skema BUSINESS SCHOOL | AI R&D CENTER CANADA

Explainable Artificial Intelligence

and

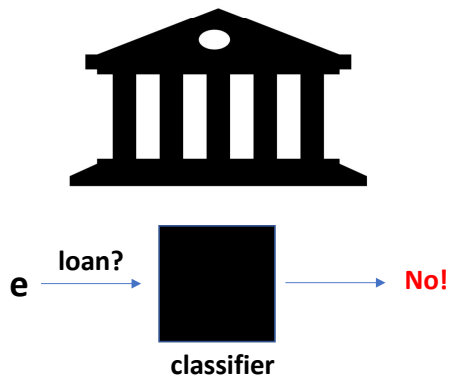
Classification Systems

Leopoldo Bertossi
leopoldo.bertossi@skema.edu

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

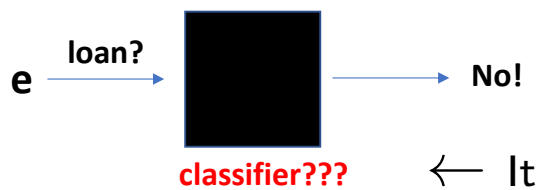
- This problem is representative of a more general situation in applications of AI systems
- Users and those affected by results from AI systems, the stakeholders, request explanations
- A whole new area of AI has emerged: *Explainable AI* (XAI)
- Part of AI because:
  - AI systems should be extended with the capability to provide explanations
  - AI researchers and professionals are those who understand these systems

    So as mathematical logicians study the methods and scope of Math (with the methods of Math)
  - Humans give explanations as part of their intelligent activities

    Hence, explanation building should be a capability of AI agents

    Then, explanations have to be understood, modeled, implemented, ... as part of AI

- XAI is of interest to many other people
- We talked about stakeholders being affected by *outcomes from AI systems*
  Assessments (e.g. a credit score), classifications (good/bad client), decisions (approve/reject loan), etc.
- A whole discipline has emerged: *Ethical AI*
- It touches many others, including AI itself, but beyond: Law, Sociology, Philosophy, ..., Business, ...
- It emerges naturally, and motivated by the need for more *transparent, trustable, fair, unbiased, ...* AI systems
- Also, *interpretable* AI systems



⟵ It may really be a "black box"!

- New legislation forces (owners of) 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

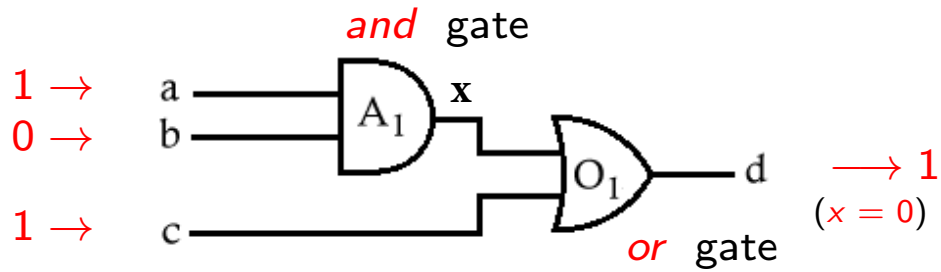
  Others have roots in already existing ones

Model-Based Diagnosis

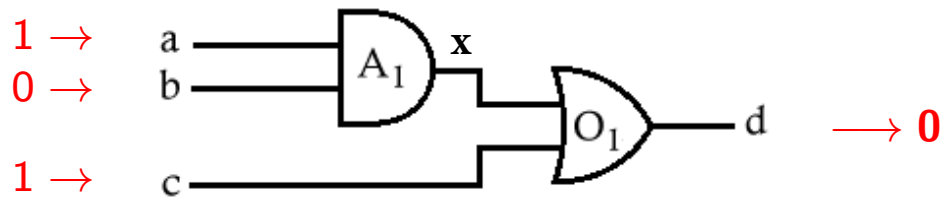
- MBD has been an area of AI for some time
- It is about doing a *diagnosis* of a system (exhibiting some unexpected behavior) using a model of the system (and possibly a bit more)

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

It should be:

