skema BUSINESS SCHOOL | AI R&D CENTER CANADA

From Database Repairs to Causality in Databases

and Beyond

Leopoldo Bertossi

leopoldo.bertossi@skema.edu

BDA, Clermont-Ferrand, Oct. 2022

Explanations in Databases

Receives	$R.1$	$R.2$
	s_2	s_1
	s_3	s_3
	s_4	s_3

Store	$S.1$
	s_2
	s_3
	s_4

- Query: Are there pairs of official stores in a receiving relationship?
- $\mathcal{Q}$: $\exists x \exists y (Store(x) \wedge Receives(x, y) \wedge Store(y))$

  The query is true in D: $D \models \mathcal{Q}$
- What tuples "cause" the query to be true?
- How strong are they as causes?
- We expect tuples $Receives(s_3, s_3)$ and $Receives(s_4, s_3)$ to be "causes"
- Explanations for a query result ...

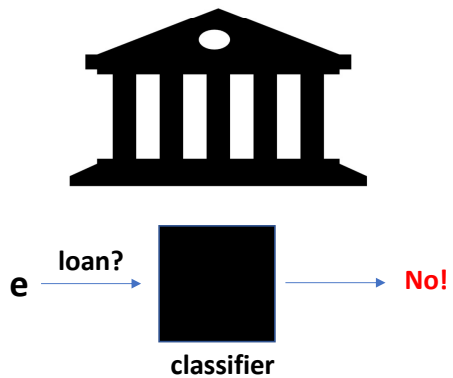
- Explanations for violation of semantic conditions (integrity constraints), etc.
- A DB system could provide *explanations*
- Explanations come in different forms
- Some of them are *causal explanations*
- Want to model, specify and compute causality
- Large part of our recent research is about the use of causality from different perspectives

  In data management and machine learning

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?

A Score-Based Approach: Responsibility

- Causality has been developed in AI for three decades or so
- In particular: Actual Causality
- Also the quantitative notion of Responsibility: a measure of causal contribution
- Both based on Counterfactual Interventions
- Hypothetical changes of values in a causal model to detect other changes: *"What would happen if we change ..."?*

  By so doing identify actual causes
- Do changes of feature values make the label change to "Yes"?
- We have investigated actual causality and responsibility in data management and ML-based classification
- Semantics, computational mechanisms, intrinsic complexity, logic-based specifications, reasoning, etc.

- There are other *local explanation scores*

  Also called "attribution scores"

- Assign numbers to, e.g., database tuples or features values to capture their causal, or, more generally, explanatory strength

- Some of them (in data management or ML)
  - Responsibility
  - The Causal Effect score
  - The Shapley value (as Shap in ML)

This Presentation

1. Causality in DBs
2. The DB repair connection
3. Responsibility
4. Causality under integrity constraints
5. Causal responsibility vs. causal effect
6. Shapley value in DBs
7. Responsibility of explanations for classification
8. Final remarks: The need for reasoning

Causality in DBs

- Causal explanations for a query result: (Meliou et al., 2010)
  - A relational instance D and a boolean conjunctive $\mathcal{Q}$
  - A tuple $\tau \in D$ is a counterfactual cause for $\mathcal{Q}$ if $D \models \mathcal{Q}$ and $D \setminus \{\tau\} \not\models \mathcal{Q}$
  - A tuple $\tau \in D$ is an actual cause for $\mathcal{Q}$ if there is a contingency set $\Gamma \subseteq D$, such that τ is a counterfactual cause for $\mathcal{Q}$ in $D \setminus \Gamma$ (Halpern and Pearl, 2001)
- The responsibility of an actual cause τ for $\mathcal{Q}$:

$$\rho_D(\tau) := \frac{1}{|\Gamma| + 1}, \; |\Gamma| = \text{size of smallest contingency set for } \tau$$

(0 otherwise)

- High responsibility tuples provide more interesting explanations (Chockler and Halpern, 2004)

Example

- Database D with relations R and S below

  $\mathcal{Q}$: $\exists x \exists y (S(x) \wedge R(x, y) \wedge S(y))$ Here: $D \models \mathcal{Q}$

