CompSci 590.6

Understanding Data:

Theory and Applications

Lecture 16

Causality in Databases

Instructor: Sudeepa Roy

Email: *sudeepa@cs.duke.edu*

Today's Reading

Meliou-Gatterbauer-Moore-Suciu

PVLDB 2010

The Complexity of Causality and Responsibility for Query Answers and Non-Answers

Optional reading:

Meliou-Gatterbauer-Nath-Suciu

SIGMOD 2011

Tracing Data Errors with View-Conditioned Causality

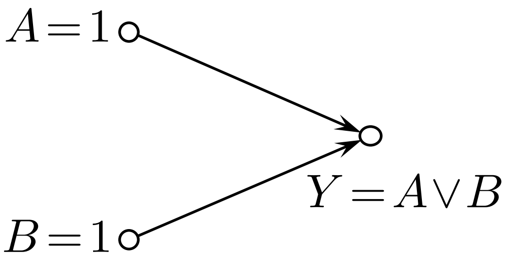
Acknowledgement:

Most of the slides in this lecture are originally due to Dr. Alexandra Meliou, University of Massachusetts-Amherst, and have been updated here

Review: Pearl's Structural Causal Model

Pearl' 2000

- Model M = (U, V, F)
  - *E.g., The house is burnt due to Fire A or Fire B*
- Endogenous variables U:
  - Variables within the model and are used as potential causes
  - Fire A reaches the house (A)
  - Fire B reaches the house (B)
  - The house is burnt (Y)



- Exogenous variables V:
  - Variables outside the model, not potential causes
  - Oxygen in the air, heavy rain
- Structural equations F
  - How endogenous variables are affected due to exogenous and other endogenous variables
  - Y = A v B

Halpern-Pearl'2001, Halpern-Pearl'2005

Review: Counterfactual vs. Actual Cause

Counterfactual Cause:

- If *not A* then *not φ*
  - In the absence of a cause, the effect doesn't occur

$$C = A \wedge B, \quad A = 1 \wedge B = 1 \longleftarrow$$ Both (A = 1) and (B = 1) are counterfactual for (C = 1)

Actual Cause:

- A variable X is an actual cause of an effect Y if there exists a contingency that makes X counterfactual for Y

$$C = A \vee B$$

$\uparrow$

(A = 1) is a cause of (C = 1) under the contingency B=0

A = 1, B = 1 $\Rightarrow$ C = 1	
A = 0, B = 1 $\Rightarrow$ C = 1	A alone does not change C
A = 0, B = 0 $\Rightarrow$ C = 0 and A = 1, B = 0 $\not\Rightarrow C = 0$	A changes C when B = 0 B = 1 to 0 does not change C

Chockler-Halpern' 2004

Review: Responsibility

$$\rho = \frac{1}{1 + \min_{\Gamma} |\Gamma|}$$

size of the contingency set

- Measures the "degree of causality"
  - Larger contingency implies a smaller degree of causality
- Counterfactual causes have the most contribution
  - empty contingency set

Example

$$Y = A \wedge (B \vee C)$$

A=1 is counterfactual for Y=1 (ρ=1)

B=1 is an actual cause for Y=1, with contingency C=0 (ρ=0.5)

Causality in Databases

- How to model the causal concepts from Pearl's model in terms of concepts in databases?
- i.e. model
  - Endogenous and exogenous variables
  - Actual and Counterfactual causes
  - Responsibility

  in terms of
  - database/relations/tuples
  - queries
  - lineage/provenance
- Why?
  - Responsibility of tuples will help in error tracing and explanations

Motivating example: IMDB dataset

IMDB Database Schema

Actor

aid	*firstName*	*lastName*

Director

did	*firstName*	*lastName*

Movie

mid	*name*	*year*	*rank*

Genre

mid	*genre*

Movie_Directors

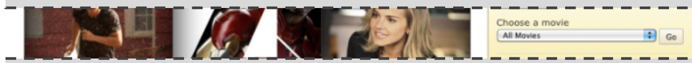
did	*mid*

Casts

aid	*mid*	*role*

Query

"What genres does Tim Burton direct?"



```
select    distinct g.genre
from      Director d, Movie_Directors md,
          Movie m, Genre g
where     d.lastName like 'Burton'
          and g. mid=m.mid
          and m. mid=md.mid
          and md. did=d.did
order by  g.genre
```

genre
...
Fantasy
History
Horror
Music
Musical
Mystery
Romance
...

?

What can databases do

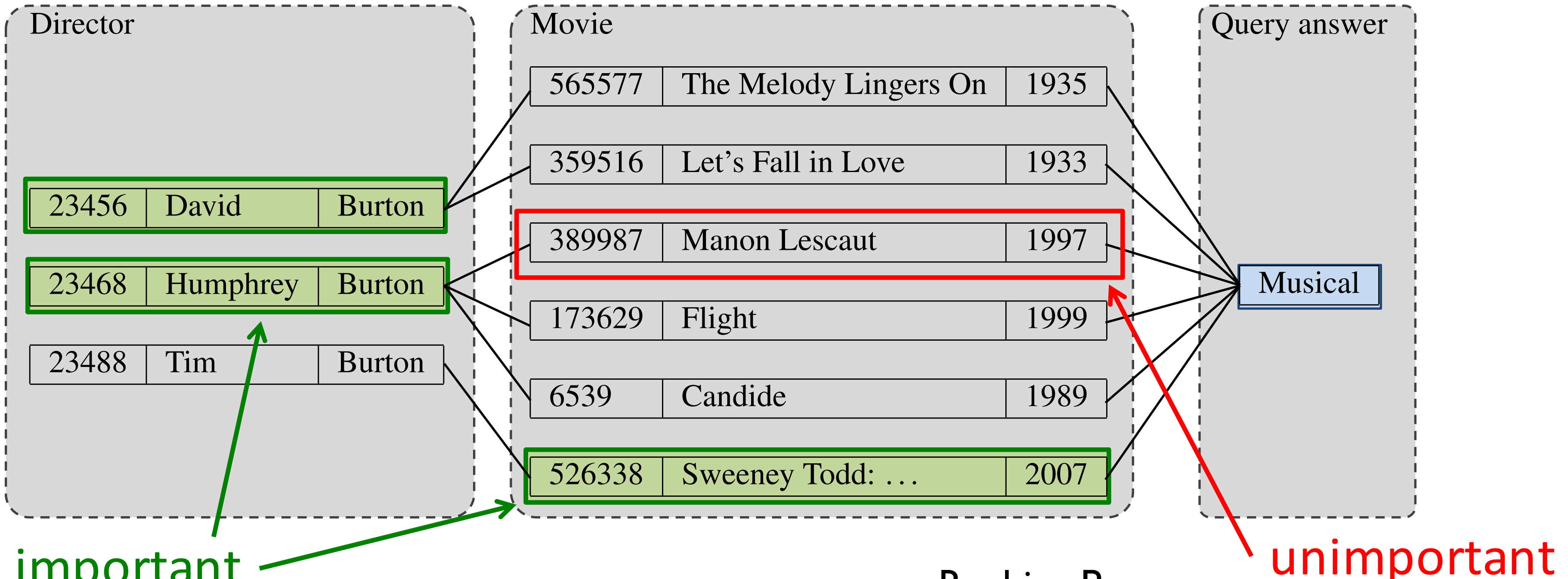
Provenance / Lineage:
The set of all tuples that contributed to a given output tuple

[Cheney et al. FTDB 2009], [Buneman et al. ICDT 2001], ...

