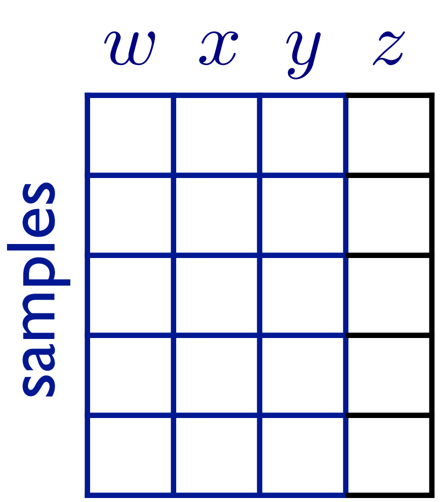
Caltech

Foundations of Causal Discovery

Frederick Eberhardt

Causal Discovery

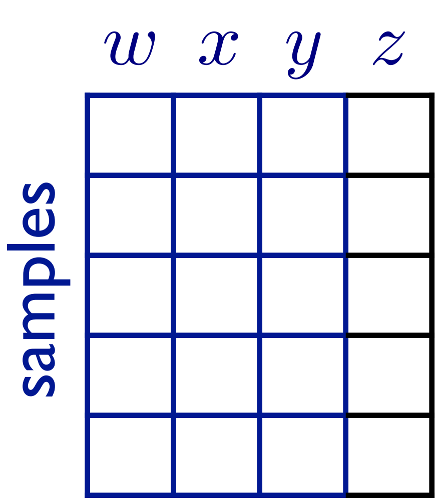
data sample



Causal Discovery

assumptions, e.g.
- causal Markov
- causal faithfulness
- functional form
- etc.

data sample

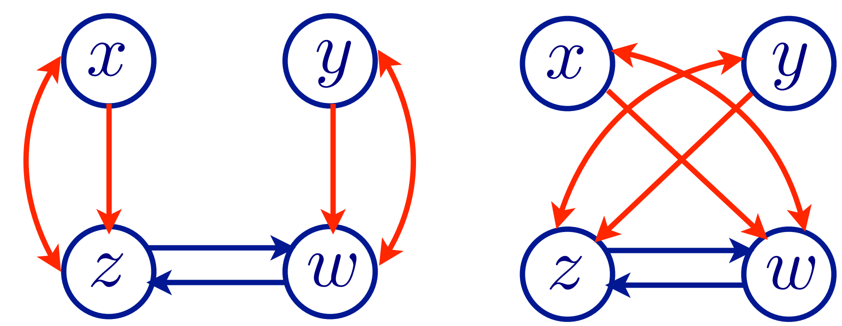


inference algorithm

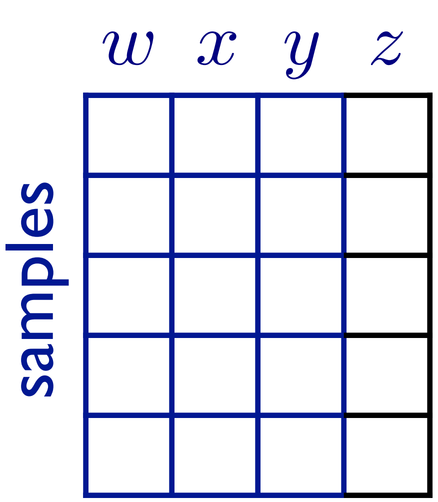
Causal Discovery

assumptions, e.g.
- causal Markov
- causal faithfulness
- functional form
- etc.

equivalence classes



data sample

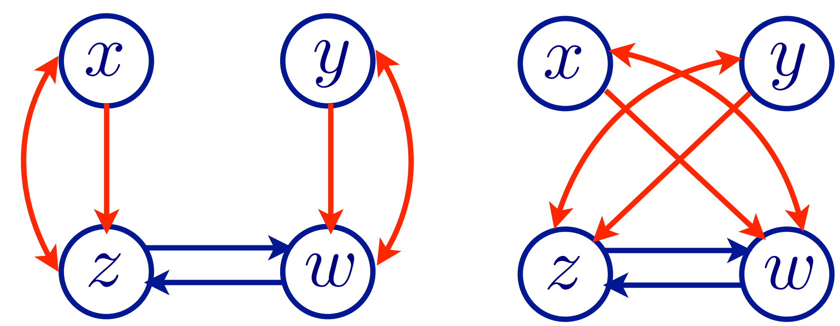


inference algorithm

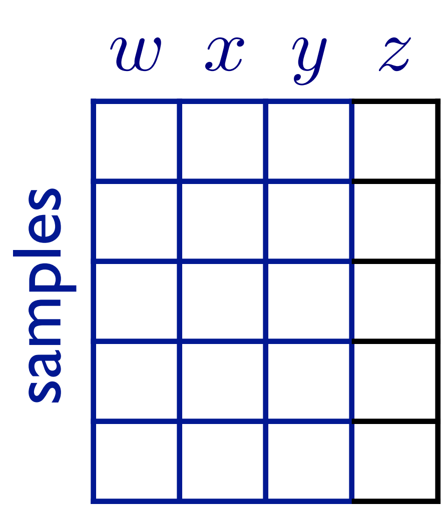
Causal Discovery

assumptions, e.g.
- causal Markov
- causal faithfulness
- functional form
- etc.

equivalence classes



data sample



inference algorithm

model specifications

	w	x	y	z
w	0	0	?	a
x	0	0	0	0
y	0	0	0	0
z	b	?	?	0

direct edges

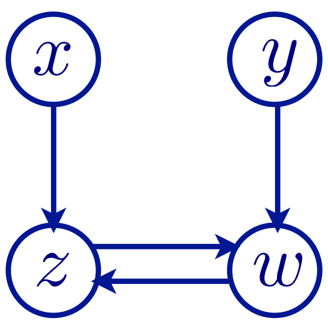
	w	x	y	z
w	0	0	?	?
x	0	0	0	?
y	?	0	0	0
z	?	?	0	0

confounders

Causal Discovery

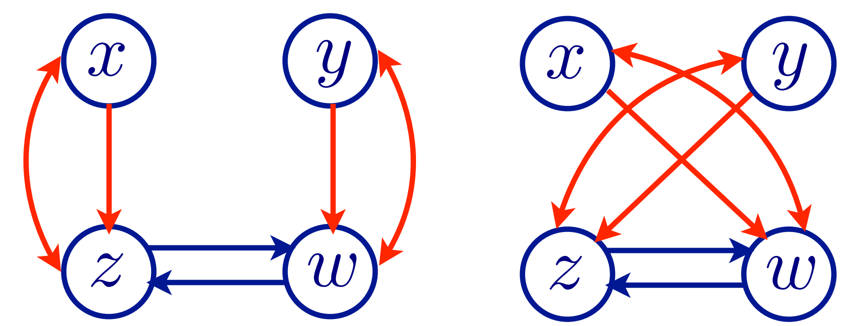
truth

(unknown)

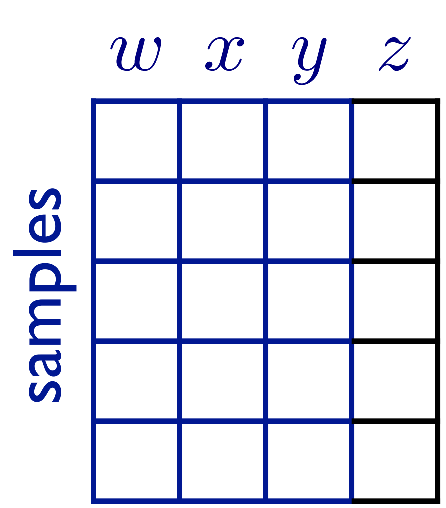


assumptions, e.g.
- causal Markov
- causal faithfulness
- functional form
- etc.

equivalence classes



data sample



inference algorithm

model specifications

	w	x	y	z
w	0	0	?	a
x	0	0	0	0
y	0	0	0	0
z	b	?	?	0

direct edges

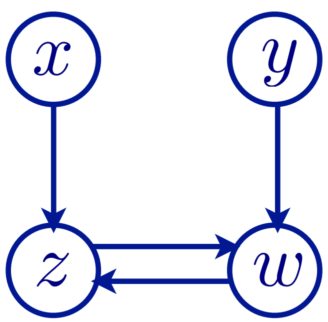
	w	x	y	z
w	0	0	?	?
x	0	0	0	?
y	?	0	0	0
z	?	?	0	0

confounders

Causal Discovery

truth

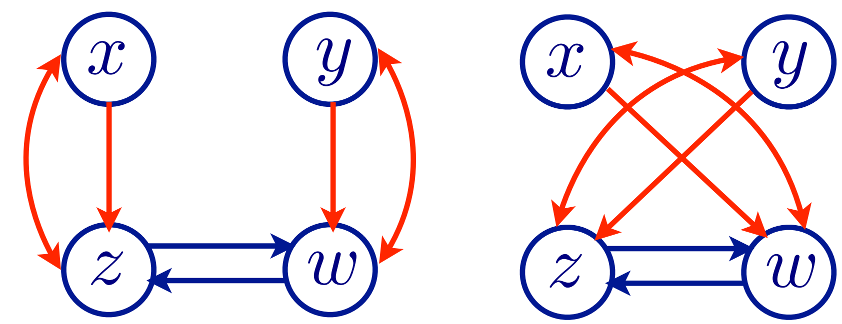
(unknown)



assumptions, e.g.

- causal Markov
- causal faithfulness
- functional form
- etc.

equivalence classes



data sample

samples

w	x	y	z

inference algorithm

model specifications

	w	x	y	z
w	0	0	?	a
x	0	0	0	0
y	0	0	0	0
z	b	?	?	0

direct edges

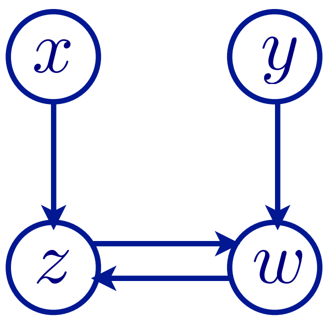
	w	x	y	z
w	0	0	?	?
x	0	0	0	?
y	?	0	0	0
z	?	?	0	0

confounders

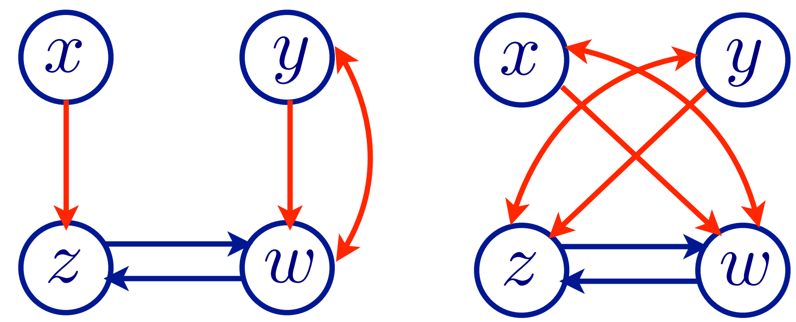
Constraint-based Causal Discovery

truth

(unknown)



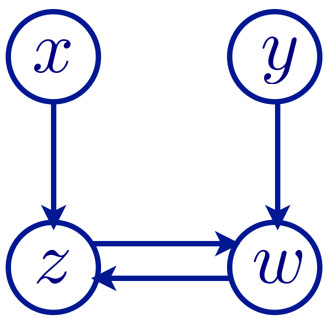
equivalence classes



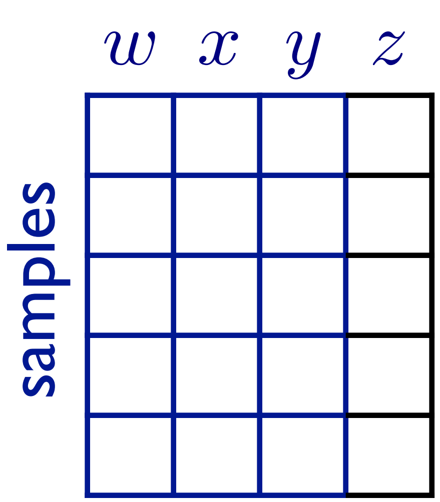
Constraint-based Causal Discovery

truth

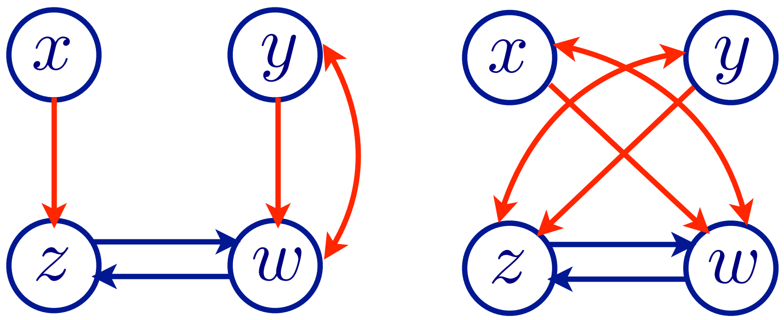
(unknown)



data sample



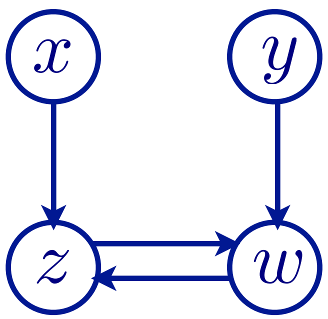
equivalence classes



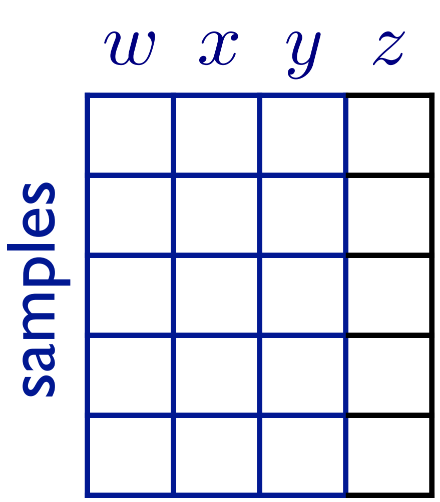
Constraint-based Causal Discovery

truth

(unknown)



data sample

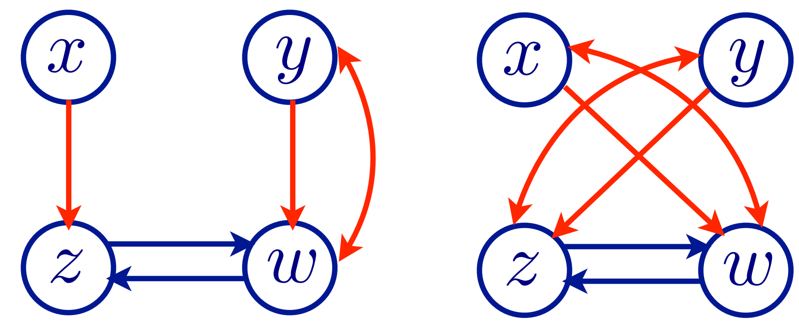


constraints

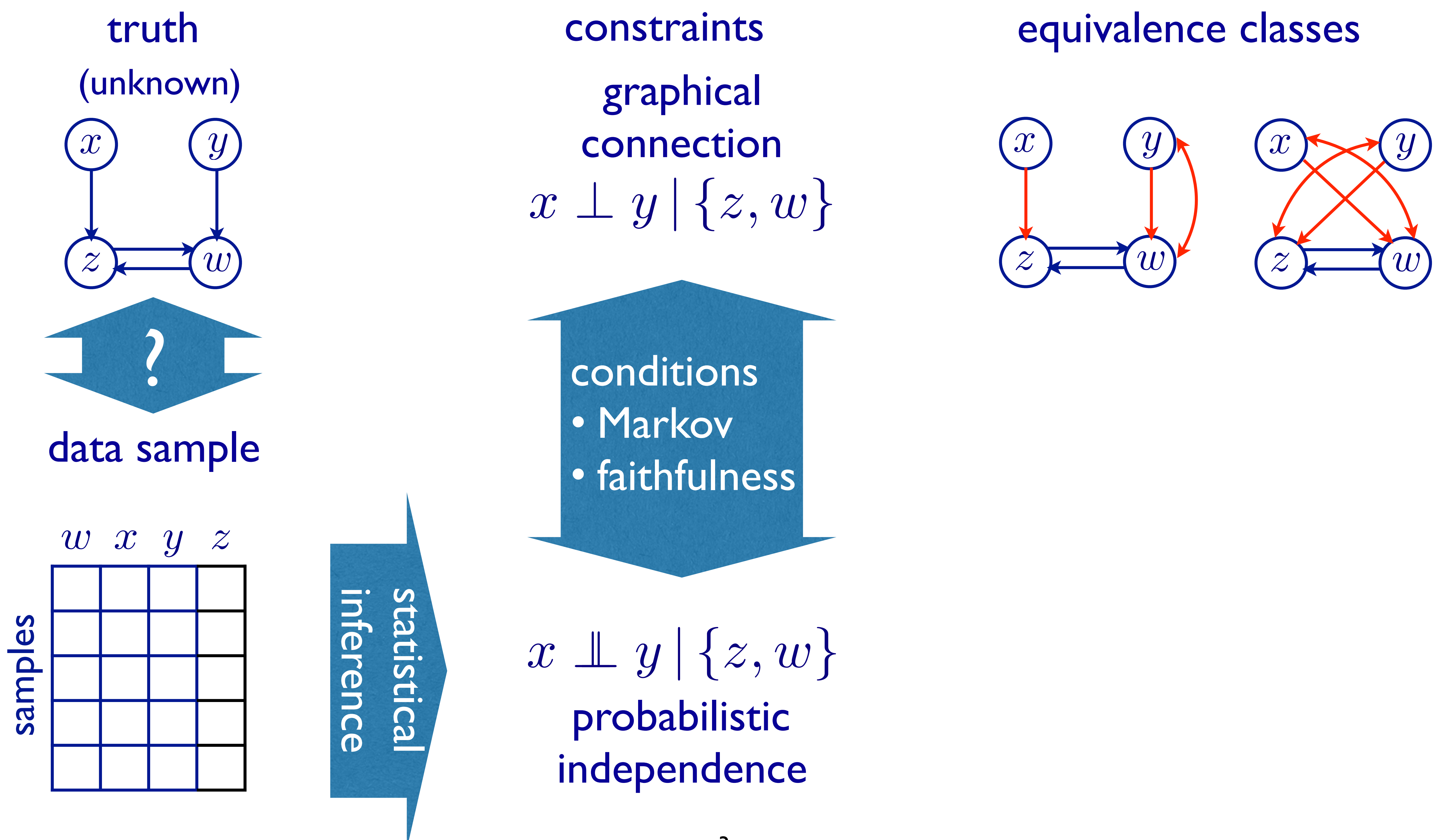
$$x \perp\!\!\!\perp y \,|\, \{z, w\}$$

probabilistic independence

equivalence classes



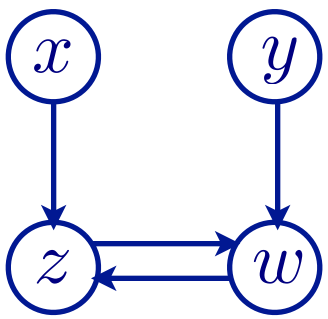
Constraint-based Causal Discovery



Constraint-based Causal Discovery

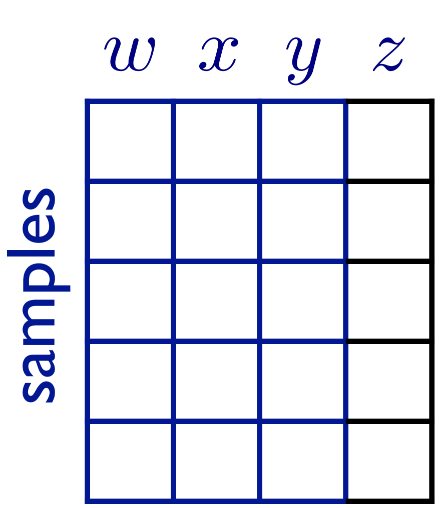
truth

(unknown)



?