R	A	B
	a_4	a_3
	a_2	a_1
	a_3	a_3

S	A
	a_4
	a_2
	a_3

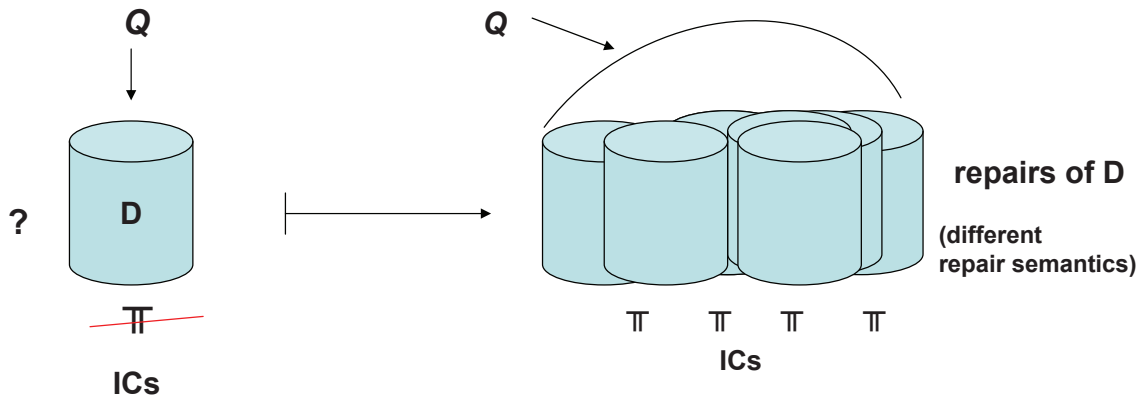
- Causes for $\mathcal{Q}$ to be true in D?
- $S(a_3)$ is counterfactual cause for $\mathcal{Q}$:
  If $S(a_3)$ is removed from D, $\mathcal{Q}$ is no longer an answer
- Its responsibility is $1 = \frac{1}{1+|\emptyset|}$
- $R(a_4, a_3)$ is an actual cause for $\mathcal{Q}$ with contingency set $\{R(a_3, a_3)\}$

  If $R(a_3, a_3)$ is removed from D, $\mathcal{Q}$ is still true, but further removing $R(a_4, a_3)$ makes $\mathcal{Q}$ false
- Responsibility of $R(a_4, a_3)$ is $\frac{1}{2} = \frac{1}{1+1}$
  Its smallest contingency sets have size 1
- $R(a_3, a_3)$ and $S(a_4)$ are actual causes, with responsibility $\frac{1}{2}$

Computational Problems

- Among many of them:
  - Computing causes
  - Deciding if a tuple is a cause
  - Computing responsibilities
  - Computing most responsible causes (MRC)
  - Deciding if a tuple has responsibility above a threshold
- Rather complete complexity picture for CQs and UCQs
- Obtained mostly via connection between: (B. & Salimi, 2017)
  - causality and database repairs, and
  - causality and consistency-based diagnosis

Database Repairs



(Arenas et al., PODS 99)

Example: Denial constraints (DCs) (in particular, FDs)

$\neg\exists x\exists y(P(x) \wedge Q(x,y))$

$\neg\exists x\exists y(P(x) \wedge R(x,y))$

P	A
	a
	e

Q	A	B
	a	b

R	A	C
	a	c

- Subset-repairs (S-repairs): (maximal consistent subinstance)

  $D_1 = \{P(e), Q(a,b), R(a,c)\}$ $D_2 = \{P(e), P(a)\}$

- Cardinality-repairs (C-repairs): (max-cardinality consistent subinstance)

  D_1

The Repair/Causality Connection

- BCQ: $\mathcal{Q}$: $\exists \bar{x}(P_1(\bar{x}_1) \wedge \cdots \wedge P_m(\bar{x}_m))$ and $\mathcal{Q}$ is true in D
  What are the causes for $\mathcal{Q}$ to be true?
- Obtain actual causes and contingency sets from DB repairs
- $\neg\mathcal{Q}$ is logically equivalent to DC

  $$\kappa(\mathcal{Q}): \neg\exists \bar{x}(P_1(\bar{x}_1) \wedge \cdots \wedge P_m(\bar{x}_m))$$

- $\mathcal{Q}$ holds in D iff D inconsistent wrt. $\kappa(\mathcal{Q})$
- S-repairs associated to causes and minimal contingency sets
- C-repairs associated to causes, minimum contingency sets, and maximum responsibilities
- Database tuple τ is actual cause with subset-minimal contingency set $\Gamma \iff D \smallsetminus (\Gamma \cup \{\tau\})$ is S-repair
  In which case, its responsibility is $\frac{1}{1+|\Gamma|}$
- τ is actual cause with min-cardinality contingency set Γ
  $\iff D \smallsetminus (\Gamma \cup \{\tau\})$ is C-repair And τ is MRAC