But

In this example, the lineage includes **137 tuples !!**

From provenance to causality



Ranking Provenance

Answer tuple	ρ_t
Movie(526338, "Sweeney Todd", 2007)	0.33
Director(23456, David, Burton)	0.33
Director(23468, Humphrey, Burton)	0.33
Director(23488, Tim, Burton)	0.33
Movie(359516, "Let's Fall in Love", 1933)	0.25
Movie(565577, "The Melody Lingers On", 1935)	0.25
Movie(6539, "Candide", 1989)	0.20
Movie(173629, "Flight", 1999)	0.20
Movie(389987, "Manon Lescaut", 1997)	0.20

Goal:

Rank tuples in order of importance

- A cause of an answer/non-answer is an input tuple
- Rank them by their responsibility

Endogenous/exogenous tuples

Partition the data D into 2 groups:

$$D = D^{[\mathtt{n}]} \cup D^{[\mathtt{x}]}$$

- Exogenous tuples: $D^{[x]}$
  - tuples that we consider correct/verified/trusted
  - not potential causes
  - E.g. the *Genre*, and *Movie_Director* tables
- Endogenous tuples: $D^{[n]}$
  - Untrusted tuples, or simply of interest to the user
  - potential causes
  - E.g. the *Director* and *Movie* tables
- This division can be application-dependent and decided during the run time
  - e.g. set movie tuples with year > 2008 to be endogenous

Meliou et al.' 2010

Causality of a query answer

Input: database D and query Q. Output: D'=Q(D)

- $D^{[n]}$ endogenous tuples, $D^{[x]}$ exogenous tuples

- $t \in D^n$ is a counterfactual cause for answer α
  - If $\alpha \in Q(D)$ and $\alpha \notin Q(D-t)$

- $t \in D^n$ is an actual cause for answer α
  - If $\exists \Gamma \subset D^n$ such that t is counterfactual in $D - \Gamma$

contingency set

Example

Lineage expression:

$r_1 s_1 + r_2 s_1$

$= s_1 (r_1 + r_2)$

Query:

$$q :- R(x, a_3), S(a_3)$$

Boolean query
answer = true

Database:

R

	X	Y
	a_1	a_5
	a_2	a_1
r_1	a_3	a_3
r_2	$\cancel{a_4}$	$\cancel{a_3}$
	a_4	a_2

S

Y	
a_1	
a_2	
$\cancel{a_3}$	s_1
a_4	
a_6	

Assume all endogenous

Responsibility: $\rho_t = \dfrac{1}{1 + \min_\Gamma |\Gamma|}$

$$\rho_{s_1} = 1 \qquad \Gamma_{s_1} = \emptyset$$

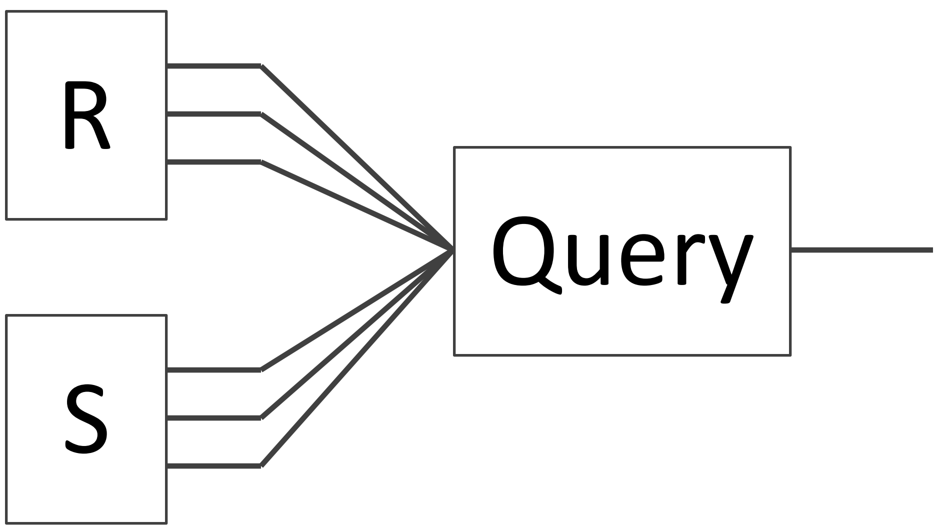
$$\rho_{r_2} = \frac{1}{2} \qquad \Gamma_{r_2} = \{r_1\}$$

NOTE: If r_1 is exogenous, r_2 is not a cause.

Causality for database queries

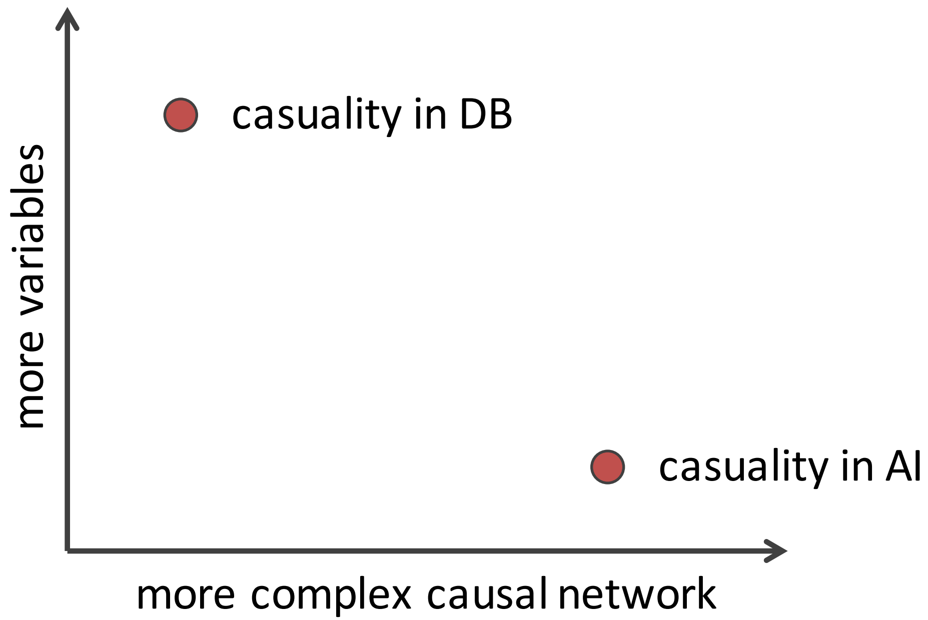
Input:	Database D and query Q
Output:	D'=Q(D)

