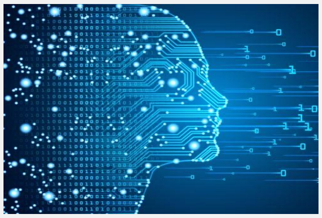
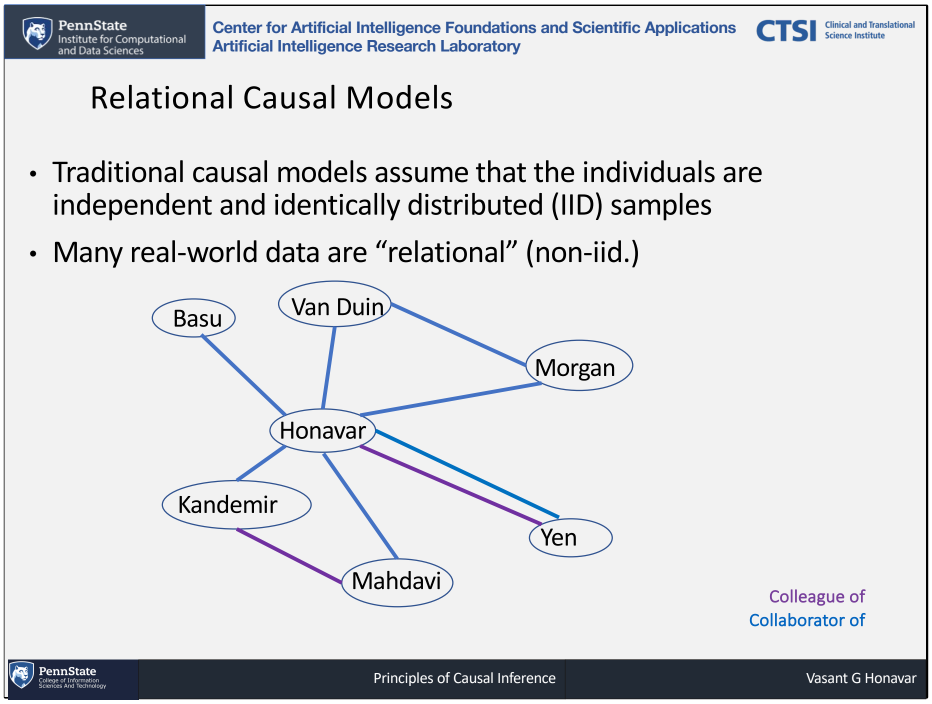


Principles of Causal Inference

Vasant G. Honavar

Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Center for Artificial Intelligence Foundations and Scientific Applications
Institute for Computational and Data Sciences
Huck Institutes of the Life Sciences
Clinical and Translational Sciences Institute
Northeast Big Data Hub
Pennsylvania State University
vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu



So causal models are useful to get insights from observational data.

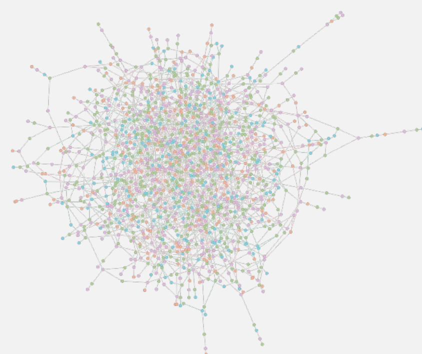
But, they assume observations are iid.

However, many real-world data are non-i.i.d.

Ubiquity of Relational Data

Relational data are

- Characterized by interconnected, heterogeneous entities
- Examples: web, Citation network, social network
- Representations
  - RDF triples
    Colored graphs
  - Entity-relationship model
  - …



Relational data is characterized by interconnected, heterogeneous entities. There are many examples including …

Also there are various representations.
(subject, predicate, object)

Modeling relations between entities

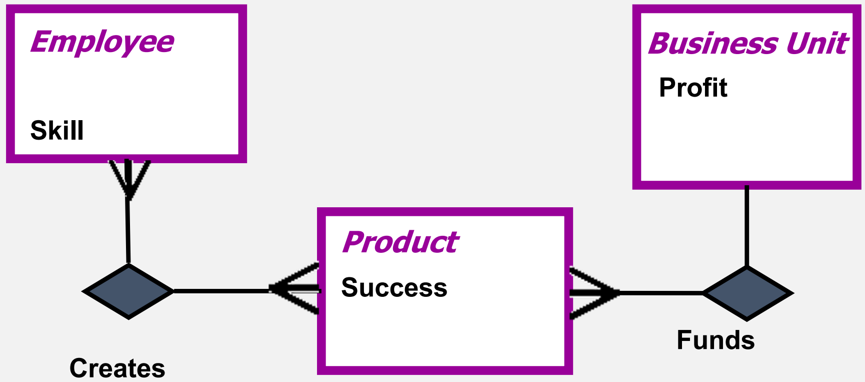
Example

- Employees, Products, Business Units are Entities
- Entities have one or more attributes
  - Skill of an employee
  - Success of a product
  - Profit of a business unit
- Entities are linked by relations
  - Employee creates product
  - Business unit funds product

Relationships between entities may induce causal relationships between their attributes

- The skill of employee(s) causally impacts the success of product(s) created by the employee(s)
- The success of product(s) causally impact the profit of the business unit(s) that funded the products

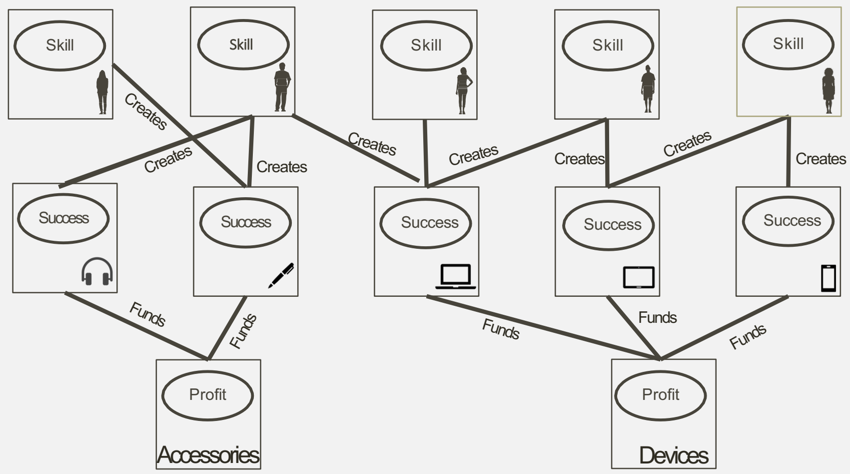
Relational Schema



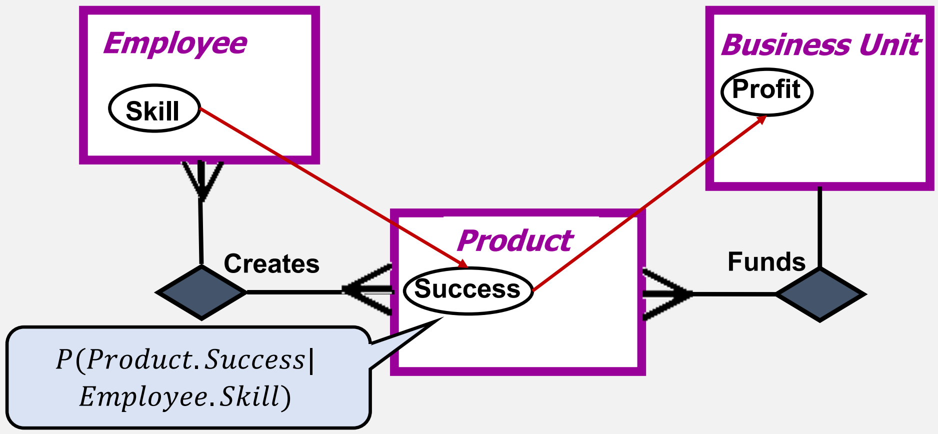
- Relational schema describes the types of objects and relations in the world that we care about

Chen, Peter Pin-Shan. "The entity-relationship model—toward a unified view of data." *ACM transactions on database systems (TODS)* 1, no. 1 (1976): 9-36.

A Relational Skeleton Instantiates a Relational Schema

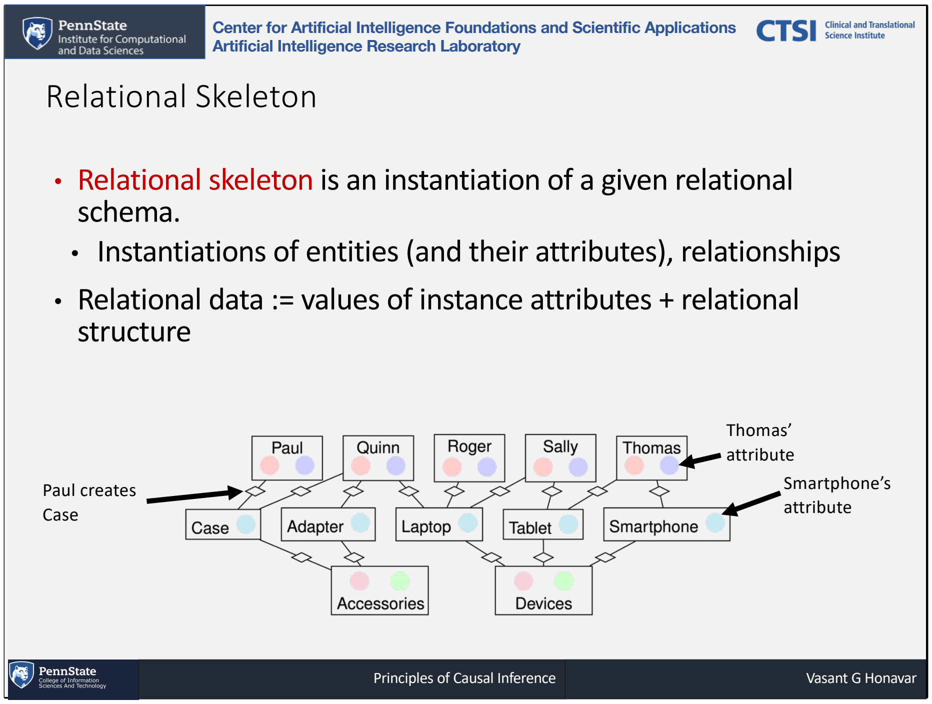


Probabilistic Relational Model



- Probabilistic relational model is the relational counterpart of a Bayesian Network

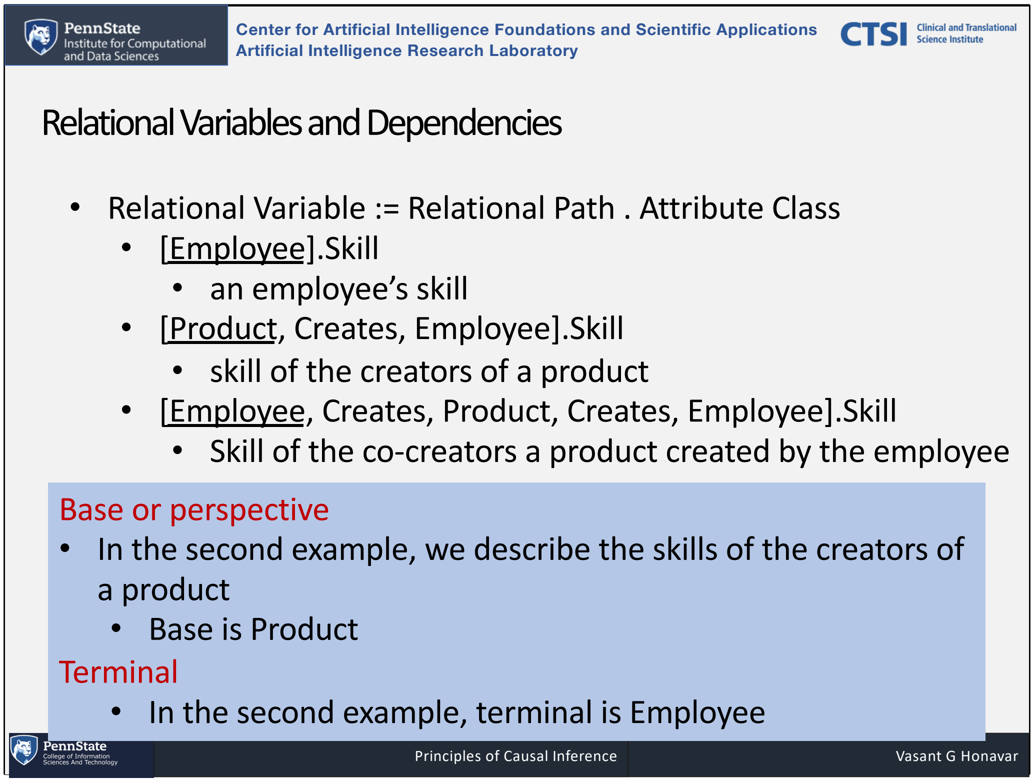
Getoor, Friedman, Koller & Pfeffer. *Learning Probabilistic Relational Models. IJCAI.* 1999.



A relational skeleton is an instantiation of the given schema.
Put in another way, relational skeletons are the ones we try to abstract by adopting a relational schema.

In this illustration, there are
5 employees developing 5 products funded by two business units.
For example, A **Case** is developed by Paul and Quinn.

Relational data is **the values of item attributes** and the structure itself.

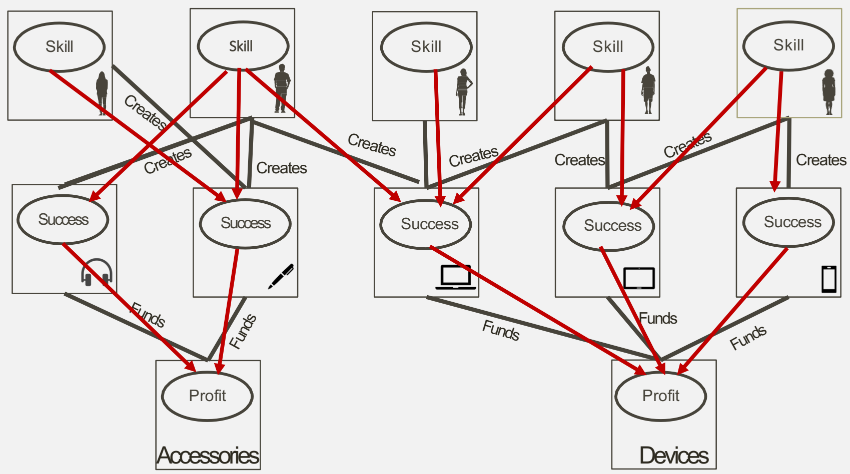


The first element of a relational path is called a base or perspective.
For example, the second one describes a set of competences from the viewpoint of a product.

