1222 · 2022
800 ANNI

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Exploiting causality methods for knowledge discovery from observational data

Dario Simionato

dario.simionato@phd.unipd.it

Why study CAUSALITY?

Why study CAUSALITY?

Because

CORRELATION IS NOT CAUSATION

Why study CAUSALITY?

CORRELATION IS NOT CAUSATION

Why study CAUSALITY?

CORRELATION IS NOT CAUSATION

..but sometimes it gives strong hints

Why study CAUSALITY?

Because

CORRELATION IS NOT CAUSATION

Why study CAUSALITY?

Because

CORRELATION IS NOT CAUSATION

... and understanding how things work (causation) is the ultimate goal of science

Why study CAUSALITY?

… and understanding how things work (causation) is the ultimate goal of science

1. Which variables are “important” in my scenario?

2. How does changing one variable affect the system?

Why study CAUSALITY?

… and understanding how things work (causation) is the ultimate goal of science

1. **Which variables are "important" in my scenario?**
   ≈ Feature selection

2. How does changing one variable affect the system?

Correlational approaches fail at discovering "important" variables

Linear regression

$$\hat{y}_i = \mathrm{w}_0 + \mathrm{w}_1\, x_1 + \cdots + \mathrm{w}_N x_N$$

Trained by minimizing the mean squared error

$$\boldsymbol{w} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \frac{1}{N} \sum_i (y_i - \hat{y}_i)^2$$

Correlational approaches fail at discovering "important" variables

Linear regression

$$\hat{y}_i = \mathrm{w}_0 + \mathrm{w}_1 x_1 + \cdots + \mathrm{w}_N x_N$$

LASSO focuses on $\boldsymbol{w}$ such that

$$\boldsymbol{w} = \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{N} \sum_i (y_i - \hat{y}_i)^2 - \lambda \sum_i ||w_i||$$

Correlational approaches fail at discovering "important" variables

Solution obtained minimizing MSE

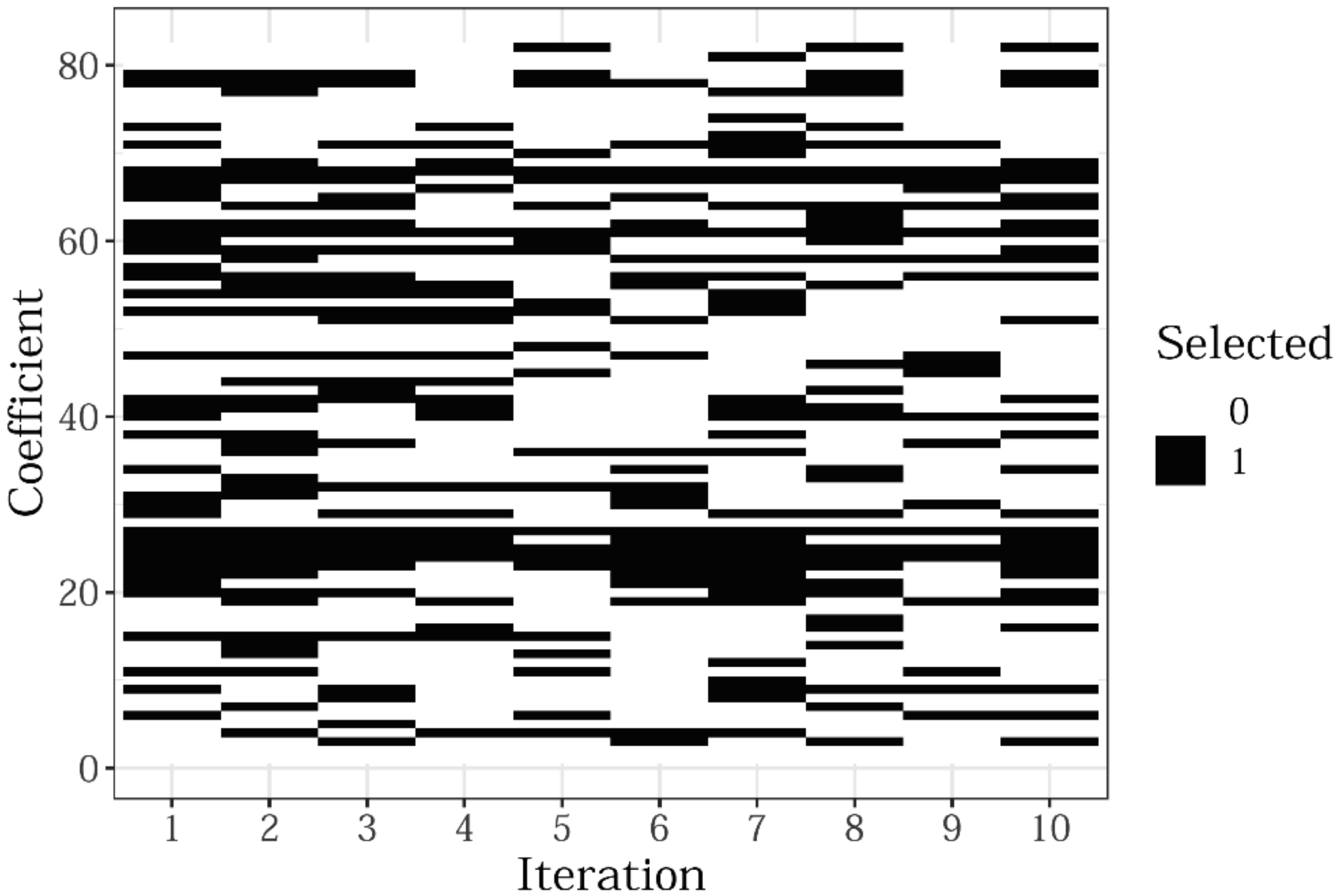
$$\hat{y}_i = \mathrm{w}_0 + \mathrm{w}_1\, x_1 + \mathrm{w}_2\, x_2 + \cdots + \mathrm{w}_{N-1}\, x_{N-1} + \mathrm{w}_N x_N$$

Solution obtained minimizing LASSO

$$\hat{y}_i = \mathrm{w}_0 + \mathrm{w}_1\, x_1 + \mathrm{w}_2\, x_2 + \cdots + \mathrm{w}_{N-1}\, x_{N-1} + \mathrm{w}_N x_N$$

Correlational approaches fail at discovering "important" variables

FIGURE 2: INCLUDED PREDICTORS IN A LASSO REGRESSION ACROSS TEN SAMPLES FROM THE SAME POPULATION



Gillis, T.B. and Spiess, J.L., 2019. Big data and discrimination. *The University of Chicago Law Review*, *86*(2), pp.459-488

Why study CAUSALITY?

… and understanding how things work (causation) is the ultimate goal of science

1. **Which variables are "important" in my scenario?**
   ≈ Feature selection

2. How does changing one variable affect the system?

Why study CAUSALITY?

… and understanding how things work (causation) is the ultimate goal of science

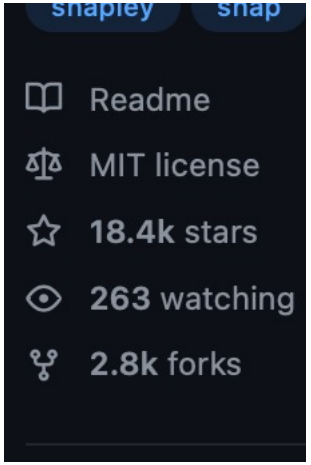
1. Which variables are "important" in my scenario?

2. How does changing one variable affect the system?

≈ **Interpretation**

Correlational approaches fail at discovering effects

SHAP



A unified approach to interpreting model predictions

[PDF] neurips.cc

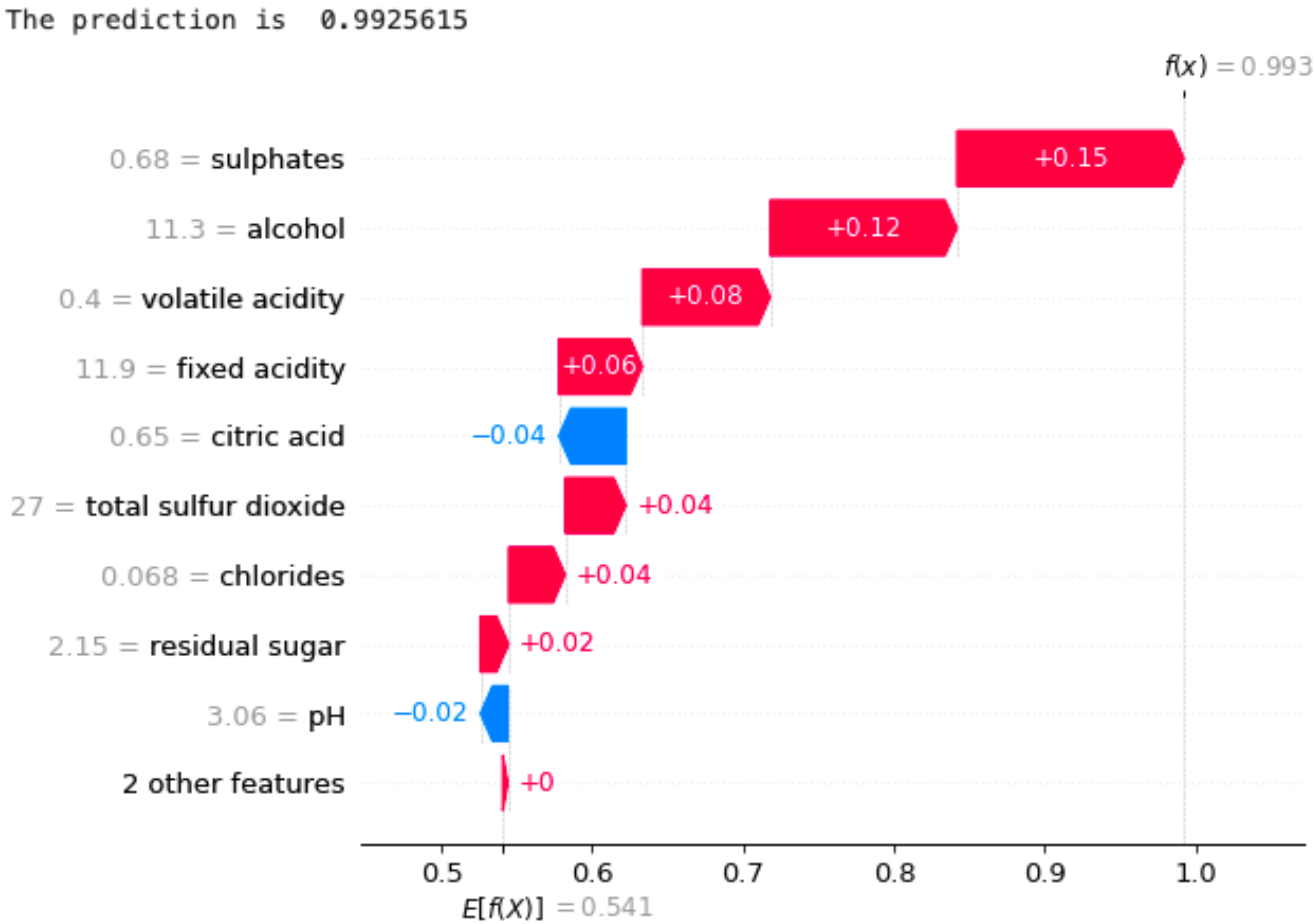
SM Lundberg, SI Lee - Advances in neural information …, 2017 - proceedings.neurips.cc

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between accuracy and interpretability. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and …

Save Cite Cited by 10824 Related articles All 17 versions

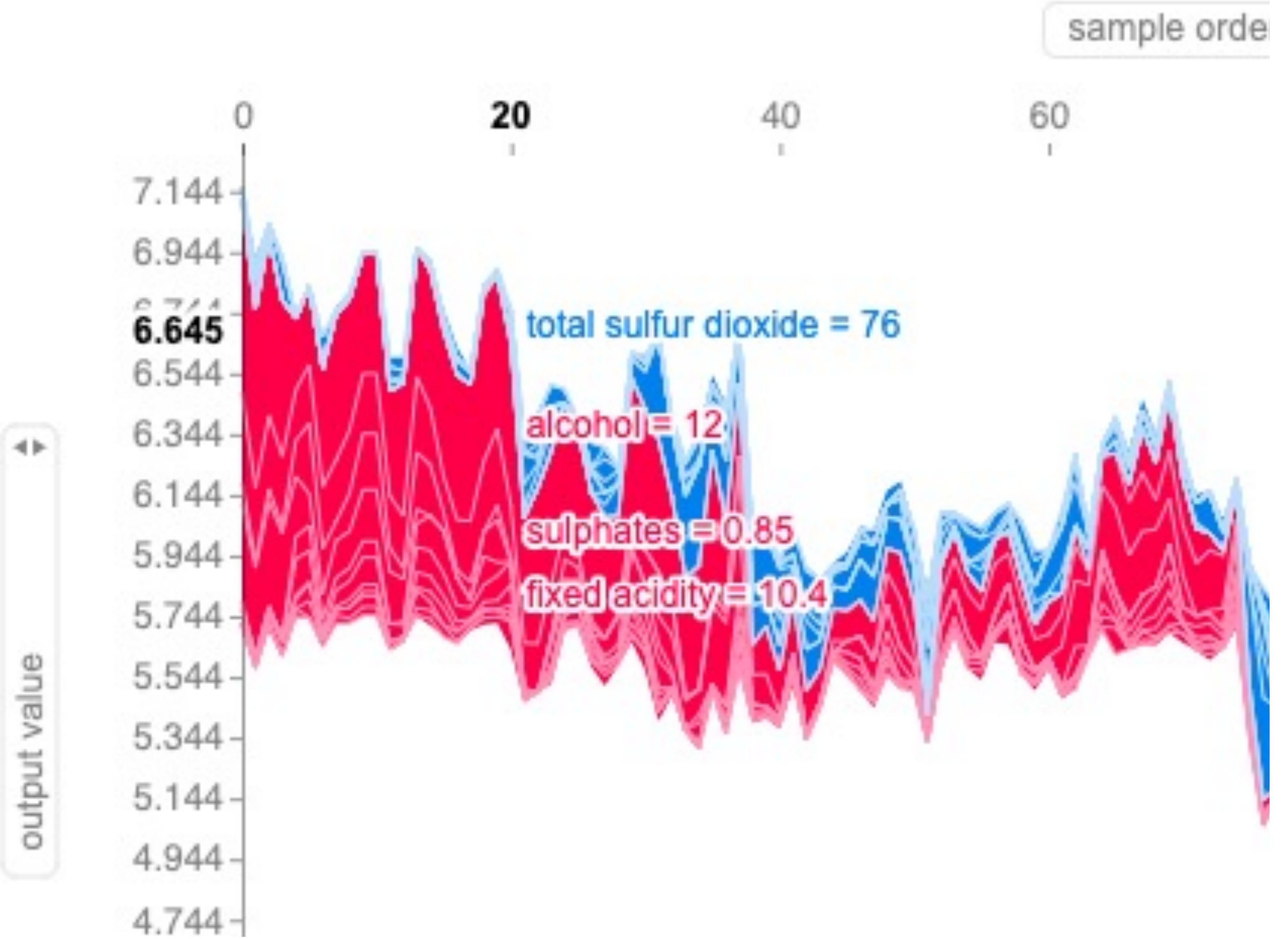
Correlational approaches fail at discovering effects

SHAP



Correlational approaches fail at discovering effects

SHAP



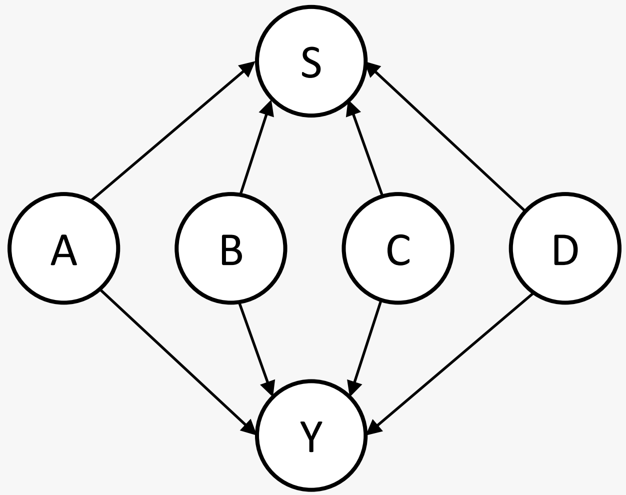
Correlational approaches fail at discovering effects

```
[96] #!pip install shap
     import shap
     import numpy as np
     import sklearn

     shap.initjs()

     n_el = 250
     p = 0.8
     a = np.random.binomial(1,p,n_el)
     b = np.random.binomial(1,p,n_el)
     c = np.random.binomial(1,p,n_el)
     d = np.random.binomial(1,p,n_el)
```

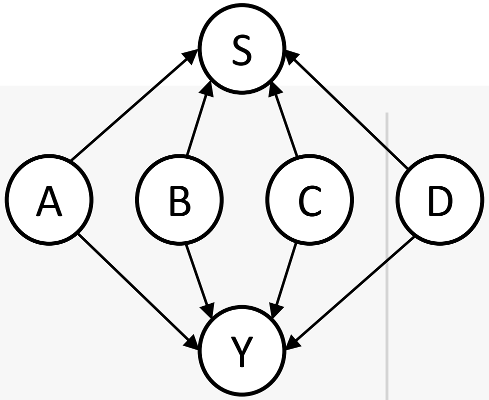
s = $(A \wedge B \wedge C \wedge D) \oplus B(0.1)$

y = $(A \wedge B \wedge C \wedge D) \oplus B(0.1)$

```
     X = np.reshape(np.array((a,b,c,d,s)),(5,-1)).T

     lr = sklearn.linear_model.LogisticRegression().fit(X,y)
```



Correlational approaches fail at discovering effects




```
n_el = 1000
p = 0.8
a2 = np.random.binomial(1,p,n_el)
b2 = np.random.binomial(1,p,n_el)
c2 = np.random.binomial(1,p,n_el)
d2 = np.random.binomial(1,p,n_el)
```

s2 = $(A_2 \wedge B_2 \wedge C_2 \wedge D_2) \oplus B(0.1)$

y2 = $(A_2 \wedge B_2 \wedge C_2 \wedge D_2) \oplus B(0.1)$

```
X2 = np.reshape(np.array((a2,b2,c2,d2,s2)),(5,-1)).T
lr.score(X2,y2)
```

```
0.842
```

```
lr.coef_
```

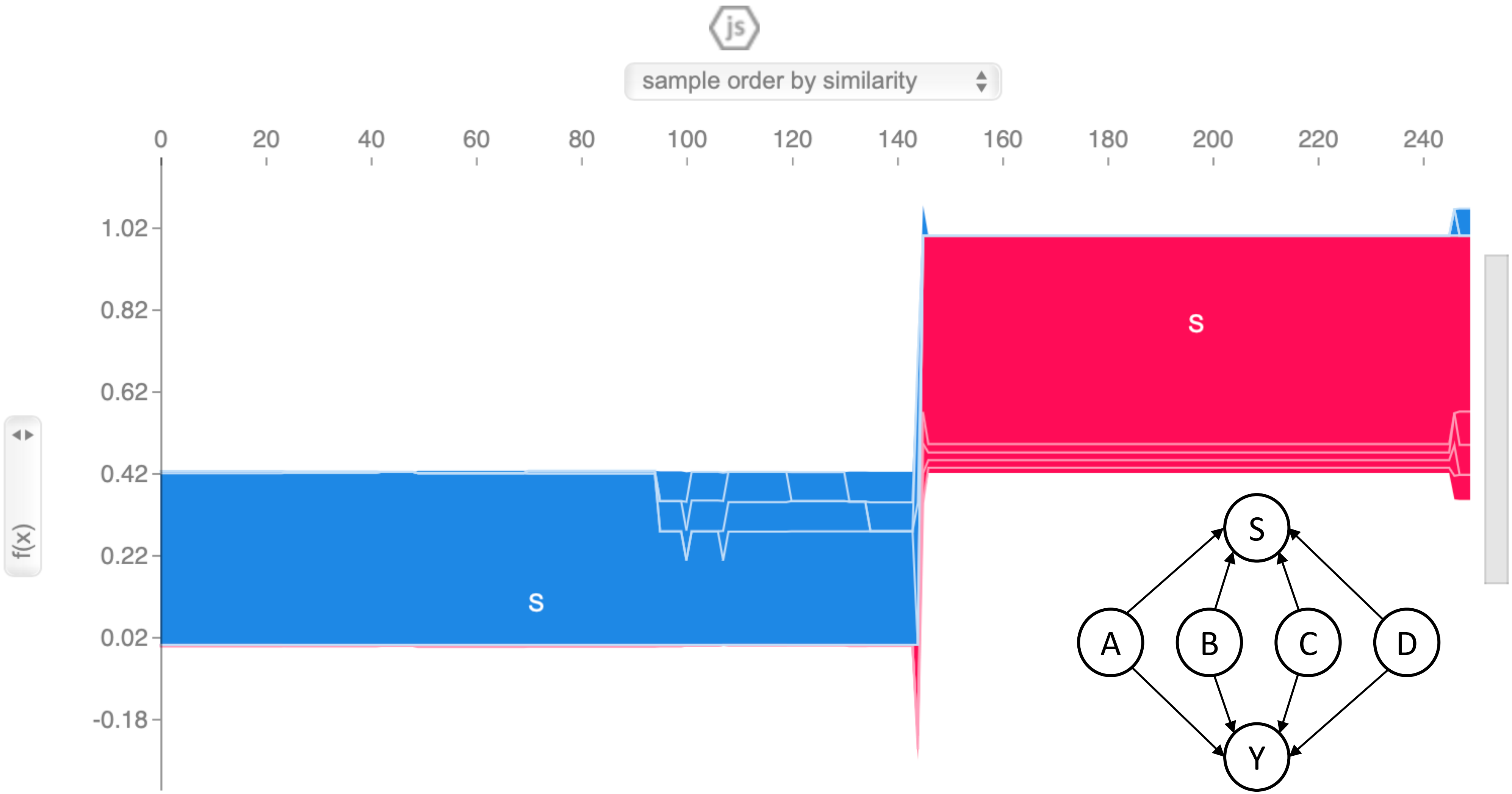
Coefficients of A B C D S

```
array([[1.11347288, 1.28590981, 0.98832866, 1.25790585, 2.09644493]])
```

Correlational approaches fail at discovering effects

[100]

```
shap.initjs()
shap.force_plot(explainer.expected_value, shap_values, pd.DataFrame(X,columns=["A","B'
```



Università degli Studi di Padova

Why study CAUSALITY?

