# Score-Based Explanations in Data Management and Machine Learning


Leopoldo Bertossi*

Universidad Adolfo Ibáñez

Faculty of Engineering and Sciences

Santiago, Chile

*: Also: RelationalAI Inc. and IMFD, Chile

# Explanations in Databases

- In data management (DM), we need to understand and compute ***why*** certain results are obtained or not

  E.g. query answers, violation of semantic conditions, ...

- A DB system should provide *explanations*

  In our case, causality-based explanations (Halpern and Pearl, 2001)

  There are other (related) approaches, e.g. *lineage*, *provenance*

- Our interest: model, specify and compute causes

  Understand causality in DM from different perspectives; and profit from the connections

# Causality in DBs

Example: DB $D$ as below

Boolean conjunctive query (BCQ):

$\mathcal{Q}\colon\ \exists x \exists y (S(x) \land R(x,y) \land S(y))$

$D \models \mathcal{Q}$ Causes?

| $R$ | $A$ | $B$ |
|---|---|---|
| | $a$ | $b$ |
| | $c$ | $d$ |
| | $b$ | $b$ |

| $S$ | $A$ |
|---|---|
| | $a$ |
| | $c$ |
| | $b$ |

(Meliou, Gatterbauer, Moore, Suciu; 2010)

- Tuple $\tau \in D$ is counterfactual cause for $\mathcal{Q}$ if $D \models \mathcal{Q}$ and $D \smallsetminus \{\tau\} \not\models \mathcal{Q}$

  $S(b)$ is counterfactual cause for $\mathcal{Q}$: if $S(b)$ is removed from $D$, $\mathcal{Q}$ is no longer an answer

# Causality in DBs

Example: DB $D$ as below

Boolean conjunctive query (BCQ):

$\mathcal{Q}$: $\exists x \exists y (S(x) \land R(x,y) \land S(y))$

$D \models \mathcal{Q}$ Causes?

| $R$ | $A$ | $B$ |
|---|---|---|
| | $a$ | $b$ |
| | $c$ | $d$ |
| | $b$ | $b$ |

| $S$ | $A$ |
|---|---|
| | $a$ |
| | $c$ |
| | $b$ |

(Meliou, Gatterbauer, Moore, Suciu; 2010)

- Tuple $\tau \in D$ is counterfactual cause for $\mathcal{Q}$ if $D \models \mathcal{Q}$ and $D \smallsetminus \{\tau\} \not\models \mathcal{Q}$

  $S(b)$ is counterfactual cause for $\mathcal{Q}$: if $S(b)$ is removed from $D$, $\mathcal{Q}$ is no longer an answer

- Tuple $\tau \in D$ is actual cause for $\mathcal{Q}$ if there is a contingency set $\Gamma \subseteq D$, such that $\tau$ is a counterfactual cause for $\mathcal{Q}$ in $D \smallsetminus \Gamma$

  $R(a,b)$ is an actual cause for $\mathcal{Q}$ with contingency set $\{R(b,b)\}$: if $R(a,b)$ is removed from $D$, $\mathcal{Q}$ is still true, but further removing $R(b,b)$ makes $\mathcal{Q}$ false

- How strong are these as causes? (Chockler and Halpern, 2004)

  - The responsibility of an actual cause $\tau$ for $\mathcal{Q}$:

    $$\rho_D(\tau) \ := \ \frac{1}{|\Gamma| \ + \ 1}$$

    $|\Gamma| =$ size of smallest contingency set for $\tau$ (0 otherwise)

    Responsibility of $R(a,b)$ is $\frac{1}{2} = \frac{1}{1+1}$ (its several smallest contingency sets have all size $1$)

    $R(b,b)$ and $S(a)$ are also actual causes with responsibility $\frac{1}{2}$

    $S(b)$ is actual (counterfactual) cause with responsibility $1 = \frac{1}{1+0}$

- How strong are these as causes? (Chockler and Halpern, 2004)

  - The responsibility of an actual cause $\tau$ for $\mathcal{Q}$:

    $$\rho_{_D}(\tau) \ := \ \frac{1}{|\Gamma| \ + \ 1}$$

    $|\Gamma| =$ size of smallest contingency set for $\tau$ (0 otherwise)

    Responsibility of $R(a,b)$ is $\frac{1}{2} = \frac{1}{1+1}$ (its several smallest contingency sets have all size $1$)

    $R(b,b)$ and $S(a)$ are also actual causes with responsibility $\frac{1}{2}$

    $S(b)$ is actual (counterfactual) cause with responsibility $1 = \frac{1}{1+0}$

  High responsibility tuples provide more interesting explanations

- Causes in this case are tuples that come with their responsibilities as "scores"

  All tuples can be seen as actual causes and only the non-zero scores matter

- Causality can be extended to attribute-value level (Bertossi, Salimi; TOCS 2017)

# Connections: Repairs and Diagnosis

- There is a connection with repairs of DBs wrt. integrity constraints (ICs)

  A connection to consistency-based diagnosis and abductive diagnosis

  $\rightsquigarrow$ new complexity and algorithmic results for causality and responsibility

  (Bertossi, Salimi; TOCS, IJAR, 2017)

- Causality under ICs $\rightsquigarrow$ Causality under semantic, domain knowledge (op. cit.)

- Model-Based Diagnosis is an older area of Knowledge Representation

  A logic-based model is used

  Elements of the model are identified as explanations

- Causality-based explanations are newer

  Still a model is used, representing a more complex scenario than a DB and a query

  Pearl's causality: Perform counterfactual *interventions* on a structural, logico/probabilistic model

  *What would happen if we change ...?*

- Causality-based explanations are newer

  Still a model is used, representing a more complex scenario than a DB and a query

  Pearl's causality: Perform counterfactual *interventions* on a structural, logico/probabilistic model

  *What would happen if we change ...?*

- In the case of DBs the underlying logical model is *query lineage* (coming ...)

- Much newer in "explainable AI": Provide explanations in the possible absence of a model

- Explainability scores have become popular (coming ...)

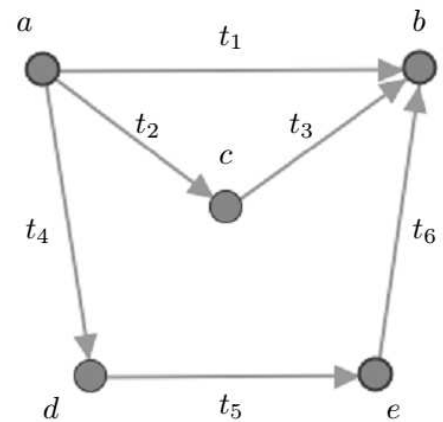
  They usually have a counterfactual component: *What would happen if ...?*

  Responsibility can be seen as such ...