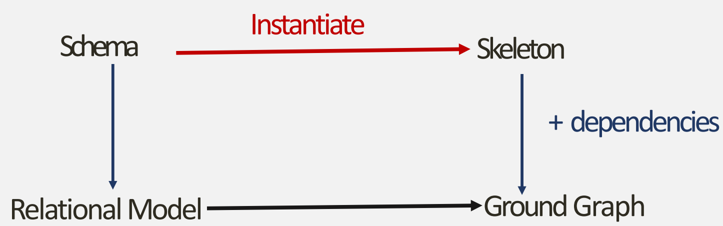
Expressing causal relations in relational domain

- The skills of the employees who created a product causally impacts the success of the product
  - [Product, Creates, Employee].Skill → [Product].Success
- The success of a product causally impacts the profit of the business unit that funded the creation of the product
  - [BizUnit, Funds, Product].Success →[BizUnit].Profit
- The profit of a business unit causally impacts its (future) budget
  - [BizUnit].Profit → [BizUnit].Budget
- Employee’s salary depends on the budget(s) of the business unit(s) that fund the products that he/she creates
  - [Employee, Creates, Product, Funds, BizUnit].Budget → [Employee].Salary

Relational Skeleton + Direct Causal Links = Ground Graph



Relational data



Getoor, Friedman, Koller & Pfeffer. *Learning Probabilistic Relational Models. IJCAI.* 1999.

Modeling causal relations in relational domains

In our simple model,

- An employee's skill causally impacts the success of the products he or she creates
- The success of a product causally impacts the profit of the business unit that funded the creation of the product

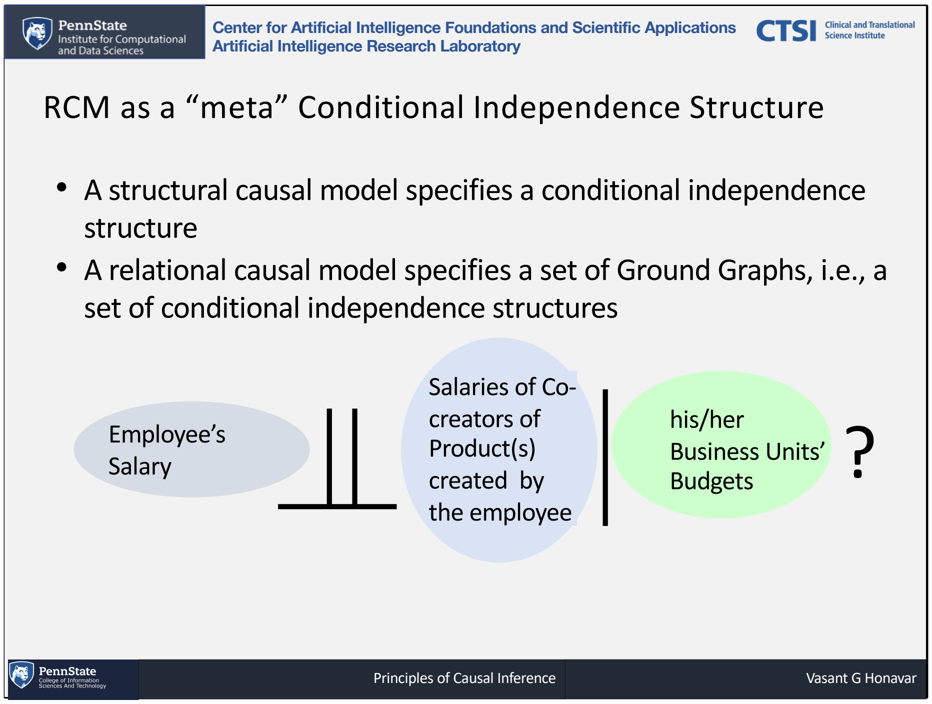
In a more complex model, it is possible that

- The profit of a business unit causally impacts its budget
- The budget of a business unit causally impacts the salary of the employees who create products funded by the unit

  - Relational causal models considered so far are limited to modeling causal relations between relational random variables

Modeling causal relations in relational domains

- For every acyclic relational causal model structure (defined by a relational schema and a set of direct causal dependencies) and a relational skeleton, the corresponding ground graph is a DAG.

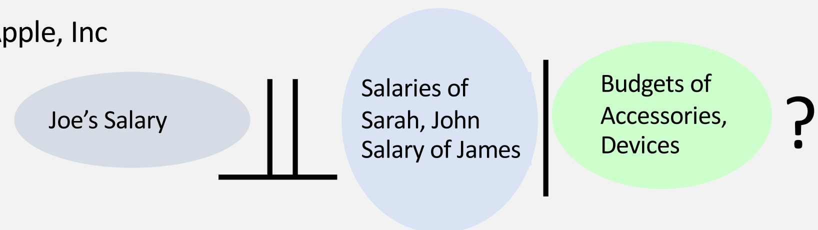


One can ask a Conditional Independence query with relational variables.
"Is an Employee's Salary independent to Coworkers' Salaries
given the his/her business units' Budgets?"

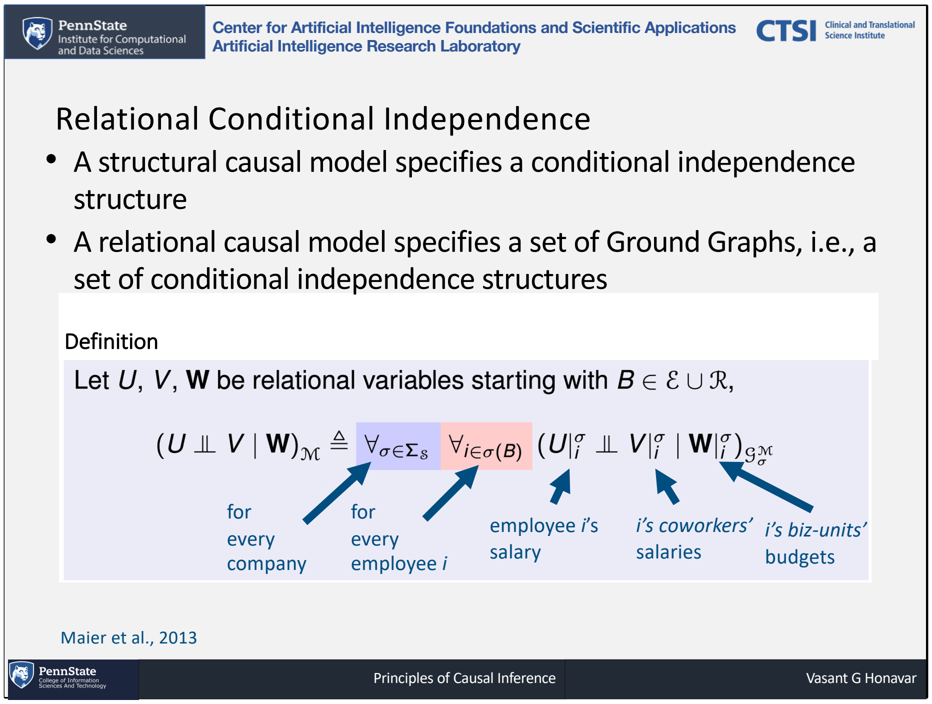
RCM as a "meta" Conditional Independence Structure

- A structural causal model specifies a conditional independence structure
- A relational causal model specifies a set of Ground Graphs, i.e., a set of conditional independence structures

Apple, Inc

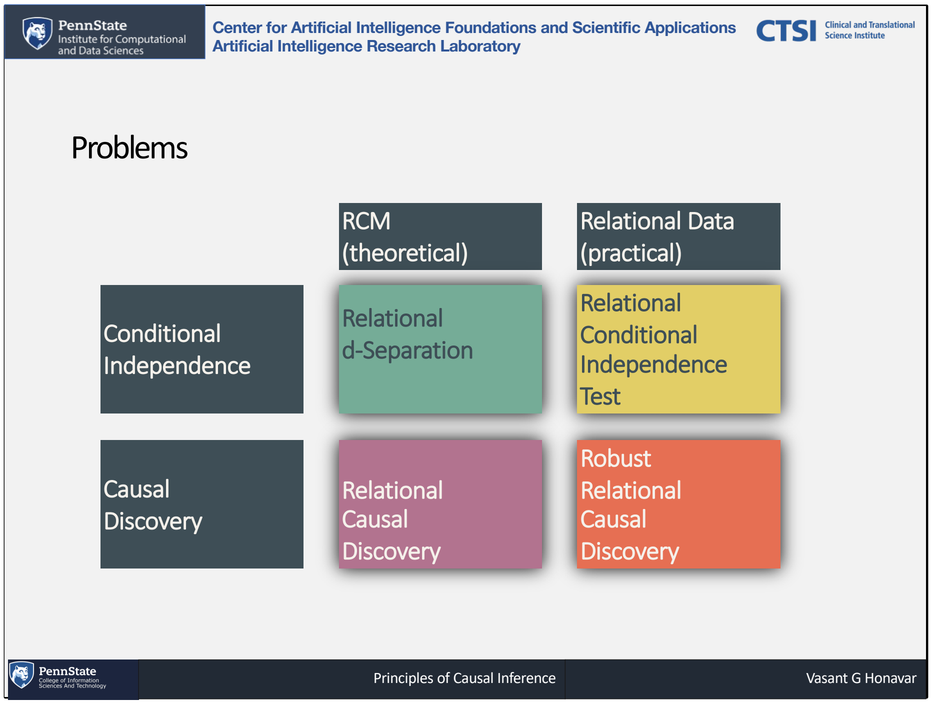


One can ask a Conditional Independence query with relational variables.
"Is an Employee's Salary independent to Coworkers' Salaries
given the his/her business units' Budgets?"



This is the definition of
Relational Conditional Independence
given by Maier and his colleagues

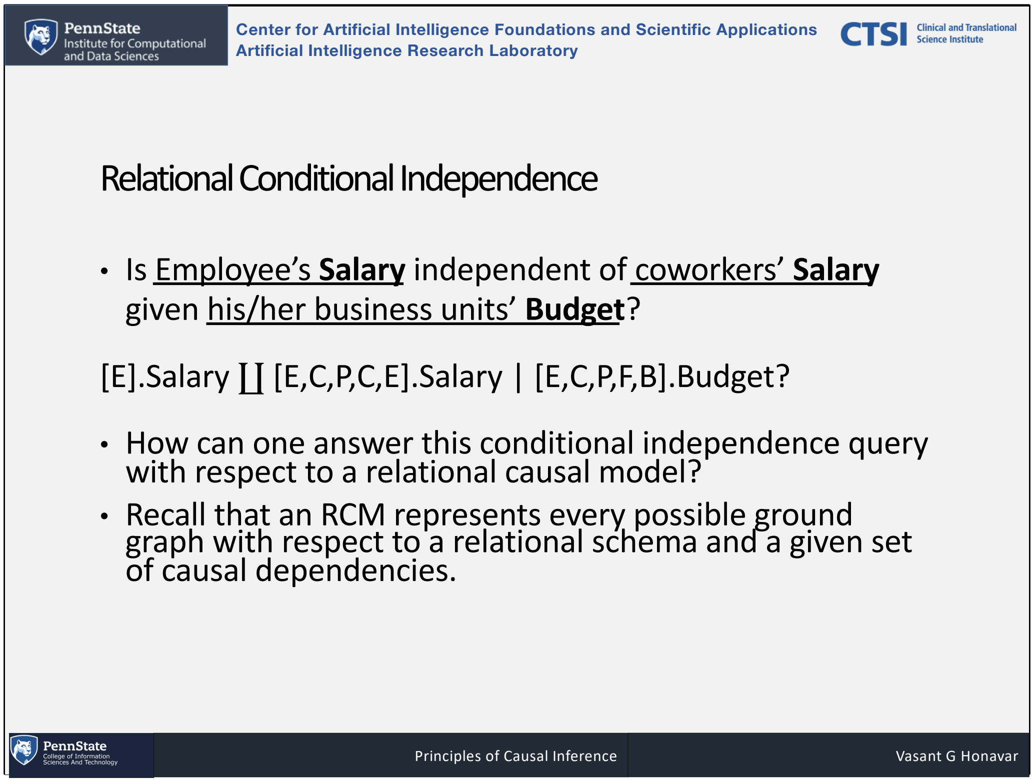
a query with relational variables **is equivalent to the logical conjunction** of individualized CI queries.



These four research questions can be laid out in a two-by-two table.

All these problems are **in fact** addressed by a group of researchers
but somewhat partly, narrowly, or incorrectly.

My dissertation begins by examining definitions of RCM and relevant concepts, and their implications.
Then, each of four problems



This is from the previous example — how can we say true or false.

We are able to say **true** if this independence holds **true**

for every employee in every possible company.

In other words, we can say no if salaries of an employee and his/her coworker are dependent in some company.

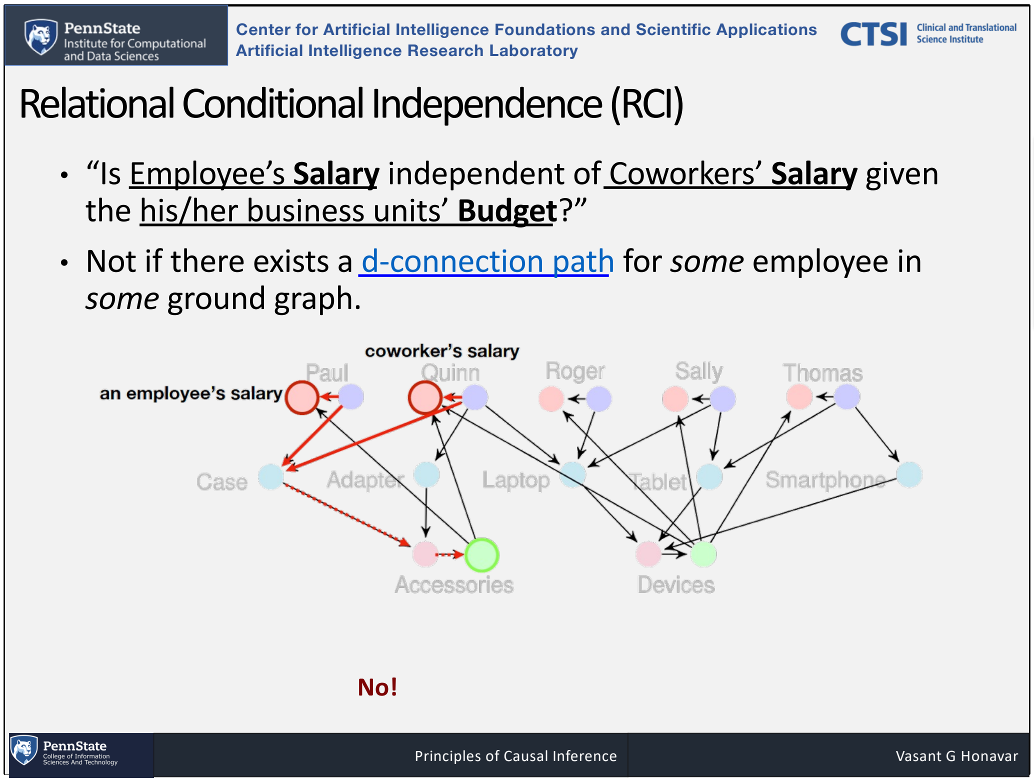
Relational D-Separation

- "Is Employee's **Salary** independent of Coworkers' **Salary** given the his/her business units' **Budget**?"
- [E].Salary $\perp\!\!\!\perp$ [E,C,P,C,E].Salary | [E,C,P,F,B].Budget?
  - Check if there exists a d-connection path for some employee in some ground graph.
  - If there is, such a path serves as witness of dependence

Relational D-separation is the generalization of d-separation to a relational setting.

if we find a d-connection path in a ground graph corresponding to a given query,

then we can say two relational variables are dependent given the conditionals.



