Causal Data Integration

Brit Youngmann

11.17.2023

Joint work with Michael Caffarela, Babak Salimi, and Anna Zeng (thanks for slides!)

TECHNION
Israel Institute
of Technology

In this talk

- Correlation Explanation

- Causal Explanations

- Causal Data Integration (vision)

Explaining Aggregate Query Result

- Aggregate SQL queries expose correlations.
- *Confounding bias*.

[**Youngmann**-Cafarella-Moskovitch-Salimi ICDE’22, SIGMOD’22 paper & demo]

Hungry Judges?

RESEARCH ARTICLE | SOCIAL SCIENCES |

Extraneous factors in judicial decisions

Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso Authors Info & Affiliations

Edited* by Daniel Kahneman, Princeton University, Princeton, NJ, and approved February 25, 2011 (received for review December 8, 2010)

April 11, 2011 | 108 (17) 6889-6892 | https://doi.org/10.1073/pnas.1018033108

LETTER |

Overlooked factors in the analysis of parole decisions

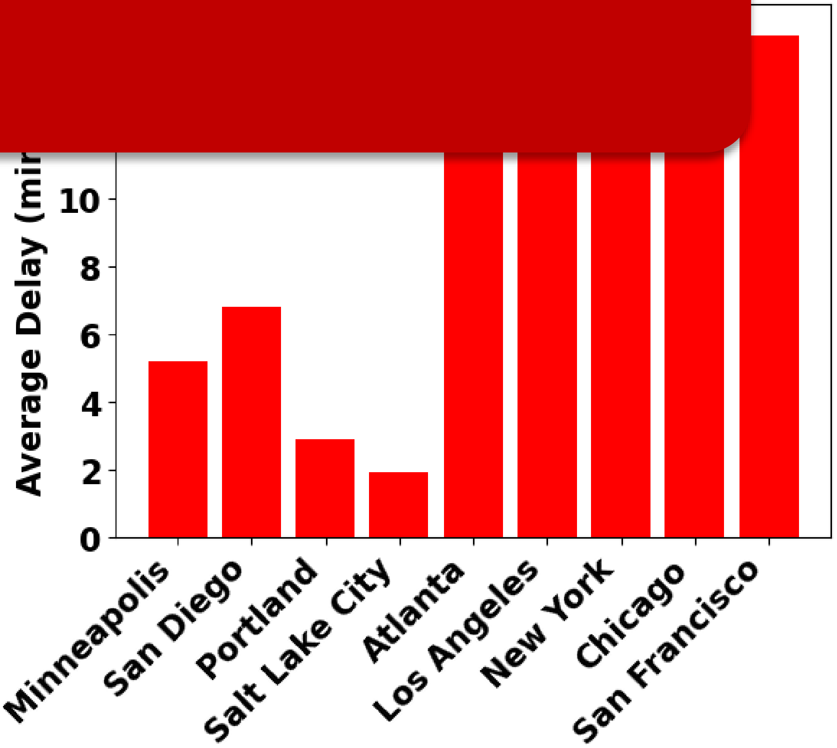
Keren Weinshall-Margel and John Shapard Authors Info & Affiliations

October 10, 2011 | 108 (42) E833 | https://doi.org/10.1073/pnas.1110910108

Departure City	Airport	Airline	Date	Arrival City	Duration	Departure Delay	Arrival Delay
New York	EWR	AA	1/1/2015	Chicago	181	14.26	80.64
New York	EWR	UA	2/1/2015	Boston	68	13.71	36.86
Chicago	ORD	WN	4/6/2015	Seattle	540	7.13	76.82
Atlanta	ATL	B6	7/1/2015	Huston	164	4.41	91.89

Why does the choice of departure city have such a substantial effect on the departure delay?

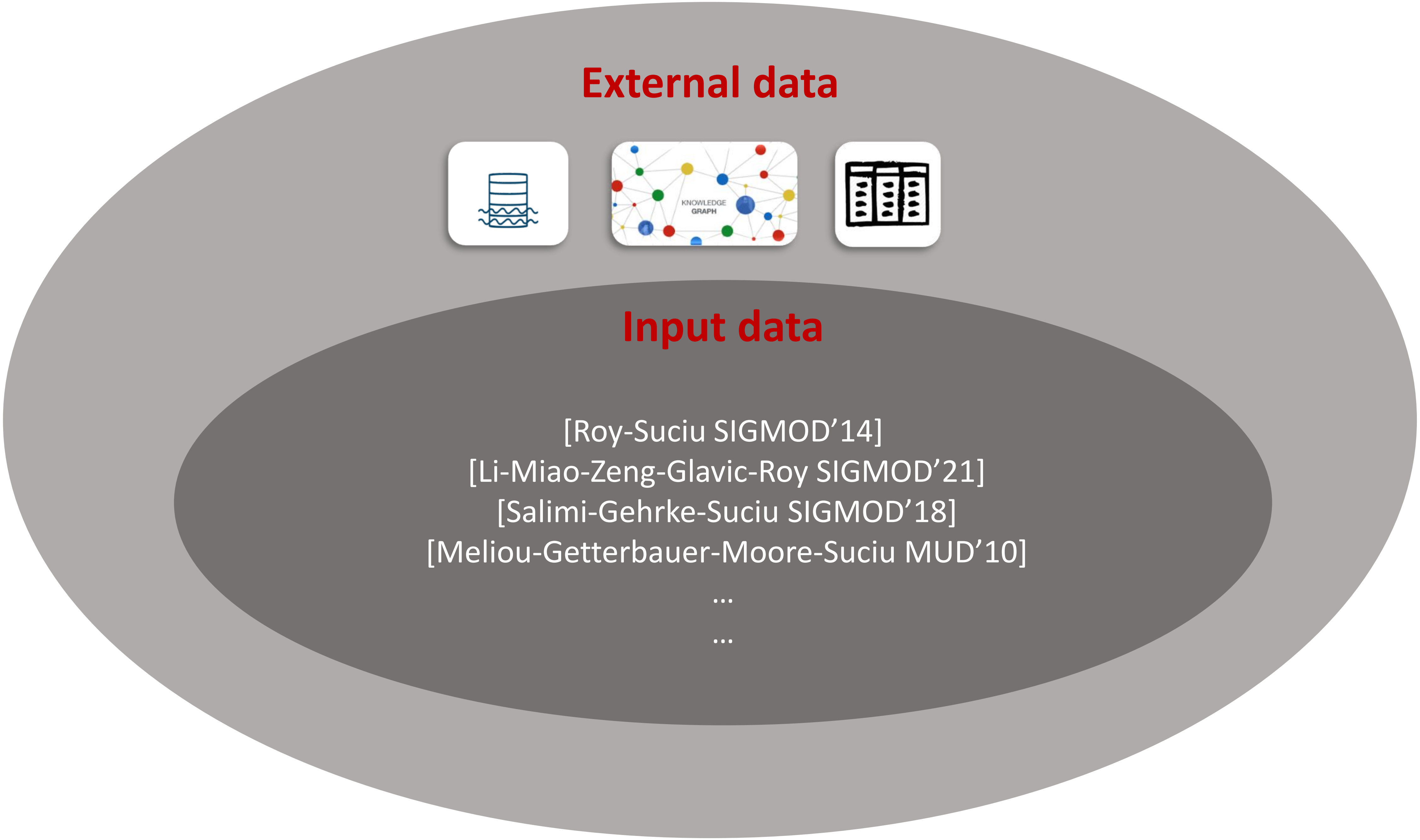
```
SELECT City, AVG(Departure-delay)
FROM Flight-Data
GROUP-BY City
```



Key Observation

Important variables **are missing** from the data:

- Weather-related variables (Temperature, Precipitation, Visibility)
- Airport Traffic (Population Density, Population Size)
- Airline Carrier (Fleet Size, Num of Employees)
- …


External data
KNOWLEDGE GRAPH
Input data
[Roy-Suciu SIGMOD'14]
[Li-Miao-Zeng-Glavic-Roy SIGMOD'21]
[Salimi-Gehrke-Suciu SIGMOD'18]
[Meliou-Getterbauer-Moore-Suciu MUD'10]
…
…

Explaining Unexpected Correlations

- We automatically mine unobserved variables from knowledge graphs.
- We then identify a subset of variables that minimize the partial correlation.