data sample



d-separation

statistical inference

constraints

graphical connection

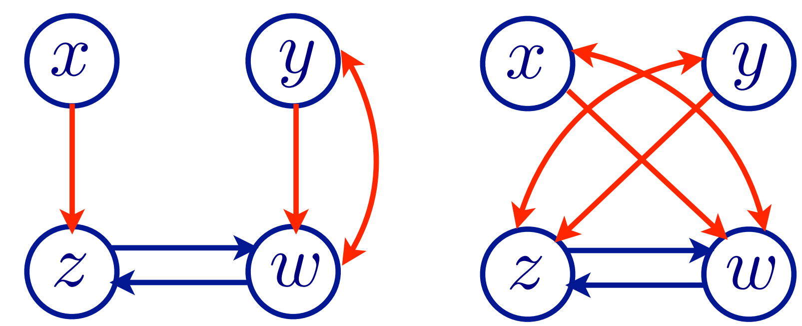
$$x \perp y \,|\, \{z, w\}$$

conditions
- Markov
- faithfulness

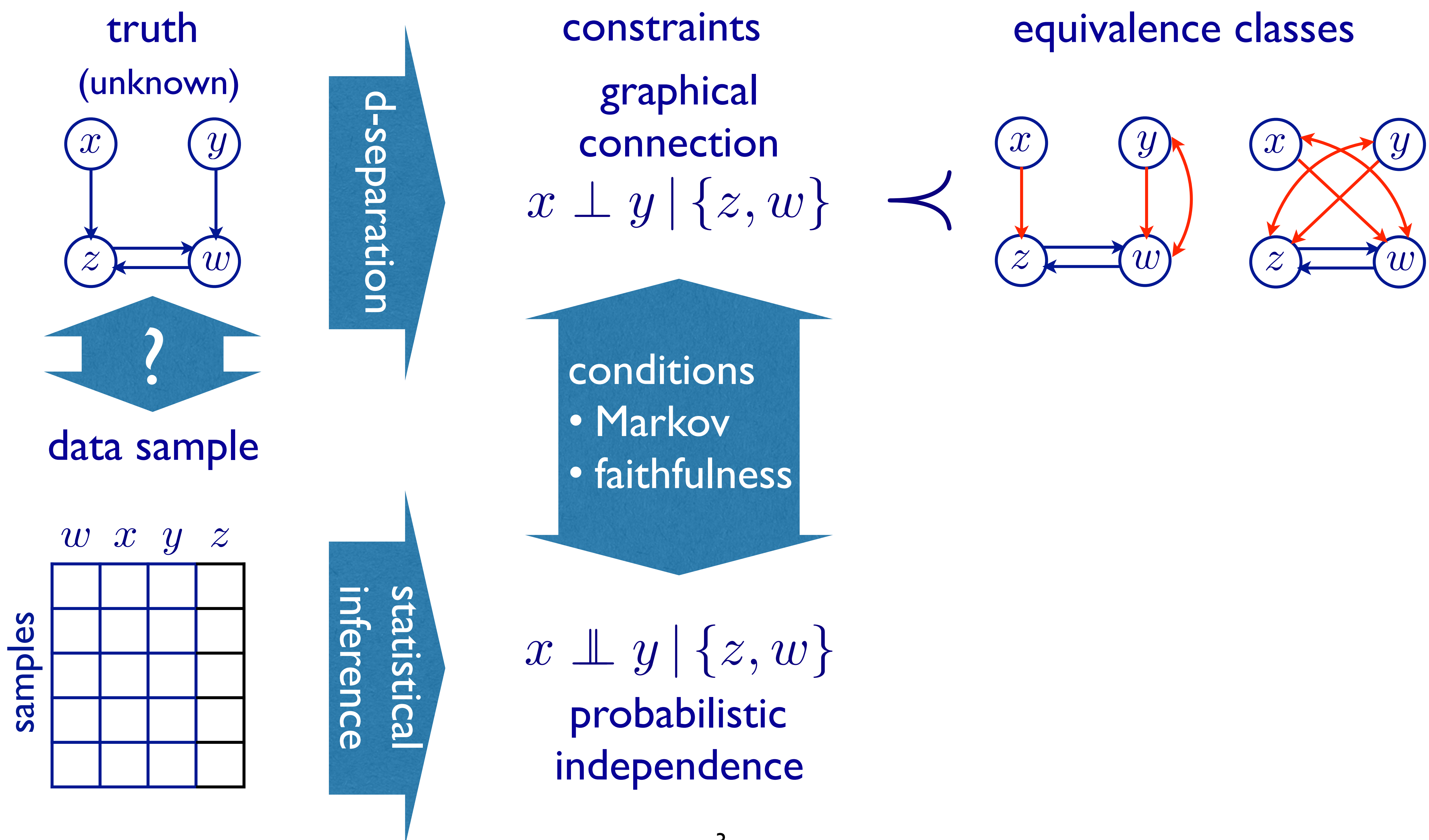
$$x \perp\!\!\!\perp y \,|\, \{z, w\}$$

probabilistic independence

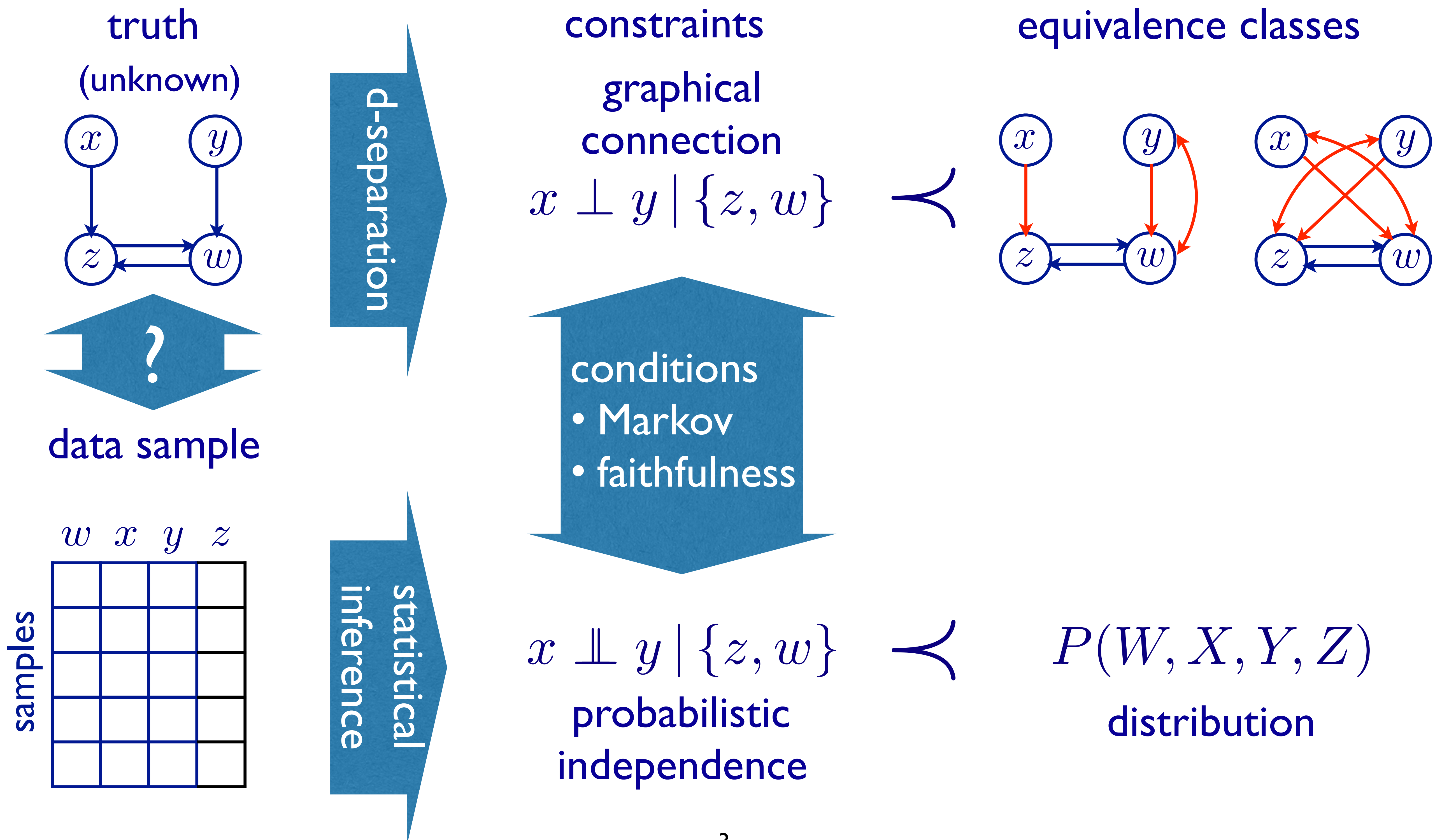
equivalence classes



Constraint-based Causal Discovery

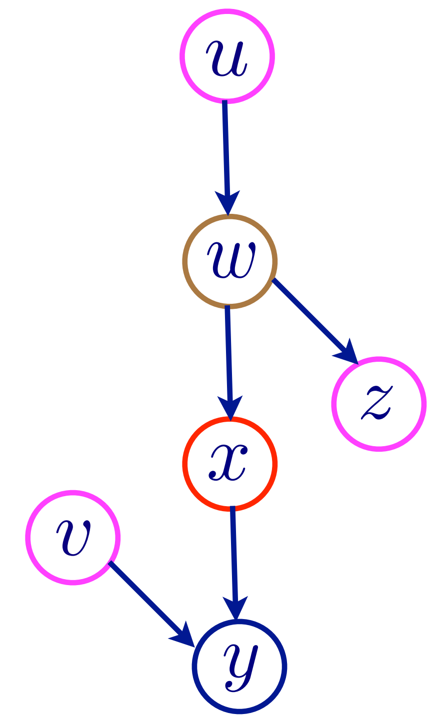


Constraint-based Causal Discovery



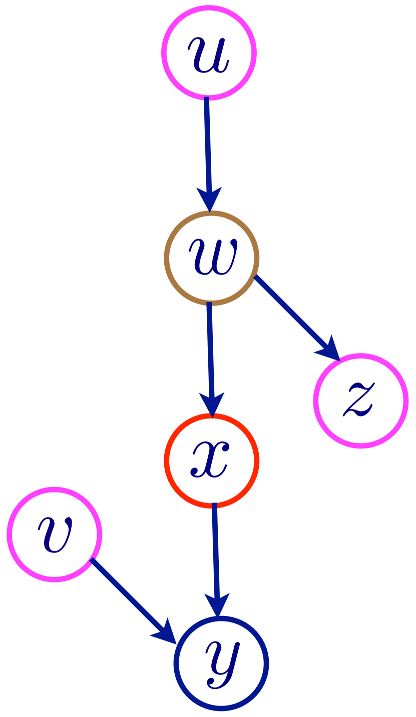
Causal Markov

x is independent of its non-descendents given its parents in the causal graph



Causal Markov

x is independent of its non-descendents given its parents in the causal graph



Violations of Causal Markov

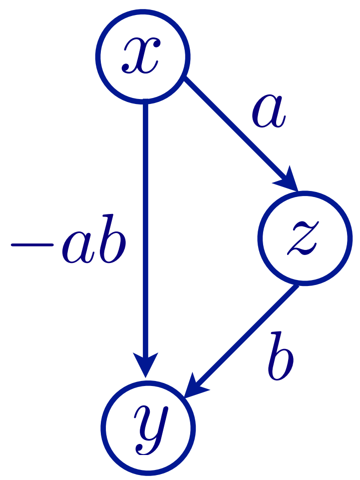
- quantum mechanics
- [unmeasured common causes]
- [mixtures of populations]
- [variables are not distinct, or too coarsely grained]

Causal Faithfulness

If x is independent of y given **C** in the probability distribution then x is d-separated from y given **C** in the graph.

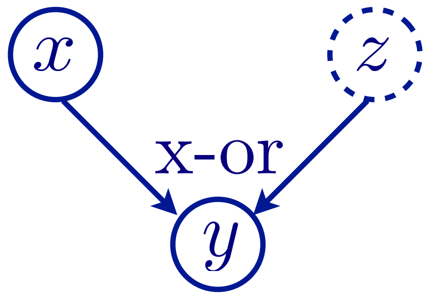
Causal Faithfulness

If x is independent of y given **C** in the probability distribution then x is d-separated from y given **C** in the graph.



Violations of Causal Faithfulness

- canceling pathways
- matching pennies cases
- [small sample sizes and near violations of faithfulness]

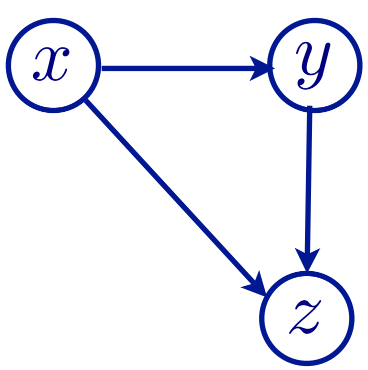


Assumptions

- **causal Markov**: permits inference from probabilistic dependence to causal connection

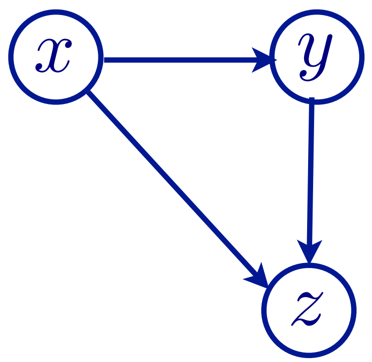
Assumptions

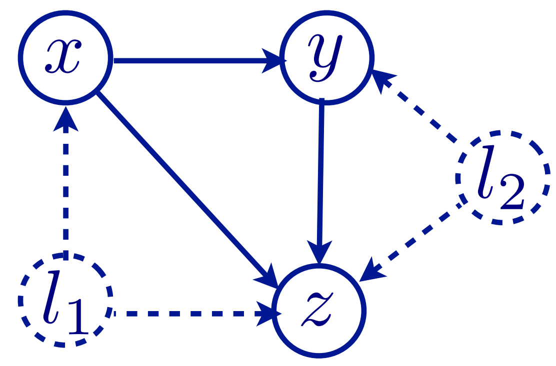
- **causal Markov**: permits inference from probabilistic dependence to causal connection
- **causal faithfulness**: permits inference from probabilistic independence to causal separation



Assumptions

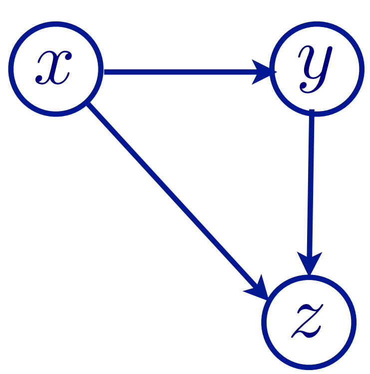
- **causal Markov**: permits inference from probabilistic dependence to causal connection
- **causal faithfulness**: permits inference from probabilistic independence to causal separation
- **causal sufficiency**: there are no unmeasured common causes

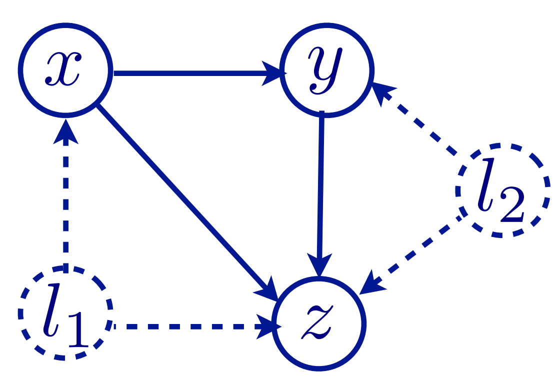


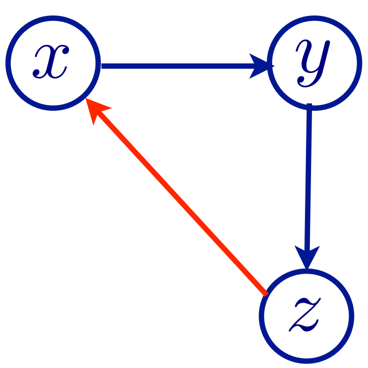


Assumptions

- **causal Markov**: permits inference from probabilistic dependence to causal connection
- **causal faithfulness**: permits inference from probabilistic independence to causal separation
- **causal sufficiency**: there are no unmeasured common causes
- **acylicity**: no variable is an (indirect) cause of itself







Assumptions

- **causal Markov**: permits inference from probabilistic dependence to causal connection
- **causal faithfulness**: permits inference from probabilistic independence to causal separation
- **causal sufficiency**: there are no unmeasured common causes
- **acylicity**: no variable is an (indirect) cause of itself

All graphs in an equivalence class have:

- same adjacencies (“skeleton”)
- same unshielded colliders

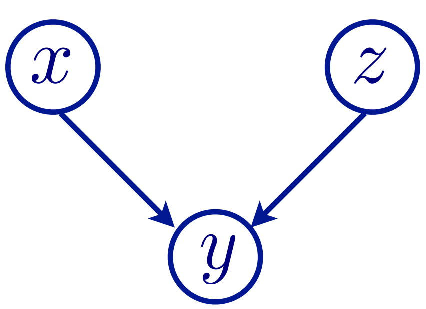
[Verma & Pearl 1990, Frydenberg 1990]

Assumptions

- **causal Markov**: permits inference from probabilistic dependence to causal connection
- **causal faithfulness**: permits inference from probabilistic independence to causal separation
- **causal sufficiency**: there are no unmeasured common causes
- **acylicity**: no variable is an (indirect) cause of itself

All graphs in an equivalence class have:
- same adjacencies ("skeleton")
- same unshielded colliders

[Verma & Pearl 1990, Frydenberg 1990]