# The Causal Effect Score

Example: Boolean Datalog query $\Pi$ becomes true on $E$ if there is a path between $a$ and $b$

| $E$ | $X$ | $Y$ |
|---|---|---|
| $t_1$ | $a$ | $b$ |
| $t_2$ | $a$ | $c$ |
| $t_3$ | $c$ | $b$ |
| $t_4$ | $a$ | $d$ |
| $t_5$ | $d$ | $e$ |
| $t_6$ | $e$ | $b$ |



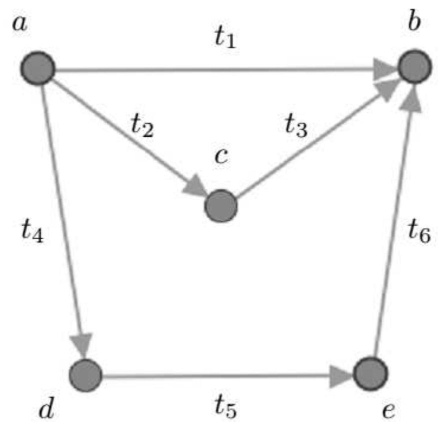

$$\begin{array}{rcl} yes & \leftarrow & P(a,b) \\ P(x,y) & \leftarrow & E(x,y) \\ P(x,y) & \leftarrow & P(x,z), E(z,y) \end{array}$$

$E \cup \Pi \models yes$

# The Causal Effect Score

Example: Boolean Datalog query $\Pi$ becomes true on $E$ if there is a path between $a$ and $b$

| $E$ | $X$ | $Y$ |
|---|---|---|
| $t_1$ | $a$ | $b$ |
| $t_2$ | $a$ | $c$ |
| $t_3$ | $c$ | $b$ |
| $t_4$ | $a$ | $d$ |
| $t_5$ | $d$ | $e$ |
| $t_6$ | $e$ | $b$ |




$$
\begin{aligned}
yes &\leftarrow P(a,b) \\
P(x,y) &\leftarrow E(x,y) \\
P(x,y) &\leftarrow P(x,z), E(z,y)
\end{aligned}
$$

$E \cup \Pi \models yes$

All tuples are actual causes: every tuple appears in a path from $a$ to $b$

All the tuples have the same causal responsibility: $\frac{1}{3}$

Maybe counterintuitive: $t_1$ provides a direct path from $a$ to $b$

- Alternative notion to responsibility: *causal effect* (Salimi et al., TaPP'16)
- Causal responsibility has been criticized for other reasons and from different angles
- Retake question: How answer to $\mathcal{Q}$ changes if $\tau$ deleted from $D$? (inserted)

  An *intervention* on a *structural causal model*

  In this case provided by the the *lineage of the query*

Example: Database $D$

| $R$ | $A$ | $B$ |
|---|---|---|
| | $a$ | $b$ |
| | $a$ | $c$ |
| | $c$ | $b$ |

| $S$ | $B$ |
|---|---|
| | $b$ |
| | $c$ |

BCQ $\mathcal{Q}: \exists x \exists y (R(x,y) \land S(y))$ True in $D$

Query lineage instantiated on $D$ given by propositional formula:

$$\Phi_{\mathcal{Q}}(D) = (X_{R(a,b)} \land X_{S(b)}) \lor (X_{R(a,c)} \land X_{S(c)}) \lor (X_{R(c,b)} \land X_{S(b)}) \qquad (*)$$

$X_\tau$: propositional variable that is true iff $\tau \in D$

$\Phi_{\mathcal{Q}}(D)$ takes value $1$ in $D$

Example: Database $D$

| $R$ | $A$ | $B$ |
|---|---|---|
| | $a$ | $b$ |
| | $a$ | $c$ |
| | $c$ | $b$ |

| $S$ | $B$ |
|---|---|
| | $b$ |
| | $c$ |

BCQ $\mathcal{Q}: \ \exists x \exists y (R(x,y) \wedge S(y))$ True in $D$

Query lineage instantiated on $D$ given by propositional formula:

$$\Phi_{\mathcal{Q}}(D) = (X_{R(a,b)} \wedge X_{S(b)}) \vee (X_{R(a,c)} \wedge X_{S(c)}) \vee (X_{R(c,b)} \wedge X_{S(b)}) \qquad (*)$$

$X_\tau$: propositional variable that is true iff $\tau \in D$

$\Phi_{\mathcal{Q}}(D)$ takes value $1$ in $D$

- Want to quantify contribution of a tuple to a query answer, say, $S(b)$

  Assign probabilities, uniformly and independently, to the tuples in $D$

- A probabilistic database $D^p$ (tuples outside $D$ get probability $0$)

| $R^p$ | $A$ | $B$ | prob |
|---|---|---|---|
| | $a$ | $b$ | $\frac{1}{2}$ |
| | $a$ | $c$ | $\frac{1}{2}$ |
| | $c$ | $b$ | $\frac{1}{2}$ |

| $S^p$ | $B$ | prob |
|---|---|---|
| | $b$ | $\frac{1}{2}$ |
| | $c$ | $\frac{1}{2}$ |

- The $X_\tau$'s become independent, identically distributed random variables; and $\mathcal{Q}$ is Bernouilli random variable

  What's the probability that $\mathcal{Q}$ takes truth value $1$ (or $0$) when an intervention is done on $D$?

- A probabilistic database $D^p$ (tuples outside $D$ get probability $0$)

| $R^p$ | $A$ | $B$ | prob |
|---|---|---|---|
| | $a$ | $b$ | $\frac{1}{2}$ |
| | $a$ | $c$ | $\frac{1}{2}$ |
| | $c$ | $b$ | $\frac{1}{2}$ |

| $S^p$ | $B$ | prob |
|---|---|---|
| | $b$ | $\frac{1}{2}$ |
| | $c$ | $\frac{1}{2}$ |

- The $X_\tau$'s become independent, identically distributed random variables; and $\mathcal{Q}$ is Bernouilli random variable

  What's the probability that $\mathcal{Q}$ takes truth value $1$ (or $0$) when an intervention is done on $D$?

