Relational Causal Inference & Hypothetical Reasoning

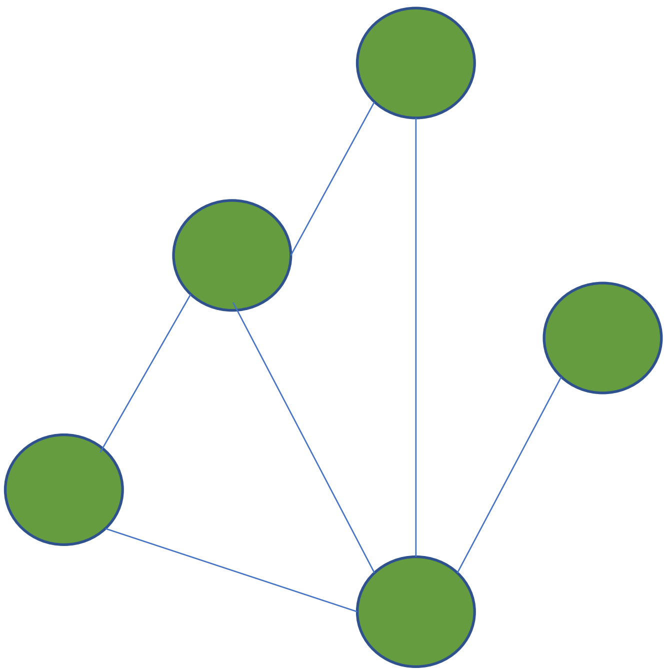
Sudeepa Roy

Papers:

- Causal Relational Learning.
  Babak Salimi, Harsh Parikh, Moe Kayali, Lise Getoor, Sudeepa Roy, and Dan Suciu
  ACM SIGMOD International Conference on Management of Data (SIGMOD), 2020.
- HypeR: Hypothetical Reasoning With What-If and How-To Queries Using a Probabilistic Causal Approach.
  Sainyam Galhotra*, Amir Gilad*, Sudeepa Roy, and Babak Salimi)
  ACM SIGMOD International Conference on Management of Data (SIGMOD), 2022.

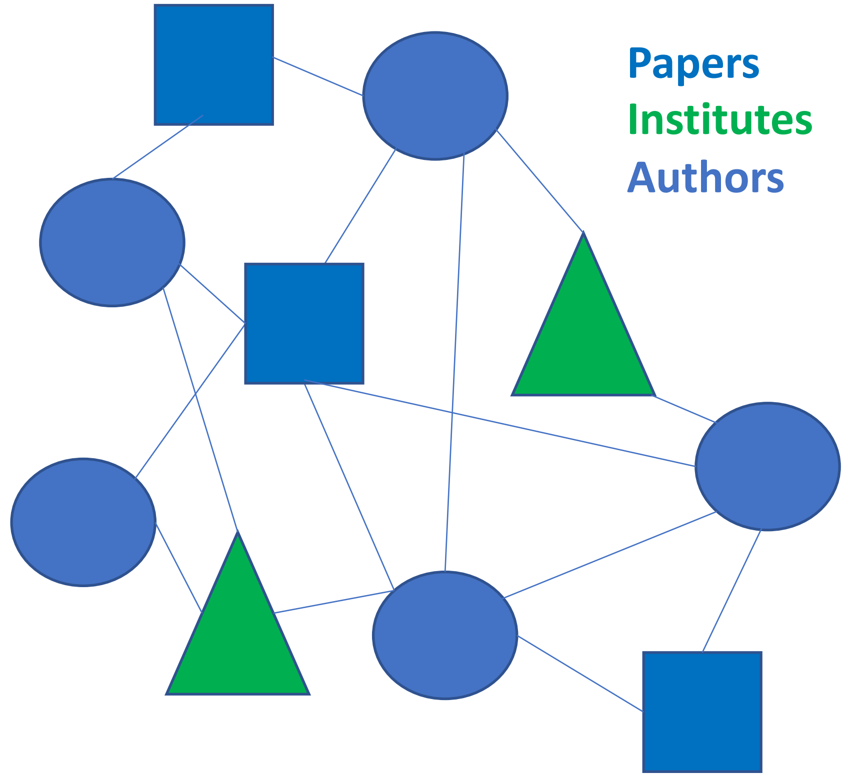
Thanks to Harsh Parikh and Sainyam Galhotra for some slides!

Units with Interference



Student sharing rooms in college dorms

"homogenous units"



"heterogenous units"

Network data

Relational data

[Sherman-Shpitser, UAI'19]
[Bhattacharya-Malinsky-Shpitser, UAI'19]
[Morucci-Awan-Orlandi-Roy-Rudin-Volfovsky UAI '19]

- Treatment of one unit may affect outcome of another
- Basic assumptions fail

Heterogenous "relational" data

Authors

person	prestige	qualification (h-index)
Bob	1	50
Carlos	0	20
Eva	1	2

Submissions

sub	score
s1	0.75
s2	0.4
s3	0.1

Authorship

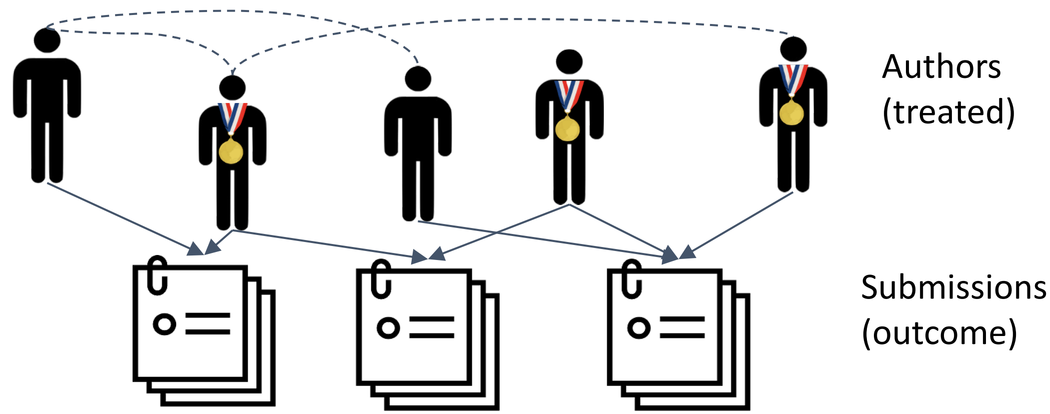
person	sub
Bob	s1
Eva	s1
Eva	s2
Eva	s3
Carlos	s3

Submitted

sub	conf
s1	ConfDB
s2	ConfAI
s3	ConfAI

Conferences

conf	blind
ConfDB	Single
ConfAI	Double



Does institutional rank (prestige) causally affect Scores received by papers in reviews?

- For single-blind reviews?
- For double-blind reviews?

