Tutorial: Provenance and Causality

Wolfgang Gatterbauer,
Alexandra Meliou,
Dan Suciu

Overview

- Let Q(D) = D’ be a database transformation
- Let t’ $\in$ D’ be an output tuple
- Which tuples t $\in$ D *caused* t’ $\in$ D’?

- This talk: review causality, define causality in database transformation, give applications

Credits

- *Causality,* Judea Pearl, 2000
- *Causes and Explanations: A Structural-Model Approach,* Halpern and Pearl, 2001
- *Complexity results for explanations in the structural-model approach*, Eiter and Lukasiewicz, 2004
- *Some Topics in Analysis of Boolean Functions*, Ryan O'Donnell, 2008
- *Scalable Techniques for Mining Causal Structures*, Silverstein, Brin, Motwani, Ullman, 2000
- Responsibility and blame: A structural-model approach, Chockler, Halpern, 2004
- Y. Crama, P. L. Hammer, Boolean Functions: Theory, Algorithms, and Applications, Cambridge University Press, 2011 (preprint 2009).
- *Causality in Databases,* Meliou et al., 2010
- *Tracing Data Errors with View-Conditioned Causality,* Meliou et al., 2011
- *The Complexity of Causality and Responsibility for Query Answers and non-Answers,* Meliou et al., 2011b
- *WHY SO? or WHY NO? Functional Causality for Explaining Query Answers*, Meliou, et al., 2010
- *Causality and the Semantics of Provenance,* James Cheney, DCM 2010

Outline

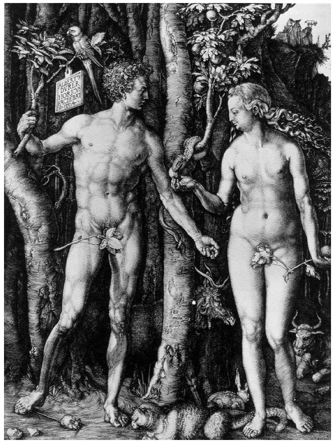
- Brief History
- Causality for a Boolean function
- Causality for a query
- Applications
- Summary

History of Causality

- What is the mathematical equation of cause?
- Surprisingly difficult to give
- The following brief "history" is based mostly on [Pearl'2000] – an advertisement for this great book

Antiquity

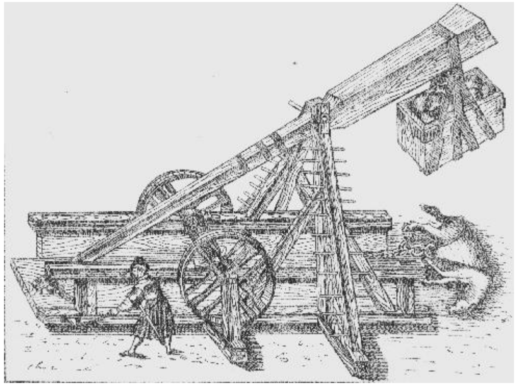
- Causality used to pass responsibility, attributing intent, and blame: only gods, humans, animals are agents of cause

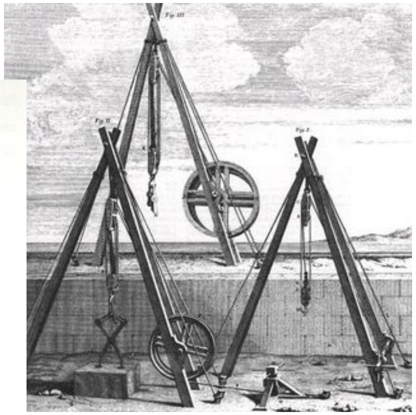


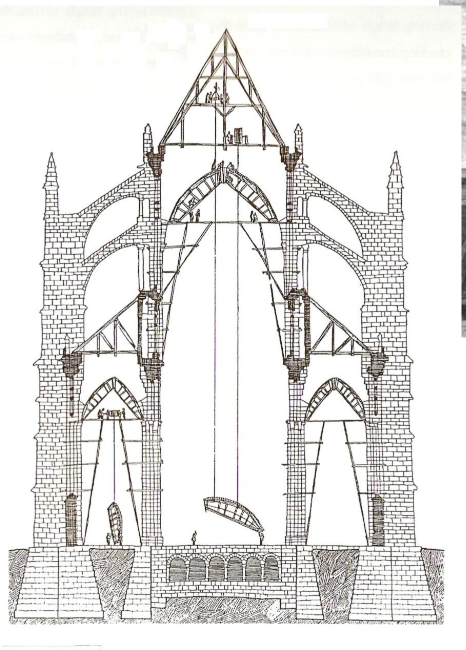
- Aristotle viewed causes in terms of a purpose; no definition

The Dawn of Science&Engineering

- Find objective causes rather than passing responsibility
- Questions of interest:
  - *Why* doesn't the wheel turn?
  - *What if* I make the beam half as thick, will it carry the load?
  - *How* do I shape the beam so it will carry the load?
- Same questions today in DB!

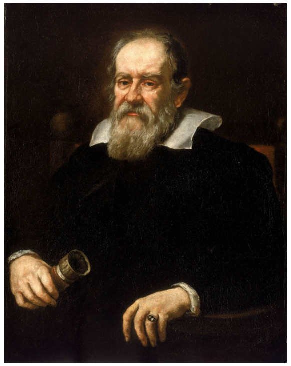






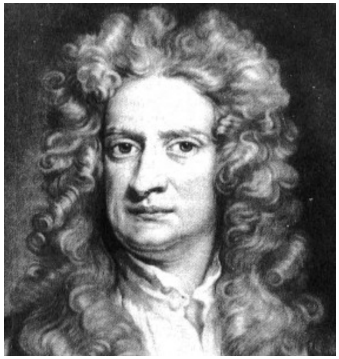
But what *IS causality*?