- Interventions of the form $do(X = x)$: In the *structural equations* make $X$ take value $x$

  For $\{y, x\} \subseteq \{0, 1\}$: $P(\mathcal{Q} = y \mid do(X_\tau = x))$? (i.e. make $X_\tau$ false/true)

  E.g. with $do(X_{S(b)} = 0)$ lineage $(*)$ becomes: $\Phi_{\mathcal{Q}}(D)\frac{X_{S(b)}}{0} := (X_{R(a,c)} \wedge X_{S(c)})$

- A probabilistic database $D^p$ (tuples outside $D$ get probability $0$)

| $R^p$ | $A$ | $B$ | prob |
|---|---|---|---|
| | $a$ | $b$ | $\frac{1}{2}$ |
| | $a$ | $c$ | $\frac{1}{2}$ |
| | $c$ | $b$ | $\frac{1}{2}$ |

| $S^p$ | $B$ | prob |
|---|---|---|
| | $b$ | $\frac{1}{2}$ |
| | $c$ | $\frac{1}{2}$ |

- The $X_\tau$'s become independent, identically distributed random variables; and $\mathcal{Q}$ is Bernouilli random variable

  What's the probability that $\mathcal{Q}$ takes truth value $1$ (or $0$) when an intervention is done on $D$?

- Interventions of the form $do(X = x)$: In the *structural equations* make $X$ take value $x$

  For $\{y, x\} \subseteq \{0, 1\}$: $P(\mathcal{Q} = y \mid do(X_\tau = x))$? (i.e. make $X_\tau$ false/true)

  E.g. with $do(X_{S(b)} = 0)$ lineage $(*)$ becomes: $\Phi_{\mathcal{Q}}(D)\frac{X_{S(b)}}{0} := (X_{R(a,c)} \wedge X_{S(c)})$

- The *causal effect* of $\tau$: $\mathcal{CE}^{D,\mathcal{Q}}(\tau) := \mathbb{E}(\mathcal{Q} \mid do(X_\tau = 1)) - \mathbb{E}(\mathcal{Q} \mid do(X_\tau = 0))$

$$\mathcal{CE}^{D,\mathcal{Q}}(\tau) \ := \ \mathbb{E}(\mathcal{Q} \mid do(X_\tau = 1)) - \mathbb{E}(\mathcal{Q} \mid do(X_\tau = 0))$$

Example: (cont.) With $D^p$, when $X_{S(b)}$ is made false, probability that instantiated lineage becomes true in $D^p$:

$$P(\mathcal{Q} = 1 \mid do(X_{S(b)} = 0)) = P(X_{R(a,c)} = 1) \times P(X_{S(c)} = 1) = \tfrac{1}{4}$$

When $X_{S(b)}$ is made true, probability of lineage becoming true in $D^p$:

$$\Phi_{\mathcal{Q}}(D)\frac{X_{S(b)}}{1} := X_{R(a,b)} \vee (X_{R(a,c)} \wedge X_{S(c)}) \vee X_{R(c,b)}$$

$$P(\mathcal{Q} = 1 \mid do(X_{S(b)} = 1)) = P(X_{R(a,b)} \vee (X_{R(a,c)} \wedge X_{S(c)}) \vee X_{R(c,b)} = 1)$$
$$= \cdots \ = \ \tfrac{13}{16}$$

$$\mathcal{CE}^{D,\mathcal{Q}}(\tau) \ := \ \mathbb{E}(\mathcal{Q} \mid do(X_\tau = 1)) - \mathbb{E}(\mathcal{Q} \mid do(X_\tau = 0))$$

Example: (cont.) With $D^p$, when $X_{S(b)}$ is made false, probability that instantiated lineage becomes true in $D^p$:

$$P(\mathcal{Q} = 1 \mid do(X_{S(b)} = 0)) = P(X_{R(a,c)} = 1) \times P(X_{S(c)} = 1) = \tfrac{1}{4}$$

When $X_{S(b)}$ is made true, probability of lineage becoming true in $D^p$:

$$\Phi_{\mathcal{Q}}(D)\frac{X_{S(b)}}{1} := X_{R(a,b)} \vee (X_{R(a,c)} \wedge X_{S(c)}) \vee X_{R(c,b)}$$

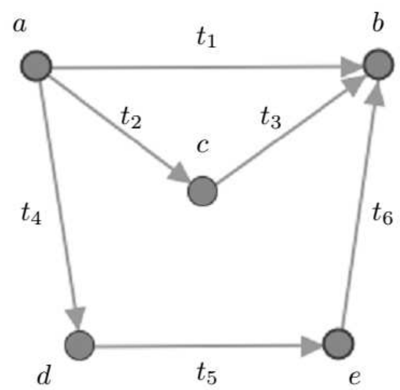
$$P(\mathcal{Q} = 1 \mid do(X_{S(b)} = 1)) = P(X_{R(a,b)} \vee (X_{R(a,c)} \wedge X_{S(c)}) \vee X_{R(c,b)} = 1)$$
$$= \cdots = \tfrac{13}{16}$$

$$\mathbb{E}(\mathcal{Q} \mid do(X_{S(b)} = 0)) = P(\mathcal{Q} = 1 \mid do(X_{S(b)} = 0)) \ = \ \tfrac{1}{4}$$

$$\mathbb{E}(\mathcal{Q} \mid do(X_{S(b)} = 1)) = \tfrac{13}{16}$$

$\mathcal{CE}^{D,\mathcal{Q}}(S(b)) = \frac{13}{16} - \frac{1}{4} = \frac{9}{16} \ > \ 0$ causal effect for actual cause $S(b)$!

Example: (cont.) The Datalog query (here as a union of BCQs) has the lineage:




$$\Phi_{\mathcal{Q}}(D) = X_{t_1} \vee (X_{t_2} \wedge X_{t_3}) \vee (X_{t_4} \wedge X_{t_5} \wedge X_{t_6})$$

$$\mathcal{CE}^{D,\mathcal{Q}}(t_1) = 0.65625$$

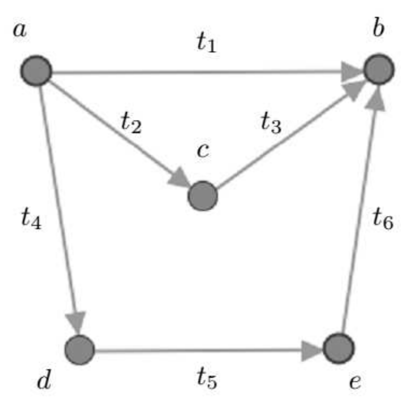
$$\mathcal{CE}^{D,\mathcal{Q}}(t_2) = \mathcal{CE}^{D,\mathcal{Q}}(t_3) = 0.21875$$

$$\mathcal{CE}^{D,\mathcal{Q}}(t_4) = \mathcal{CE}^{D,\mathcal{Q}}(t_5) = \mathcal{CE}^{D,\mathcal{Q}}(t_6) = 0.09375$$

The causal effects are different for different tuples!