Treatment

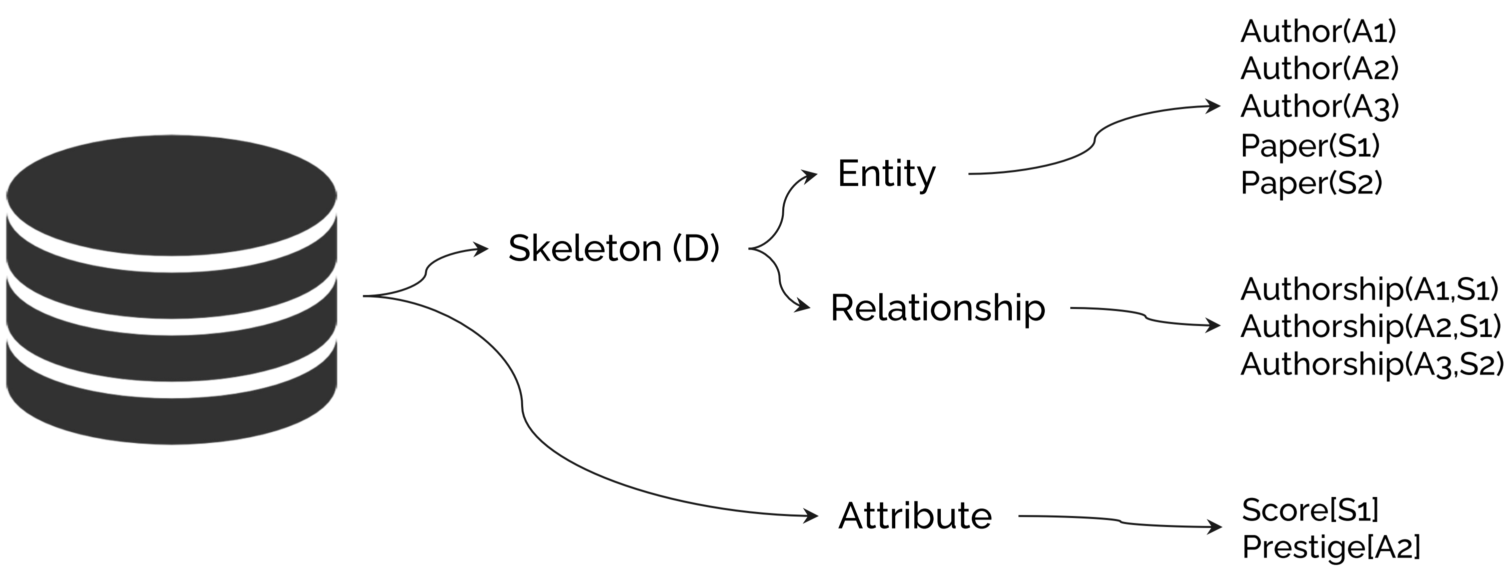
Authors(person, name, position, **f(inst-rank)**)
Authorship(person, sub)
Submission-reviews(sub, **score**) **Outcome**
Submitted(sub, conf)
Conferences(conf, is-single-blind)

Relational DB
- Multiple Tables with heterogeneous entities and relationships
- Non-uniform Many-to-Many connections between Authors & Papers

More examples?