Science Seeks to Explain Causes, but *Lacks* A Language For Causality



Second law of motion says this:

$F = m a$



- What we know, but the law doesn't say it:

The force *causes* acceleration: $a = F/m$

- We also know, but the law doesn't say it:

Force + acceleration *determine* mass: $m = F/a$

They do *not cause* the mass.

David Hume

- Causality is a matter of perception:
  - *"we remember seeing the flame, and feeling a sensation called heat; without farther ceremony, we call the one cause and the other effect".*
- Opens door to finding causes from empirical observations
- But correlation is not causation

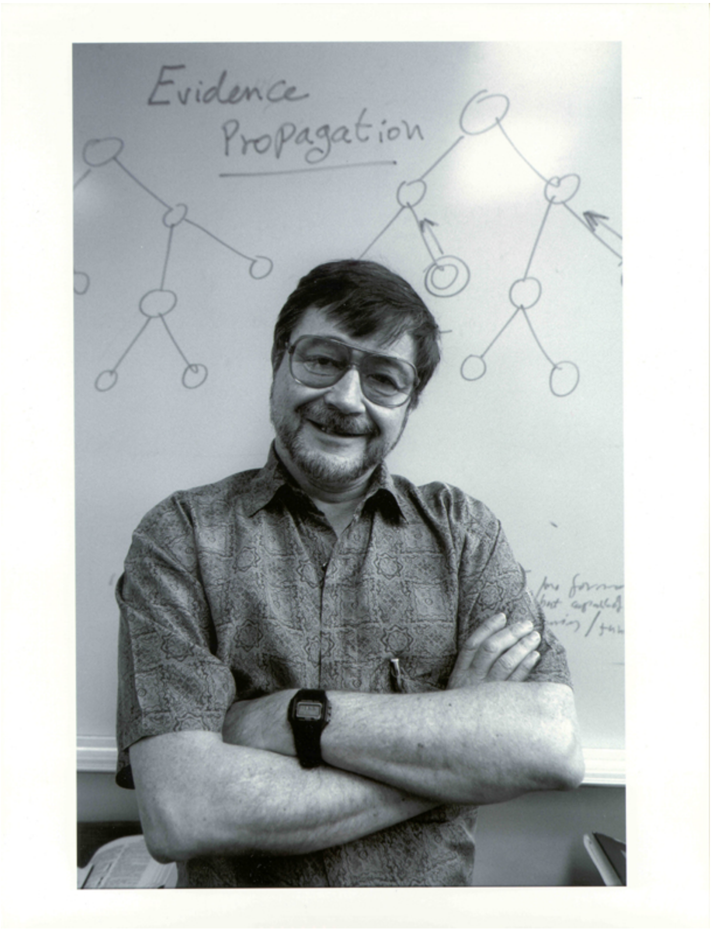
Karl Pearson

- Co-founder of modern statistics
- Forget causation! Correlation is all you should ask for
- Statistical machine learning (ML) relies on the principle of seeking only correlations
- Mantra: don't attempt to find causation!
- Very few dissenters dare to look for causations in data, e.g. [Silverstein'2000]

Judea Pearl

- Forget empirical observations
- Start from a <u>*Causal Network,*</u> consisting of known, physical causation relationships
- Substitute randomness with <u>*Exogenous Variables*</u>

- Result: a mathematical definition of causality
- Caveat: must build the causal network first



In database transformations:
causal network = provenance

Outline

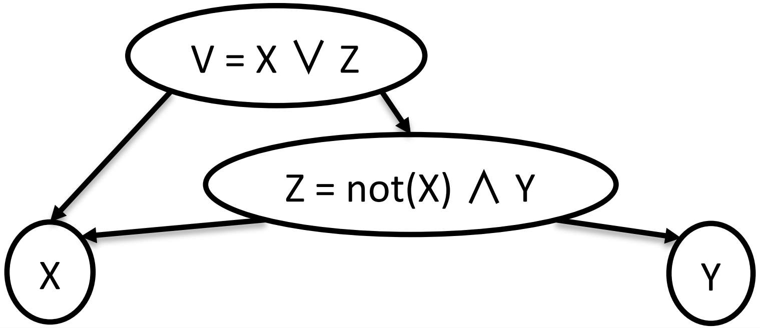
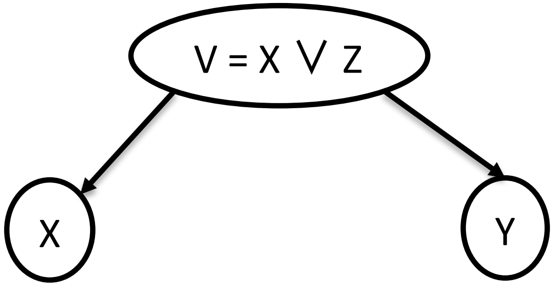
- Brief History
- Causality for a Boolean function
- Causality for a query
- Applications
- Summary

[Pearl&Halpern’2001, Eiter&Lukasiewicz 2004]

Causal Model

Definition A *causal model* (causal network) consists of:

- *Exogenous* Boolean variables -- external; fixed
- *Endogenous* Boolean variables -- internal; modifiable
- Boolean functions defining some endogenous variables

AI: “recursive model” (DAG)	DB: single Boolean function F
 	

This talk: Causal Model is a Boolean function $Y = F(X_1,\ldots,X_n)$

Cause of a Boolean Function $Y = F(X_1,\ldots,X_n)$

- Fix θ = an assignment (world); $y = \theta(F)$
- Fix X_i (endogenous variable); $x_i = \theta(X_i)$
- $[\theta\text{-}X_i](F)$ = assign all variables in F *except* X_i

Definition The event $X_i = x_i$ is a *counterfactual cause* for $Y = y$ in θ, if $[\theta\text{-}X_i](F)$ depends on X_i

Equivalently: changing the value of X_i in θ causes Y to change

Definition The event $X_i = x_i$ is an *actual cause* for $Y=y$ if $\exists\, \theta'$ s.t. it is counterfactual for $Y = y$ in θ'

θ, θ' must agree on X_i, on exogenous variables, on output Y

Three Simple Examples

Assume all variables X_1, X_2, X_3 are endogenous

$Y = X_1 \wedge X_2$

$\theta(X_1)=\theta(X_2)=1 \rightarrow Y=1$

$X_1=1$ is a counterfactual cause for $Y=1$

$[\theta\text{-}X_1](F) = X_1 \wedge 1=X_1$

$Y = X_1 \vee X_2$

$\theta(X_1)=\theta(X_2)=1 \rightarrow Y=1$

$X_1=1$ is no counterfactual cause for $Y=1$;

$[\theta\text{-}X_1](F) = X_1 \vee 1=1$

$X_1=1$ is an actual cause for $Y=1$

$\theta'(X_1)=1,\theta'(X_2)=0 \rightarrow [\theta'\text{-}X_1](F)=X_1$

$Y = [\mathrm{not}(X_1) \wedge X_2] \vee X_3$

$\theta(X_1)=\theta(X_2)=\theta(X_2)=1 \rightarrow Y=1$

$X_1=1$ is not an actual cause for $Y=1$

$[\theta'\text{-}X_1](F) =\mathrm{not}(X_1)$

Complexity of Causality

[Eiter&Lukasiewicz 2004]

Counterfactual cause	Actual cause
PTIME	NP-complete

Proof: Reduction from SAT.

Given F,   F is satisfiable iff X is an actual cause for $X \wedge F$

Related Concepts 1/3

Definition Fix F, θ.
X_i is a *critical* for F in θ if [θ-X_i](F) depends on X_i

"X_i is counterfactual cause for F in θ"
"X_i is a critical voter (swing voter)"

Definition Fix F.
X_i is a *critical* for F if ∃ θ s.t X_i is critical for F in θ

"F depends on X_i" "X_i is in the support of F"

Applications to data privacy:
if X_i is not critical, then F reveals nothing about X_i

[O'Donnell 2008]

Related Concepts 2/3

Probability over random choices of θ

Definition Fix F.

The *<u>influence</u>* of X_i: $\mathrm{Inf}(X_i)$ = Prob[X_i is critical for F, θ]

Examples:

Majority function: $\mathrm{Inf}(X_i) = \binom{n-1}{\frac{n-1}{2}}\frac{1}{2^{n-1}} \approx \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}}$

Parity function: $\mathrm{Inf}(X_i) = 1$

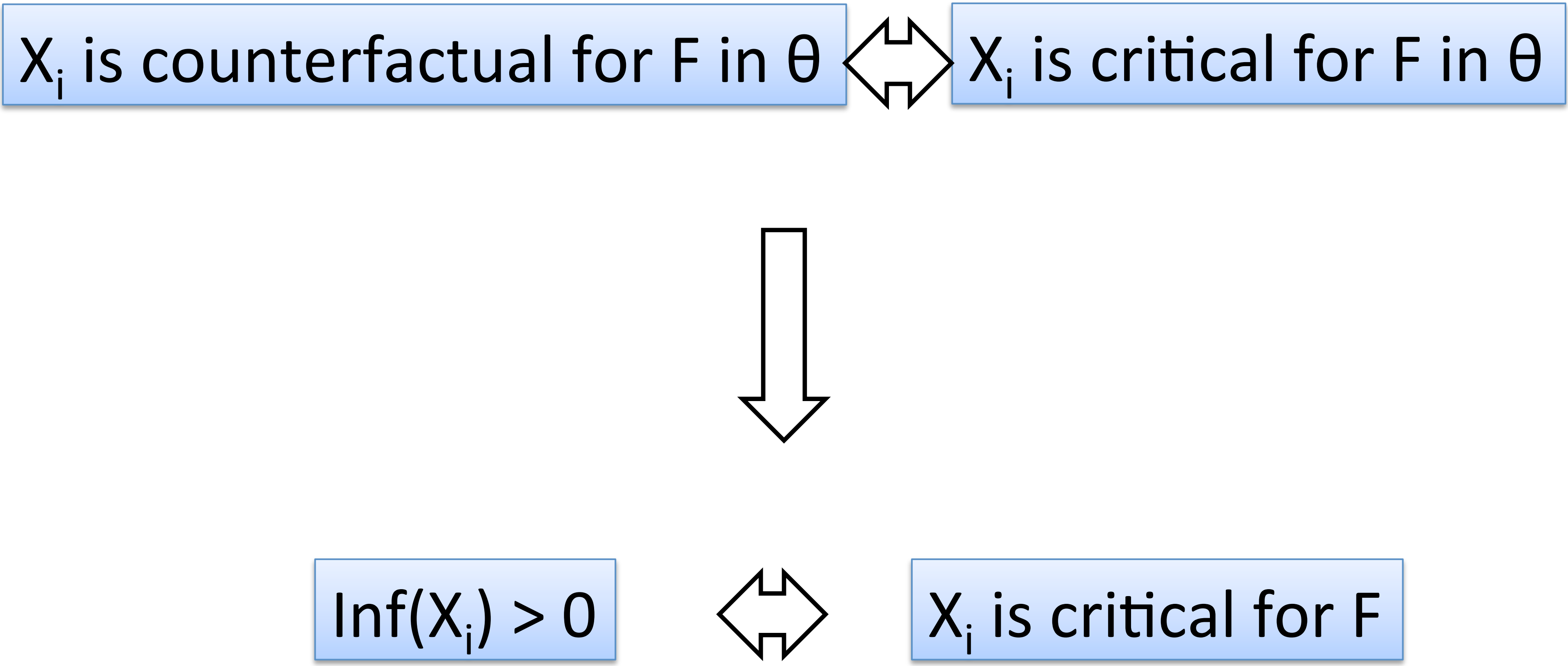
Application: The *influence* (or *power*) of voter X_i

