skema BUSINESS SCHOOL | AI R&D CENTER CANADA

Explainable Artificial Intelligence

for

Classification and Decision Systems

Leopoldo Bertossi

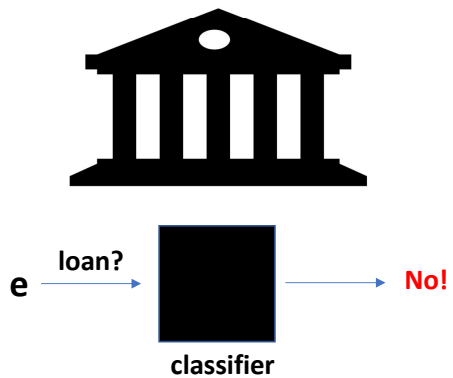
leopoldo.bertossi@skema.edu

"Carleton University Institute for Data Science" Seminar, March 2023

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, ...
- **e** requests a loan from a bank, which uses a classifier



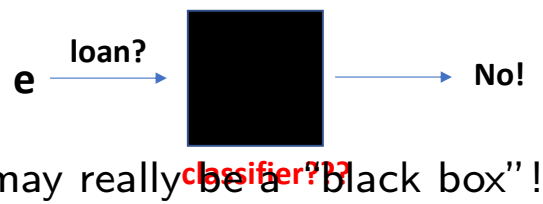
- The client asks *Why?*
- What kind of *explanation?*
  How?
  From what?

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



Explanations (in AI)

- 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, e.g. Logic, Statistics

  Some forms of explanations are new in AI

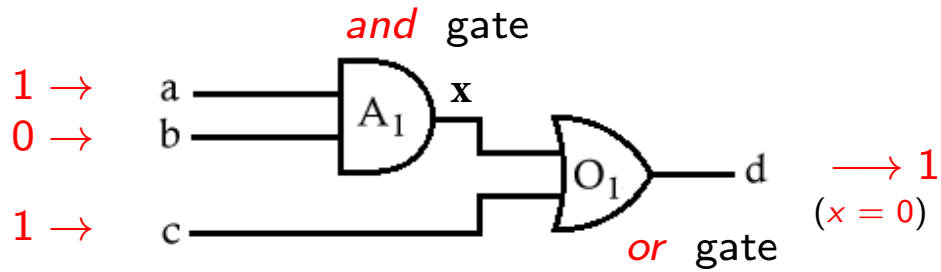
  Others have roots in already existing ones

Model-Based Diagnosis

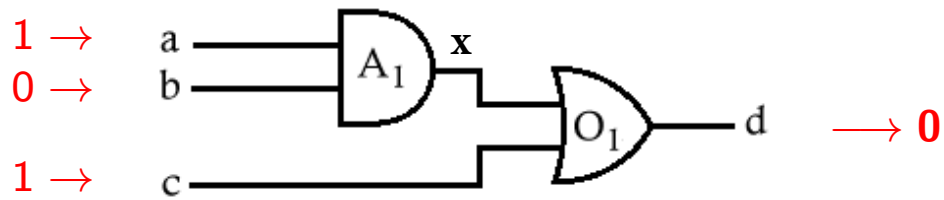
- MBD has been an area of AI for some time
- It is about doing a *diagnosis* of a system (exhibiting some unexpected behavior) using a model of the system (and possibly a bit more)

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

It should be:

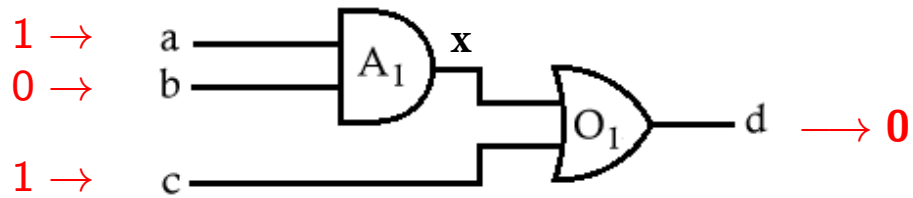


However:



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 perfectly consistent
- But $Obs \cup \mathcal{M} \cup \{\neg AbA, \neg AbO\}$ is inconsistent (as before)
- So, something has to be abnormal ...
- $D = \{abO\}$ is a diagnosis, because making gate O abnormal restores consistency