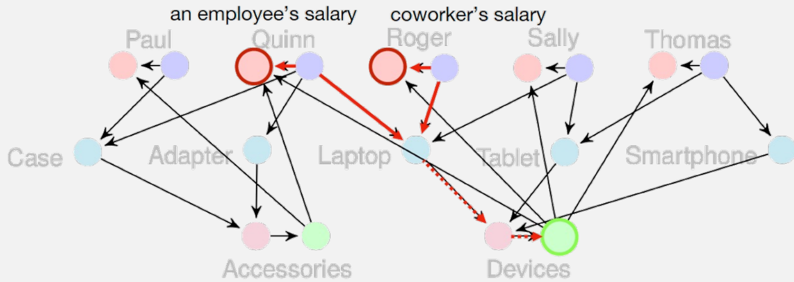
add link to d-separation

Whether we have an imaginary company or real company,
if we have a sample relational skeleton,
we can examine whether there exists a d-connection path.
and there it is.

In this example,
Paul's salary is dependent to Quinn's salary given Accessories' Budget

Relational Conditional Independence (RCI)

- "Is Employee's **Salary** independent of Coworkers' **Salary** given the his/her business units' **Budget**?"
- dependent if there exists a d-connection path for *some* employee in *some* ground graph.



No!

add link to d-separation

Here with Quinn and Roger.

Challenges in Relational D-Separation

- All ground graphs semantics
  - d-separation holds for every item in every ground graph (infinite)
  - Relational variables are defined by paths that end in an instance attribute
  - Multiple paths can represent the same random variable

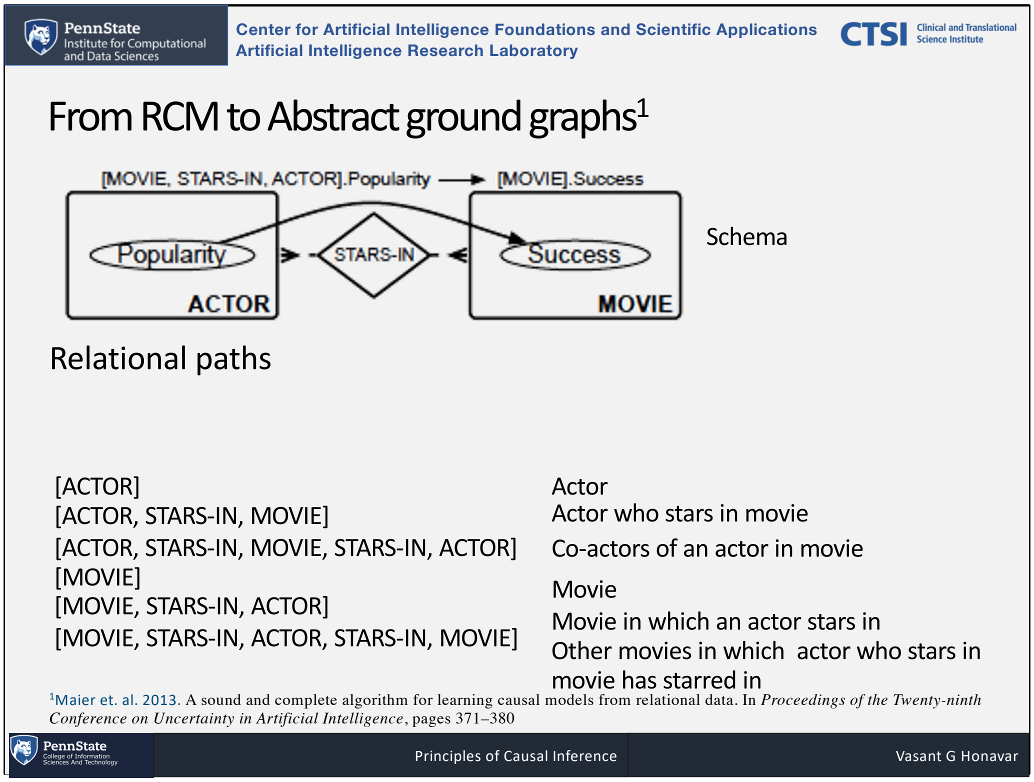
[ECPCECPCE].Salary & [ECPCE].Salary

So why is this problem challenging?

First,
we can't simply say that two relational variables are independent
after examining a few ground graphs.
We have to examine **all-possible-ground graphs.**

There are, in general, infinite number of relational skeletons, and, thus, infinite number of ground graphs.
Hence, a naive brute-force algorithm would not work.

Second,
Two different relational paths may represent common entities.
For example, a coworker can be a coworkers' coworker.
IF salaries of coworkers coworkers are given,
it is the case that
salaries of some of coworkers are ALSO given.

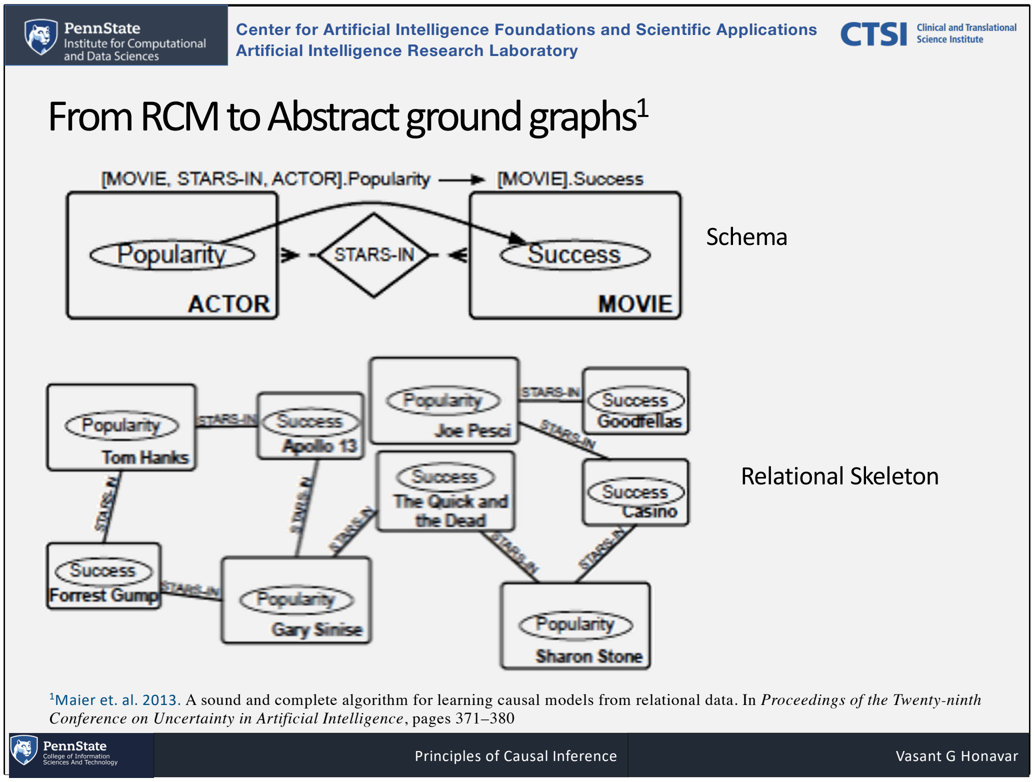


In 2013, Maier and coauthors devised a graph named “abstract ground graph”

where they intended to represent all the edges in all possible ground graphs as a directed acyclic graph.

There are ….

They claimed that an RCI query can be answered by applying traditional d-separation in this graph with the query modified.



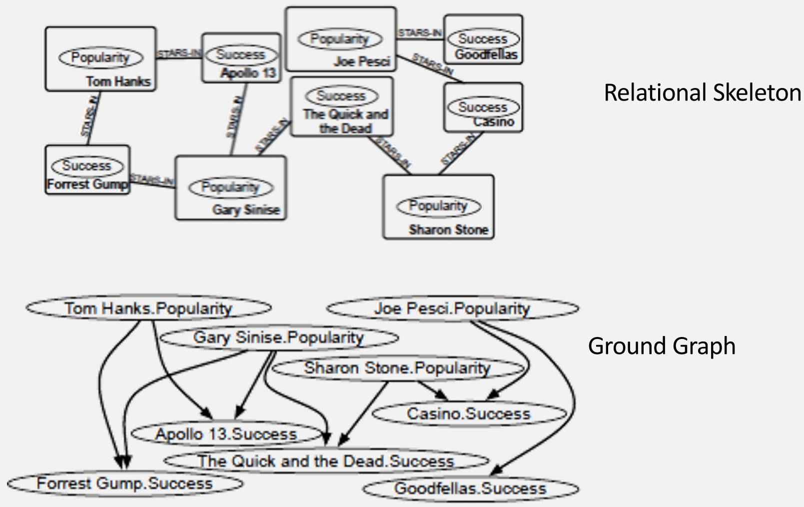
In 2013, Maier and coauthors devised a graph named “abstract ground graph”

where they intended to represent all the edges in all possible ground graphs as a directed acyclic graph.

There are ….

They claimed that an RCI query can be answered by applying traditional d-separation in this graph with the query modified.

From RCM to Abstract ground graphs[1]



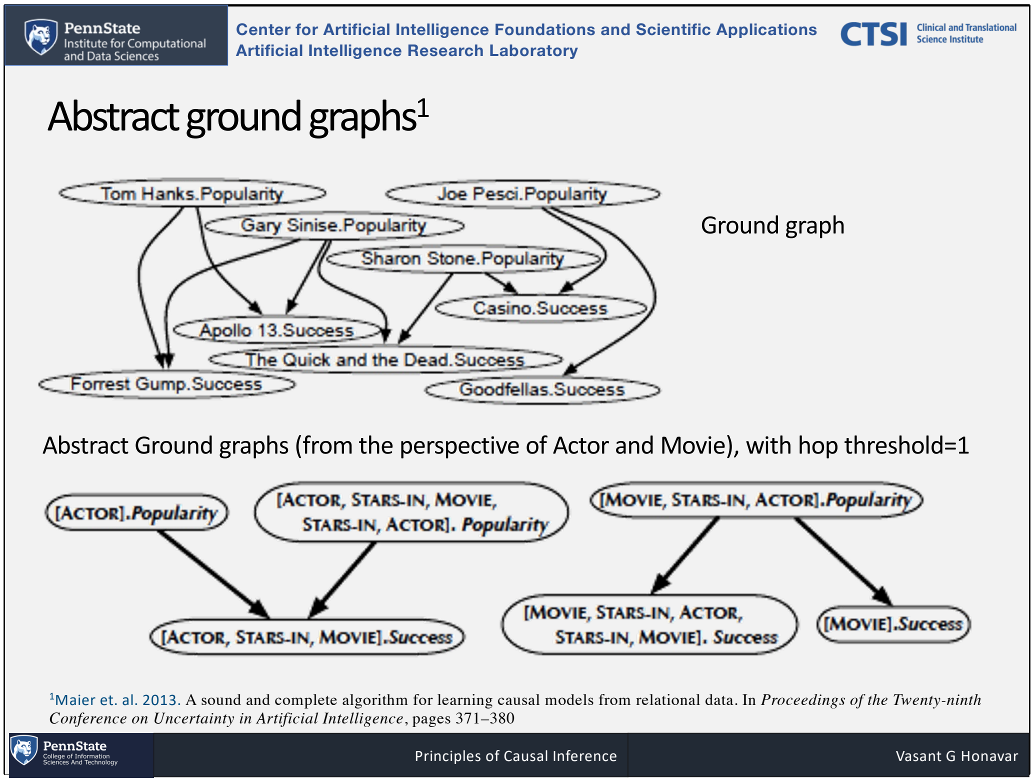
[1]Maier et. al. 2013. A sound and complete algorithm for learning causal models from relational data. In *Proceedings of the Twenty-ninth Conference on Uncertainty in Artificial Intelligence*, pages 371–380

In 2013, Maier and coauthors devised a graph named “abstract ground graph”

where they intended to represent all the edges in all possible ground graphs as a directed acyclic graph.

There are ….

They claimed that an RCI query can be answered by applying traditional d-separation in this graph with the query modified.



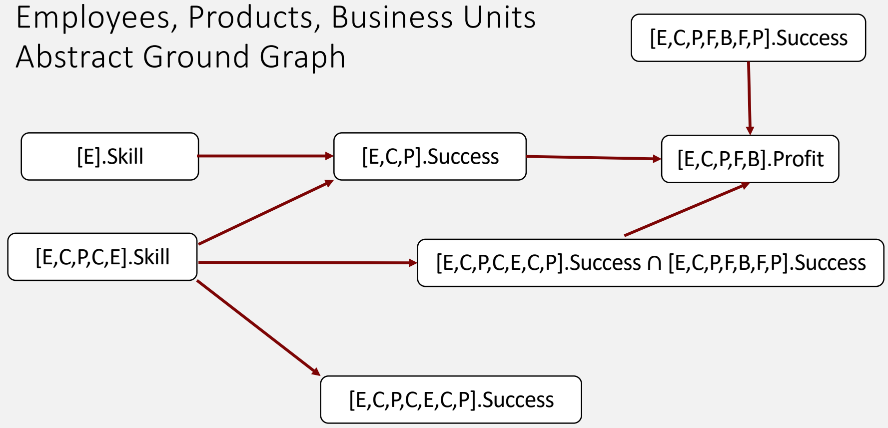
In 2013, Maier and coauthors devised a graph named “abstract ground graph”

where they intended to represent all the edges in all possible ground graphs as a directed acyclic graph.

There are ….

They claimed that an RCI query can be answered by applying traditional d-separation in this graph with the query modified.

Employees, Products, Business Units Abstract Ground Graph



Hop cutoff 6
Perspective: Employee

Abstract ground graphs

- Abstract Ground Graphs 'abstract' all edges in all ground graphs[1]
  - Claim: d-separation can be checked using the abstract ground graph[1]
  - However, Lee and Honavar found a counter example that invalidates the claim[2]
  - Ahsan et al[3]. repair Maier et. al claim by modifying the semantics and imposing some additional conditions on permitted relations
  - Lee and Honavar, 2016 proposed an alternative formulation

[1]Maier et. al. 2013. A sound and complete algorithm for learning causal models from relational data. In *Proceedings of the Twenty-ninth Conference on Uncertainty in Artificial Intelligence*, pages 371–380

[2]Lee, S. and Honavar, V., 2015, Lifted representation of relational causal models revisited: implications for reasoning and structure learning. In *Proceedings of the UAI 2015 Conference on Advances in Causal Inference-Volume 1504* (pp. 56-65).