- Causal network:
  - Lineage of the query



$$r_1 s_1 \vee r_2 s_1$$

Causality in AI vs. databases



So far “why-so” causality – explain an answer

Dual : “why-no” causality – explain a non-answer

Why-no causality

- Given database $D^{[x]}$
- Query answer $Q(D^{[x]})$
- Non-answer $p \notin Q(D^{[x]})$

- Real database $D = D^{[x]} \cup D^{[n]}$
  - $D^{[n]}$ = missing endogenous tuples (recall missing answers)

- Counterfactual cause $t \in D^{[n]}$
  - if $p \in Q(D^{[x]} \cup \{t\})$

- Actual cause t with contingency $\Gamma \subseteq D^{[n]}$
  - if t is a counterfactual cause for $D^{[x]} \cup \Gamma$

Problems to solve

Given $D = D^{[x]} \cup D^{[n]}$, query q, a potential answer/non-answer *p*

- Causality
  - Compute the set $C \subseteq D^{[n]}$ of actual causes for p
- Responsibility
  - For each actual cause $t \in C$, compute its responsibility

Consider Boolean query without loss of generality
e.g. q() :- R(x, y), S(y)

Causes: that can change “true” to “false”

Overview: Complexity Results

Causality	answers *Why So?*	non-answers *Why No?*
w/o SJ	PTIME (CQ)	PTIME (FO)
with SJ	PTIME (FO)	

Responsibility		*Why So?*	*Why No?*
w/o SJ	*linear*	PTIME	PTIME
	non-linear	NP-hard	
with SJ		NP-hard	

dichotomy

Data complexity

Problem 1: Causality

- Goal: compute all actual causes by a Boolean query q
- Let φ be the lineage (provenance) of q
- $\varphi^{[n]}$ = set all exogenous tuples to true (= 1) in φ
  - n-lineage
  - depends only on endogenous tuples
  - apply absorption: $r + rs = r$

Theorem:

The following three conditions are equivalent

1. An endogenous tuple t is an actual cause for q
2. There are endogenous tuples Γ such that
   - $\varphi\,[u = 0,\, u \in \Gamma]$ is satisfiable
   - $\varphi\,[u = 0,\, u \in \Gamma;\, t = 0]$ is unsatisfiable
3. There is a conjunct (after absorption) in $\varphi^{[n]}$ containing t

Example

Query:

$$q :- R(x, a_3), S(a_3)$$

Database:

R

	X	Y
r_1	a_1	a_5
r_2	a_2	a_1
r_3	a_3	a_3
r_4	a_4	a_3
r_5	a_4	a_2

S

Y	
a_1	s_1
a_2	s_2
a_3	s_3
a_4	s_4
a_6	s_5

Provenance/Lineage?

$\varphi = r_3s_3 + r_4s_3$

Ex 1: The set of actual cause

$C = \{r_3, r_4, s_3\}$

Ex 2: Suppose r_4 is exogenous

- Then $\varphi^{[n]}$

$= r_3s_3 + s_3$

$= s_3$ (absorption)

The only actual cause is

$C = \{s_3\}$

Further, the actual causes C can be computed by a SQL query

Responsibility: PTIME Queries

- Assume conjunctive queries with no self joins

$$q :- R(a, y)$$

- A simple case:

The lineage of q will be of the form:

$R(a, a) \vee R(a, b) \vee R(a, c) \vee \ldots$

What is the responsibility of $t = R(a, b)$

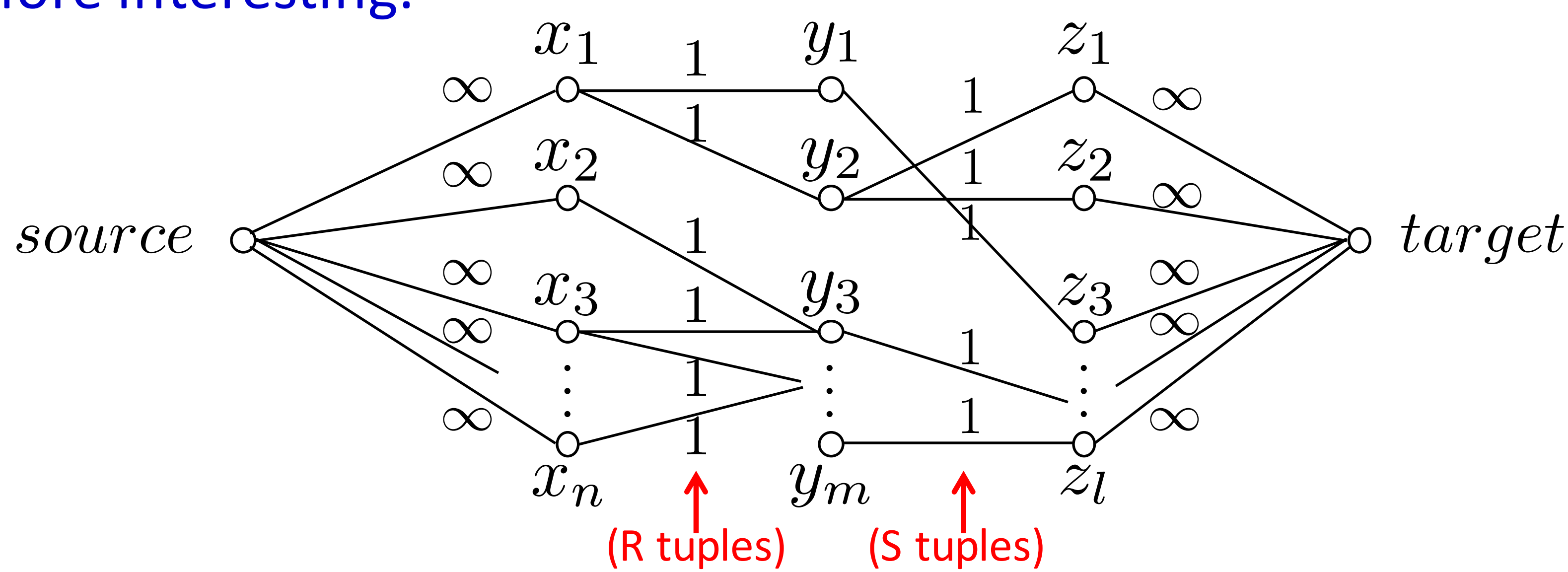
$$\Gamma_t = \{R(a, y) | y \neq b\}$$

PTIME

Responsibility: PTIME Queries

More interesting:

$$q :- R(x,y), S(y,z)$$



A cut in the graph : interrupts the s-t flow.
Min-cut : a cut with min capacity

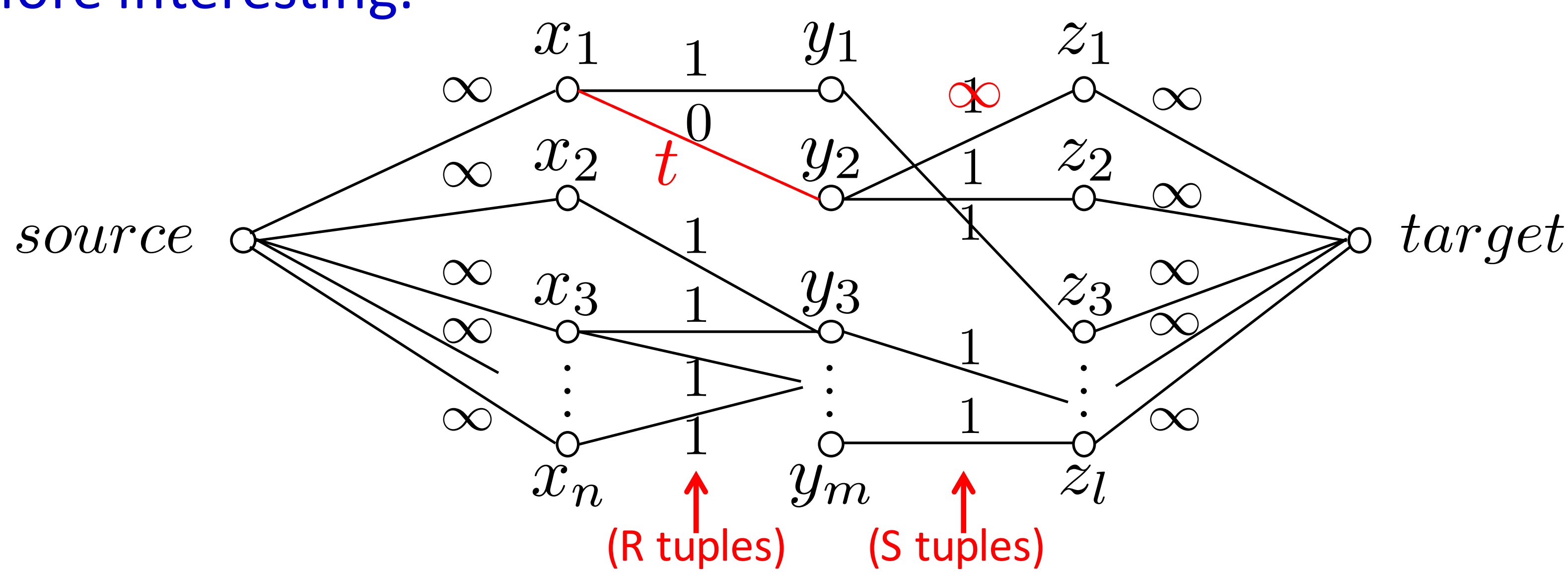
- can be computed in PTIME (e.g. Ford-Fulkerson)
- never includes the edges from s or to t (capacity = ∞)

Any mincut corresponds to a minimal set of tuples Γ' so that q is false on D – Γ'

Responsibility: PTIME Queries

More interesting:

$$q :- R(x,y), S(y,z)$$



To compute responsibility of t:

- The mincut Γ' must include t, i.e. Γ' = {t} ∪ Γ
- Set the capacity of t to 0

For all s-t paths p that go through t

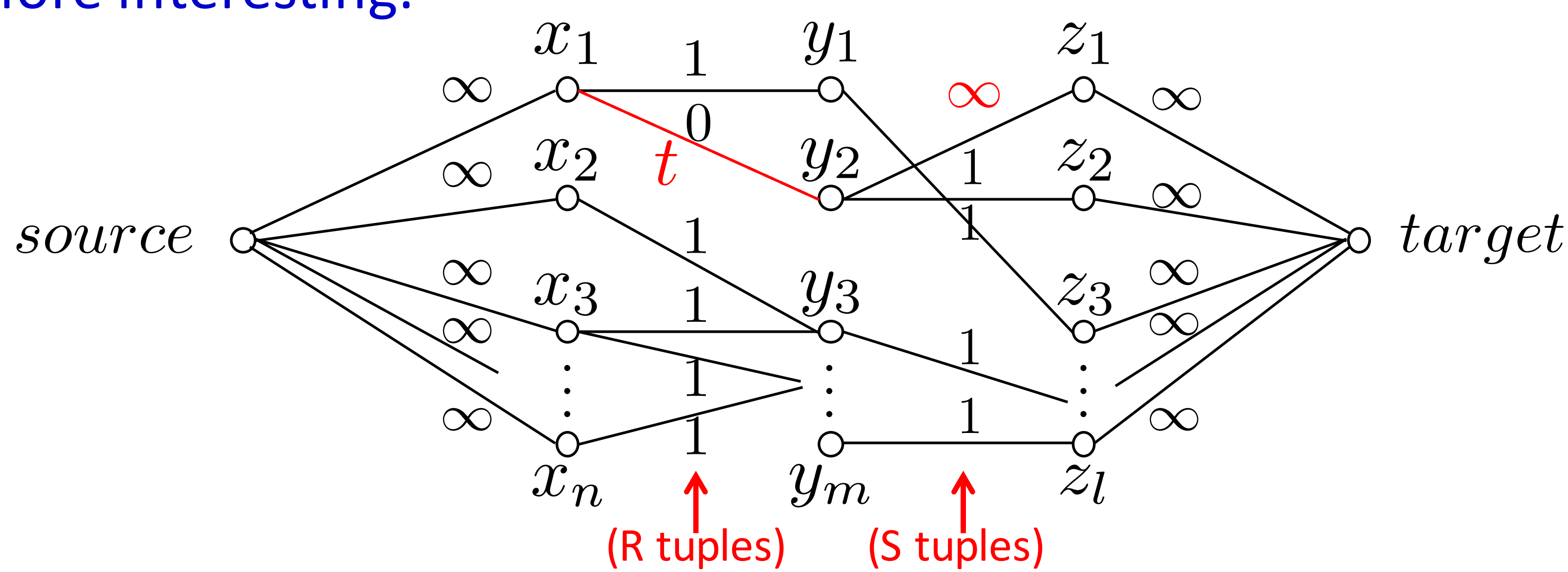
- set the capacities of all edges in p – {t} to ∞
- compute the size of the mincut
- reset the capacity back to 1
- here two paths $x_1y_2z_1$ and $x_1y_2z_2$

Poly-time?

