Carleton UNIVERSITY

Millennium Institute Foundational Research on Data

Explanations in AI and Explainable AI:

(Some of) The Classic and The New

Prof. Leopoldo Bertossi

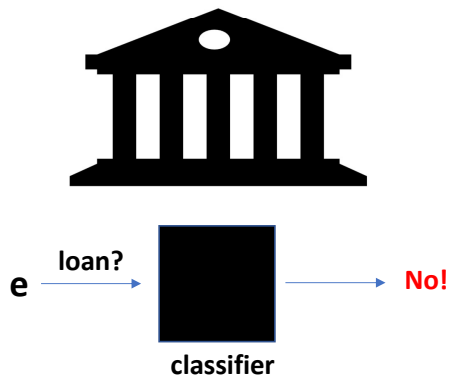
Algeria, April 2026

www.scs.carleton.ca/∼bertossi

Explanations in Machine Learning

- Bank client $\mathbf{e} = \langle \text{john}, 18, \text{plumber}, 70\text{K}, \text{harlem}, \ldots \rangle$

  As an entity represented as a record of values for features Name, Age, Activity, Income, ...
- $\mathbf{e}$ requests a loan from a bank, which uses a classifier



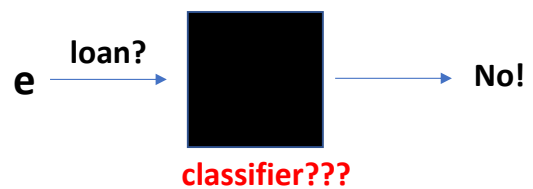
- The client asks *Why?*
- What kind of *explanation?*
  How?
  From what?

Explanations (in AI)

- Users and those affected by results from AI systems, the stakeholders, request explanations

  Assessments (e.g. a credit score), classifications (good/bad client), decisions (approve/reject loan), etc.
- A whole new area of AI has emerged: *Explainable AI* (XAI)

  A whole discipline has emerged: *Ethical AI*
- It touches Law, Sociology, Philosophy, ...
- Motivated by the need for more *transparent, trustable, fair, unbiased, ...* and *interpretable* AI systems



It may really be a "black box"!

- New legislation forces AI systems affecting users to provide explanations and guarantee all the above

- Search for explanations belongs to the nature of human beings
- The quest has been around since the inception of humans
- Ancient Greeks already concerned with *causes* (and effects)
- Studied as such by Philosophers, Logicians, Physicists, ...
- Are explanations a new subject in AI?
- Yes and No
- Explanations have been studied in AI for some decades by now, and in related disciplines, such as Logic, Statistics

  Some forms of explanations are new in AI

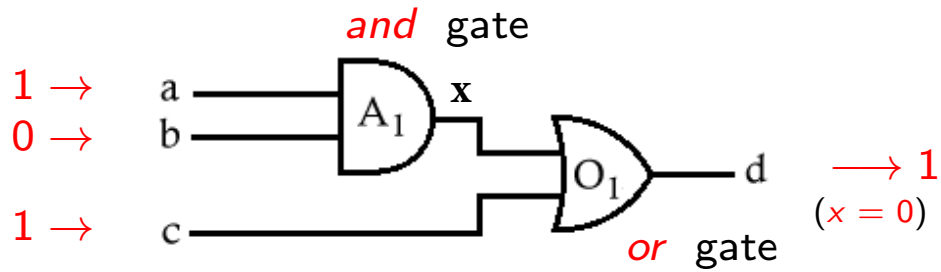
  Others have roots in already existing ones

Model-Based Diagnosis

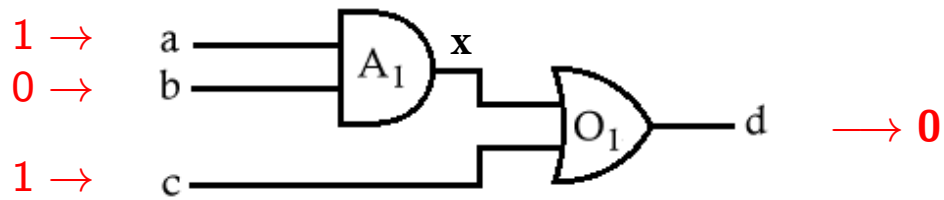
- MBD has been an area of AI for some time
- It is about doing a *diagnosis* of a system (exhibiting some unexpected behavior) using a model of the system (and possibly a bit more)

Example: A very simple Boolean circuit (a classifier?)