Departure Delay

Departure City

Given a set of candidate attributes $\mathcal{A}$ and a query Q, find a minimal-size set of attributes $\mathbf{E} \subseteq \mathcal{A}$ that minimizes $\mathrm{I}(\mathrm{O};\mathrm{T}|\mathbf{E})$

A Greedy Algorithm

- Min-Conditional-mutual-Information:

$$E_k = argmin_{E_i \in \mathcal{A}}[I(O;T|E_i)]$$

Individual contribution

- Min-Redundancy:

$$E_k = argmin_{E_i \in \mathcal{A}}[\frac{\sum_{\{E_j \in \boldsymbol{E}\}} I(E_i;E_j)}{k-1}]$$

Redundancy

- The *k*-th attribute to be added:

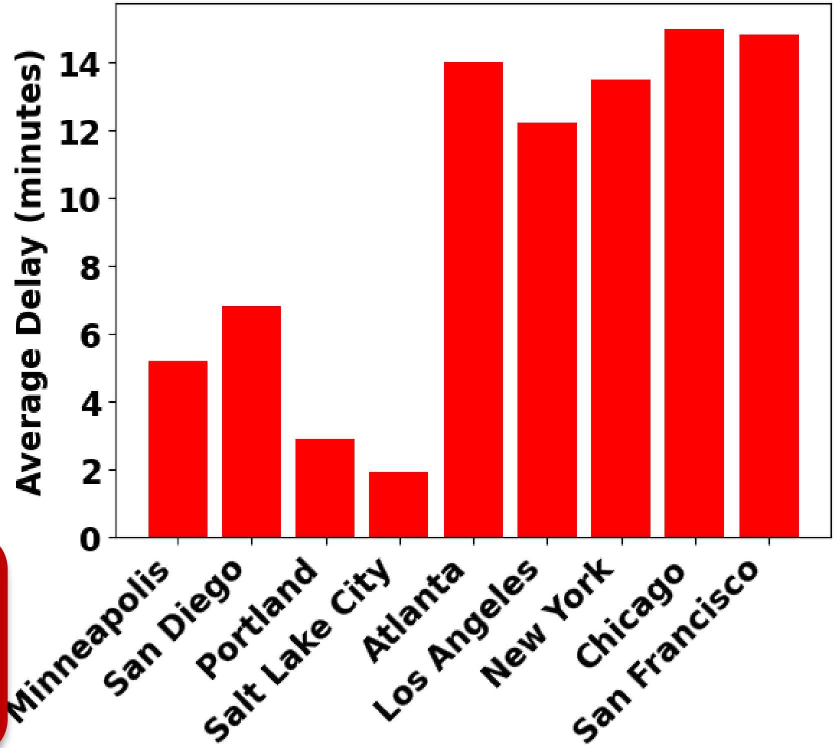
$$E_k = argmin_{E_i \in \mathcal{A}}\left[I(O;T|E_i) + \frac{\sum_{\{E_j \in \boldsymbol{E}\}} I(E_i;E_j)}{k-1}\right]$$

Departure City	Airport	Airline	Date	Arrival City	Duration	Departure Delay	Arrival Delay
New York	EWR	AA	1/1/2015	Chicago	181	14.26	80.64
New York	EWR	UA	2/1/2015	Boston	68	13.71	36.86
Chicago	ORD	WN	4/6/2015	Seattle	540	7.13	76.82
Atlanta	ATL	B6	7/1/2015	Huston	164	4.41	91.89
Portland	PDX	AA	2/2/2015	New York	612	3.45	30.9

SELECT City, AVG(Departure-delay)

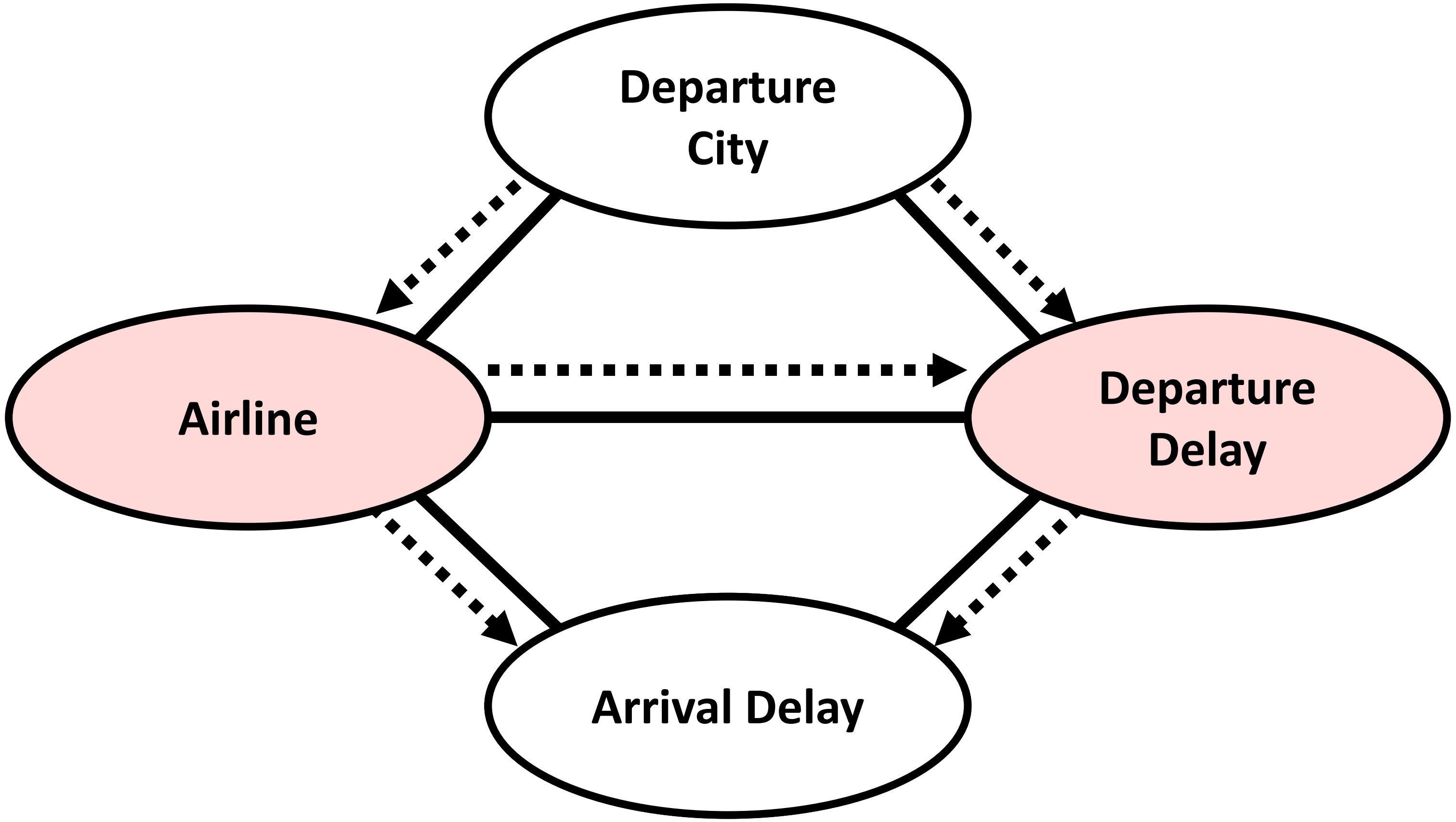
FROM Flight-Data

GROUP-BY City



Explanation: Population size, Year Low F, Airline

Association vs. Causation



Causal Explanations

What factors influence the salaries of developers in various countries?

