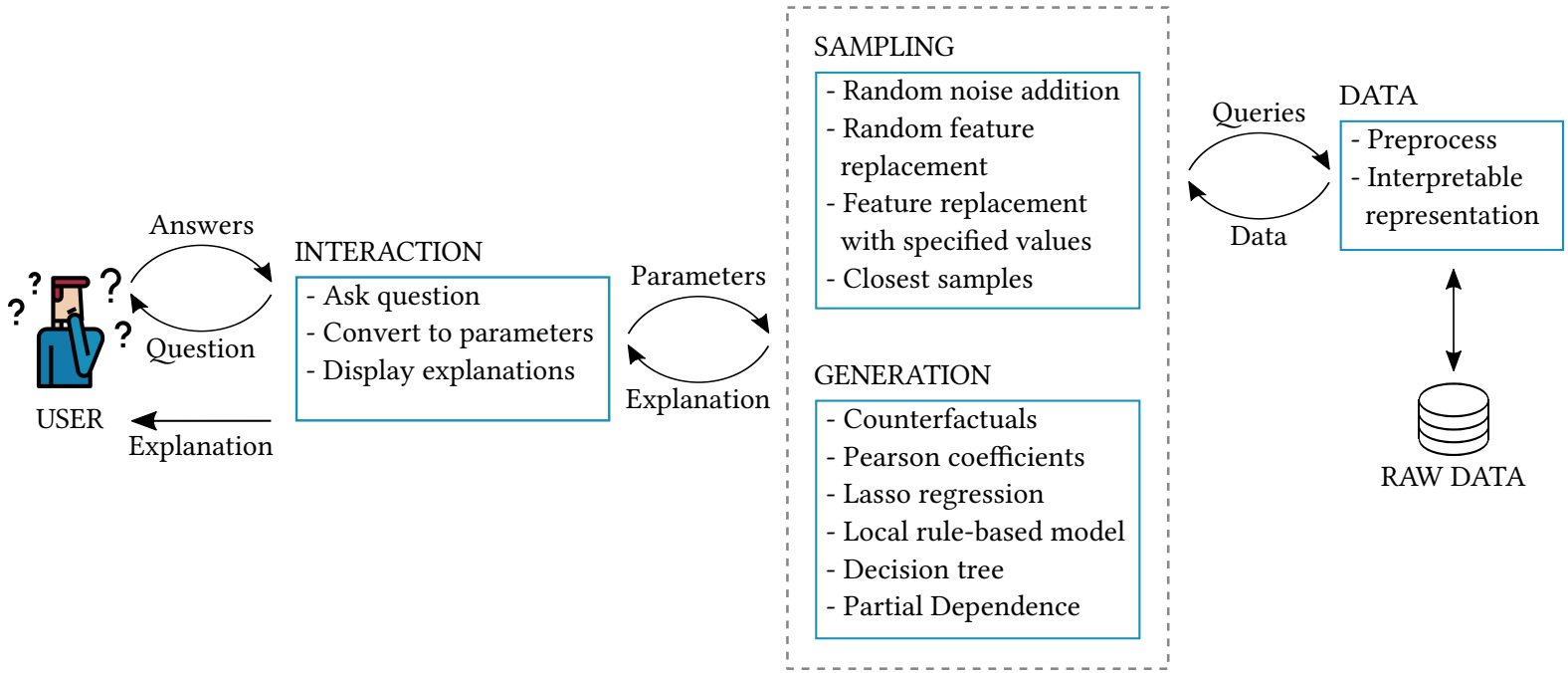


Figure: Architecture of the IBEX system

IBEX: Interaction

Q1 What form of explanations do you prefer? Examples, rules or plots?

Q2 Do you want a local explanation (concerning a specific input) or a global explanation (about the whole model)?

Q3 Why do you ask for explanations? To understand the model (emphasis on high generality and low complexity) or to evaluate it or challenge its results (emphasis on high precision) ?

Q4 Do you want to impose constraints on certain features (choose values of some features that should be considered as fixed)?

Q5 Do you prefer that the system considers any possible input value (emphasis on high precision) or focuses on realistic values (emphasis on understandability) ?