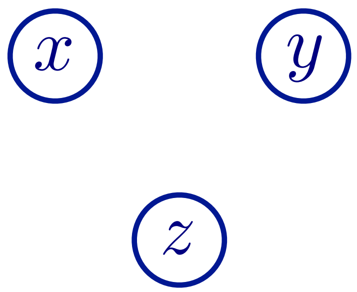
unshielded collider

assumptions

- Markov
- faithfulness
- acyclicity
- causal sufficiency

equivalence class

- same adjacencies (“skeleton”)
- same unshielded colliders

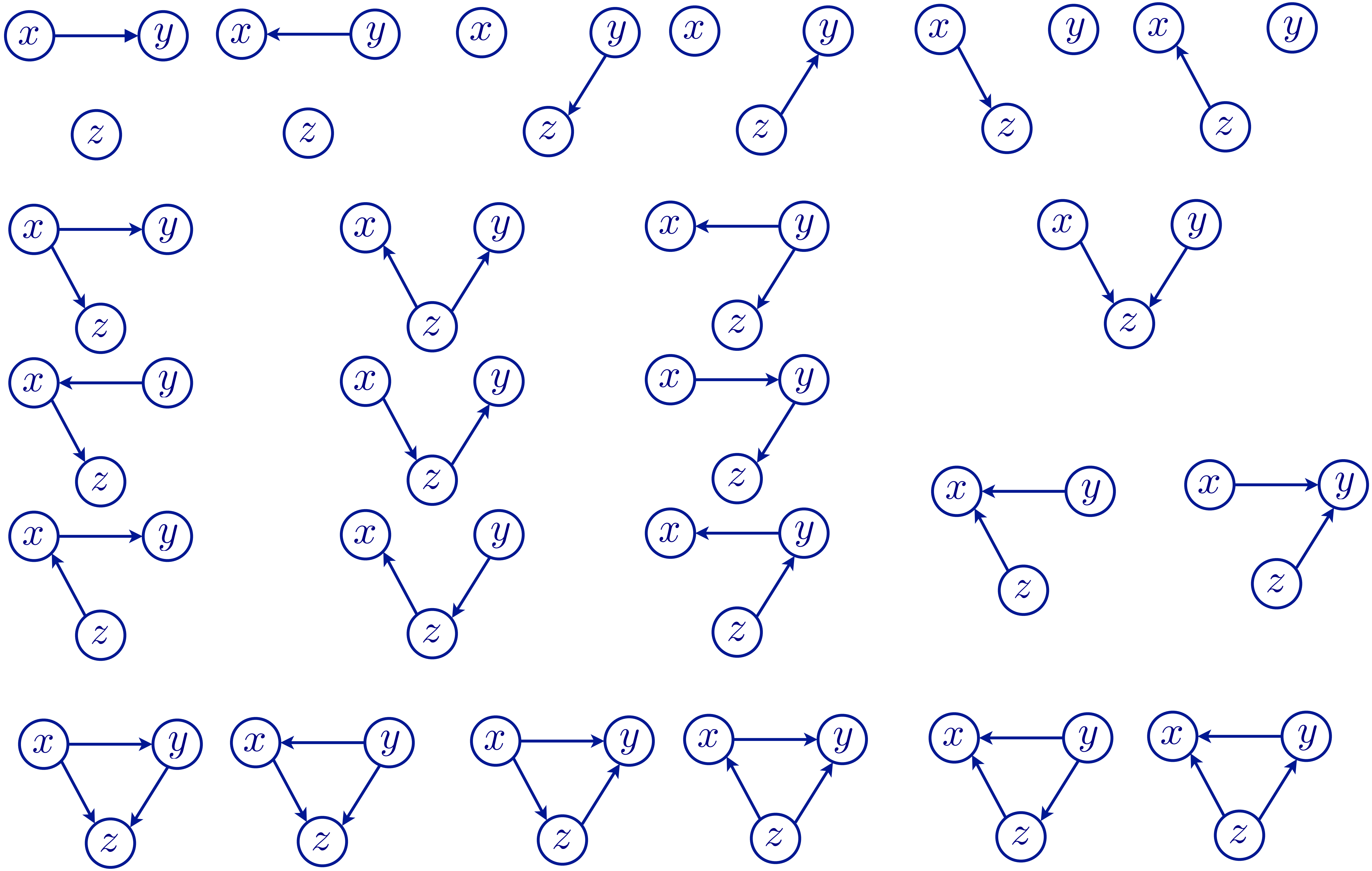


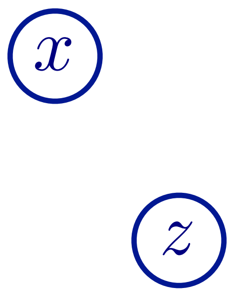
assumptions

- Markov
- faithfulness
- acyclicity
- causal sufficiency

equivalence class

- same adjacencies (“skeleton”)
- same unshielded colliders



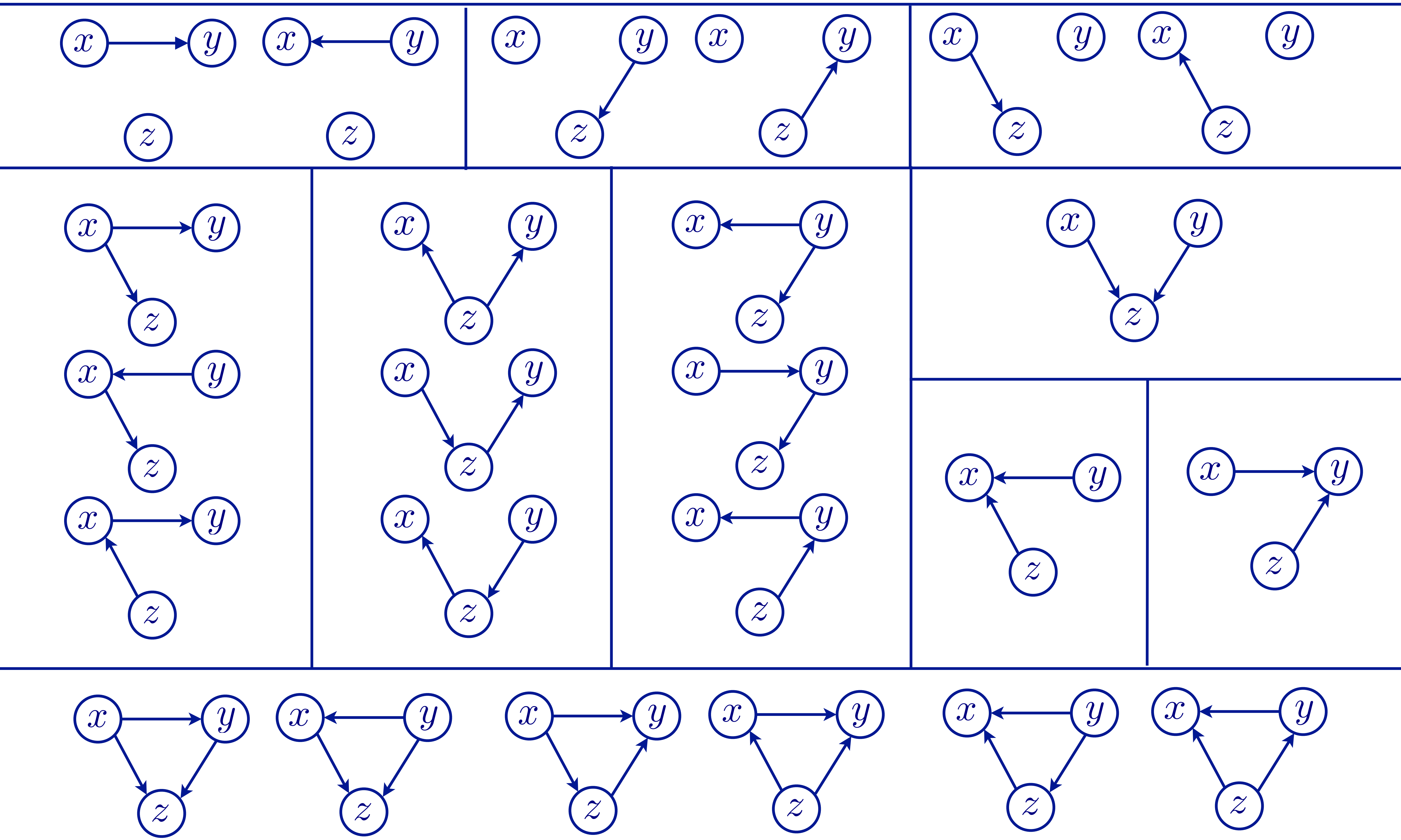


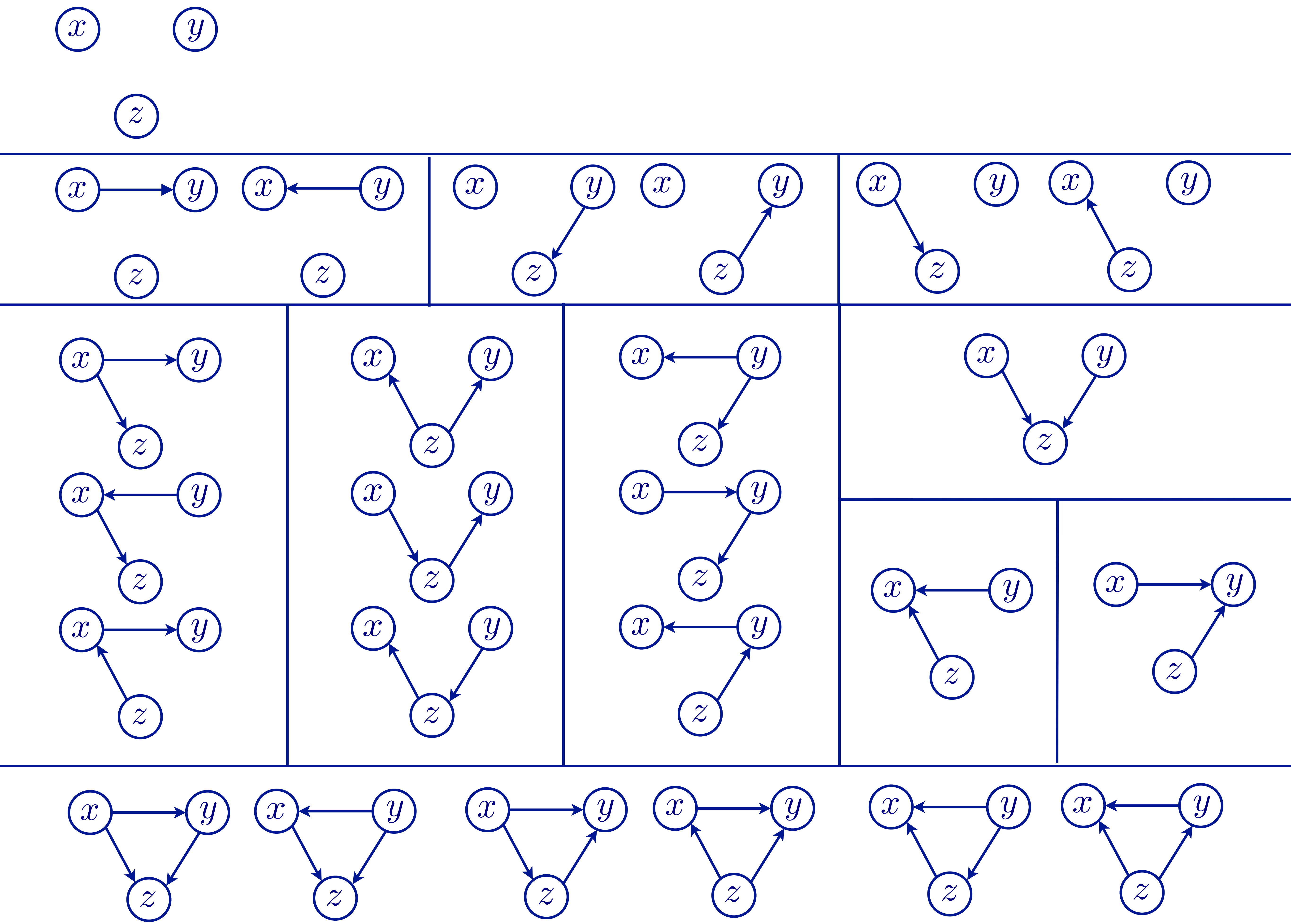
assumptions

- Markov
- faithfulness
- acyclicity
- causal sufficiency

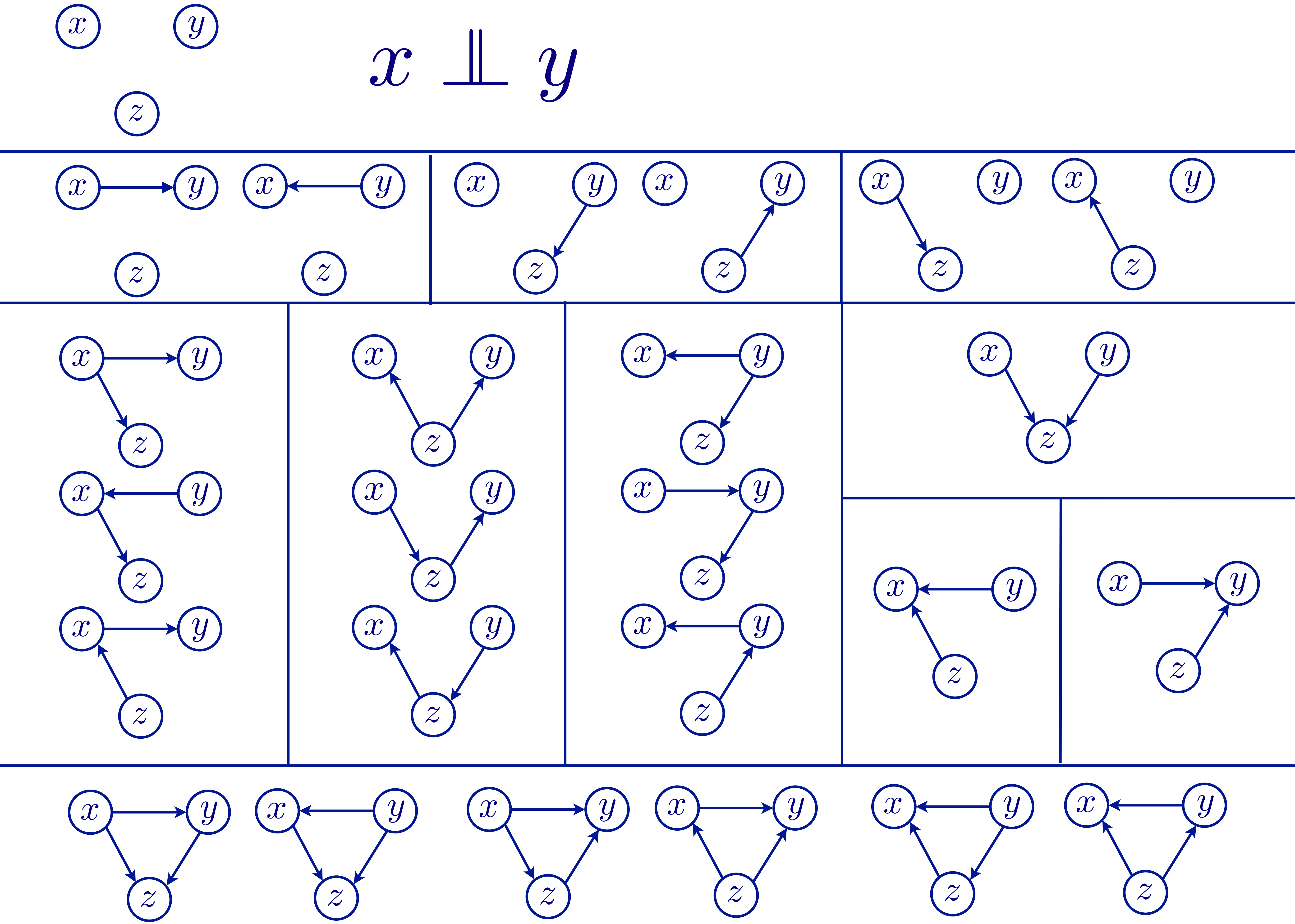
equivalence class

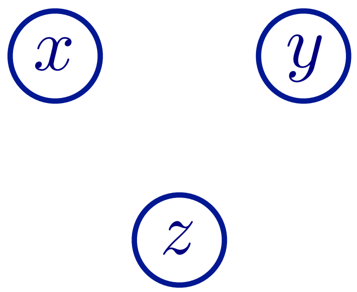
- same adjacencies ("skeleton")
- same unshielded colliders



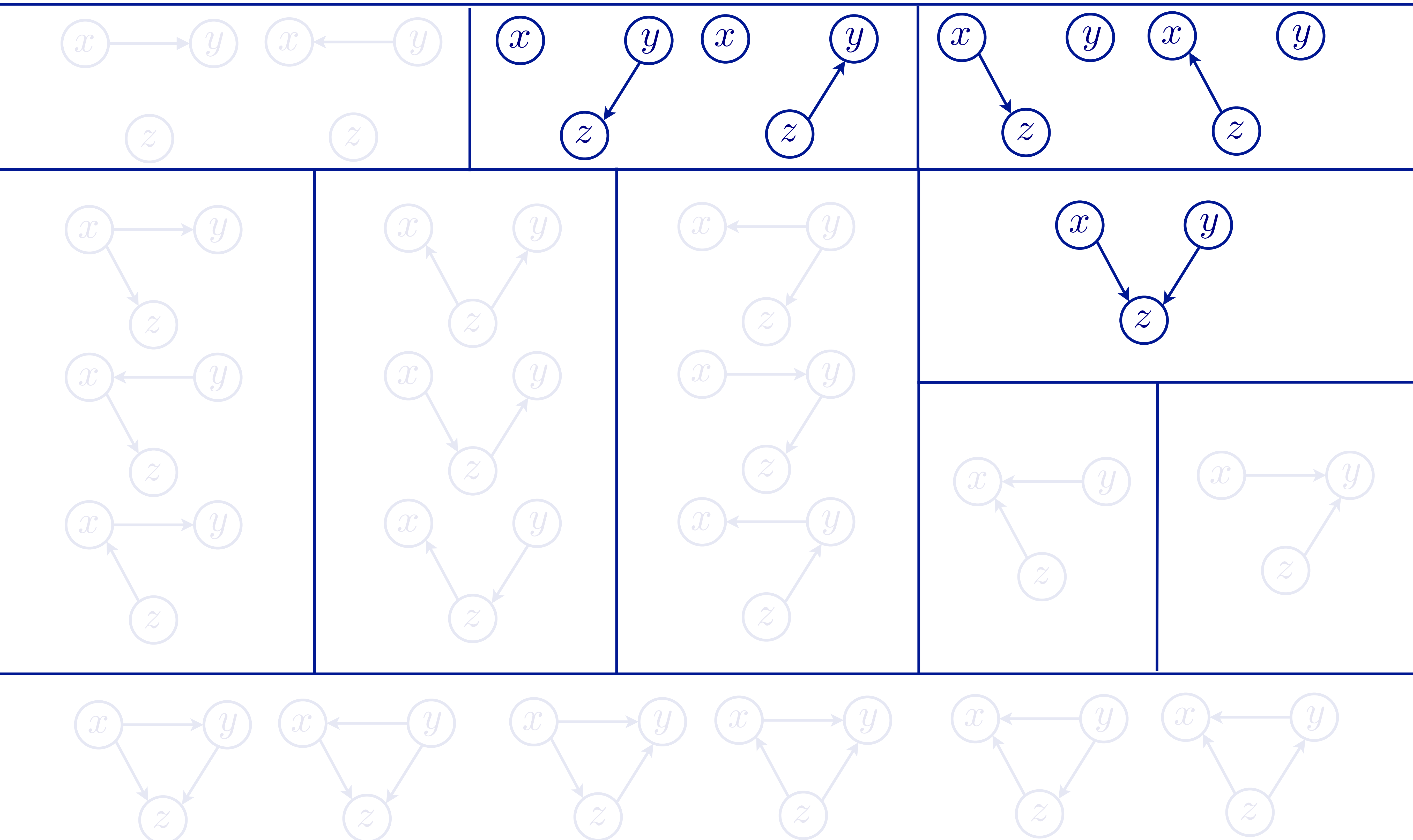


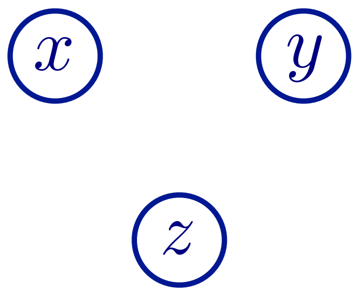
$$x \perp\!\!\!\perp y$$



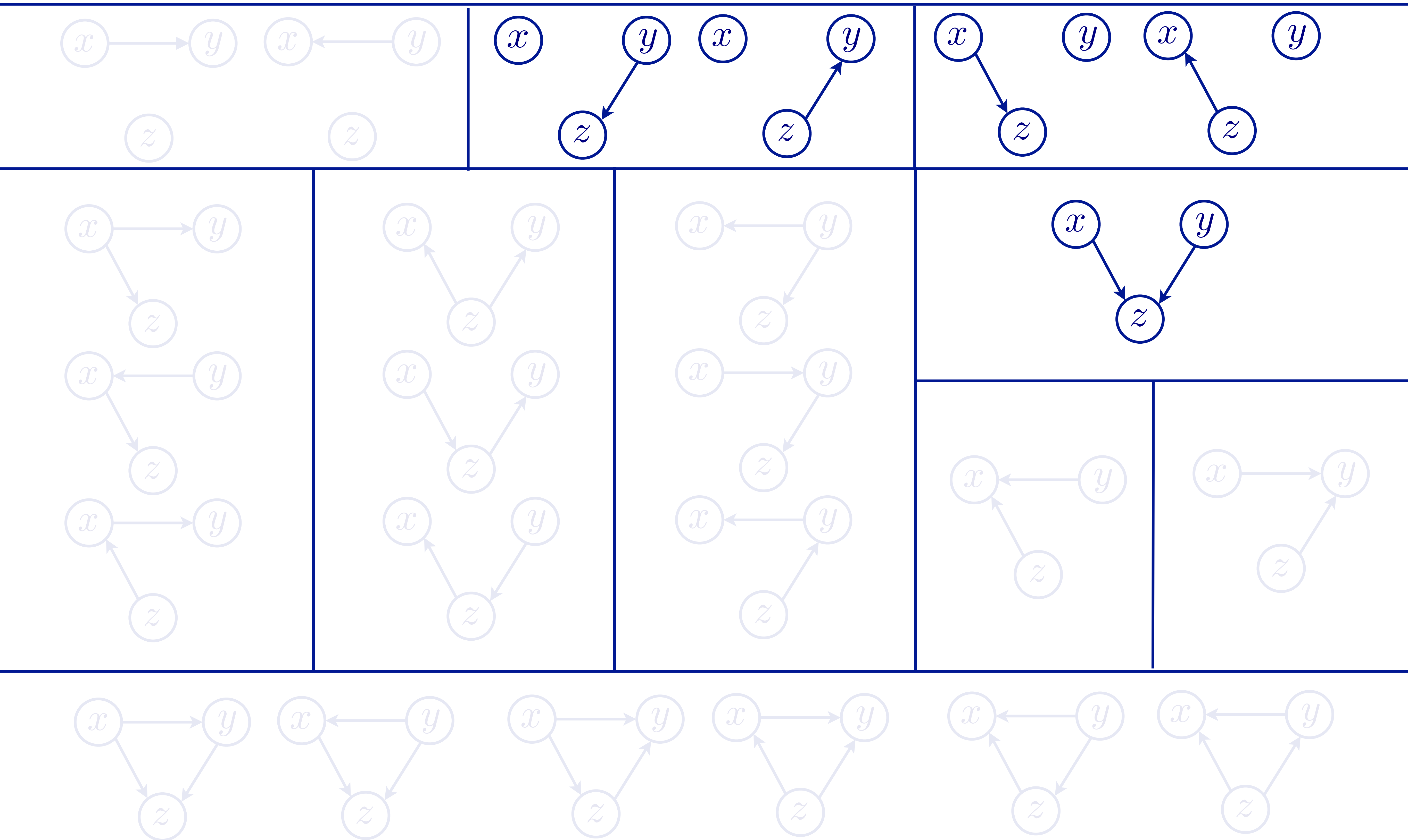


$$x \perp\!\!\!\perp y$$

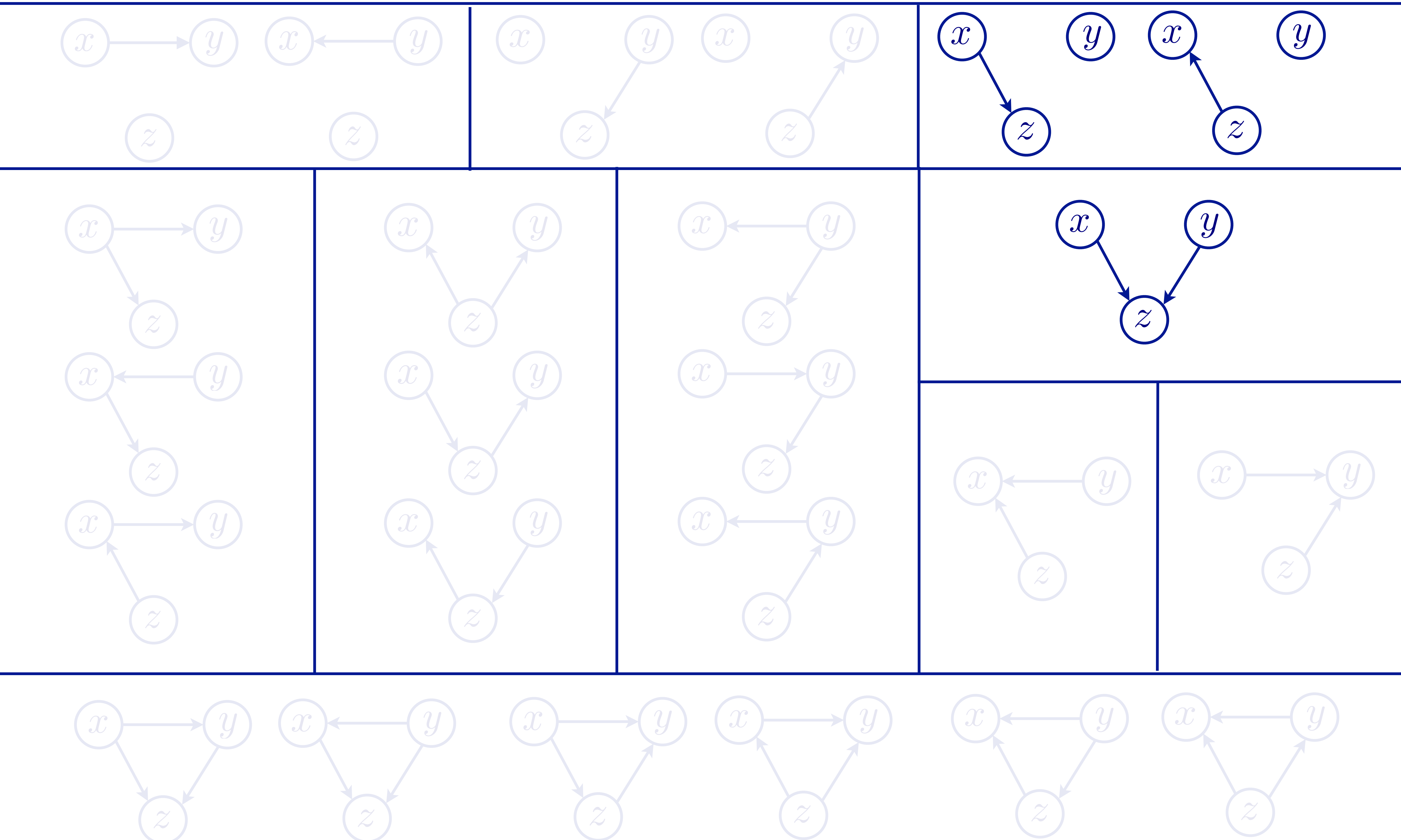




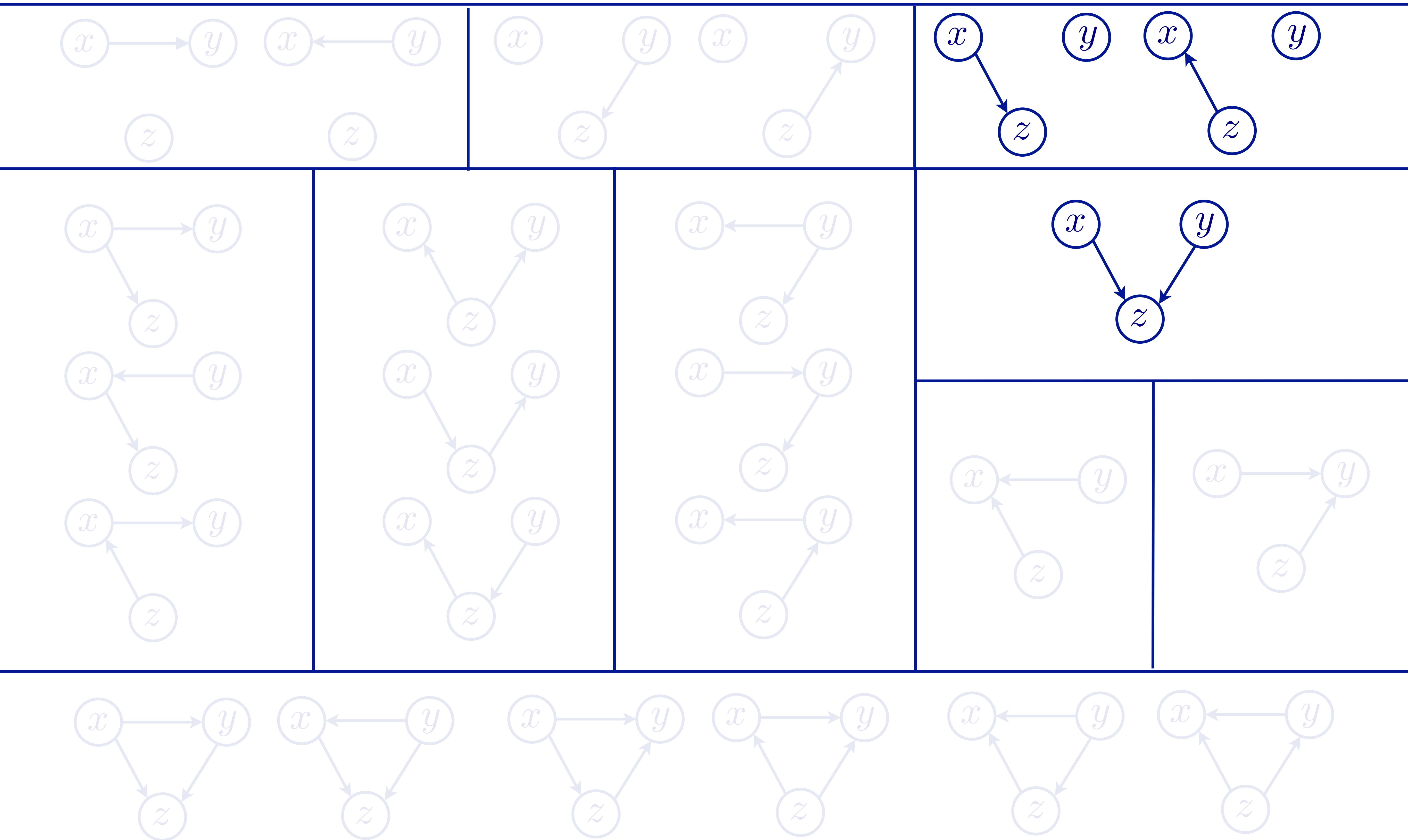
$$x \perp\!\!\!\perp y \qquad x \not\!\perp\!\!\!\perp z$$



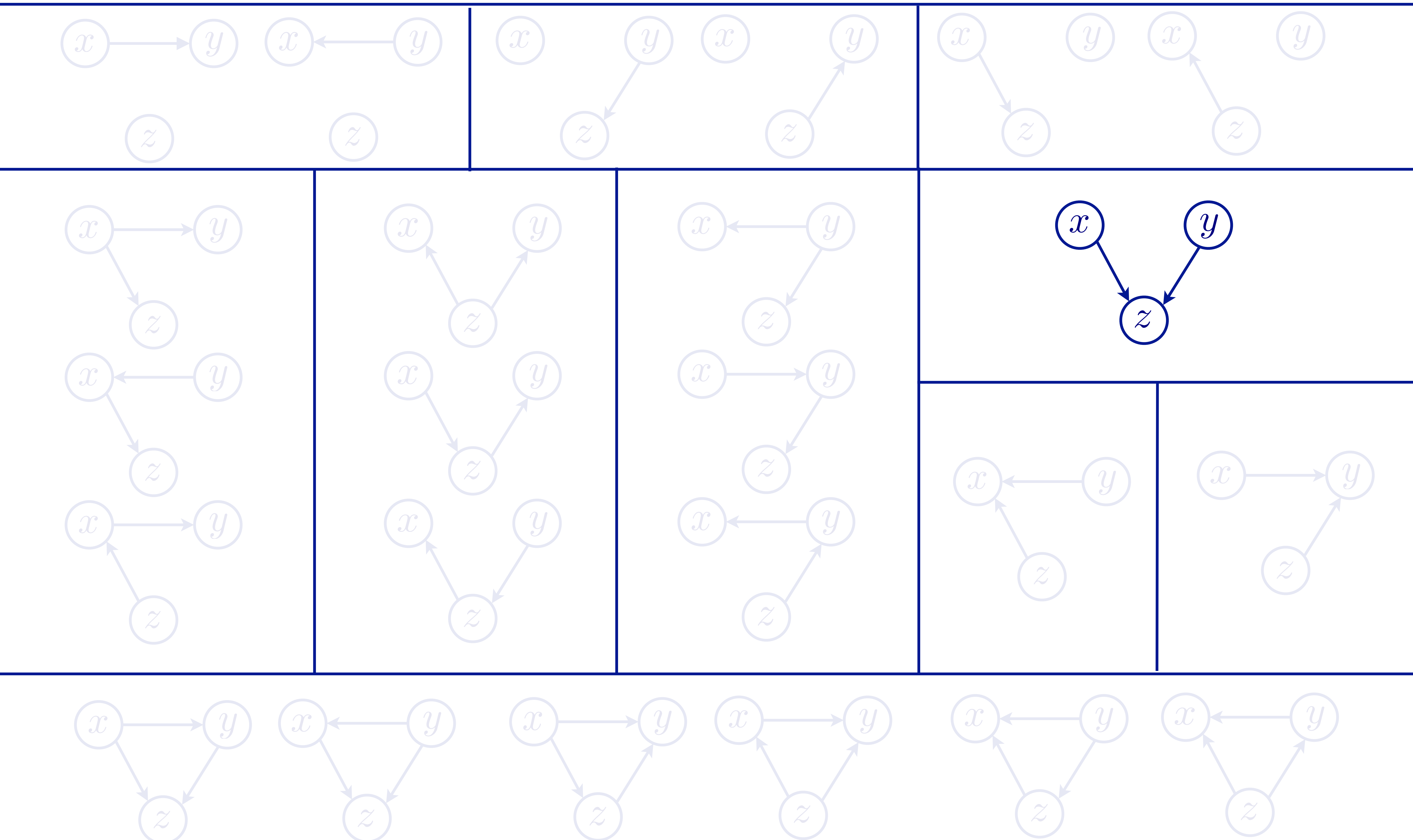
$$x \perp\!\!\!\perp y \qquad x \not\perp\!\!\!\perp z$$



$$x \perp\!\!\!\perp y \qquad x \not\perp\!\!\!\perp z \qquad y \not\perp\!\!\!\perp z$$