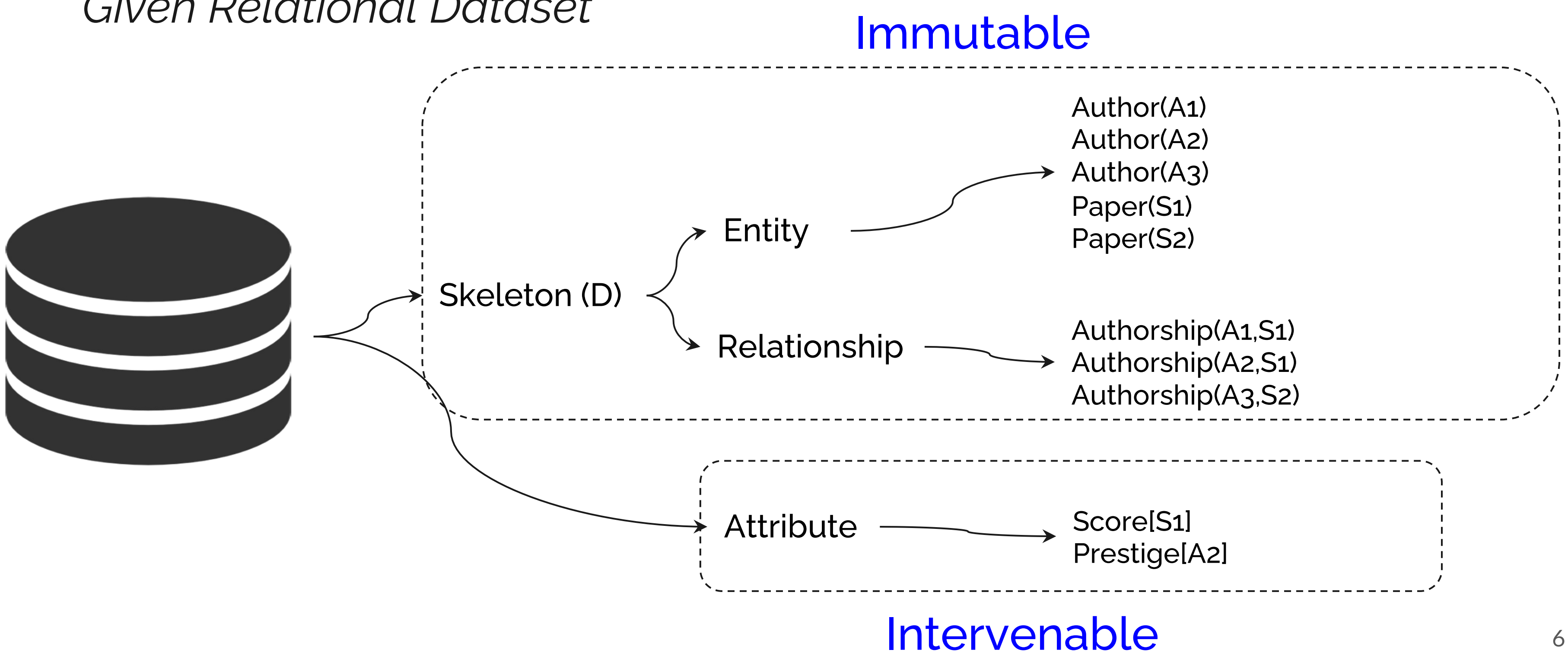
Relational DB

Given Relational Dataset



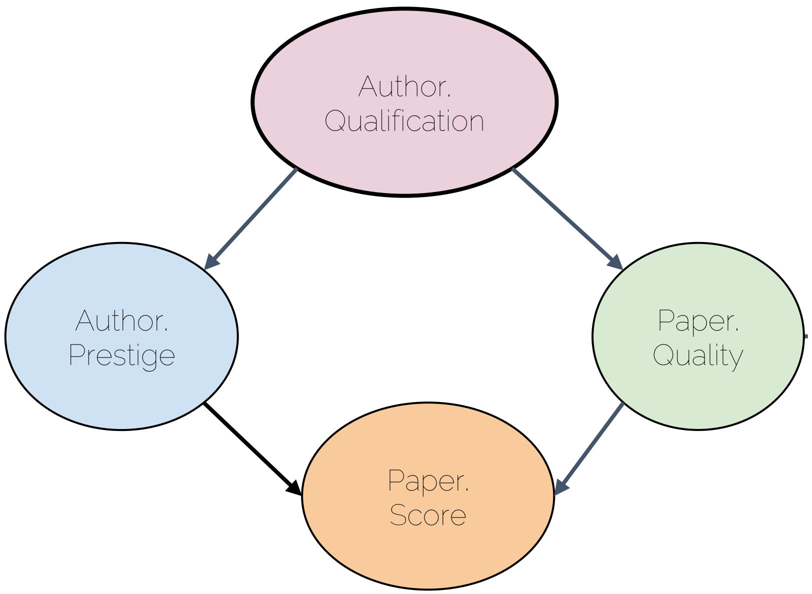
Relational DB

Given Relational Dataset



Encode Background Knowledge by Relational Causal Graphs

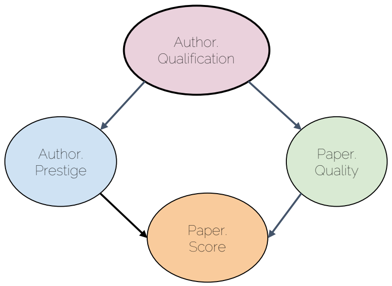
Potential Causal Links



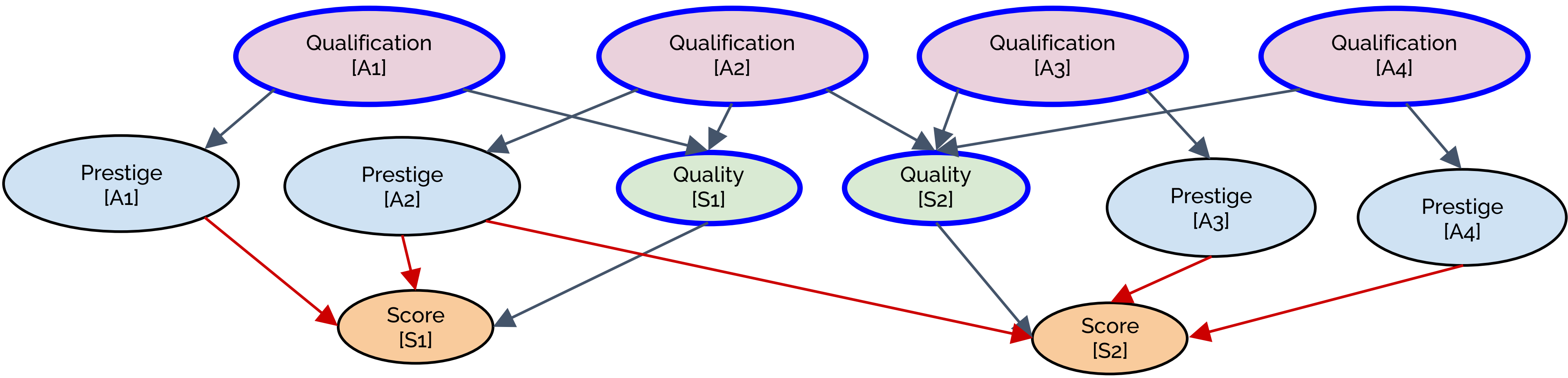

```
Prestige[A] <= Qualification[A]
           WHERE Author(A)

Quality[S] <= Qualification[A]
           WHERE Authorship(A, S)

Score[S] <= Prestige[A]
           WHERE Authorship(A, S)

Score[S] <= Quality[S]
           WHERE Paper(S)
```

Similar to Pearl’s Graphical Causal Model but Parameterized

Grounded Relational Causal Graphs using data




```
Prestige[A] <= Qualification[A] WHERE Author(A)

Quality[S] <= Qualification[A] WHERE Authorship(A, S)

Score[S] <= Prestige[A] WHERE Authorship(A, S)

Score[S] <= Quality[S] WHERE Paper(S)
```



	Authors		
	person	prestige	qualification (h-index)
A1	Bob	1	50
A3	Carlos	0	20
A2	Eva	1	2

A4 Mary 1 30

Submissions	
sub	score
s1	0.75
s2	0.4
s3	0.1

Authorship	
person	sub
Bob	s1
Eva	s1
Eva	s2
Carlos	s3 s2

Mary s2

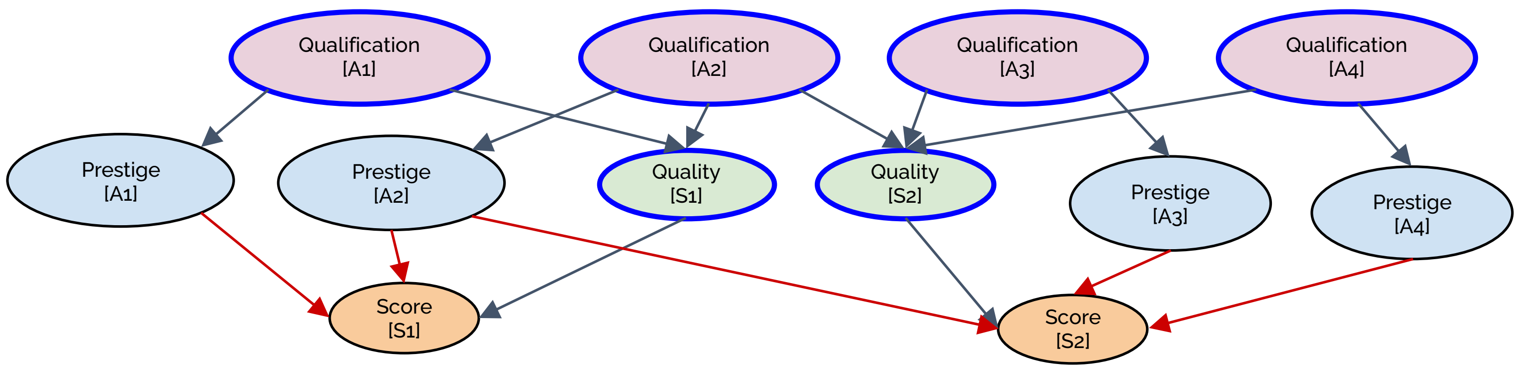
Submitted	
sub	conf
s1	ConfDB
s2	ConfAI
s3	ConfAI

Conferences	
conf	blind
ConfDB	Single
ConfAI	Double

Assumptions

- **Structural Homogeneity:** All grounded attributes $\mathtt{A}[\mathbf{x}] \in \mathtt{A}^{\Delta}$ of the same attribute $\mathtt{A} \in \mathbf{A}$ share the same structural equation and, hence, the same conditional probability distribution in equation (10).

$$\Pr\left(\mathtt{A}[\mathbf{x}] \mid \mathbf{Pa}(\mathtt{A}[\mathbf{x}])\right) \tag{10}$$



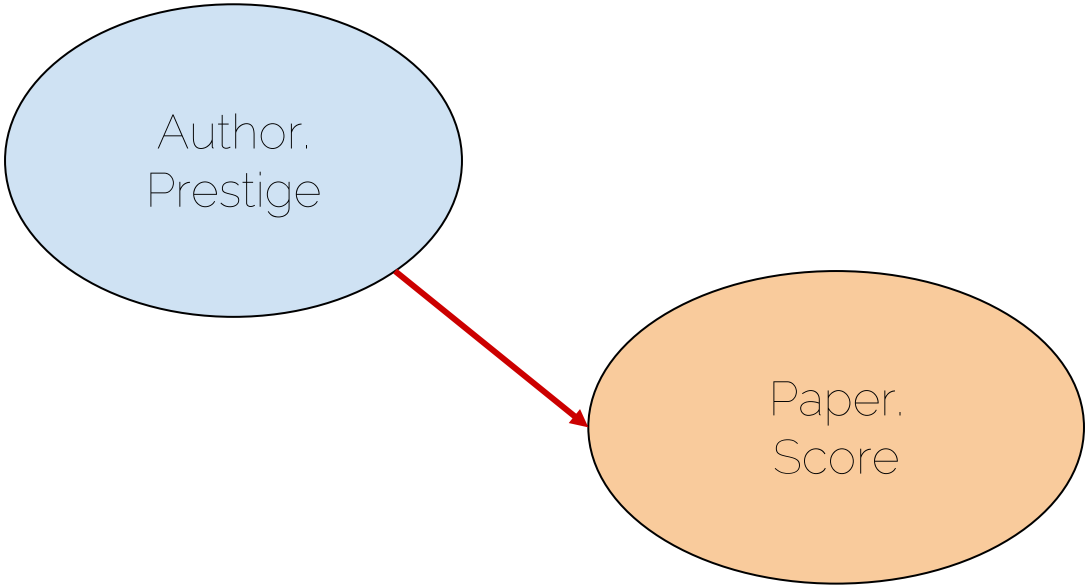
Problem: different number of parents – not easily captured – so use “embeddings”

$$\Pr\left(\mathtt{A}[\mathbf{x}] \mid \Psi^{\mathtt{A}}(\mathbf{Pa}(\mathtt{A}[\mathbf{x}]))\right) \tag{17}$$