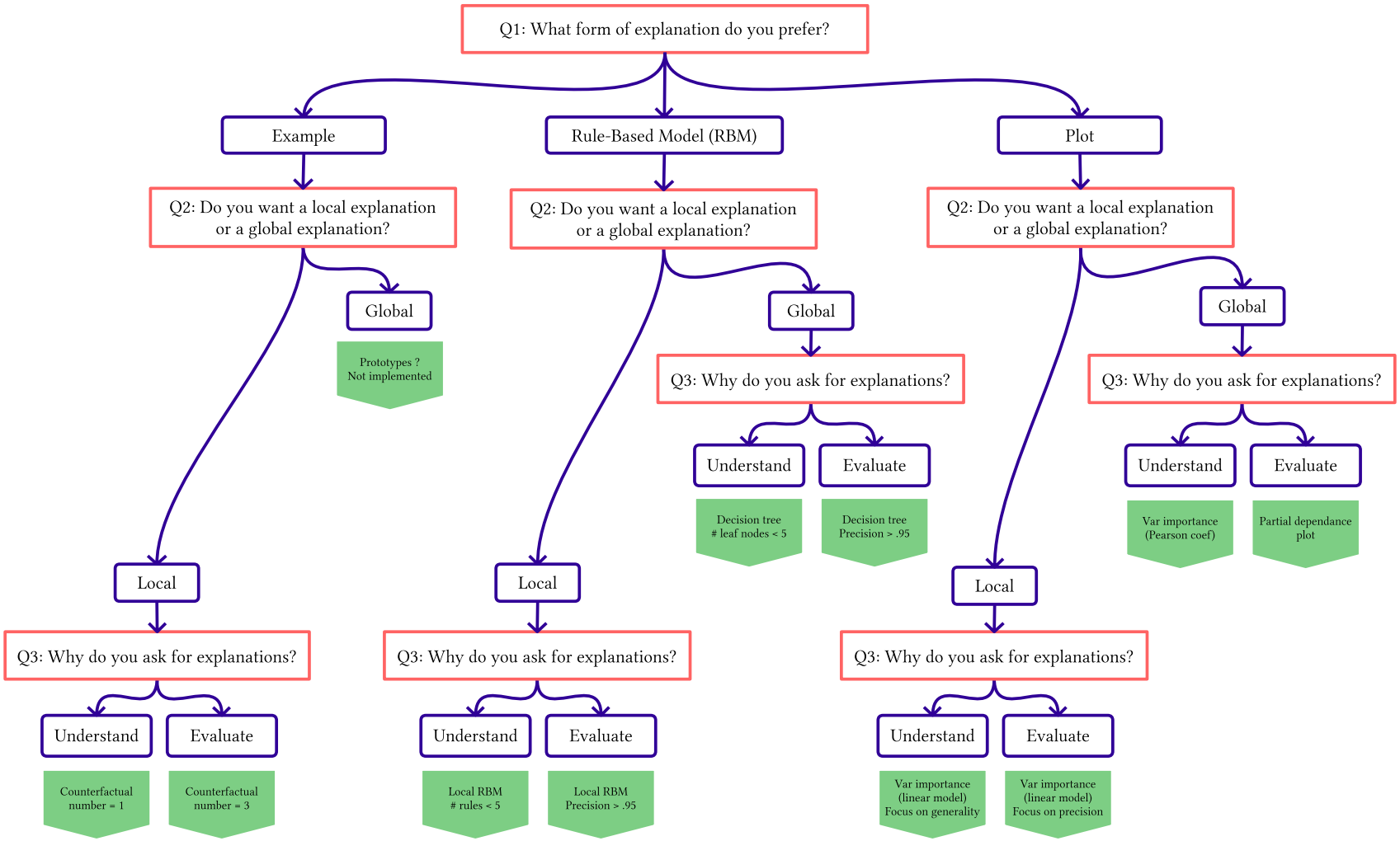


Figure: Flowchart interaction Q1, Q2, Q3

IBEX: interaction task

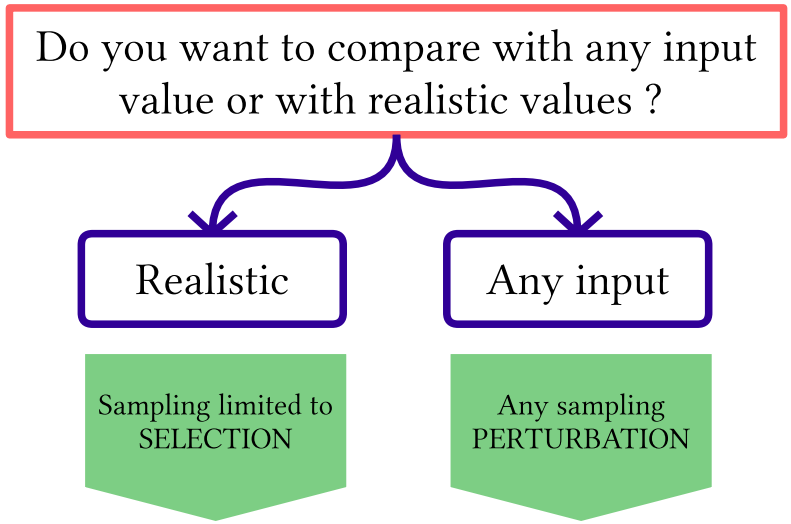


Figure: Flowchart interaction Q5

IBEX: examples from POC

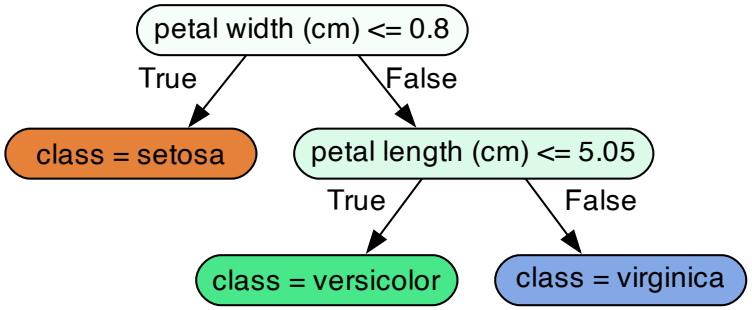


Figure: Simple explanation for a new user

IBEX: examples from POC

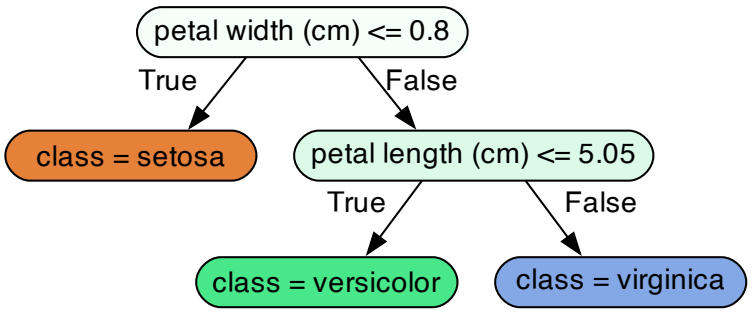


Figure: Simple explanation for a new user

Sepal		Petal	
length (cm)	width (cm)	length (cm)	width (cm)
6.3	2.7	4.9	1.8

Table: Values for the features for an unusual Iris Virginica

IBEX: examples from POC

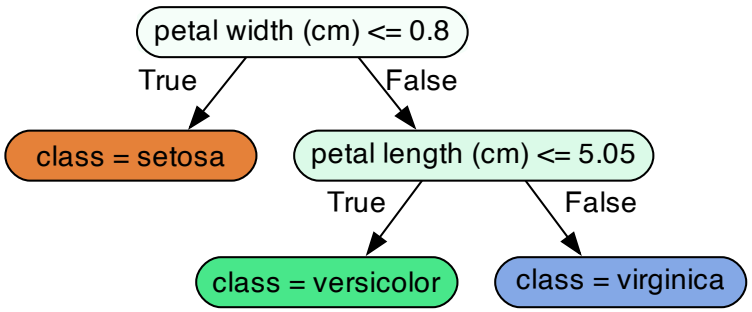


Figure: Simple explanation for a new user

Sepal		Petal	
length (cm)	width (cm)	length (cm)	width (cm)