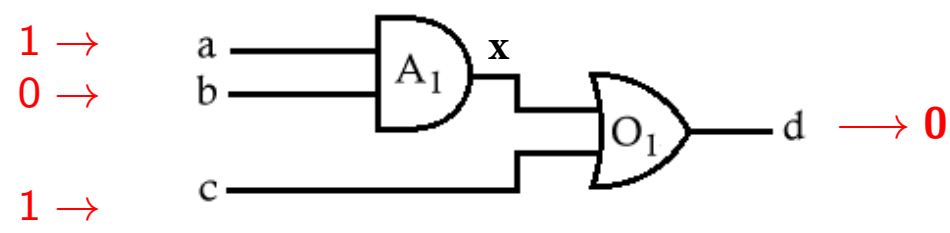
  $Obs \cup \mathcal{M} \cup \{\neg AbA, AbO\}$ is consistent

  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

  $Obs \cup \mathcal{M} \cup \{AbA, AbO\}$ is consistent
- 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.)



We had:

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

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

Logically equivalent to:

$$\{a, \neg b, c\} \cup \mathcal{M} \cup \{\underline{\neg AbA, \neg AbO}\} \Longrightarrow d \qquad (*)$$

Counterfactuals: hypothetical changes of non-abnormalities into abnormalities, to see if implication changes

$$\{a, \neg b, c\} \cup \mathcal{M} \cup \{\neg AbA, \underbrace{AbO}_{\text{changed}}\} \underbrace{\not\Longrightarrow}_{\text{switched}} d$$

abO is a *counterfactual cause* for the observation

However: $\{a, \neg b, c\} \cup \mathcal{M} \cup \{\underbrace{AbA}_{\text{changed}}, \neg AbO\} \underbrace{\Longrightarrow}_{\text{not switched}} d$

AbA is not a counterfactual cause

Extra counterfactual changes may be necessary:

$$\{a, \neg b, c\} \cup \mathcal{M} \cup \{\underbrace{AbA}_{\text{changed}}, \underbrace{AbO}_{\text{contingent change}}\} \underbrace{\not\Longrightarrow}_{\text{switched}} d$$

Had it not been abO a counterfactual cause, AbA would have been an *actual cause with Contingency Set* $\{abO\}$

AbA is neither counterfactual nor actual cause

- Actual Causality: J. Halpern & J. Pearl (2001)
- Actual causality provides *counterfactual explanations*
- Correspondences with both forms of MBD
- Numerical scores to quantify strength of a cause?

Causal Responsibility (Chokler & Halpern, 2004)

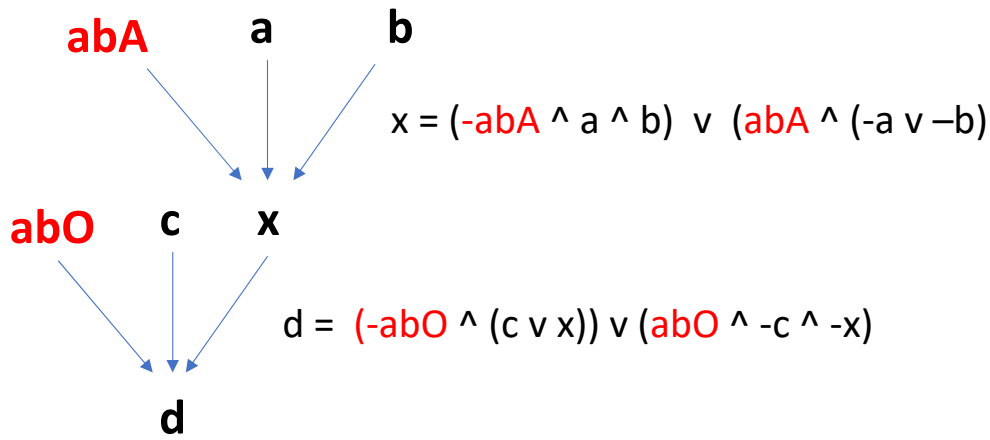
$Resp(abO) := \frac{1}{1+\text{min. cardinality of CS}} = \frac{1}{1+0} = 1$ (max. responsibility)

$Resp(abA) := 0$

The Causal Networks Connection

- Actual causality as presented may not look like the *Causal Networks* and *Structural Models* used in AI

  It can be cast in those terms



- Here *abA*, *abO* are endogenous variables, which can be subject to counterfactual changes

  The others are exogenous variables

- Links have *structural equations*

Some Applications of Actual Causality

- We have applied AC to explanations for query answers from databases
- Explanations are DB tuples that contribute to a query answer