Responsibility: PTIME Queries

More interesting:

$$q :- R(x, y), S(y, z)$$



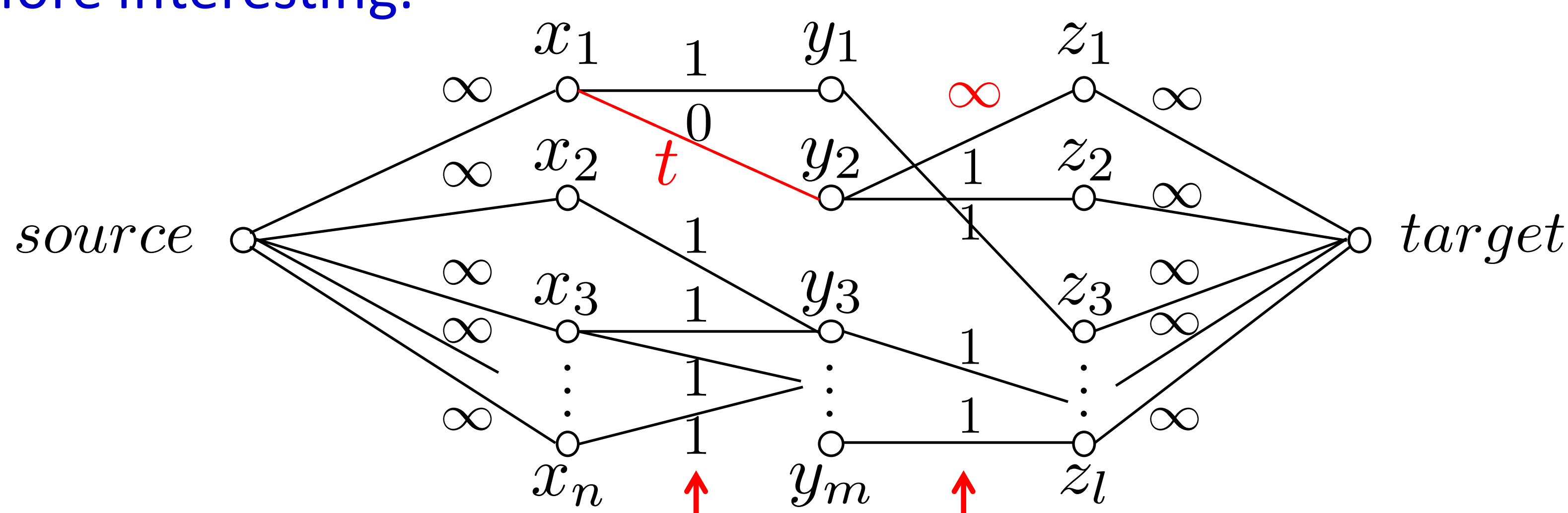
Claim: if Γ' is a mincut, Γ = Γ' – {t} is a contingency for t

- q is false on D - Γ'
  - s and t are disconnected
- q is true on D - Γ' ∪ {t}
  - Add t back, along with the edges in path p, a path from s to t is restored
  - the edges on p have ∞ capacity, cannot belong to Γ'

Responsibility: PTIME Queries

More interesting:

$$q : -R(x,y), S(y,z)$$



Claim: if Γ' is a mincut, Γ = Γ' – {t} is a contingency for t

Therefore, repeating over all paths, we can compute the minimum contingency set and responsibility for t

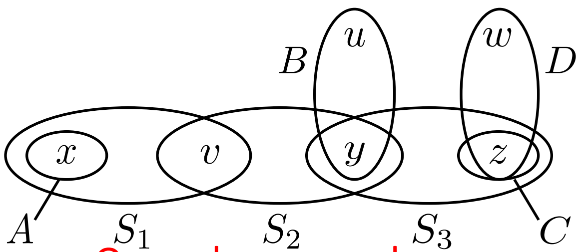
Q. what are other queries for which this trick works?

A. Linear queries

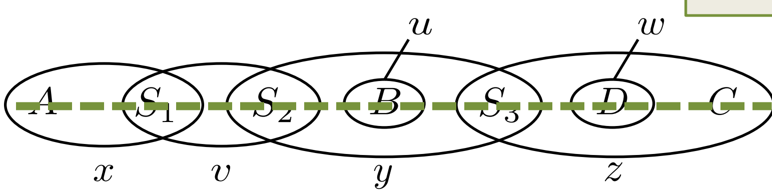
$$q :- R_1(x_1, x_2), R_2(x_2, x_3), R_3(x_3, x_4), \ldots$$

Linear Queries and
Query Dual Hypergraph

$$q :- A(x)S_1(x,v)S_2(v,y)B(y,u)S_3(y,z)D(z,w)C(z)$$



Query hypergraph



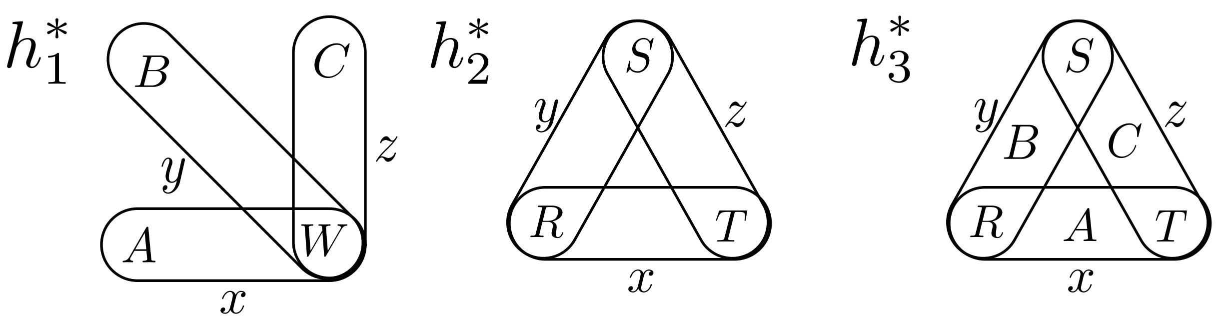
Query dual hypergraph

Definition: Linear Queries

There exists an ordering of the nodes (relation names) of the dual hypergraph, such that every hyperedge is a consecutive subsequence.

Theorem:

Computing responsibility for all linear queries is in PTIME.