It should be:

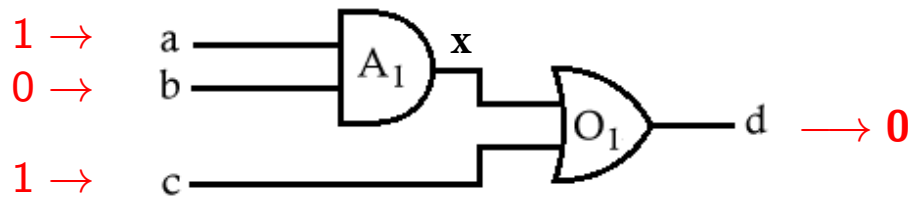


However:



Why? What's wrong? A diagnosis?

- What is a diagnosis? We need a characterization ...



- A logical model of the *ideal circuit*:

$$\{(x \longleftrightarrow (a \wedge b)),\ (d \longleftrightarrow (x \vee c))\}$$

- The observation Obs: $a \wedge \neg b \wedge c \wedge \neg d$
- What can be get from the combination? Logically?

  Since the combination is inconsistent, everything!
  Trivial, irrelevant, useless conclusion ...
- Need flexible model that allows failures: (a "weak model of failure", specifying things under normality)

$$\mathcal{M} = \{\neg AbA \longrightarrow (x \leftrightarrow (a \wedge b)),\ \neg AbO \longrightarrow (d \leftrightarrow (x \vee c))\}$$

"when A is not abnormal, it works as an and *gate"*, etc.

Now gates could be abnormal (faulty)

- Now, $Obs \cup \mathcal{M}$ is consistent, but, as before:

$$Obs \cup \mathcal{M} \cup \{\neg AbA, \neg AbO\} \text{ is inconsistent} \qquad (*)$$

- So, something has to be abnormal ...
- $D = \{abO\}$ is a diagnosis, because making gate O abnormal restores consistency

$$Obs \cup \mathcal{M} \cup \{\neg AbA, AbO\} \text{ is consistent} \qquad (**)$$

Abnormality of gate O is an explanation for the malfuction of the circuit

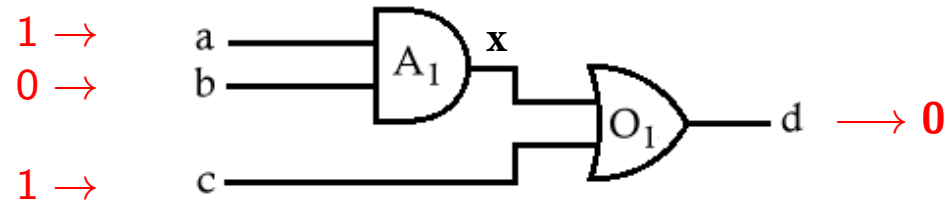
- $D' = \{abO, abA\}$ is a diagnosis, because making every gate abnormal restores consistency

$$Obs \cup \mathcal{M} \cup \{AbA, AbO\} \text{ is consistent} \qquad (***)$$

- D is "better" than D': fewer assumptions, narrower, more focused and informative
- This is *Consistency-Based Diagnosis* (CBD, Ray Reiter, 1987)
- Can we assign scores to diagnoses? (coming)