What factors affect departure delays in various cities?

How much does having a Master's degree impact the salary of developers in the US?

How much does the choice of an airline carrier affect flight departure delay?

[**Youngmann**-Cafarella-Gilad-Roy, Under-Submission]

Table 1: A subset of the Stack Overflow dataset.

ID	Country	Continent	Gender	Ethnicity	Age	Role	Education	YrsCoding	Undergrad Major	Salary
1	US	North America	Male	White	26	Data Scientist	PhD	10	Computer Science	180k
2	US	North America	Non-binary	White	32	QA developer	Bachelor's degree	5	Mechanical Eng.	83k
3	India	Asia	Male	South Asian	29	C-suite executive	Bachelor's degree	8	Computer Science	24k
4	India	Asia	Female	South Asian	25	Back-end developer	Master's degree	9	Mathematics	7.5k
5	China	Asia	Male	East Asian	21	Back-end developer	Bachelor's degree	7	Computer Science	19k

```
SELECT Country, AVG(Salary)   FROM Stack-Overflow
GROUP BY Country
```

- For countries in Europe, the most substantial positive effect on salaries (effect size of 36K) is observed for individuals under 35 with a Master's degree. Conversely, being a student has the strongest adverse impact on income (effect size -39K)
- For countries with a high GDP level….
- …

Causal Explanations

Given a database *D*, a group-by-avg query *Q*, a number *k,* and a threshold θ, find a set of explanation Φ such that:

- Φ contains no more than *k* explanations
- Φ explains at least θ-fraction of the groups in *Q(D)*
- No redundancy in Φ
- The overall explainability of Φ is maximized

[**Youngmann**-Cafarella-Gilad-Roy, Under-Submission]

Causal Explanations

How much does the choice of an <u>airline carrier</u> affect <u>flight departure delay</u>?

[Pearl. 2009. *Causality*. Cambridge University Press.]

Causal Effect

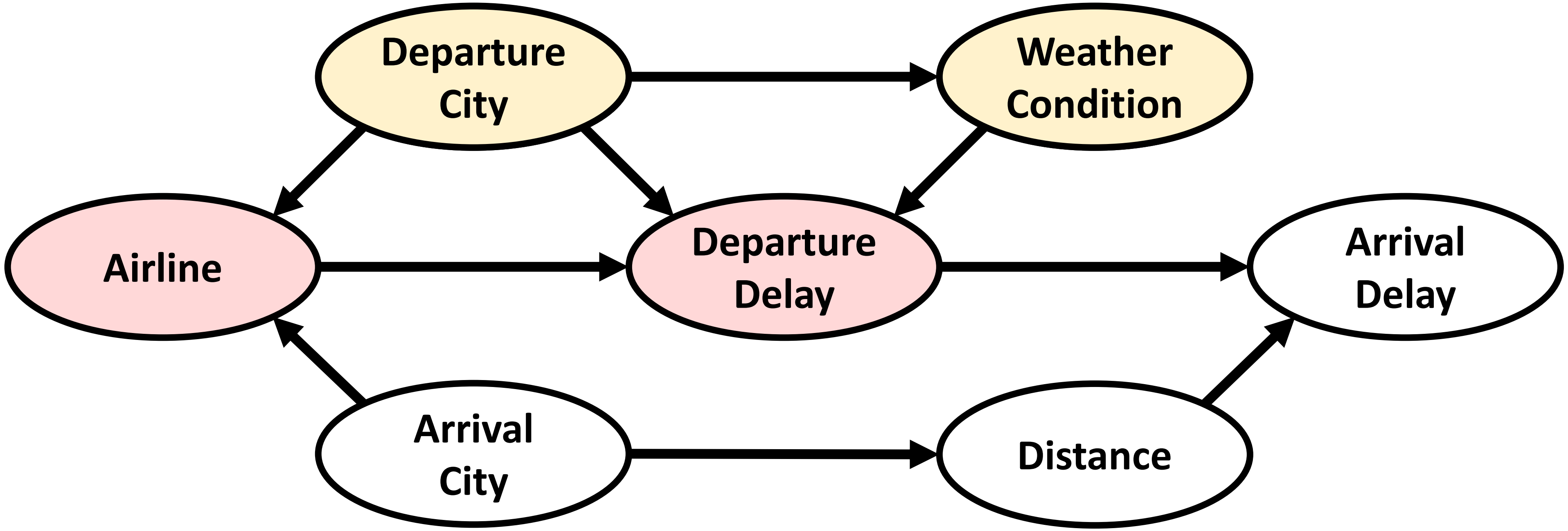
Randomized experiment

$ATE(T, O)$ = E[O | do(T = 1)] − E[O | do(T = 0)]

Observational data

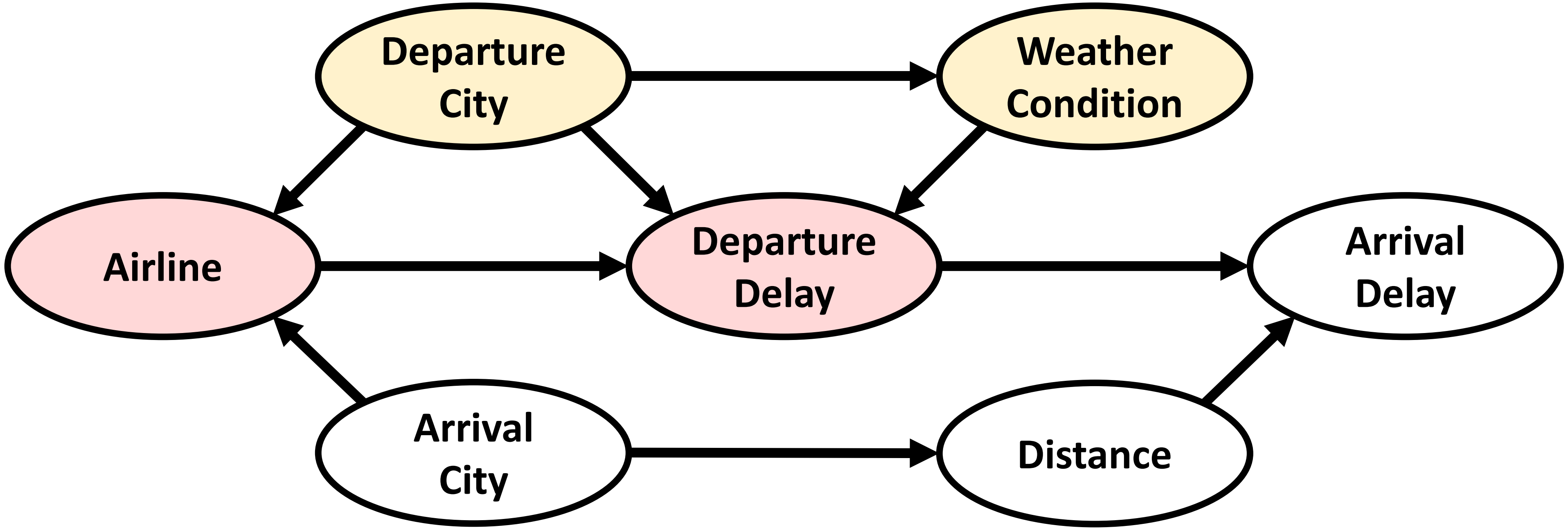
$ATE(T, O)$ = E[O | T = 1, Z = z] − E[O | T = 0, Z = z]

How much does the choice of an airline carrier affect flight departure delay?