…and understanding how things work (causation) is the ultimate goal of science

1. Which variables are “important” in my scenario?

2. How does changing one variable affect the system?

Causal framework (à la Pearl)

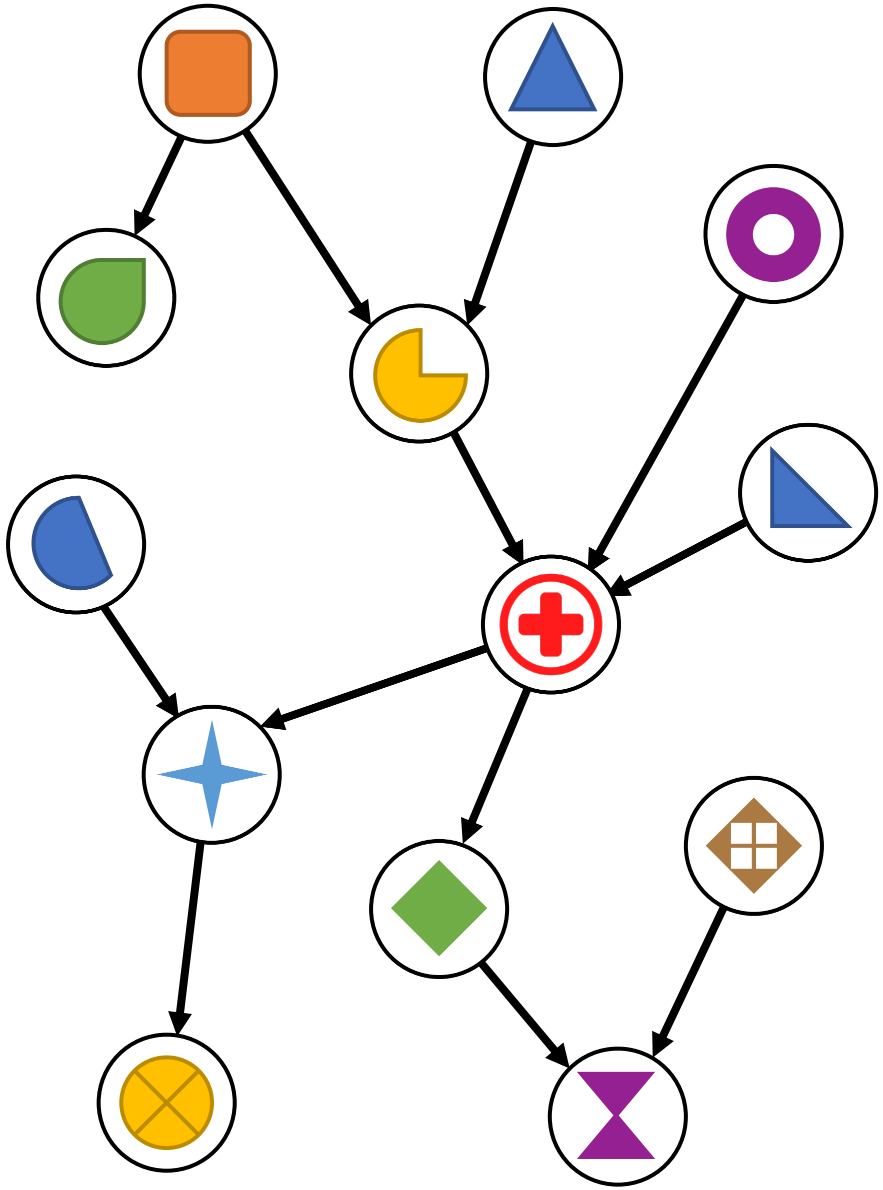
Causal Bayesian Networks are convient ways to model causal relationships among variables

$$BN = < G, p >$$

where

$$\underline{p}(X_i) = f(pa(X_i), \epsilon)$$

Interventional probability

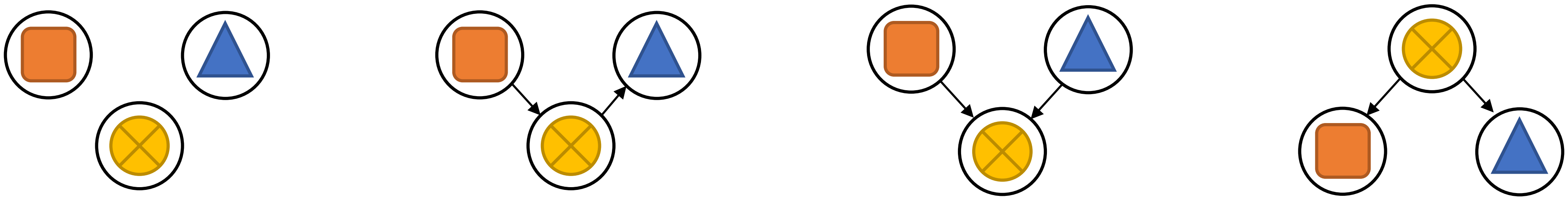


Causal inference objectives

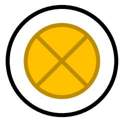
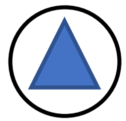
(Some) main objectives of causal techniques:

- Which variables are "important" in my scenario?

  Structure discovery: How are variables linked?



- How does changing one variable affect the system?

  Effect estimation: How changing  affects  ?

My focus

Develop algorithms to:

- (Structure discovery) Discover causally related variables to a target
- (Effect estimation) Evaluate effect of causal rules

From observational data and providing guarantees on the results

Observational vs. Interventional

Causal BNs are built using **interventional data** (e.g. setting variable $X_i = x_i$)

Observational data is much more common

Th. [informal] If spurious dependencies are removed, observational and interventional probability distributions are equivalent

My focus

Develop algorithms to:

- (Structure discovery) Discover causally related variables to a target
- (Effect estimation) Evaluate effect of causal rules

From observational data and providing guarantees on the results

Providing guarantees

Statistical guarantees are NOT just evaluation of perfomances

Guarantees are fundamental to gain users trust

Typically, algorithms with guarantees focus on False Discoveries (or False Positives)

My focus

Develop algorithms to:

- (Structure discovery) Discover causally related variables to a target
- (Effect estimation) Evaluate effect of causal rules

From observational data and providing guarantees on the results

My focus

Develop algorithms to:

- (Structure discovery) **Discover causally related variables to a target**
- (Effect estimation) Evaluate effect of causal rules

From observational data and providing guarantees on the results

My focus

Develop algorithms to:

- (Structure discovery) **Discover causally related variables to a target**

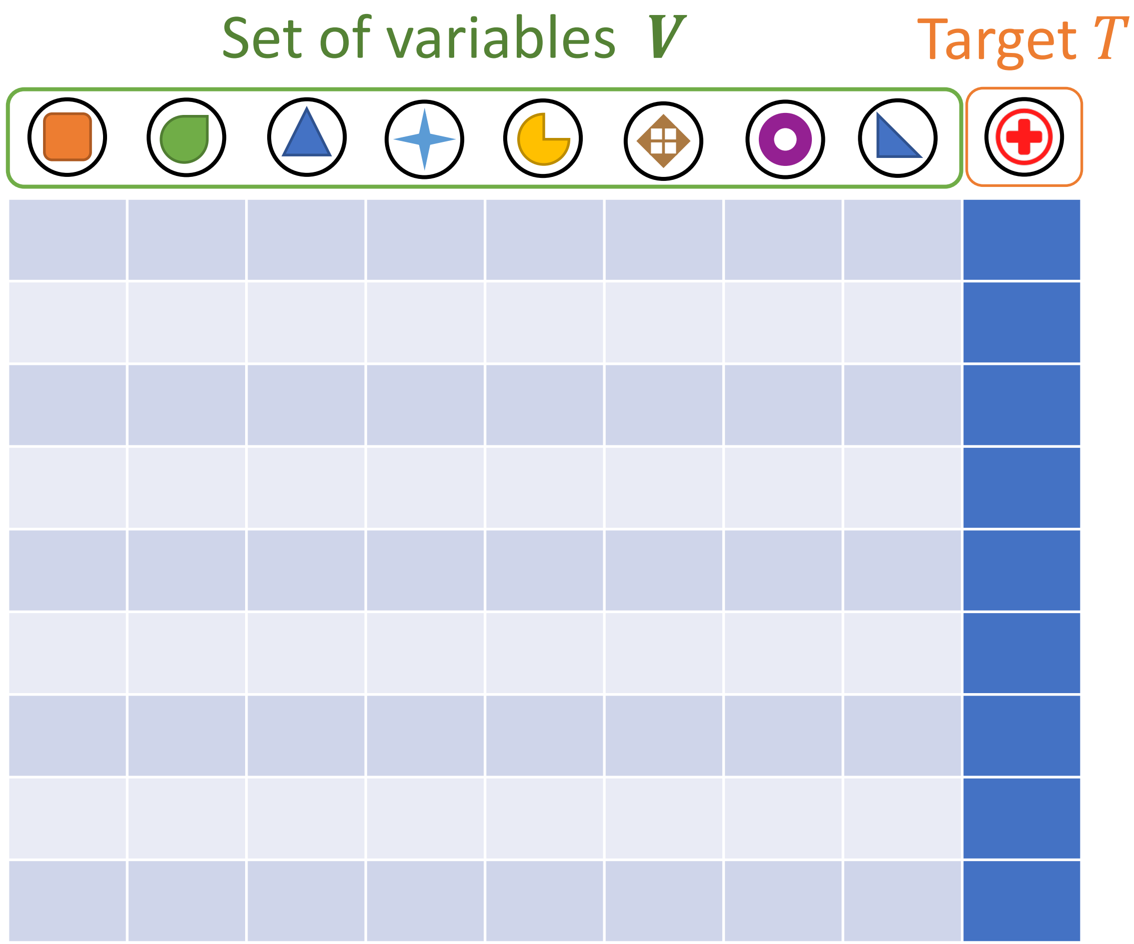
Simionato Dario, and Fabio Vandin. "Bounding the Family-Wise Error Rate in Local Causal Discovery using Rademacher Averages."

Accepted at ECML PKDD 2022 – Best paper award

Problem Definition

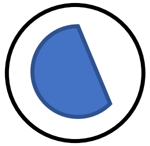
Task: Given a dataset of observations of variables $\boldsymbol{V}$, find those **causally** related to T with **guarantees** on the result (e.g. no false positives)

Useful in: Biology, medicine, neuroscience

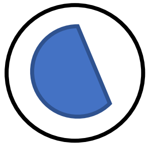


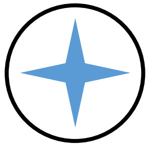
Problem Definition

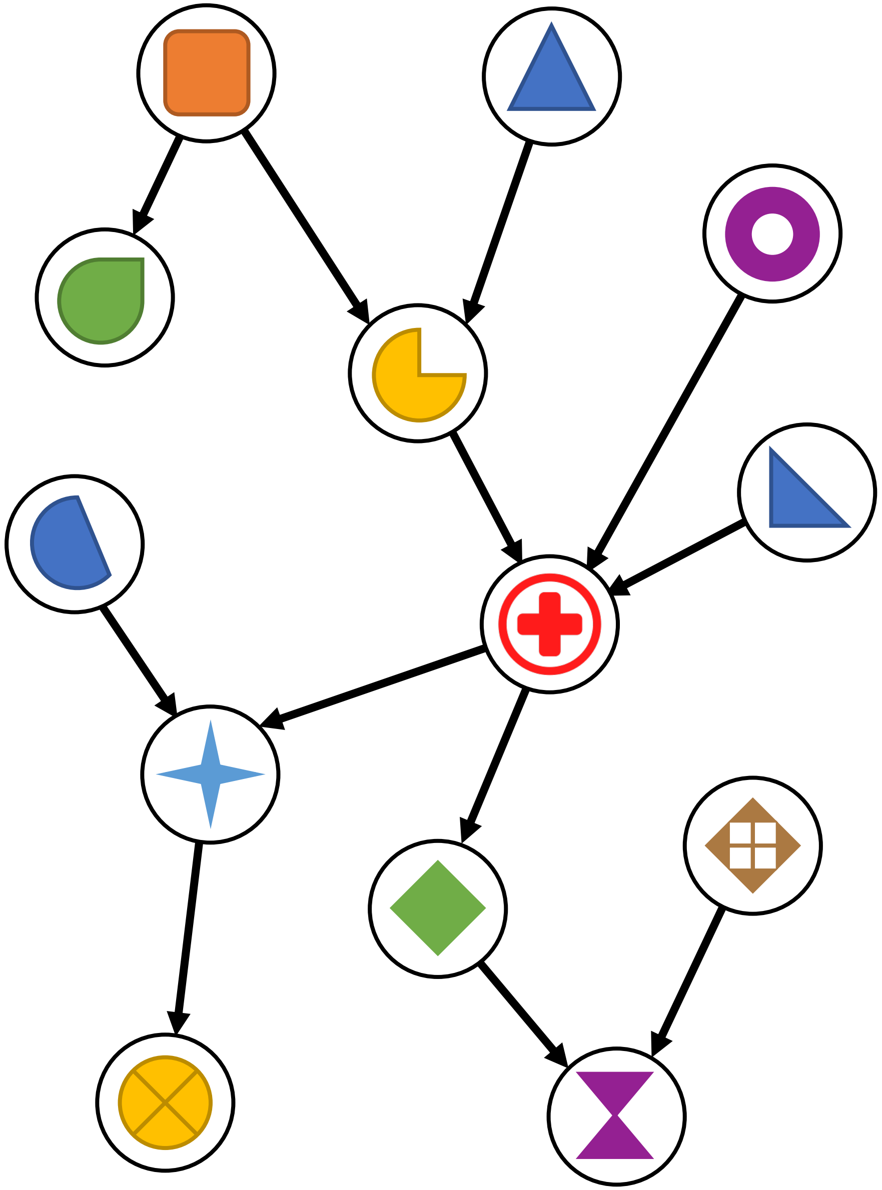
Causal Bayesian Networks represent cause-consequence relations between variables

Informally, if  is a **cause** of  ,

then fixing all variables values and changing

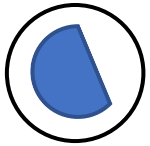
the value of  leads to a variation on

the values of 

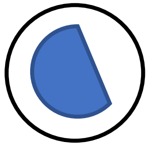


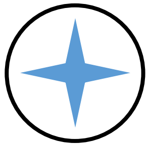
Problem Definition

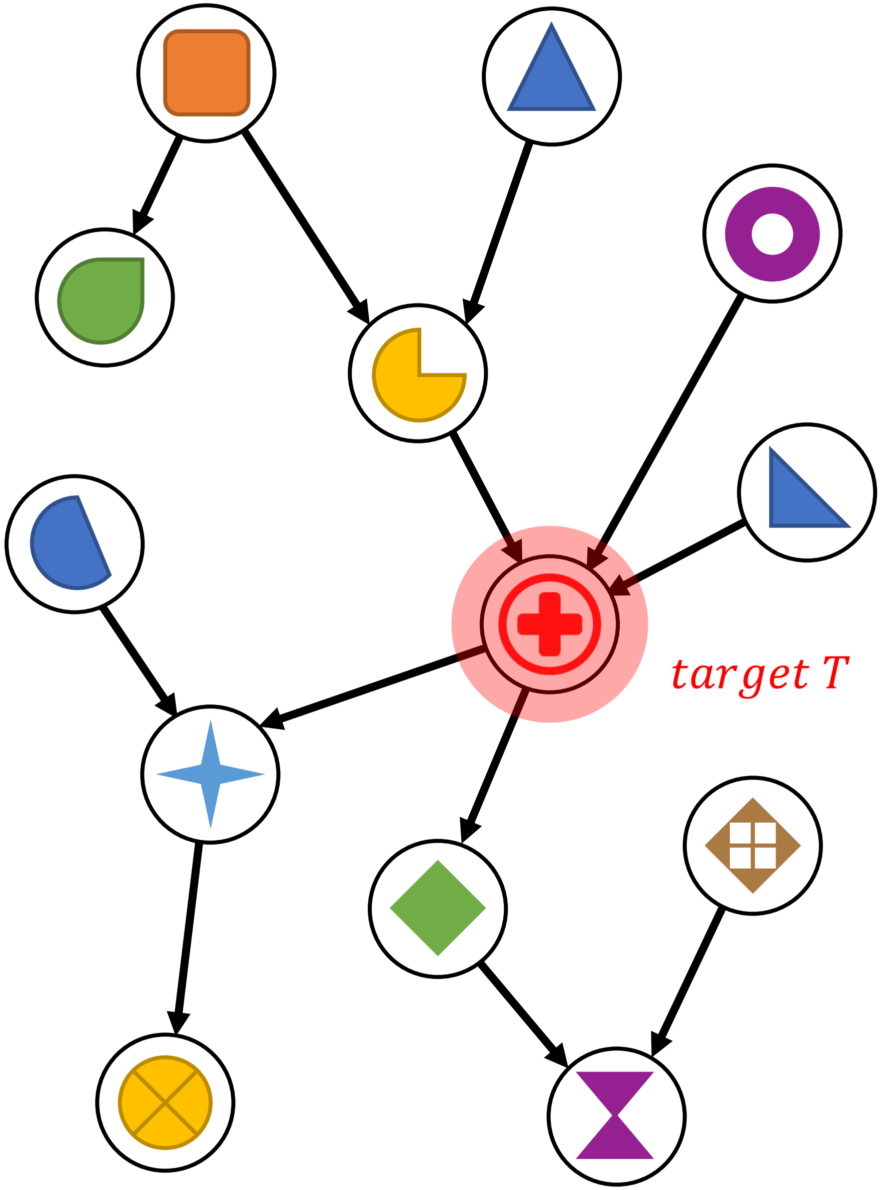
Causal Bayesian Networks represent cause-consequence relations between variables

Informally, if  is a **cause** of ,

then fixing all variables values and changing

the value of  leads to a variation on

the values of 



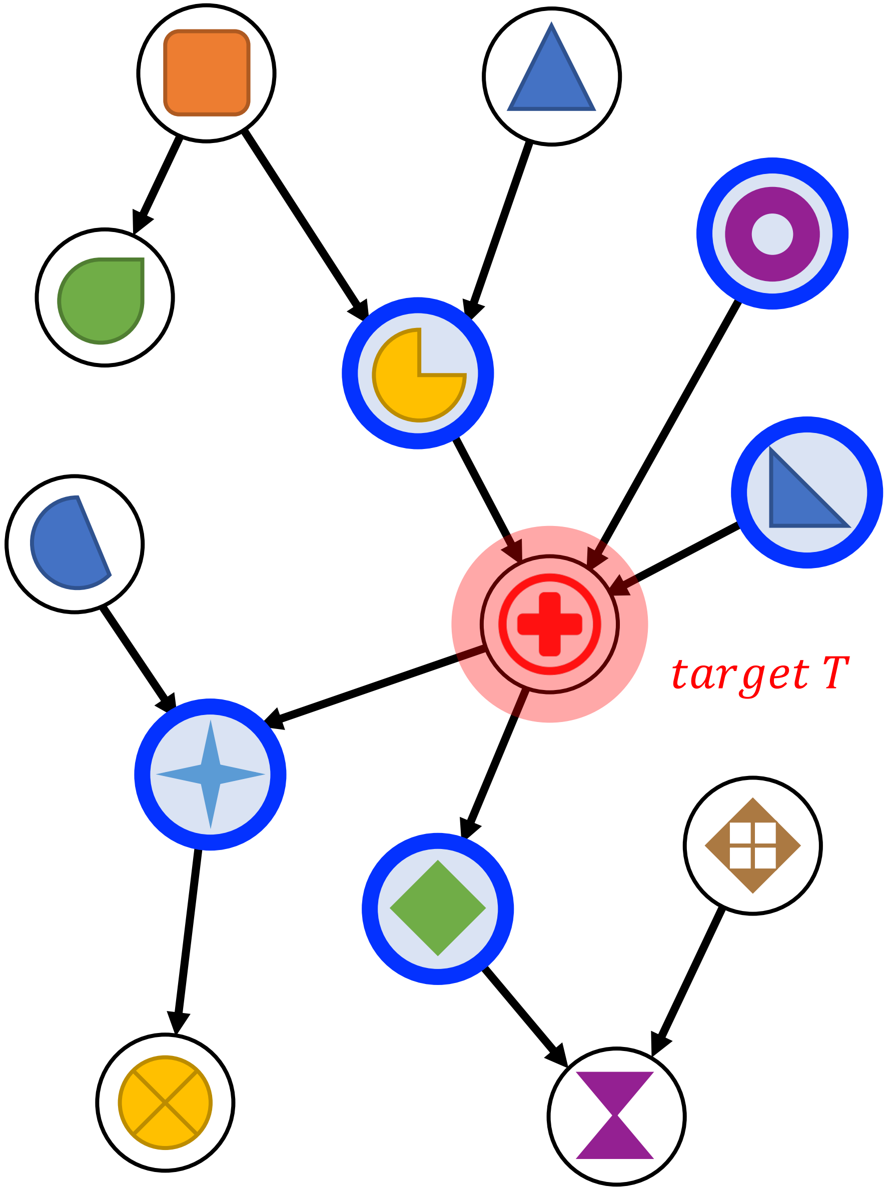
Problem Definition

Local causal discovery focuses on:

- **Parent – Children set** of T $\boldsymbol{PC(T)}$

  Parents(T) + Children(T)

  Contains direct causes and consequences of T



Problem Definition

Local causal discovery focuses on:

- **Parent – Children set** of T $\boldsymbol{PC(T)}$

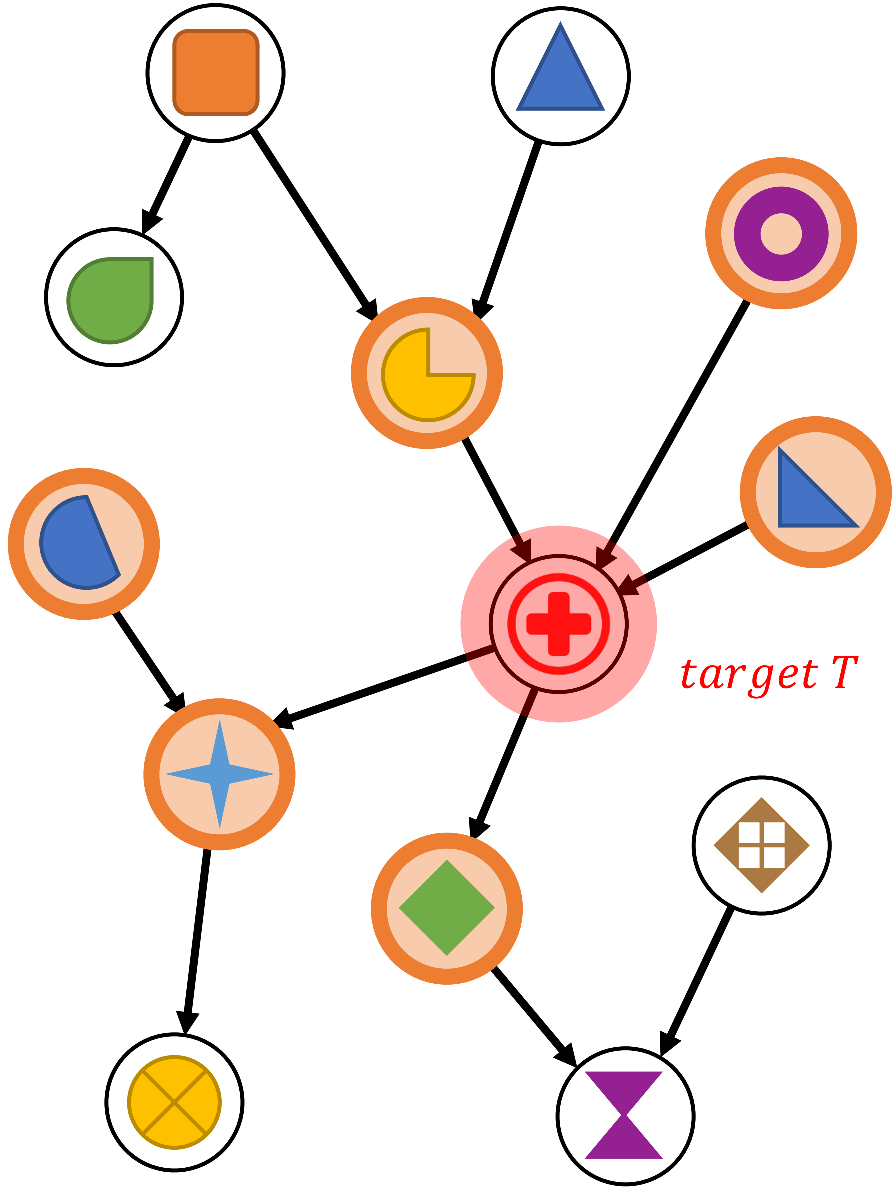
  Parents(T) + Children(T)

  Contains direct causes and consequences of T

- **Markov Boundary** of T $\boldsymbol{MB(T)}$

  $PC(T)$ + Spouses(T)

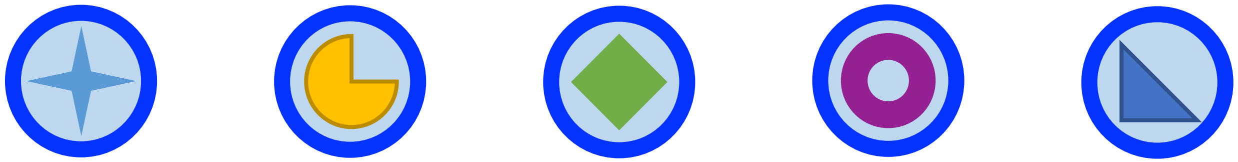
  Optimal set for the prediction of T



Problem Definition

Given a dataset and a target T, our task is to discover:

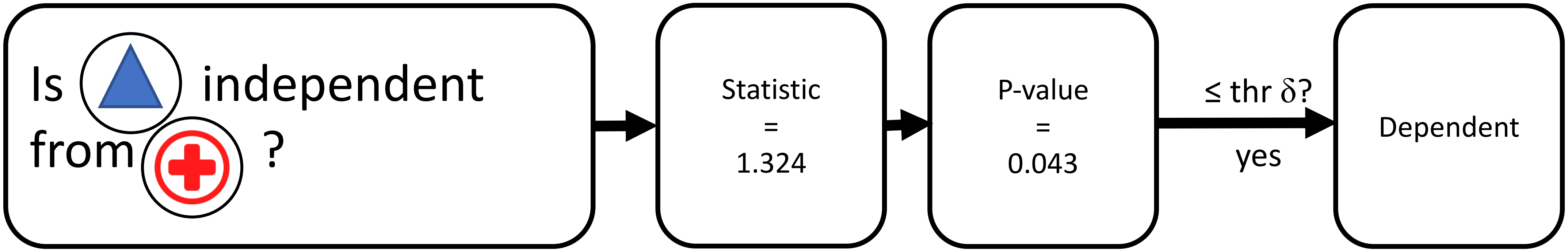
- **Parent – Children set** of T $\boldsymbol{PC(T)}$



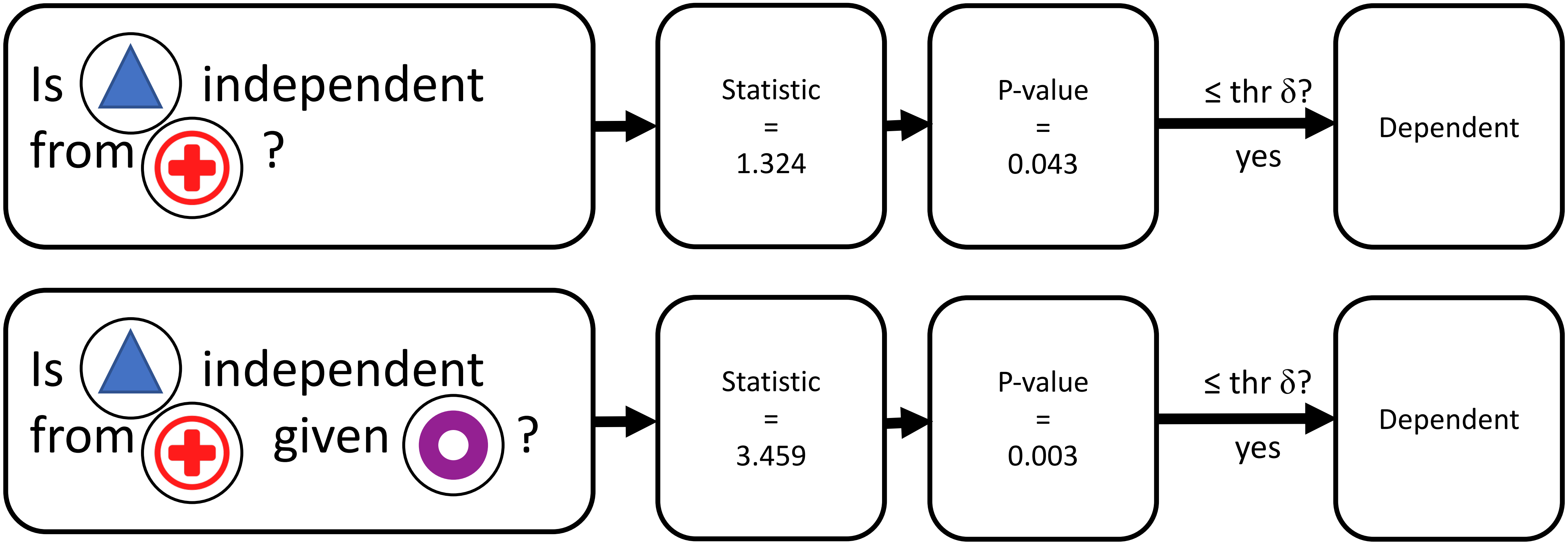
- (or) **Markov Boundary** of T $\boldsymbol{MB(T)}$

Several approaches are proposed in the literature
[Pena et al. '07], [Aliferis et al. '10], [Aliferis et al. '03]

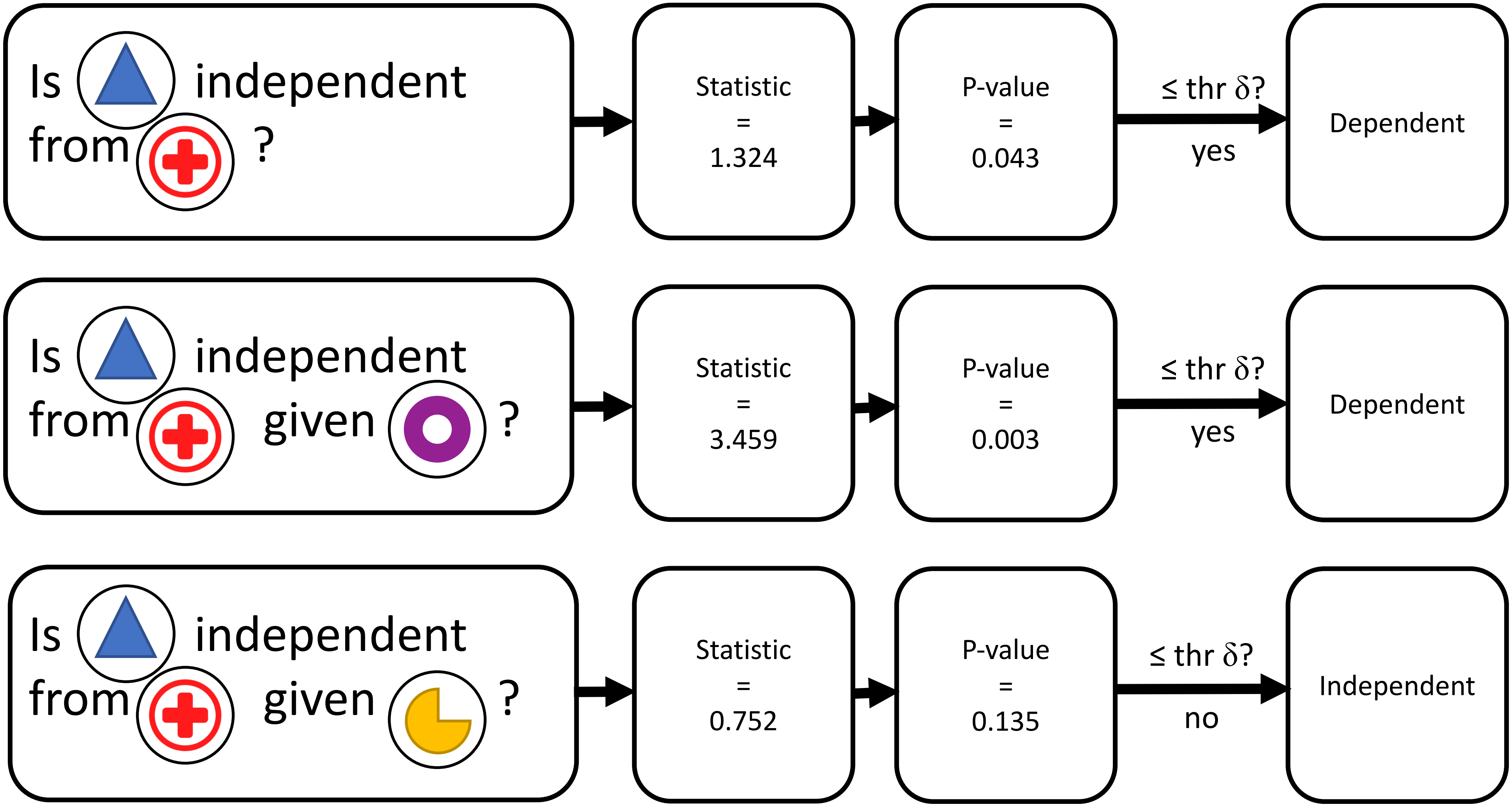
Statistical Testing and Guarantees



Statistical Testing and Guarantees

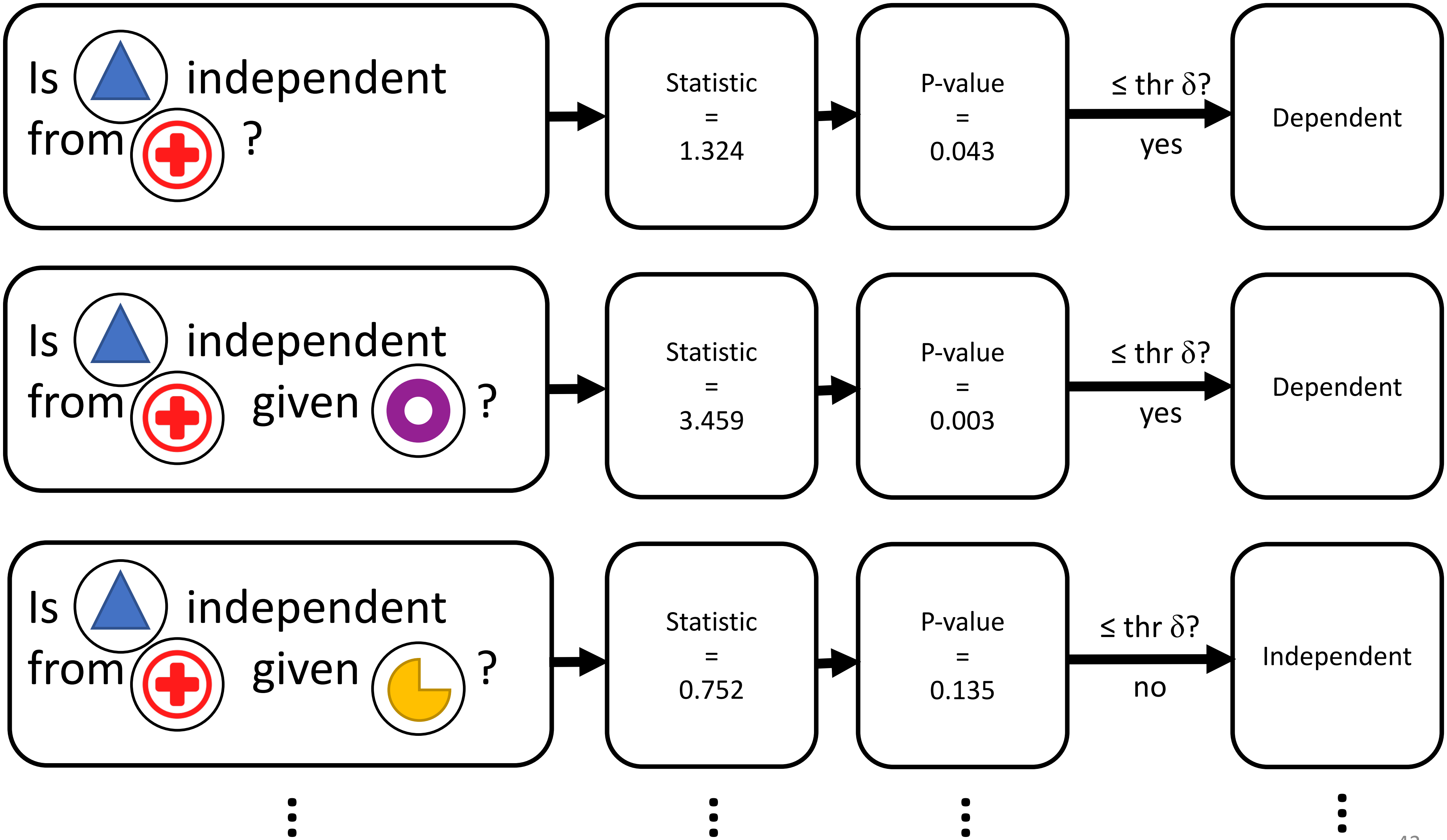


Statistical Testing and Guarantees



UNIVERSITÀ DEGLI STUDI DI PADOVA

Statistical Testing and Guarantees



Statistical Testing and Guarantees

When testing N hypotheses

$$Number\ of\ false\ positives \sim Bin(N, \delta)$$

$$P(At\ least\ one\ false\ positive) = 1 - (1 - \delta)^N \geq \delta$$

In our problem $N = |\boldsymbol{V}|(|\boldsymbol{V}| + 1)2^{|V|-2}$ is the number of possible conditional independence tests.

Statistical Testing and Guarantees

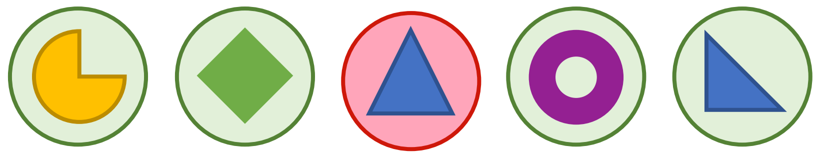
Approaches with guarantees typically focus on bounding

Correct solution

False Discovery Rate

$$FDR = E\left[\frac{Number\ false\ discoveries}{Total\ discoveries}\right]$$

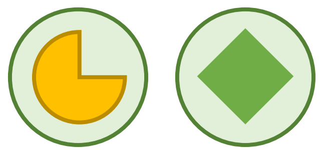
Typical solution (controlling the FDR)



Family-Wise Error Rate

$$FWER = P(Return\ at\ least\ one\ false\ positive)$$

Typical solution (controlling the $FWER$)



Statistical Testing and Guarantees

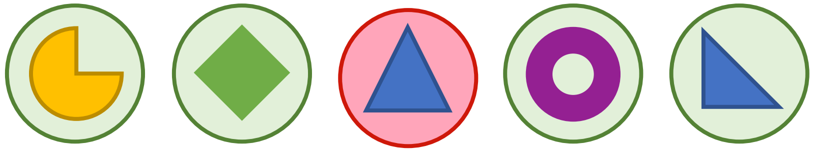
Approaches with guarantees typically focus on bounding

Correct solution

False Discovery Rate

$$FDR = E\left[\frac{Number\ false\ discoveries}{Total\ discoveries}\right]$$

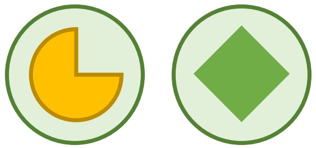
Typical solution (controlling the FDR)



Family-Wise Error Rate

$$FWER = P(Return\ at\ least\ one\ false\ positive)$$

Typical solution (controlling the $FWER$)



Our focus

Previous Works

Assume perfect detection of dependencies and independencies

[Pena et al. '07], [Aliferis et al. '10], [Aliferis et al. '03]

- Unfeasible and untestable assumptions

Bound the **false discovery rate**

[Tsamardinos and Brown '08]

- May still return false positives

Our Contributions

- Proved that **SoA algorithms cannot control the $FWER$** by correcting for multiple hypothesis testing

- Developed **RAveL-PC and RAveL-MB**: the first algorithms for local causal discovery with **guarantees on the $FWER$**

- Implemented bounds on $FWER$ exploiting **classical corrections** and **data-dependent** bounds based on **Rademacher averages**

- **Tested** RAveL-PC and RAveL-MB both on **synthetic** and **real-world** datasets

Università degli Studi di Padova

SoA Algorithm Analysis

Th. [informal] State of the art algorithms (GetPC, PCMB and IAMB) control the $FWER$ if they correct for multiple hypothesis testing and some strong assumptions (infinite power) are met.

We showed some examples in which **removing such assumptions** may lead to returning **false positives in output**

RAveL Pipeline

Our algorithms:

- Formulate the local discovery task using **only independence tests** (and not dependence tests)
- **Apply suitable corrections** (Rademacher, Bonferroni) to control for the $FWER$

RAveL Pipeline

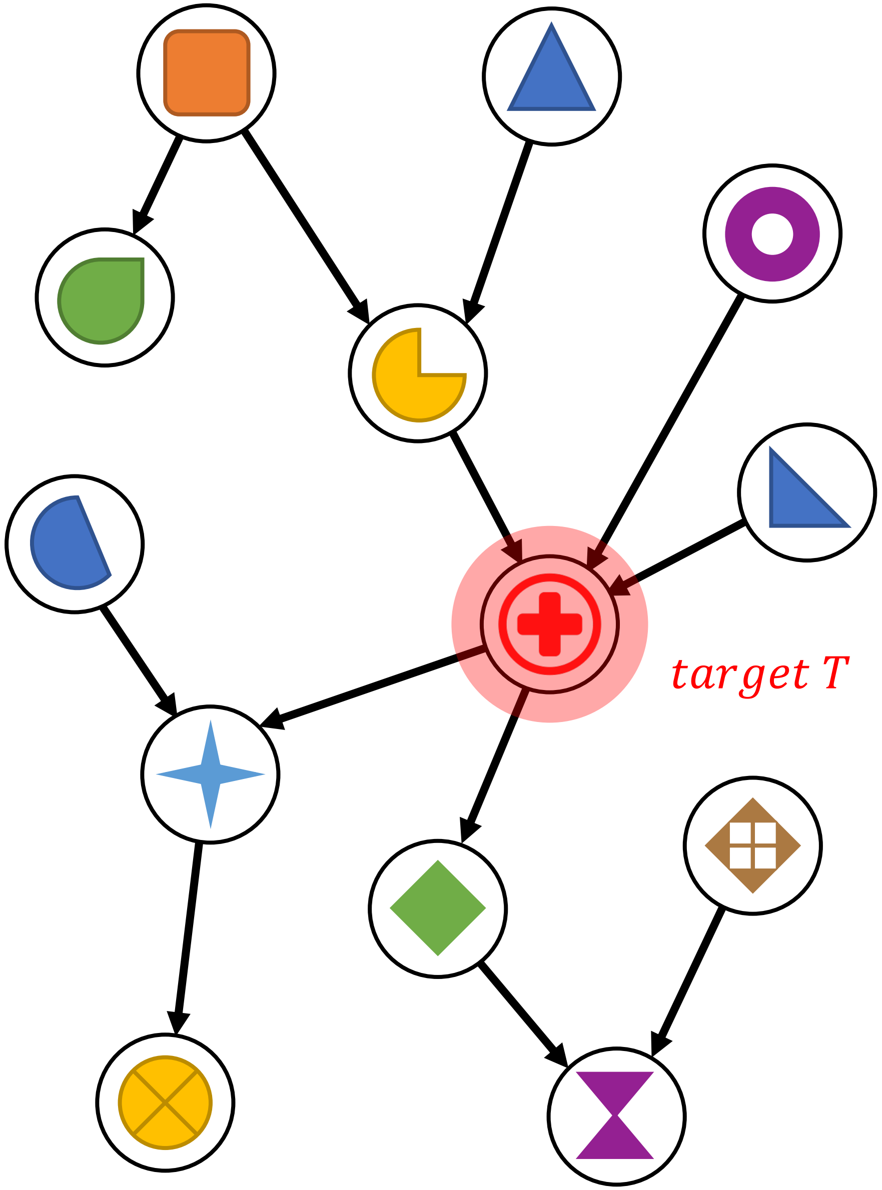
Our algorithms:

- Formulate the local discovery task using **only independence tests** (and not dependence tests)
- **Apply suitable corrections** (Rademacher, Bonferroni) to control for the $FWER$

Th. [informal] RAveL-PC and RAveL-MB effectively control the $FWER$ below a given threshold δ.