None of these are linear

Responsibility: Hard Queries

Theorem: The following queries are NP-hard:

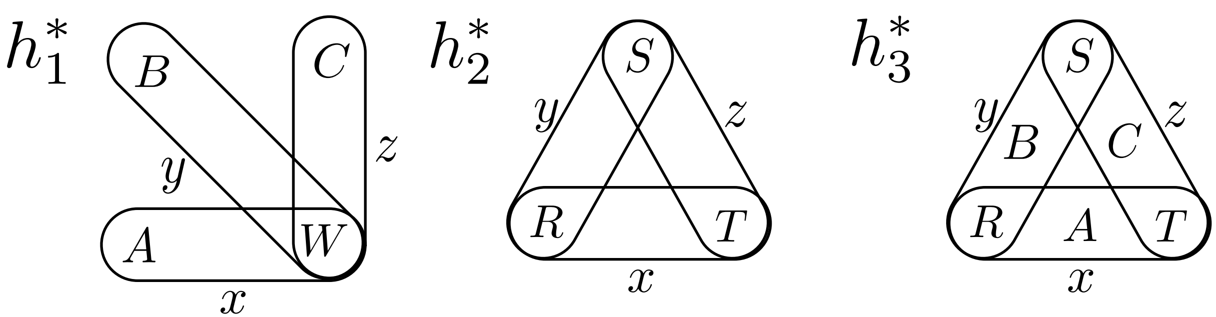
$$h_1^* \coloneq A^{[\mathtt{n}]}(x), B^{[\mathtt{n}]}(y), C^{[\mathtt{n}]}(z), W(x, y, z)$$

$$h_2^* \coloneq R^{[\mathtt{n}]}(x, y), S^{[\mathtt{n}]}(y, z), T^{[\mathtt{n}]}(z, x)$$

$$h_3^* \coloneq A^{[\mathtt{n}]}(x), B^{[\mathtt{n}]}(y), C^{[\mathtt{n}]}(z), R(x, y), S(y, z), T(z, x)$$

endogenous

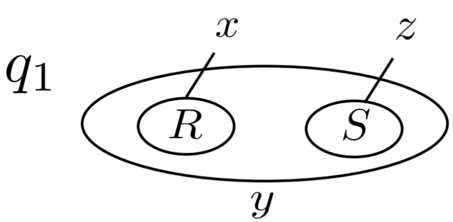
If unspecified, it could be either

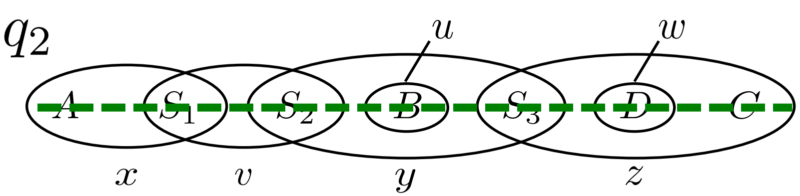


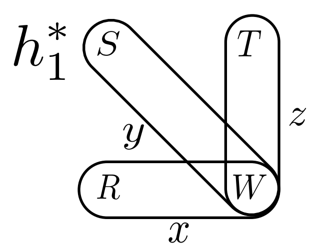
None of these are linear

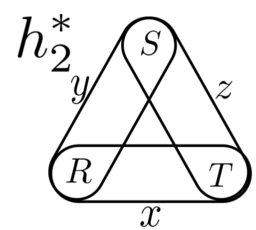
Responsibility dichotomy

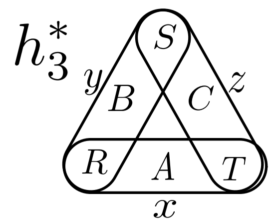
PTIME	NP-hard
q_1 :– $R(x,y), S(y,z)$ q_2 :– $A(x)S_1(x,v), S_2(v,y),$ $B(y,u), S_3(y,z), D(z,w), C(z)$	h_1^* :– $A(x), B(y), C(z), W(x,y,z)$ h_2^* :– $R(x,y), S(y,z), T(z,x)$ h_3^* :– $A(x), B(y), C(z),$ $R(x,y), S(y,z), T(z,x)$











Any query w/o self-join either reduces to an easy query
or has a reduction from a hard query by weakening

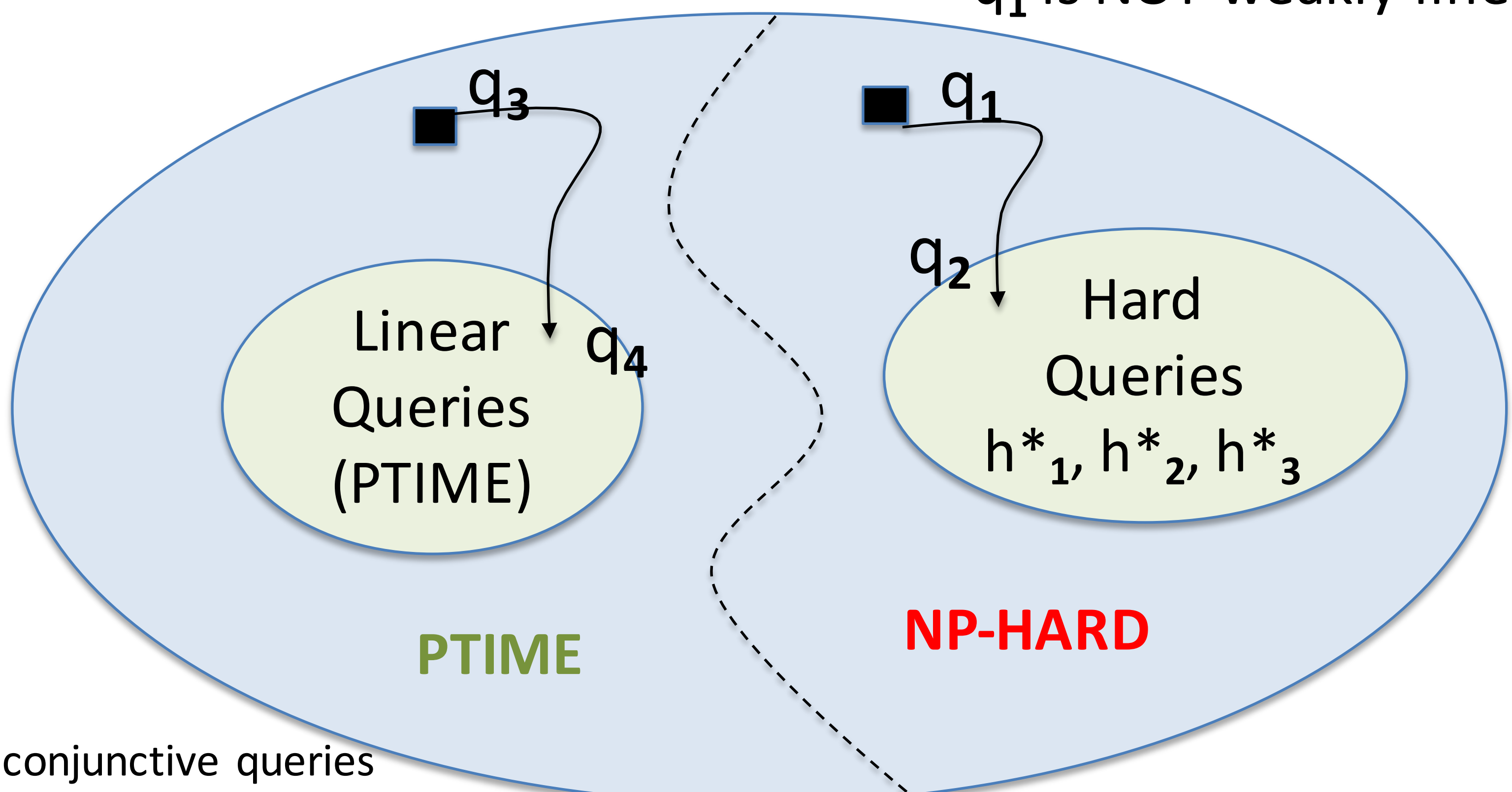
Proof Sketch: Dichotomy

Weakening:

- if q_4 is PTIME, so is q_3
- q_3 is "weakly linear"

Rewriting:

- if q_2 is hard, so is q_1
- if no more rewriting possible, then one of h^*_1, h^*_2, h^*_3
- q_1 is NOT weakly-linear



Set of conjunctive queries without self-joins

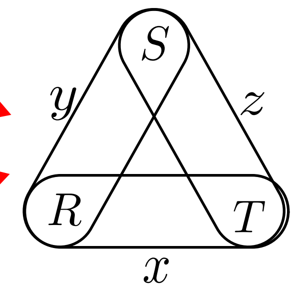
Example: Weakenings (for PTIME)

NP-hard

$$h_2^* :- R^{[\mathtt{n}]}(x,y), S^{[\mathtt{n}]}(y,z), T^{[\mathtt{n}]}(z,x)$$

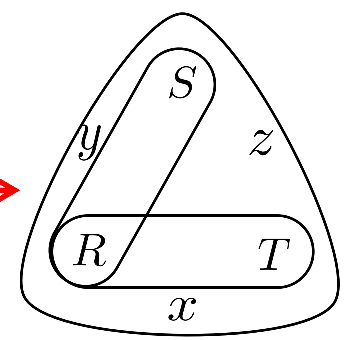
PTIME

$$q_2 :- R^{[\mathtt{x}]}(x,y), S^{[\mathtt{x}]}(y,z), \boxed{T^{[\mathtt{n}]}(z,x)}$$



R is exogenous, and therefore its tuples cannot be part of the contingency set

$$q_2' :- R^{[\mathtt{x}]}(x,y,z), S^{[\mathtt{x}]}(y,z), T^{[\mathtt{n}]}(z,x)$$



Expand R with the domain of z.
Responsibility of T tuples is not affected!

Dissociation

There are other rules for weakenings

Example: Rewriting (for NP-hardness)

q :- R(x, y), S(y, z), T(z, u), K(u, x)

→ R(x, y), S(y, z), T(x, z, u), K(u, x)

add x: add variable x to all atoms that contain u provided there is an atom containing both x and u

→ R(x, y), S(y, z), T(x, z, u), K(u, x, z)

add z: ….. u ….   (as above)

→ R(x, y), S(y, z), T(x, z, u)

delete K: if K is exogenous or if there is an atom T (here) such that var(T) $\subseteq$ var(K)

→ R(x, y), S(y, z), T(x, z)

delete u: delete u from all atoms containing u

= h^*_2

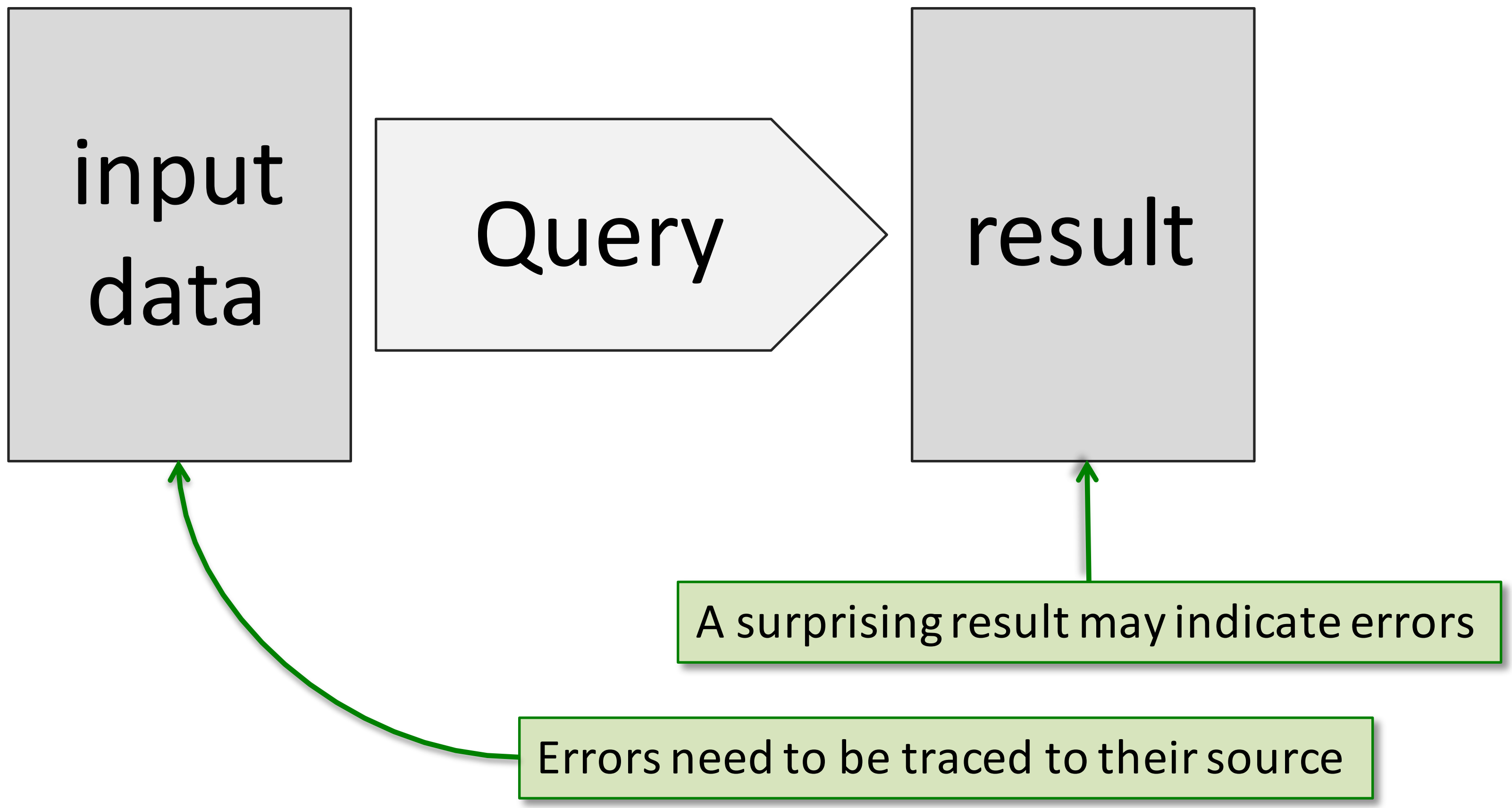
Responsibility for Why-No causality

- What to add along with a tuple t to make a non-answer p an answer
- Much easier (PTIME)
- If query has m subgoals, the size of the contingency set is at most m-1
  - e.g. q:- R(x, y) T(y, z) has 2 subgoals