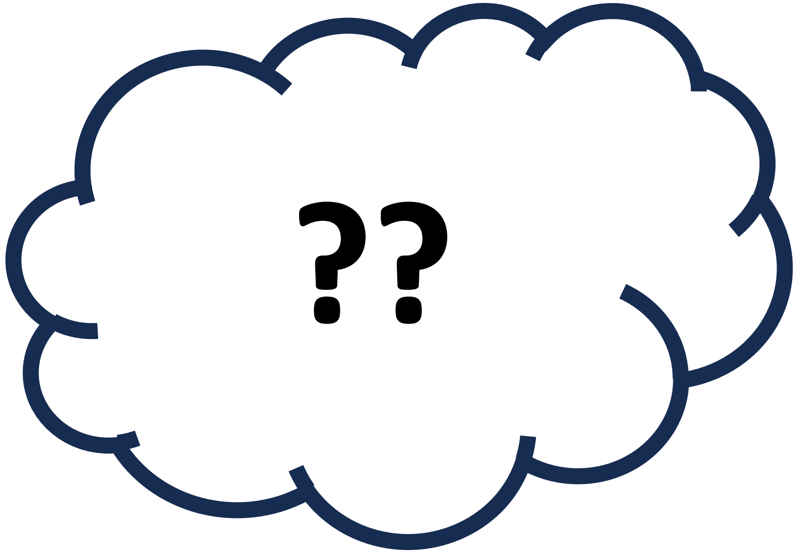
Providing Causal Explanations

Flight No.	Departure City	Arrival City	Airline	Distance	Departure Delay	Arrival Delay	Weather Condition
12343	Boston	New York	AA	232	15	11	Heavy snow
77654	Phoenix	Seattle	UA	1415	5	0	rain
86658	New York	Boston	B6	232	4	0	clear
66553	Chicago	New York	UA	790	18	9	clear
77642	Boston	Seattle	AA	3045	28	15	thunderstorm

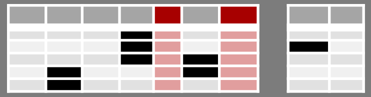


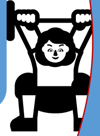
Providing Causal Explanations *In Practice*

Flight No.	Departure City	Arrival City	Airline	Departure Delay	Arrival Delay
12343	Boston	New York	AA	15	11
77654	Phoenix	Seattle	UA	5	
86658	New York		B6	4	
66553			UA	18	9
77642	Boston			28	15

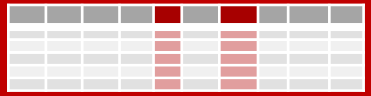
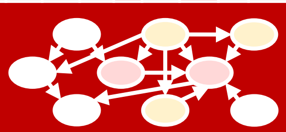


Usually, in practice, we have:

1.  ~~A single and complete~~ relational database(s)
2. ~~A correct causal DAG~~

Causal Data Integration 

To answer causal questions, we need:

1.  A single and complete relational database
2.  A correct causal DAG

Challenges

1. ***Completeness:*** missing columns can lead to confounding bias

2. ***Robustness:*** handling data issues conventionally can result in erroneous conclusions

   Missing values; outliers; wrong values; dependencies;…

3. ***Conciseness:*** the causal DAG can't be verified if it is too complex

Completeness: Datasets might be missing columns

Flight No.	Departure City	Arrival City	Airline	Distance	Departure Delay	Arrival Delay	Passenger Sentiment
12343	Boston	New York	AA	232	15	11	low
77654	Phoenix	Seattle	UA	1415	5	0	medium
86658	New York	Boston	B6	232	4	0	medium
66553	Chicago	New York	UA	790	18	9	high
77642	Boston	Seattle	AA	3045			

Provenance matters!

- **What is the source?**
- **How the extractor operates?**

Completeness: Datasets might be missing columns

Flight No.	Departure City	Arrival City	Airline	Distance	Departure Delay	Arrival Delay	Weather Condition
12343	Boston	New York	AA	232	15	11	Heavy snow
77654	Phoenix	Seattle	UA	1415	5	0	rain
86658	New York	Boston	B6	232	4	0	clear
66553	Chicago	New York	UA	790	18	9	clear
77642	Boston	Seattle	AA	3045	28	15	thunderstorm

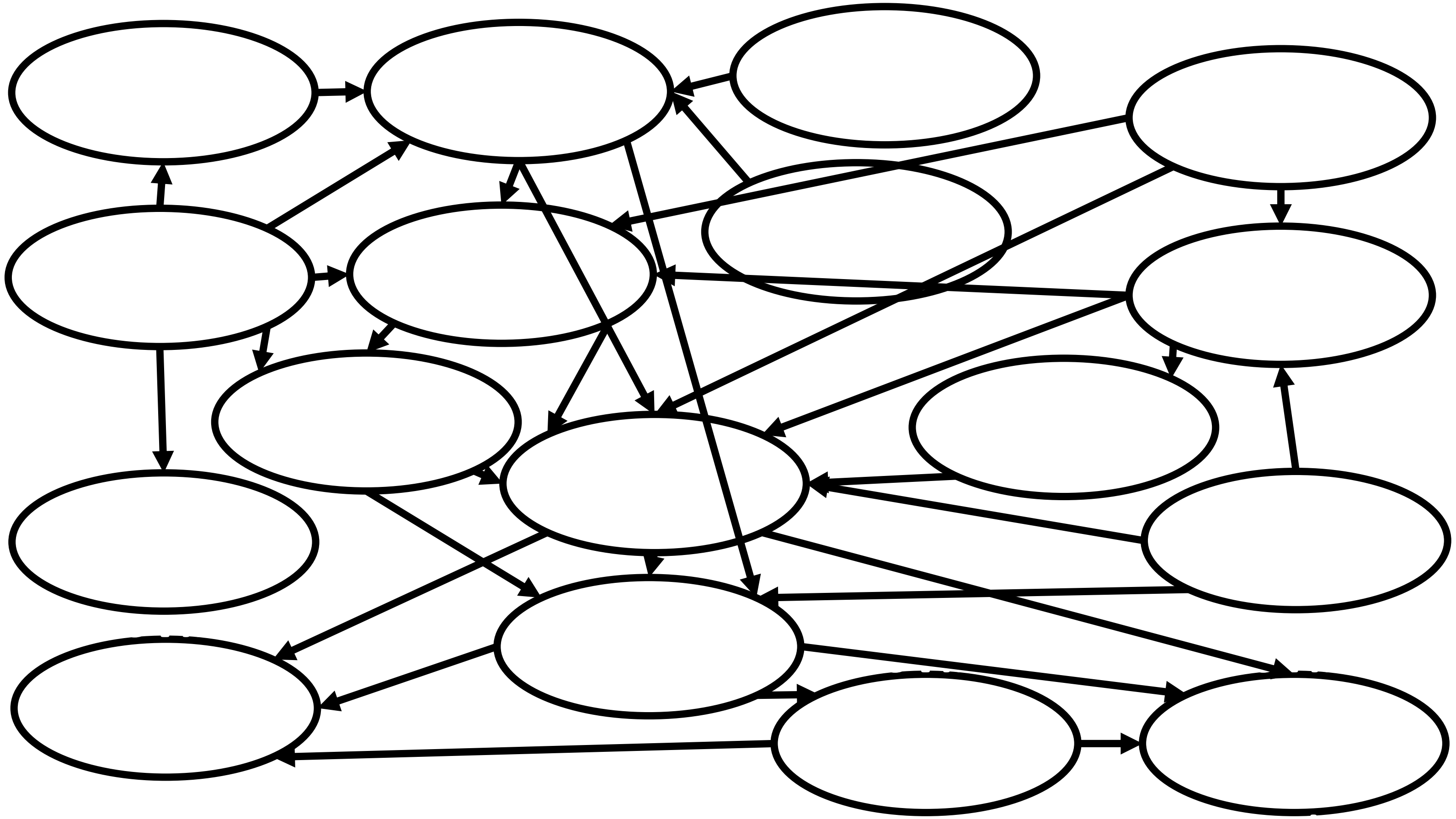
Robustness: Resolving data issues can result with wrong conclusions