e.g., average,
Or use a GNN

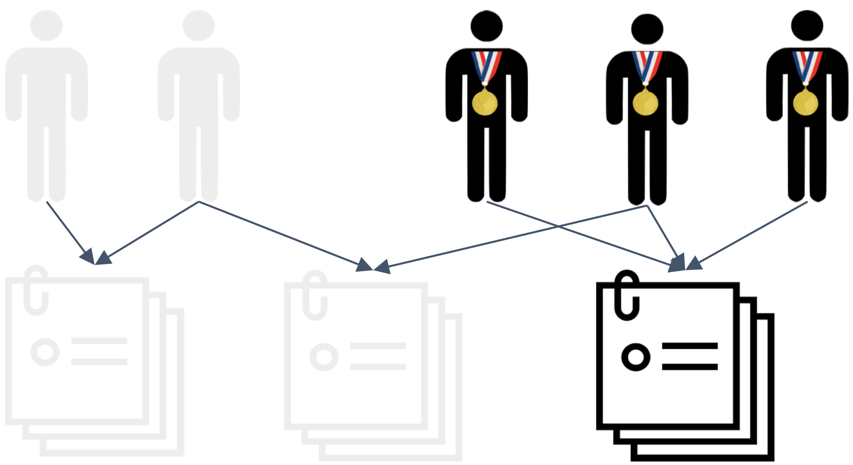
Causal Query

The Question of Interest



Make the Causal Query Well-defined

The Question of Interest



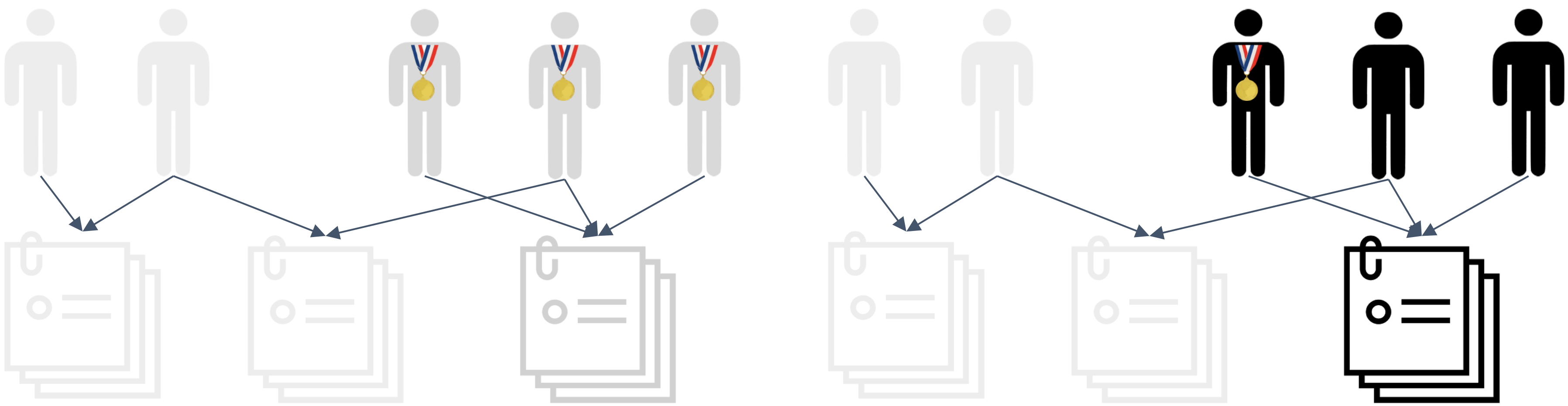
`Score [`S` ] ⇐ Prestige [`A`] ?`

`WHEN ALL AUTHOR TREATED`

Make the Causal Query Well-defined

Treatment and control "vectors" instead of scalars

The Question of Interest



Score [S] ⇐ Prestige [A] ?
WHEN ALL AUTHOR TREATED

Score [S] ⇐ Prestige [A] ?
WHEN AT LEAST 1 AUTHOR TREATED

$$\mathrm{ATE}(T, Y) \overset{\text{def}}{=} \sum_{\mathbf{x}' \in \mathbb{U}_Y} \frac{1}{m} (\mathbb{E}[Y[\mathbf{x}'] \mid \mathbf{do}(T[\mathbb{U}_T] = \vec{0})] - \mathbb{E}[Y[\mathbf{x}'] \mid \mathbf{do}(T[\mathbb{U}_T] = \vec{1})])$$

Collapsing multiple tables to single unit table

- Relational paths: Connect treated and response units using entities/relationships
  - At least one such path must exist

$$\textbf{Author}(A) \xleftrightarrow{\text{Author}(A,S)} \textbf{Submission}(S)$$

- Suppose
  - Treatment T[X] = Prestige[A] (on author)
  - outcome Y[X'] = Score[S] (on submission)

- Use aggregate rule:
- AVG_Score[S] <= Score[S] WHERE Author(A, S)
- New "attribute" in authors
- Similarly embed the other covariates

Relational, Isolated, and Overall Effects

- Given two intervention strategies

$(t, \vec{t})$ and $(t', \vec{t'})$ over n response units

$$\mathrm{AIE}(t; t' \mid \vec{t}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\mathbf{x} \in \mathrm{U}_{(T,Y)}} Y_{\mathbf{x}}(t, \vec{t}) - Y_{\mathbf{x}}(t', \vec{t})$$

Neighbors have same treatment

$$\mathrm{ARE}(\vec{t}; \vec{t'} \mid t) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\mathbf{x} \in \mathrm{U}_{(T,Y)}} Y_{\mathbf{x}}(t, \vec{t}) - Y_{\mathbf{x}}(t, \vec{t'})$$

Unit has same treatment

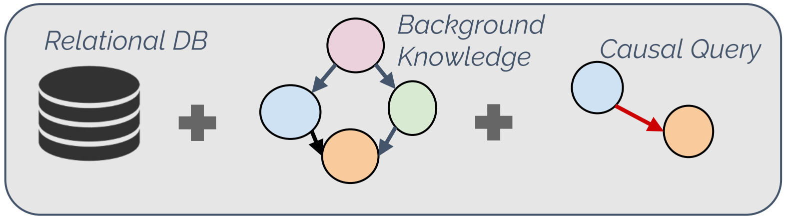
$$\mathrm{AOE}(t, \vec{t}; t', \vec{t'}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\mathbf{x} \in \mathrm{U}_{(T,Y)}} Y_{\mathbf{x}}(t, \vec{t}) - Y_{\mathbf{x}}(t', \vec{t'})$$