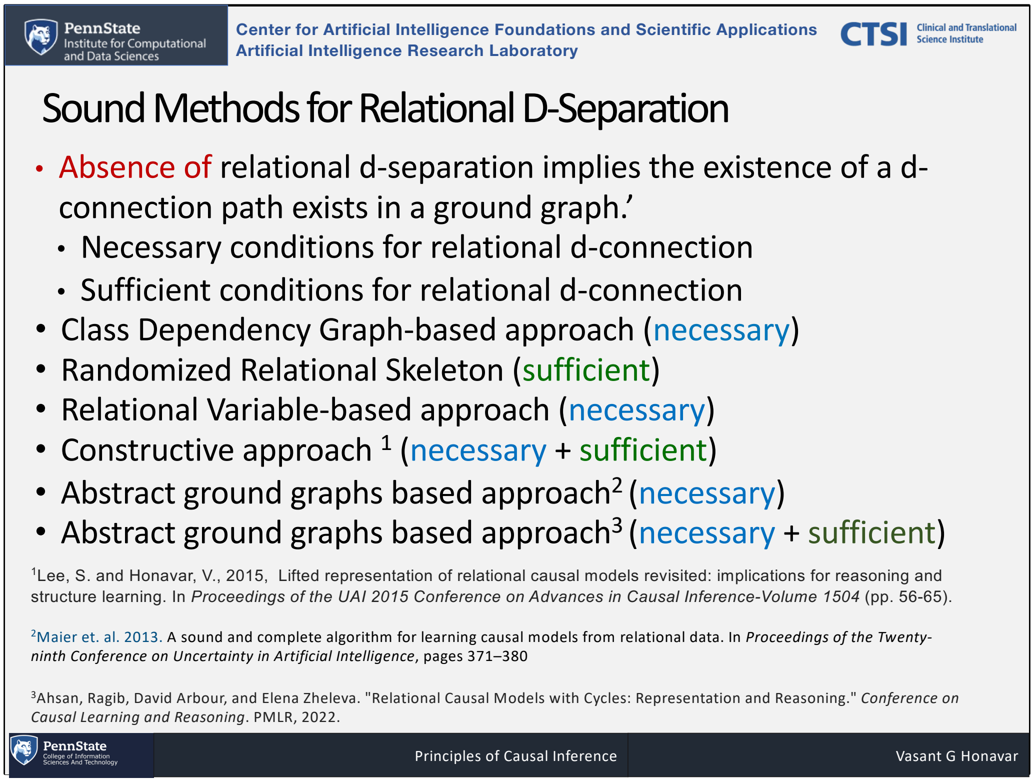
[3]Ahsan, Ragib, David Arbour, and Elena Zheleva. "Relational Causal Models with Cycles: Representation and Reasoning." *Conference on Causal Learning and Reasoning*. PMLR, 2022.

In 2013, Maier and coauthors devised a graph named "abstract ground graph"

where they intended to represent all the edges in all possible ground graphs as a directed acyclic graph.

There are ....

They claimed that an RCI query can be answered by applying traditional d-separation in this graph with the query modified.



I proposed sound algorithms for relational d-separation.

Basically, I investigated what are necessary OR sufficient conditions for the existence of a d-connection path in some ground graph.

I proposed four methods —

some are necessary some are sufficient conditions.
But none is a necessary and sufficient condition.

I will illustrate two methods.

Constructive Approach

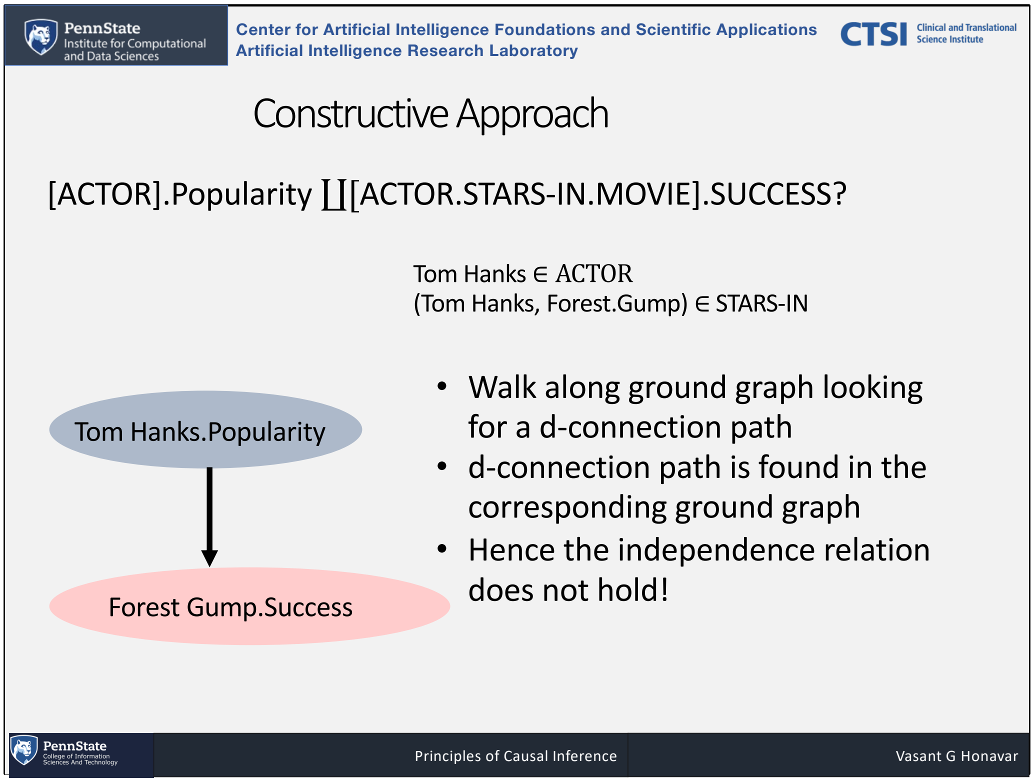
- Incrementally build a relational skeleton, until
  - A suitable d-connection path is found in the corresponding ground graph - in which case, we have a witness for lack of d-separation; or
  - The relational skeleton cannot be expanded further and no suitable d-connection path has been found – in which case, we can claim d-separation; or
  - There is a timeout in which case we cannot determine for sure if d-separation holds or not, but we can rule out violation of d-separation by virtue of d-connection paths that are shorter than a certain length

Next, I would like to describe a constructive approach.

In this approach,
we try to build a relational skeleton,
where the resulting ground graph contains a d-connection path.

This is also a search problem.

If we are failed to find one, it implies relational d-separation.
If we find one, you know it is a concrete d-connection path, so we can conclude relational d-connection.
The method may not stop within a user-specified computational budget, then it can return undecided.



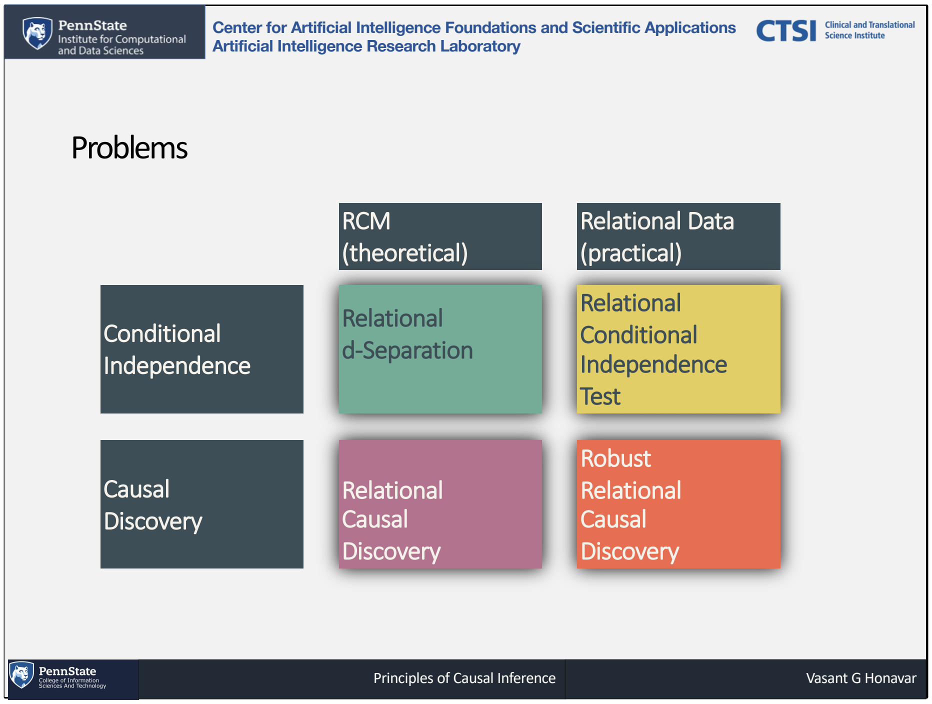
We have a start node.

someone’s salary

Abstract Ground Graph Based d-separation

- Abstract ground graphs
  - are directed acyclic whenever the the ground graphs are acyclic
  - abstract all (and only) causal dependencies in the ground graphs
  - D-separation holds in all ground graphs that match a schema and a skeleton if and only if it holds in the abstract ground graph
    - Not true[2] in the case of the original definition[1]
    - True in the case of modified definition[2]
- If abstract ground graphs abstract all (and only) causal dependencies in the ground graphs (within a specified hop cutoff), then
  - We can efficiently check for d-separation using abstract ground graphs

[1]Maier et. al. 2013. A sound and complete algorithm for learning causal models from relational data. In *Proceedings of the Twenty-ninth Conference on Uncertainty in Artificial Intelligence*, pages 371–380
[2]Lee, S. and Honavar, V., 2015, Lifted representation of relational causal models revisited: implications for reasoning and structure learning. In *Proceedings of the UAI 2015 Conference on Advances in Causal Inference-Volume 1504* (pp. 56-65).
[3]Ahsan, Ragib, David Arbour, and Elena Zheleva. "Relational Causal Models with Cycles: Representation and Reasoning." *Conference on Causal Learning and Reasoning*. PMLR, 2022.



These four research questions can be laid out in a two-by-two table.

All these problems are **in fact** addressed by a group of researchers
but somewhat partly, narrowly, or incorrectly.

My dissertation begins by examining definitions of RCM and relevant concepts, and their implications.
Then, each of four problems

Relational causal discovery

Learning the structure of Relational Causal Models Using a Relational Conditional Independence Oracle[1,2,3]

- Independence queries offer a way to learn the equivalence class of causal bayes networks from data
- Replace variables by relational variables
- Replace conditional independence oracle by relational conditional independence oracle
  - How can we be sure that we have answers to a sufficient set of conditional independence queries? (completeness)
  - What queries should we ask? (efficiency, non-redundancy)
- Related work[4]

[1]Lee, S. and Honavar, V., 2015. Lifted representation of relational causal models revisited: implications for reasoning and structure learning. In *Proceedings of the UAI 2015 Conference on Advances in Causal Inference-Volume 1504*(pp. 56-65).
[2]Lee S, Honavar V. 2016. A characterization of Markov equivalence classes of Relational Causal Models under path semantics. In32nd Conference on Uncertainty in Artificial Intelligence (pp. 387-396).
[3]Lee S, Honavar V. 2016. On learning causal models from relational data. In30th AAAI Conference on Artificial Intelligence, AAAI 2016 2016 Jan 1 (pp. 3263-3270). AAAI press.
[4]Maier M, Marazopoulou K, Arbour D, Jensen D. 2013. A sound and complete algorithm for learning causal models from relational data. In: Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (pp. 371-380).

Learning the structure of Relational Causal Models Using a Relational Conditional Independence Oracle

Key observations

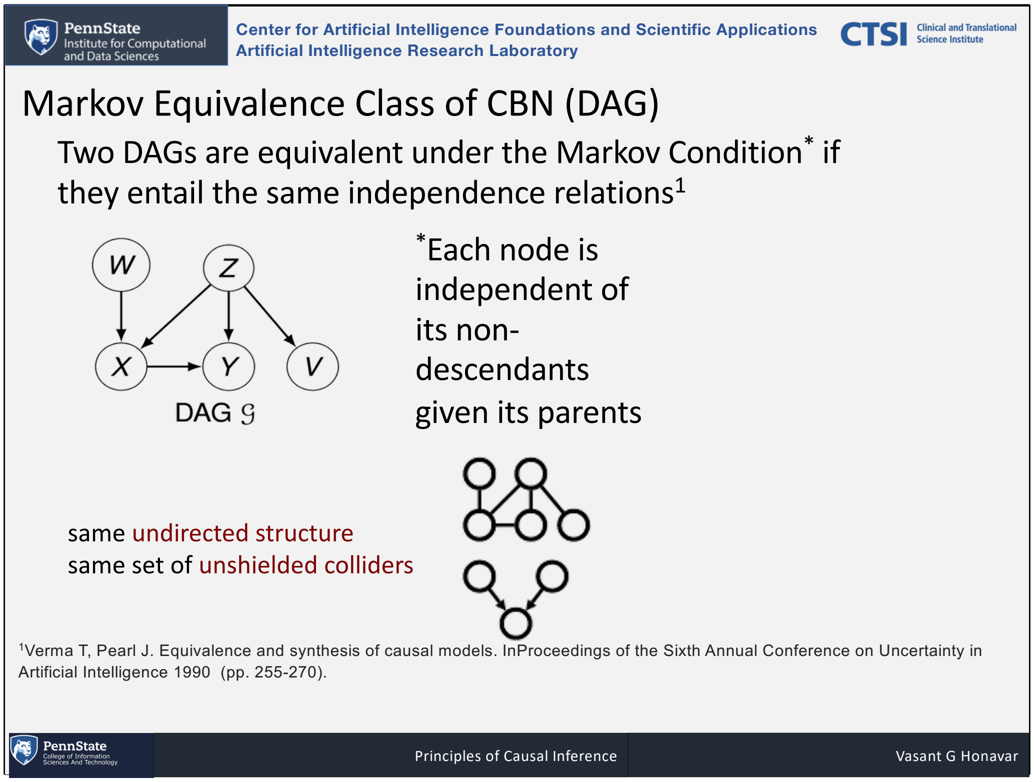
- A probability distribution over a set of variables is said to be faithful to a DAG over the same set of variables if and only if every conditional independence relation that is valid in the probability distribution is entailed by the DAG.
- A relational causal model is not faithful to Abstract Ground Graphs
- Relational Causal Models satisfy adjacency-faithfulness and orientation-faithfulness[1]

[3]Lee S, Honavar V. 2016. On learning causal models from relational data. In30th AAAI Conference on Artificial Intelligence, AAAI 2016 2016 Jan 1 (pp. 3263-3270). AAAI press.