Modified Chow Index = $\mathrm{Inf}(X_i) * 2^n$

Banzhaf Index = $\mathrm{Inf}(X_i)$ / [$\mathrm{Inf}(X_1)$ +…+$\mathrm{Inf}(X_n)$]

[Crama'2010 p.87 and p78]

Related Concepts 3/3



X_i is counterfactual for F in θ

X_i is critical for F in θ

$\text{Inf}(X_i) > 0$

X_i is critical for F

Outline

- Brief History
- Causality for a Boolean function
- Causality for a query
- Applications
- Summary

Endogenous/exogenous Tuples

Input database D, Query Q, output D' = Q(D)

Two kinds of tuples can go into D:

- Exogenous tuples: D^x
  - External, from sources that are certain; not causes

- Endogenous tuples: D^n
  - Tuples that affect outcome; potential causes

Database $D \subseteq D^x \cup D^n$

Causality of a Query Answer

Fix database D, query Q, output D'=Q(D)

symmetric difference

- Input $t \in D^n$ is counterfactual cause in D for t' if t' occurs in exactly one of Q(D) or $Q(D \otimes t)$
  - **Why-so cause**: when $t' \in Q(D)$ and $t' \notin Q(D \otimes t)$
  - **Why-no cause**: when $t' \notin Q(D)$ and $t' \in Q(D \otimes t)$
- Input $t \in D^n$ is an actual cause in D for output t' if $\exists \Gamma \subseteq D^n$ s.t. t is a counterfactual cause in $D \otimes \Gamma$

Contingency set

[Meliou'2011b]

Responsibility of a Query Answer

Definition: Responsibility of t for t'

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

Here $\Gamma \subseteq D^n$ ranges over contingency sets

- If $\rho_t=1$ then t is a counterfactual cause
- If $0 < \rho_t < 1$ then t is an actual cause
- If $\rho_t=0$ then t is not a cause

Responsibility introduced for *causal networks* [Chockler'2004]

[Meliou’2011b]

Complexity

Causality of CQ queries;

Why-so?	Why-no?
PTIME	PTIME

In FO

Responsibility of CQ queries without self-joins;

Why-so?	Why-no?
PTIME/NP-hard dichotomy	PTIME

Some queries in PTIME
are NL-hard (hence not in FO)

Related: hardness of responsibility in causal networks [Chockler’2004]

[Meliou’2011b]

Responsibility: Dichotomy

Theorem Data complexity of the responsibility:
- If Q is *weakly linear*, then Q is in PTIME
- If Q is not *weakly linear*, then it is NP-hard

See [Meliou’2011b] for a definition of “weakly linear”.

Will give examples next.

[Meliou'2011b]

Responsibility: Easy and Hard Queries

Example Responsibility for the following query is in PTIME

Q :- R(x,y), S(y,z), T(z,u), M(u,v), …

Weakly linear

Non-weakly linear

Example: Responsibility for these queries is NP-hard:

$$h_1^* :- A^n(x), B^n(y), C^n(z), W(x,y,z)$$

$$h_2^* :- R^n(x,y), S^n(y,z), T^n(z,x)$$

$$h_3^* :- A^n(x), B^n(y), C^n(z), R(x,y), S(y,z), T(z,x)$$

endogenous

If unspecified, it could be either

Related Concept

The *Deletion Propagation Problem* [Buneman'2002]

- Fix D, Q, and $t' \in Q(D)$
- Problem: find $\Gamma \subseteq D$ such that
  - $t' \notin Q(D - \Gamma)$
  - The side effects $|Q(D) \otimes Q(D - \Gamma)|$ are minimal