6.3	2.7	4.9	1.8

Table: Values for the features for an unusual Iris Virginica

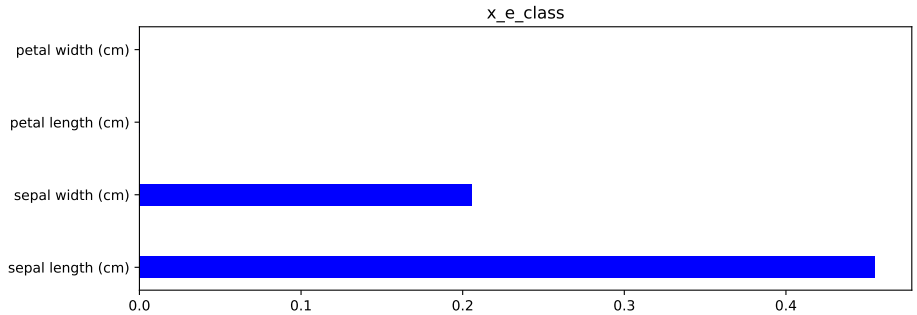


Figure: Explanation for a surprising classification

IBEX: example from POC

Selected sample:

sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
6.0	2.7	5.1	1.6

Predicted as : Versicolor

Local model :
If:
petal length (cm) <= 5.1499998569488525
and petal length (cm) > 2.449999988079071
then the output of the model equals the output of the selected samples with high precision among samples (>.95).

Figure: Explanation for technical user

Conclusion and future work

Conclusion

- Improving the POC
- Assessment of explanation
- User study

Thank you for your attention