Flight No.	Departure City	Arrival City	Airline	Distance	Departure Delay	Arrival Delay	Weather Condition
12343	Boston	New York	AA	232	15	11	Heavy snow
77654	Phoenix	Seattle	UA	1415	5	0	
86658	New York	Boston	B6	232	4	0	
66553	Chicago	New York	UA	790	18	9	
77642	Boston	Seattle	AA	3045	28	15	thunderstorm

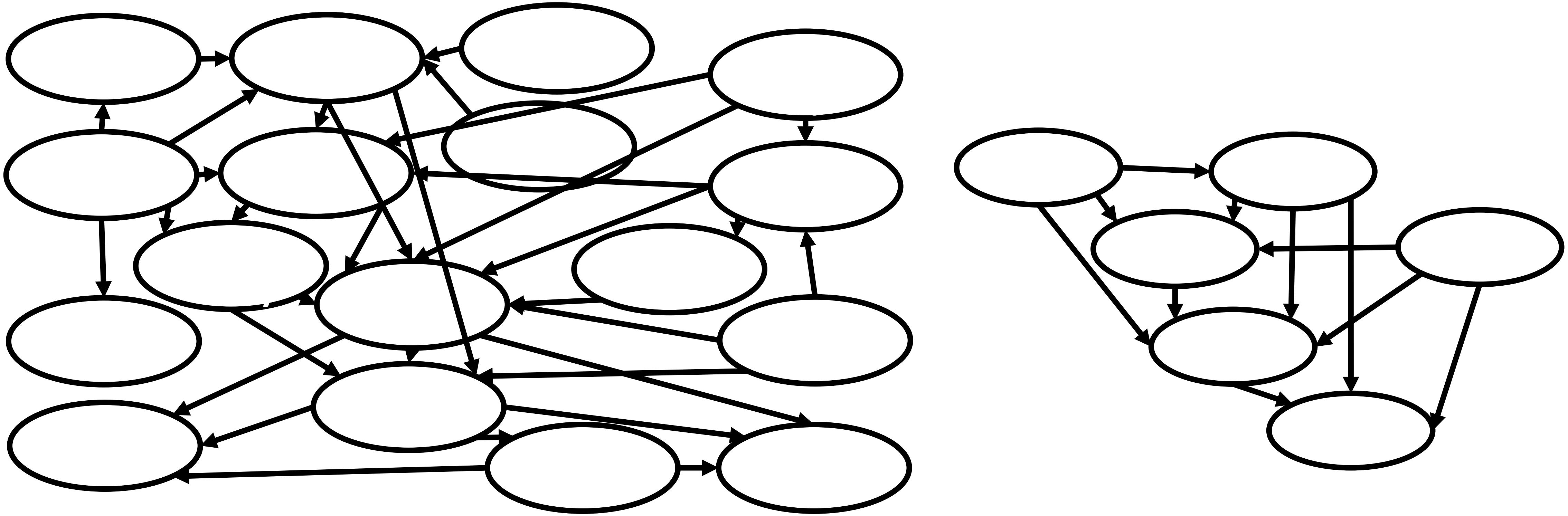
Selection bias: a skewed result due to a non-representative observed population

[Fariha-Moskovitch-**Youngmann**, …]

Conciseness: The causal DAG might be incorrect



Conciseness: so let's summarize the causal DAG



[Zeng-Cafarella-Salimi-**Youngmann**, …]

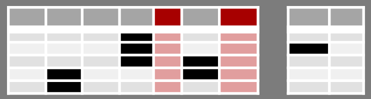
Causal DAG Summarization

Given a causal DAG $\mathrm{G} = (V, E)$, a number k, and a threshold τ find a *summarized causal DAG* G_P such that:

- [**size constraint**] G_P has no more then k nodes
- [**semantic constraint**] the inter-cluster semantic similarity of every node (cluster) in G_P is above τ
- [**causal information**] G_P retains the maximum amount of causal information from G (in terms of *recoverable* conditional independence statements).

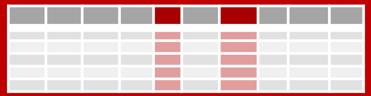
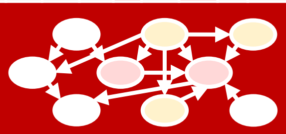
- **The number of CIs is exponential**
- **How to read off a summarized DAG all the CIs**

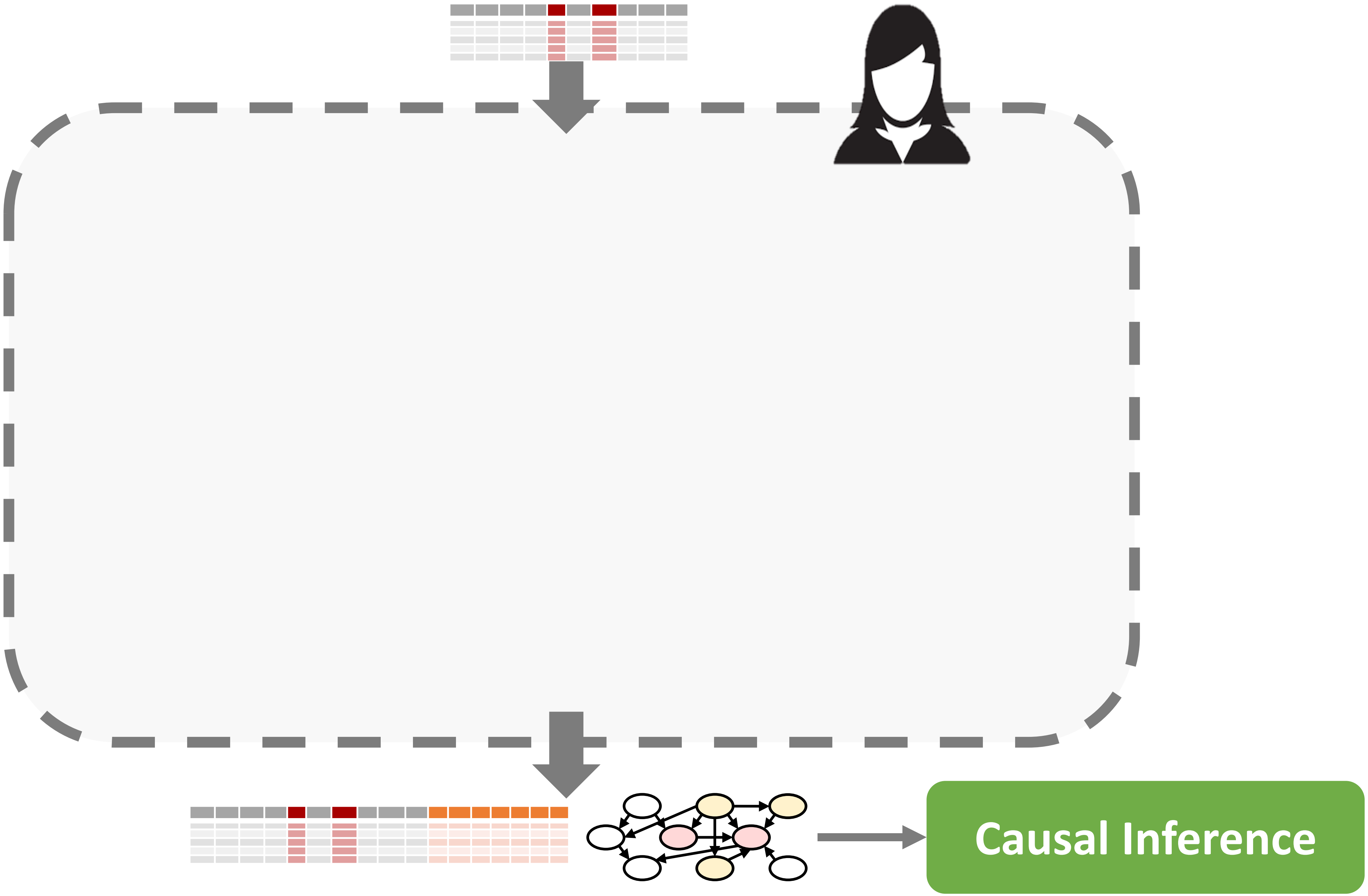
Usually, in practice, we have:

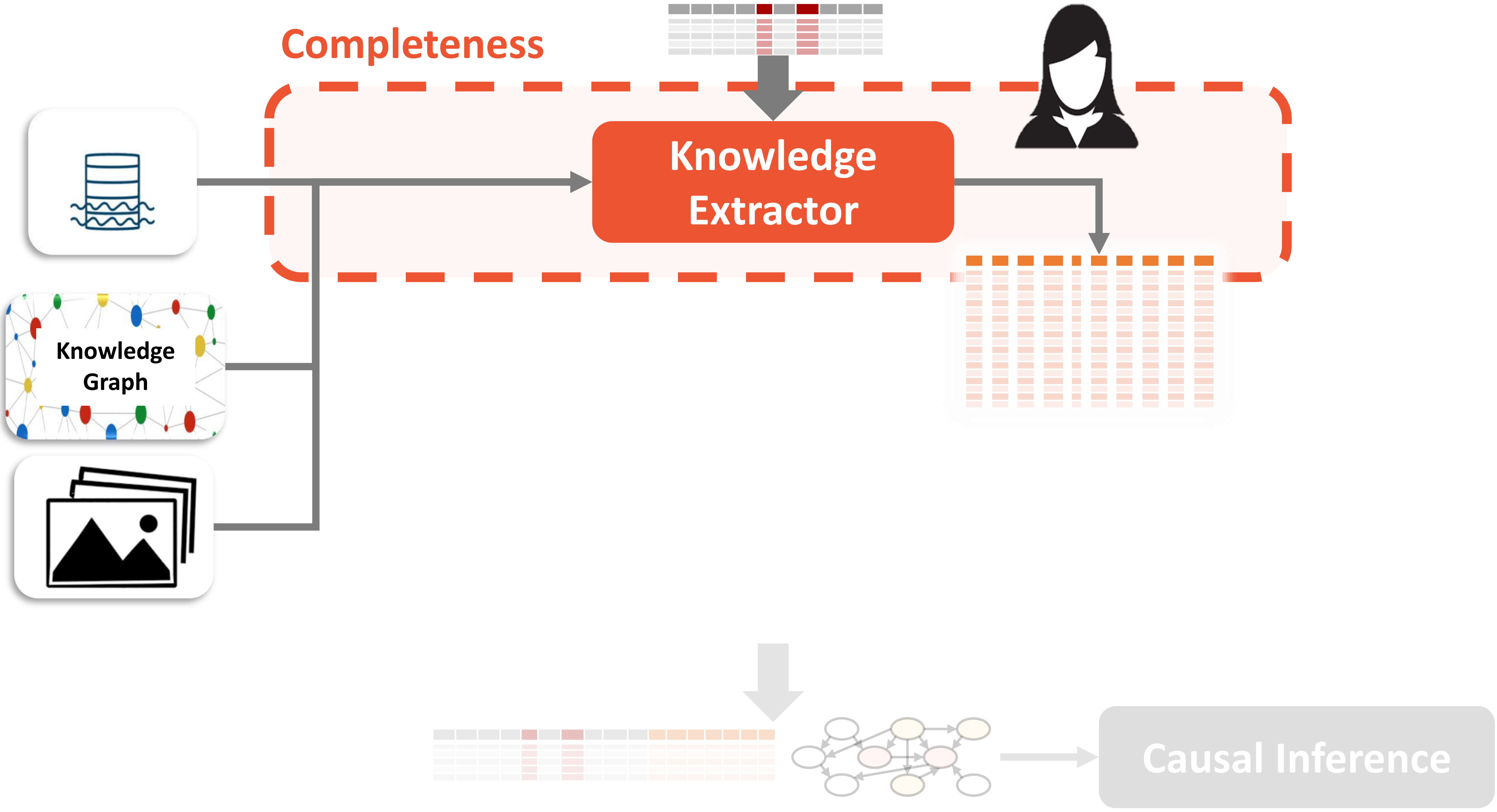
1.  ~~A single and complete~~ relational database(s)
2. ~~A correct causal DAG~~