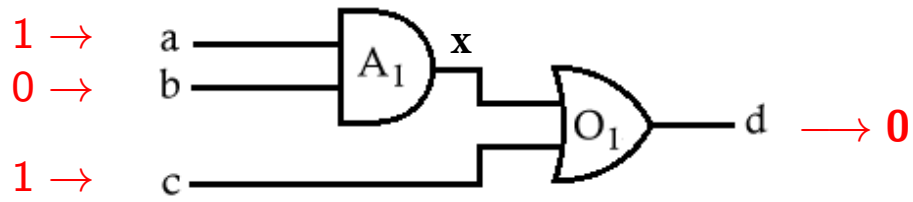


However:



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)

- Abductive Diagnosis (AD) is also "classic" MBD in AI

  Example: (typical)

  $$Covid19 \longrightarrow Fever \quad \text{(the model)}$$
  $$Obs: \ Fever$$

  W/O other information, we would like to "infer" *Covid19* as an explanation

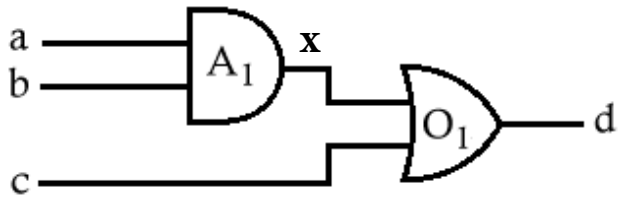
  Not classic logic-based deduction, but *abduction* of an explanation, in that:

  $$\{Covid19\} \cup \{Covid19 \longrightarrow Fever\} \Longrightarrow Obs$$

  Sort of *backward reasoning* in search for explanations that support implications (forward reasoning)

- Abductive or sufficient explanations are relevant in XAI
- There are connections between these forms of MBD

Example: (cont., different ways)



$$abO \wedge c \rightarrow \neg d \quad (1)$$

$$abO \wedge x \rightarrow \neg d \quad (2)$$

$$abA \wedge a \wedge b \rightarrow \neg x \quad (3)$$

$$a,\ \neg b,\ c,\ \neg d \quad (4)$$

- Start from $\neg d$ in (1), reason backwards trying to find what is needed to prove it (according to classical logic)

  Hopefully an "abductible variable" will be reached

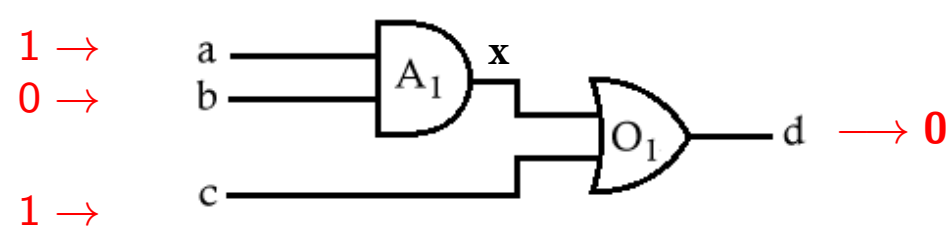
- With (1) and $\neg d$ in (4), reach: $abO \wedge c$

  Since c is given in (4), we reach abO as all is needed to obtain $\neg d$

- Now: $\{abO\} \cup \{(1), (2), (3)\} \cup \{a, \neg b, c\} \Rightarrow \neg d$

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 conterfactual 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)

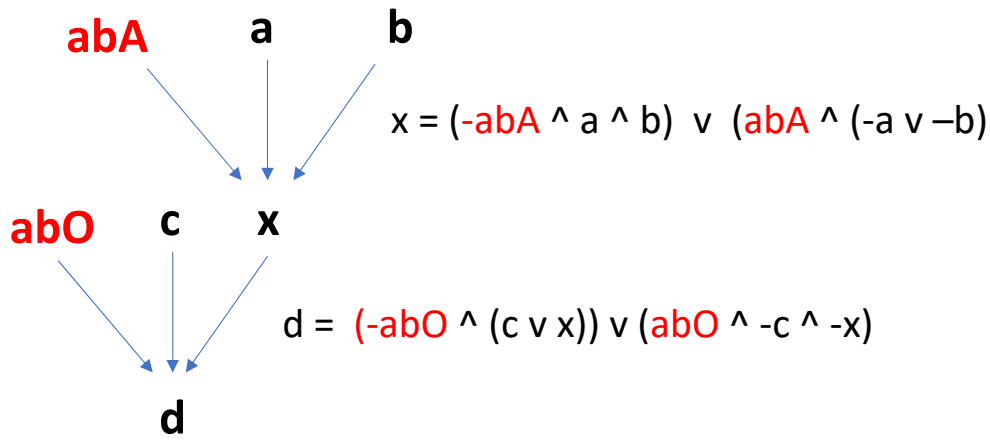
$Resp(abO) := \frac{1}{1+\text{min. cardinality of CS}} = \frac{1}{1+0} = 1$ (max. responsibility)