Both vary

AOE = AIE + ARE

CaRL – Causal Relational Learning

CaRL Methodology

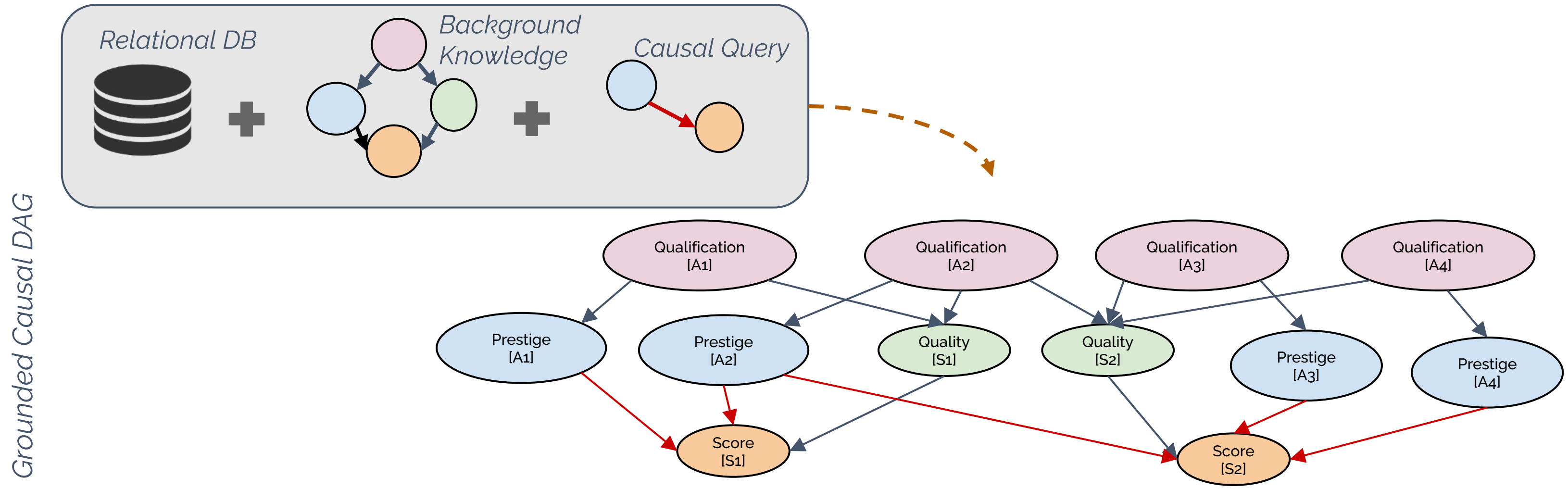


CaRL – Causal Relational Learning

CaRL Methodology

Skeleton Traversal

Grounding



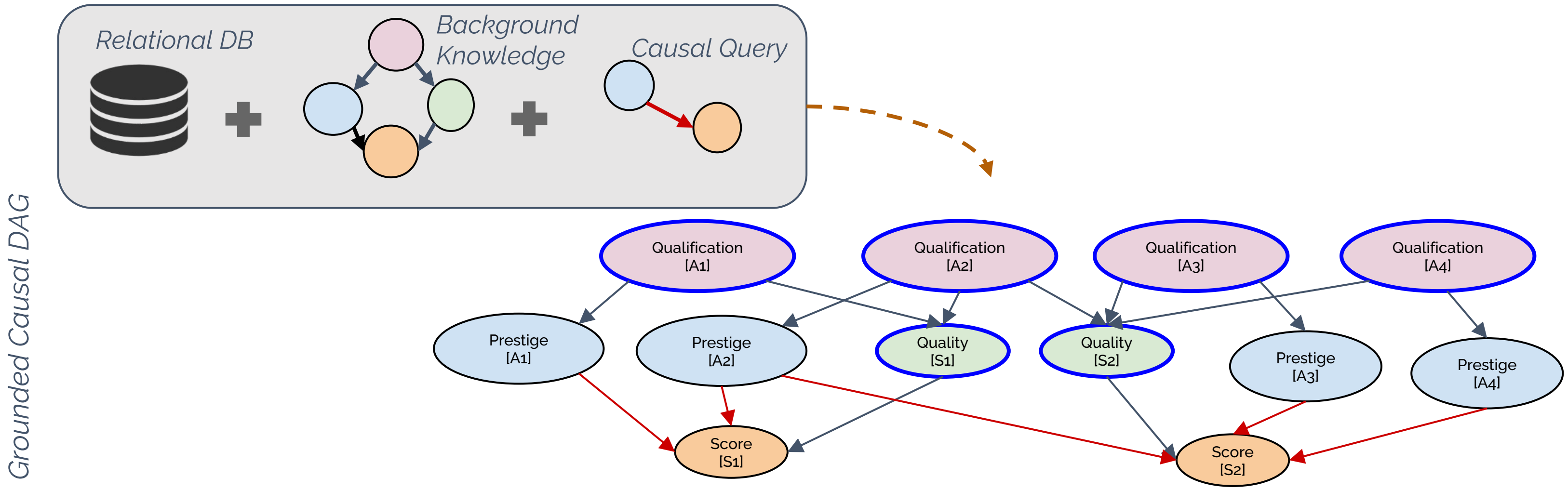
CaRL – Causal Relational Learning

CaRL Methodology

Skeleton Traversal

Grounding

Confounder Identification



CaRL – Causal Relational Learning

CaRL Methodology

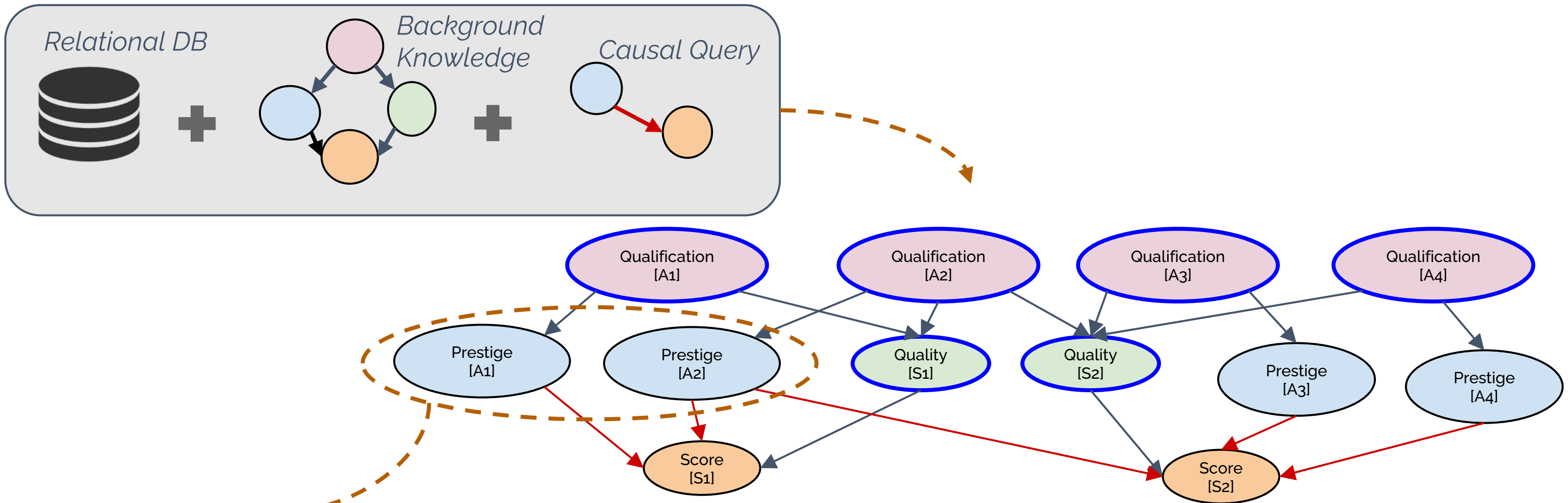
Skeleton Traversal

Grounding

Confounder Identification

Summary Functions

Grounded Causal DAG



Flat Table

Index	Summary-Prestige	Summary-Qualification	Score	Quality
S1	$g_{prestige}$(Prestige[A_1,A_2])	g_{qual}(Qualification[A_1,A_2])	Score[S1]	Quality[S1]
S2	$g_{prestige}$(Prestige[A_1,A_2,A_3])	g_{qual}(Qualification[A_1,A_2,A_3])	Score[S2]	Quality[S2]