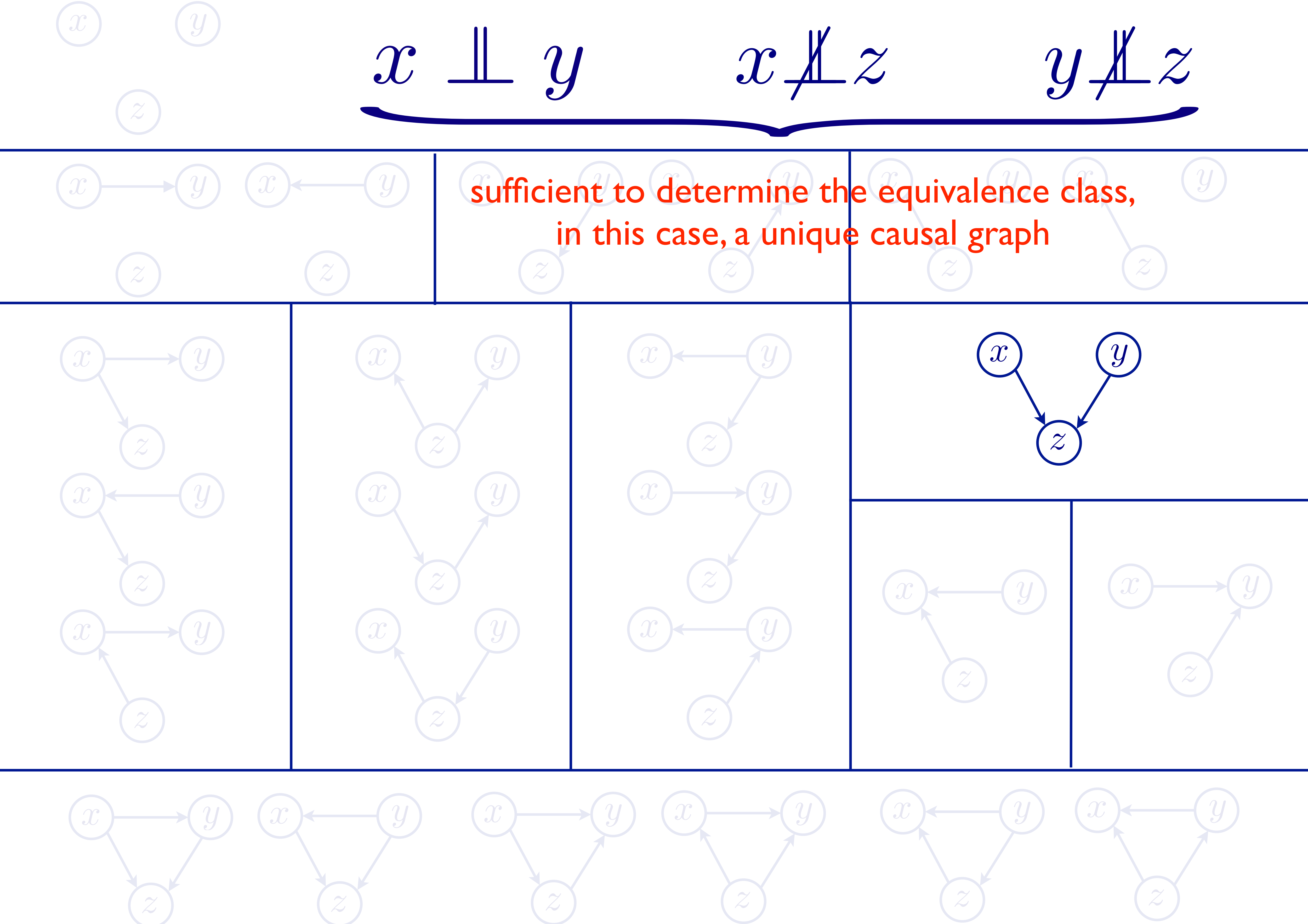


$x \perp\!\!\!\perp y \qquad x \not\perp\!\!\!\perp z \qquad y \not\perp\!\!\!\perp z$

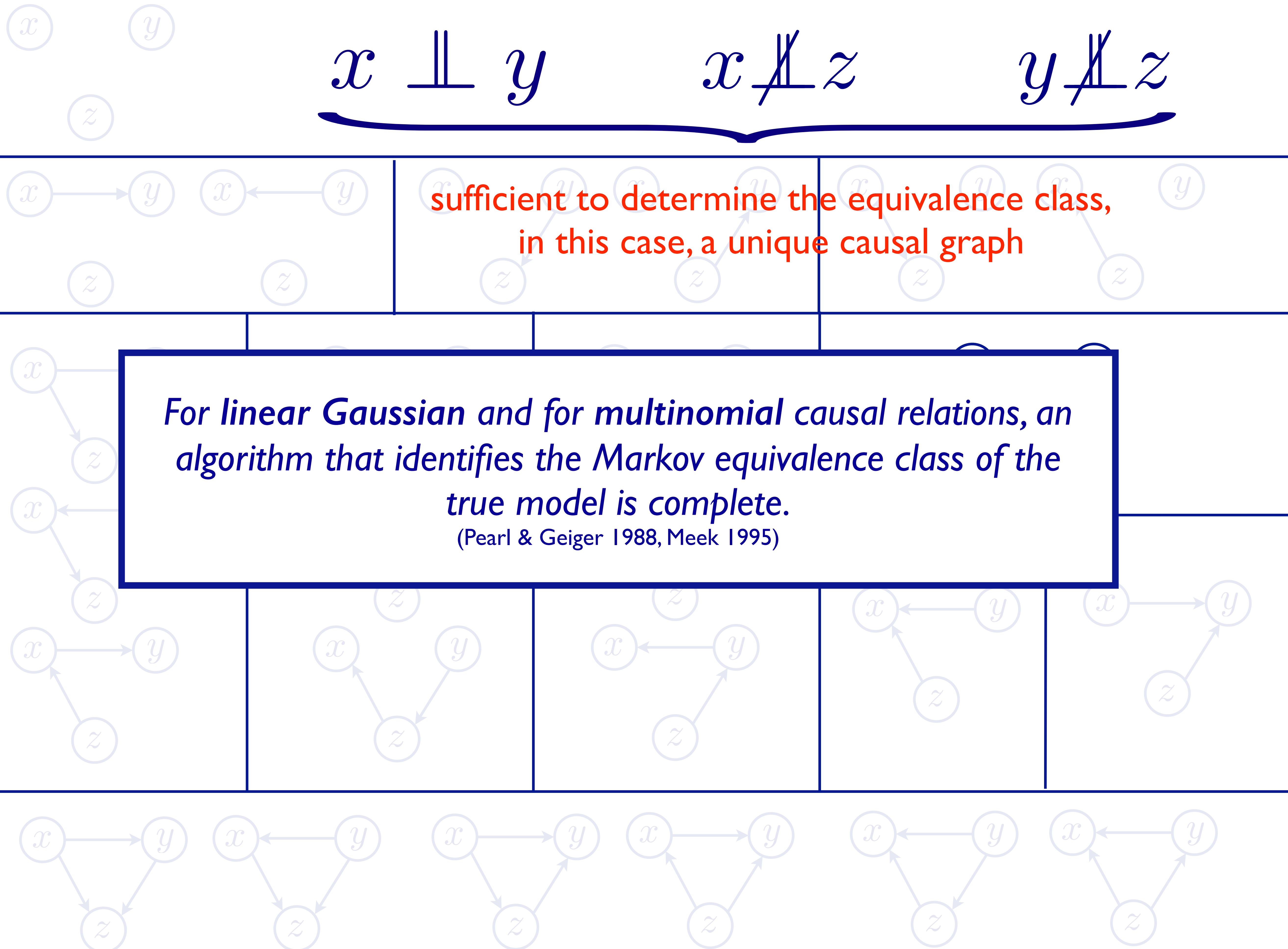


$$\underbrace{x \perp\!\!\!\perp y \qquad x \not\perp\!\!\!\perp z \qquad y \not\perp\!\!\!\perp z}$$

sufficient to determine the equivalence class,
in this case, a unique causal graph



$$\underbrace{x \perp\!\!\!\perp y \qquad x \not\perp\!\!\!\perp z \qquad y \not\perp\!\!\!\perp z}$$



sufficient to determine the equivalence class,
in this case, a unique causal graph

*For **linear Gaussian** and for **multinomial** causal relations, an algorithm that identifies the Markov equivalence class of the true model is complete.*

(Pearl & Geiger 1988, Meek 1995)

Staying in business

- **Weaken the assumptions** (and increase the equivalence class)
  - allow for unmeasured common causes
  - allow for cycles
  - weaken faithfulness

Staying in business

- **Weaken the assumptions** (and increase the equivalence class)
  - allow for unmeasured common causes
  - allow for cycles
  - weaken faithfulness
- **Exclude the limitations** (and reduce the equivalence class)
  - restrict to non-Gaussian error distributions
  - restrict to non-linear causal relations
  - restrict to specific discrete parameterizations

Staying in business

- **Weaken the assumptions** (and increase the equivalence class)
  - allow for unmeasured common causes
  - allow for cycles
  - weaken faithfulness
- **Exclude the limitations** (and reduce the equivalence class)
  - restrict to non-Gaussian error distributions
  - restrict to non-linear causal relations
  - restrict to specific discrete parameterizations
- **Include more general data collection set-ups** (and see how assumptions can be adjusted and what equivalence class results)
  - experimental evidence
  - multiple (overlapping) data sets
  - relational data

Staying in business

- **Weaken the assumptions** (and increase the equivalence class)
  - allow for unmeasured common causes
  - allow for cycles
  - weaken faithfulness

Zhalama talk

- **Exclude the limitations** (and reduce the equivalence class)
  - restrict to non-Gaussian error distributions

Tank talk

  - restrict to non-linear causal relations
  - restrict to specific discrete parameterizations
- **Include more general data collection set-ups** (and see how assumptions can be adjusted and what equivalence class results)
  - experimental evidence
  - multiple (overlapping) data sets
  - relational data

Limitations

*For **linear Gaussian** and for **multinomial** causal relations, an algorithm that identifies the Markov equivalence class of the true model is complete.*

(Pearl & Geiger 1988, Meek 1995)

Linear non-Gaussian method (LiNGaM)

- Linear causal relations:

$$x_i = \sum_{x_j \in \mathbf{Pa}(x_i)} \beta_{ij} x_j + \epsilon_j$$

- Assumptions:
  - causal Markov
  - causal sufficiency
  - acyclicity

[Shimizu et al., 2006]

Linear non-Gaussian method (LiNGaM)

- Linear causal relations:

$$x_i = \sum_{x_j \in \mathbf{Pa}(x_i)} \beta_{ij} x_j + \epsilon_j$$

- Assumptions:
  - causal Markov
  - causal sufficiency
  - acyclicity

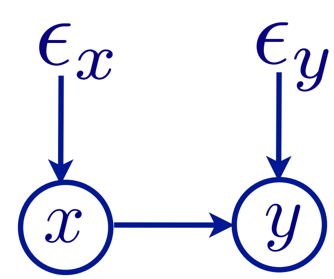
▸ If $\epsilon_j \sim$ non-Gaussian, then the true graph is **uniquely identifiable** from the joint distribution.

[Shimizu et al., 2006]

Two variable case

True model

$$y = \beta x + \epsilon_y$$



Two variable case

True model

$$y = \beta x + \epsilon_y$$

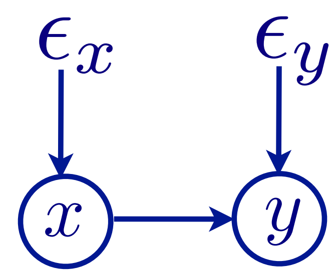
$$\begin{tikzcd} \epsilon_x \arrow[d] & \epsilon_y \arrow[d] \\ x \arrow[r] & y \end{tikzcd}$$

$$x \perp\!\!\!\perp \epsilon_y$$

Two variable case

True model

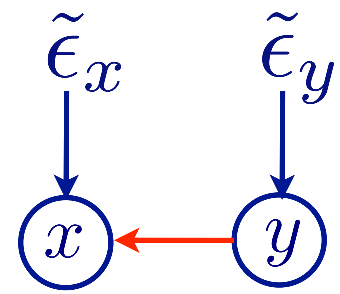
$$y = \beta x + \epsilon_y$$



$$x \perp\!\!\!\perp \epsilon_y$$

Backwards model

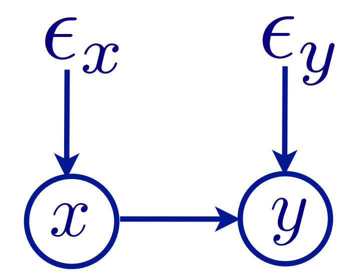
$$x = \theta y + \tilde{\epsilon}_x$$



Two variable case

True model

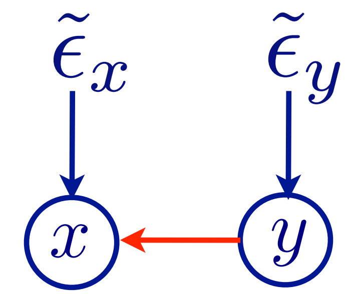
$$y = \beta x + \epsilon_y$$



$$x \perp\!\!\!\perp \epsilon_y$$

Backwards model

$$x = \theta y + \tilde{\epsilon}_x$$

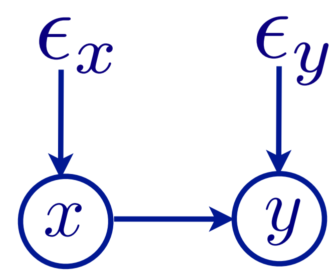


$$y \perp\!\!\!\perp \tilde{\epsilon}_x$$

Two variable case

True model

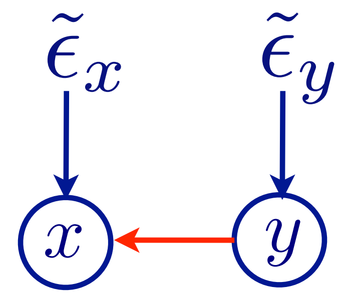
$$y = \beta x + \epsilon_y$$



$$x \perp\!\!\!\perp \epsilon_y$$

Backwards model

$$x = \theta y + \tilde{\epsilon}_x$$



$$y \perp\!\!\!\perp \tilde{\epsilon}_x$$

$$\begin{aligned}
\tilde{\epsilon}_x &= x - \theta y \\
&= x - \theta(\beta x + \epsilon_y) \\
&= (1 - \theta\beta)x - \theta\epsilon_y
\end{aligned}$$

Two variable case

True model

$$y = \beta x + \epsilon_y$$

$$\begin{tikzcd} \epsilon_x \arrow[d] & \epsilon_y \arrow[d] \\ x \arrow[r] & y \end{tikzcd}$$

$$x \perp\!\!\!\perp \epsilon_y$$

Backwards model

$$x = \theta y + \tilde{\epsilon}_x$$

$$\begin{tikzcd} \tilde{\epsilon}_x \arrow[d] & \tilde{\epsilon}_y \arrow[d] \\ x & y \arrow[l] \end{tikzcd}$$

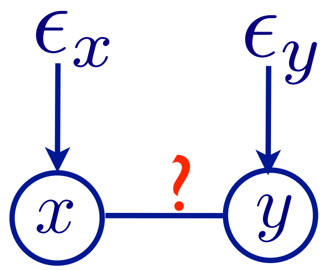
$$y \perp\!\!\!\perp \tilde{\epsilon}_x$$

$$\begin{aligned} \tilde{\epsilon}_x &= x - \theta y \\ &= x - \theta(\beta x + \epsilon_y) \\ &= (1 - \theta\beta)x - \theta\epsilon_y \end{aligned}$$

Why Normals are unusual

Forwards model $y = \beta x + \epsilon_y$

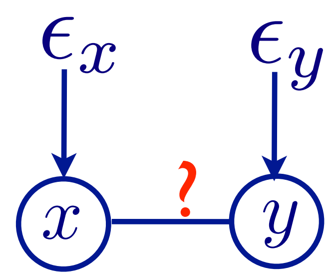
For backwards model $\tilde{\epsilon}_x = (1 - \theta\beta)x - \theta\epsilon_y$



Why Normals are unusual

Forwards model $y = \beta x + \epsilon_y$

For backwards model $\tilde{\epsilon}_x = (1 - \theta\beta)x - \theta\epsilon_y$



Theorem 1 (Darmois-Skitovich) *Let $X_1, \ldots, X_n$ be independent, non-degenerate random variables. If for two linear combinations*

$$
\begin{aligned}
l_1 &= a_1 X_1 + \ldots + a_n X_n, \quad a_i \neq 0 \\
l_2 &= b_1 X_1 + \ldots + b_n X_n, \quad b_i \neq 0
\end{aligned}
$$

are independent, then each X_i is normally distributed.

algorithm/ assumption
Markov faithfulness causal sufficiency acyclicity parametric assumption
output

algorithm/ assumption	PC / GES	FCI	CCD
Markov	✓	✓	✓
faithfulness	✓	✓	✓
causal sufficiency	✓	✗	✓
acyclicity	✓	✓	✗
parametric assumption	✗	✗	✗
output	Markov equivalence	PAG	PAG

algorithm/ assumption	PC / GES	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM
Markov	✓	✓	✓	✓	✓	✓
faithfulness	✓	✓	✓	✗	✓	~
causal sufficiency	✓	✗	✓	✓	✗	✓
acyclicity	✓	✓	✗	✓	✓	✗
parametric assumption	✗	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian
output	Markov equivalence	PAG	PAG	unique DAG	set of DAGs	set of graphs

Limitations

*For **linear Gaussian** and for **multinomial** causal relations, an algorithm that identifies the Markov equivalence class of the true model is complete.*

(Pearl & Geiger 1988, Meek 1995)

Limitations

*For linear **Gaussian** and for **multinomial** causal relations, an algorithm that identifies the Markov equivalence class of the true model is complete.*

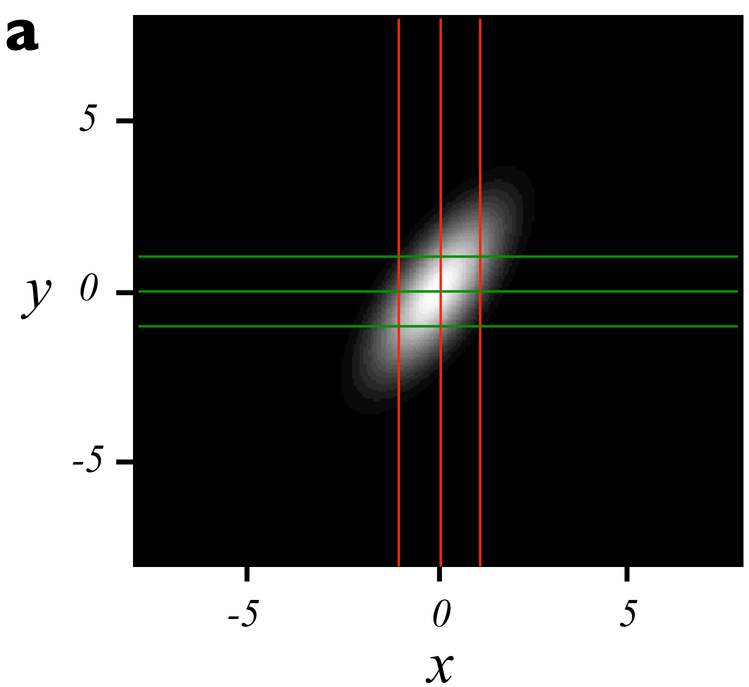
(Pearl & Geiger 1988, Meek 1995)

Bivariate Linear Gaussian case

True model

$$x = \epsilon_x$$
$$y = x + \epsilon_y$$

$$\epsilon_x, \epsilon_y \sim \text{indep. Gaussian}$$



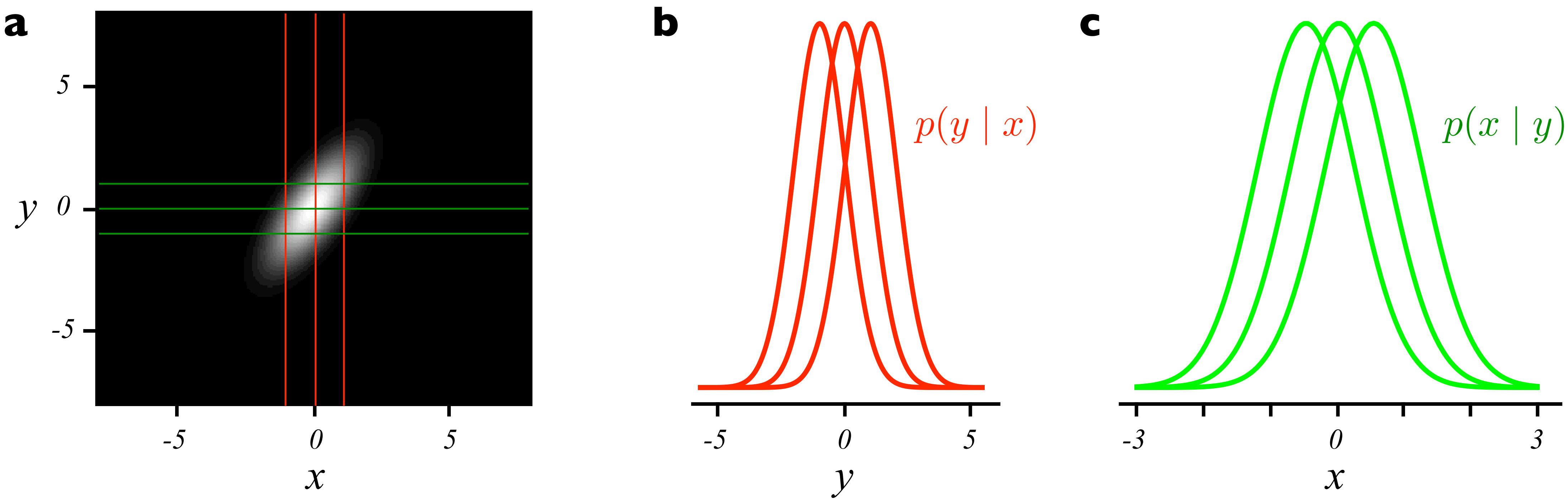
(graphics from Hoyer et al. 2009)

Bivariate Linear Gaussian case

True model

$$x = \epsilon_x$$
$$y = x + \epsilon_y$$

$$\epsilon_x, \epsilon_y \sim \text{indep. Gaussian}$$



Forwards (true) model

Backwards model

(graphics from Hoyer et al. 2009)

Continuous additive noise models

$$x_j = f_j(pa(x_j)) + \epsilon_j$$

Continuous additive noise models

$$x_j = f_j(pa(x_j)) + \epsilon_j$$

- If $f_j(.)$ is linear, then non-Gaussian errors are required for identifiability

Continuous additive noise models

$$x_j = f_j(pa(x_j)) + \epsilon_j$$

- If $f_j(.)$ is linear, then non-Gaussian errors are required for identifiability

➡ What if the errors are Gaussian, but $f_j(.)$ is non-linear?

Continuous additive noise models

$$x_j = f_j(pa(x_j)) + \epsilon_j$$

- If $f_j(.)$ is linear, then non-Gaussian errors are required for identifiability

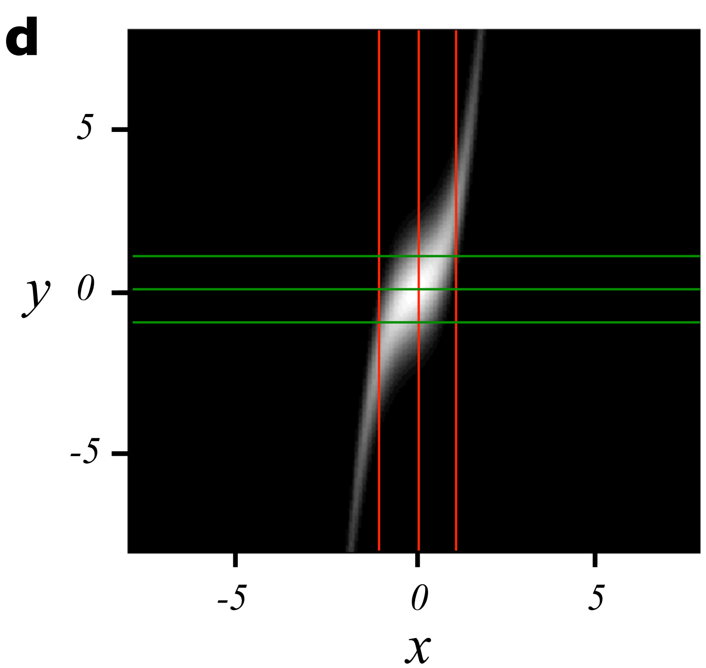
➡ What if the errors are Gaussian, but $f_j(.)$ is non-linear?

➡ More generally, under what circumstances is the causal structure represented by this class of models identifiable?