Learning the structure of Relational Causal Models Using a Relational Conditional Independence Oracle

- Adjacency Faithfulness[1]: Given a set of variables V whose causal structure can be represented by a DAG G, if two variables X and Y are adjacent in G, then they are dependent conditional on any subset of V \ {X, Y} in G.
- Orientation Faithfulness[1]: Given a set of variables V whose causal structure can be represented by a DAG G, let X, Y, Z be any unshielded triple in G.
  - if $X \rightarrow Y \leftarrow Z$, then X and Z are dependent given any subset of $V \setminus \{X, Z\}$ that contains Y ;
  - otherwise, X and Z are dependent conditional on any subset of $V \setminus \{X, Z\}$ that does not contain Y .
- Given adjacency and orientation faithfulness, it is possible to learn the correct Markov equivalence class of RCM from data[2]

[1] Ramsey, J., Spirtes, and Zhang 2006. Adjacency-faithfulness and conservative causal inference. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI-06)* (pp. 401-408).
[2]Lee S, Honavar V. 2016. On learning causal models from relational data. In30th AAAI Conference on Artificial Intelligence, AAAI 2016 2016 Jan 1 (pp. 3263-3270). AAAI press.



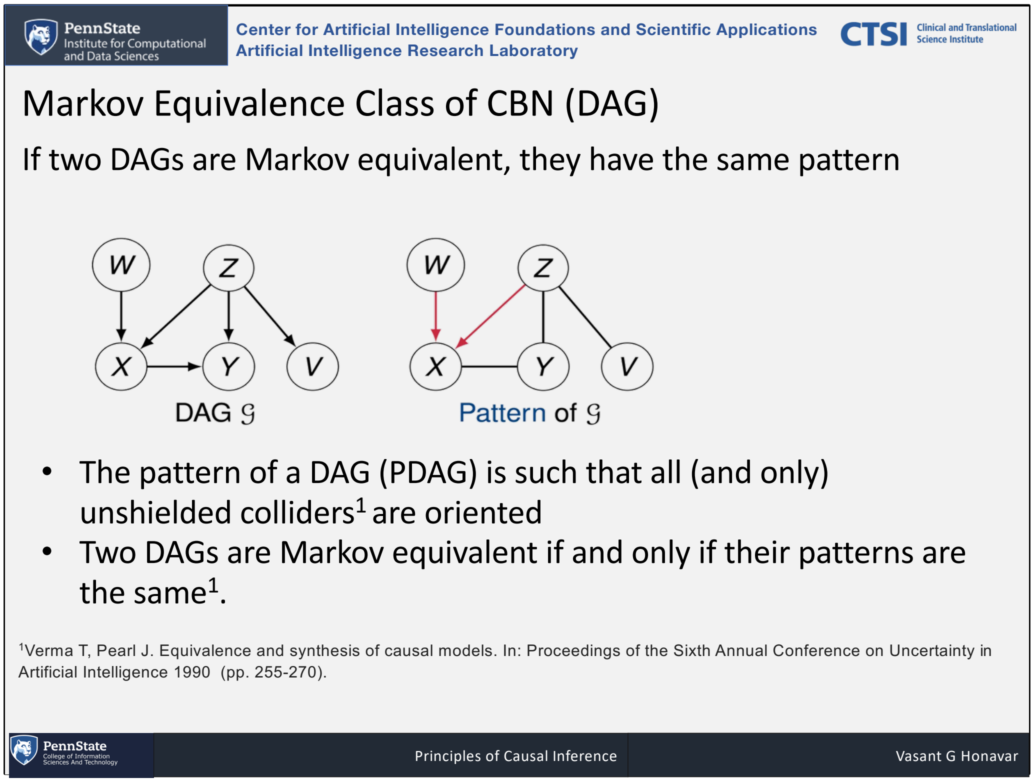
In the case of
researchers studied Markov equivalence class.

(definition)
Two DAGs are Markov equivalent if they entail the same set of independence relations.

(Markov = pattern)
It is well-known that
Two Markov equivalent DAGs share the same pattern.

(pattern …)
The pattern of G has the same undirected structure of G,
and the same set of unshielded colliders.

(An unshielded collider is a tuple of three nodes where two nodes are connected towards a node but not to each other)



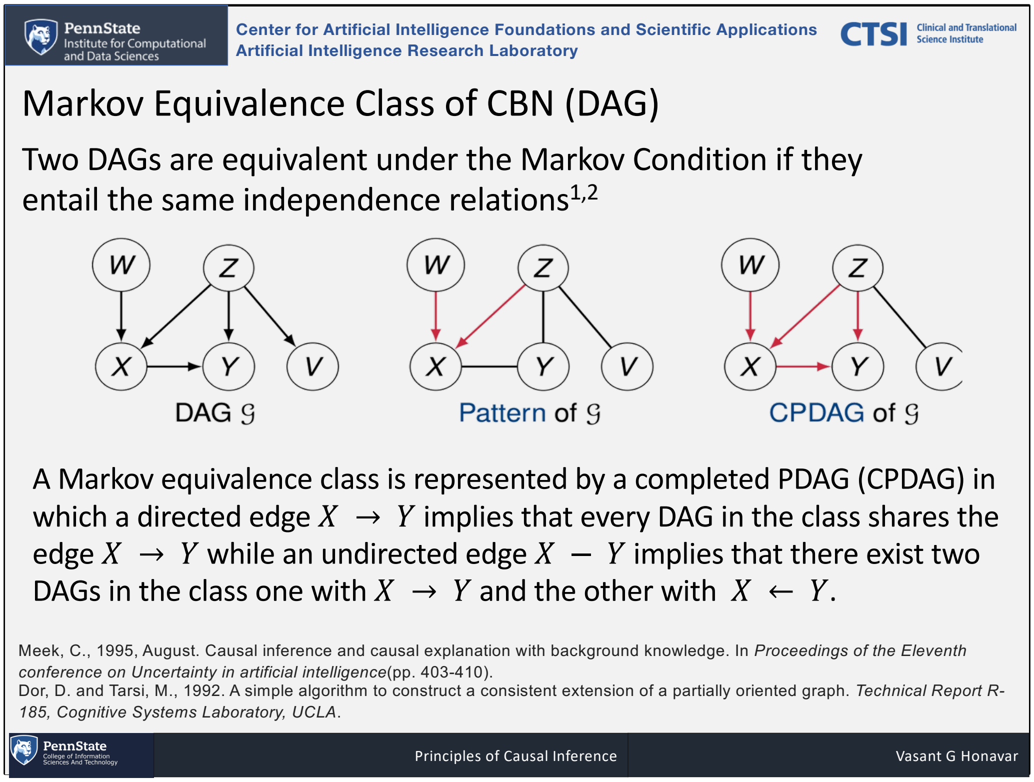
In the case of
researchers studied Markov equivalence class.

(definition)
Two DAGs are Markov equivalent if they entail the same set of independence relations.

(Markov = pattern)
It is well-known that
Two Markov equivalent DAGs share the same pattern.

(pattern …)
The pattern of G has the same undirected structure of G,
and the same set of unshielded colliders.

(An unshielded collider is a tuple of three nodes where two nodes are connected towards a node but not to each other)



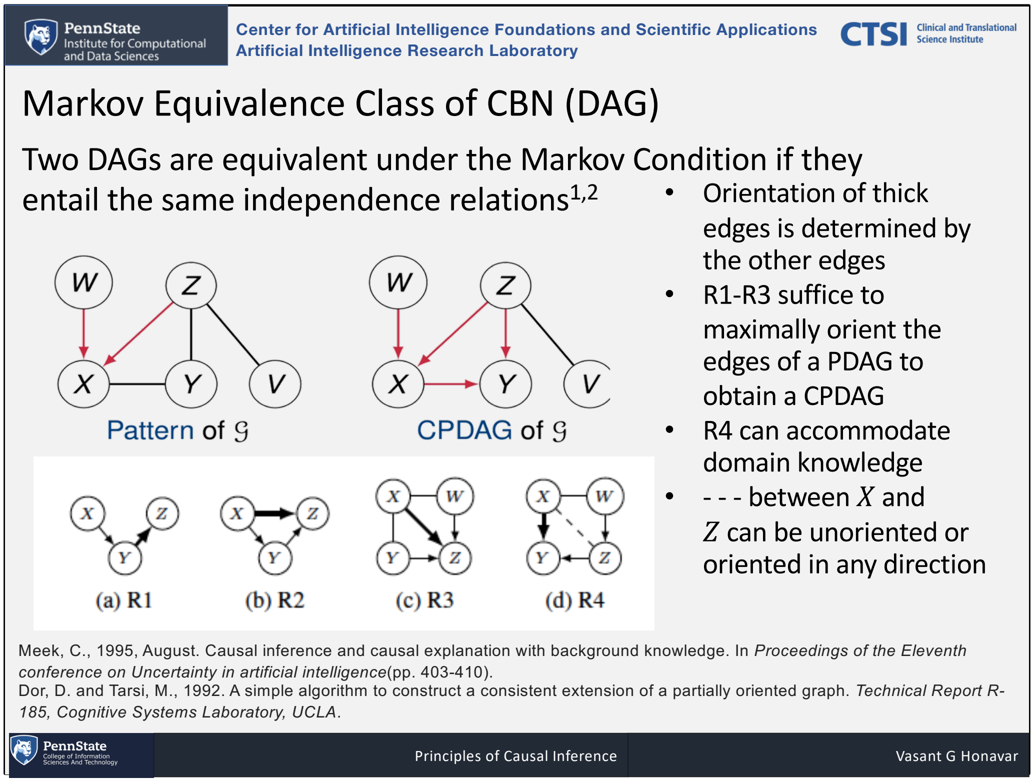
In the case of
researchers studied Markov equivalence class.

(definition)
Two DAGs are Markov equivalent if they entail the same set of independence relations.

(Markov = pattern)
It is well-known that
Two Markov equivalent DAGs share the same pattern.

(pattern …)
The pattern of G has the same undirected structure of G,
and the same set of unshielded colliders.

(An unshielded collider is a tuple of three nodes where two nodes are connected towards a node but not to each other)



In the case of
researchers studied Markov equivalence class.

(definition)
Two DAGs are Markov equivalent if they entail the same set of independence relations.

(Markov = pattern)
It is well-known that
Two Markov equivalent DAGs share the same pattern.

(pattern …)
The pattern of G has the same undirected structure of G,
and the same set of unshielded colliders.

(An unshielded collider is a tuple of three nodes where two nodes are connected towards a node but not to each other)

Markov equivalence class of Relational Causal Models

Theorem * *Two RCMs defined over the same relational schema are Markov equivalent if and only if their ground graphs are Markov equivalent for every relational skeleton of the relational schema:*

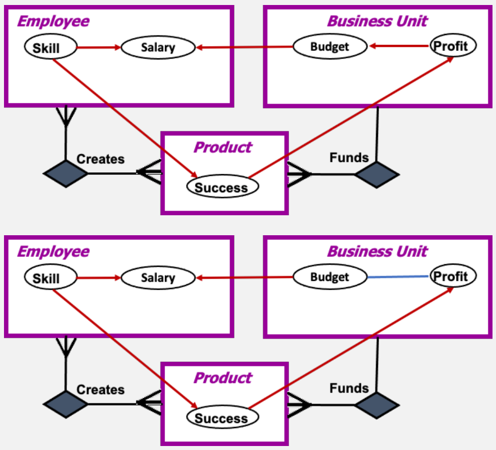
$$[\mathcal{M}] = [\mathcal{M}'] \Leftrightarrow \forall_{\sigma \in \Sigma_{\mathcal{S}}} [\mathcal{G}_{\sigma}^{\mathcal{M}}] = [\mathcal{G}_{\sigma}^{\mathcal{M}'}].$$

Two RCMs are Markov equivalent if they share the
- same undirected structure
- same set of canonical unshielded colliders+

+ Generalization of unshielded colliders to the relational setting

*Lee S, Honavar V. 2016. A characterization of Markov equivalence classes of Relational Causal Models under path semantics. In 32nd Conference on Uncertainty in Artificial Intelligence (pp. 387-396).

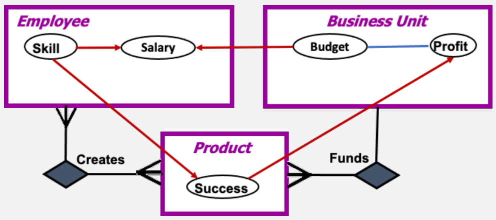
Pattern of the DAG of Relational Causal Model (PRDAG)

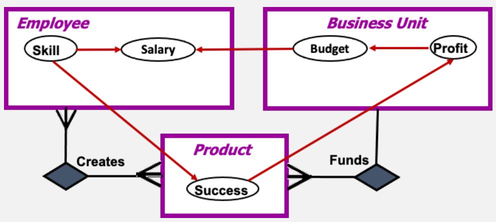


- PRDAG is such that all (and only) canonical unshielded colliders[1] are oriented
- Two PRAAGs are Markov equivalent if and only if their patterns are the identical[1].

*Lee S, Honavar V. 2016. A characterization of Markov equivalence classes of Relational Causal Models under path semantics. In 32nd Conference on Uncertainty in Artificial Intelligence (pp. 387-396).

From PRDAG to CPRDAG (via Meek's orienting rules)





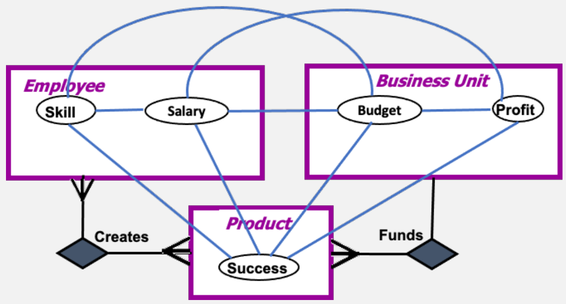
Lee, S. and Honavar, V., 2016, March. On learning causal models from relational data. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30, No. 1).