CaRL – Causal Relational Learning

CaRL Methodology

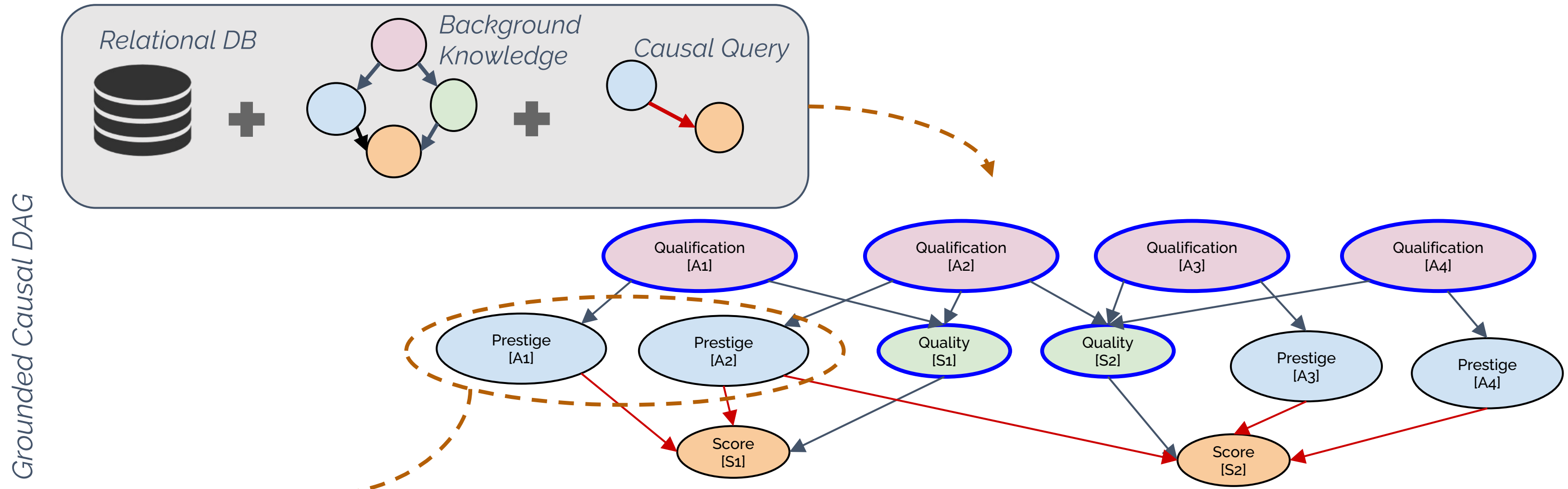
Skeleton Traversal

Grounding

Confounder Identification

Summary Functions

Causal Inference using standard methods



Flat Table

Index	Summary-Prestige	Summary-Qualification	Score	Quality
S1	$g_{prestige}$(Prestige[A_1,A_2])	g_{qual}(Qualification[A_1,A_2])	Score[S1]	Quality[S1]
S2	$g_{prestige}$(Prestige[A_1,A_2,A_3])	g_{qual}(Qualification[A_1,A_2,A_3])	Score[S2]	Quality[S2]

Causal Effect (ATE)

$$\sum_{\mathbf{x}' \in \mathbb{U}_Y} \frac{1}{m}(\mathbb{E}[Y[\mathbf{x}'] \mid \mathtt{do}(T[\mathbb{U}_T] = \vec{0})] - \mathbb{E}[Y[\mathbf{x}'] \mid \mathtt{do}(T[\mathbb{U}_T] = \vec{1})])$$

Data

OpenReview.net (Paper Review Data)

#Tables	3
#Attributes	7
#Row	6000
Time to Construct Unit-table	10.6s
Time to Answer Causal Query	1.2s

MIMIC (Hospital Stay Data)

#Tables	26
#Attributes	324
#Row	400 Million
Time to Construct Unit-table	6h
Time to Answer Causal Query	4.5h

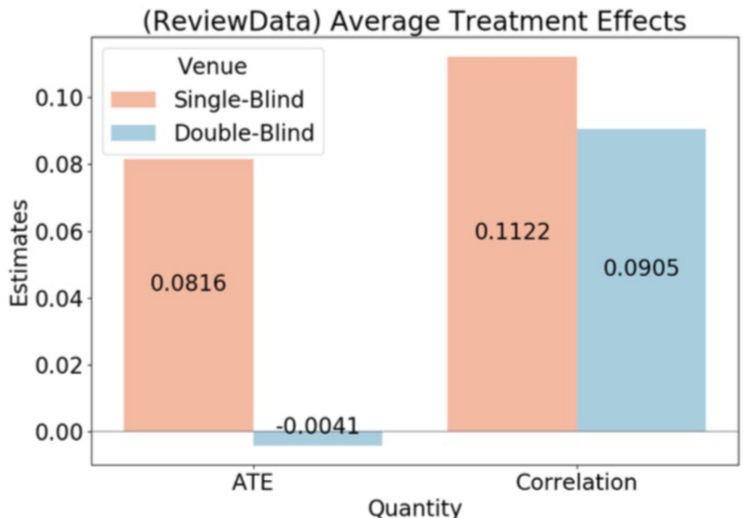
*More datasets and experiments in the paper

Sample Results: Correlation vs. Causation

Are reviewers influenced by authors' prestige?

OpenReview.net

≥ ⅓rd Authors Prestigious → Reviewer Score



(a)

Causation vs. Correlation

High correlation in both single and double blind
High causation only in single blind

(Maybe) Double Blind conferences, unlike Single Blind conferences, are successful in ensuring that the reviewers are not influenced by the prestige of the authors.

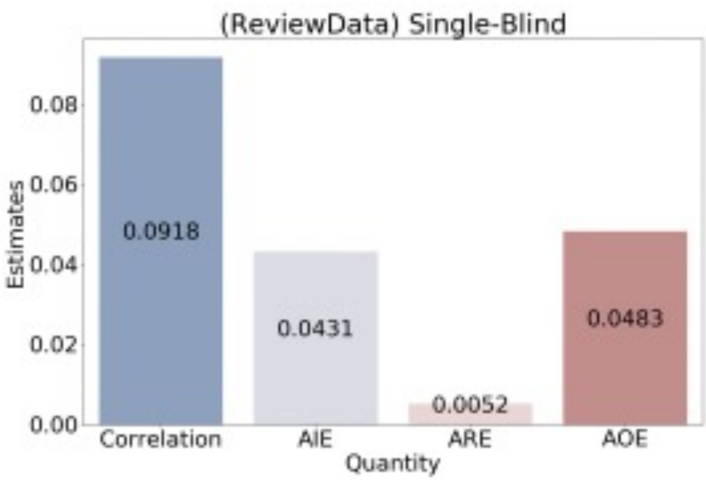
(control for authors' qualification)

Sample Results: AIE, ARE, AOE

OpenReview.net

Are reviewers influenced by authors' prestige?

≥ ⅓rd Authors Prestigious → Reviewer Score



Causation vs. Correlation

ARE less than AIE

he isolated effect (AIE) is more significant than the elational effect (ARE), meaning that an author's own estige has a stronger effect on his or her average ıbmission score than their collaborators' prestige has