Bivariate non-linear Gaussian additive noise model

True model

$$x = \epsilon_x \qquad \epsilon_x, \epsilon_y \sim \text{indep. Gaussian}$$
$$y = x + x^3 + \epsilon_y$$

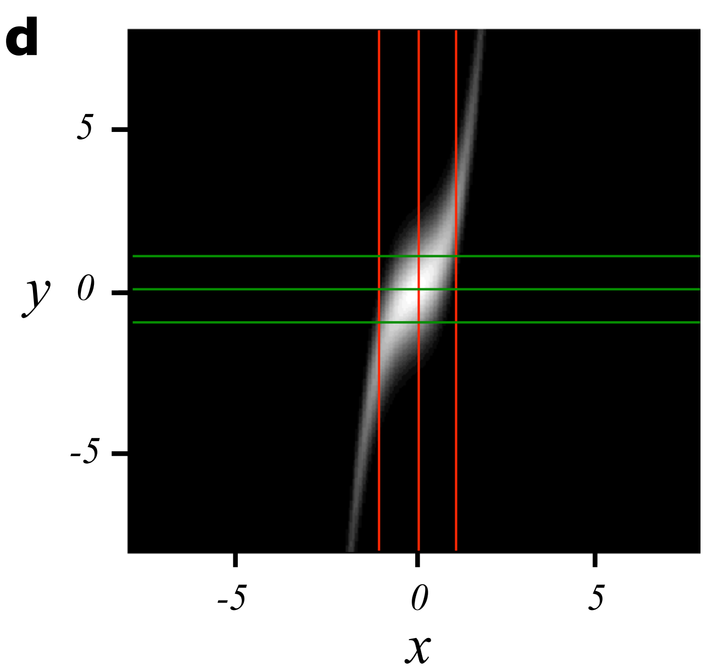


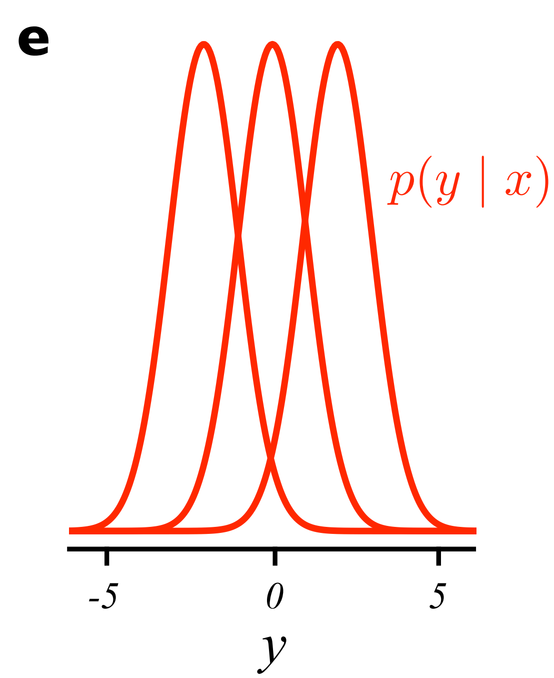
(graphics from Hoyer et al. 2009)

Bivariate non-linear Gaussian additive noise model

True model

$$x = \epsilon_x \qquad\qquad \epsilon_x, \epsilon_y \sim \text{indep. Gaussian}$$
$$y = x + x^3 + \epsilon_y$$



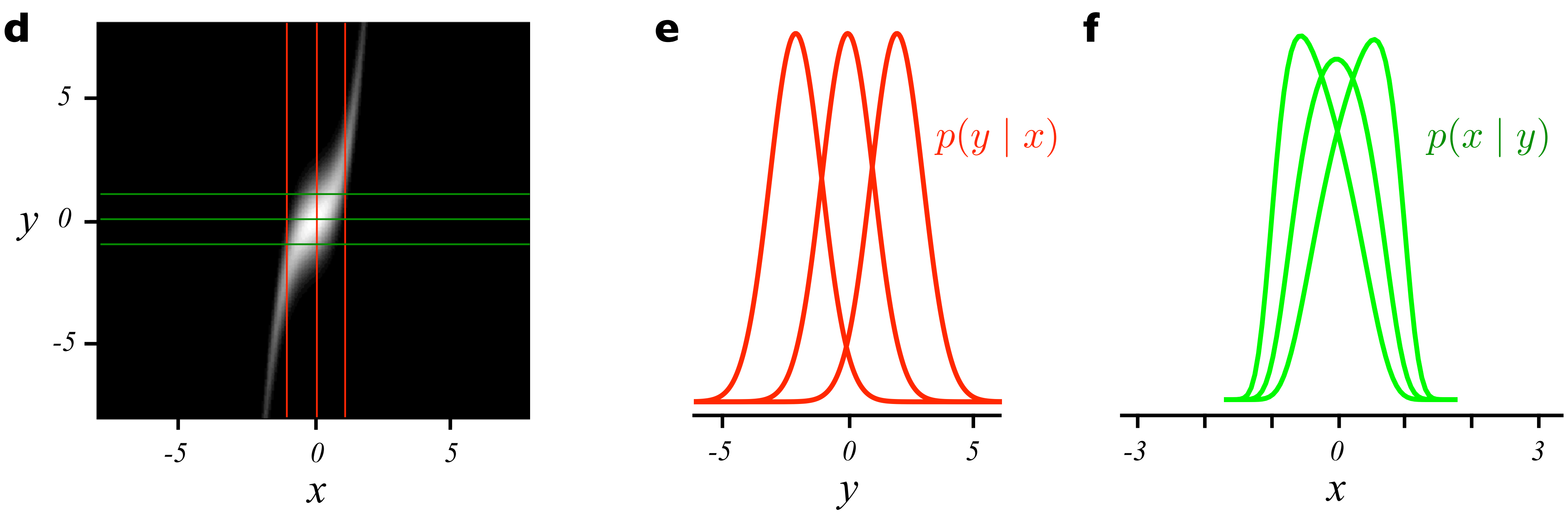


Forwards
(true) model

(graphics from Hoyer et al. 2009)

Bivariate non-linear Gaussian additive noise model

True model $$x = \epsilon_x \qquad \epsilon_x, \epsilon_y \sim \text{indep. Gaussian}$$
$$y = x + x^3 + \epsilon_y$$



Forwards (true) model

Backwards model

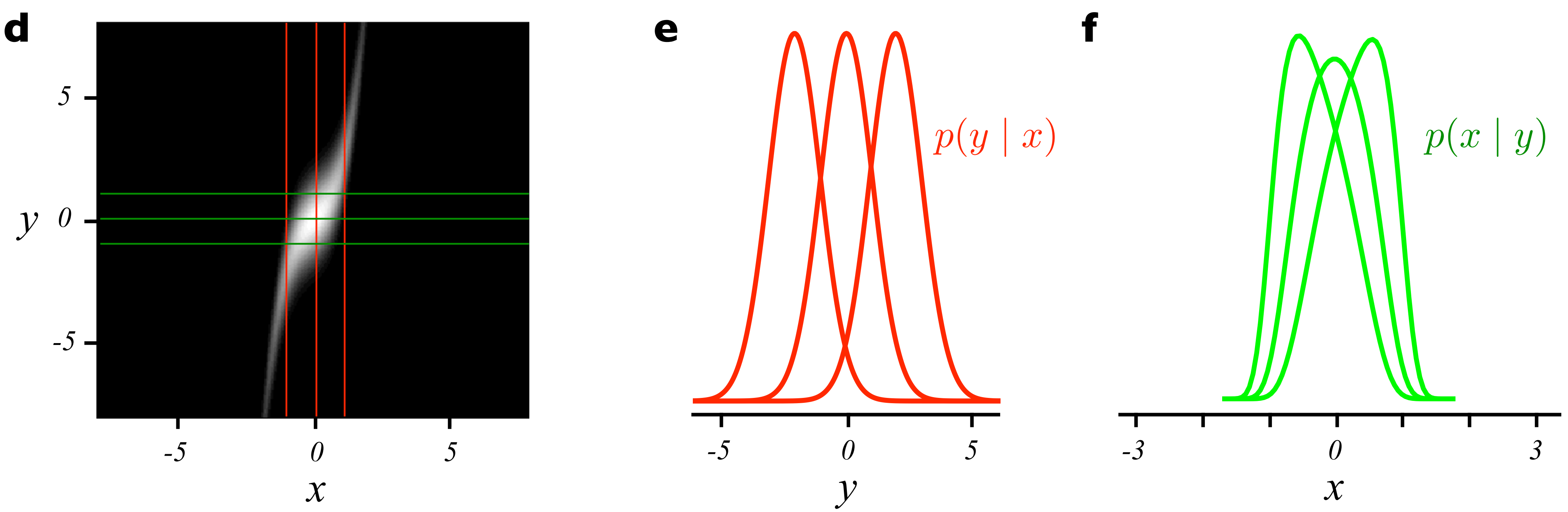
(graphics from Hoyer et al. 2009)

Bivariate non-linear Gaussian additive noise model

True model

$$x = \epsilon_x$$
$$y = x + x^3 + \epsilon_y$$

$\epsilon_x, \epsilon_y \sim$ indep. Gaussian



Forwards (true) model

Backwards model

$$x = g(y) + \tilde{\epsilon}_x$$
$$y \not\perp\!\!\!\perp \tilde{\epsilon}_x$$

(graphics from Hoyer et al. 2009)

General non-linear additive noise models

Hoyer Condition (HC): *Technical condition on the relation between the function, the noise distribution and the parent distribution that, if satisfied, permits a backward model.*

General non-linear additive noise models

Hoyer Condition (HC): *Technical condition on the relation between the function, the noise distribution and the parent distribution that, if satisfied, permits a backward model.*

- If the error terms are **Gaussian**, then the **only** functional form that **satisfies** HC is **linearity**, otherwise the model is **identifiable**.

General non-linear additive noise models

Hoyer Condition (HC): *Technical condition on the relation between the function, the noise distribution and the parent distribution that, if satisfied, permits a backward model.*

- If the error terms are **Gaussian**, then the **only** functional form that **satisfies** HC is **linearity**, otherwise the model is **identifiable**.
- If the errors are **non-Gaussian**, then there are (rather contrived) functions that satisfy HC, but **in general identifiability is guaranteed**.

General non-linear additive noise models

Hoyer Condition (HC): *Technical condition on the relation between the function, the noise distribution and the parent distribution that, if satisfied, permits a backward model.*

- If the error terms are **Gaussian**, then the **only** functional form that **satisfies** HC is **linearity**, otherwise the model is **identifiable**.
- If the errors are **non-Gaussian**, then there are (rather contrived) functions that satisfy HC, but **in general identifiability is guaranteed**.
  - this generalizes to multiple variables (assuming minimality*)!

General non-linear additive noise models

Hoyer Condition (HC): *Technical condition on the relation between the function, the noise distribution and the parent distribution that, if satisfied, permits a backward model.*

- If the error terms are **Gaussian**, then the **only** functional form that **satisfies** HC is **linearity**, otherwise the model is **identifiable**.
- If the errors are **non-Gaussian**, then there are (rather contrived) functions that satisfy HC, but **in general identifiability is guaranteed**.
  - this generalizes to multiple variables (assuming minimality*)!
  - extension to discrete additive noise models

General non-linear additive noise models

Hoyer Condition (HC): *Technical condition on the relation between the function, the noise distribution and the parent distribution that, if satisfied, permits a backward model.*

- If the error terms are **Gaussian**, then the **only** functional form that **satisfies** HC is **linearity**, otherwise the model is **identifiable**.
- If the errors are **non-Gaussian**, then there are (rather contrived) functions that satisfy HC, but **in general identifiability is guaranteed**.
  - this generalizes to multiple variables (assuming minimality*)!
  - extension to discrete additive noise models
- If the function is **linear**, but the error terms **non-Gaussian**, then one can't fit a linear backwards model (Lingam), but there are cases where **one can fit a non-linear backwards model**

algorithm/ assumptions	PC / GES	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM
Markov	✓	✓	✓	✓	✓	✓
faithfulness	✓	✓	✓	✗	✓	~
causal sufficiency	✓	✗	✓	✓	✗	✓
acyclicity	✓	✓	✗	✓	✓	✗
parametric assumption	✗	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian
output	Markov equivalence	PAG	PAG	unique DAG	set of DAGs	set of graphs

algorithm/ assumptions	PC / GES	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM	non-linear additive noise
Markov	✓	✓	✓	✓	✓	✓	✓
faithfulness	✓	✓	✓	✗	✓	~	minimality
causal sufficiency	✓	✗	✓	✓	✗	✓	✓
acyclicity	✓	✓	✗	✓	✓	✗	✓
parametric assumption	✗	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian	non-linear additive noise
output	Markov equivalence	PAG	PAG	unique DAG	set of DAGs	set of graphs	unique DAG

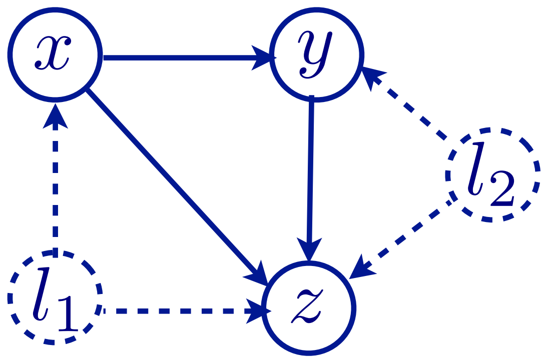
Experiments, Background Knowledge and all the other Jazz

Experiments, Background Knowledge and all the other Jazz

- how to integrate data from experiments?

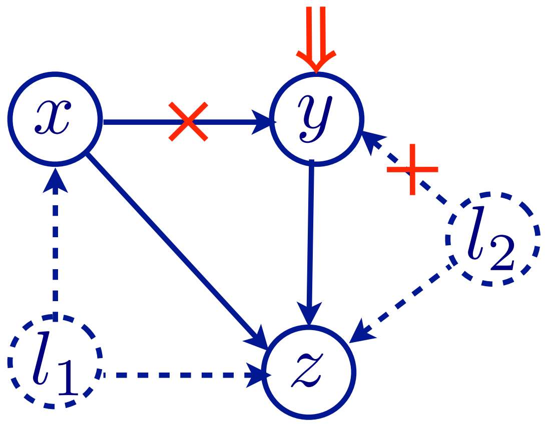
Experiments, Background Knowledge and all the other Jazz

- how to integrate data from experiments?

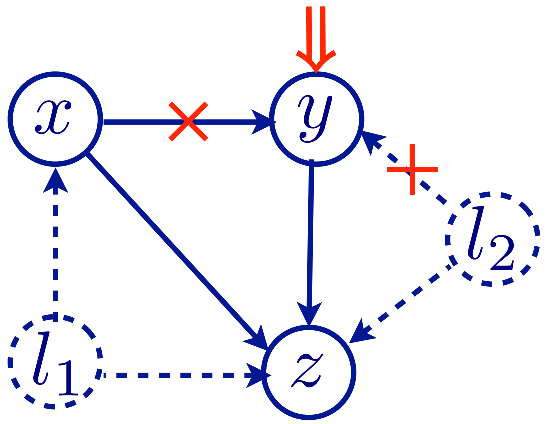


Experiments, Background Knowledge and all the other Jazz

- how to integrate data from experiments?



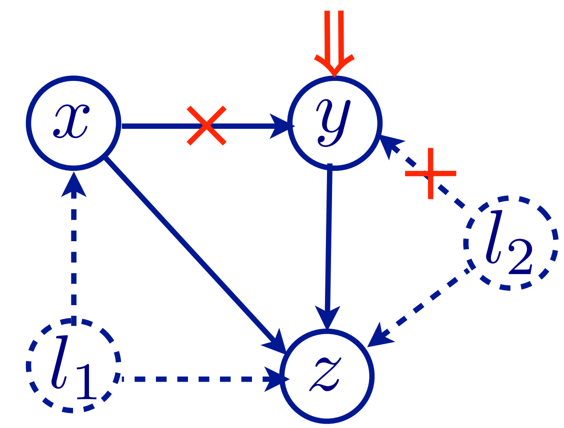
Experiments, Background Knowledge and all the other Jazz



- how to integrate data from experiments?

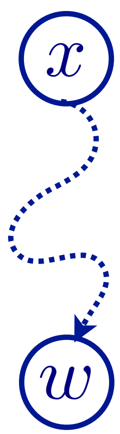
- how to include background knowledge?

Experiments, Background Knowledge and all the other Jazz



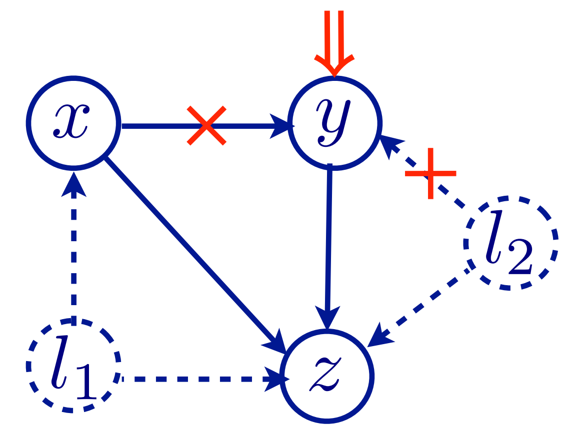
- how to integrate data from experiments?

- how to include background knowledge?



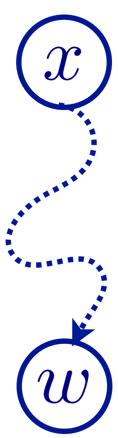
pathways

Experiments, Background Knowledge and all the other Jazz

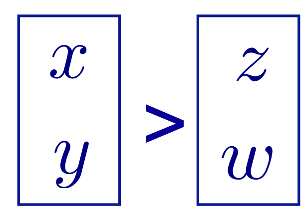


- how to integrate data from experiments?

- how to include background knowledge?



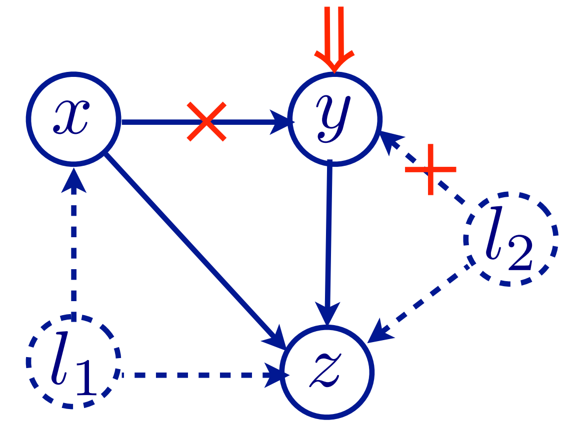
pathways



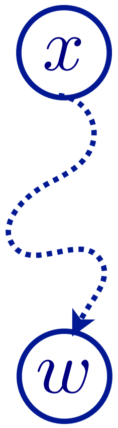
tier orderings

Experiments, Background Knowledge and all the other Jazz

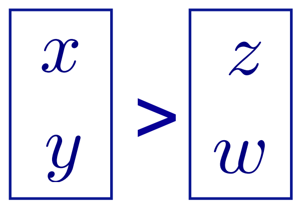
- how to integrate data from experiments?



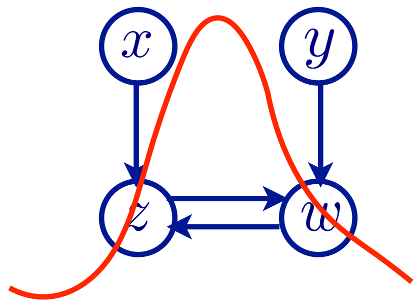
- how to include background knowledge?



pathways



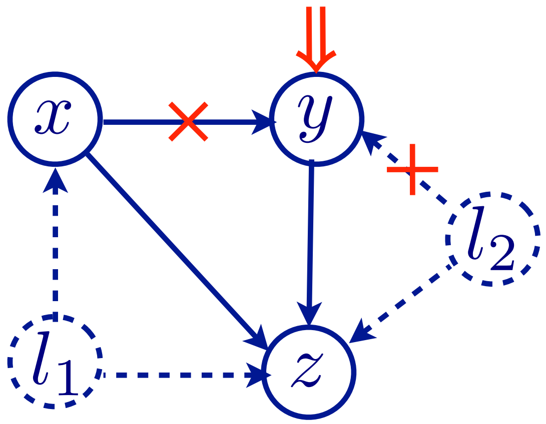
tier orderings



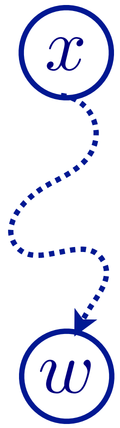
"priors"

Experiments, Background Knowledge and all the other Jazz

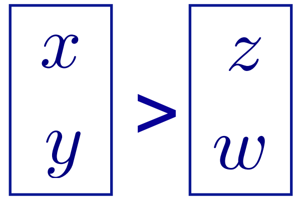
- how to integrate data from experiments?



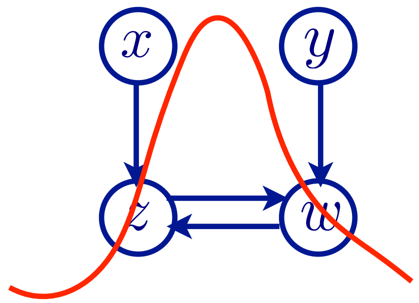
- how to include background knowledge?



pathways



tier orderings

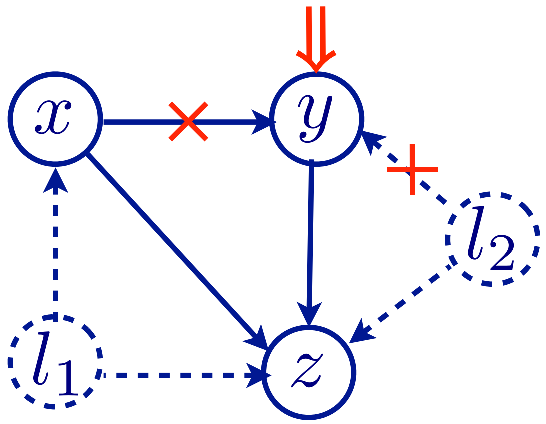


"priors"

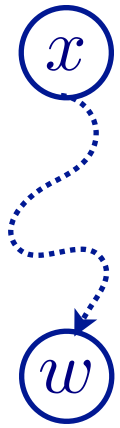
- specific search space restrictions

Experiments, Background Knowledge and all the other Jazz

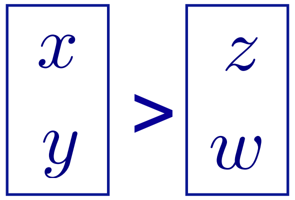
- how to integrate data from experiments?



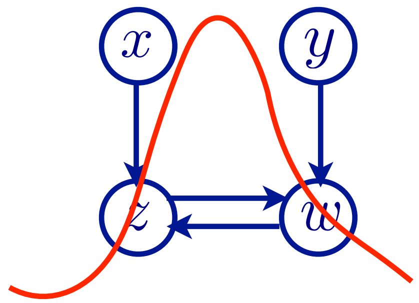
- how to include background knowledge?



pathways

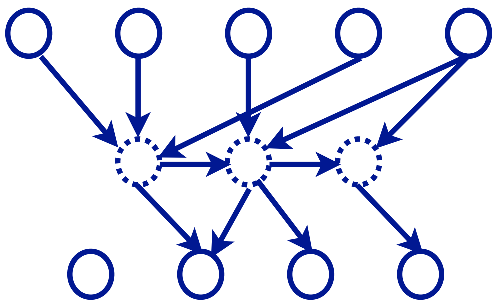


tier orderings



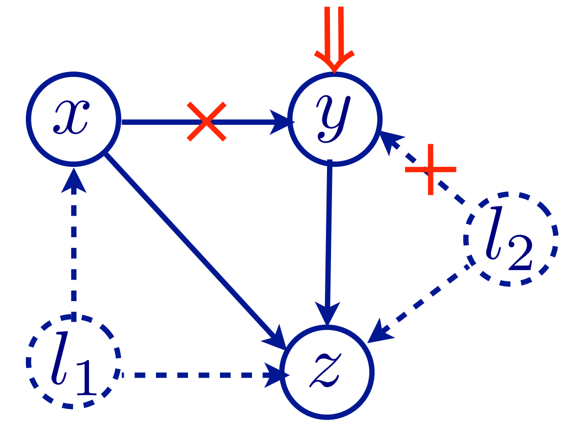
"priors"

- specific search space restrictions

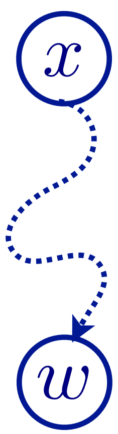

biological settings

Experiments, Background Knowledge and all the other Jazz

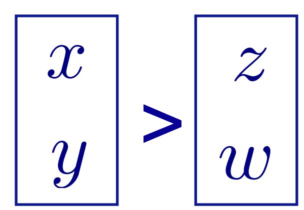
- how to integrate data from experiments?