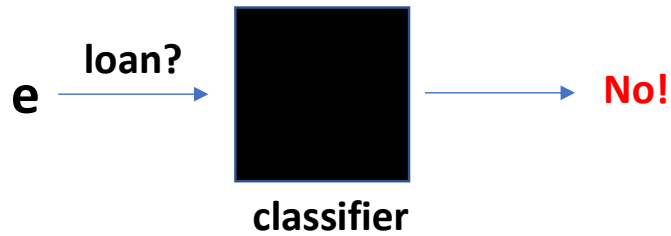
  Or attribute values in them
- Tuples get *responsibility scores*, quantifying *how much* they contribute
- We have established some connections with MBD

  Profiting from those connections
- We have applied AC to explanations for outcomes from ML classification systems $\longrightarrow$ XAI
- 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

Resp and Explanations (gist and simple case)



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

- Counterfactual versions:

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

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

- For the gist:
  1. Value for feature *Age* is counterfactual cause with explanatory responsibility $\mathit{Resp}(\mathbf{e}, \mathit{Age}) = 1$
  2. Value for *Income* is actual cause with $\mathit{Resp}(\mathbf{e}, \mathit{Income}) = \frac{1}{2}$

     This one needs additional (contingent) changes ...

Causality and Responsibility -4

- For binary features the previous definition of responsibility (as for DBs) 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 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: $Resp(\mathbf{e}, F^\star) := \max_{\langle \Gamma, \bar{w} \rangle,\ |\Gamma| \text{ min.},\ (*) > 0} Resp(\mathbf{e}, F^\star, \Gamma, \bar{w})$

Generalized Responsibility -2

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

  Several probability distributions can be used (B. et al., Deem@SIGMOD20)
- In our experiments, *Resp* score computed with empirical product distribution

  This is quite a relevant issue ...

Generalized Responsibility -3

- 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*

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: $Shap(\mathcal{F}, \mathcal{G}_{\mathbf{e}}, F^\star)$

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

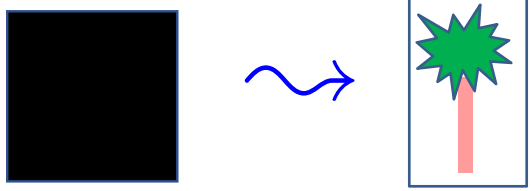
- *Shap* score has become popular (Lee & Lundberg, 2017)
- Assumes a probability distribution on entity population
- Requires multiple passes through classifier ...

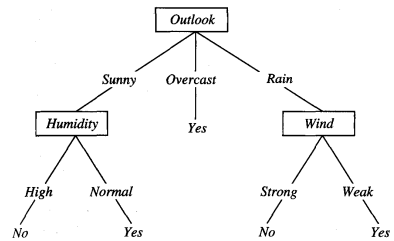
Shap: **Tractability** -1

- Both *Resp* and *Shap* may end up considering exponentially many combinations

  And multiple passes through the black-box classifier

- Both provably intractable in the general case
- Can we do better with an open-box classifier?





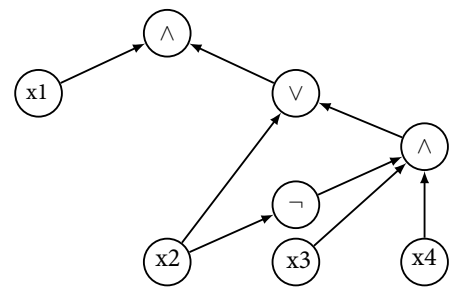
  Exploiting its elements and internal structure?

- What if we have a decision tree, or a random forest, or a Boolean circuit?
- Can we compute *Shap* in polynomial time?

Shap: **Tractability** -2

- We investigated this problem in detail (Arenas, Barcelo, B., Monet; AAAI21)
- Tractable and intractable cases, with algorithms for the former

  Investigated existence (or not) of good approximation algorithms
- Choosing the right abstraction (model) is crucial
- We used Boolean classifiers (BCs), i.e. propositional formulas with (binary) output gate
- We established early on that computing *Shap* is at least as hard as counting the satisfying truth assignments of the BC (intractable in general)



- So, it has to be a broad and interesting class of BCs for which the latter problem is not intractable