$Resp(abA) := 0$

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

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

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

- We are working on a precise and general connection with MBD (see above for the gist)

- We have experimentally compared responsibility scores with other *local attribution scores*: *Causal-Effect*, *Shap*

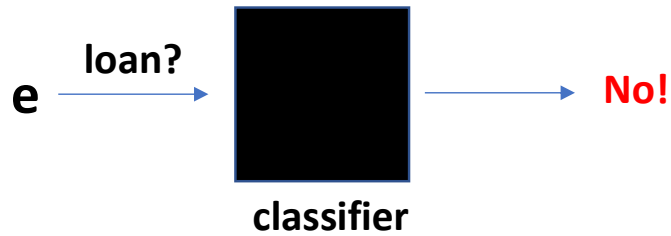
  And other scores based on (used with) "open models" (e.g. connected logistic regressions)

  With financial data

- We have established that score computations "behave better" when applied with an open classifier

- There is still much research to do in all these fronts ...

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:

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

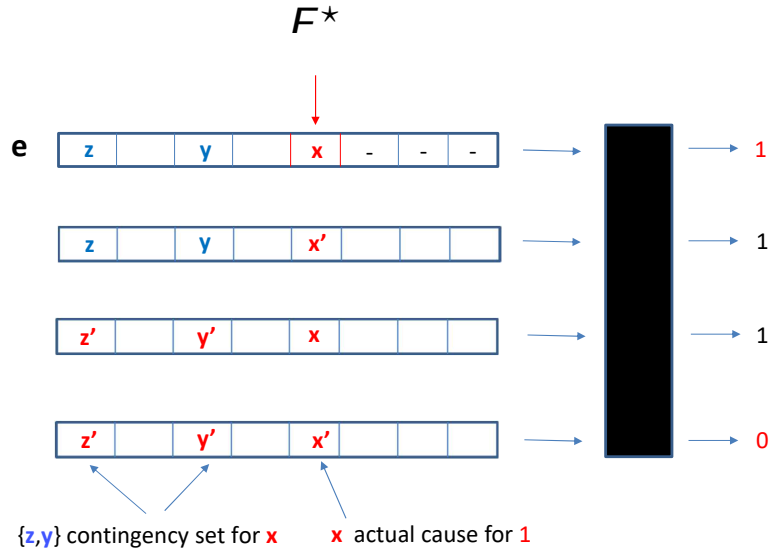
$$\mathbf{e}'' = \langle \text{john}, 18, \text{plumber}, 80\text{K}, \text{brooklyn}, \ldots \rangle \quad \text{Yes}$$

- 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 ...

$Resp(\mathbf{e}, F^\star)$ score for value of feature $F^\star$ in $\mathbf{e}$:

- Want explanation for label "1"
- Through value changes for feature $F^\star$, try to get "0"
- Feature value $\mathbf{x} = \mathbf{e}_{F^\star}$



- $\mathbf{x}$ counterfactual explanation for $L(\mathbf{e}) = 1$ if $L(\mathbf{e}\frac{\mathbf{x}}{\mathbf{x}'}) = 0$, for some $\mathbf{x}' \in Dom(F^\star)$
- $\mathbf{x}$ actual explanation for $L(\mathbf{e}) = 1$ if there are values $\mathbf{Y}$ in $\mathbf{e}$, $\mathbf{x} \notin \mathbf{Y}$, and new values $\mathbf{Y}' \cup \{\mathbf{x}'\}$:

  (a) $L(\mathbf{e}\frac{\mathbf{Y}}{\mathbf{Y}'}) = 1$ (b) $L(\mathbf{e}\frac{\mathbf{x}\mathbf{Y}}{\mathbf{x}'\mathbf{Y}'}) = 0$

- For minimum-size contingency set $\mathbf{Y}$: $Resp(\mathbf{e}, F^\star) := \frac{1}{1+|\mathbf{Y}|}$

The *Resp* Score: Towards a General Definition

- For binary features the previous definition works fine
- Otherwise, there may be many values for a feature that do not change the label: original value not great explanation
- First attempt: Consider all possible values for a fixed feature, w/o contingent changes (of other values)

  Consider the average label obtained this way, i.e. *Resp* is expressed as an expected value (Bertossi et al.; 2020)

  - Entity $\mathbf{e} = \langle \ldots, \mathbf{e}_F, \ldots \rangle$, $F^\star \in \mathcal{F}$ (set of features)

$$Counter(\mathbf{e}, F^\star) := \underbrace{L(\mathbf{e})}_{1} - \mathbb{E}(L(\mathbf{e}') \mid \underbrace{\mathbf{e}'_{\mathcal{F} \setminus \{F^\star\}}}_{\text{(coincides with } \mathbf{e} \text{ outside } F^\star)} = \mathbf{e}_{\mathcal{F} \setminus \{F^\star\}})$$

  - Easy to compute, worth trying ...
  - Experimentally, gives reasonable results
  - Requires (estimated) probability on entity population

The *Resp* Score: General Definition

- Changing one value (no contingencies) may not switch label

  No explanations are obtained

  Better consider both contingencies and average labels!

- $\mathbf{e}$ entity under classification, $L(\mathbf{e}) = 1$, $F^\star \in \mathcal{F}$

- "Local" *Resp*-score: for fixed contingent assignment $\Gamma := \bar{w}$

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

  - $\Gamma \subseteq \mathcal{F} \setminus \{F^\star\}$ (potential contingent set of features)
  - $\mathbf{e}' := \mathbf{e}[\Gamma := \bar{w}]$, $L(\mathbf{e}') = L(\mathbf{e})$ (potential contingent values)
  - $\mathbf{e}'' := \mathbf{e}[\Gamma := \bar{w}, F^\star := v]$, with $v \in dom(F^\star)$
  - When $F^\star(\mathbf{e}) \neq v$, $L(\mathbf{e}'') \neq L(\mathbf{e})$, $F^\star(\mathbf{e})$ is *actual causal explanation* for $L(\mathbf{e}) = 1$ with contingency $\langle \Gamma, \mathbf{e}_\Gamma \rangle$

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

Some Remarks

- 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*

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?

Beyond Explanations

- Explainability in AI is related to other dimensions of Ethical AI

  Especially in combination with reasoning

- In particular, causality and explanations are related to *Fairness*

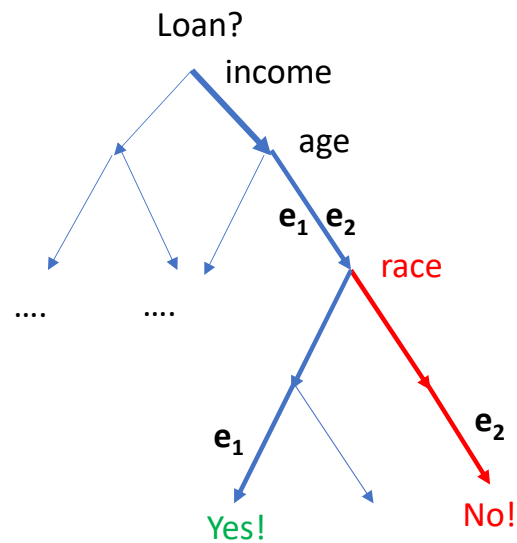
  We want AI and ML systems to be *fair*

- Reasoning and query answering can help specify and detect unfair situations or behaviors ...

  For example, about decisions related to protected features, e.g. *Race* here

  Paths in Decision Tree for two entities diverge at that point, getting different labels

- We can keep track of counterfactual "histories" and compare them



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.