Exploiting the Connection

- Causality problem (CP): Computing/deciding actual causes can be done in polynomial time in data for CQs and UCQs

  (Meliou et al., 2010; B&S, 2017)

- Most computational problems related to repairs, in particular, C-repairs, are provably hard (data complexity)

  (Lopatenko & B., 2007)

  Techniques and results for repairs can be leveraged

- Responsibility problem: Deciding if a tuple has responsibility above a certain threshold is NP-complete for UCQs (B&S, 2017)

- Computing $\rho_D(\tau)$ is $FP^{NP(log(n))}$-complete for BCQs
  The *functional* version of the responsibility problem

- Deciding if τ is a most responsible cause is $P^{NP(log(n))}$-complete for BCQs

Causality under Integrity Constraints

- Want to take ICs into account
  Instances obtained from D by tuple deletions should satisfy the ICs (B. & Salimi; 2017)
- In this case, we start assuming that $D \models \Sigma$
- For τ to be actual cause for $\mathcal{Q}(\bar{a})$, the contingency set Γ must satisfy:

$$D \smallsetminus \Gamma \models \Sigma \qquad D \smallsetminus \Gamma \models \mathcal{Q}(\bar{a})$$
$$D \smallsetminus (\Gamma \cup \{\tau\}) \models \Sigma \qquad D \smallsetminus (\Gamma \cup \{\tau\}) \not\models \mathcal{Q}(\bar{a})$$

- Responsibility $\rho^{D,\Sigma}_{\mathcal{Q}(\bar{a})}(\tau)$ defined as before

- Example: DB instance D and CQ, $\mathcal{Q}$ below

	Dep	*DName*	*TStaff*
t_1		Computing	John
t_2		Philosophy	Patrick
t_3		Math	Kevin

	Course	*CName*	*TStaff*	*DName*
t_4		COM08	John	Computing
t_5		Math01	Kevin	Math
t_6		HIST02	Patrick	Philosophy
t_7		Math08	Eli	Math
t_8		COM01	John	Computing

(A) $\mathcal{Q}(x)\colon \exists y \exists z(\mathit{Dep}(y,x) \wedge \mathit{Course}(z,x,y))$

$\langle \text{John} \rangle \in \mathcal{Q}(D)$

(a) t_1 is counterfactual

(b) t_4 with single minimal contingency set $\Gamma_1 = \{t_8\}$

(c) t_8 with single minimal contingency set $\Gamma_2 = \{t_4\}$

- Under IND ψ: $\forall x \forall y\ (\mathit{Dep}(x,y) \rightarrow \exists u\ \mathit{Course}(u,y,x))$
- t_4 t_8 not actual causes anymore: $D \smallsetminus \Gamma_1 \models \psi$, but (satisfied)
  $D \smallsetminus (\Gamma_1 \cup \{t_4\}) \not\models \psi$
- t_1 still is counterfactual cause

(B) $\mathcal{Q}_1(x)\colon\ \exists y\ \mathit{Dep}(y,x)$ $\langle \text{John} \rangle \in \mathcal{Q}_1(D)$

- Under IND: same causes as $\mathcal{Q}$: $\mathcal{Q} \equiv_\psi \mathcal{Q}_1$

(C) $\mathcal{Q}_2(x)$: $\exists y \exists z\, Course(z, x, y)$ $\langle \text{John} \rangle \in \mathcal{Q}_2(D)$

- W/O ψ: t_4 and t_8 only actual causes, with $\Gamma_1 = \{t_8\}$ and $\Gamma_2 = \{t_4\}$, resp.
- Under IND: t_4 and t_8 still actual causes
- Contingency sets?
- We lose Γ_1 and Γ_2

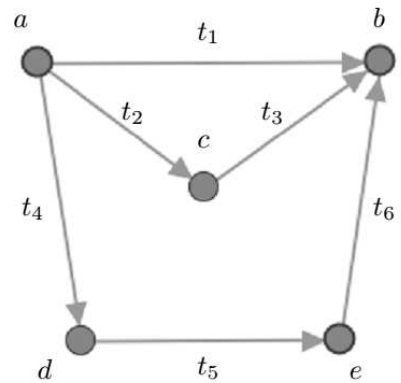
  $D \smallsetminus (\Gamma_1 \cup \{t_4\}) \not\models \psi$, $D \smallsetminus (\Gamma_2 \cup \{t_8\}) \not\models \psi$
- Smallest contingency set for t_4: $\Gamma_3 = \{t_8, t_1\}$
  Smallest contingency set for t_8: $\Gamma_4 = \{t_4, t_1\}$
- Responsibilities of t_4, t_8 decrease: $\rho^{D}_{\mathcal{Q}_2(\text{John})}(t_4) = \frac{1}{2}$, but $\rho^{D,\psi}_{\mathcal{Q}_2(\text{John})}(t_4) = \frac{1}{3}$
- t_1 is still not an actual cause, but affects the responsibility of actual causes