Actual Causality

Example: (cont.) Can be cast as a causality detection problem



Model: $\mathcal{M} = \{\neg AbA \longrightarrow (x \leftrightarrow (a \wedge b)),\ \neg AbO \longrightarrow (d \leftrightarrow (x \vee c))\}$

- Meta-Observation:

  $\{a, \neg b, c, \neg d\} \cup \mathcal{M} \cup \{\neg AbA, \neg AbO\}$ is inconsistent

- *Counterfactuals*: hypothetical changes of (interventions on) non-abnormalities into abnormalities, to see if this changes

- $\{a, \neg b, c, \neg d\} \cup \mathcal{M} \cup \{\neg AbA, \underbrace{AbO}_{\text{counterfactual change}}\}$ is consistent

Meta-Observation changed: abO is *counterfactual cause* for observation

- However: $\{a, \neg b, c, \neg d\} \cup \mathcal{M} \cup \{\underbrace{AbA}_{\text{changed}}, \neg AbO\}$ is inconsistent

  Meta-Observation not changed: AbA not counterfactual cause

  Extra, contingent counterfactual changes may be necessary:

$$\{a, \neg b, c\} \cup \mathcal{M} \cup \{\underbrace{AbA}_{\text{changed}}, \quad \underbrace{AbO}_{\text{contingent change}} \quad\} \text{ is consistent}$$

- AbA is neither counterfactual nor actual cause

  By definition, because contingency AbO is already a counterfactual cause (J. Halpern & J. Pearl, 2001)

- Numerical Attribution Score quantifying strength of a cause?

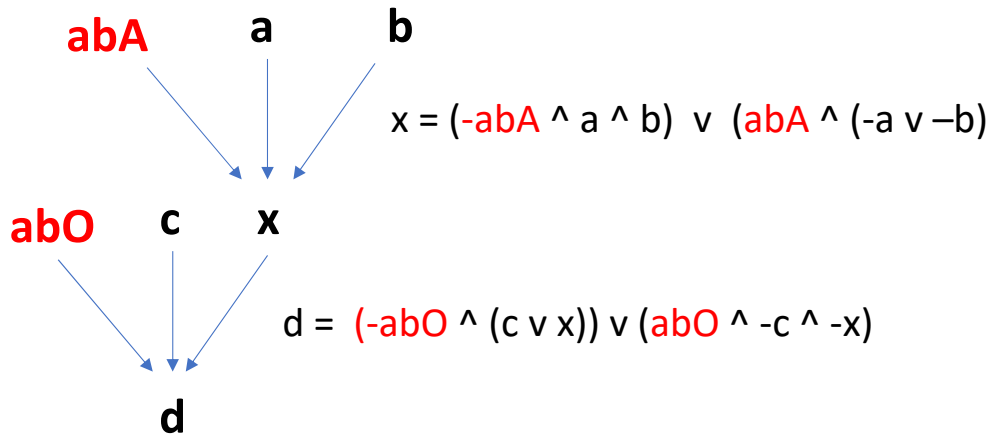
- Causal Responsibility: (Chokler & Halpern, 2004)

$$Resp(abO) := \frac{1}{1+\text{min. cardinality of CS}} = \frac{1}{1+0} = 1 \quad \text{(max. responsibility)}$$

$$Resp(abA) := 0 \quad \text{(no responsibility)}$$

- We will concentrate on attribution scores as explanations

- There is a connection with causality based on Causal Networks and *Structural Models* used in AI
- Actual causality can be cast in those terms



- Here *abA*, *abO* are endogenous variables

  They can be subject to counterfactual intereventions

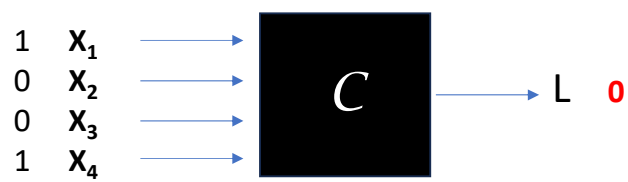
  The others are exogenous variables
- The CN above involves some *structural equations*