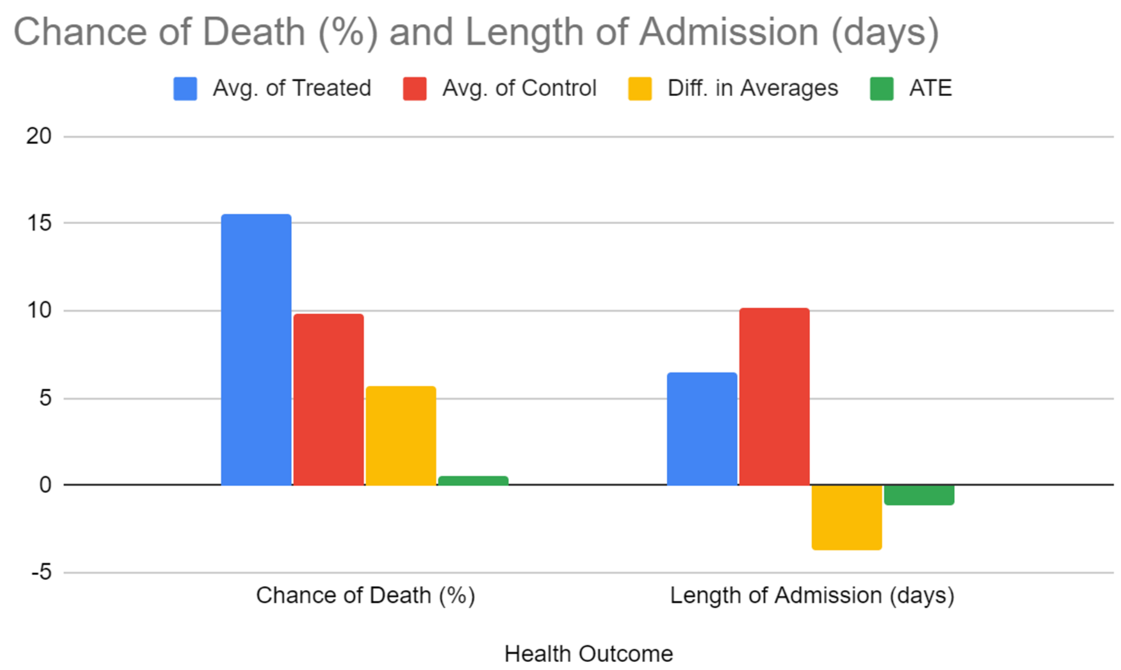
Sample Results: Correlation vs. Causation

MIMIC

Hospital Stay data

Does patients’ insurance plan affect health outcome?

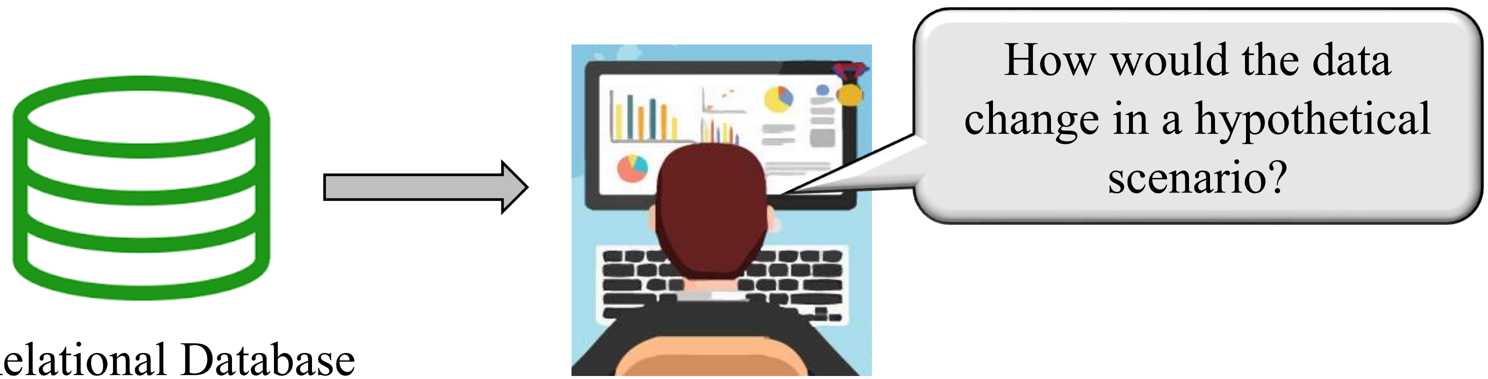
No Insurance → Health Outcome for Admitted Patient



(May be) Health outcomes of an admitted patient doesn’t depends on their insurance plan.
(control for severity and complications)

Application to hypothetical reasoning

Exploratory Data Analysis



- (What if) **What** will happen to X **if** Y changes in this way ....
- (How to) **How to** optimize X by tuning Y given some constraints…

e.g., [Deutch et al., CIDR'13, SIGMOD'19]

A Provenance Tracking / View Update Problem?

Suppose "Rating" could be computed by a query on Products involving "Price"

Products

PID	Category	Price	Brand	Color	Quality
1	Laptop	1099	Asus	Silver	0.7
2	Laptop	582	Asus	Black	0.65
3	Laptop	599	HP	Silver	0.5
4	DSLR	549	Canon	Black	0.75
5	eBook	15.99	Fantasy Press	Blue	0.4

Reviews

PID	RID	Sentiment	Rating
1	1	-0.95	?
2	1	-0.7	?
2	2	-0.2	?
3	1	0.23	3
3	3	0.95	5
4	4	0.7	4

Avg=?

What would be the average rating of **Asus** laptops if **Asus** price is increased by 10%?

amazon

- Isolate the attributes mentioned in the query Q(D)
- Use provenance of the query to update D → D' (what if)
- Recompute Q(D') as efficiently as possible

What about what-if on input data itself?
Price → Rating

There may be "causal dependencies" of attributes and tuples

Suppose "Rating" could be computed by a query on Products involving "Price"

Products

PID	Category	Price	Brand	Color	Quality
1	Laptop	1099	Asus	Silver	0.7
2	Laptop	582	Asus	Black	0.65
3	Laptop	599	HP	Silver	0.5
4	DSLR	549	Canon	Black	0.75
5	eBook	15.99	Fantasy Press	Blue	0.4

Reviews

PID	RID	Sentiment	Rating
1	1	-0.95	2
2	1	-0.7	4
2	2	-0.2	1
3	1	?	?
3	3	?	?
4	4	0.7	4

Avg=?

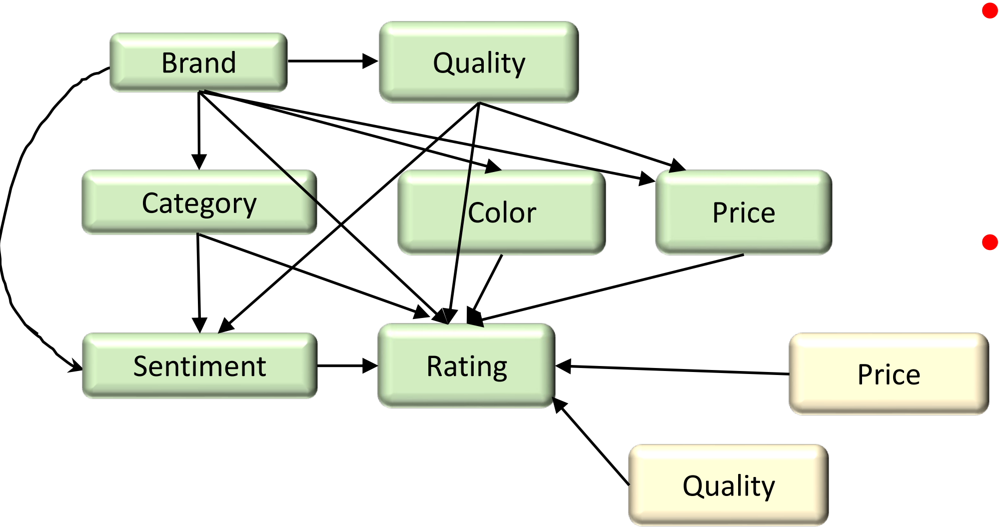
What would be the average ratings of **HP** laptop reviews **if** **Asus** price increases by 15%?

Even if Rating = Q(Price,..)