More intuitive result than responsibility

Example: (cont.) The Datalog query (here as a union of BCQs) has the lineage:




$$\Phi_{\mathcal{Q}}(D) = X_{t_1} \vee (X_{t_2} \wedge X_{t_3}) \vee (X_{t_4} \wedge X_{t_5} \wedge X_{t_6})$$

$$\mathcal{CE}^{D,\mathcal{Q}}(t_1) = 0.65625$$

$$\mathcal{CE}^{D,\mathcal{Q}}(t_2) = \mathcal{CE}^{D,\mathcal{Q}}(t_3) = 0.21875$$

$$\mathcal{CE}^{D,\mathcal{Q}}(t_4) = \mathcal{CE}^{D,\mathcal{Q}}(t_5) = \mathcal{CE}^{D,\mathcal{Q}}(t_6) = 0.09375$$

The causal effects are different for different tuples!

More intuitive result than responsibility

- Rather *ad hoc* or arbitrary? (we'll be back ...)

# Scores and Coalition Games

- A starting point for a research direction: By how much a database tuple contributes to the inconsistency of a DB? (violation of an IC)

  $\rightsquigarrow$ Contribution of a DB tuple to a query answer?

# Scores and Coalition Games

- A starting point for a research direction: By how much a database tuple contributes to the inconsistency of a DB? (violation of an IC)

  $\rightsquigarrow$ Contribution of a DB tuple to a query answer?

- There had been research in KR on the Shapley-value to measure the inconsistency of a propositional KB

- The Shapley-value is firmly established in Game Theory, and used in several areas

  Why not investigate its application to query answering in DBs?

  (Livshits et al.; ICDT'20)

# Scores and Coalition Games

- A starting point for a research direction: By how much a database tuple contributes to the inconsistency of a DB? (violation of an IC)

  $\rightsquigarrow$ Contribution of a DB tuple to a query answer?

- There had been research in KR on the Shapley-value to measure the inconsistency of a propositional KB

- The Shapley-value is firmly established in Game Theory, and used in several areas

  Why not investigate its application to query answering in DBs?

  (Livshits et al.; ICDT'20)

- *Several tuples together* are necessary to violate an IC or produce a query result

  Like players in a coalition game, some may contribute more than others

  The Shapley-value of a tuple will be a score for its contribution

## The Shapley Value

- Consider a set of players $D$, and a wealth-distribution (game) function $\mathcal{G}: \ \mathcal{P}(D) \longrightarrow \mathbb{R}$ ($\mathcal{P}(D)$ the power set of $D$)

# The Shapley Value

- Consider a set of players $D$, and a wealth-distribution (game) function $\mathcal{G}: \mathcal{P}(D) \longrightarrow \mathbb{R}$ ($\mathcal{P}(D)$ the power set of $D$)

- The Shapley value of player $p$ among a set of players $D$:

$$Shapley(D, \mathcal{G}, p) := \sum_{S \subseteq D \setminus \{p\}} \frac{|S|!(|D| - |S| - 1)!}{|D|!} (\mathcal{G}(S \cup \{p\}) - \mathcal{G}(S))$$

($|S|!(|D| - |S| - 1)!$ is number of permutations of $D$ with all players in $S$ coming first, then $p$, and then all the others)

# The Shapley Value

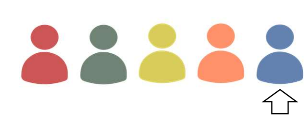
- Consider a set of players $D$, and a wealth-distribution (game) function $\mathcal{G}: \mathcal{P}(D) \longrightarrow \mathbb{R}$ ($\mathcal{P}(D)$ the power set of $D$)

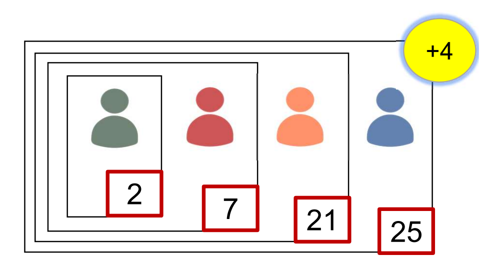
- The Shapley value of player $p$ among a set of players $D$:

$$Shapley(D, \mathcal{G}, p) := \sum_{S \subseteq D \setminus \{p\}} \frac{|S|!(|D| - |S| - 1)!}{|D|!} (\mathcal{G}(S \cup \{p\}) - \mathcal{G}(S))$$

($|S|!(|D| - |S| - 1)!$ is number of permutations of $D$ with all players in $S$ coming first, then $p$, and then all the others)

Expected contribution of player $p$ under all possible additions of $p$ to a partial random sequence of players followed by a random sequence of the rests of the players





- Shapley value is the only function that satisfy certain natural properties

  A result of a categorical set of axioms/conditions

- Shapley difficult to compute; provably #P-hard in general

- Counterfactual flavor: What happens having $p$ vs. not having it?

# Shapley as Score for QA

- Back to QA in DBs, players are tuples in DB $D$

  Boolean query $\mathcal{Q}$ becomes game function: for $S \subseteq D$

$$\mathcal{Q}(S) = \begin{cases} 1 & \text{if } S \models \mathcal{Q} \\ 0 & \text{if } S \not\models \mathcal{Q} \end{cases}$$

# Shapley as Score for QA

- Back to QA in DBs, players are tuples in DB $D$

  Boolean query $\mathcal{Q}$ becomes game function: for $S \subseteq D$

  $$\mathcal{Q}(S) = \begin{cases} 1 & \text{if } S \models \mathcal{Q} \\ 0 & \text{if } S \not\models \mathcal{Q} \end{cases}$$

- Concentrated on BCQs (and some aggregation on CQs)

  $$Shapley(D, \mathcal{Q}, \tau) := \sum_{S \subseteq D \setminus \{\tau\}} \frac{|S|!(|D|-|S|-1)!}{|D|!}(\mathcal{Q}(S \cup \{\tau\}) - \mathcal{Q}(S))$$

  Quantifies the contribution of tuple $\tau$ to query result (Livshits et al.; ICDT'20)

# Shapley as Score for QA

- Back to QA in DBs, players are tuples in DB $D$

  Boolean query $\mathcal{Q}$ becomes game function: for $S \subseteq D$

  $$\mathcal{Q}(S) = \begin{cases} 1 & \text{if } S \models \mathcal{Q} \\ 0 & \text{if } S \not\models \mathcal{Q} \end{cases}$$