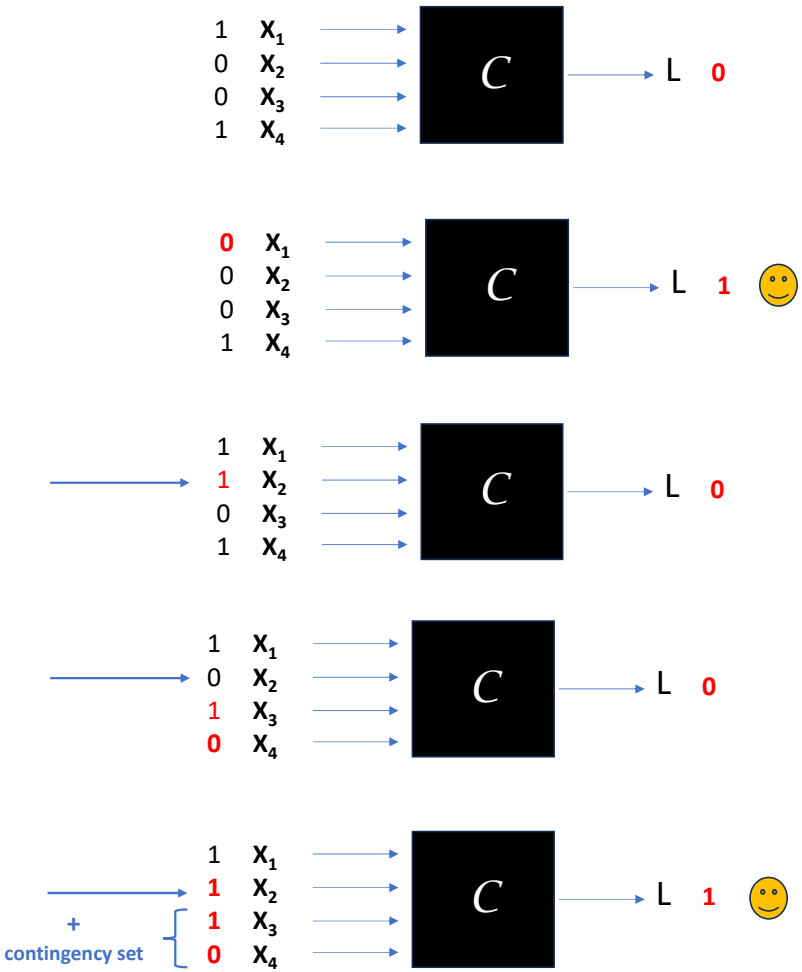
Resp and Explanations for Classification

- Feature values get *responsibility scores*, quantifying *how much* they contribute to the label
- These methods can be applied without necessarily knowing "the internals" of the classifier

  The latter is treated (or is) a "black box" system

  Only input/output relation is needed
- Given an entity **e** that gets a (binary) label, say 0
- Do changes of (interventions on) feature values change the label from 0 to 1?
- The idea: Could be initial circuit (malfunctioning or not), or any classifiers for binary entities





Counterfactual cause for $L = 0$:

- $\mathbf{x_1} = 1$
- $Resp(\mathbf{x_1}) := 1$
  maximum responsibility

concentrate on $\mathbf{x_2}$: not counterfactual cause

changes on $\mathbf{x_3}, \mathbf{x_4}$ do not change label

change on $\mathbf{x_2}$ *accompanied by* changes on $\mathbf{x_3}, \mathbf{x_4}$ does change label!