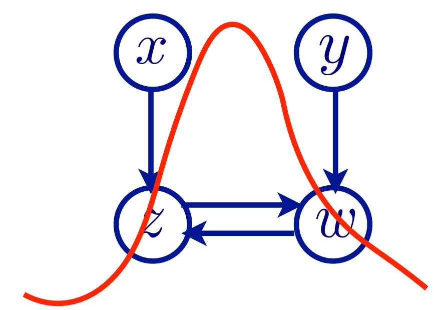
- how to include background knowledge?



pathways

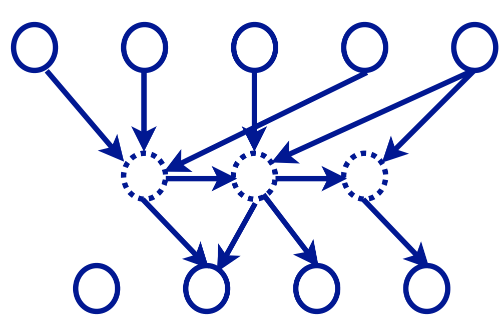


tier orderings

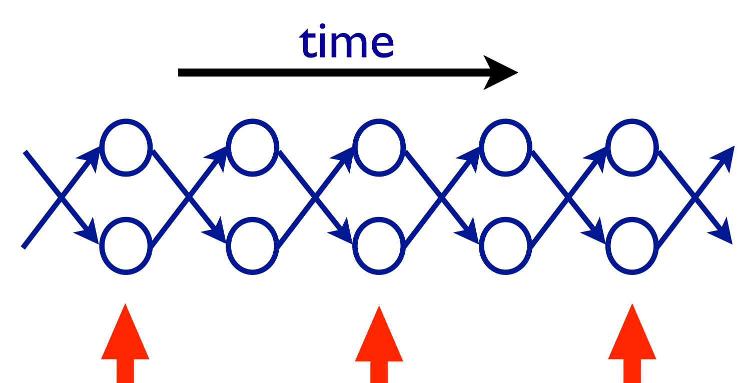


“priors”

- specific search space restrictions



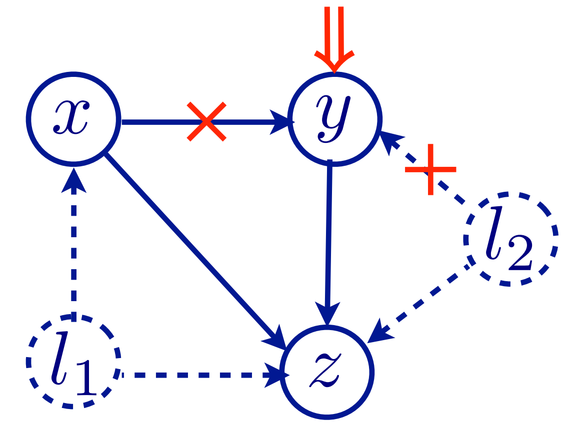
biological settings



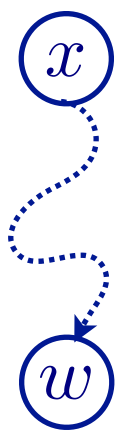
subsampled time series

Experiments, Background Knowledge and all the other Jazz

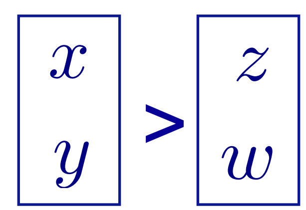
- how to integrate data from experiments?



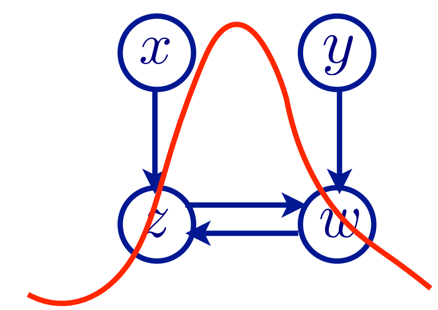
- how to include background knowledge?



pathways

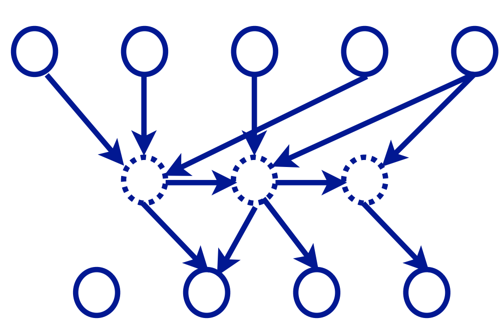


tier orderings



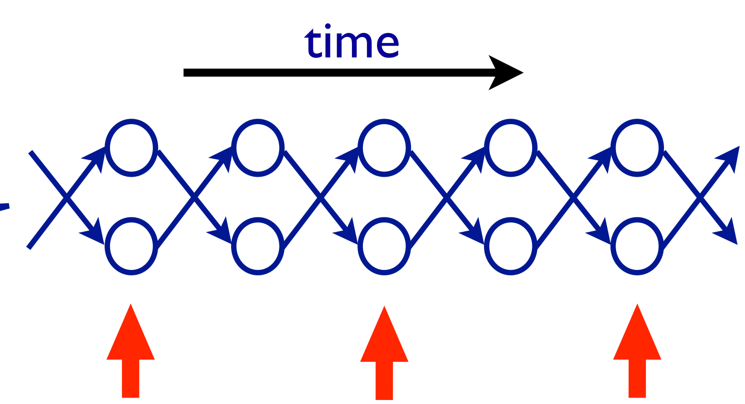
"priors"

- specific search space restrictions



biological settings

Tank talk

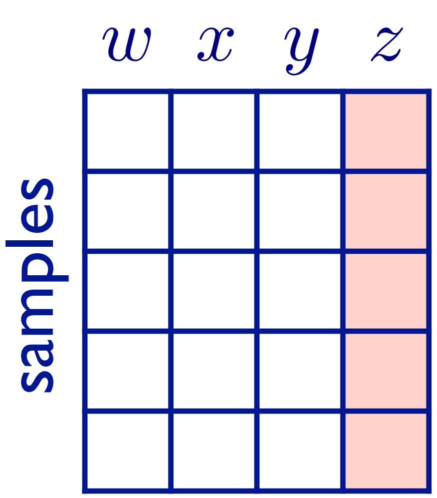


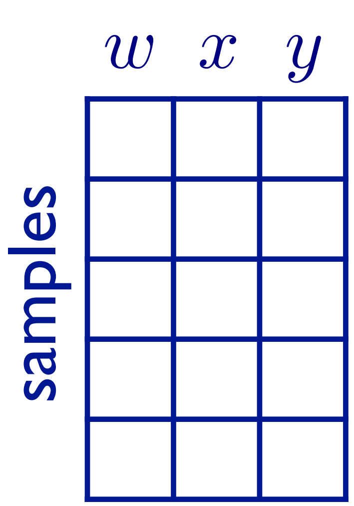
subsampled time series

High-Level

High-Level

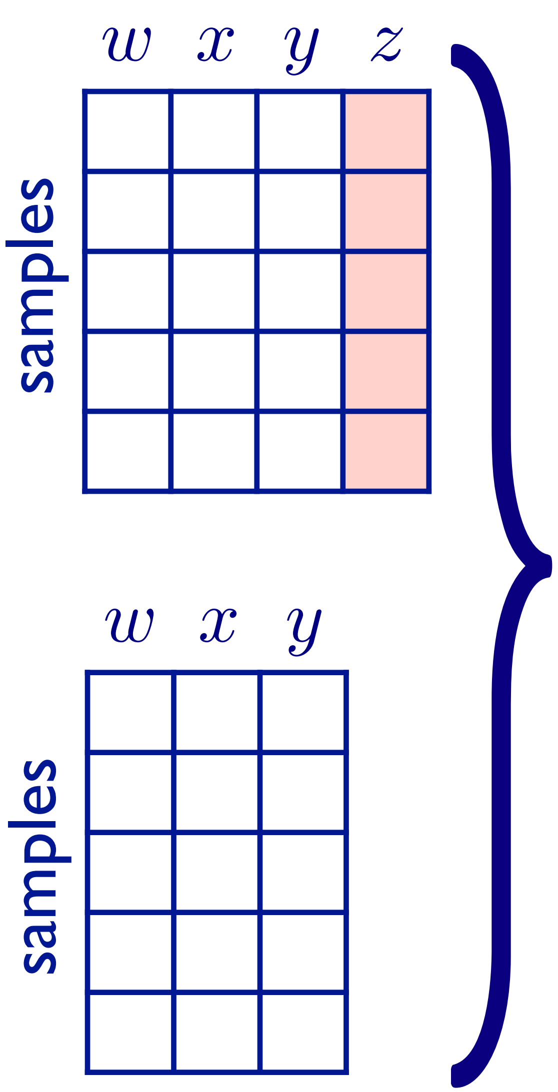
data
sample





High-Level

data
sample



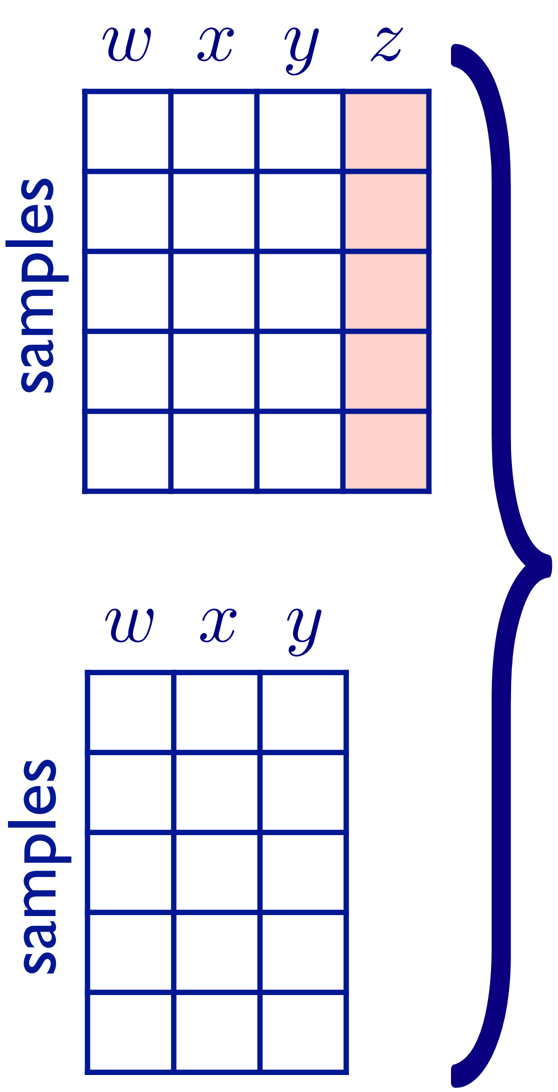
(in)dependence
constraints

$$x \not\perp\!\!\!\perp y \mid \mathbf{C} \,||\, \mathbf{J}$$

High-Level

data
sample

assumptions, e.g.
- causal Markov
- causal faithfulness
- etc.

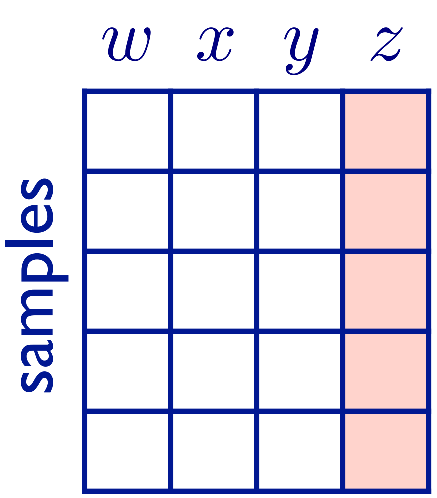


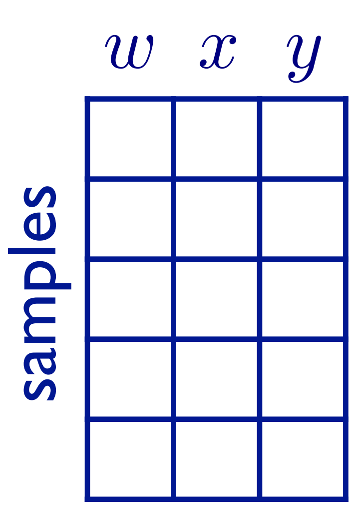
(in)dependence
constraints

$$x \not\perp\!\!\!\perp y \,|\, \mathbf{C} \,||\, \mathbf{J}$$

High-Level

data sample





assumptions, e.g.

- causal Markov
- causal faithfulness
- etc.

background knowledge, e.g.

- pathways
- tier ordering
- "priors"
- etc.

(in)dependence constraints

$$x \not\perp\!\!\!\perp y \,|\, \mathbf{C} \,||\, \mathbf{J}$$

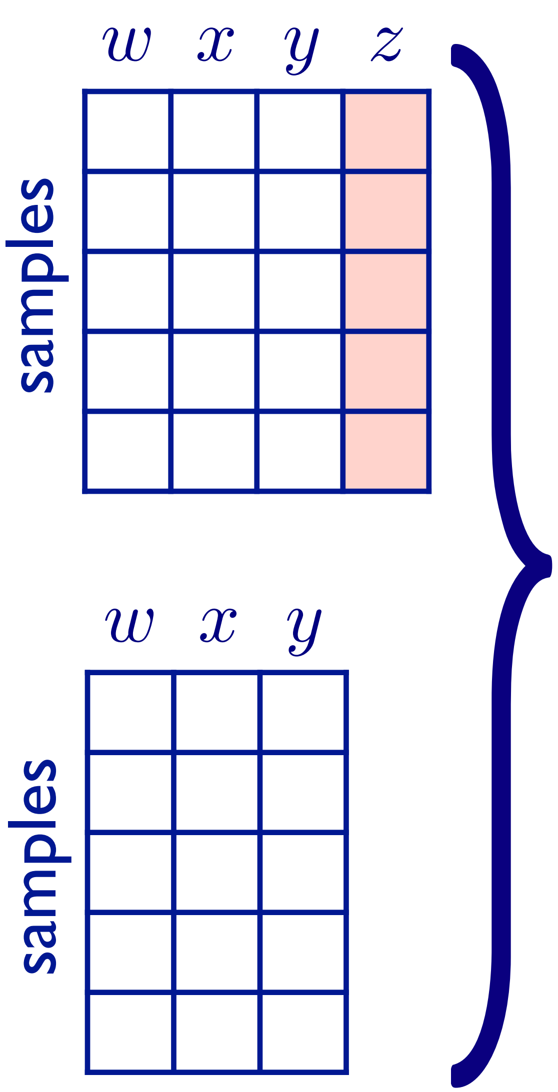
High-Level

data sample

assumptions, e.g.
- causal Markov
- causal faithfulness
- etc.

background knowledge, e.g.
- pathways
- tier ordering
- “priors”
- etc.

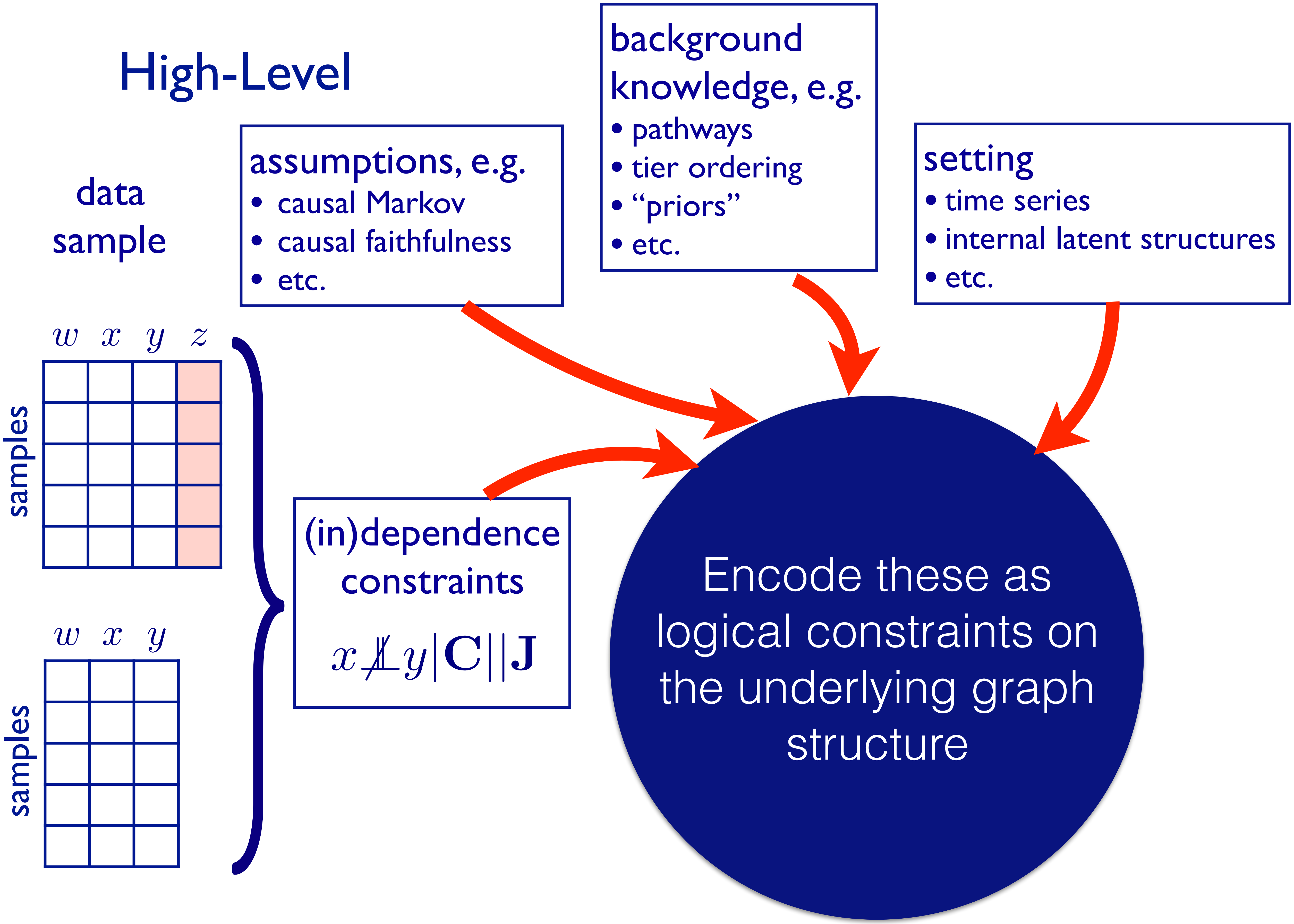
setting
- time series
- internal latent structures
- etc.



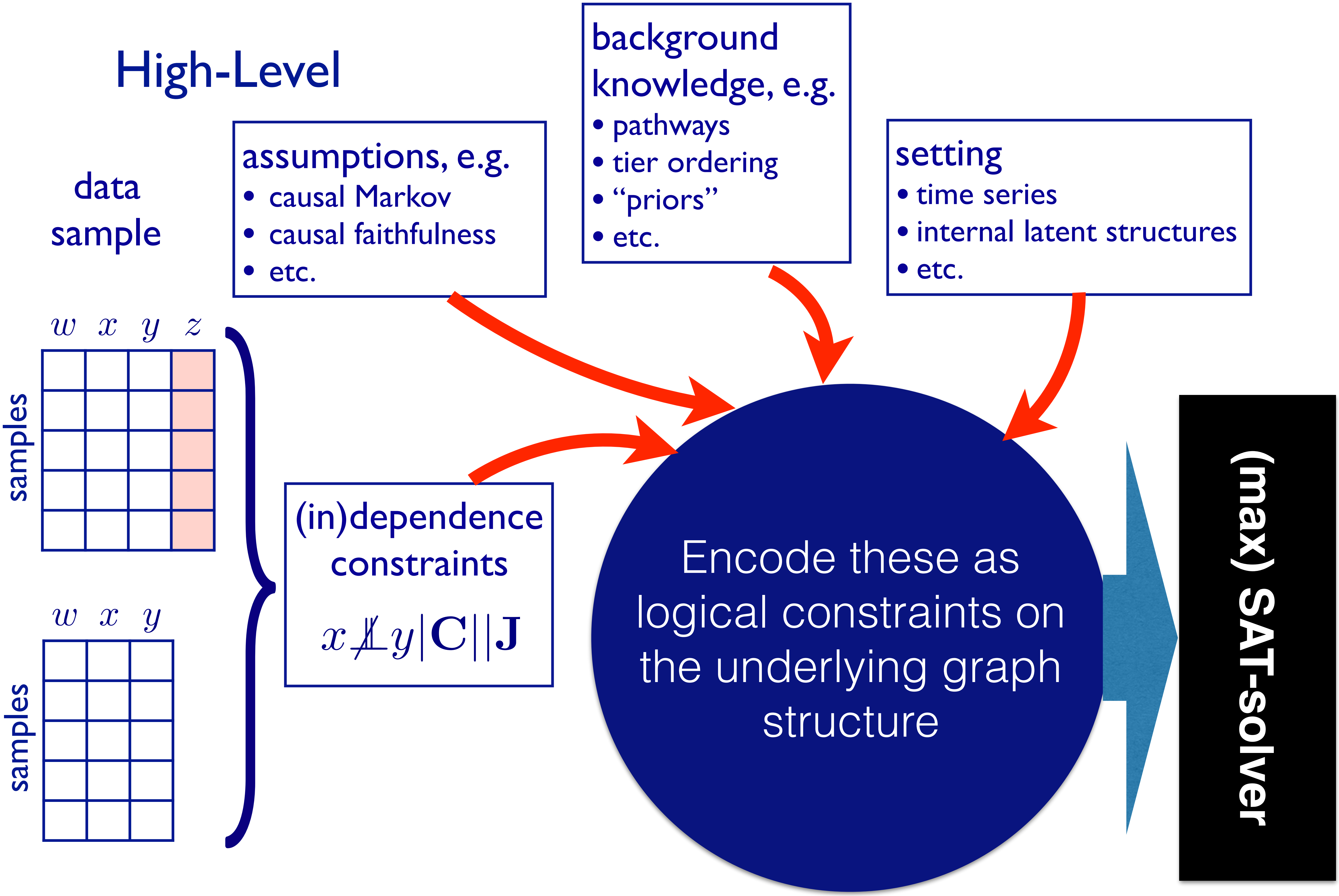
(in)dependence constraints

$$x \not\perp\!\!\!\perp y \,|\, \mathbf{C} \,||\, \mathbf{J}$$

High-Level


data sample
assumptions, e.g.
• causal Markov
• causal faithfulness
• etc.
background knowledge, e.g.
• pathways
• tier ordering
• “priors”
• etc.
setting
• time series
• internal latent structures
• etc.
w x y z
samples
w x y
samples
(in)dependence constraints
$x \not\perp\!\!\!\perp y \,|\, \mathbf{C} \,||\, \mathbf{J}$
Encode these as logical constraints on the underlying graph structure

High-Level


data sample
w x y z
samples
w x y
samples
assumptions, e.g.
• causal Markov
• causal faithfulness
• etc.
background knowledge, e.g.
• pathways
• tier ordering
• "priors"
• etc.
setting
• time series
• internal latent structures
• etc.
(in)dependence constraints
$x \not\perp\!\!\!\perp y \mid \mathbf{C} \mid\mid \mathbf{J}$
Encode these as logical constraints on the underlying graph structure
(max) SAT-solver

SAT-based Causal Discovery

- **Formulate the independence constraints in propositional logic**

$$x \perp\!\!\!\perp y \iff \neg A \wedge \neg B \ldots$$

$$A = \text{`}x \rightarrow y \text{ is present'}$$

SAT-based Causal Discovery

- **Formulate the independence constraints in propositional logic**

$$x \perp\!\!\!\perp y \iff \neg A \wedge \neg B \ldots$$

$$A = \text{`}x \to y \text{ is present'}$$

- **Encode the constraints into one formula.**

$$\neg A \wedge \neg B \wedge \neg(C \wedge D) \wedge \neg \ldots$$

SAT-based Causal Discovery

- **Formulate the independence constraints in propositional logic**

$$x \perp\!\!\!\perp y \iff \neg A \wedge \neg B \ldots$$

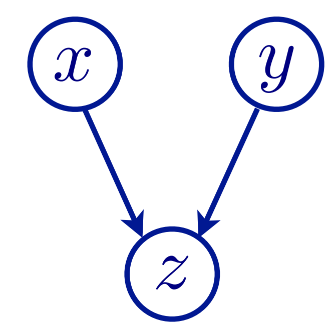
$$A = \text{`}x \rightarrow y \text{ is present'}$$

- **Encode the constraints into one formula.**

$$\neg A \wedge \neg B \wedge \neg (C \wedge D) \wedge \neg \ldots$$

- **Find satisfying assignments using a SAT-solver**

$$A = false$$
$$B = false \iff$$
$$\ldots$$



SAT-based Causal Discovery

- **Formulate the independence constraints in propositional logic**

$$x \perp\!\!\!\perp y \iff \neg A \wedge \neg B \ldots$$

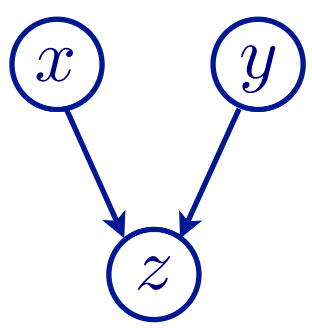
$$A = \text{`}x \rightarrow y \text{ is present'}$$

- **Encode the constraints into one formula.**

$$\neg A \wedge \neg B \wedge \neg (C \wedge D) \wedge \neg \ldots$$

- **Find satisfying assignments using a SAT-solver**

$$\begin{array}{l} A = false \\ B = false \\ \ldots \end{array} \iff$$



➡ very general setting (allows for cycles and latents) and trivially complete

SAT-based Causal Discovery

- **Formulate the independence constraints in propositional logic**

$$x \perp\!\!\!\perp y \iff \neg A \wedge \neg B \dots$$

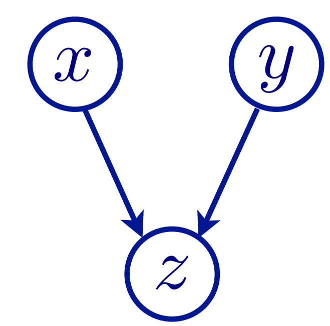
$$A = \text{‘}x \rightarrow y \text{ is present’}$$