- Concentrated on BCQs (and some aggregation on CQs)

  $$\mathit{Shapley}(D, \mathcal{Q}, \tau) := \sum_{S \subseteq D \setminus \{\tau\}} \frac{|S|!(|D|-|S|-1)!}{|D|!}(\mathcal{Q}(S \cup \{\tau\}) - \mathcal{Q}(S))$$

  Quantifies the contribution of tuple $\tau$ to query result (Livshits et al.; ICDT'20)

- So as with actual causality/responsibility, players (tuples) can split into endogenous and exogenous tuples

  E.g. the former are those in a specific table

  One wants to measure the contribution of endogenous tuples

- Dichotomy Theorem: $\mathcal{Q}$ BCQ without self-joins

  If $\mathcal{Q}$ hierarchical, then $Shapley(D, \mathcal{Q}, \tau)$ can be computed in PTIME

  Otherwise, the problem is $FP^{\#P}$-complete

- Dichotomy Theorem: $\mathcal{Q}$ BCQ without self-joins

  If $\mathcal{Q}$ hierarchical, then $Shapley(D, \mathcal{Q}, \tau)$ can be computed in PTIME

  Otherwise, the problem is $FP^{\#P}$-complete

- $\mathcal{Q}$ is hierarchical if for every two existential variables $x$ and $y$:

  - $Atoms(x) \subseteq Atoms(y)$, or
  - $Atoms(y) \subseteq Atoms(x)$, or
  - $Atoms(x) \cap Atoms(y) = \emptyset$

- Dichotomy Theorem: $\mathcal{Q}$ BCQ without self-joins

  If $\mathcal{Q}$ hierarchical, then $\mathit{Shapley}(D, \mathcal{Q}, \tau)$ can be computed in PTIME

  Otherwise, the problem is $FP^{\#P}$-complete

- $\mathcal{Q}$ is hierarchical if for every two existential variables $x$ and $y$:
  - $\mathit{Atoms}(x) \subseteq \mathit{Atoms}(y)$, or
  - $\mathit{Atoms}(y) \subseteq \mathit{Atoms}(x)$, or
  - $\mathit{Atoms}(x) \cap \mathit{Atoms}(y) = \emptyset$

Example: $\mathcal{Q}: \ \exists x \exists y \exists z (R(x, y) \wedge S(x, z))$

$\mathit{Atoms}(x) = \{R(x, y), \ S(x, z)\}, \ \mathit{Atoms}(y) = \{R(x, y)\}, \ \mathit{Atoms}(z) = \{S(x, z)\}$

Hierarchical!

Example: $\mathcal{Q}^{nh}: \ \exists x \exists y (R(x) \wedge S(x, y) \wedge T(y))$

$\mathit{Atoms}(x) = \{R(x), \ S(x, y)\}, \ \mathit{Atoms}(y) = \{S(x, y), T(y)\}$ Not hierarchical!

- Same criteria as for QA over probabilistic DBs (Dalvi & Suciu; 2004)
- Positive case: reduced to counting subsets of $D$ of fixed size that satisfy $\mathcal{Q}$

  A dynamic programming approach works
- Negative case: requires a fresh approach (not from probabilistic DBs)

  Use query $\mathcal{Q}^{nh}$ above

  Reduction from counting independent sets in a bipartite graph

- Same criteria as for QA over probabilistic DBs (Dalvi & Suciu; 2004)
- Positive case: reduced to counting subsets of $D$ of fixed size that satisfy $\mathcal{Q}$

  A dynamic programming approach works
- Negative case: requires a fresh approach (not from probabilistic DBs)

  Use query $\mathcal{Q}^{nh}$ above

  Reduction from counting independent sets in a bipartite graph
- Dichotomy extends to summation over CQs; same conditions and cases

  Shapley value is an expectation, that is linear
- Hardness extends to aggregate non-hierarchical queries: max, min, avg
- What to do in hard cases?

- Approximation: For every fixed BCQ $\mathcal{Q}$, there is a multiplicative fully-polynomial randomized approximation scheme (FPRAS)

$$P(\tau \in D \mid \frac{Sh(D,\mathcal{Q},\tau)}{1+\epsilon} \leq A(\tau,\epsilon,\delta) \leq (1+\epsilon)Sh(D,\mathcal{Q},\tau)\}) \geq 1-\delta$$

Also applies to summations

- Approximation: For every fixed BCQ $\mathcal{Q}$, there is a multiplicative fully-polynomial randomized approximation scheme (FPRAS)

$$P(\tau \in D \mid \frac{Sh(D, \mathcal{Q}, \tau)}{1+\epsilon} \leq A(\tau, \epsilon, \delta) \leq (1+\epsilon) Sh(D, \mathcal{Q}, \tau)\}) \geq 1 - \delta$$

  Also applies to summations

- A related and popular score is the Bahnzhaf Power Index (order ignored)

$$Banzhaf(D, \mathcal{Q}, \tau) := \frac{1}{2^{|D|-1}} \cdot \sum_{S \subseteq (D \setminus \{\tau\})} (\mathcal{Q}(S \cup \{\tau\}) - \mathcal{Q}(S))$$

  Bahnzhaf also difficult to compute; provably #P-hard in general

- We proved "Causal Effect" coincides with the Banzhaf Index (op. cit.)

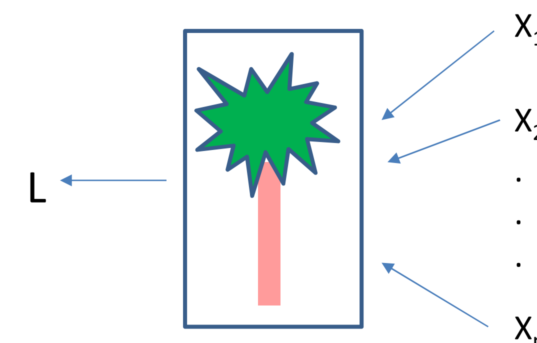
# Score-Based Explanations for Classification


$\mathbf{e} = \langle x_1, \ldots, x_n \rangle$ entity requesting a loan

- Black-box binary classification model returns label $L(\mathbf{e}) = 1$, i.e. rejected

  Why???!!!