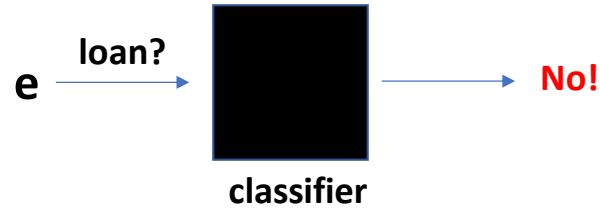
$\mathbf{x_2}$ is actual cause for $L = 0$

$\Gamma = \{\mathbf{x_3}, \mathbf{x_4}\}$ is contingency set for $\mathbf{x_2}$

If Γ is minimum-size contingency set for $\mathbf{x_2}$: $Resp(\mathbf{x_2}) := \frac{1}{1+|\Gamma|} = \frac{1}{3}$

We call $\langle 1, 1, 1, 0 \rangle$ a counterfactual (version) of original entity

- Back to our banking example ...



$$\mathbf{e} = \langle \text{john}, 18, \text{plumber}, 70\text{K}, \text{harlem}, \ldots \rangle \quad \text{No}$$

- Counterfactual versions:

$$\mathbf{e}' = \langle \text{john}, 25, \text{plumber}, 70\text{K}, \text{harlem}, \ldots \rangle \quad \text{Yes}$$

Value for feature *Age* is counterfactual cause with explanatory responsibility $\mathit{Resp}(\mathbf{e}, \mathit{Age}) = 1$

$$\mathbf{e}'' = \langle \text{john}, 18, \text{plumber}, 80\text{K}, \text{brooklyn}, \ldots \rangle \quad \text{Yes}$$

Value for *Income* is actual cause with $\mathit{Resp}(\mathbf{e}, \mathit{Income}) = \frac{1}{2}$

This one needs additional (contingent) changes ...

- Really?

- For binary features the previous definition of responsibility works fine
- In the case of the classifier, possibly many new values for a feature do not change the label, and few of them do
- Then, the original value is not a great explanation
- Responsibility score has to be generalized (B. et al., Deem@SIGMOD20)
- Better consider contingent features and values for them, and average labels!
- We are considering binary classifiers, with labels 1 or 0

  Assume label 1 is the one we want to explain
- *Resp* is a "local" explanation score: for a feature value in a particular entity

Generalized Responsibility -1

- $\mathbf{e}$ classified entity, $L(\mathbf{e}) = 1$, $F^\star \in \mathcal{F}$ (set of features)
- "Local" *Resp*-score: for fixed contingent assignment $\Gamma := \bar{w}$
  $\Gamma \subseteq \mathcal{F} \setminus \{F^\star\}$ (potential contingent set of features)
- $\mathbf{e}' := \mathbf{e}[\Gamma := \bar{w}]$ (potential contingent values), with $L(\mathbf{e}') = L(\mathbf{e})$

$$Resp(\mathbf{e}, F^\star, \Gamma, \bar{w}) := \frac{L(\mathbf{e}) - \mathbb{E}[L(\mathbf{e}'') \mid \mathbf{e}''_{\mathcal{F}\setminus\{F^\star\}} = \mathbf{e}'_{\mathcal{F}\setminus\{F^\star\}}]}{1+|\Gamma|} \quad (*)$$

  - $\mathbf{e}'' := \mathbf{e}[\Gamma := \bar{w}, F^\star := v]$, with $v \in dom(F^\star)$
  - $\mathbf{e}_S$ is projection of $\mathbf{e}$ on $S \subseteq \mathcal{F}$
  - When $(*) > 0$, $F^\star(\mathbf{e})$ is *actual causal explanation* for $L(\mathbf{e}) = 1$ with contingency $\langle \Gamma, \mathbf{e}_\Gamma \rangle$

- Global score: $Resp(\mathbf{e}, F^\star) := \max_{\langle \Gamma, \bar{w} \rangle,\ |\Gamma| \text{ min.},\ (*) > 0} Resp(\mathbf{e}, F^\star, \Gamma, \bar{w})$

- *Resp* requires multiple passes through the classifier ...
- *Resp* requires (assumes) a probability distribution on the entity population $\mathcal{E}$

  Several probability distributions can be used

  Quite a relevant issue ... (B. et al., Deem@SIGMOD20)
- In our experiments, *Resp* score computed with empirical product distribution
- We are usually interested in max-*Resp* feature values

  Associated to minimum (cardinality) contingency sets

  Their computation is in some cases provably intractable

- *Resp* does not require the internals of a classifier

  Can we compute it faster when we have access to the internals?