RAveL-MB at a Glance

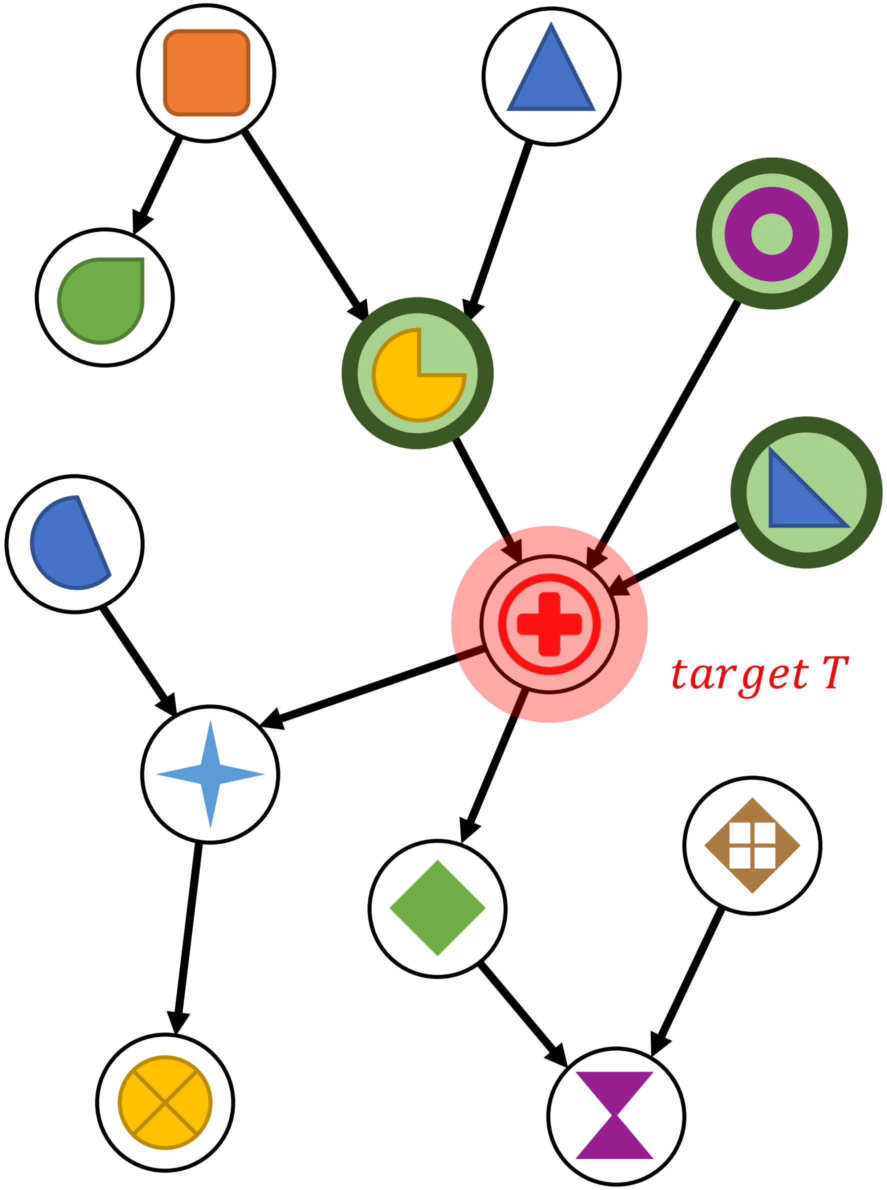
Task: Discover a subset of $\boldsymbol{MB}(\boldsymbol{T})$ without returning any false positive



RAveL-MB at a Glance

Task: Discover a subset of $MB(T)$ without returning any false positive

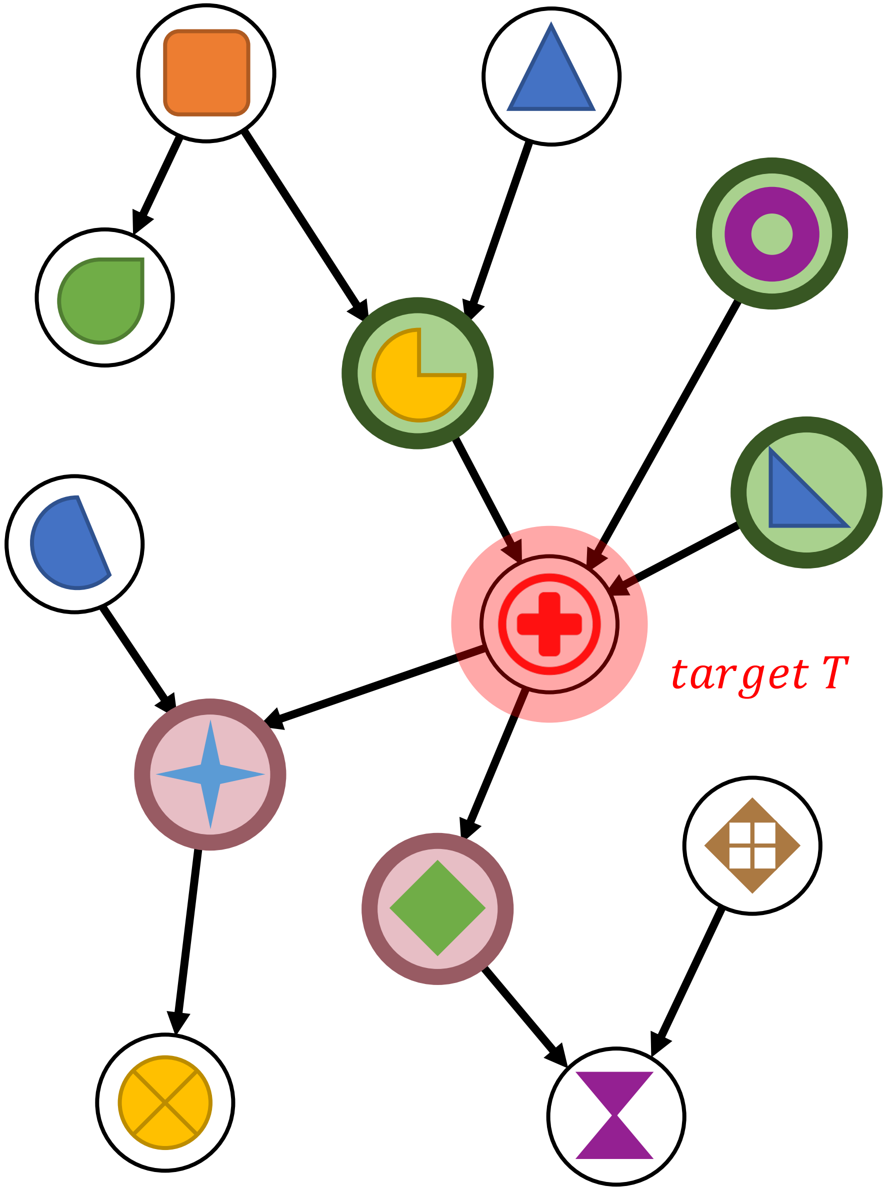
$Parents(T)$ } $MB(T)$



RAveL-MB at a Glance

Task: Discover a subset of $\boldsymbol{MB(T)}$ without returning any false positive

$\boldsymbol{Parents(T)}$
$\boldsymbol{Children(T)}$
} $\boldsymbol{MB(T)}$

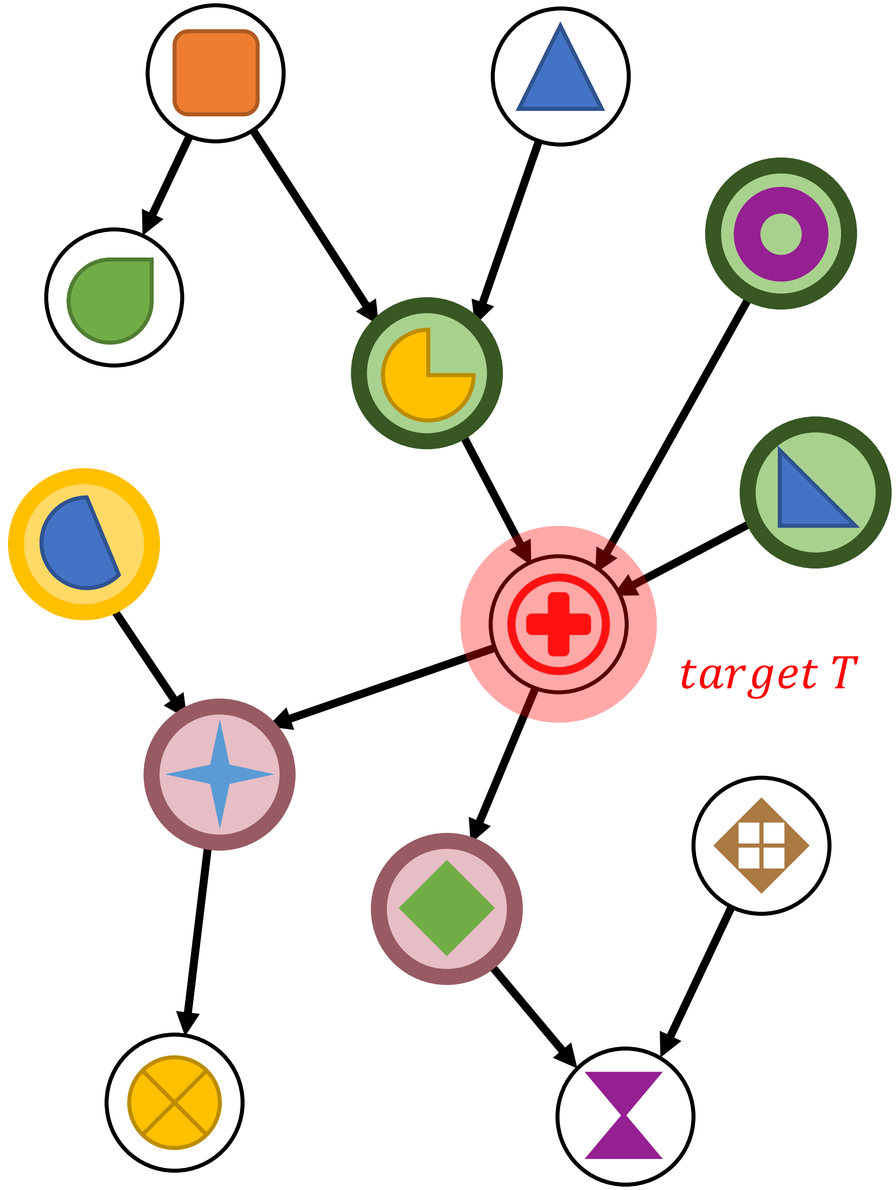


Università degli Studi di Padova

RAveL-MB at a Glance

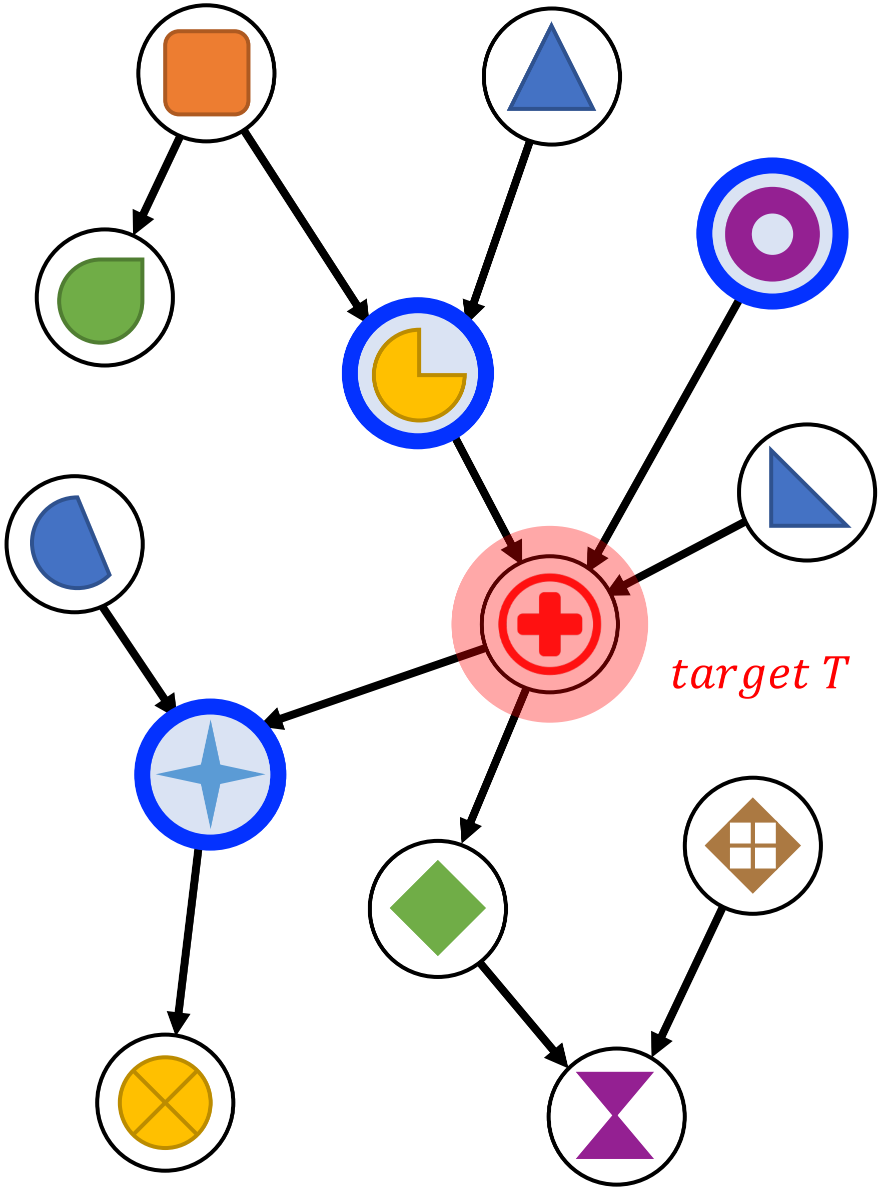
Task: Discover a subset of $\boldsymbol{MB}(\boldsymbol{T})$ without returning any false positive

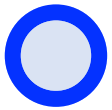
$\boldsymbol{Parents}(\boldsymbol{T})$, $\boldsymbol{Children}(\boldsymbol{T})$, $\boldsymbol{Spouses}(\boldsymbol{T})$ } $\boldsymbol{MB}(\boldsymbol{T})$



RAveL-MB at a Glance

1- Discover a subset $\boldsymbol{S}$ of elements in $PC(T)$

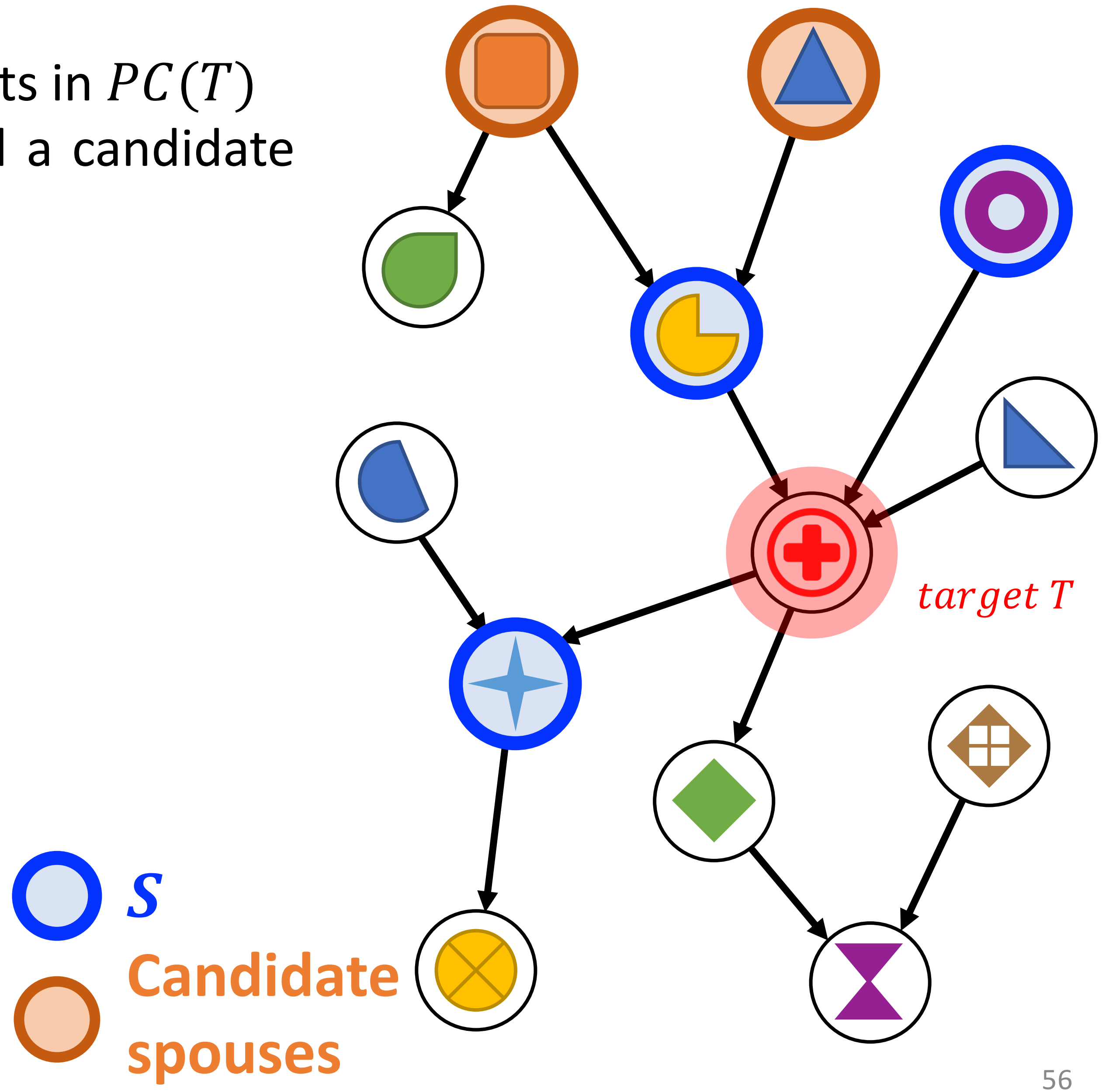




RAveL-MB at a Glance

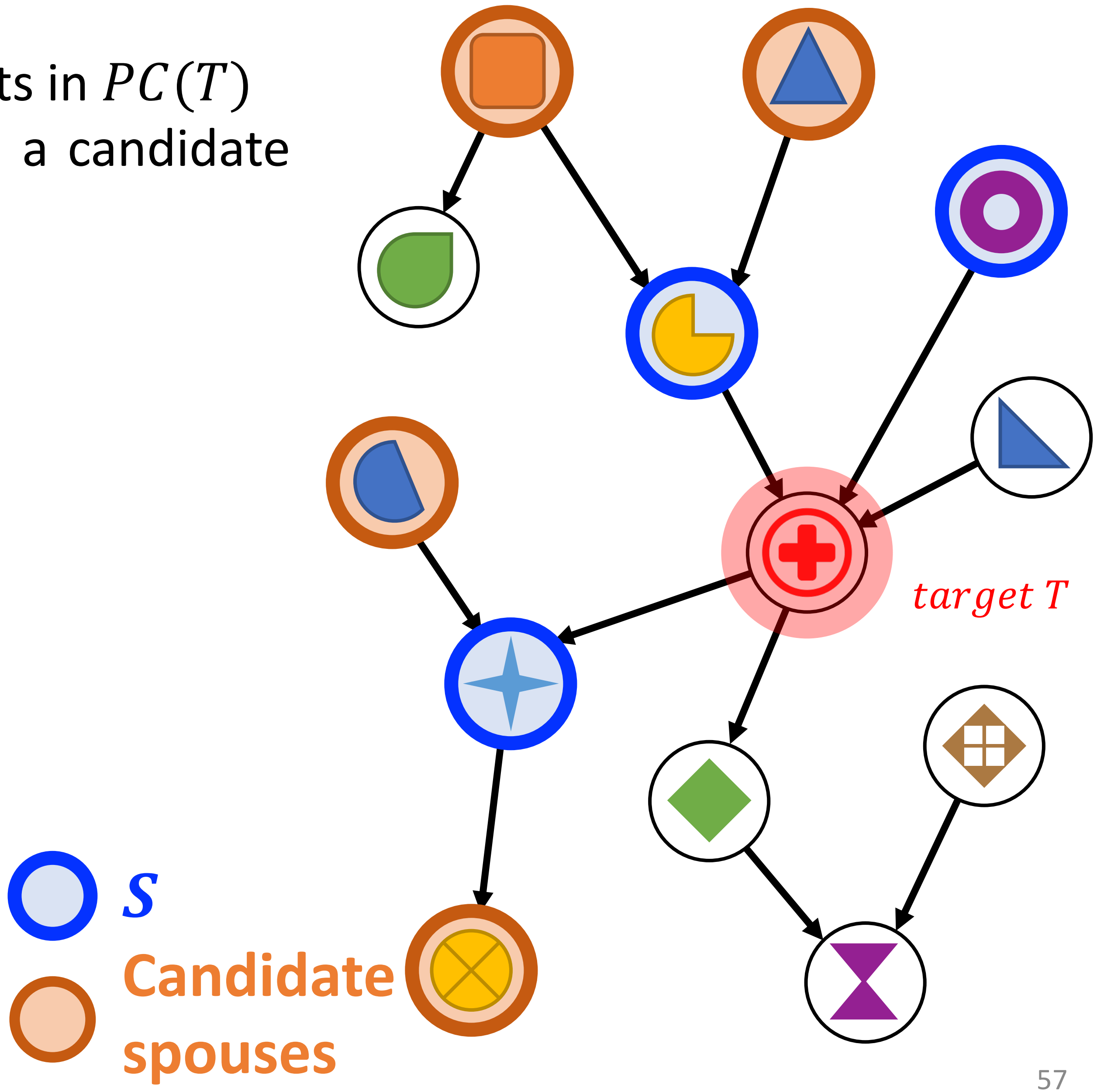
1- Discover a subset $\boldsymbol{S}$ of elements in $PC(T)$

2- For each element X in $\boldsymbol{S}$ find a candidate set of spouses in $PC(X)$



RAveL-MB at a Glance

1- Discover a subset $\boldsymbol{S}$ of elements in $PC(T)$
2- For each element X in $\boldsymbol{S}$ find a candidate set of spouses in $PC(X)$



RAveL-MB at a Glance

1- Discover a subset $\boldsymbol{S}$ of elements in $PC(T)$
2- For each element X in $\boldsymbol{S}$ find a candidate set of spouses in $PC(X)$
3- For each candidate spouse Y with children X, test the spouse condition

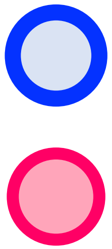
Previous approaches:

Dependence test — Independence test

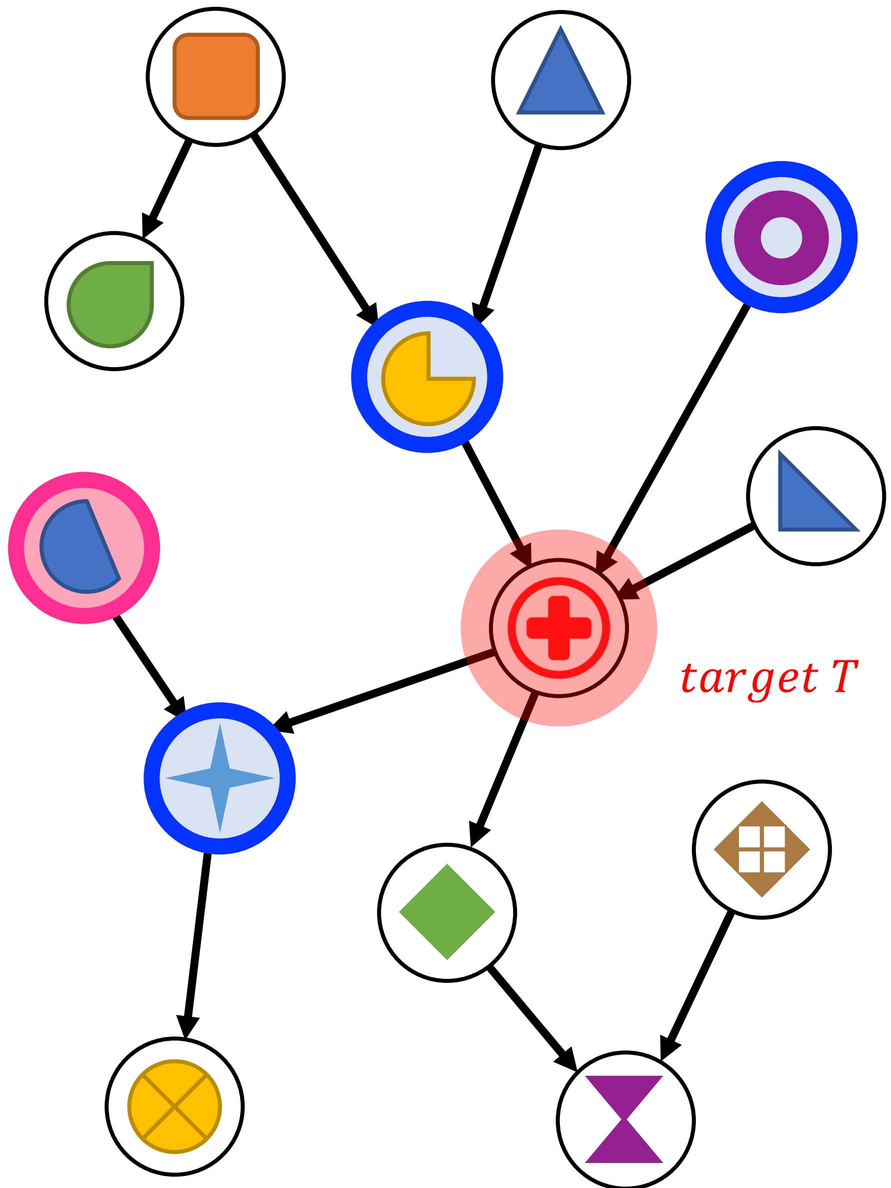
$$Y \perp\!\!\!\perp T|\mathbf{Z} \; and \; Y \not\perp\!\!\!\perp T|\mathbf{Z} \cup \{X\}$$
$$for \; all \; \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y, T\}$$

Our formulation (equivalent):

$$Y \not\perp\!\!\!\perp T|\mathbf{V} \setminus \{Y\}$$



$\boldsymbol{S}$
Actual spouses



RAveL-MB at a Glance

1- Discover a subset $\boldsymbol{S}$ of elements in $PC(T)$
2- For each element X in $\boldsymbol{S}$ find a candidate set of spouses in $PC(X)$
3- For each candidate spouse Y with children X, test the spouse condition

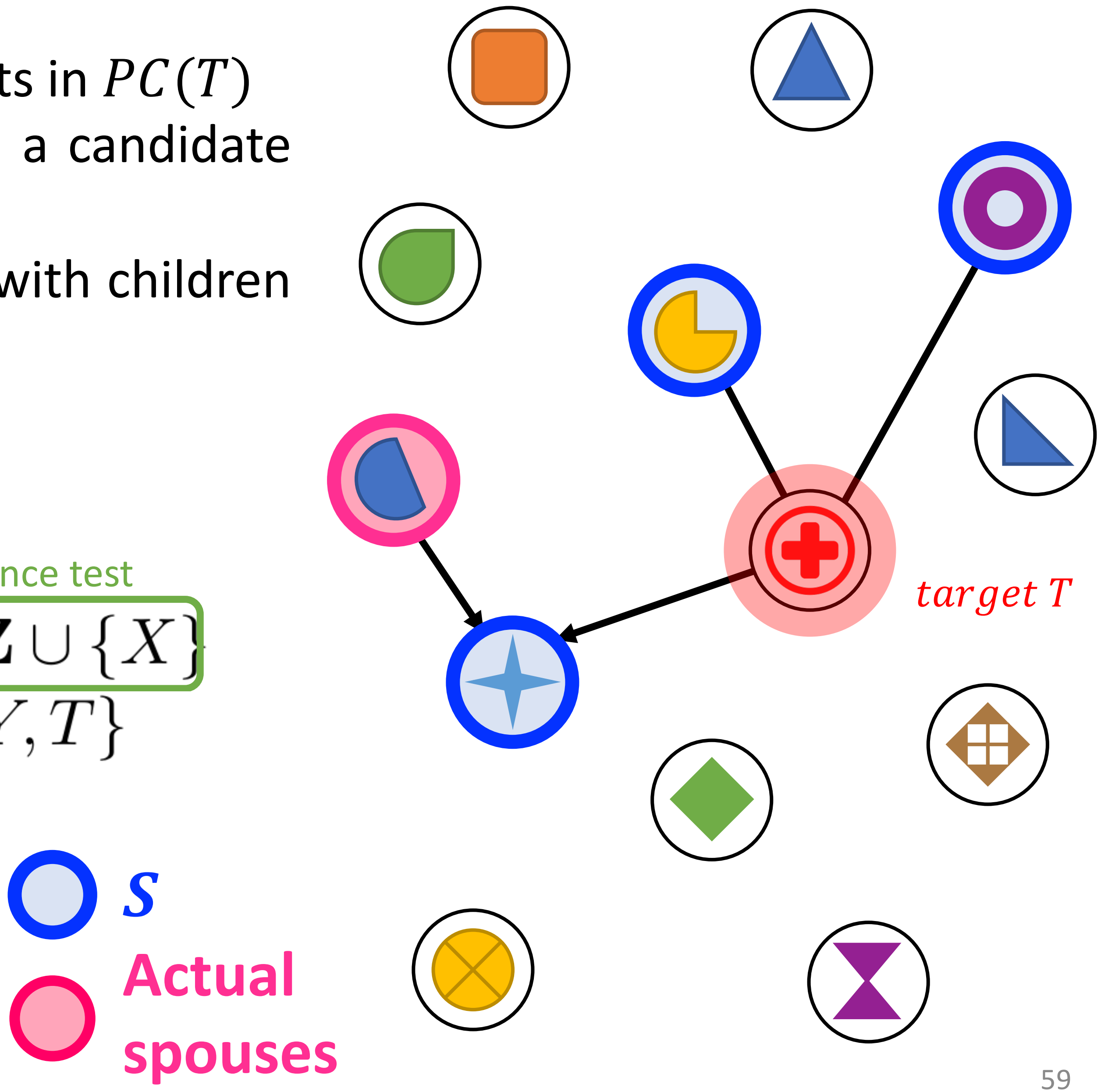
Previous approaches:

Dependence test — Independence test

$$Y \perp\!\!\!\perp T|\mathbf{Z} \; and \; Y \not\perp\!\!\!\perp T|\mathbf{Z} \cup \{X\}$$
$$for \; all \; \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y, T\}$$

Our formulation (equivalent):

$$Y \not\perp\!\!\!\perp T|\mathbf{V} \setminus \{Y\}$$



Multiple Hypothesis Testing Corrections

- **Bonferroni correction** on the threshold δ

  Classical correction for multiple hypotheses testing uses a modified threshold δ/N on each test

  May be too strict if N is big (and in our case N is exponential on $|\boldsymbol{V}|$)

- **Rademacher averages** to bound each test statistic

  Provide data-dependent bounds

Rademacher Averages - Idea

Given a family of functions $\mathcal{F}$ and a dataset S, Rademacher averages upper bound with high probability the Supremum Deviation $D(\mathcal{F}, S)$

Empirical sample mean

Rademacher estimate

$$D(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} \left| \hat{E}_S[f] - E[f] \right| \leq 2\tilde{R} + O\left(\frac{1}{\sqrt{m}}\right)$$

Expectation

Dataset size

Rademacher Averages - Idea

Given a family of functions $\mathcal{F}$ and a dataset S, Rademacher averages upper bound with high probability the Supremum Deviation $D(\mathcal{F}, S)$

Empirical sample mean

Rademacher estimate

$$D(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} \left| \hat{E}_S[f] - E[f] \right| \leq 2\tilde{R} + O\left(\frac{1}{\sqrt{m}}\right)$$

Expectation

Dataset size

They can lower bound **simultaneously** each independence test statistic for providing guarantees on the $FWER$

Rademacher Averages - Idea

Given two normalized vectors of observations $\boldsymbol{x}, \boldsymbol{y}$

$$Pearson's\ r\ coefficient \qquad r_{x,y} = \frac{\sum_{i=0}^{m-1} x_i y_i}{(m-1)}$$

Rademacher Averages - Idea

Given two normalized vectors of observations $\boldsymbol{x}, \boldsymbol{y}$

$$Pearson's\, r\, coefficient \qquad r_{x,y} = \frac{\sum_{i=0}^{m-1} x_i y_i}{(m-1)}$$

By defining

$$r_{x,y}(s_i) = \frac{m}{(m-1)} x_i y_i$$

We get

$$r_{x,y} = \hat{\mathbb{E}}_S[r_{x,y}(s_i)]$$

Rademacher Averages - Idea

Given two normalized vectors of observations $\boldsymbol{x}, \boldsymbol{y}$

$$Pearson's\ r\ coefficient \qquad r_{x,y} = \frac{\sum_{i=0}^{m-1} x_i y_i}{(m-1)}$$

By defining

$$\mathcal{F} \ni r_{x,y}(s_i) = \frac{m}{(m-1)} x_i y_i$$

We get

$$r_{x,y} = \hat{\mathbb{E}}_S[r_{x,y}(s_i)]$$

Rademacher Averages - Idea

Given two normalized vectors of observations $\boldsymbol{x}, \boldsymbol{y}$

$$Pearson's\ r\ coefficient \qquad r_{x,y} = \frac{\sum_{i=0}^{m-1} x_i y_i}{(m-1)}$$

By defining

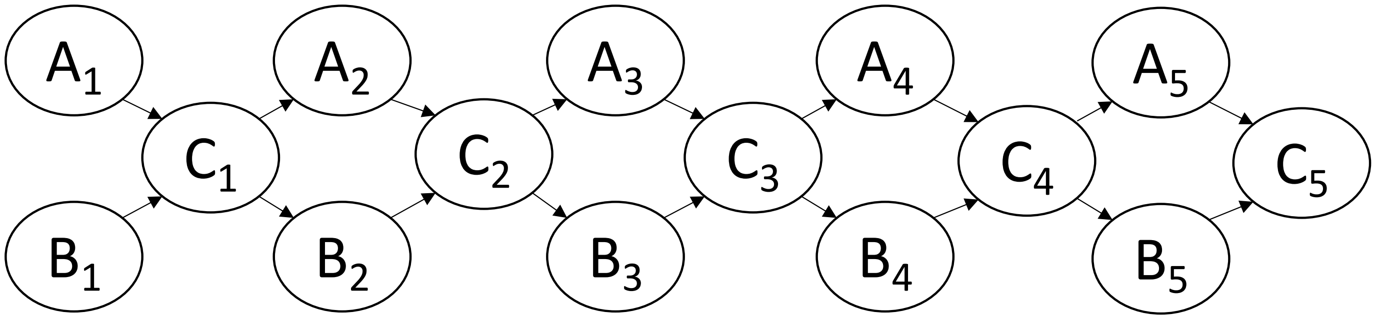
$$\mathcal{F} \ni r_{x,y}(s_i) = \frac{m}{(m-1)} x_i y_i$$

We get

$$r_{x,y} = \hat{\mathbb{E}}_S[r_{x,y}(s_i)]$$

The same approach works with other test statistics

Experimental Evaluation



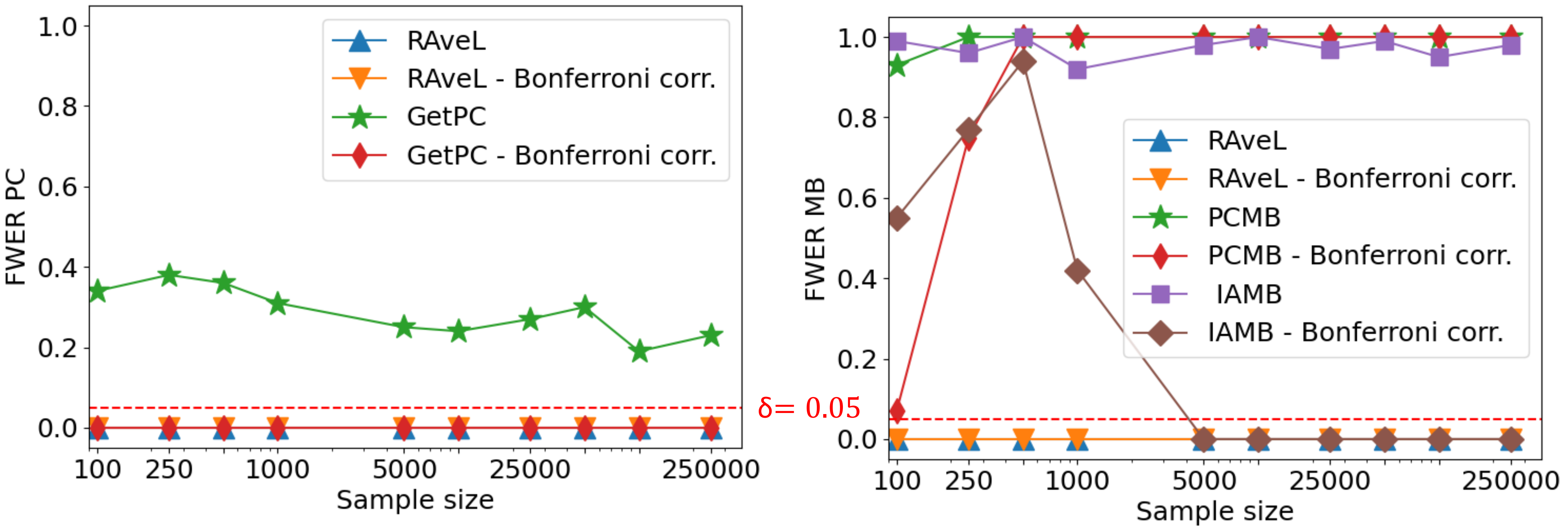
$$A_1 \sim N(0,1)$$
$$B_1 \sim N(0,1)$$
$$C_1 \sim -3 + 2A_1 + 3B_1 + N(0,1)$$
$$\forall i \in [2,3,4,5]$$
$$A_i \sim +3 + 3C_{i-1} + N(0,1)$$
$$B_i \sim -1 - 2C_{i-1} + N(0,1)$$
$$C_i \sim -3 + 2A_i + 3B_i + N(0,1)$$

- On each run, test the SoA and RAveL algorithms on every variable

- Analyse results of each iteration:
  - No False Positives only if **all the outputs** (one per variable) **of the run** do not contain false positives
  - Count mean percentage of false negatives

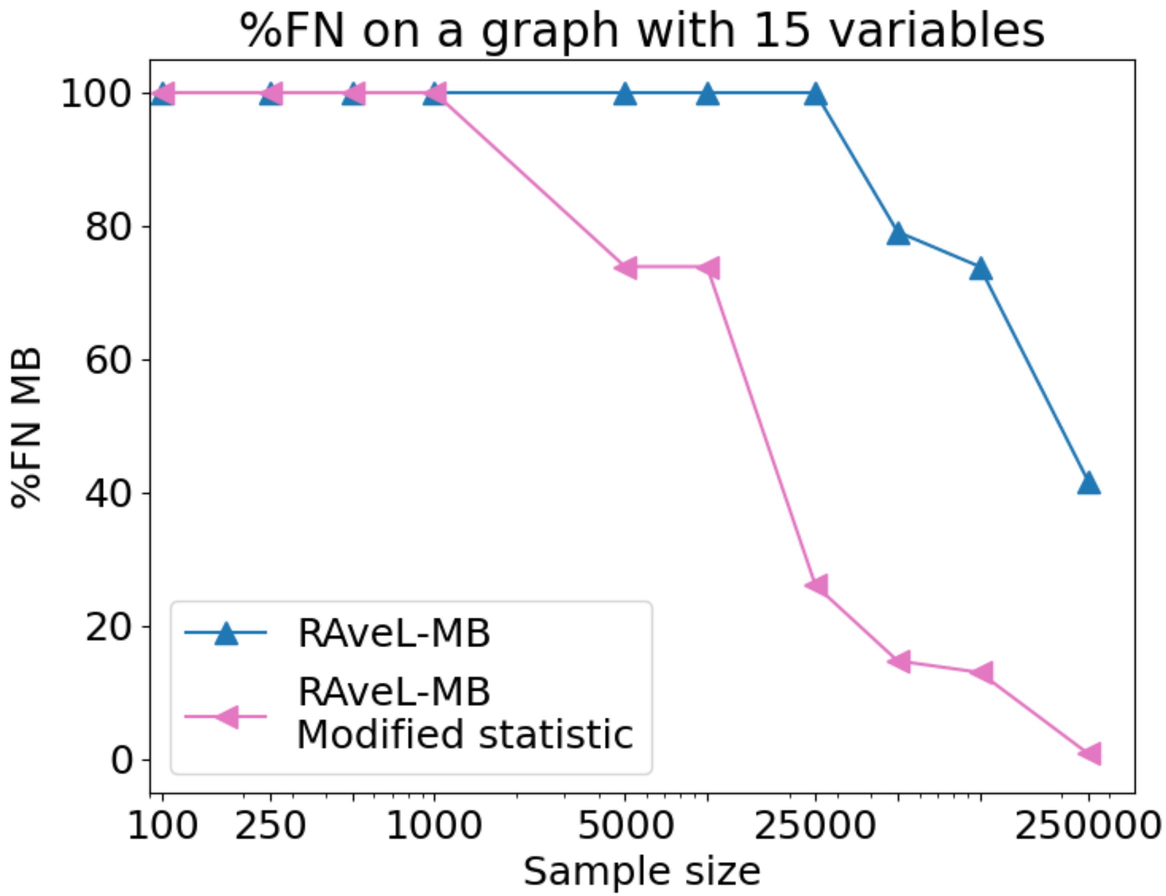
Experimental Evaluation

RAveL-PC and RAveL-MB effectively control the $FWER$



Experimental Evaluation

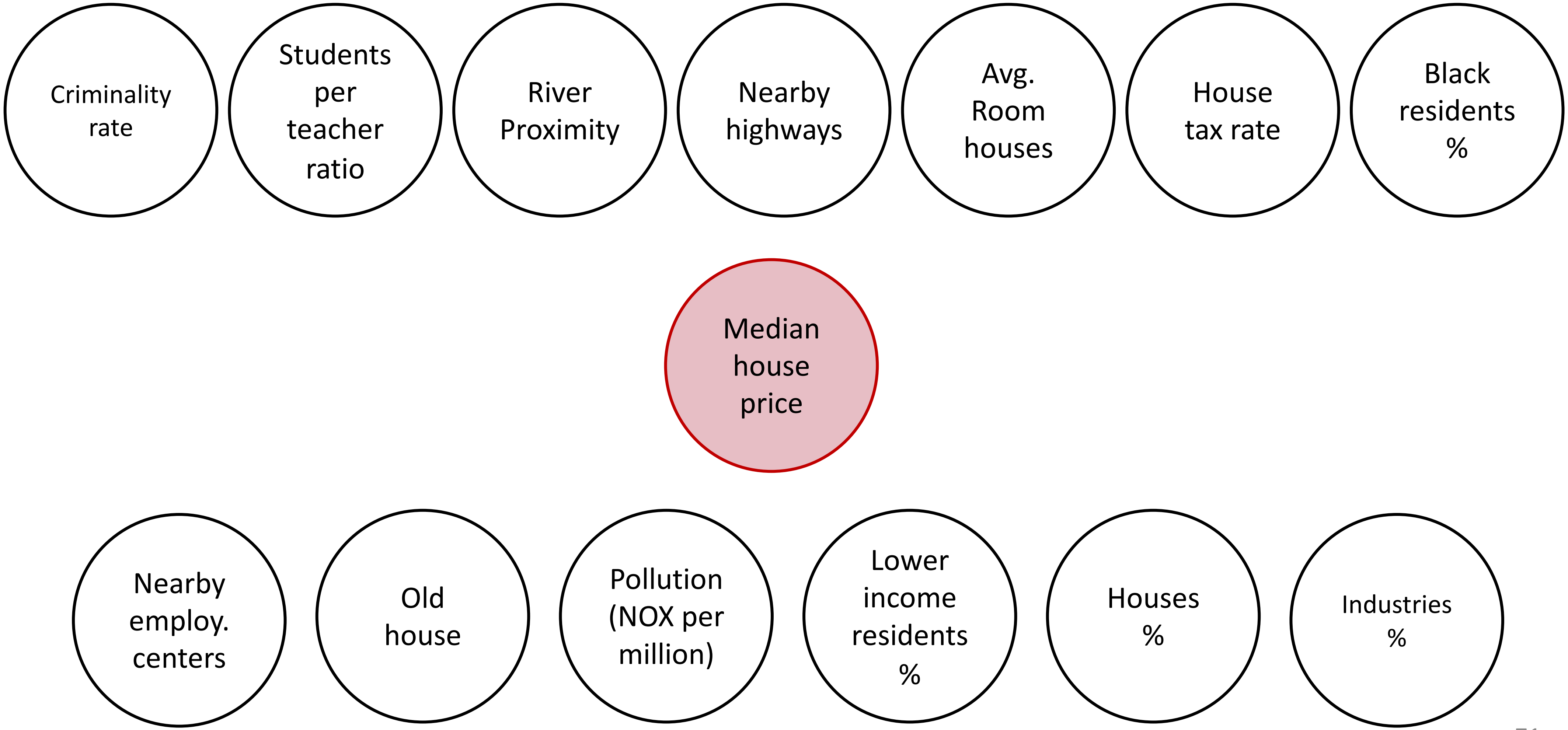
Tests with increasing number of variables