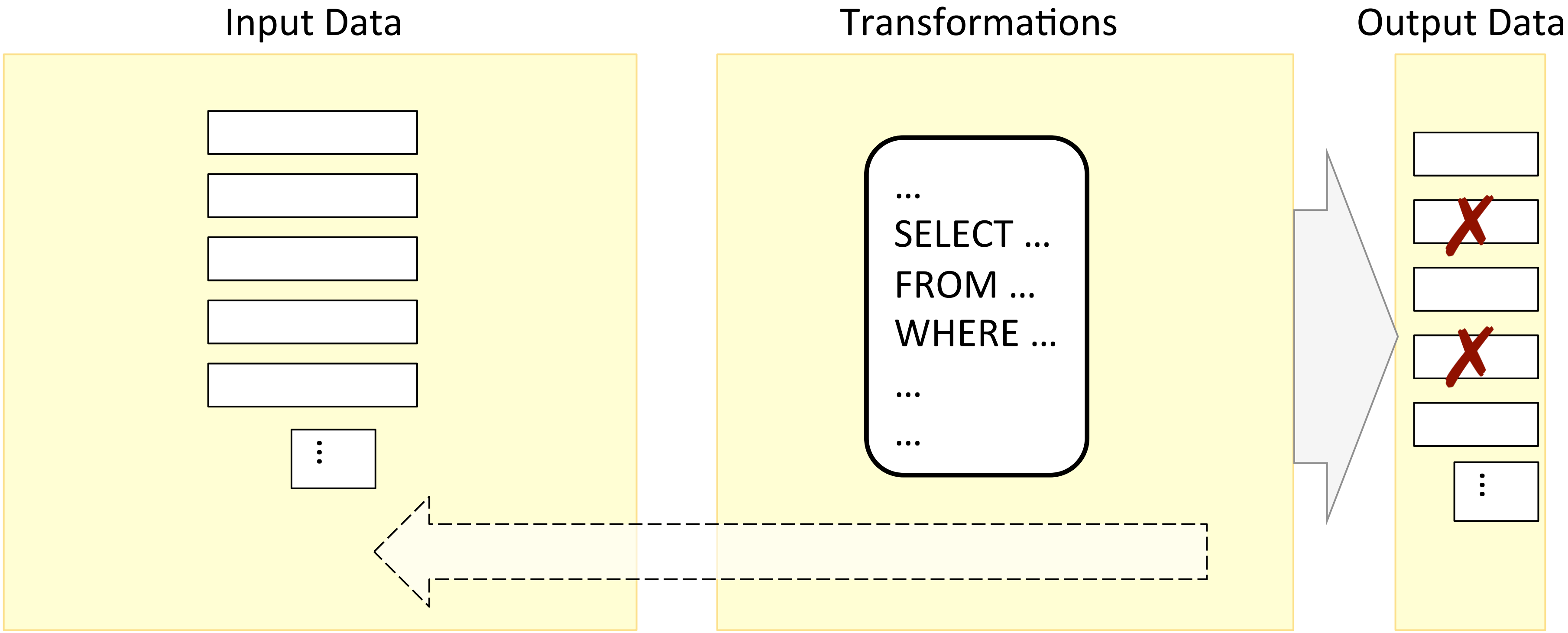
[Kimelfeld'2011] prove a dichotomy into PTIME and NP-hard for the Deletion Propagation Problem

Intuitively, Γ acts like a contingency set, but precise connection to causality has not been studied

Outline

- Brief History
- Causality for a Boolean function
- Causality for a query
- Applications
- Summary