- Also relevant: doing something with a high-responsibility explanation

  Some counterfactuals may not "make sense" or be "useful"

- In the example, changing the age (waiting for 7 years) may not be feasible

  But maybe changing job and neighborhood could be done ...

- We may want an *actionable* explanation

  We may want the explanation to be a *resource*

- Specification of- and reasoning with counterfactuals can be useful (coming back)

Shapley Values: *Shap*

- Based on the general Shapley value of coalition game theory
- For each application of Shapley one needs an appropriate game function that maps (sub)sets of players to real numbers
- Our case: Set of players $\mathcal{F}$ contain features, but relative to $\mathbf{e}$
- Game function: For $S \subseteq \mathcal{F}$, and $\mathbf{e}_S$ the projection of $\mathbf{e}$ on S

$$\mathcal{G}_{\mathbf{e}}(S) := \mathbb{E}(L(\mathbf{e}') \mid \mathbf{e}' \in \mathcal{E} \;\&\; \mathbf{e}'_S = \mathbf{e}_S)$$

- For a feature $F^\star \in \mathcal{F}$, compute: $Shap(\mathcal{F}, \mathcal{G}_{\mathbf{e}}, F^\star)$

$$\sum_{S \subseteq \mathcal{F} \setminus \{F^\star\}} \frac{|S|!(|\mathcal{F}|-|S|-1)!}{|\mathcal{F}|!} [\underbrace{\mathbb{E}(L(\mathbf{e}'|\mathbf{e}'_{S\cup\{F^\star\}} = \mathbf{e}_{S\cup\{F^\star\}})}_{\mathcal{G}_{\mathbf{e}}(S\cup\{F^\star\})} - \underbrace{\mathbb{E}(L(\mathbf{e}')|\mathbf{e}'_S = \mathbf{e}_S)}_{\mathcal{G}_{\mathbf{e}}(S)}]$$

- *Shap* score has become popular (Lee & Lundberg, 2017)
- Assumes a probability distribution on entity population
- Requires multiple passes through classifier ...

Tractability of SHAP on Open-Box Models

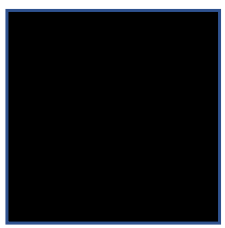
- SHAP has become the most popular explanation score in ML

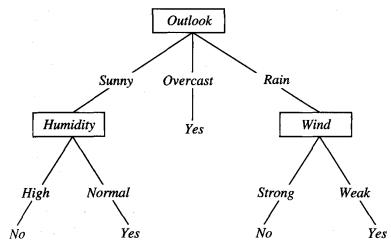
  Shapley value, in particular SHAP, is provably intractable

  Can we do better at computing SHAP with an open classifier?

  Having access to its structure and components?

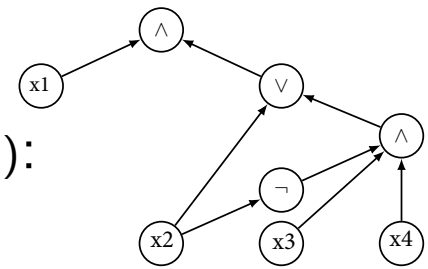
SHAP is tractable for
Deterministic and Decomposable
Boolean Circuits (dDBCs)
(uniform or product distribution)





We get tractability of SHAP for a vast
collection of classifiers (via Knowledge Compilation):