- Similarly if we have the model, e.g. a classification tree or a logistic regression model



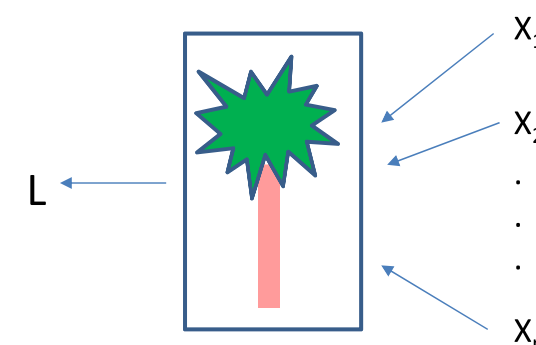
# Score-Based Explanations for Classification


$\mathbf{e} = \langle x_1, \ldots, x_n \rangle$ entity requesting a loan

- Black-box binary classification model returns label $L(\mathbf{e}) = 1$, i.e. rejected

  Why???!!!

- Similarly if we have the model, e.g. a classification tree or a logistic regression model




- Which feature values $x_i$ contribute the most?

  Assign numerical scores to feature values in $\mathbf{e}$

  Capturing the relevance of the feature value for the outcome

- In general (but not always) they are based on counterfactual interventions

- Some scores can be applied with both black-box and open models

  E.g. Shapley $\rightsquigarrow$ SHAP has become popular (Lee& Lundberg; 2017)

- Some scores can be applied with both black-box and open models

  E.g. Shapley $\rightsquigarrow$ SHAP has become popular (Lee& Lundberg; 2017)

  - Players are feature values in $\mathbf{e}$: $D = \{x_i = F_i(\mathbf{e}) \mid$ for some $F_i \in \mathcal{F}\}$
  - Game function: $\mathcal{G}_{\mathbf{e}}(S) := \mathbb{E}(L(\mathbf{e}') \mid \mathbf{e}'_S = \mathbf{e}_S)$ ($\mathbf{e}_S$: projection on $S$)
  - For a feature $F \in \mathcal{F}$, compute: $\mathit{Shapley}(\mathcal{F}, \mathcal{G}_{\mathbf{e}}, F(\mathbf{e}))$

- Some scores can be applied with both black-box and open models

  E.g. Shapley $\rightsquigarrow$ SHAP has become popular (Lee& Lundberg; 2017)

  - Players are feature values in $\mathbf{e}$: $D = \{x_i = F_i(\mathbf{e}) \mid \text{ for some } F_i \in \mathcal{F}\}$
  - Game function: $\mathcal{G}_{\mathbf{e}}(S) := \mathbb{E}(L(\mathbf{e}') \mid \mathbf{e}'_S = \mathbf{e}_S)$ ($\mathbf{e}_S$: projection on $S$)
  - For a feature $F \in \mathcal{F}$, compute: $Shapley(\mathcal{F}, \mathcal{G}_{\mathbf{e}}, F(\mathbf{e}))$

- This requires computing

  $\sum_{S \subseteq D \setminus \{F(\mathbf{e})\}} \frac{|S|!(|D|-|S|-1)!}{|D|!} (\mathbb{E}(L(\mathbf{e}'|\mathbf{e}'_{S \cup \{F(\mathbf{e})\}} = \mathbf{e}_{S \cup \{F(\mathbf{e})\}}) - \mathbb{E}(L(\mathbf{e}')|\mathbf{e}'_S = \mathbf{e}_S))$

  Assuming one has the probability space of possible entities $\mathbf{e}'$

  Then $L$ acts as a Bernoulli random variable

  Using the classifier many times, and computing the weighted averages

- In practice? (we'll be back ...)

# Yet Another Score: RESP

- Same classification setting (Bertossi, Li, Schleich, Suciu, Vagena; DEEM@SIGMOD'20)
- COUNTER$(\mathbf{e}, F) := L(\mathbf{e}) - \mathbb{E}(L(\mathbf{e}') \mid \mathbf{e}'_{\mathcal{F} \smallsetminus \{F\}} = \mathbf{e}_{\mathcal{F} \smallsetminus \{F\}})$, $F \in \mathcal{F}$

  This score can be applied to same scenarios, it is easy to compute

  Gives reasonable results, intuitively and in comparison to other scores

# Yet Another Score: RESP

- Same classification setting (Bertossi, Li, Schleich, Suciu, Vagena; DEEM@SIGMOD'20)

- $\mathsf{COUNTER}(\mathbf{e}, F) := L(\mathbf{e}) - \mathbb{E}(L(\mathbf{e}') \mid \mathbf{e}'_{\mathcal{F}\smallsetminus\{F\}} = \mathbf{e}_{\mathcal{F}\smallsetminus\{F\}}), \quad F \in \mathcal{F}$

  This score can be applied to same scenarios, it is easy to compute

  Gives reasonable results, intuitively and in comparison to other scores

- So as with SHAP: underlying probability space? (if any)

  No need to access the internals of the classification model

# Yet Another Score: RESP

- Same classification setting (Bertossi, Li, Schleich, Suciu, Vagena; DEEM@SIGMOD'20)
- $\textsf{COUNTER}(\mathbf{e}, F) := L(\mathbf{e}) - \mathbb{E}(L(\mathbf{e}') \mid \mathbf{e}'_{\mathcal{F}\smallsetminus\{F\}} = \mathbf{e}_{\mathcal{F}\smallsetminus\{F\}}), \quad F \in \mathcal{F}$

  This score can be applied to same scenarios, it is easy to compute

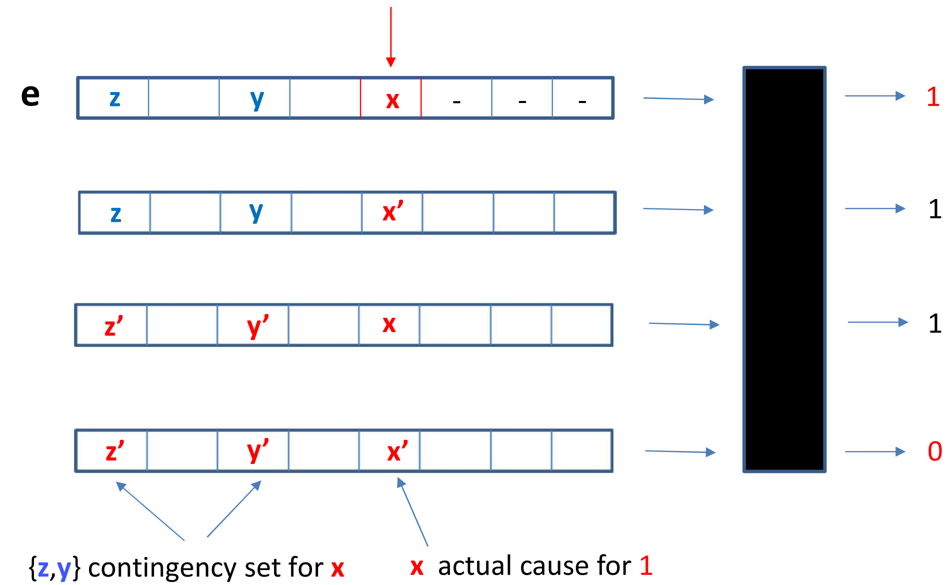
  Gives reasonable results, intuitively and in comparison to other scores
- So as with SHAP: underlying probability space? (if any)

  No need to access the internals of the classification model
- One problem: changing one value may not switch the label

  No explanations are obtained
- Extend this score bringing in contingency sets of feature values!