Disjoint provenance – but indirect effect

[Galhotra-Gilad-Roy-Salimi, SIGMOD'22]

Capture dependencies by Causal Graphs



- Intra-tuple dependencies
  - Brand ⟶ Price, Rating, ….
- Inter-tuple dependencies
  - Product's Rating depends on competitor's pricing

(Again) Relational/grounded Causal graph

HypeR - Hypothetical Reasoning: What-if Syntax

SQL query to compute a "view" for what-if (with multiple tables)

$\text{USE } \textit{RelevantView} \text{ AS}$

$\quad (\text{SELECT } T1.PID, T1.Category, T1.Price, T1.Brand,$

$\quad\quad \text{AVG}(Sentiment) \text{ AS } Senti, \text{AVG}(T2.Rating) \text{ AS } Rtng$

$\quad \text{FROM } Product \text{ AS } T1, Review \text{ AS } T2$

$\quad \text{WHERE } T1.PID = T2.PID$

$\quad \text{GROUP BY } T1.PID, T1.Category, T1.Price, T1.Brand)$

$\text{WHEN } Brand =' \, Asus'$

$\text{UPDATE}(Price) = 1.1 \times \text{PRE}(Price)$

$\text{OUTPUT AVG}(POST(Rtng))$

$\text{FOR PRE}(Category) = \text{`}Laptop\text{'} \text{ AND PRE}(Brand) = \text{`}Asus\text{'}$

$\text{AND POST}(Senti) > 0.5$

Hypothetical update in the database

Condition to select which attributes to update

Output of what-if query

Selection conditions for which the output of the query will be computed

Pre: pre-update, Post = post-update

How-to query has a similar syntax
An optimization on satisfying what-if queries

HypeR - Hypothetical Reasoning: What-if Semantics

Use concepts from Probabilistic Databases "Possible Worlds" – all possible database instances from the domains
Each possible world W has a post-update distribution $Pr_{D,U}(W)$

```
USE RelevantView AS
    (SELECT T1.PID, T1.Category, T1.Price, T1.Brand,
       AVG(Sentiment) AS Senti, AVG(T2.Rating) AS Rtng
    FROM Product AS T1, Review AS T2
    WHERE T1.PID = T2.PID
    GROUP BY T1.PID, T1.Category, T1.Price, T1.Brand)
WHEN Brand =' Asus'
UPDATE(Price) = 1.1 × PRE(Price)
OUTPUT AVG(POST(Rtng))
FOR PRE(Category) = 'Laptop' AND PRE(Brand) = 'Asus'
AND POST(Senti) > 0.5
```

For tuples satisfying pre-update "WHEN" condition, apply hypothetical "UPDATE" U in the causal model

For possible worlds W satisfying "FOR" condition, compute the attribute value specified in "OUTPUT": val(W)

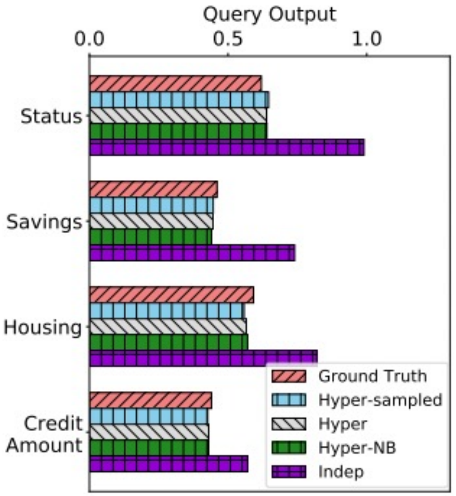
Answer of what-if: $\sum_W val(W) * Pr_{D,U}(W)$

Reduce to observed probabilities by using Relational Causal Methods (control for confounders)

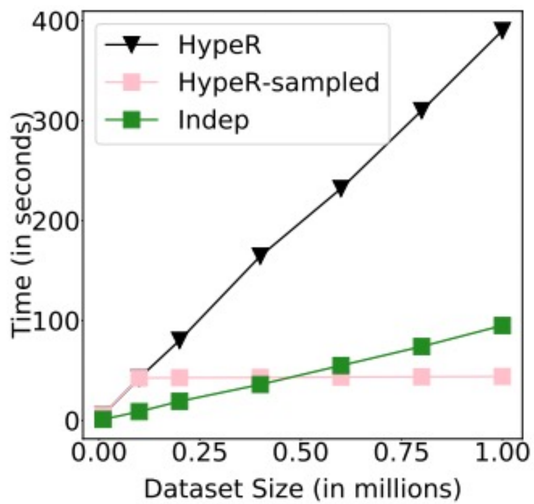
Naïve computation inefficient – reduces to simpler (poly-time) formulas for many what-if queries
Also use optimizations like "block-independent decomposition" in the causal graph

Sample Results: What-if analysis with HyPeR

Using causal graphs estimates "what-if" better
Even if background knowledge is unavailable
HypeR-Sampled is highly efficient



Solution quality



Running time

Reducing prices increase product ratings:

If set to lower price, higher rating

Amazon review dataset

German semi-synthetic dataset

Reviewing paper reviews!

Time-series causal analysis

- how such the synthetic control method would need to change with non-stationary covariates. My initial thoughts would be that we should consider trends and seasonality within the covariates to determine whether these could be modeled as part of the synthetic generation process
- their variance is not guaranteed to be low
- It is impossible to account for all the factors that could contribute to tobacco use, so it is difficult to say with confidence that the states used to create synthetic California would continue to reliably replicate trends in California post-1988
- The synthetic control method is highly data-driven and specific to the characteristics and policies of the state being studied.One potential area of improvement is the selection of control variables
- Perhaps, determining whether the gap in post-intervention outcomes between affected and unaffected groups is statistically significant is a better way of testing robustness.

Time-series causal analysis

- Although it is likely that fast food chains in a state are faced with similar business conditions, it is also highly likely that each store face different local business conditions. Controlling variables associated with their local business conditions such as the income distribution of neighbors and the density of competitors would make a better comparison.
- there is no intuition for why a reader should believe that the combination of Nevada, Utah, Colorado, and Vermontthat replicated California pre-1989 should generalize. The culture of Utah is vastly different than California And Vermont is further away geographically than parts of Mexico and Canada. Giving their model access to so many degrees of freedom without penalization, I am unconfident in their justification of these states has the correct "synthetic California”
- this paper did not examine the long-term effects of the policy.
- this study is not broadly generalizable
- Furthermore, future studies must also consider that the economies of Eastern Pennsylvania andNew Jersey are connected, as actors in these economies can act in both economies and I think there should be more proof demonstrated that the control group will not be affected by the treatment

Almost Exact Matching

- Categorical variable only
- High dimensionality not scalable or effective
- What happens when covariates are highly correlated
- More study on stopping condition
- Computationally expensive (DAME)

Instrumental Variable

- Continuous IV
- Automate selection of IV
- Robustness to skewness and outliers
- The LATE approach assumes that the IV has a monotonic relationship with the treatment. However, thisassumption may not hold in some cases
- One limitation of this paper is that the result is using the additive interpretability model like SHAP to calculatethe marginal contribution of features in each unit. However, this kind of interpretable method is not showingthe causal effect in the model like LEWIS as we mentioned in class. If the covariates are not independent, likeage and gender are related to some of the middle hidden layers for the other covariates. The explanation of thecovariates is not reasonable

GNN & GNNExplainer

- No baseline yet
- The GNNExplainer claims to beinterpretable before the conclusion, but it does not induce interpretable GNN modeling, its explanations are not guaranteed to be human interpretable, and does not even aim to explainbroader relationships in the graph structure (beyond individual connections)
- it requires access to internalmodel parameters and computations of the GNN, so it may not be applicable toGNN architectures or models which do not expose these internal details
- High computational cost
- Not heterogenous graph friendly

Interpretability vs. Explainability

- it assumes thatthe robust and non-robust features (or causal and spurious correlations) can be decomposedfrom an image X. This is probably not always the case, as differences in image domain extendbeyond simply the background of an image and may be more subtle in nature

- it assumes the availability of high-quality causalmodels for each data source. In practice, constructing such models can be challenging,especially when dealing with complex systems or limited data

- heavy dependence on the assumption that thelatent representation of an image learnt by a machine learning model contains all the causal informationfrom the input image.