- Try all possible options
- If the active domain size is N, at most N^m options
- PTIME data complexity (m = constant)

Meliou et al.' 2011

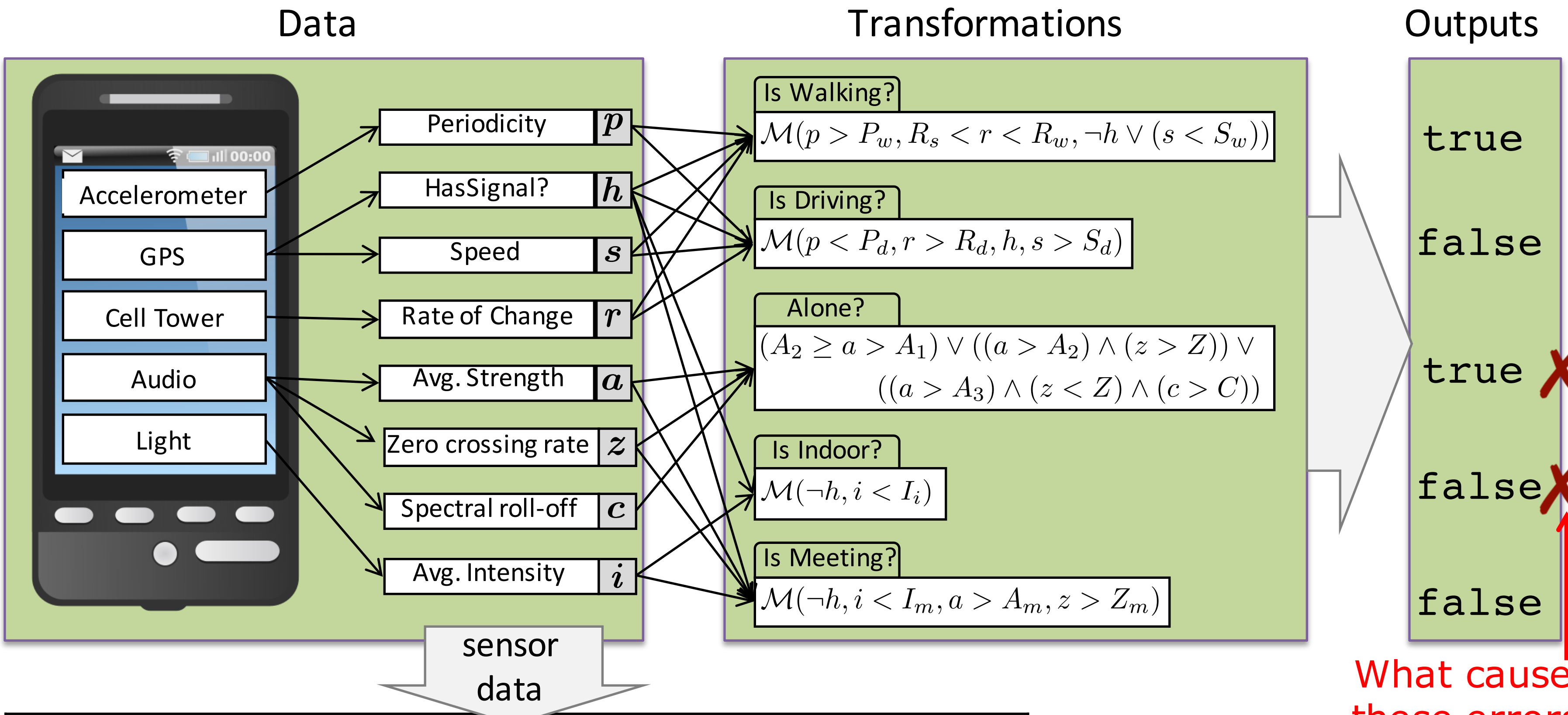
Responsibility in practice



Post-factum data cleaning

Meliou et al.' 2011

Context Aware Recommendations



0.016	True	0.067	0	0.4	0.004	0.86	0.036	10
0.0009	False	0	0	0.2	0.0039	0.81	0.034	68
0.005	True	0.19	0	0.03	0.003	0.75	0.033	17
0.0008	True	0.003	0	0.1	0.003	0.8	0.038	18

What caused these errors?

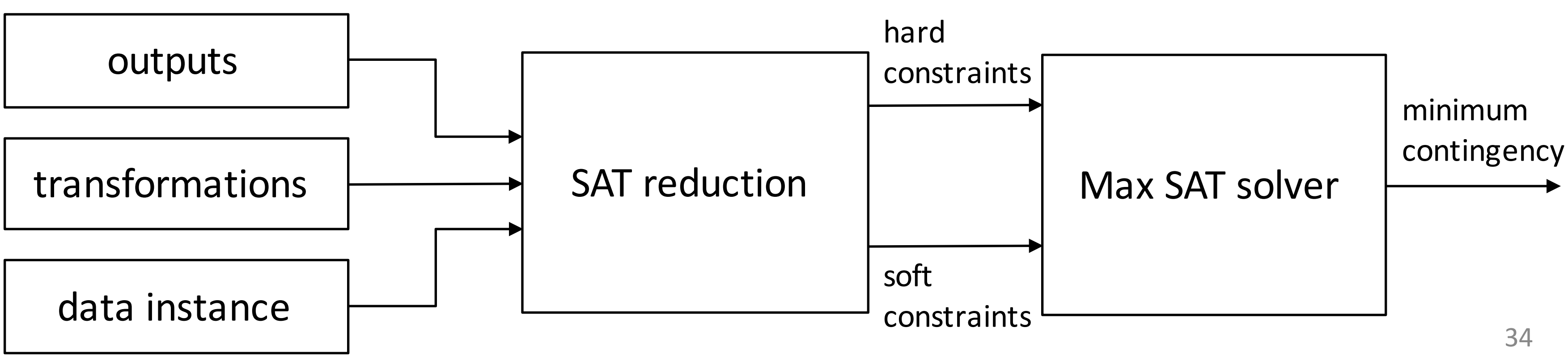
Sensors may be faulty or inhibited

It is not straightforward to spot such errors in the provenance

Meliou et al.' 2011

Solution

- Extension to view-conditioned causality
  - Ability to condition on multiple correct or incorrect outputs
- Reduction of computing responsibility to a Max SAT problem
  - Use state-of-the-art tools



Summary

- Pearl's causality model in AI can be adopted in DB
  - Causal network = provenance/lineage
  - Tuples are potential causes
  - Both for answers and non-answers
- However,
  - This does not reveal causal inferences in practice
  - e.g. whether smoking causes cancer
- We need to infer causal relationships among variables in the presence of other variables
  - confounding covariates
- Causality in Statistics and Rubin's potential outcome model
  - next lecture