  The RESP-score (c.f. paper, simplified version follows)

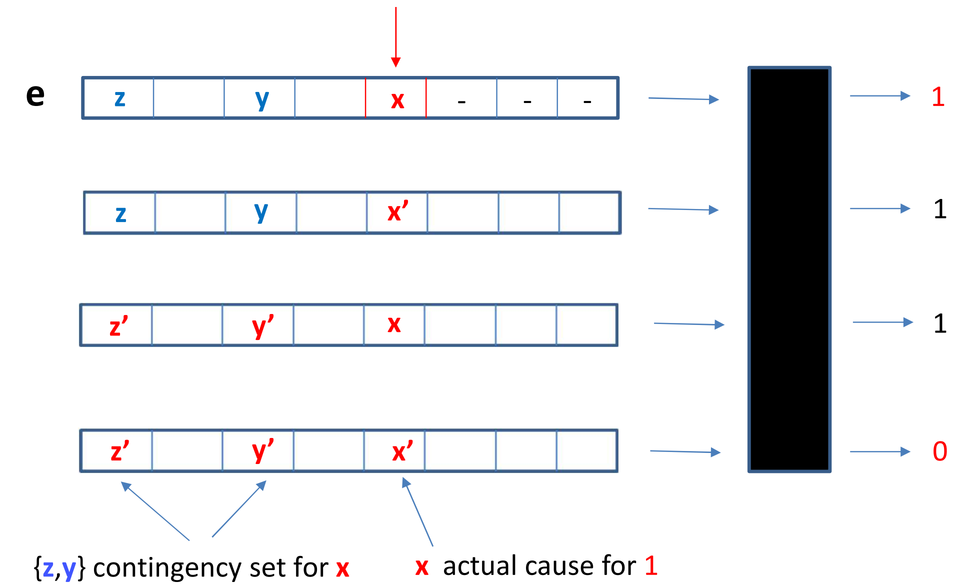
- Want explanation for classification "$1$" for $\mathbf{e}$
- Through interventions, changes of feature values, try to change it to "0"
- Fix a feature value $\mathbf{x} = F(\mathbf{e})$



- Want explanation for classification "1" for $\mathbf{e}$

- Through interventions, changes of feature values, try to change it to "0"

- Fix a feature value $\mathbf{x} = F(\mathbf{e})$

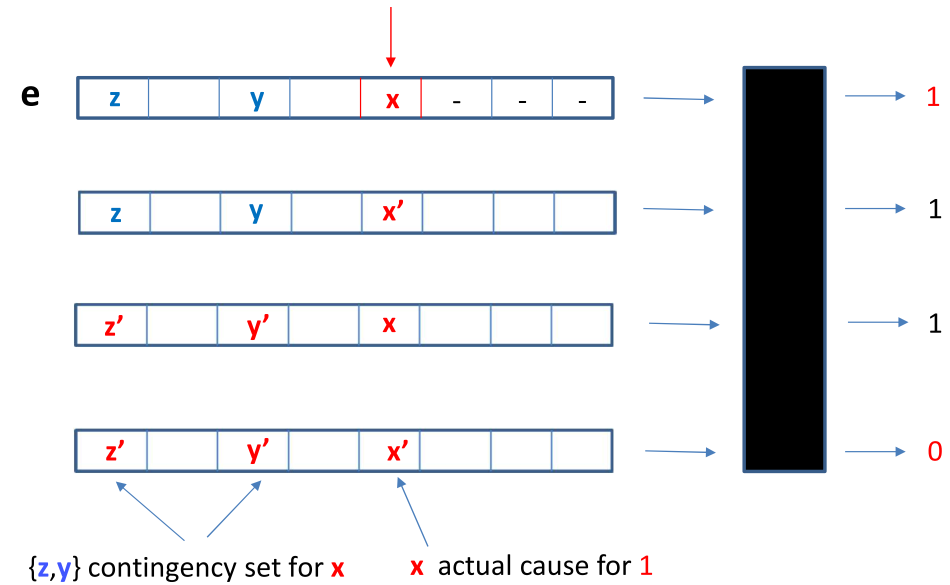



- $\mathbf{x}$ counterfactual explanation for $L(\mathbf{e}) = 1$ if $L(\mathbf{e}\frac{\mathbf{x}}{\mathbf{x}'}) = 0$, for $\mathbf{x}' \in Dom(F)$

- $\mathbf{x}$ actual explanation for $L(\mathbf{e}) = 1$ if there is a set of values $\mathbf{Y}$ in $\mathbf{e}$, $\mathbf{x} \notin \mathbf{Y}$, and (all) different values $\mathbf{Y}' \cup \{\mathbf{x}'\}$:

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

- Want explanation for classification "1" for $\mathbf{e}$
- Through interventions, changes of feature values, try to change it to "0"
- Fix a feature value $\mathbf{x} = F(\mathbf{e})$




- $\mathbf{x}$ counterfactual explanation for $L(\mathbf{e}) = 1$ if $L(\mathbf{e}\frac{\mathbf{x}}{\mathbf{x}'}) = 0$, for $\mathbf{x}' \in Dom(F)$
- $\mathbf{x}$ actual explanation for $L(\mathbf{e}) = 1$ if there is a set of values $\mathbf{Y}$ in $\mathbf{e}$, $\mathbf{x} \notin \mathbf{Y}$, and (all) different values $\mathbf{Y}' \cup \{\mathbf{x}'\}$:

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

- If $\mathbf{Y}$ is minimum in size, $\text{RESP}(\mathbf{x}) := \frac{1}{1+|\mathbf{Y}|}$ (can be formulated with expected values)

Example:

$\mathcal{C}$

| entity (id) | $F_1$ | $F_2$ | $F_3$ | $L$ |
|---|---|---|---|---|
| $\mathbf{e}_1$ | 0 | 1 | 1 | 1 |
| $\mathbf{e}_2$ | 1 | 1 | 1 | 1 |
| $\mathbf{e}_3$ | 1 | 1 | 0 | 1 |
| $\mathbf{e}_4$ | **1** | **0** | 1 | 0 |
| $\mathbf{e}_5$ | 1 | 0 | 0 | 1 |
| $\mathbf{e}_6$ | 0 | 1 | 0 | 1 |
| $\mathbf{e}_7$ | 0 | **0** | 1 | 0 |
| $\mathbf{e}_8$ | 0 | 0 | 0 | 0 |