Learning the Structure of Relational Causal Models from Conditional Independence Queries

- Generalizes algorithms for learning the structure of causal Bayesian Networks (IID setting) to work with Relational Causal Models
  - First use independence queries to eliminate as many undirected edges as possible
  - Then maximally orient the edges using Meek's rules

Learning the Structure of Relational Causal Models from Conditional Independence Queries

- Start with a fully connected undirected graph
- Use (conditional) independence queries to eliminate the edges

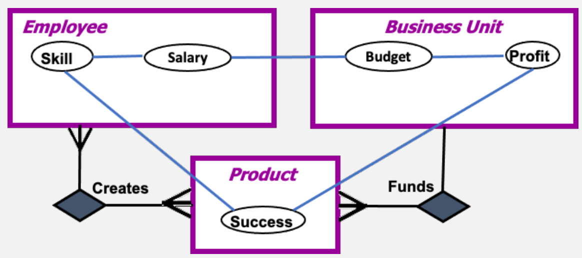


Sample Query:

$$[\mathrm{E, S, P, S, E}].\mathrm{Salary} \perp\!\!\!\perp [\mathrm{E}].\mathrm{Salary} \mid \{[\mathrm{E}].\mathrm{Salary}, [\mathrm{E, S, P, F, B}].\mathrm{Budget}\}$$

Learning the Structure of Relational Causal Models from Conditional Independence Queries

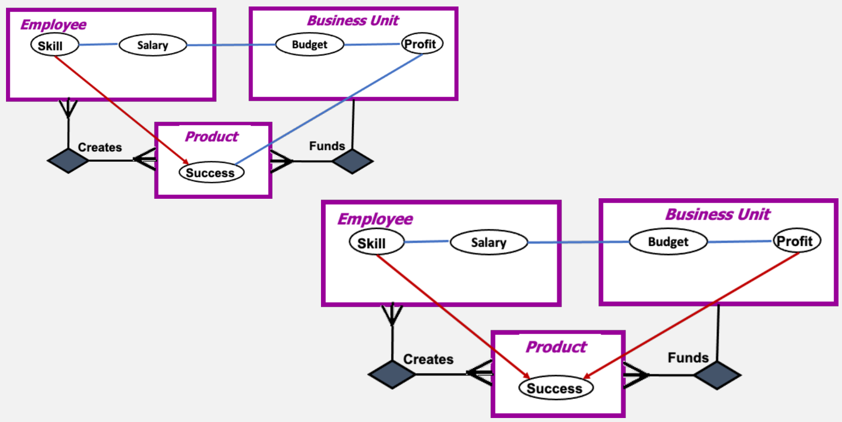
- Undirected structure consistent with independence oracle



PennState Institute for Computational and Data Sciences
Center for Artificial Intelligence Foundations and Scientific Applications
Artificial Intelligence Research Laboratory
CTSI Clinical and Translational Science Institute

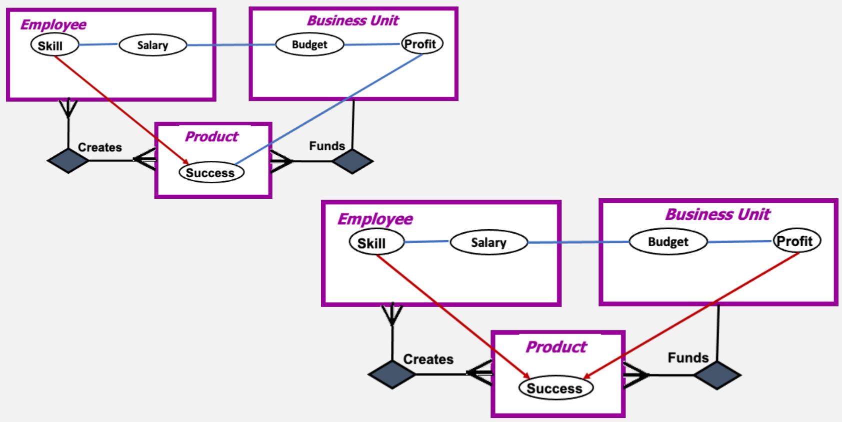
Learning the Structure of Relational Causal Models from Conditional Independence Queries

- Orient edges using canonical unshielded colliders (triples)

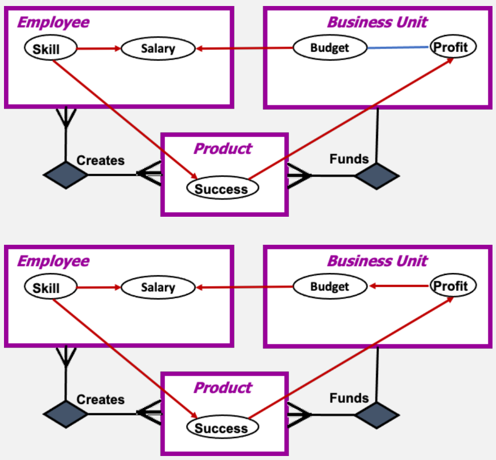


Learning the Structure of Relational Causal Models from Conditional Independence Queries

- Orient edges using canonical unshielded colliders (triples)



Learning the Structure of Relational Causal Models from Conditional Independence Queries



PRDAG obtained by orienting additional relational dependencies using constraints

- canonical unshielded non-colliders
- acyclicity at an attribute class level

CPRDAG via Meek's rules and generalized PDAG extensibility

*Lee S, Honavar V. 2016. A characterization of Markov equivalence classes of Relational Causal Models under path semantics. In 32nd Conference on Uncertainty in Artificial Intelligence (pp. 387-396).

Learning the Structure of Relational Causal Models from Conditional Independence Queries

Theorem: Given access to the conditional independence oracle for an RCMM, the RCDL algorithm offers, under the adjacency faithfulness and orientation faithfulness assumptions, a sound and complete procedure for learning the structure of the RCM whose maximum number of hops of dependencies is bounded by h.[1]

*Lee S, Honavar V. 2016. A characterization of Markov equivalence classes of Relational Causal Models under path semantics. In 32nd Conference on Uncertainty in Artificial Intelligence (pp. 387-396).

Learning the Structure of Relational Causal Models from Conditional Independence Queries

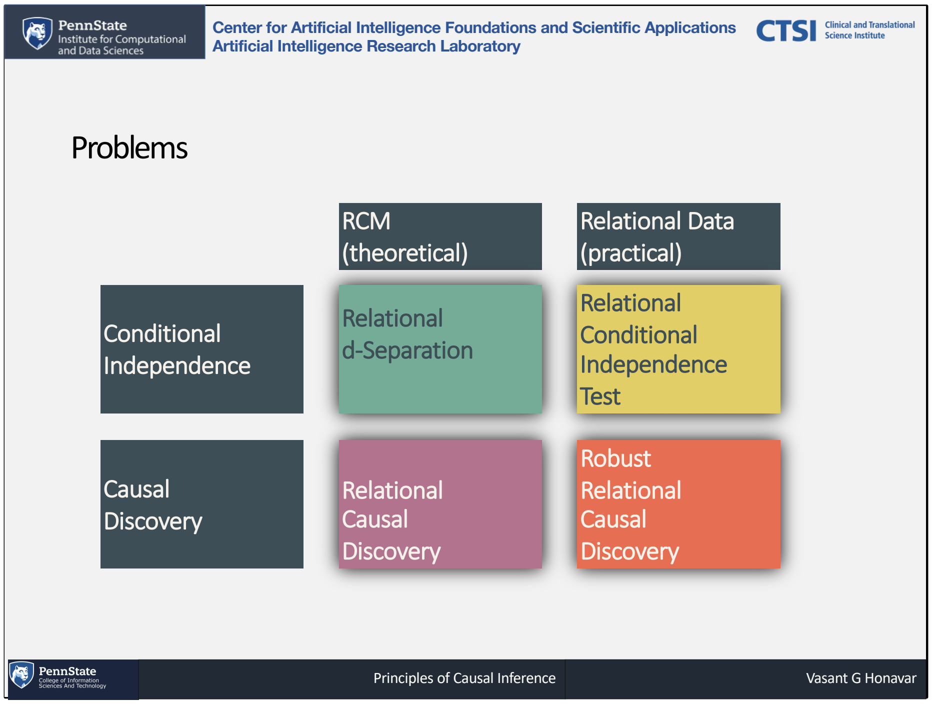
Theorem: Given access to the conditional independence oracle for an RCMM, the RCDL algorithm offers, under the adjacency faithfulness and orientation faithfulness assumptions, a sound and complete procedure for learning the structure of the RCM whose maximum number of hops of dependencies is bounded by h.[1]

*Lee, S. and Honavar, V., 2016, On learning causal models from relational data. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 3263-3270).

Learning the Structure of Relational Causal Models from Conditional Independence Queries

- RCDL algorithm[4] is provably sound and complete for learning the structure of relational causal models from independence queries
- Because Abstract Ground Graphs[1] do not represent all conditional independences in all ground graphs, RPC algorithm[1] is provably not complete[2,3,4] for learning the structure of relational causal models from independence queries
- Modified definition[5] of abstract ground graphs restores completeness of RPC[6]

[1]Maier et. al. 2013. A sound and complete algorithm for learning causal models from relational data. In *Proceedings of the Twenty-ninth Conference on Uncertainty in Artificial Intelligence*, pages 371–380
[2]Lee, S. and Honavar, V., 2015, Lifted representation of relational causal models revisited: implications for reasoning and structure learning. In *Proceedings of the UAI 2015 Conference on Advances in Causal Inference-Volume 1504* (pp. 56-65).
[3]Lee S, Honavar V. 2016. A characterization of Markov equivalence classes of Relational Causal Models under path semantics. In 32nd Conference on Uncertainty in Artificial Intelligence (pp. 387-396).
Learning the Structure of Relational Causal Models from Conditional Independence Queries
[5]Ahsan, Ragib, David Arbour, and Elena Zheleva. "Relational Causal Models with Cycles: Representation and Reasoning." *Conference on Causal Learning and Reasoning*. PMLR, 2022.
[6]Ahsan R, Arbour D, Zheleva E. Learning Relational Causal Models with Cycles through Relational Acyclification. arXiv preprint arXiv:2208.12210. 2022

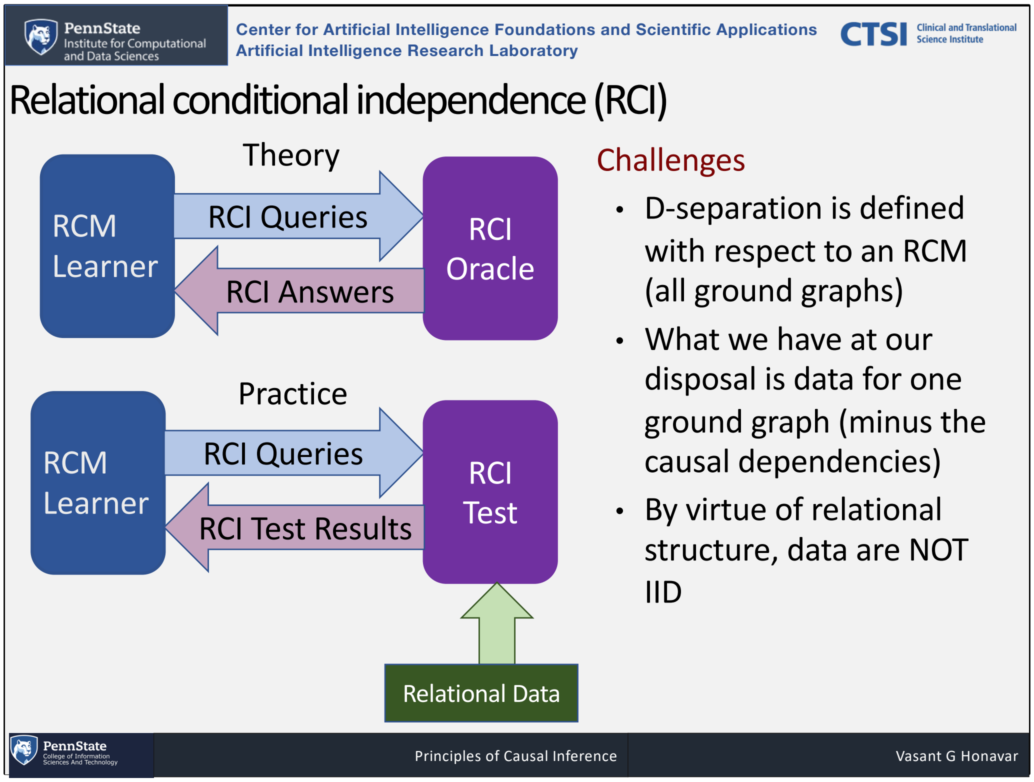


These four research questions can be laid out in a two-by-two table.

All these problems are **in fact** addressed by a group of researchers
but somewhat partly, narrowly, or incorrectly.

My dissertation begins by examining definitions of RCM and relevant concepts, and their implications.
Then, each of four problems

Relational conditional independence tests

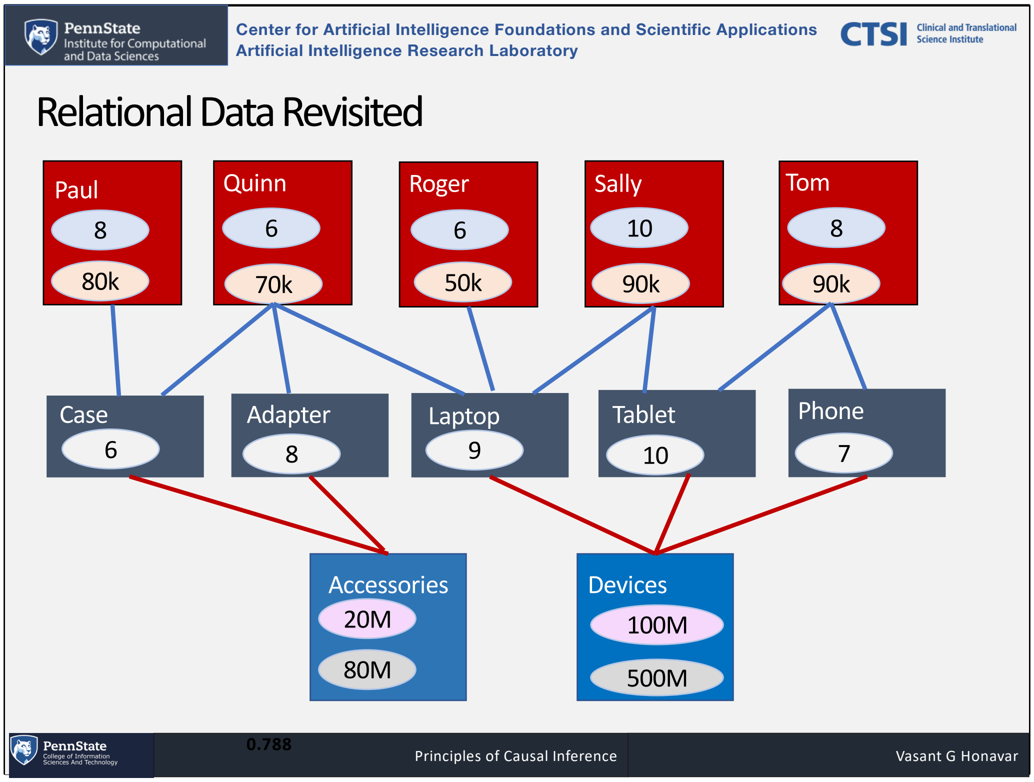


Can we test

RCI test is different from relational d-separation,
which is about examining RCI implied by a given model. which can be done in a qualitative manner.