- Some Results:
  - Causes preserved under logical equivalence of queries under ICs
  - Without ICs, deciding causality for CQs is tractable, but their presence may make complexity grow
  - There are a CQ $\mathcal{Q}$ and an inclusion dependency ψ, for which deciding causality is NP-complete (B & S, 2017)
  - ASPs for computation of causes and responsibilities under ICs can be produced
- Beyond CQs:
  - What about causality for Datalog queries?
  - For Datalog queries, cause computation can be NP-complete
  - Through a connection to Datalog abduction (B. & S., 2017)

Causal Responsibility and Causal Effect

- Causal responsibility can be seen as an explanation score for database tuples in relation to query results
- It is not the only possible score
- Example: Boolean query Π is true 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



$$yes \leftarrow P(a,b)$$
$$P(x,y) \leftarrow E(x,y)$$
$$P(x,y) \leftarrow P(x,z), E(z,y)$$

- $E \cup \Pi \models yes$ (query in Datalog, also union of CQs)
- All tuples are actual causes: every tuple 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

- We proposed an alternative to the notion of causal responsibility: *Causal Effect* (Salimi et al., 2016)
- Causal responsibility has been questioned for other reasons and from different angles
- Retake question about how answer to query $\mathcal{Q}$ changes if τ is deleted/inserted from/into D
- An *intervention* on a *structural causal model*
- In this case provided by the the *lineage* of the query
- Example: $D = \{R(a,b), R(a,c), R(c,b), S(b), S(c)\}$

  BCQ $\mathcal{Q}: \exists x(R(x,y) \wedge S(y))$
- True in D, with lineage instantiated on D given by propositional formula:

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

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

- Want to quantify contribution of a tuple to a query answer
- Assign probabilities uniformly and independently to tuples in D

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	$\frac{1}{2}$

D^p Probabilistic database (tuples outside D get probability 0)

- The X_τ's become independent, identically distributed random variables; and $\mathcal{Q}$ is Bernouilli random variable
- What's the probability that $\mathcal{Q}$ takes a particular truth value when an intervention is done on D?
- Interventions of the form $do(X = x)$: In the *structural equations* make X take value x
- For $y, x \in \{0, 1\}$: $P(\mathcal{Q} = y \mid do(X_\tau = x))$?
- Corresponding to making X_τ false or true
- E.g. $do(X_{S(b)} = 0)$ leaves lineage in the form:

$$\Phi_{\mathcal{Q}}(D)\frac{X_{S(b)}}{0} := (X_{R(a,c)} \wedge X_{S(c)})$$

- The *causal effect* of τ:

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

- Example: (cont.) When X_τ is made false, probability that the instantiated lineage above becomes true in D^p:

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

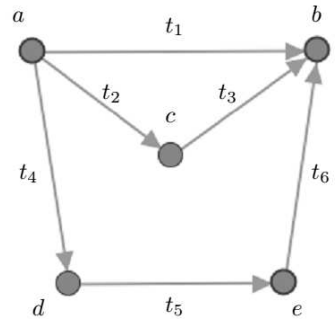
- When X_τ is made true, is probability of this lineage becoming true in D^p:

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

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

- $\mathbb{E}(\mathcal{Q} \mid do(X_{S(b)} = 0)) = P(\mathcal{Q} = 1 \mid do(X_{S(b)} = 0)) = \frac{1}{4}$
  $\mathbb{E}(\mathcal{Q} \mid do(X_{S(b)} = 1)) = \frac{13}{16}$
- $\mathcal{CE}^{D,\mathcal{Q}}(S(b)) = \frac{13}{16} - \frac{1}{4} = \frac{9}{16} > 0$, an actual cause with this causal effect!

- Example: (cont.) The Datalog query, 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 ...)

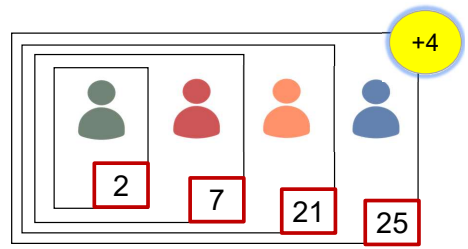
Coalition Games and the Shapley Value

- Initial motivation: By how much a database tuple contributes to the inconsistency of a DB? To the violation of ICs
- Similar ideas can be applied to the contribution to query results (Livshits et al., 2020)
- Usually *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
- Apply standard measures used in game theory: the Shapley value of tuple
- Implicitly based on counterfactual intervention: What would happen if we change ...?