Test statistic choice heavily affects Rademacher based corrections

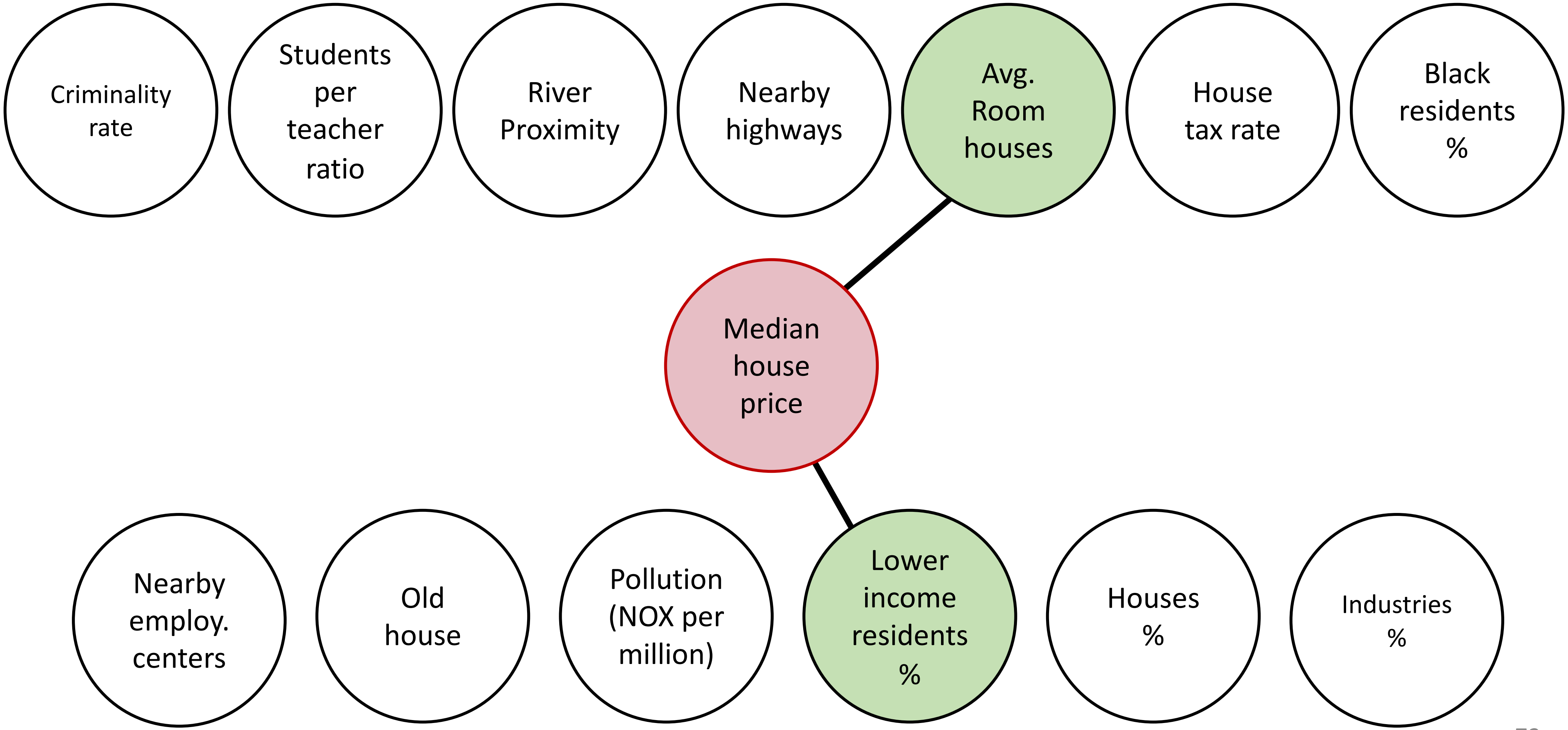
Experimental Evaluation

Tested RAveL-PC and RAveL-MB on **Boston housing** dataset with $\delta = 0{,}01$.



Experimental Evaluation

Tested RAveL-PC and RAveL-MB on **Boston housing** dataset with $\delta = 0{,}01$.



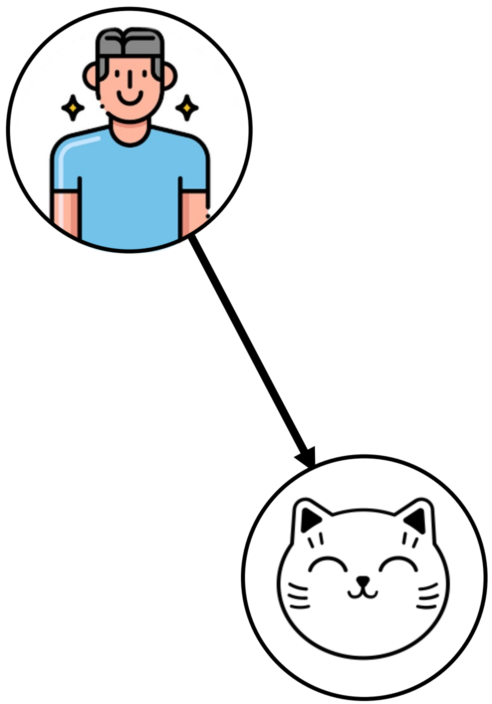
My focus

Develop algorithms to:

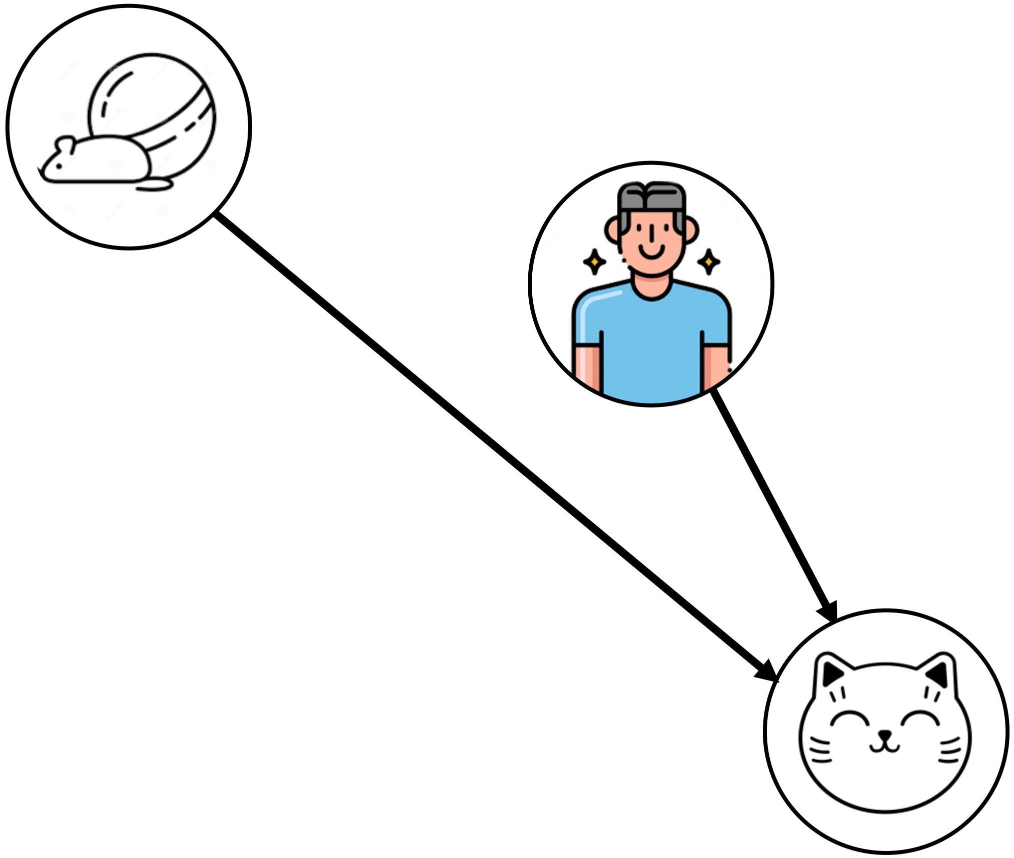
- (Structure discovery) **Discover causally related variables to a target**
- (Effect estimation) Evaluate effect of causal rules

From observational data and providing guarantees on the results

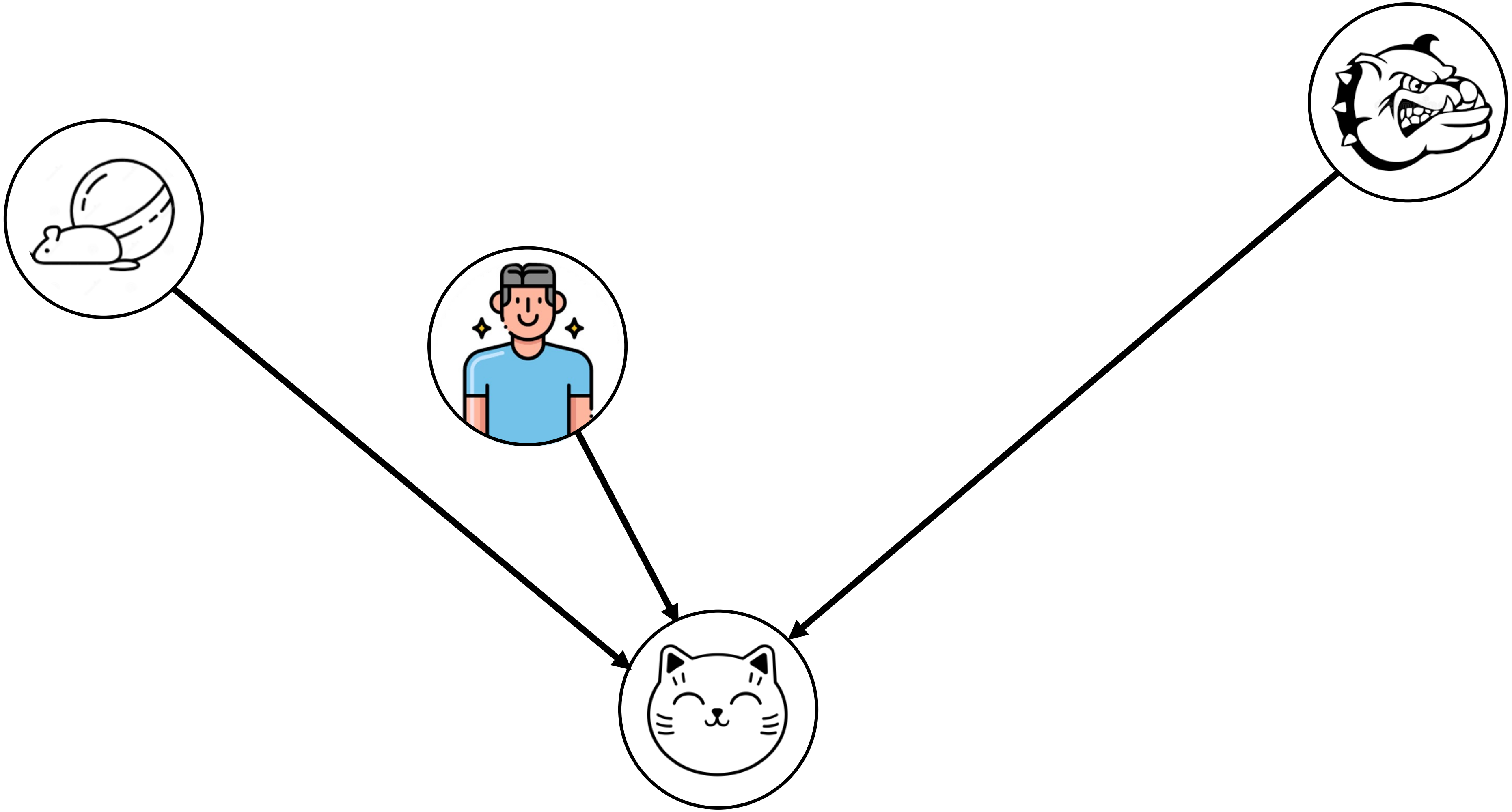
Very serious research question



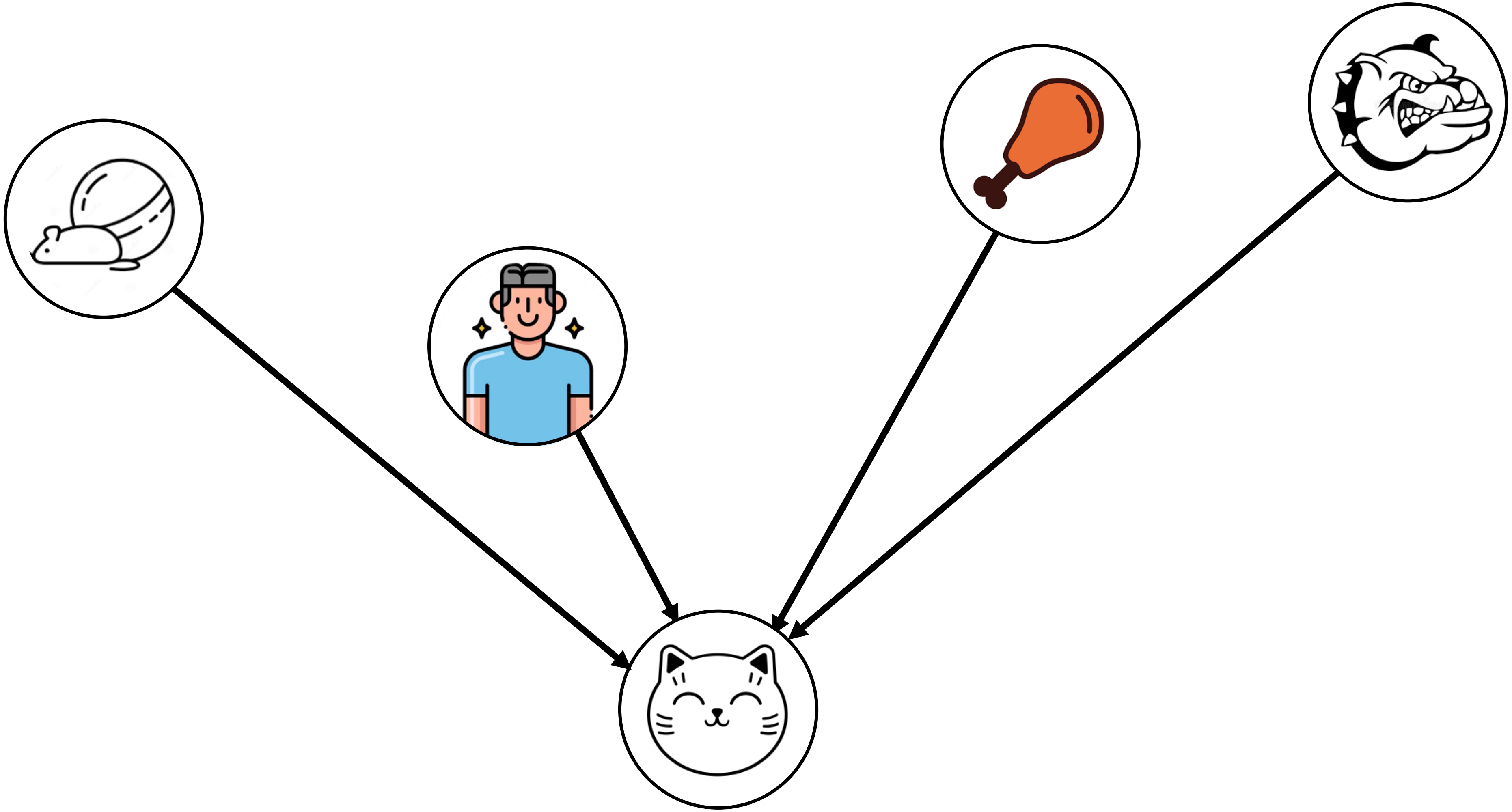
Very serious research question



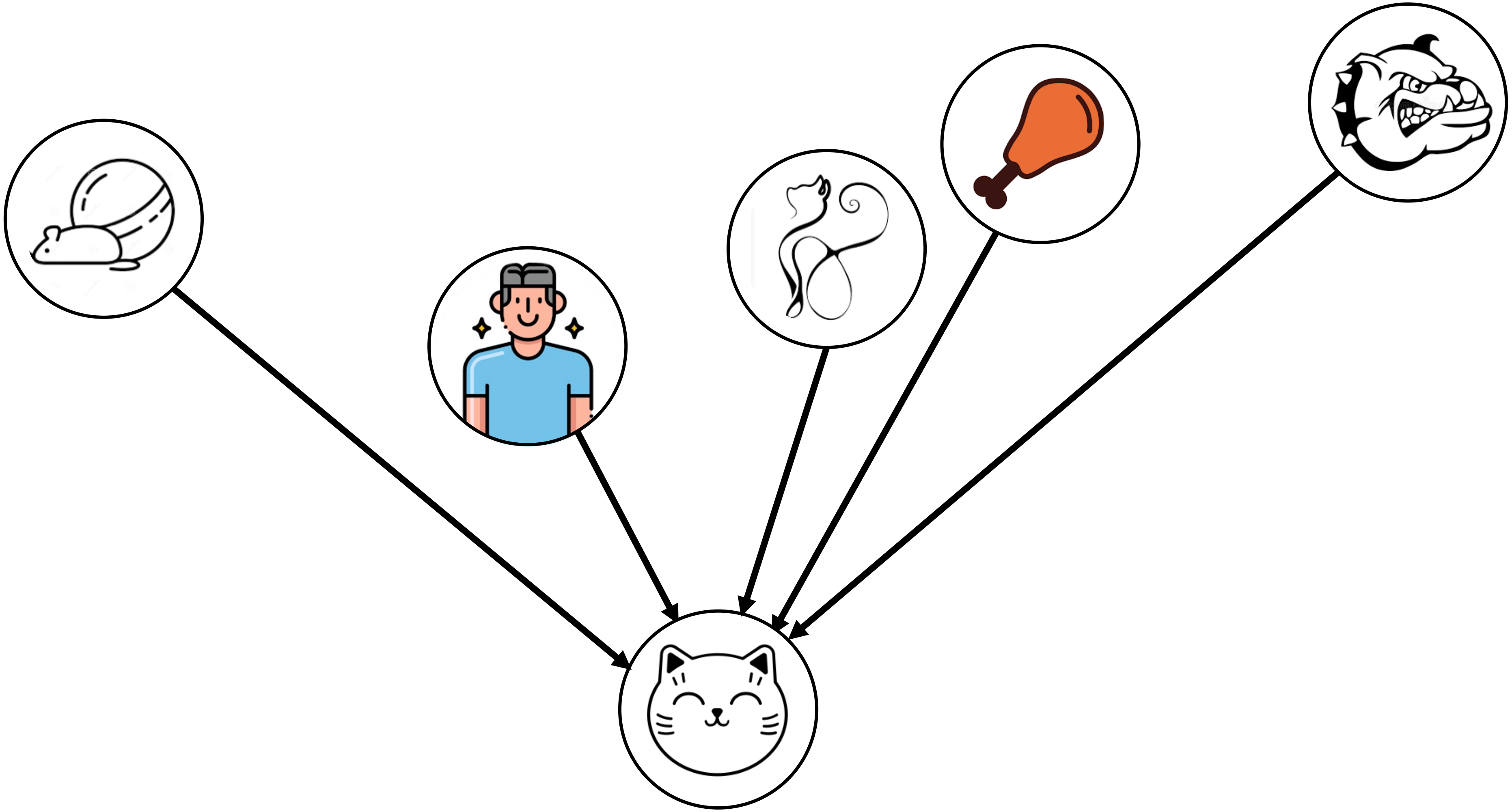
Very serious research question



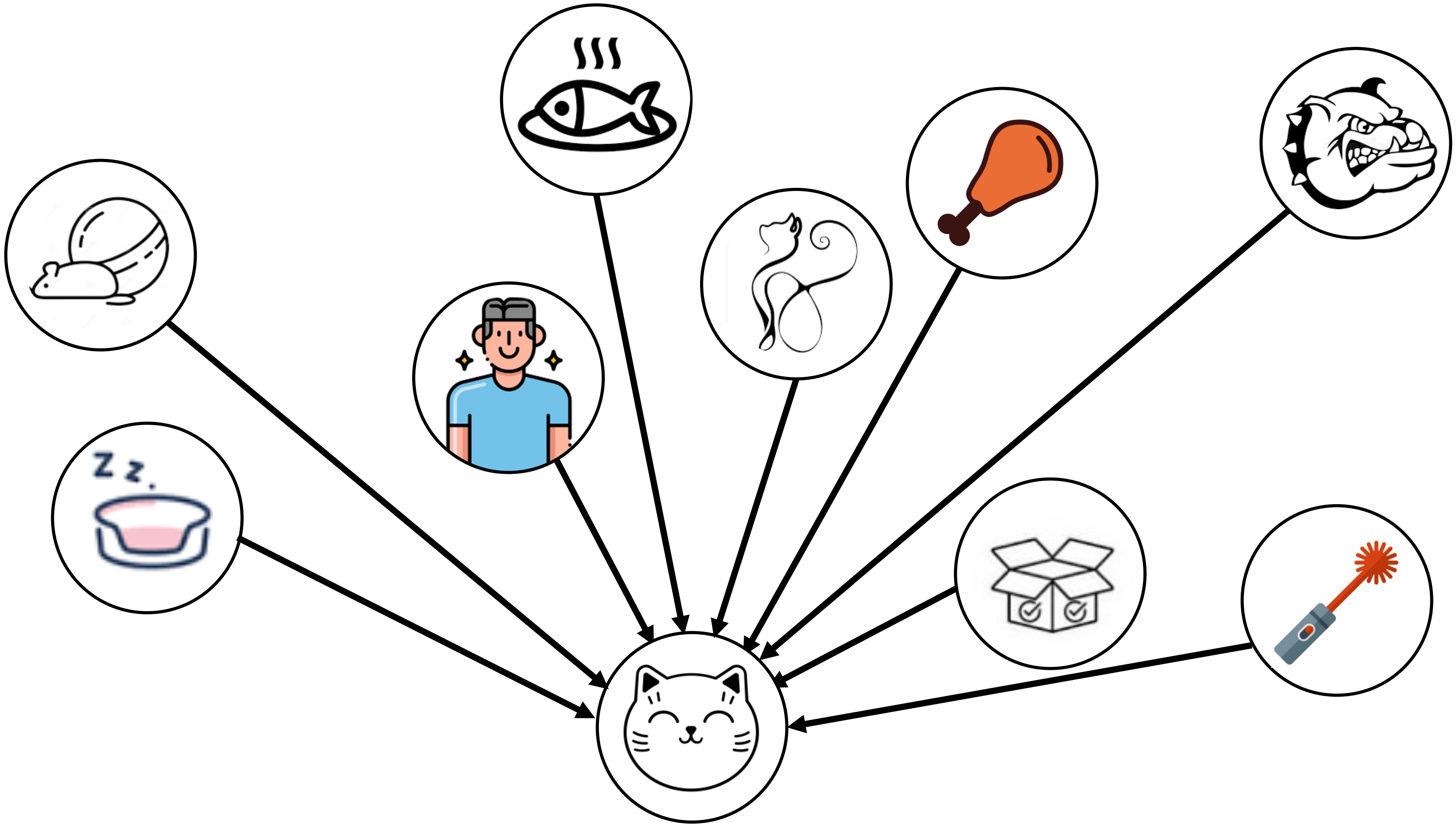
Very serious research question



Very serious research question



Very serious research question



My focus

Develop algorithms to:

- (Structure discovery) Discover causally related variables to a target
- **(Effect estimation) Evaluate effect of causal rules**

Submitted paper at ECCB 2023 and work in progress in collaboration with:

Antonio Collesei, PhD s. (IOV)

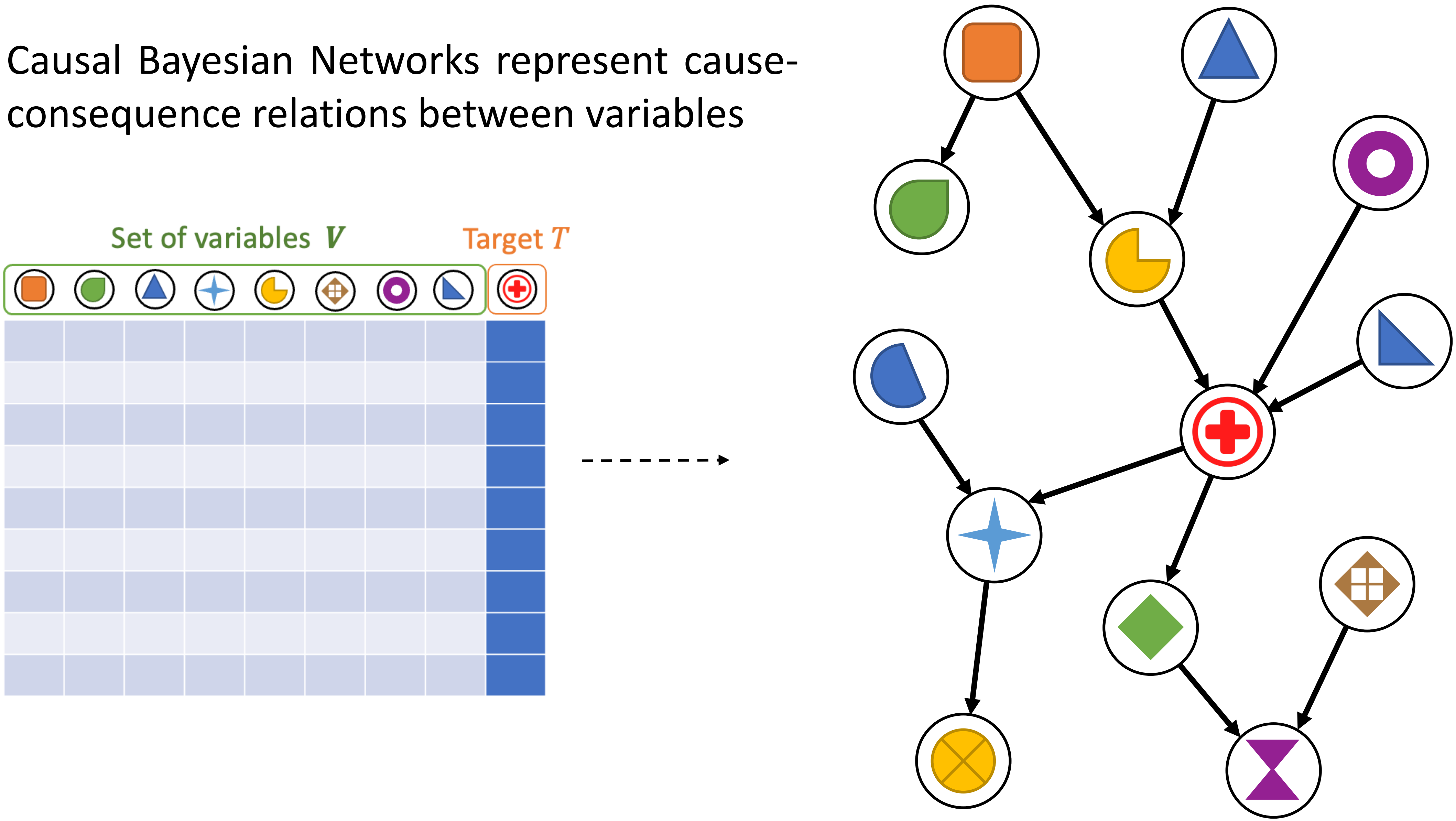
Paola Donolato, MS

Federica Miglietta, MD and PhD s. (IOV)

Fabio Vandin, Full Professor

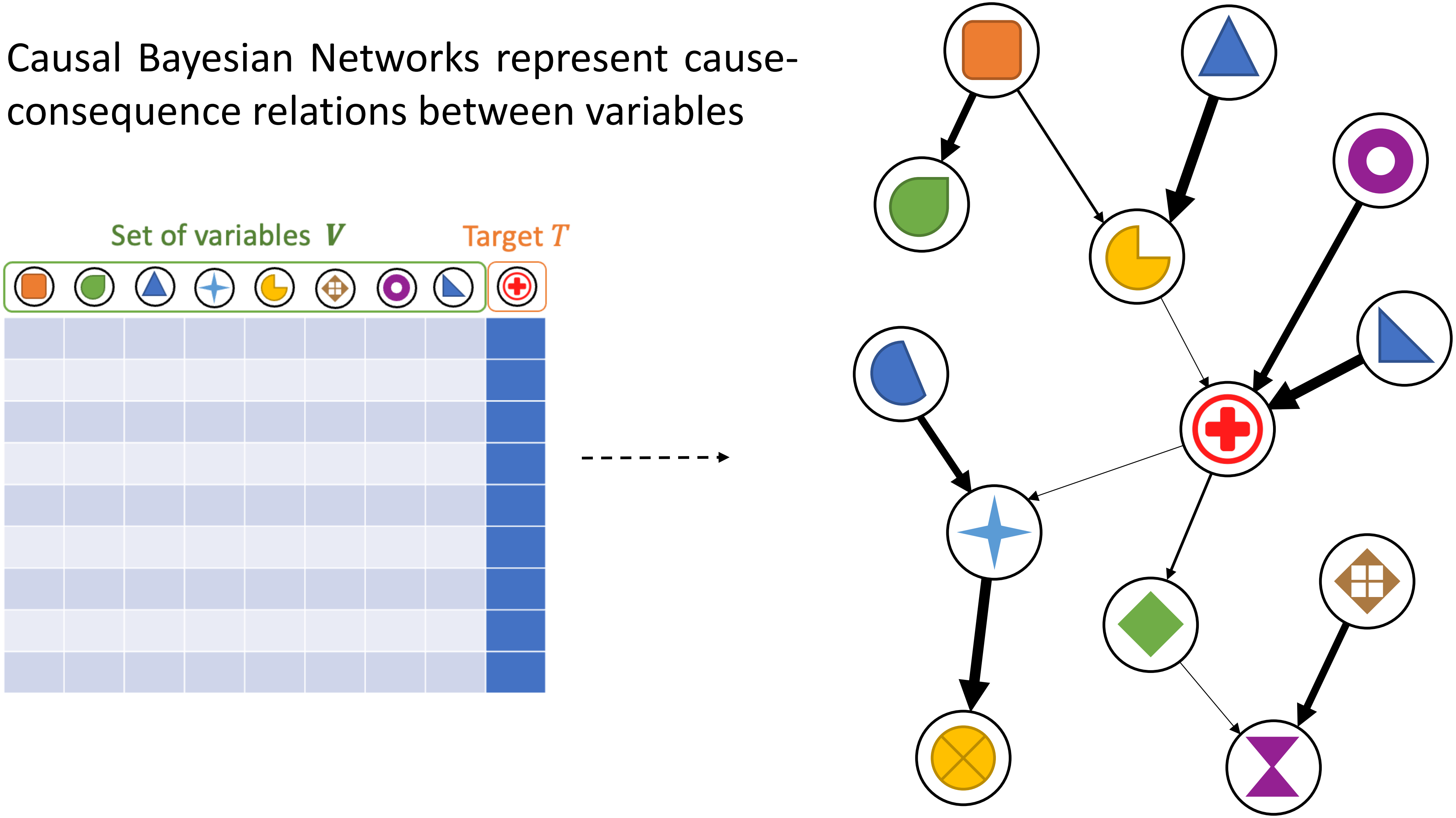
Problem Setup

Causal Bayesian Networks represent cause-consequence relations between variables



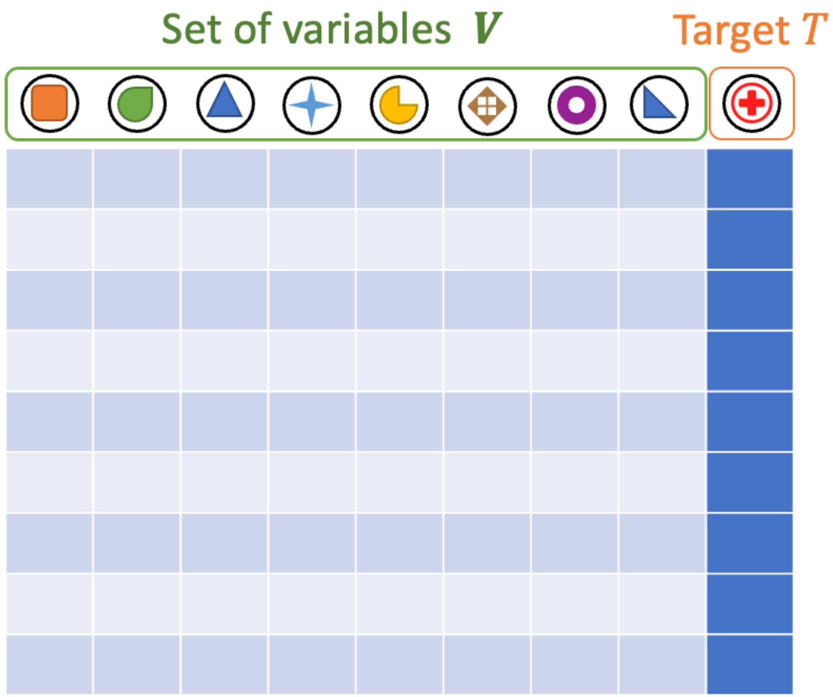
Problem Setup

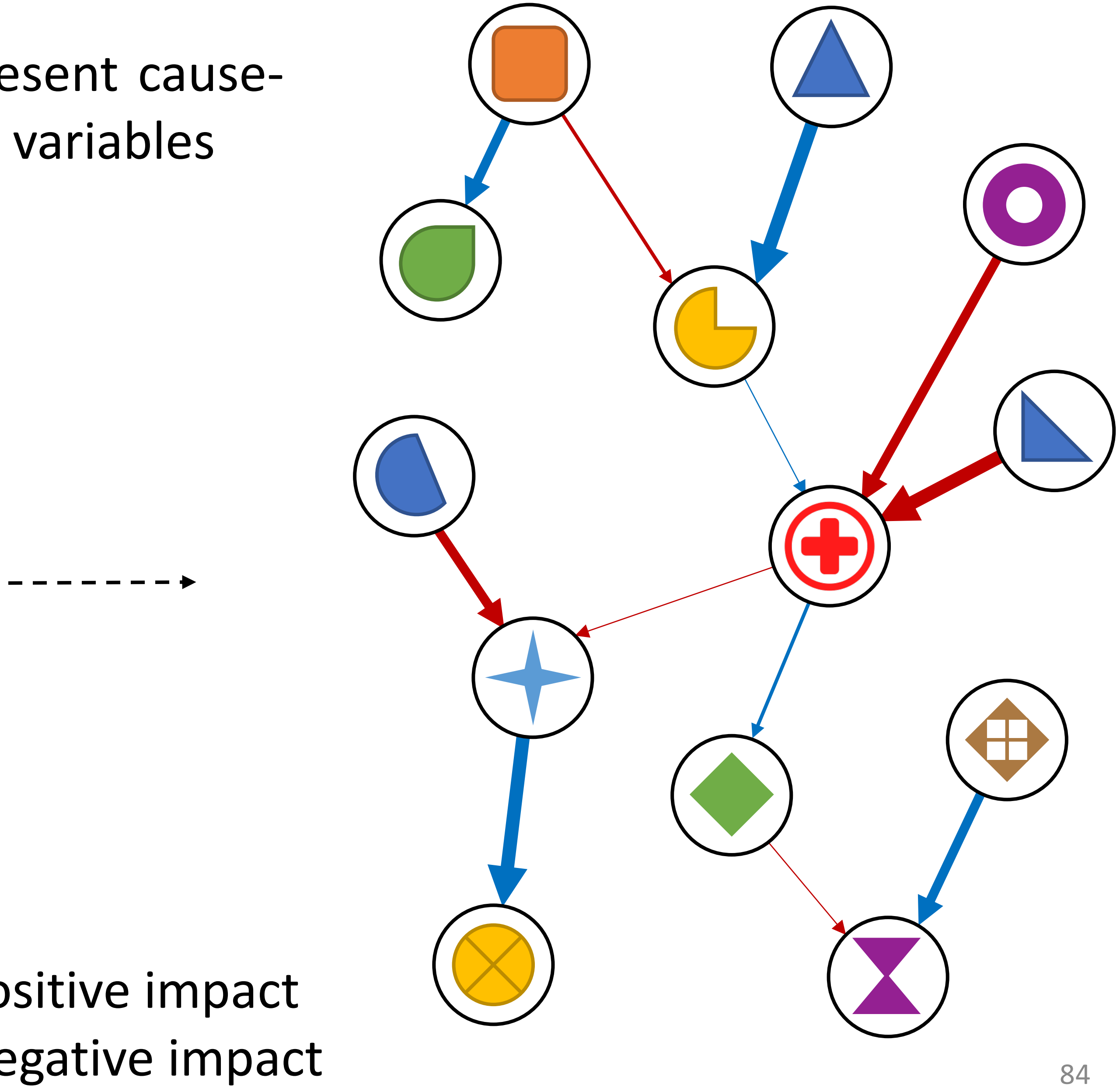
Causal Bayesian Networks represent cause-consequence relations between variables



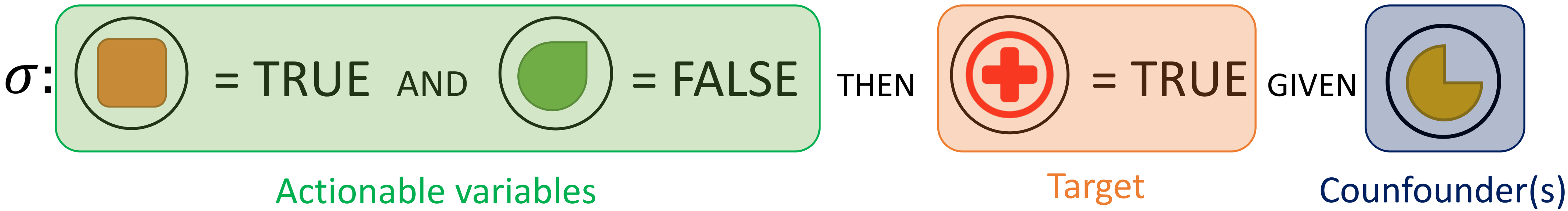
Problem Setup

Causal Bayesian Networks represent cause-consequence relations between variables

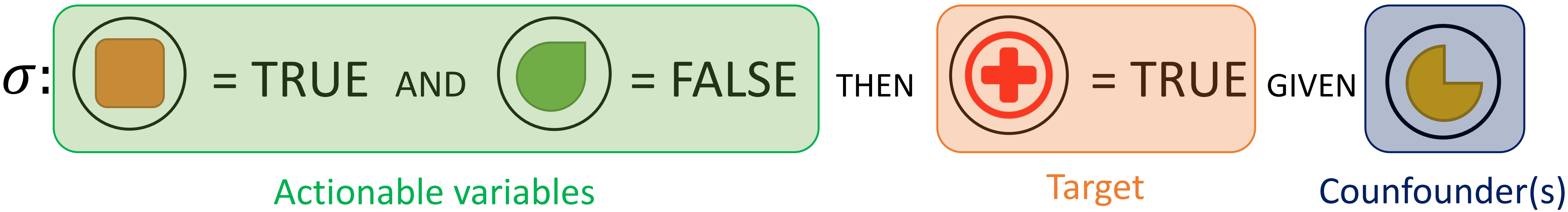




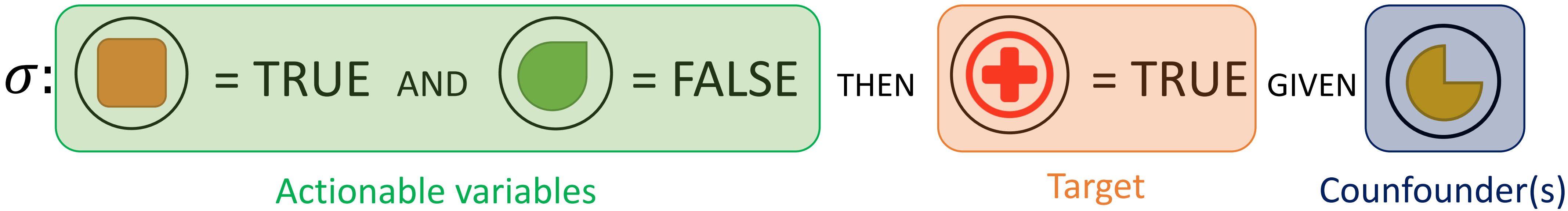
Problem Definition



Problem Definition



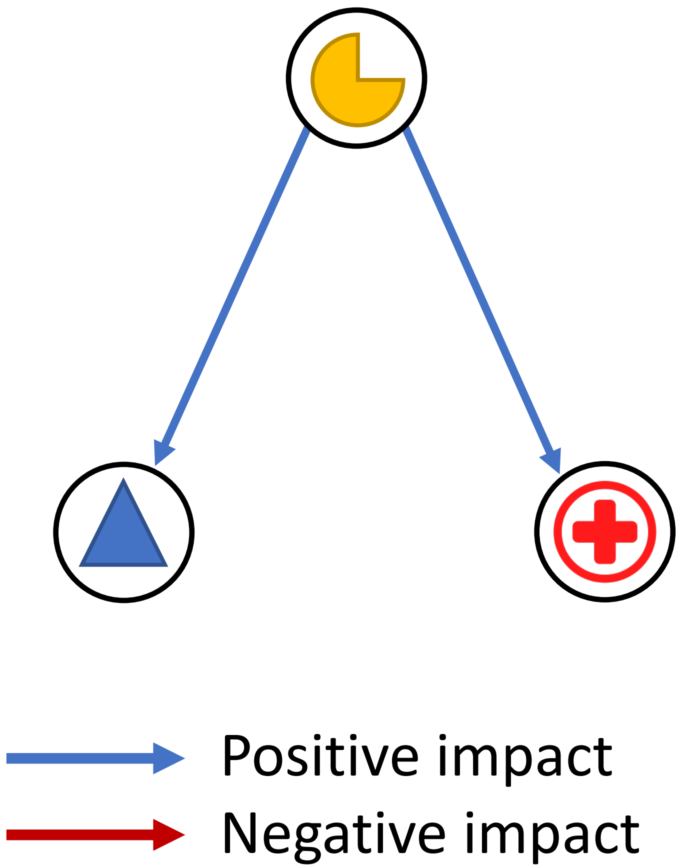
Problem Definition



Task: Given a dataset of observations of variables $\boldsymbol{V}$, **find** the top-k rules $\sigma_1^*, \ldots, \sigma_k^*$ with **the highest causal effect** on T with **guarantees** on the result (e.g. no false positives)

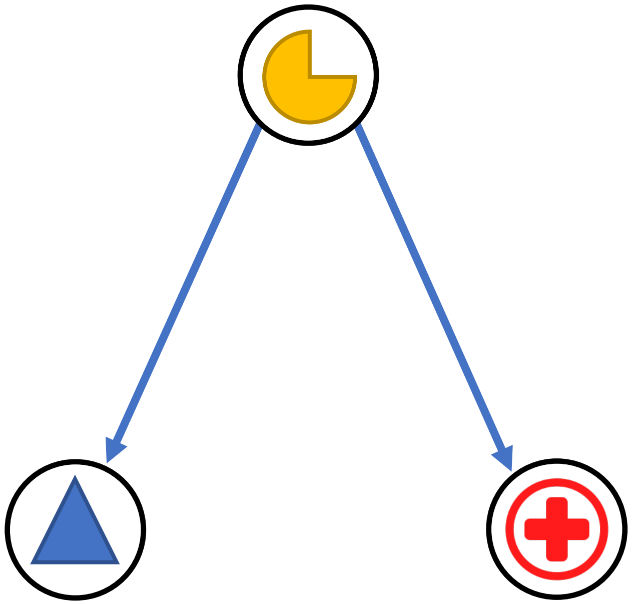
Example

We have to deal with confounder variables

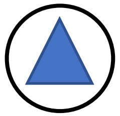


Example

We have to deal with confounder variables

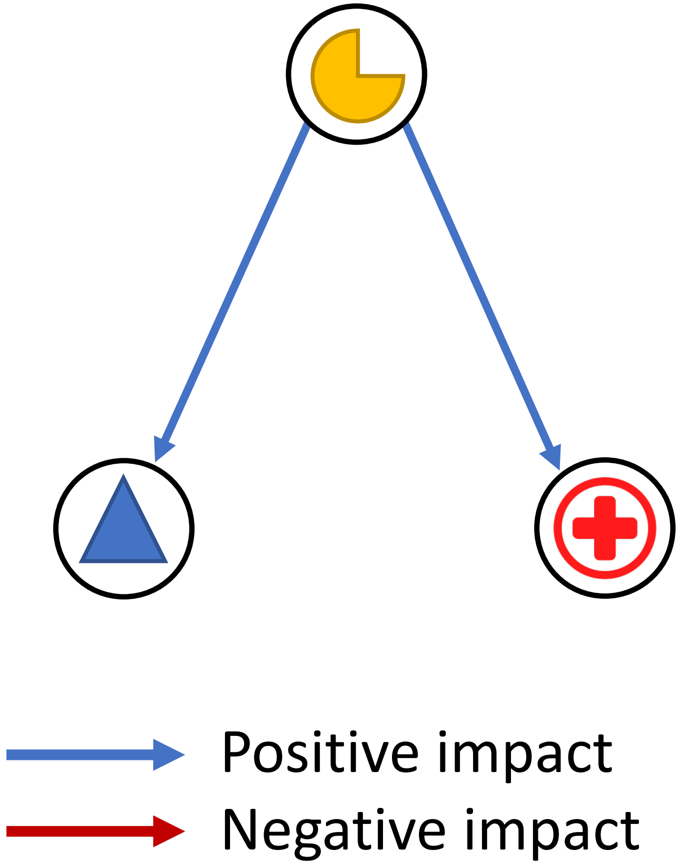


Positive impact
Negative impact

		
0.5	0.7	1.5
1.4	1.8	2.0
2.7	3.0	3.7
2.3	2.9	3.2
1.8	1.9	2.3
1.6	1.9	2.2
0.6	0.9	1.4
0.3	0.5	1.0
1.1	1.3	1.4