Causal Data Integration

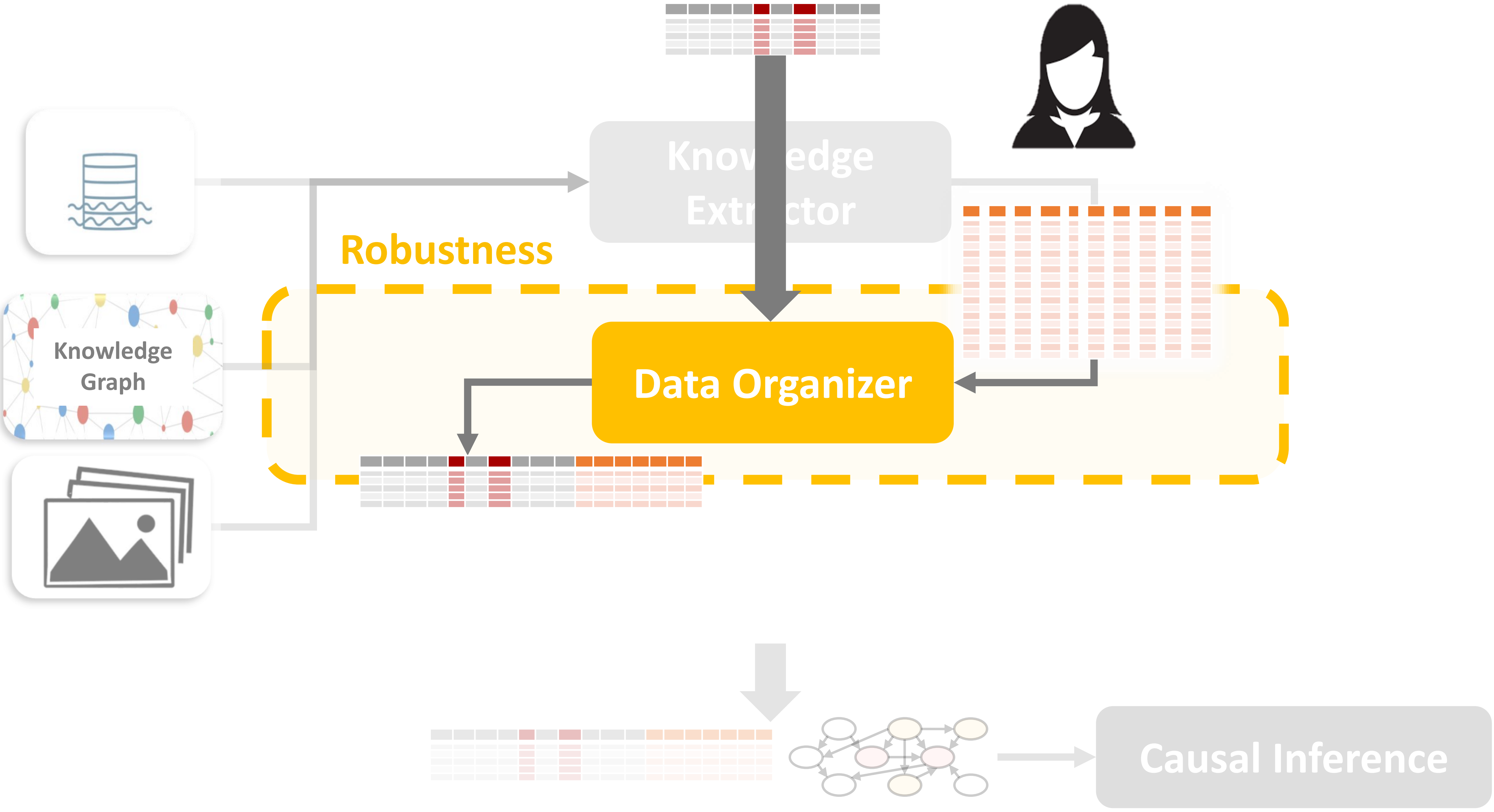
To answer causal questions, we need:

1.  A single and complete relational database
2.  A correct causal DAG

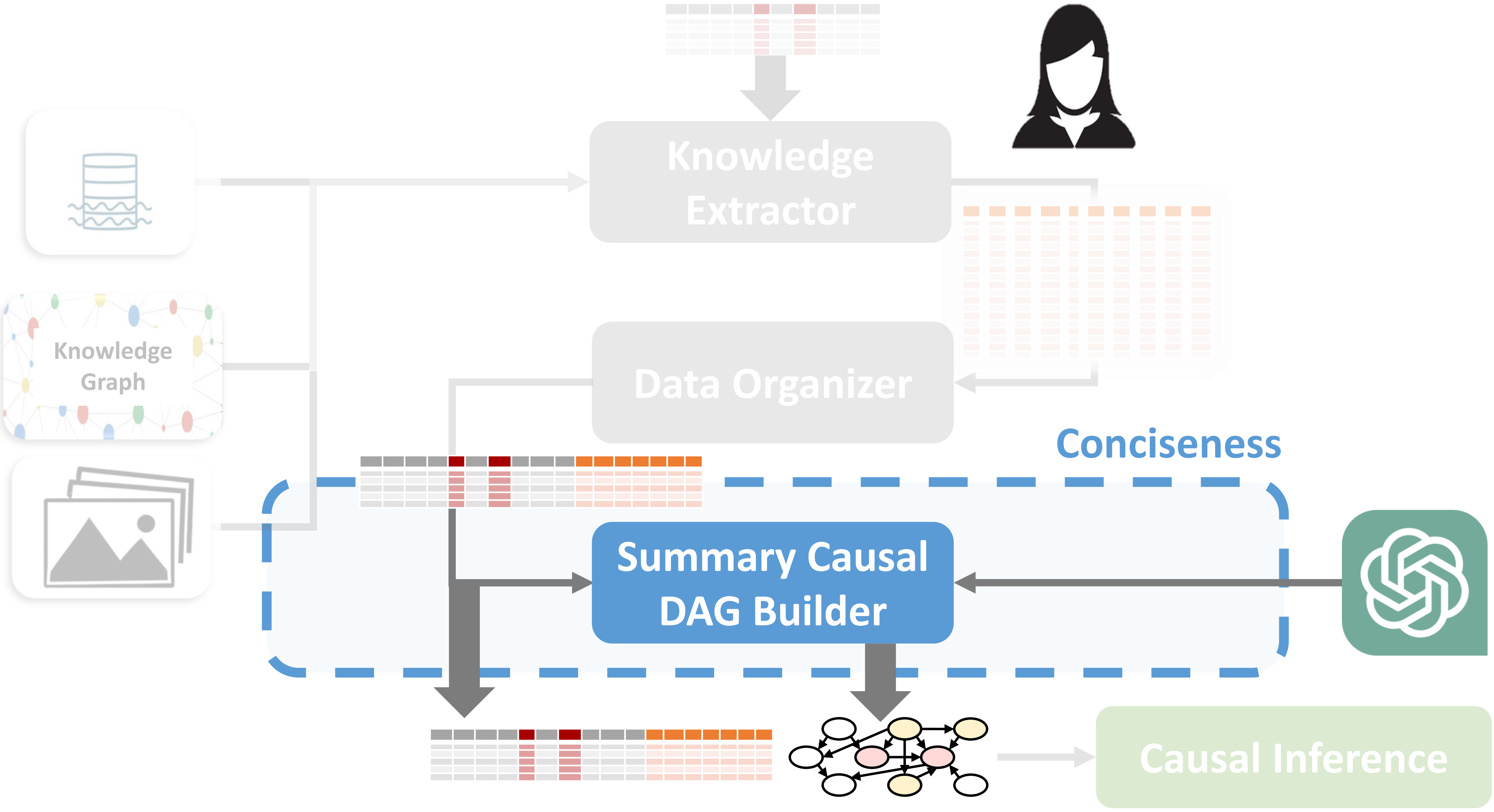

Causal Inference

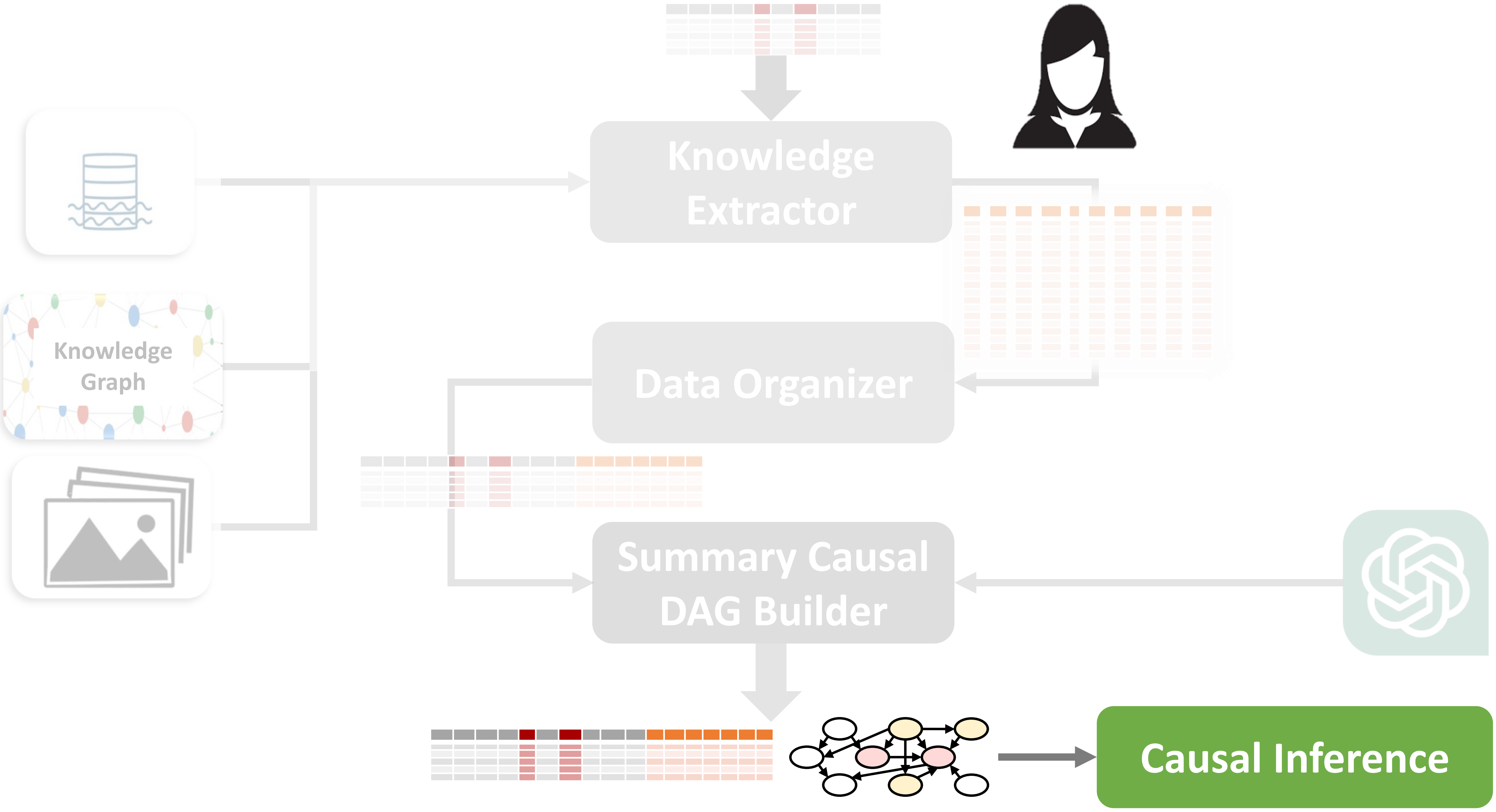


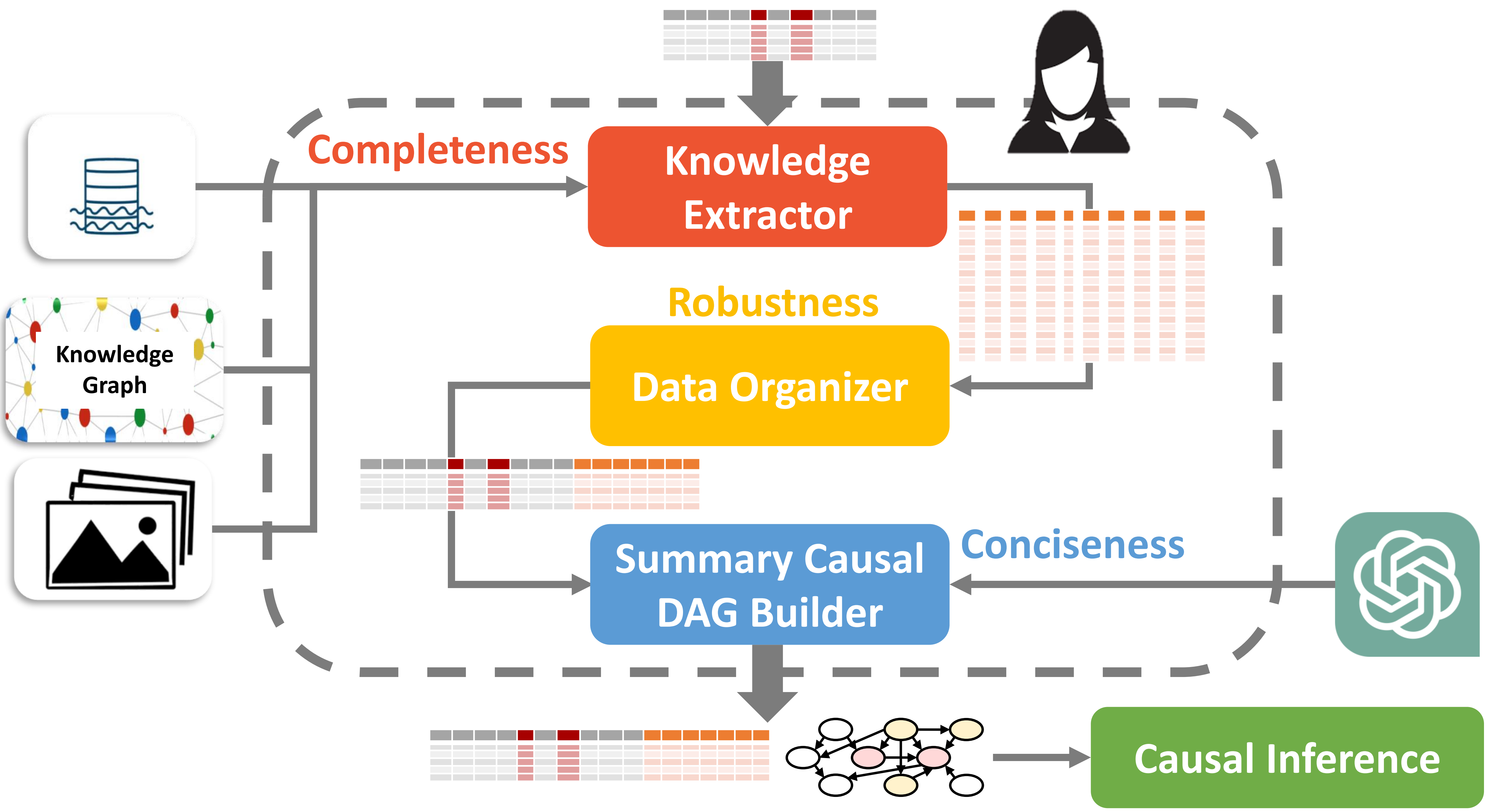
Completeness
Knowledge Extractor
Knowledge Graph
Causal Inference



Robustness
Data Organizer
Knowledge Graph
Causal Inference


Knowledge
Extractor
Knowledge
Graph
Data Organizer
Conciseness
Summary Causal
DAG Builder
Causal Inference


Knowledge
Extractor
Knowledge
Graph
Data Organizer
Summary Causal
DAG Builder
Causal Inference



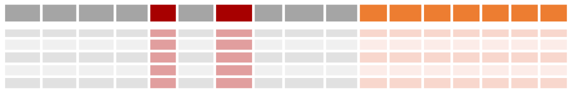
Completeness
Knowledge Extractor
Knowledge Graph
Robustness
Data Organizer
Conciseness
Summary Causal DAG Builder
Causal Inference

Summary

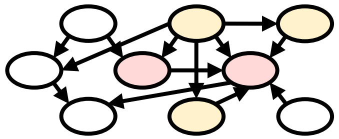
We introduce Causal Data Integration, a new research direction to mine and organize data for causal inference

Unlike conventional data integration we need to curate:

- Causally relevant variables



- Interpretable and concise causal DAG



Open Problems

- Extracting relevant and interpretable variables from text/images/videos/log-data/…
- Dealing with complex dependencies, probabilistic data,..
- Gaps between data and domain knowledge

Questions?

brity@technion.ac.il