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



- Database tuples and feature values can be seen as players in a coalition game

  Each of them contributing to a shared wealth function
- The Shapley value is a established measure of contribution by players to the wealth function
- It emerges as the only measure that enjoys certain desired properties
- For each game one defines an appropriate wealth or game function
- Shapley difficult to compute: #P-hard in general
- Evidence of difficulty: $\#SAT$ is #P-hard

  About counting satisfying assignments for propositional formulas

  At least as difficult as SAT

Shapley Values as Scores in DBs

- Database tuples can be seen as players in a coalition game
- Query $\mathcal{Q}$: $\exists x \exists y (Store(x) \wedge Receives(x, y) \wedge Store(y))$

  It takes values 0 or 1 in a database
- Game function becomes the value of the query
- A set of tuples make it true or not, with some possibly contributing more than others to making it true

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

- Quantifies the contribution of tuple τ to query result
- All possible permutations of subinstances of D
- Average of differences between having τ or not
- Counterfactuals implicitly involved and aggregated

- We investigated algorithmic, complexity and approximation problems
- A dichotomy theorem for Boolean CQs without self-joins

  Syntactic characterization: : PTIME vs. #P-hard
- Extended to aggregate queries
- It has been applied to measure contribution of tuples to inconsistency of a database
- Related and popular score: 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: #P-hard in general
- We proved "Causal Effect" coincides with the Banzhaf Index!

Explanations in AI

- The loan applicant classification example is representative of a more general problem with applications of AI systems
- 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 naturally emerged: *Ethical AI*

  Motivated by the need for more *transparent, trustable, fair, unbiased, ...* AI systems
- Also: *Interpretable* AI systems

It may really be a "black box"! $\longrightarrow$

Causality and XAI

- We have applied actual causality to explanations for outcomes from ML classification systems
- 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 have experimentally compared responsibility scores with other *local attribution scores*: *Causal-Effect*, *Shap*

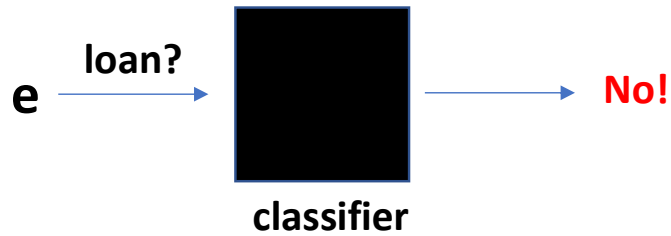
  And other scores based on (used with) "open models" (e.g. connected logistic regressions)

  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:

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

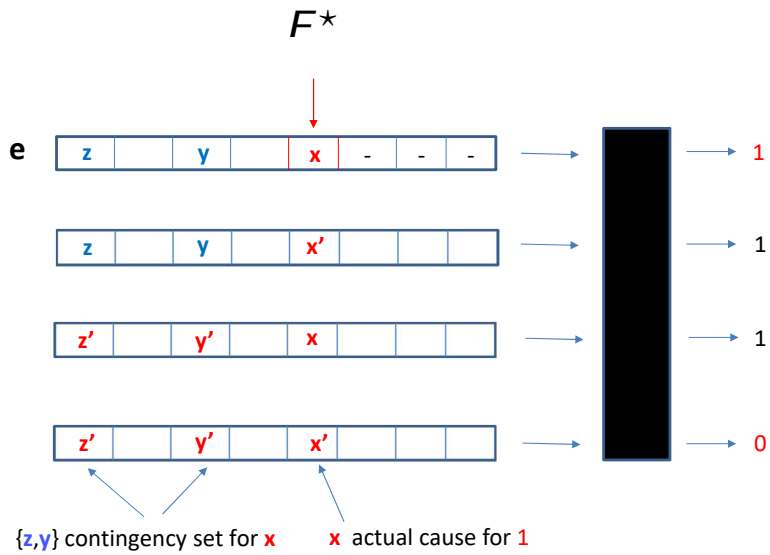
$$\mathbf{e}'' = \langle \text{john}, 18, \text{plumber}, 80\text{K}, \text{brooklyn}, \ldots \rangle \quad \text{Yes}$$

- 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 ...

x-$Resp(\mathbf{e}, F^\star)$ 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*

Final Remarks: 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.