Shap: **Tractability** -3

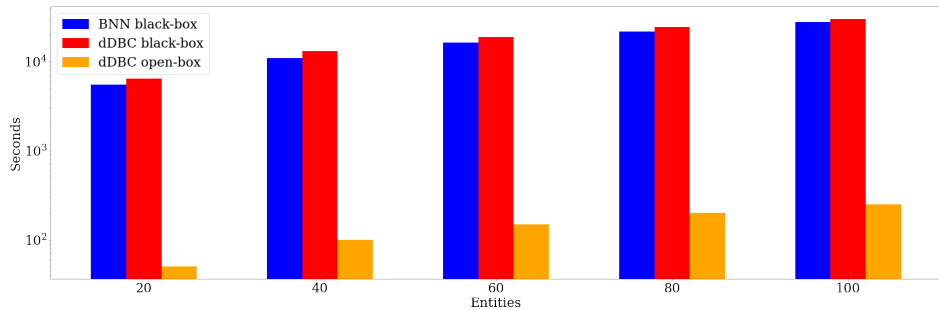
- We concentrated on the class of deterministic and decomposable Boolean circuits (dDBCs) (example above)
  - Input gates are variables (features) or constants
  - An $\vee$-gate never has both inputs true (determinism)
  - An $\wedge$-gate do not has inputs sharing variables (decomposability)
- A class of BCs that includes -possibly via efficient compilation- many interesting ones, syntactic and not ...
  - Decision trees (and random forests)
  - Ordered binary decision diagrams (OBDDs)
  - Sentential decision diagrams (SDDs)
  - Deterministic-decomposable negation normal-form (dDNNFs)
- Theorem: For dDBCs, under the uniform or product distribution, *Shap* can be computed in polynomial time

Shap: **Tractability** -4

- Binary Neural Networks (BNNs) -usually considered black-box models- can be compiled into OBDDs (Shi et al., KR20)
- Opening the ground for efficient *Shap* computation for BNNs (via additional compilation into dDBC)
- We have experimented with *Shap* computation with a black-box BNN and with its compilation into an open-box dDBC

  Considerable efficiency gain

  (this is logarithmic scale)



- A BNN with 14 gates was compiled into a dDBC with 18,670 nodes

  A one-time computation that fully replaces the BNN

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?

References (some publications for this presentation)

- L. Bertossi, L. and B. Salimi. "From Causes for Database Queries to Repairs and Model-Based Diagnosis and Back". *Theory of Computing Systems*, 2017, 61(1):191-232.

- L. Bertossi and B. Salimi. "Causes for Query Answers from Databases: Datalog Abduction, View-Updates, and Integrity Constraints". *International Journal of Approximate Reasoning*, 2017, 90:226-252.

- L. Bertossi. "Specifying and Computing Causes for Query Answers in Databases via Database Repairs and Repair Programs". *Knowledge and Information Systems*, 2021, 63(1):199-231.

- E. Livshits, L. Bertossi, B. Kimelfeld and M. Sebag. "The Shapley Value of Tuples in Query Answering". *Logical Methods in Computer Science*, 17(3):22.1-22.33.

- E. Livshits, L. Bertossi, B. Kimelfeld, M. Sebag. "Query Games in Databases". *ACM Sigmod Record*, 2021, 50(1):78-85.

- L. Bertossi, J. Li, M. Schleich, D. Suciu and Z. Vagena. "Causality-based Explanation of Classification Outcomes". Proc. 4th International Workshop on "Data Management for End-to-End Machine Learning" (DEEM) at ACM SIGMOD/PODS, 2020, pp. 6.1-6.10.

- Leopoldo Bertossi. "Score-Based Explanations in Data Management and Machine Learning: An Answer-Set Programming Approach to Counterfactual Analysis". In *Reasoning Web. Declarative Artificial Intelligence*. Reasoning Web 2021. Springer LNCS 13100, 2022, pp. 145-184.

- M. Arenas, P. Barcelo, L. Bertossi, M. Monet. "The Tractability of SHAP-scores over Deterministic and Decomposable Boolean Circuits". To appear in *Journal of Machine Learning Research*. Extended version of AAAI 2021 paper. arXiv Paper 2104.08015, 2021

- L. Bertossi. "Declarative Approaches to Counterfactual Explanations for Classification". *Theory and Practice of Logic Programming*, 2022. (forthcoming) arXiv Paper 2011.07423, 2021.

- L. Bertossi. "Score-Based Explanations in Data Management and Machine Learning". Proc. Int. Conf. Scalable Uncertainty Management (SUM 20), Springer LNCS 2322, pp. 17-31.

- L. Bertossi and G. Reyes. "Answer-Set Programs for Reasoning about Counterfactual Interventions and Responsibility Scores for Classification". In Proc. 1st International Joint Conference on Learning and Reasoning (IJCLR'21), Springer LNAI 13191, 2022, pp. 41-56.