- **Encode the constraints into one formula.**

$$\neg A \wedge \neg B \wedge \neg(C \wedge D) \wedge \neg \dots$$

- **Find satisfying assignments using a SAT-solver**

$$\begin{array}{l} A = false \\ B = false \\ \dots \end{array} \iff$$



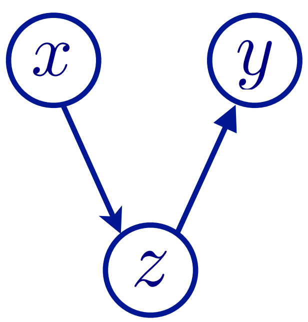
➡ very general setting (allows for cycles and latents) and trivially complete

➡ **BUT**: erroneous test results induce conflicting constraints: UNsatisfiable

Conflicts and Errors

- **Statistical independence tests produce errors**

➡ **Conflict:** **no graph can produce the set of constraints**

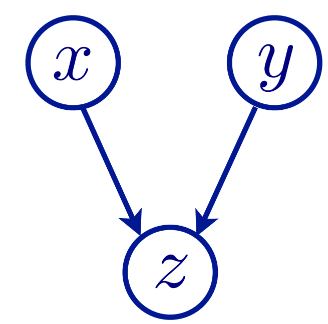


constraints

$$\Longleftarrow \boxed{x \not\perp\!\!\!\perp y}$$

$$x \perp\!\!\!\perp y$$
$$x \not\perp\!\!\!\perp z$$
$$y \not\perp\!\!\!\perp z$$
$$x \perp\!\!\!\perp y \mid z$$

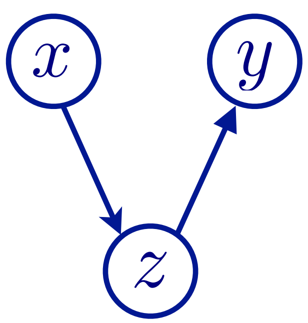
$$\boxed{x \not\perp\!\!\!\perp y \mid z} \Longrightarrow$$



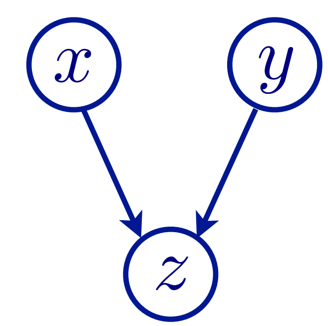
Conflicts and Errors

- **Statistical independence tests produce errors**

  ➡ **Conflict: no graph can produce the set of constraints**



constraints	weight
$x \perp\!\!\!\perp y$	500
$x \not\perp\!\!\!\perp z$	3000
$y \not\perp\!\!\!\perp z$	2500
$x \perp\!\!\!\perp y \mid z$	250

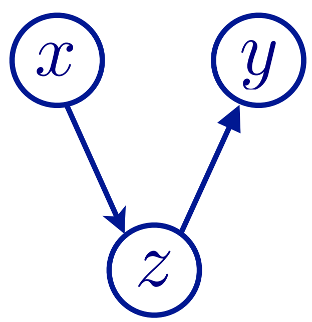


Conflicts and Errors

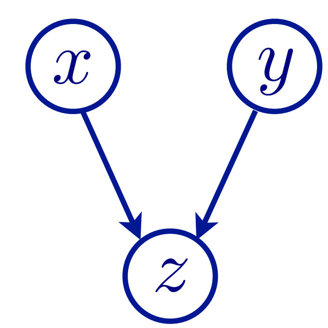
Sridhar talk

- **Statistical independence tests produce errors**

➡ **Conflict: no graph can produce the set of constraints**



constraints	weight
$x \perp\!\!\!\perp y$	500
$x \not\perp\!\!\!\perp z$	3000
$y \not\perp\!\!\!\perp z$	2500
$x \perp\!\!\!\perp y \mid z$	250



Constraint Satisfaction Approach

- **INPUT: (in)dependence constraints weighted according to reliability**

$$\min_{G} \sum_{k\,:\,\text{constraint } k \text{ is } \textbf{not} \text{ satisfied by } G} w(k)$$

- **OUTPUT: a graph G that minimizes the cost**

[Hyttinen et al. 2014]

Constraint Satisfaction Approach

- **INPUT: (in)dependence constraints weighted according to reliability**

$$\min_G \sum_{k\,:\,\text{constraint } k \text{ is } \textbf{not} \text{ satisfied by } G} w(k)$$

- **OUTPUT: a graph G that minimizes the cost**
- Answer Set Programming (ASP)
  - solver used: Clingo
  - finds globally optimal weighted maxSAT solution

[Hyttinen et al. 2014]

Constraint Satisfaction Approach

- **INPUT: (in)dependence constraints weighted according to reliability**

$$\min_{G} \sum_{k\,:\,\text{constraint } k \text{ is } \textbf{not} \text{ satisfied by } G} w(k)$$

- **OUTPUT: a graph G that minimizes the cost**
- Answer Set Programming (ASP)
  - solver used: Clingo
  - finds globally optimal weighted maxSAT solution

What are suitable weights?

[Hyttinen et al. 2014]

Weighting Schemes

- Constant weights
  - unit weights for all constraint

[Hyttinen et al. 2014]

Weighting Schemes

- Constant weights
  - unit weights for all constraint
- Hard dependencies
  - only treat rejections of the null-hypothesis as hard constraints, in line with classical statistics
  - give dependences infinite weight, maximize the independences (unit weight) in light of these dependences

[Hyttinen et al. 2014]

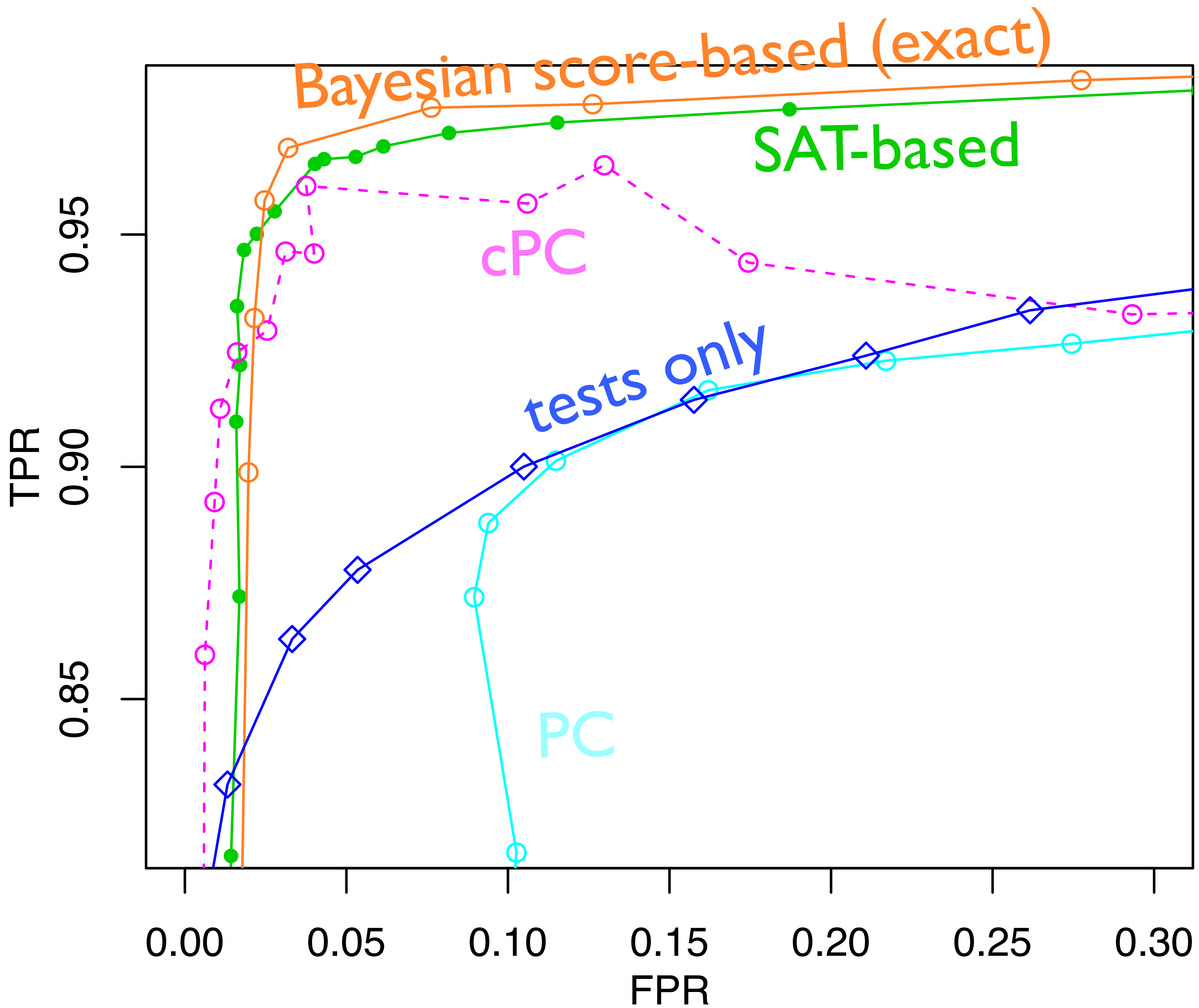
Weighting Schemes

- Constant weights
  - unit weights for all constraint
- Hard dependencies
  - only treat rejections of the null-hypothesis as hard constraints, in line with classical statistics
  - give dependences infinite weight, maximize the independences (unit weight) in light of these dependences
- Log weights
  - obtain the probability of an (in)dependence and weigh it according to the log of the probability
  - Model selection with Bayes rule:

$$\begin{array}{ccc} x \not\perp\!\!\!\perp y | C & & x \perp\!\!\!\perp y | C \\ P(x|C)P(y|x,C) & \text{vs.} & P(x|C)P(y|C) \end{array}$$

[Hyttinen et al. 2014]

Simulation 1: no cycles, no latents, linear Gaussian

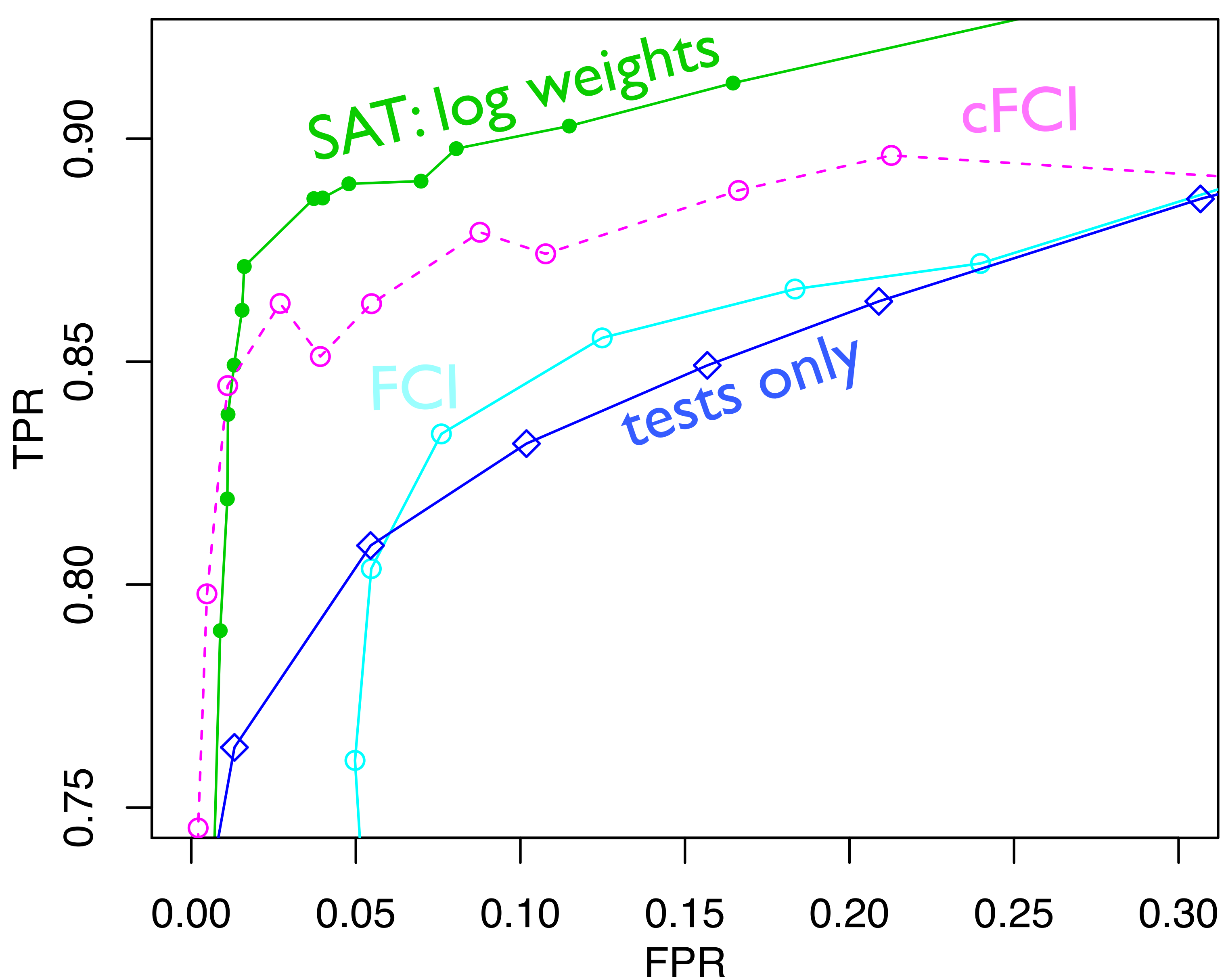


- TPR vs. FPR of all d-separation constraints of the true graph for a varying p-value cut-off
- observational data set,
  6 observed variables,
  average degree 2;
  500 samples, 200 models,
  linear Gaussian parameterization

- cPC returns a fully determined output only 58/200 times at its optimum

[Hyttinen et al. 2014]

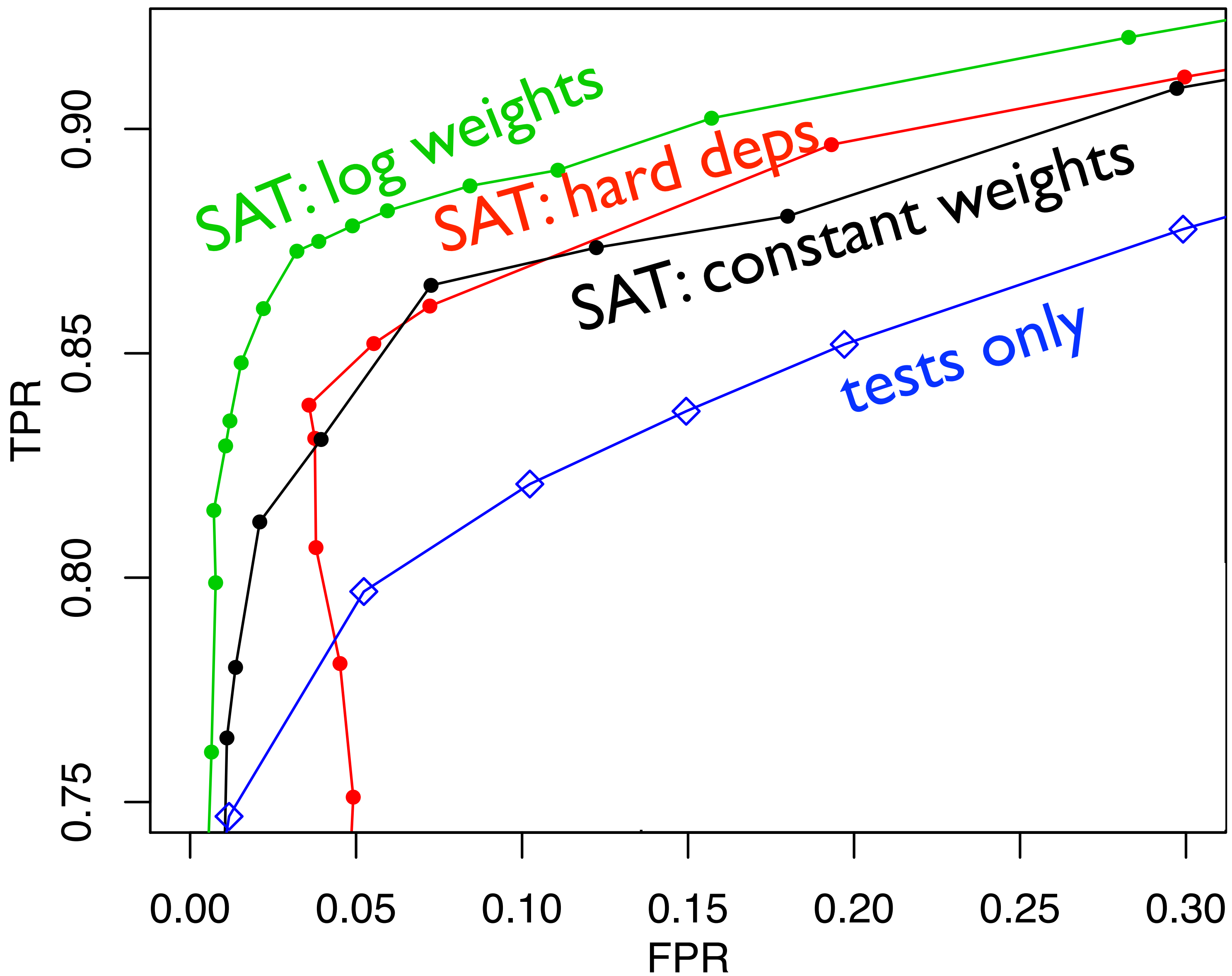
Simulation 2: no cycles, but latents



- cFCI only returns unambiguous results 61/200 times at its optimum

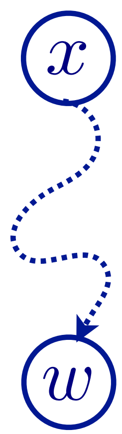
[Hyttinen et al. 2014]

Simulation 3: cycles and latents

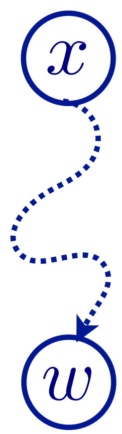


[Hyttinen et al. 2014]

Background Knowledge

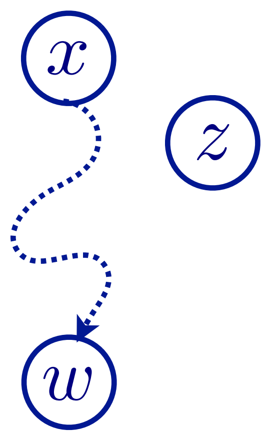


Background Knowledge



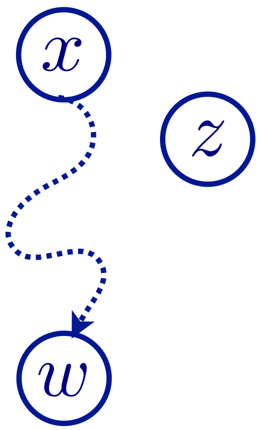
$$\iff \qquad x \not\perp\!\!\!\perp w \,||\, x$$

Background Knowledge



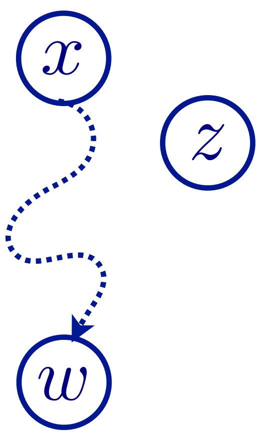
$$\Longleftrightarrow \qquad x \not\perp\!\!\!\perp w \,||\, x \; z$$

Background Knowledge



$$\Longleftrightarrow \qquad x \not\!\perp\!\!\!\perp w \,||\, x\ z \qquad weight = 0.8$$

Background Knowledge

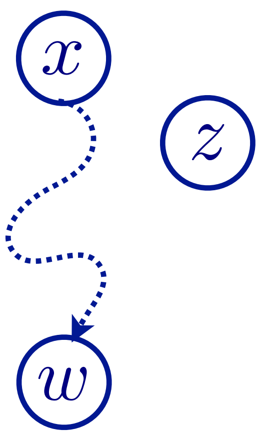


$$\iff$$

$$x \not\perp\!\!\!\perp w \,||\, x\ z \qquad weight = 0.8$$

$$\boxed{\begin{array}{c} x \\ y \end{array}} > \boxed{\begin{array}{c} z \\ w \end{array}} \iff \begin{array}{l} (x > z) \wedge (x > w) \\ \wedge\ (y > z) \wedge (y > w) \end{array}$$

Background Knowledge



$$\Longleftrightarrow \qquad x \not\perp\!\!\!\perp w \,||\, x \; z \qquad weight = 0.8$$

$$\begin{array}{|c|}\hline x \\ y \\ \hline\end{array} > \begin{array}{|c|}\hline z \\ w \\ \hline\end{array} \qquad \Longleftrightarrow \qquad (x > z) \wedge (x > w) \wedge (y > z) \wedge (y > w)$$

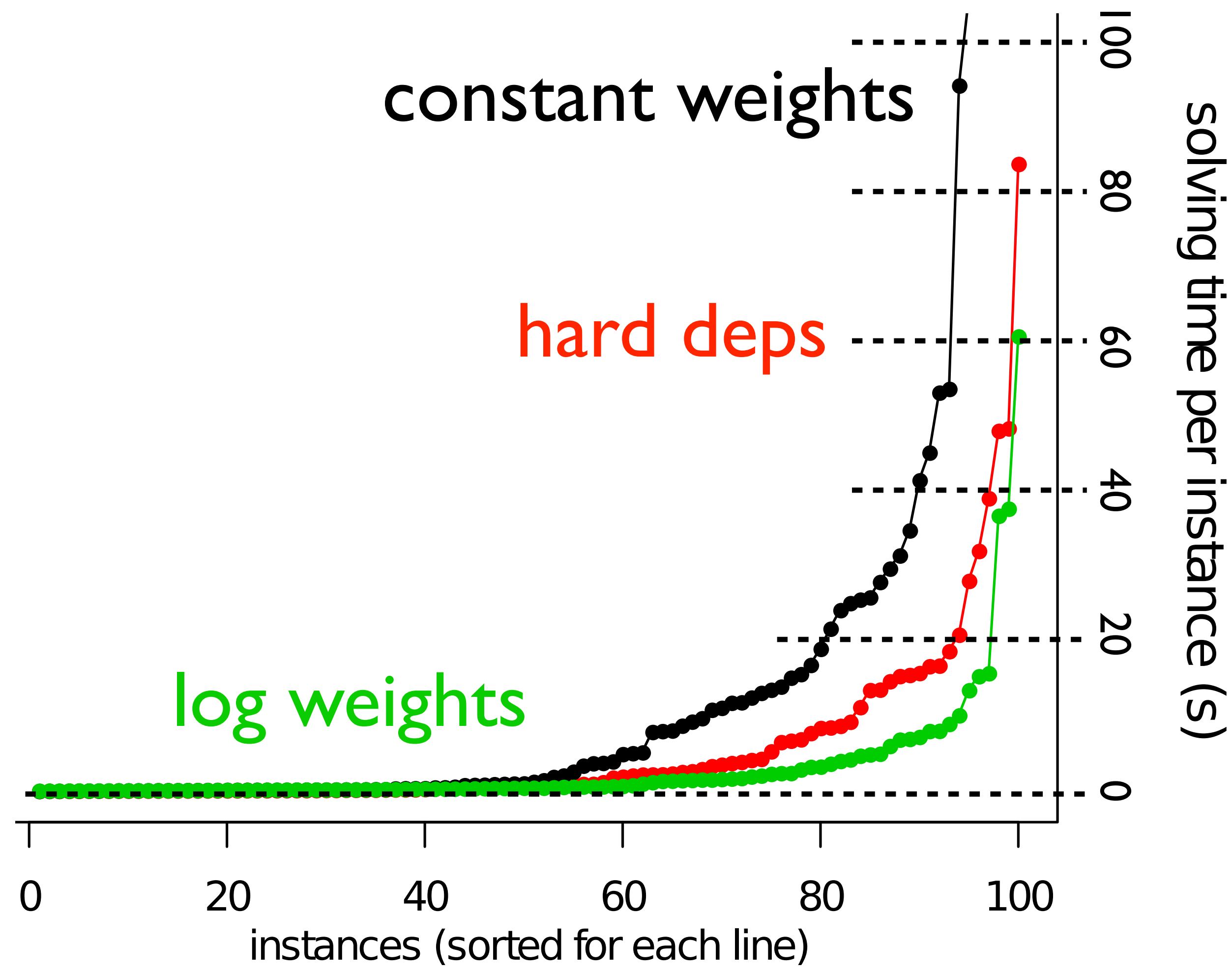
“priors” $\Longleftrightarrow$

- specific probabilities for each graph
- soft sparsity constraint
- ...