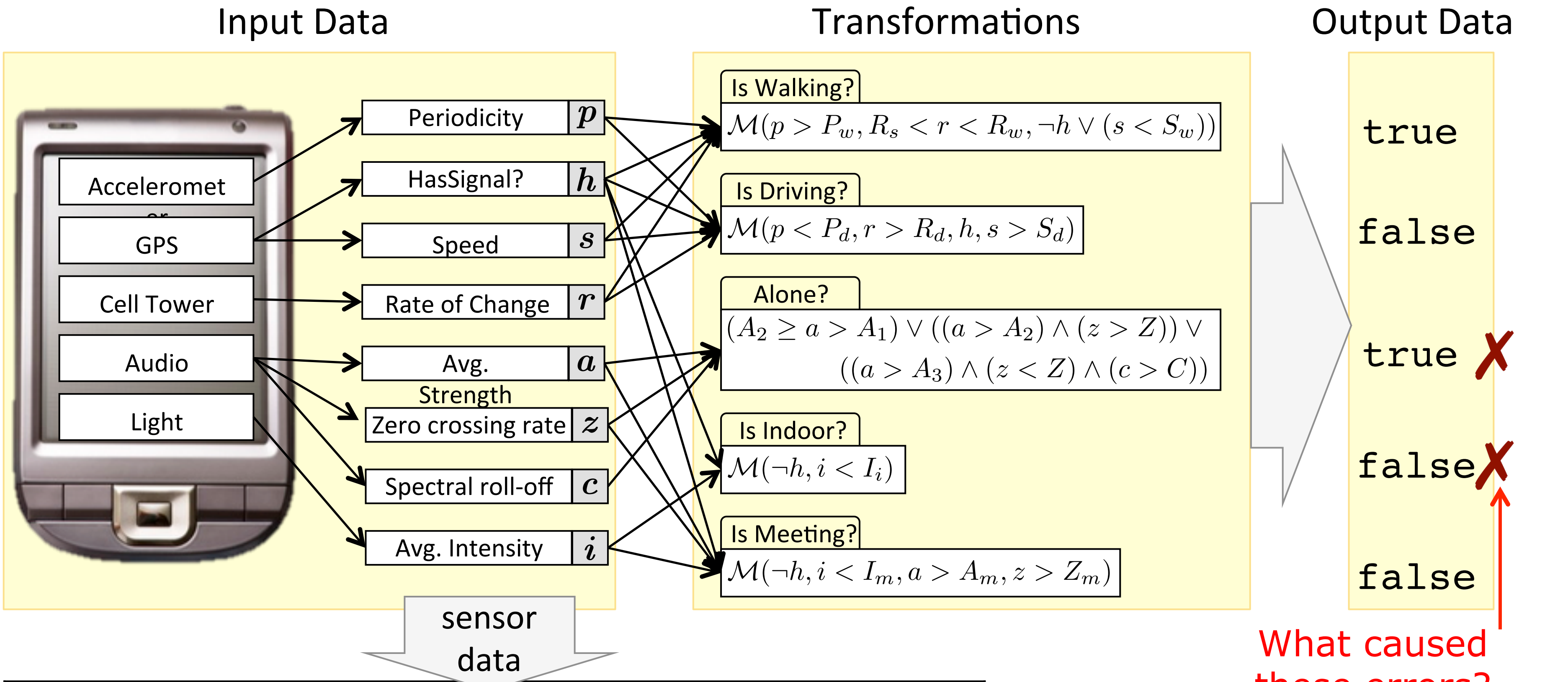
Error Correction



One or more outputs are wrong. Which inputs need to be corrected?
“Post factum” data cleaning.

Example

[Meliou'2011]



Input Data

Acceleromet

GPS

Cell Tower

Audio

Light

Periodicity p

HasSignal? h

Speed s

Rate of Change r

Avg. Strength a

Zero crossing rate z

Spectral roll-off c

Avg. Intensity i

Transformations

Is Walking?

$\mathcal{M}(p > P_w, R_s < r < R_w, \neg h \vee (s < S_w))$

Is Driving?

$\mathcal{M}(p < P_d, r > R_d, h, s > S_d)$

Alone?

$(A_2 \geq a > A_1) \vee ((a > A_2) \wedge (z > Z)) \vee$
$((a > A_3) \wedge (z < Z) \wedge (c > C))$

Is Indoor?

$\mathcal{M}(\neg h, i < I_i)$

Is Meeting?

$\mathcal{M}(\neg h, i < I_m, a > A_m, z > Z_m)$

Output Data

true

false

true ✗

false ✗

false

sensor data

What caused these errors?

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

Sensors may be faulty or inhibited
It is not straightforward to spot such errors in the provenance.

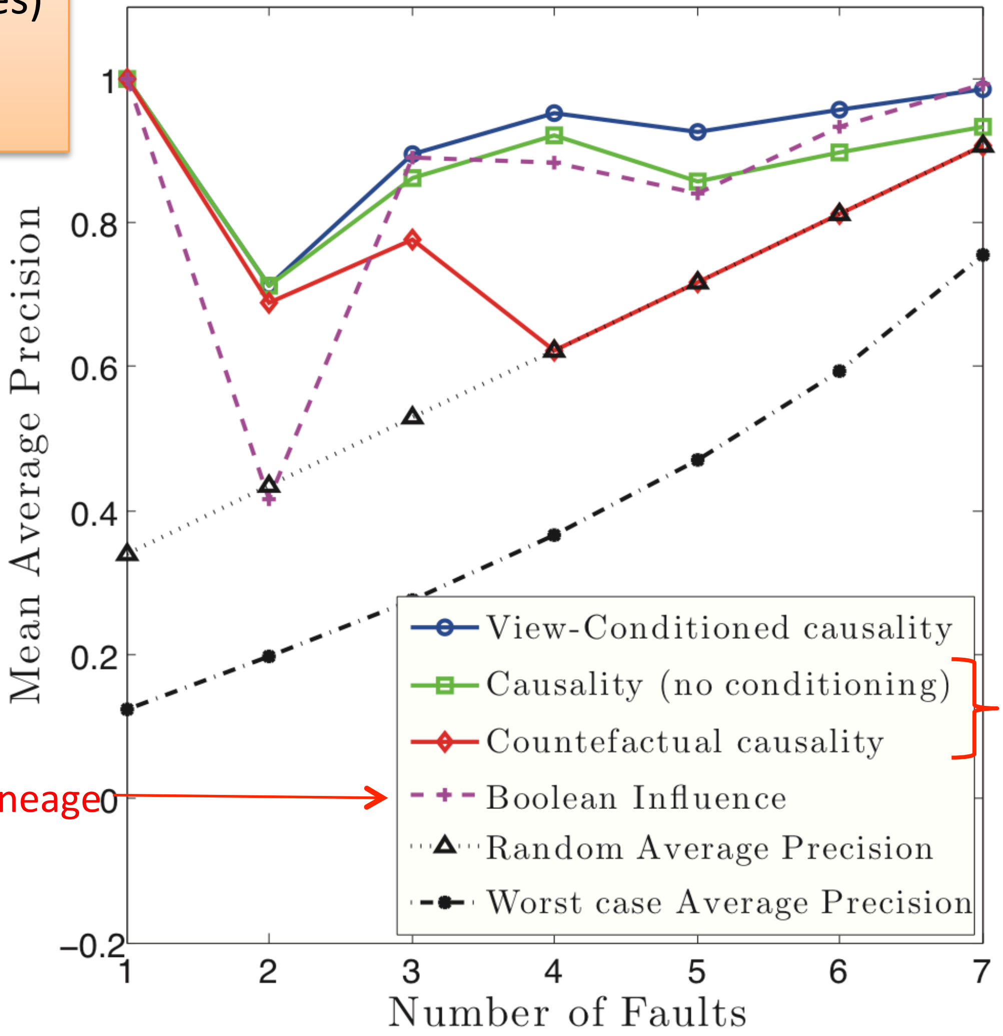
[Meliou’2011]

Precision

800 different instances
5 sensory inputs
8 extracted features (variables)
3 users
~10% observed errors

Average precision is a metric of quality of a ranking.