Example

We have to deal with confounder variables



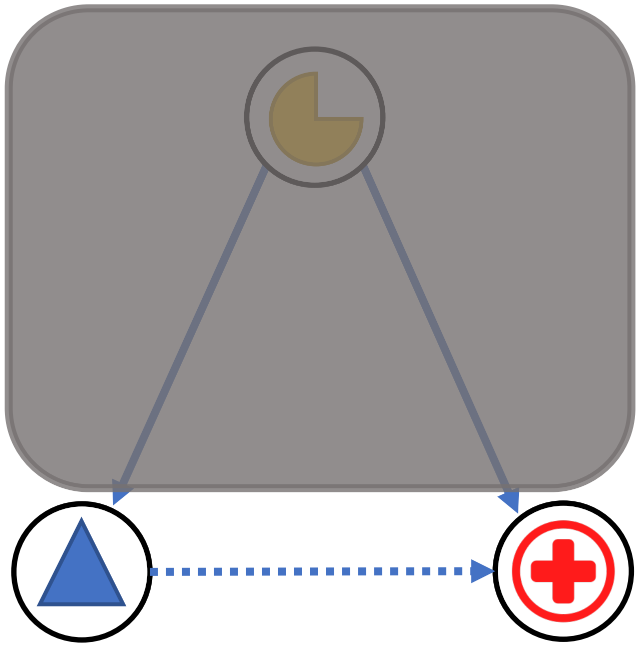
0.5	0.7	1.5
1.4	1.8	2.0
2.7	3.0	3.7
2.3	2.9	3.2
1.8	1.9	2.3
1.6	1.9	2.2
0.6	0.9	1.4
0.3	0.5	1.0
1.1	1.3	1.4

Our focus

and are correlated

Example

We have to deal with confounder variables



Positive impact

Negative impact

0.5	0.7	1.5
1.4	1.8	2.0
2.7	3.0	3.7
2.3	2.9	3.2
1.8	1.9	2.3
1.6	1.9	2.2
0.6	0.9	1.4
0.3	0.5	1.0
1.1	1.3	1.4

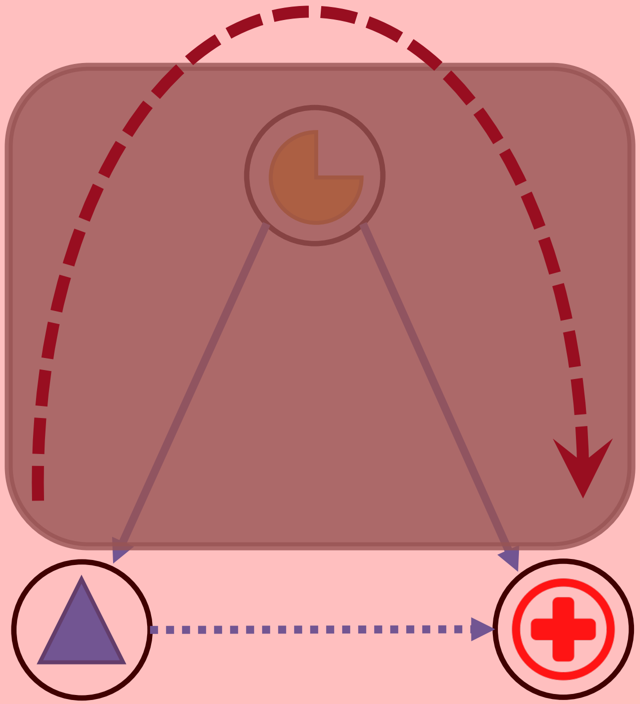
Our focus

and

are correlated

Example

We have to deal with confounder variables



Positive impact

Negative impact

0.5	0.7	1.5
1.4	1.8	2.0
2.7	3.0	3.7
2.3	2.9	3.2
1.8	1.9	2.3
1.6	1.9	2.2
0.6	0.9	1.4
0.3	0.5	1.0
1.1	1.3	1.4

Our focus

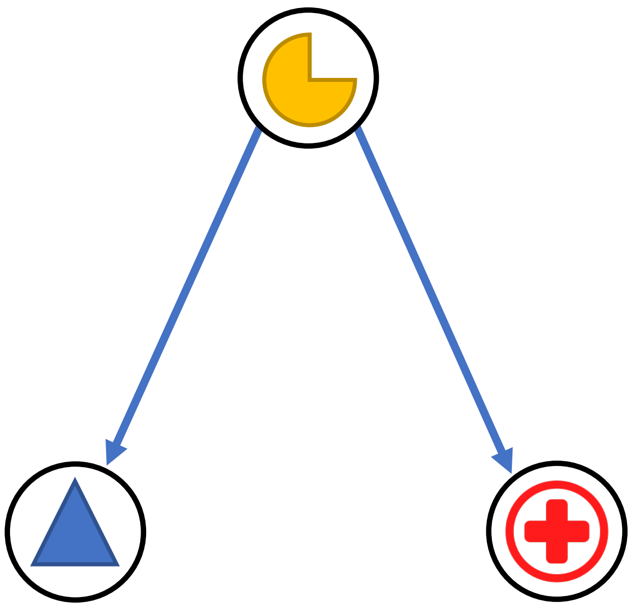
and

are correlated

CORRELATION IS NOT CAUSATION

Example

We have to deal with confounder variables



Positive impact
Negative impact

0.5	0.7	1.5
1.4	1.8	2.0
2.7	3.0	3.7
2.3	2.9	3.2
1.8	1.9	2.3
1.6	1.9	2.2
0.6	0.9	1.4
0.3	0.5	1.0
1.1	1.3	1.4

Our focus

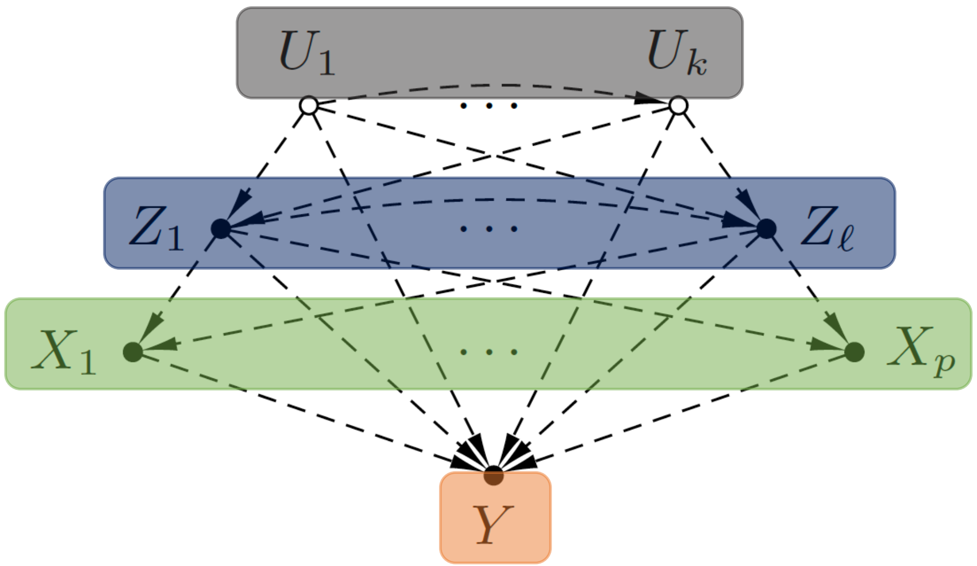
and are correlated

and are **NOT** correlated by conditioning on

Real-world data example

We are currently working with Breast Cancer Data

Set of confounders $\boldsymbol{Z}$ Set of actionables $\boldsymbol{X}$ Target T



Real-world data example

We are currently working with Breast Cancer Data

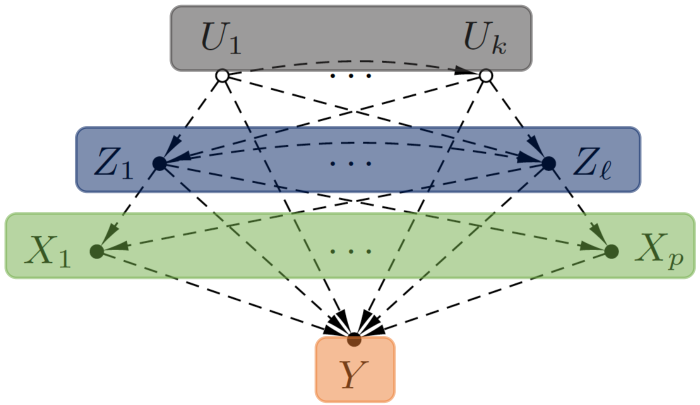
Set of confounders $\boldsymbol{Z}$ | Set of actionables $\boldsymbol{X}$ | Target T

Germline genomic alterations | Somatic genomic alterations | Cancer type

Sex

Age

…



Real-world data example

We are currently working with Breast Cancer Data

Set of confounders ***Z*** Set of actionables ***X*** Target *T*

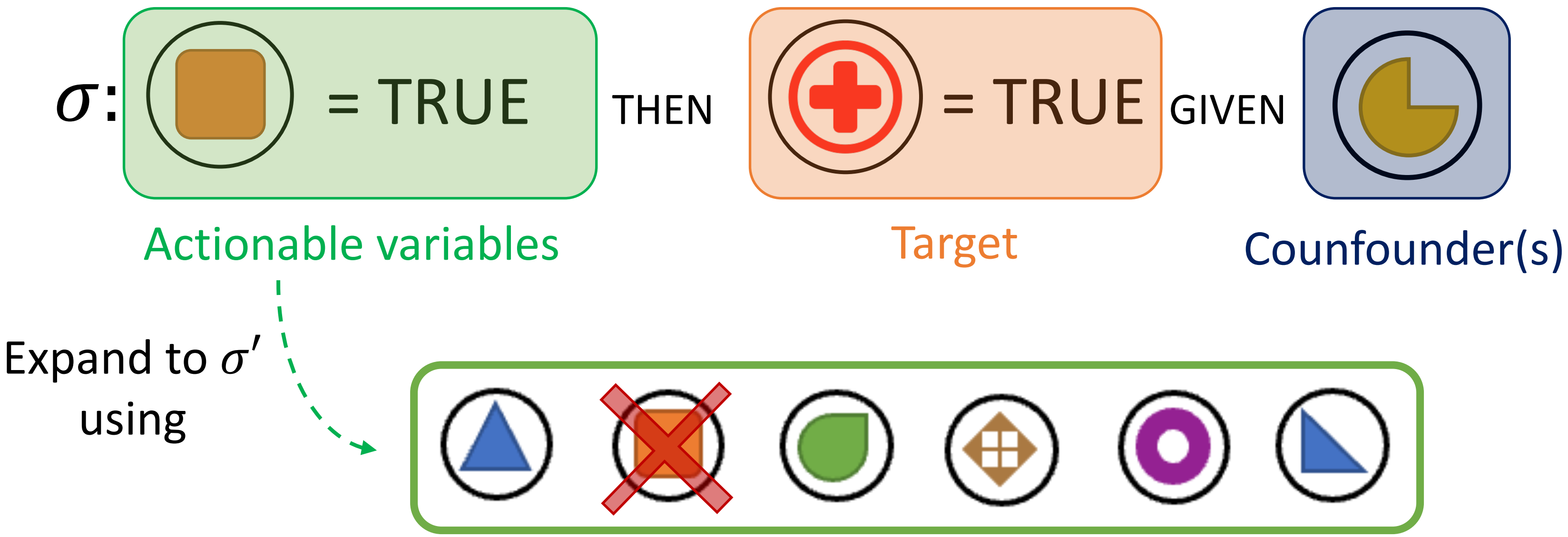
Germline genomic alterations
Sex
Age
…

Somatic genomic alterations

Cancer type

$$TP53_{som} = 1 \wedge ERBB2_{loh} = 1 \rightarrow Basal \mid history\ other\ malignancy$$

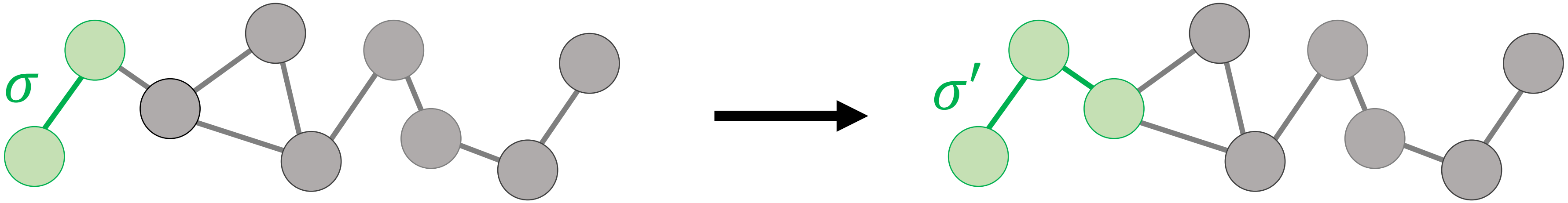
[Budhathoki et al.] proposed the **reliable** rule effect estimation framework and developed a branch and bound algorithm for the discovery task.



Budhathoki, K., Boley, M. and Vreeken, J., 2021. Discovering reliable causal rules. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)* (pp. 1-9). Society for Industrial and Applied Mathematics.

Contributions and Open Challenges

- Guarantees for multiple hypothesis testing
- Prove that the general problem is NP-hard
- Exploit dependency graph G for rule expansion



- Extensive experiments on breast cancer data
- Correction for multiple hypothesis testing issues – Currently working on data dependent corrections with Rademacher Averages

My focus

Develop algorithms to:

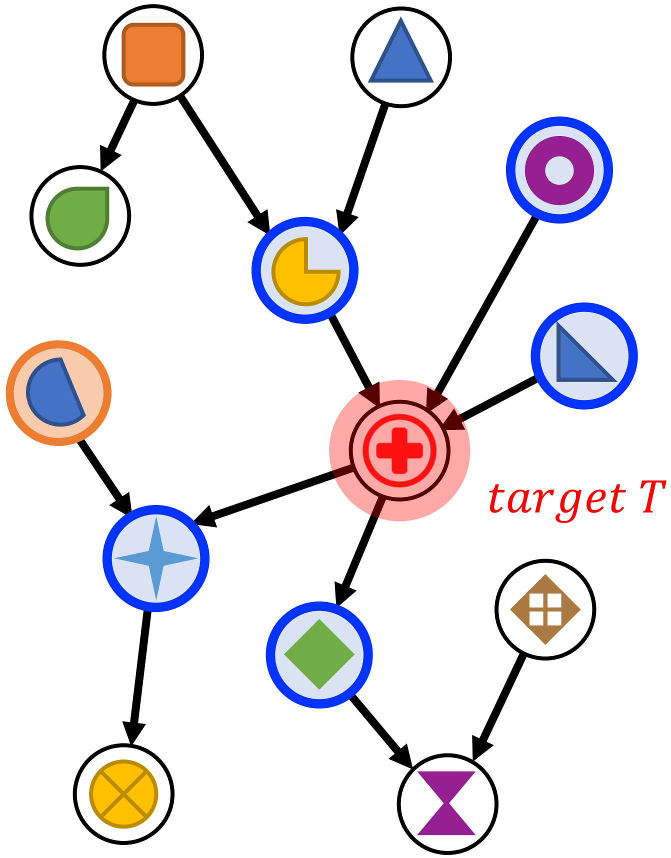
- (Structure discovery) Discover causally related variables to a target
- **(Effect estimation) Evaluate effect of causal rules**

From observational data and providing guarantees on the results

Short recap

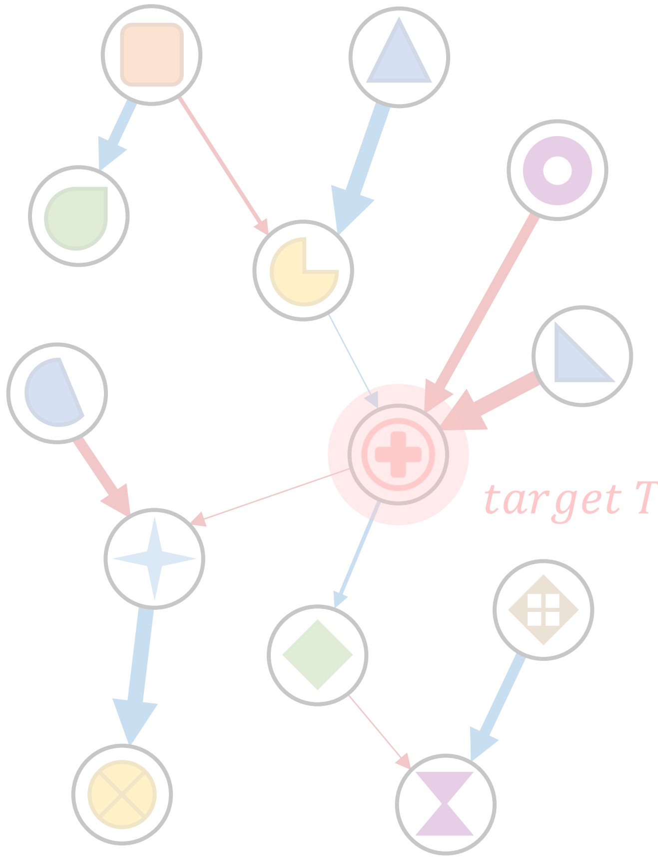
1 - Structure discovery

Discover causally related variables to a target



2 - Effect estimation

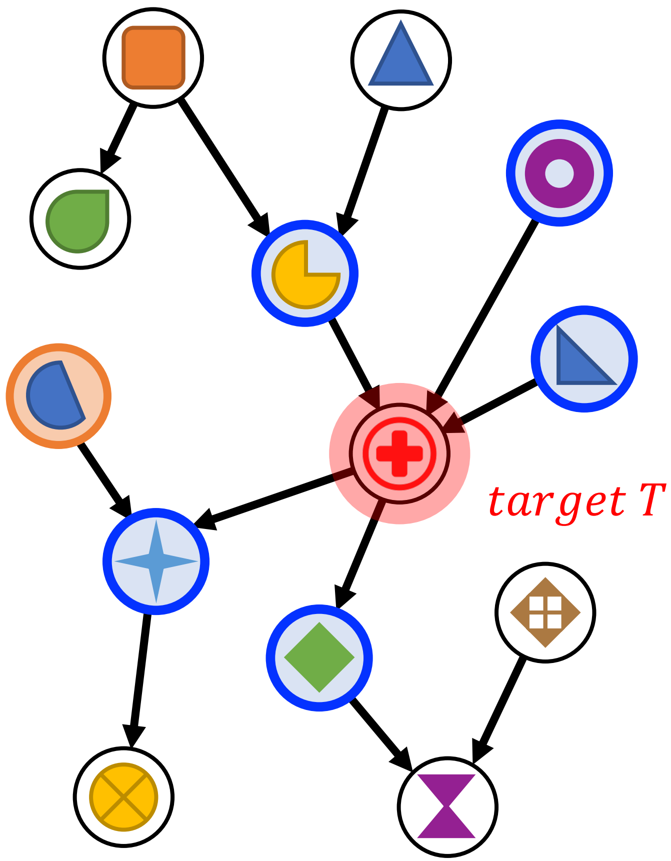
Evaluate effect of causal rules



Short recap

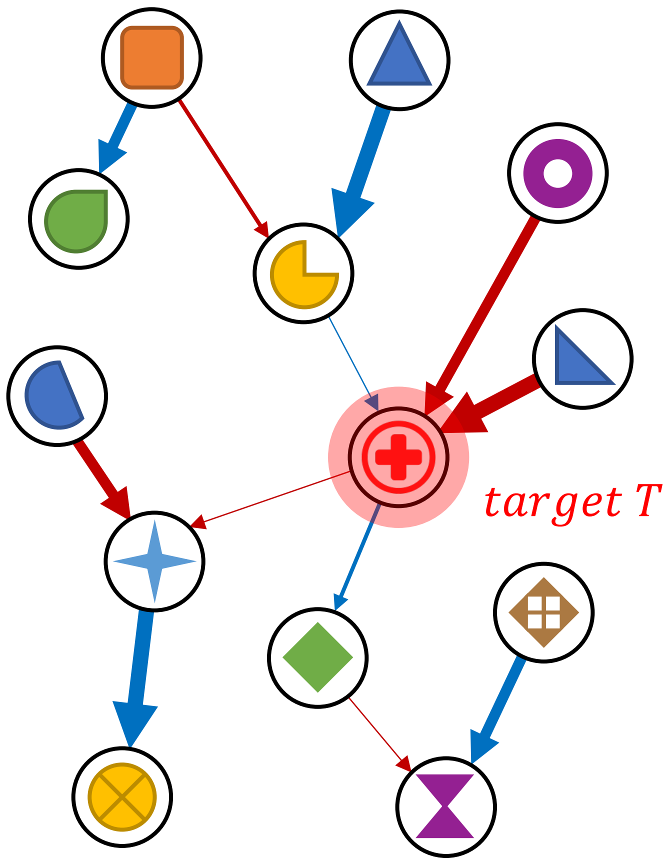
1 - Structure discovery

Discover causally related variables to a target



2 - Effect estimation

Evaluate effect of causal rules



Hopefully true correlation example

Hopefully true correlation example

Exploiting causality methods
for knowledge discovery
from observational data

Q & A TIME

Dario Simionato

dario.simionato@phd.unipd.it

Padua, Italy

Code available at: https://github.com/VandinLab/RAveL