Example:

$\mathcal{C}$

| entity (id) | $F_1$ | $F_2$ | $F_3$ | $L$ |
|---|---|---|---|---|
| $\mathbf{e}_1$ | 0 | 1 | 1 | 1 |
| $\mathbf{e}_2$ | 1 | 1 | 1 | 1 |
| $\mathbf{e}_3$ | 1 | 1 | 0 | 1 |
| $\mathbf{e}_4$ | **1** | **0** | 1 | 0 |
| $\mathbf{e}_5$ | 1 | 0 | 0 | 1 |
| $\mathbf{e}_6$ | 0 | 1 | 0 | 1 |
| $\mathbf{e}_7$ | 0 | **0** | 1 | 0 |
| $\mathbf{e}_8$ | 0 | 0 | 0 | 0 |

- Due to $\mathbf{e}_7$, $F_2(\mathbf{e}_1)$ is counterfactual explanation; with RESP$(F_2(\mathbf{e}_1)) = 1$
- Due to $\mathbf{e}_4$, $F_1(\mathbf{e}_1)$ is actual explanation; with $\{F_2(\mathbf{e}_1)\}$ as contingency set

  And RESP$(F_1(\mathbf{e}_1)) = \frac{1}{2}$

# Experiments and Foundations

- We compared COUNTER, RESP, SHAP, Banzhaf

  Kaggle loan data set, and XGBoost with Python library for classification model (opaque enough)

# Experiments and Foundations

- We compared COUNTER, RESP, SHAP, Banzhaf

  Kaggle loan data set, and XGBoost with Python library for classification model (opaque enough)

- Also comparison with Rudin's FICO-Score: model dependent, open model

  Uses outputs and coefficients of two nested logistic-regression models

  Model designed for FICO data; so, we used FICO data

# Experiments and Foundations

- We compared COUNTER, RESP, SHAP, Banzhaf

  Kaggle loan data set, and XGBoost with Python library for classification model (opaque enough)

- Also comparison with Rudin's FICO-Score: model dependent, open model

  Uses outputs and coefficients of two nested logistic-regression models

  Model designed for FICO data; so, we used FICO data

- Here we are interested more in the experimental setting than in results themselves

- RESP score: appealed to "product probability space": for $n$, say, binary features
  - $\Omega = \{0,1\}^n$, $T \subseteq \Omega$ a sample
  - $p_i = P(F_i = 1) \approx \frac{|\{\omega \in T \mid \omega_i = 1\}|}{|T|} =: \hat{p}_i$ (estimation of marginals)
  - Product distribution over $\Omega$:

$$P(\omega) := \Pi_{\omega_i=1} \hat{p}_i \times \Pi_{\omega_j=0} (1 - \hat{p}_j), \quad \text{for } \omega \in \Omega$$

- RESP score: appealed to "product probability space": for $n$, say, binary features
  - $\Omega = \{0,1\}^n$, $T \subseteq \Omega$ a sample
  - $p_i = P(F_i = 1) \approx \frac{|\{\omega \in T \ \mid \ \omega_i = 1\}|}{|T|} =: \hat{p}_i$ (estimation of marginals)
  - Product distribution over $\Omega$:

    $$P(\omega) := \Pi_{\omega_i = 1} \hat{p}_i \times \Pi_{\omega_j = 0} (1 - \hat{p}_j), \quad \text{for } \omega \in \Omega$$
- Not very good at capturing feature correlations
- RESP score computation for $\mathbf{e} \in \Omega$:
  - Expectations relative to product probability space
  - Choose values for interventions from feature domains, as determined by $T$
  - Call the classifier
  - Restrict contingency sets to, say, two features

- SHAP score appealed to "empirical probability space"

- Computing it on the product probability space is $\#P$-hard (c.f. the paper)

- SHAP score appealed to “empirical probability space”
- Computing it on the product probability space is $\#P$-hard (c.f. the paper)
- Use sample $T \subseteq \Omega$, test data

  Labels $L(\omega)$, $\omega \in T$, computed with learned classifier
- Empirical distribution: $P(\omega) := \begin{cases} \frac{1}{|T|} & \text{if } \omega \in T \\ 0 & \text{if } \omega \notin T \end{cases}$ ,for $\omega \in \Omega$

- SHAP score appealed to "empirical probability space"
- Computing it on the product probability space is $\#P$-hard (c.f. the paper)
- Use sample $T \subseteq \Omega$, test data

  Labels $L(\omega)$, $\omega \in T$, computed with learned classifier
- Empirical distribution: $P(\omega) := \begin{cases} \frac{1}{|T|} & \text{if } \omega \in T \\ 0 & \text{if } \omega \notin T \end{cases}$ ,for $\omega \in \Omega$
- SHAP value with expectations over this space, directly over data/labels in $T$
- The empirical distribution is not suitable for the RESP score (c.f. the paper)

# Final Remarks

- Explainable AI (XAI) is an effervescent area of research

  Its relevance can only grow

  Legislation around explainability, transparency and fairness of AI/ML systems

# Final Remarks

- Explainable AI (XAI) is an effervescent area of research

  Its relevance can only grow

  Legislation around explainability, transparency and fairness of AI/ML systems

- Different approaches and methodologies

  Causality, counterfactuals and scores have relevant role to play

# Final Remarks

- Explainable AI (XAI) is an effervescent area of research

  Its relevance can only grow

  Legislation around explainability, transparency and fairness of AI/ML systems

- Different approaches and methodologies

  Causality, counterfactuals and scores have relevant role to play

- Much research needed on the use of contextual, semantic and domain knowledge

  Some approaches are more appropriate, e.g. declarative (Bertossi; RuleML+RR'20)

# Final Remarks

- Explainable AI (XAI) is an effervescent area of research

  Its relevance can only grow

  Legislation around explainability, transparency and fairness of AI/ML systems

- Different approaches and methodologies

  Causality, counterfactuals and scores have relevant role to play

- Much research needed on the use of contextual, semantic and domain knowledge

  Some approaches are more appropriate, e.g. declarative (Bertossi; RuleML+RR'20)

- Still fundamental research is needed on what is a good explanation

  And the desired properties of an explanation score

  Shapley originally emerged from a list of *desiderata*