If all erroneous variables are ranked first, then average precision is 1.



Static analysis of lineage

Simpler causality schemes

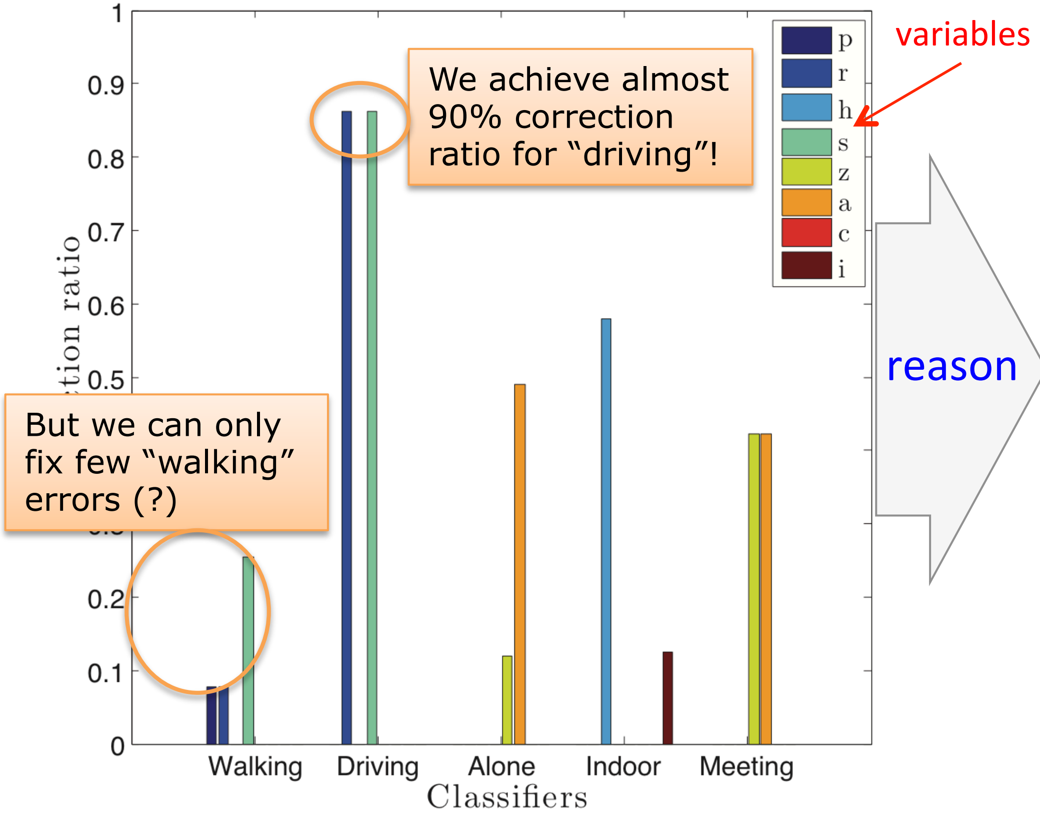
[Meliou'2011]

Correction

We select the highest responsibility variable, remove it from the evaluation of all classifiers, and record the portion of errors that get corrected per classifier

Driving has reliable features (low responsibility), means they are almost never causes of error

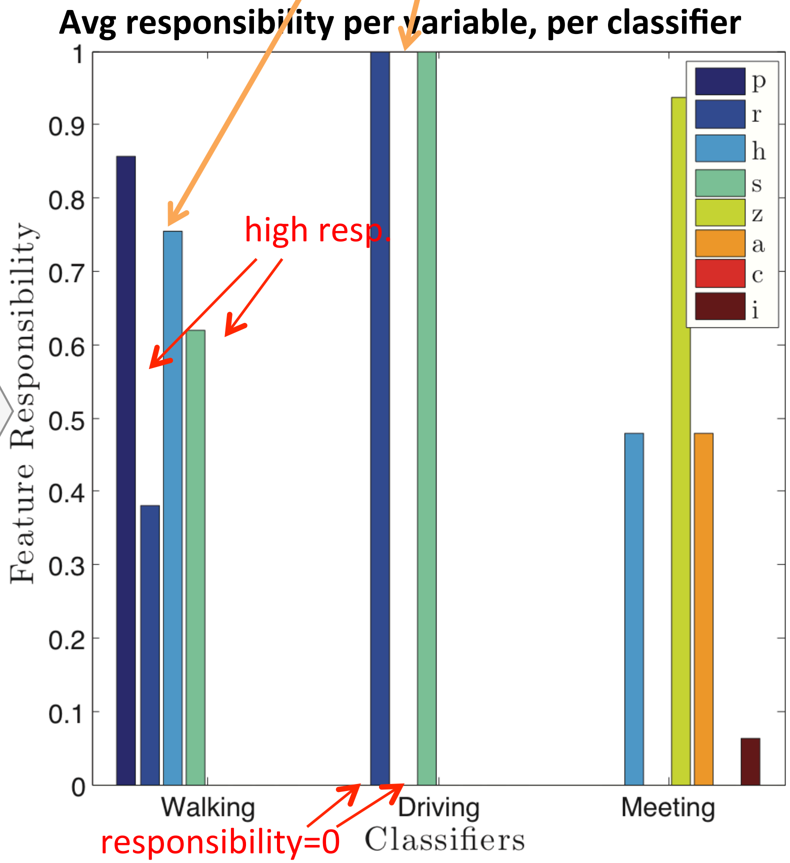
Walking has no reliable features



We achieve almost 90% correction ratio for "driving"!