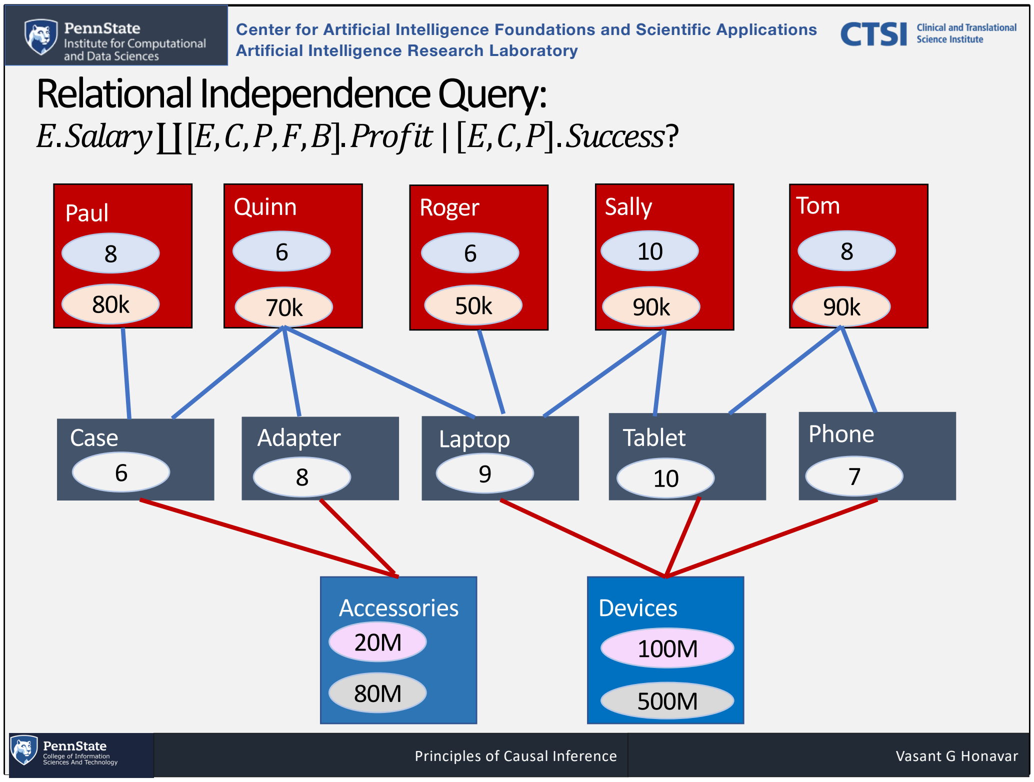
For RCI test, we have relational data
which might have been generated by a ground graph. The question is answered in a quantitative manner.

Testing RCI is not easy because
such relational data **exhibits** non-iidness
and we only have a single sample — which is the snapshot of a given network



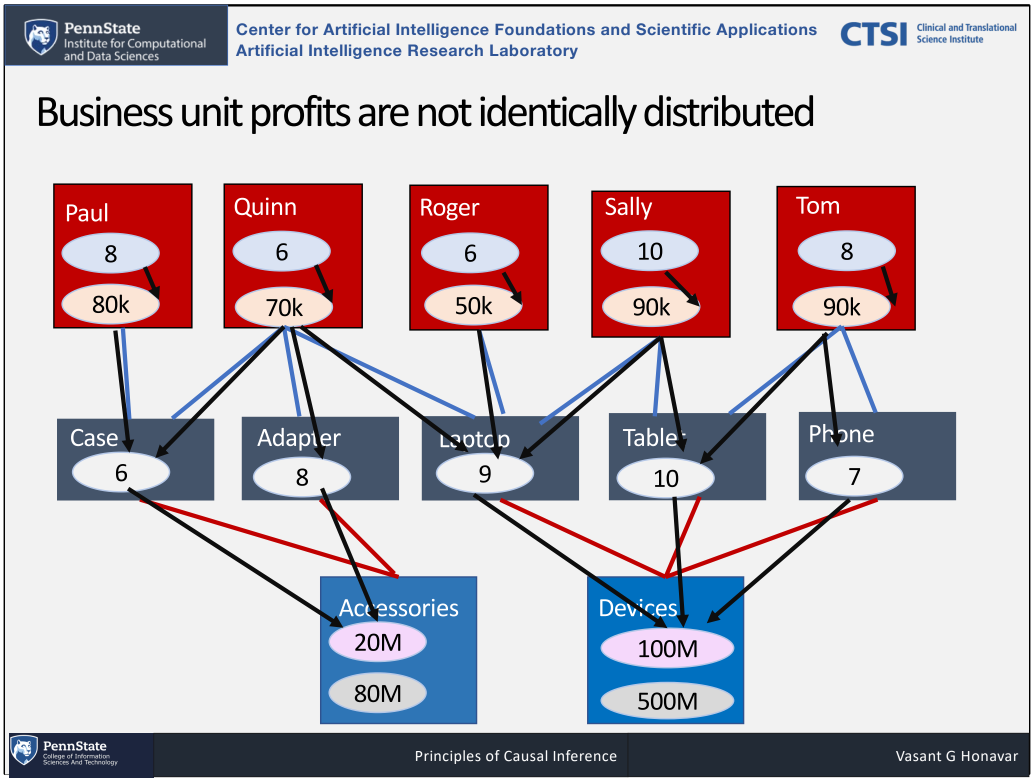
If we want to do something with relational data,
we need to understand them.

Relational data is a set of values for item attributes and the relational structure.



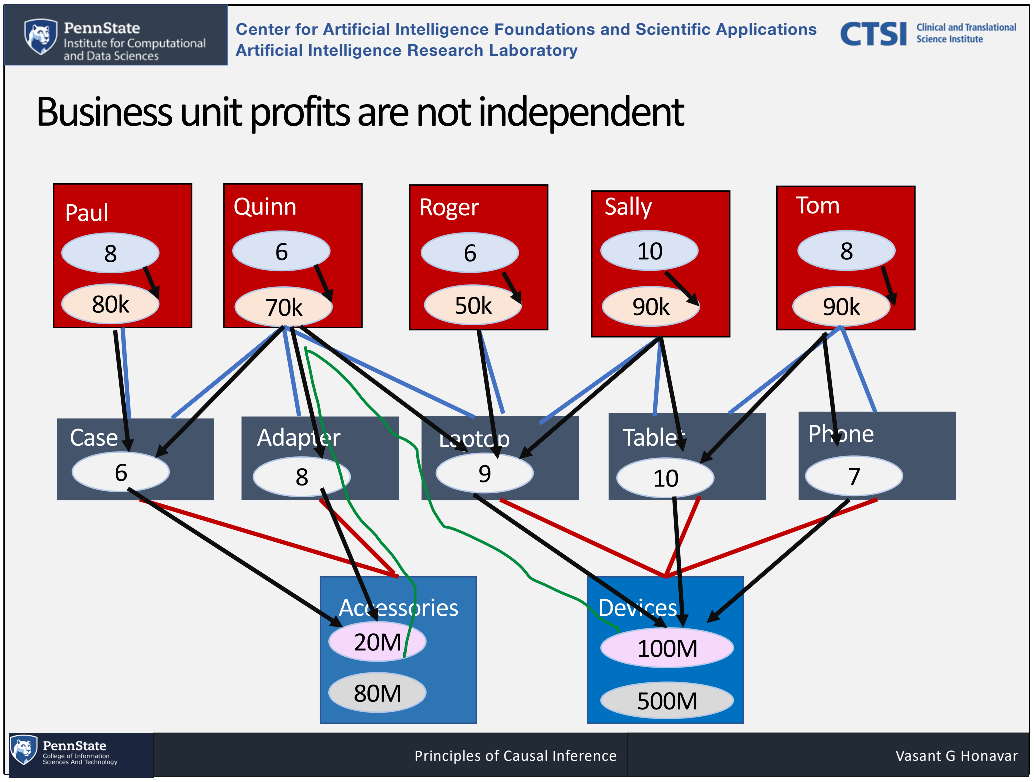
If we want to do something with relational data,
we need to understand them.

Relational data is a set of values for item attributes and the relational structure.



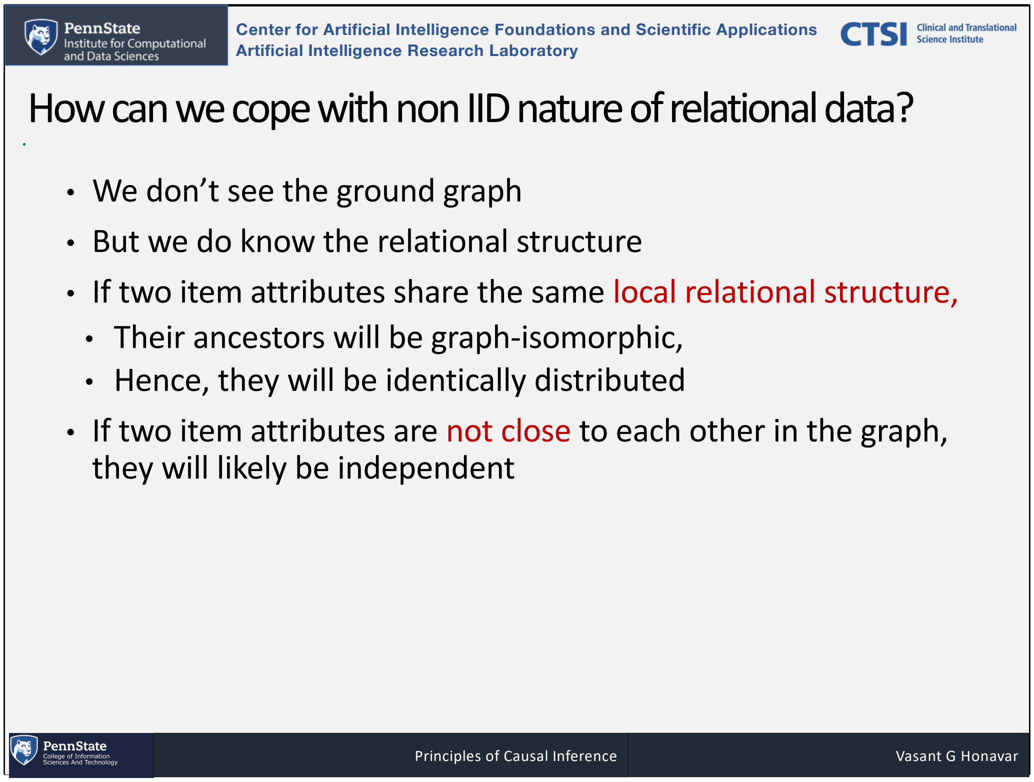
If we want to do something with relational data,
we need to understand them.

Relational data is a set of values for item attributes and the relational structure.



If we want to do something with relational data,
we need to understand them.

Relational data is a set of values for item attributes and the relational structure.



We briefly analyzed relational data **w.r.t. the underlying ground graph.**
But such ground graph is NOT observable.

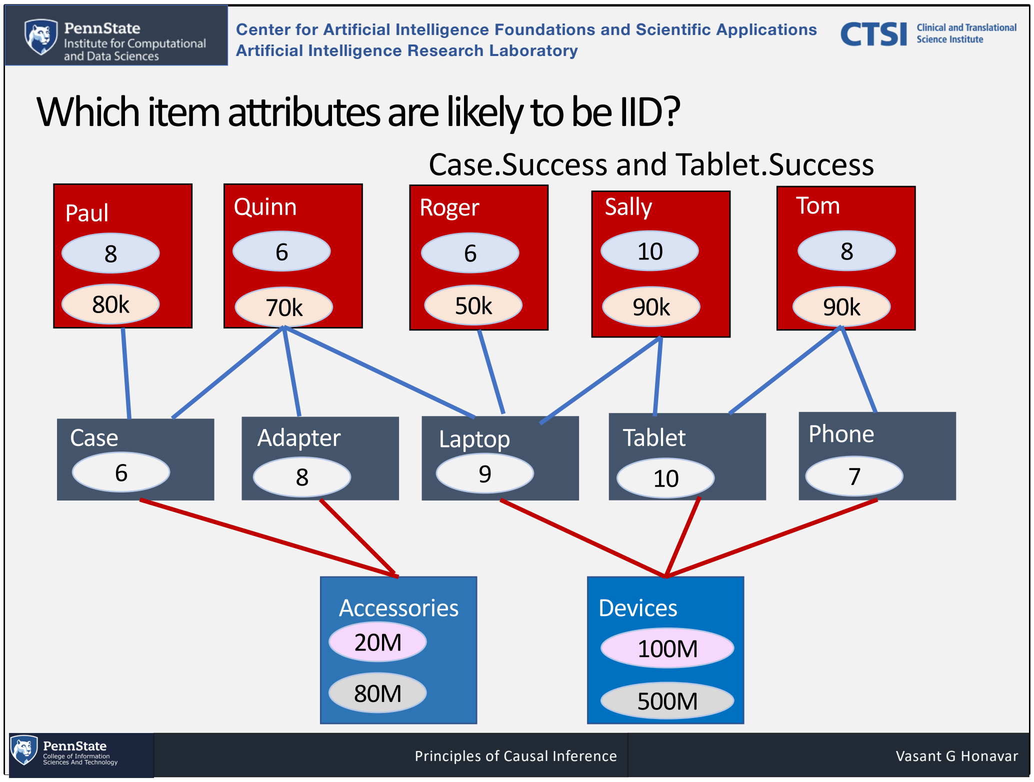
Instead, we will make use of the given relational skeleton.

With some assumptions,
IF THEN, IF THEN

take a look at success of case and success of tablet **for example**.
They are in fact iid.

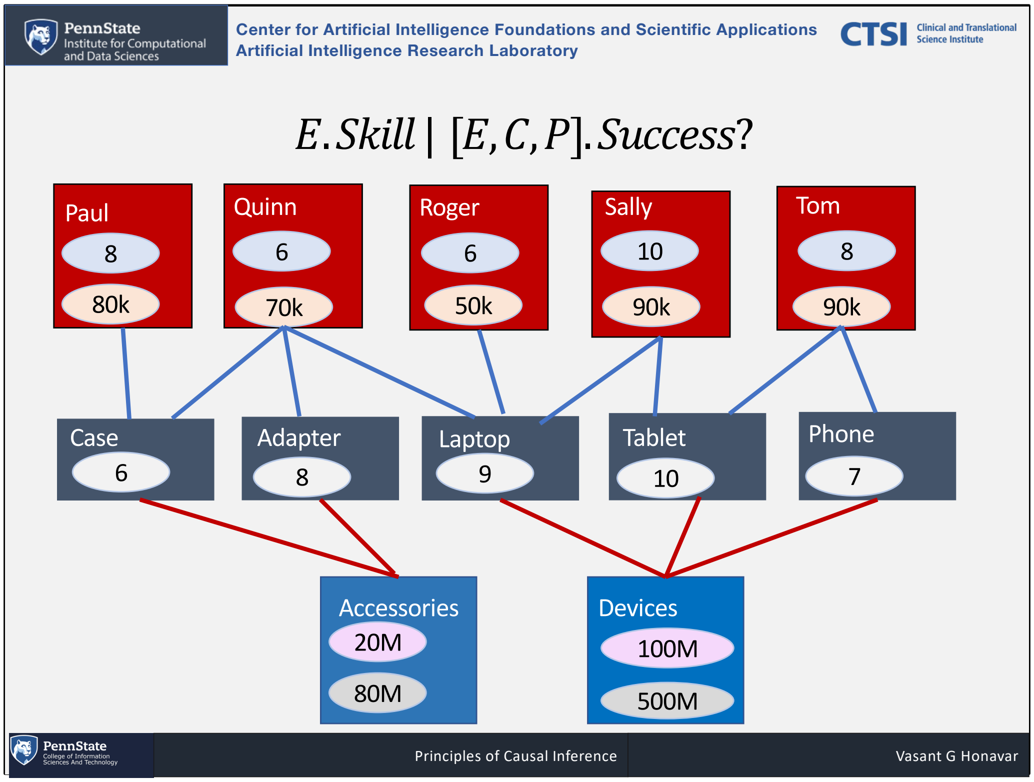
If we consider one-hop neighborhood as the choice of local relational structure,
both item attributes have two developers and one business unit.
Hence, we can conclude that they are identically distributed.