Decision Trees (even with non-binary)
Ordered Binary Decision Diagrams (OBDDs)
Sentential Decision Diagrams (SDDs)



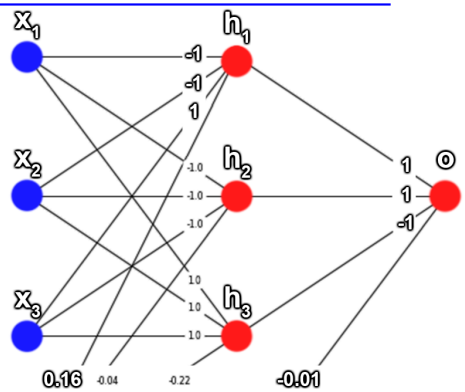
(binary dDBC classifier)

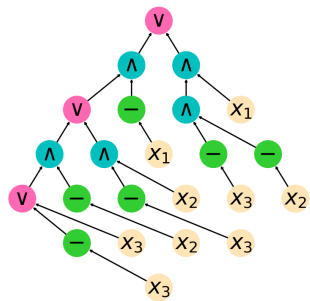
(AAAI'21, JMLR 23)

SHAP on Binary Neural Networks

- NNs considered as black-boxes
- We experimented with SHAP computation on a BNN via compilation into a dDBC
  1. BNN $\mapsto$ CNF (parsimonious and optimized)
  2. CNF $\mapsto$ SDD (non-polynomial, but FPT)
  3. SDD $\mapsto$ dDBC (straightforward)

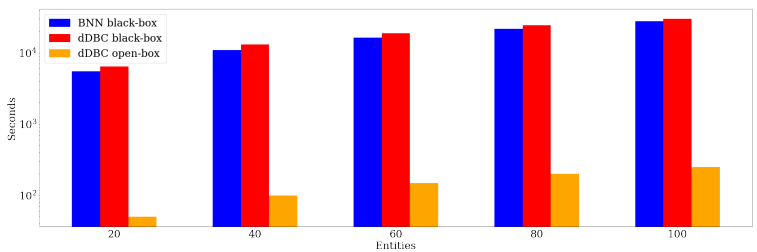
  Still worth this one-time computation
  (target dDBC may be used multiple times)





- Experiments: BNN with 14 gates, dDBC with 18,670 nodes

  Compared SHAP computation for:
  black-box BNN, open-box dDBC,
  and black-box dDBC



(logarithmic scale)

- All SHAP scores for all entities,
  with increasing numbers of them (JELIA'23)

The Need for Reasoning

- What can we do with attribution scores and counterfactual explanations? (apart from the obvious)
- We can reason about/with them, analyze them, select some of them, aggregate them, etc.

  In interaction with both attribution-score model/algorithm or classifier, for further exploration

  For global understanding of the classifier or application domain
- We need tools for conveying or imposing domain knowledge (domain semantics), e.g. an age never decreases

  Only some counterfactuals may make sense

  Some combinations of feature values may not be allowed

  Some changes may "trigger" other changes

  To impose preferences on counterfactuals

- We need tools for doing this kind of logical reasoning
- We need tools for posing and answering queries about explanations

  Are there explanations with this particular property?

  Or any two that differ by ...?
- Specification of high-score actionable explanations, and possibly computation of those only

  Or others with a different preferred property
- On-the-fly interaction with different ML models and scores

  Do I get same score with this different ML system?

  Or this other attribution score (definition, algorithm or implementation)?

- Imposing conditions on feature values

  What if I leave some feature values fixed?

  Do I get same high-score feature with this "similar" entity?

  Is there a high-score counterfactual version of the entity that changes this specific feature?

  Or never changes that one?

- We have devised *declarative* (logic-based) methods to *reason* with and about counterfactuals, and compute *Resp* scores

  In interaction with classifiers "called" from the program

  We have used *Answer-Set Programming*, a form of logic programming

  Much in the direction of Neuro-Symbolic AI!