But we can only fix few "walking" errors (?)

Avg responsibility per variable, per classifier



[Meliou'2011]

Why-So / Why-No Queries

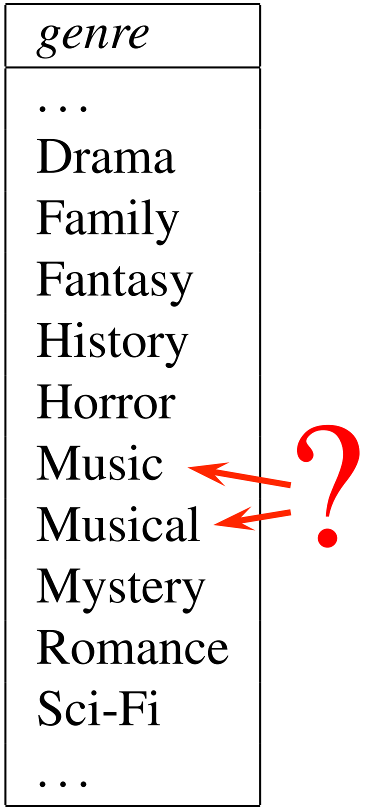
Database schema

Director(*did*, *firstName*, *lastName*)
Movie(*mid*, *name*, *year*, *rank*)
Movie_Directors(*did*, *mid*)
Genre(*mid*, *genre*)

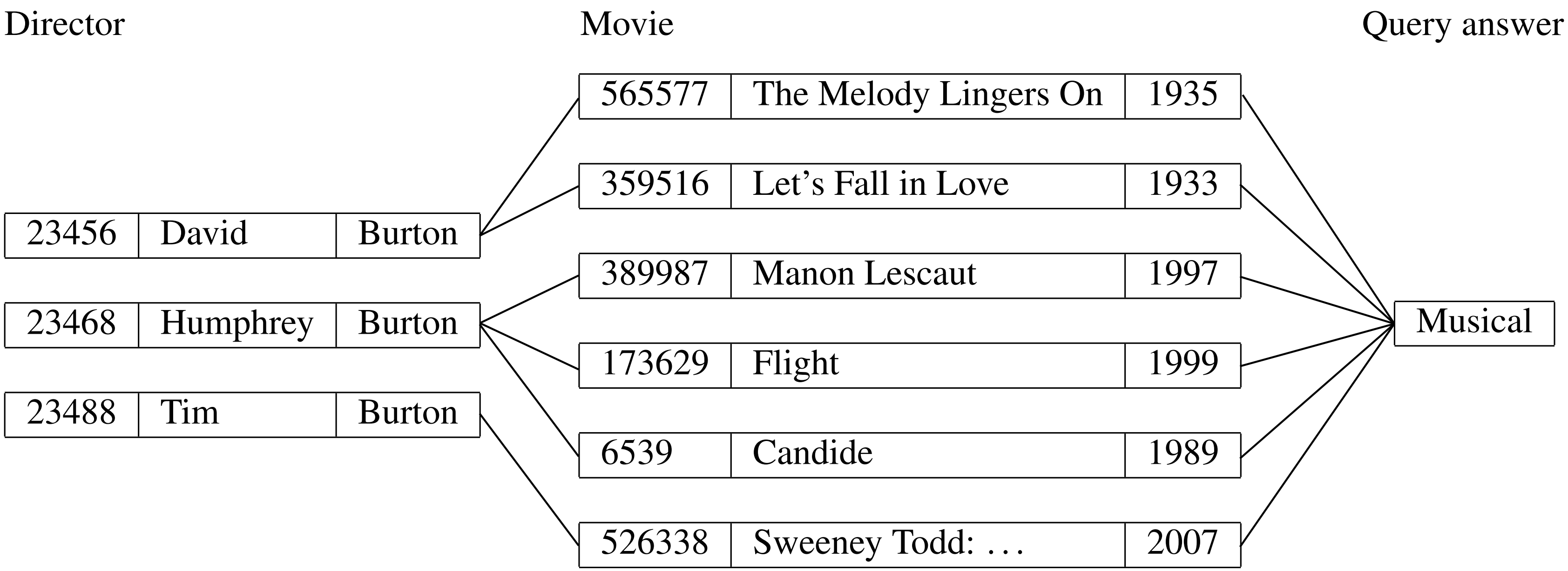
Query

```
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
```

Query answers



genre
. . .
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci-Fi
. . .



Director

23456	David	Burton
23468	Humphrey	Burton
23488	Tim	Burton

Movie

565577	The Melody Lingers On	1935
359516	Let’s Fall in Love	1933
389987	Manon Lescaut	1997
173629	Flight	1999
6539	Candide	1989
526338	Sweeney Todd: . . .	2007

Query answer

Musical

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

Outline

- Brief History
- Causality for a Boolean function
- Causality for a query
- Applications
- Summary

Summary

- Causality in Data Transformations:
  - *Which input tuple caused this output tuple?*
- Key concepts:
  - Endogenous/exogenous tuples
  - Counterfactual cause
  - Contingency set
  - Actual cause
- Very simple causal network: Boolean function
  - Avoids many complications
- Applications:
  - Error corrections
  - Why-so / why-no explanations