Further, they are relatively far away,
hence, we consider they are independent.



If we want to do something with relational data,
we need to understand them.

Relational data is a set of values for item attributes and the relational structure.



If we want to do something with relational data,
we need to understand them.

Relational data is a set of values for item attributes and the relational structure.

e item

Comp	[EDP].Succ
.247	-1.092
.958	-1.01.22.659
.238	2.659
.140	2.651.131
.508	-1.10.788

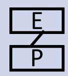
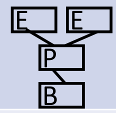
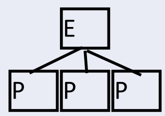
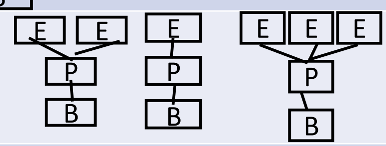
Preparing Data to Answer $[E].Skill \mid [E, C, P].Success$?

Base Item	$[E].Skill$	$[E, C, P].Success$
Paul	8	6
Quinn	6	6, 8, 9
Roger	6	9
Sally	10	9, 10
Tom	8	10, 7

- These data are not IID
  - Condition on the local structure of the graph to handle differences in distribution
  - Subsample the data to handle lack of independence

We first flatten the relational data

Preparing Data to Answer $[E].Skill \mid [E,C,P].Success, G$?

Base Item	$[E].Skill$	$[E,C,P].Success$
Paul	8 	6 
Quinn	6 	6, 8, 9 
Roger	6 ….	9 ….
Sally	10 …	9, 10 ….
Tom	8 ….	10, 7 ….

We first flatten the relational data

Conditional Independence Tests

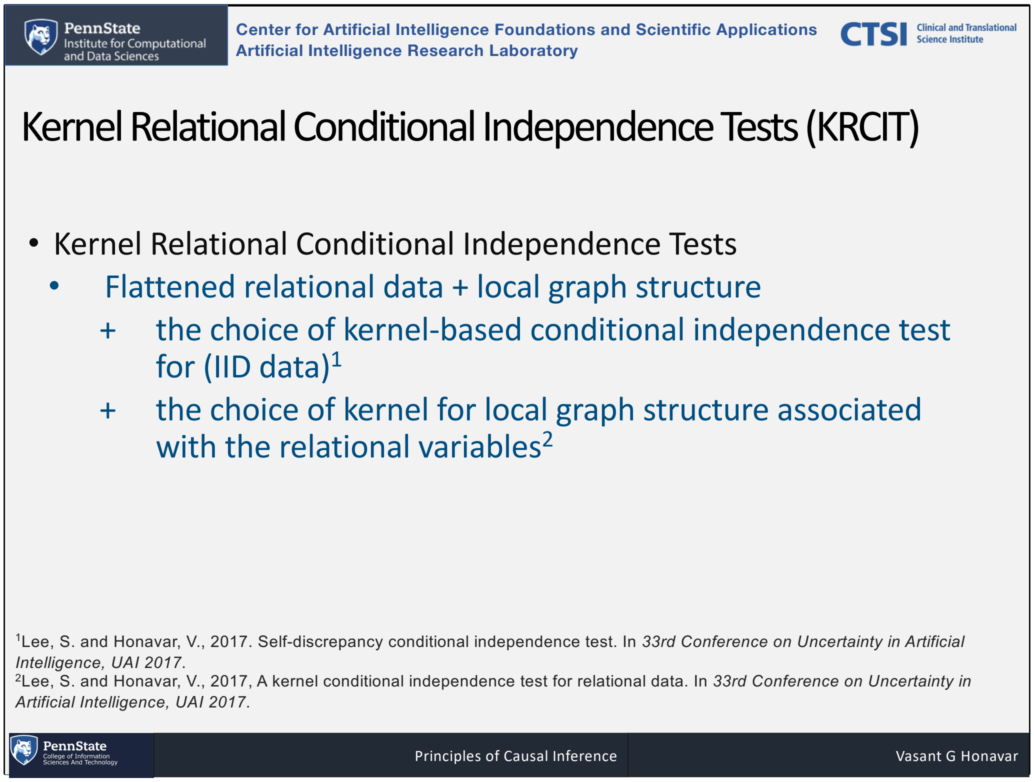
- Obtain a sample S of IID observations from a distribution $P(x, y, z)$
- Split the sample into two subsamples S_1 and S_2 of equal size. Leaving S_1 is intact, permute S_2 so as to simulate a sample from $P(x, z)P(y|z)$ - distribution where $X \coprod Y|Z$
- Apply a two-sample test to determine if S_1 and S_2 are from different distributions
- If the null hypothesis $P(x, y, z) = P(x, z)P(y|z)$ cannot be rejected, we conclude that X is independent of Y given Z
- It helps map X, Y, Z into a kernel induced feature space when their distributions are multimodal
- Several kernel independence tests exist for IID data[1,2,3]

[1]Gretton, Arthur, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. "A kernel statistical test of independence." *Advances in neural information processing systems* 20 (2007).
[2] Doran, G., Muandet, K., Zhang, K. and Schölkopf, B., 2014, A Permutation-Based Kernel Conditional Independence Test. In *UAI* (pp. 132-141).
[3]Lee, S. and Honavar, V., 2017. Self-discrepancy conditional independence test. In *33rd Conference on Uncertainty in Artificial Intelligence, UAI 2017*.

There exists a new class of conditional independence tests using kernel.
They are **powerful** and can capture **nonlinear dependencies.**
To use them, one needs to specify kernel functions — dot product in feature space.

We devised a new kernel-based CI test **which possesses many good properties.**

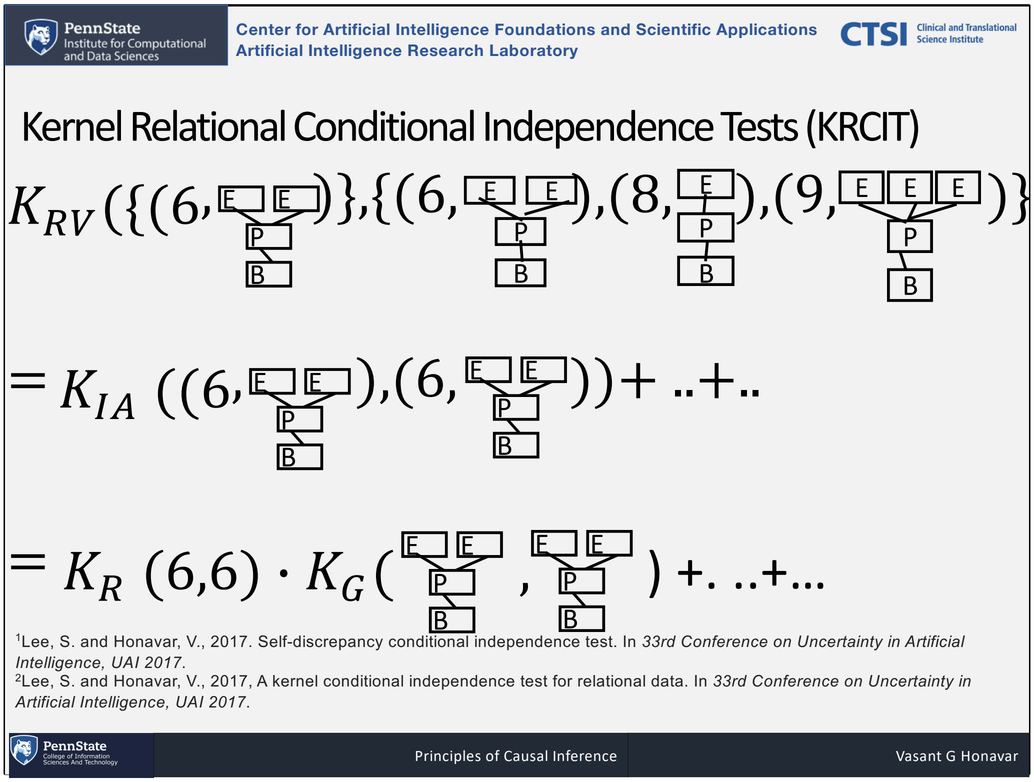
KRCIT, kernel relational conditional independence test,
is the combination of
Data representation as shown in the previous slide,
the choice of kernel-based CI test, here I use the one we proposed,
and the choice of kernel.



There exists a new class of conditional independence tests using kernel.
They are **powerful** and can capture **nonlinear dependencies.**
To use them, one needs to specify kernel functions — dot product in feature space.

We devised a new kernel-based CI test **which possesses many good properties.**

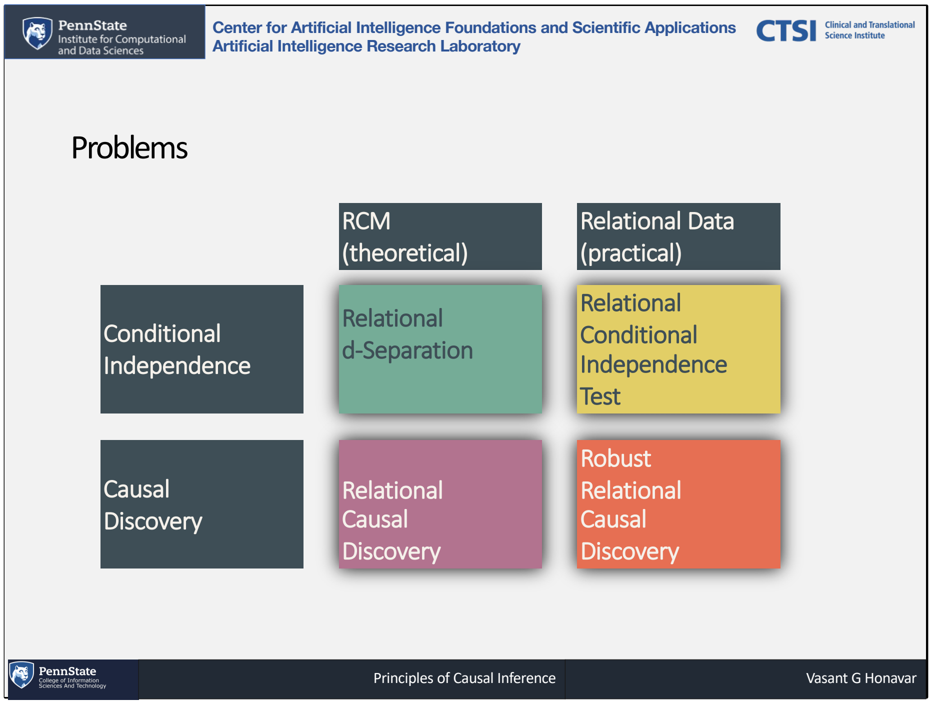
KRCIT, kernel relational conditional independence test,
is the combination of
Data representation as shown in the previous slide,
the choice of kernel-based CI test, here I use the one we proposed,
and the choice of kernel.



There exists a new class of conditional independence tests using kernel.
They are **powerful** and can capture **nonlinear dependencies.**
To use them, one needs to specify kernel functions — dot product in feature space.

We devised a new kernel-based CI test **which possesses many good properties.**

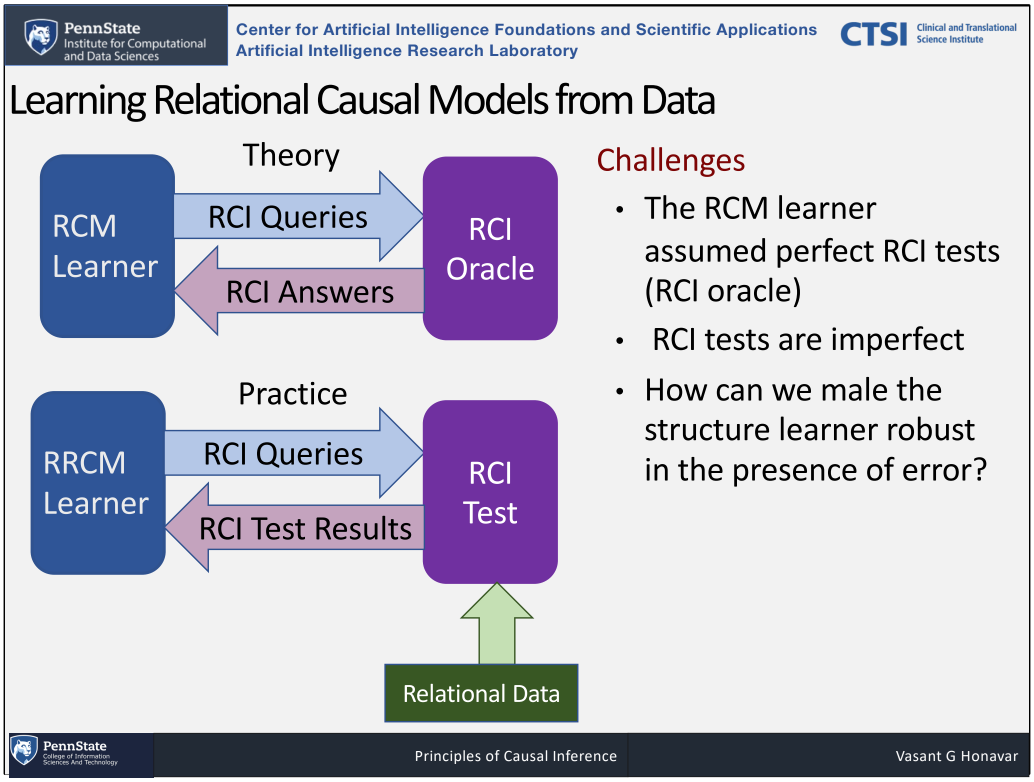
KRCIT, kernel relational conditional independence test,
is the combination of
Data representation as shown in the previous slide,
the choice of kernel-based CI test, here I use the one we proposed,
and the choice of kernel.



These four research questions can be laid out in a two-by-two table.

All these problems are **in fact** addressed by a group of researchers
but somewhat partly, narrowly, or incorrectly.

My dissertation begins by examining definitions of RCM and relevant concepts, and their implications.
Then, each of four problems



Can we test

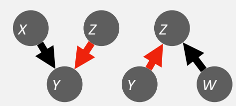
RCI test is different from relational d-separation,
which is about examining RCI implied by a given model. which can be done in a qualitative manner.

For RCI test, we have relational data
which might have been generated by a ground graph. The question is answered in a quantitative manner.

Testing RCI is not easy because
such relational data **exhibits** non-iidness
and we only have a single sample — which is the snapshot of a given network

Robust learning of RCM from Relational Data[1]

- **Phase I: Remove spurious relational dependencies**
  - Order-independence
  - Recovery from the violation of RCMC
  - Reducing variance of tests
  - …
- **Phase II: Orient relational dependencies**
  - alternatives to canonical unshielded triples
  - detecting conflicting RCI test results
  - …
- **Phase III: Resolves conflicts in orientations**
  - maximize # of consistent orientations



Now let’s talk about robustifying a ….. algorithm.

This is an overview how I made the structure learning algorithm more robust.

Among those,

I would like to introduce those three items.