assumption/ algorithm	PC / GES	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM	non-linear additive noise	maxSAT
Markov	✓	✓	✓	✓	✓	✓	✓	✓
faithfulness	✓	✓	✓	✗	✓	~	minimality	✓
causal sufficiency	✓	✗	✓	✓	✗	✓	✓	✗
acyclicity	✓	✓	✗	✓	✓	✗	✓	✗*
parametric assumption	✗	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian	non-linear additive noise	✗

assumption/ algorithm	PC / GES	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM	non-linear additive noise	maxSAT
Markov	✓	✓	✓	✓	✓	✓	✓	✓
faithfulness	✓	✓	✓	✗	✓	~	minimality	✓
causal sufficiency	✓	✗	✓	✓	✗	✓	✓	✗
acyclicity	✓	✓	✗	✓	✓	✗	✓	✗*
parametric assumption	✗	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian	non-linear additive noise	✗
output	Markov equivalence	PAG	PAG	unique DAG	set of DAGs	set of graphs	unique DAG	query based

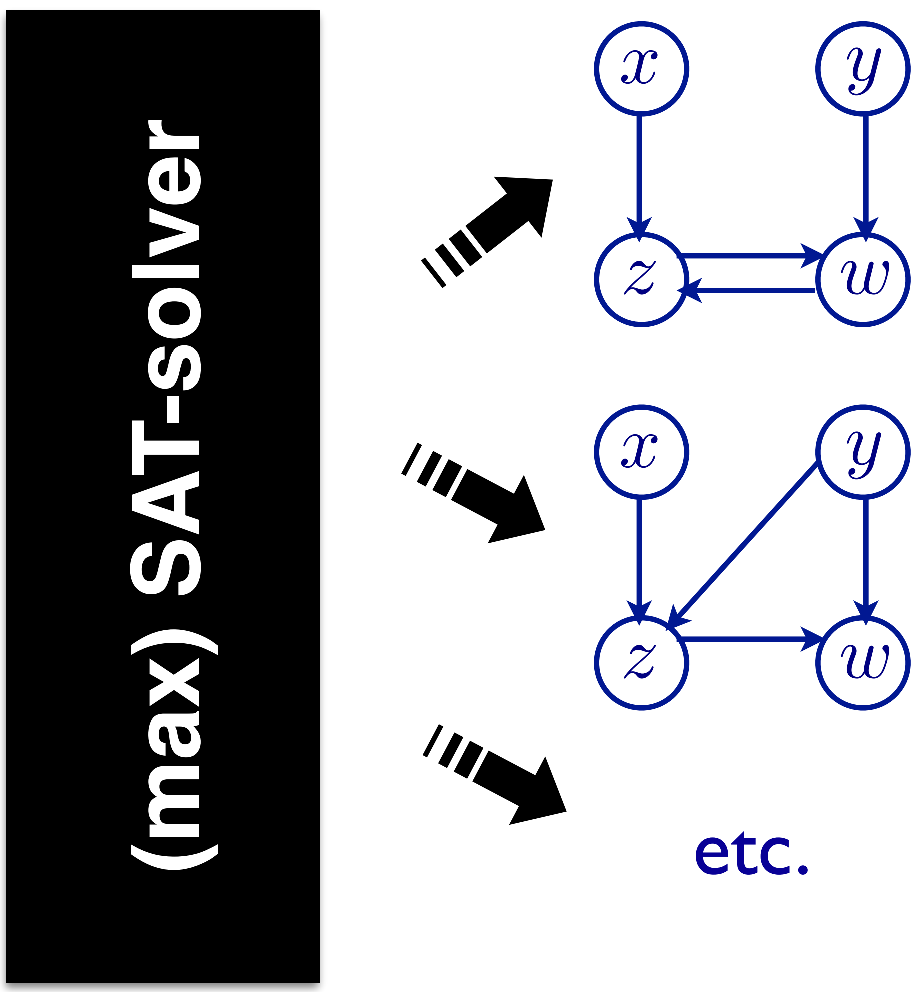
Simulation 4: Scalability



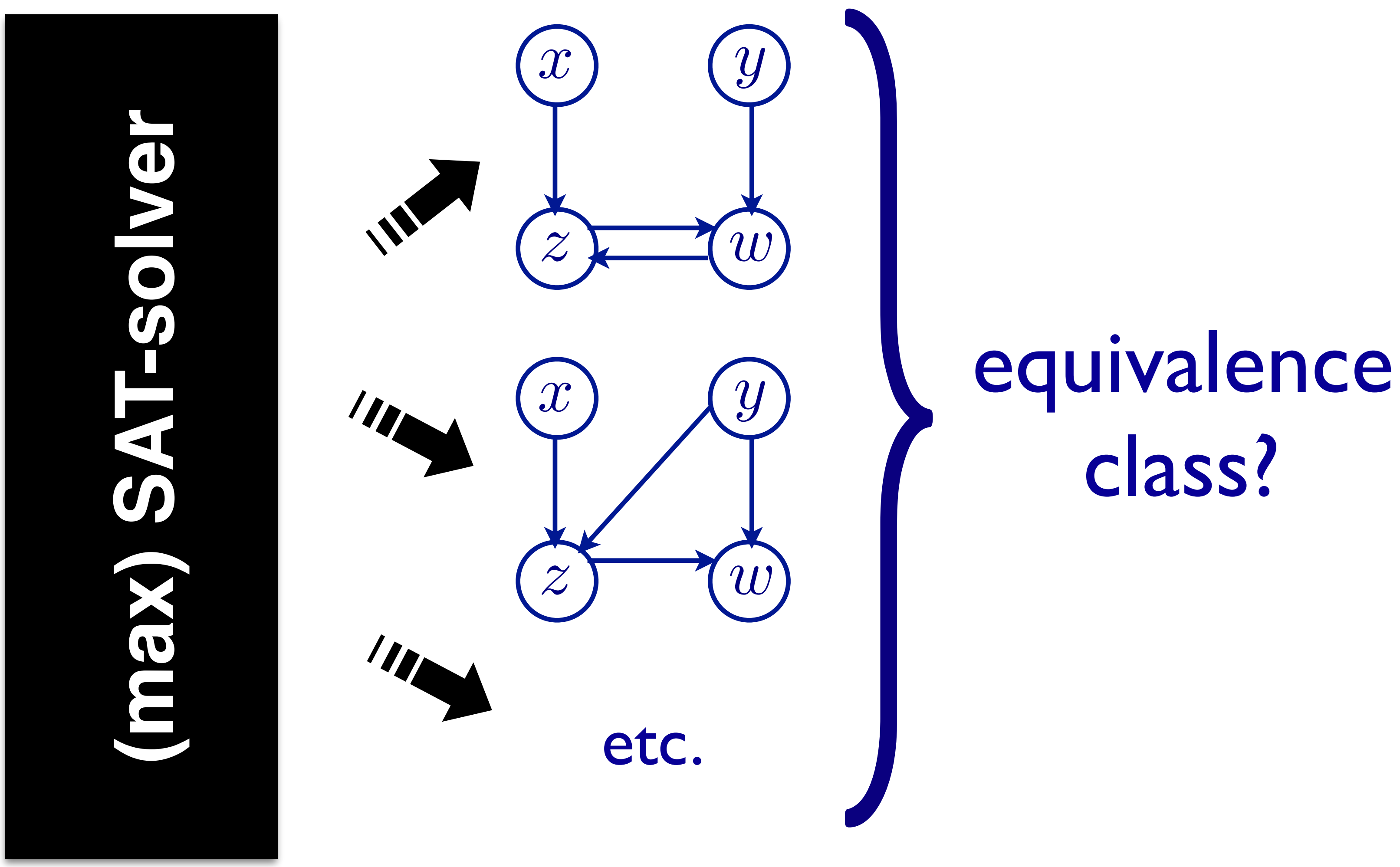
- up to 10 variables and only a few overlapping data sets for now

[Hyttinen et al. 2014]

Output of Causal Search Algorithms



Output of Causal Search Algorithms



Output of Causal Search Algorithms

Query:

(max) SAT-solver

Output of Causal Search Algorithms

Query:

- list the structures in the equivalence class

(max) SAT-solver

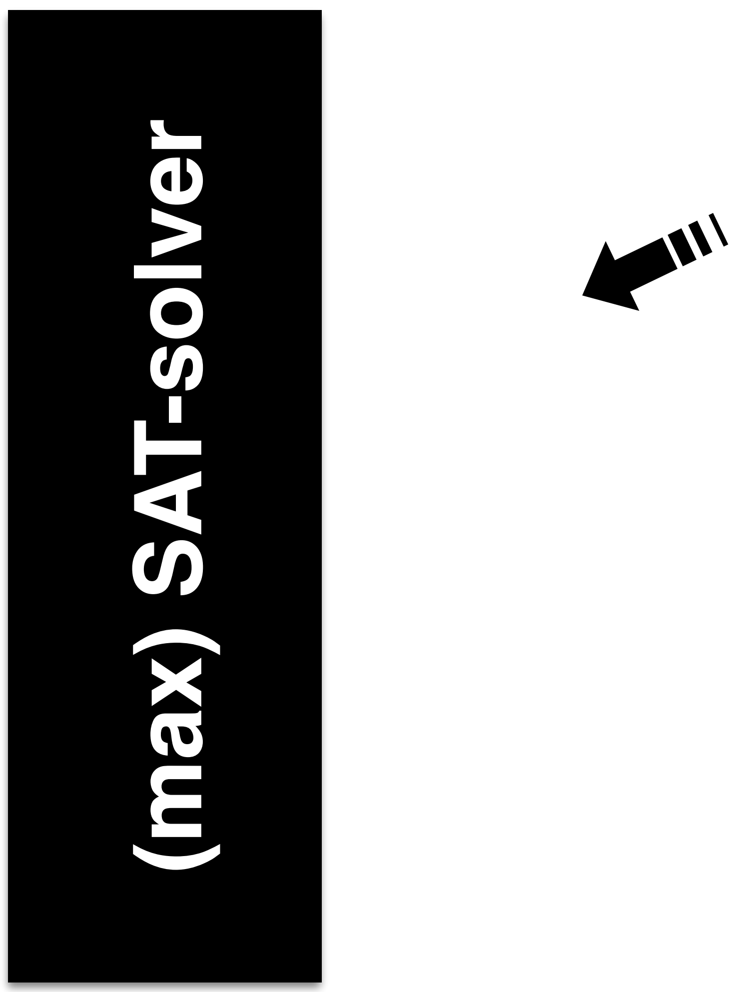
Output of Causal Search Algorithms

(max) SAT-solver

Query:

- list the structures in the equivalence class
- what structural features are determined?
  - edges, confounders
  - ancestral relations
  - pathways

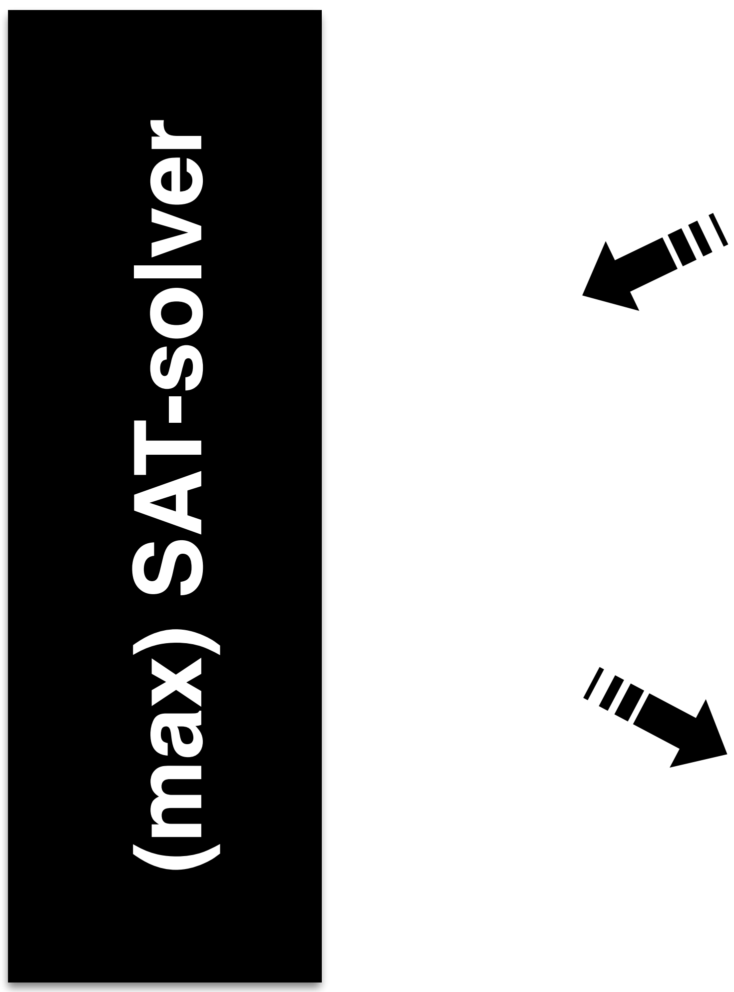
Output of Causal Search Algorithms



Query:

- list the structures in the equivalence class
- what structural features are determined?
  - edges, confounders
  - ancestral relations
  - pathways
- what are the highest scoring equivalence classes?

Output of Causal Search Algorithms



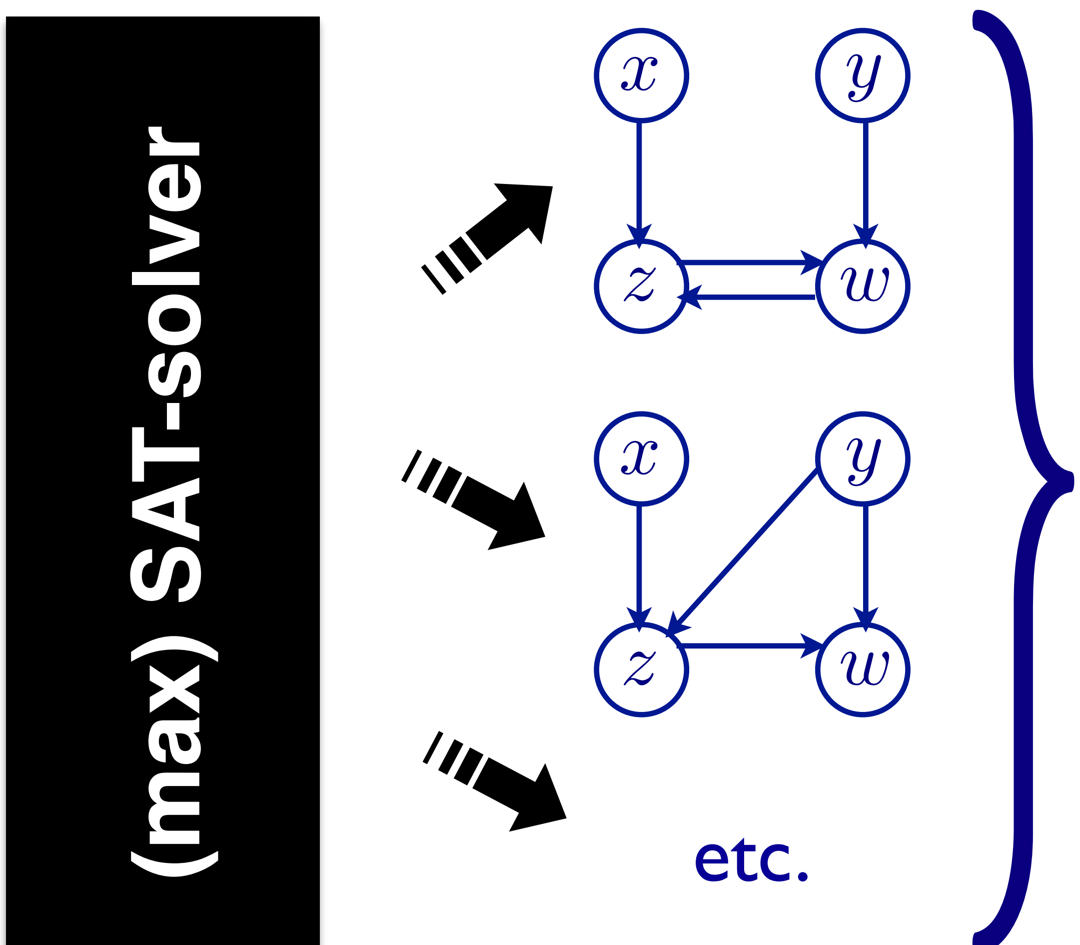
Query:

- list the structures in the equivalence class
- what structural features are determined?
  - edges, confounders
  - ancestral relations
  - pathways
- what are the highest scoring equivalence classes?

Response:

- enumeration of solutions
- “backbone” of the SAT-instance
- …

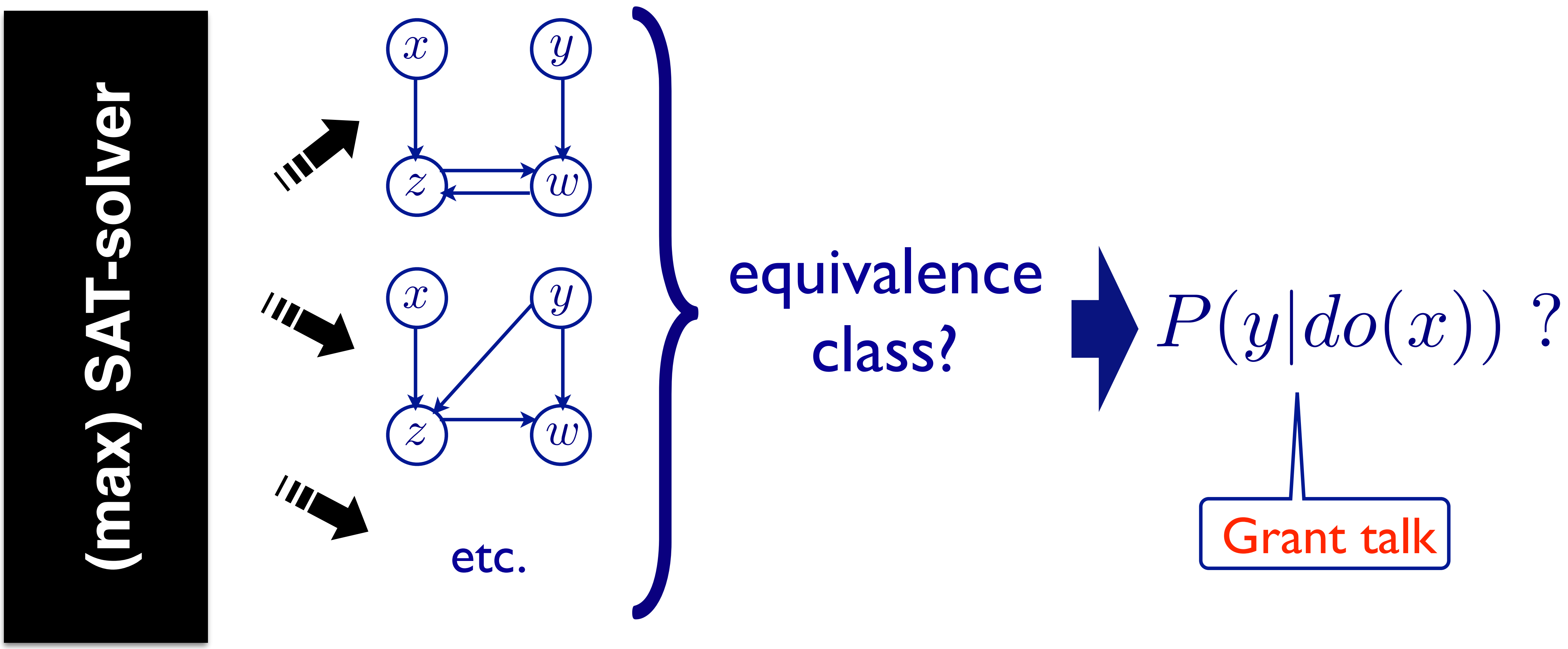
Computing Causal Effects



equivalence class?

$P(y|do(x))$?

Computing Causal Effects



equivalence class ➡ $P(y|do(x))$?

- search in the equivalence class over the possible applications of the *do*-calculus rules by *querying* the satisfaction of their d-separation conditions

[Hyttinen et al. 2015]

equivalence class $\Rightarrow$ $P(y|do(x))$?

- search in the equivalence class over the possible applications of the *do*-calculus rules by *querying* the satisfaction of their d-separation conditions

do-calculus

Rule 1 (insertion/deletion of observations)

$$P(y|do(x), z, w) = P(y|do(x), w) \text{ if } Y \perp\!\!\!\perp Z|X, W||X$$

Rule 2 (action/observation exchange)

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \text{ if } Y \perp\!\!\!\perp I_Z|X, Z, W||X$$

Rule 3 (insertion/deletion of actions)

$$P(y|do(x), do(z), w) = P(y|do(x), w) \text{ if } Y \perp\!\!\!\perp I_Z|X, W||X$$

[Hyttinen et al. 2015]

equivalence class ➡ $P(y|do(x))$?

- search in the equivalence class over the possible applications of the *do*-calculus rules by *querying* the satisfaction of their d-separation conditions

do-calculus

Rule 1 (insertion/deletion of observations)

$$P(y|do(x), z, w) = P(y|do(x), w) \text{ if } Y \perp\!\!\!\perp Z|X, W||X$$

Rule 2 (action/observation exchange)

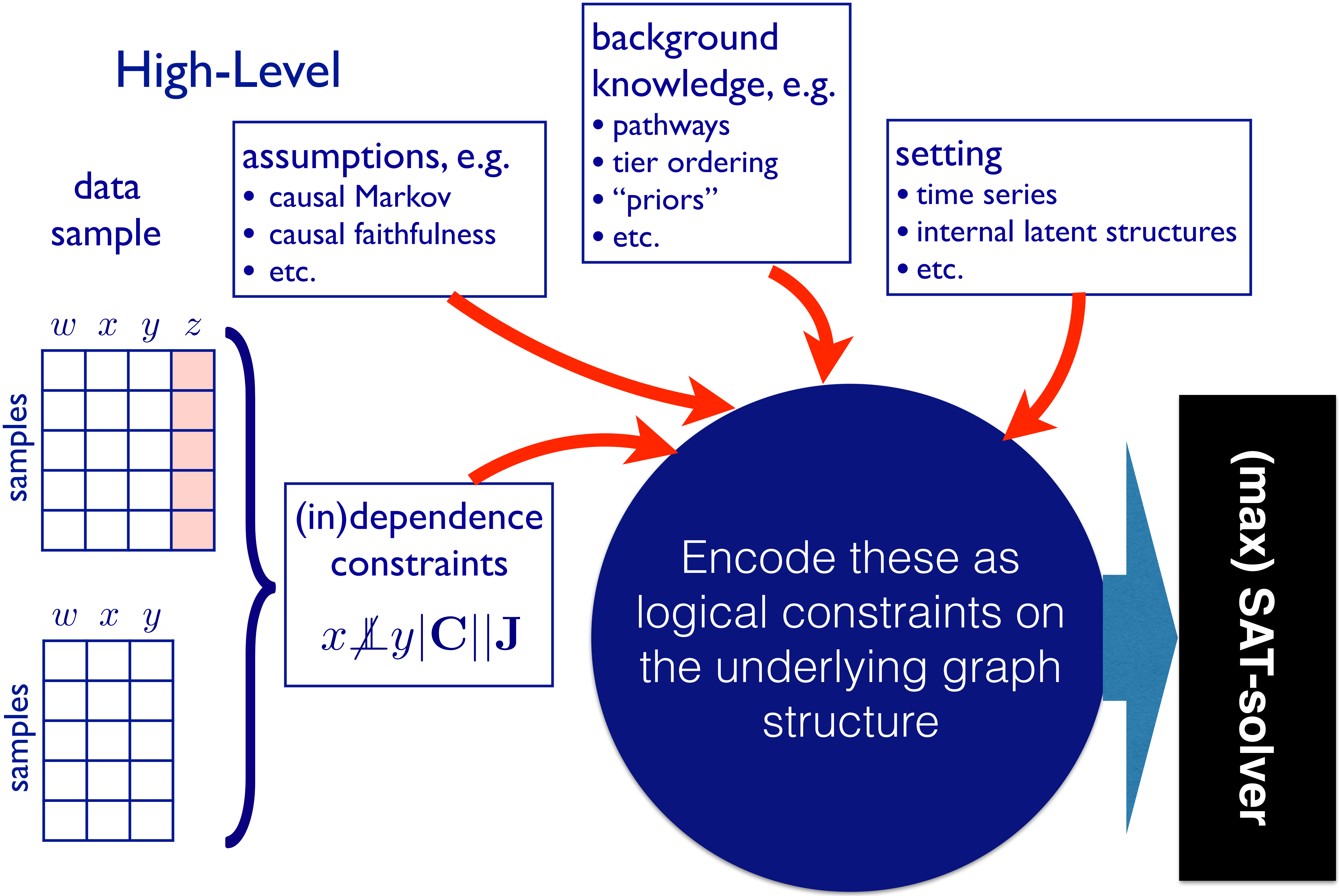
$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \text{ if } Y \perp\!\!\!\perp I_Z|X, Z, W||X$$

Rule 3 (insertion/deletion of actions)

$$P(y|do(x), do(z), w) = P(y|do(x), w) \text{ if } Y \perp\!\!\!\perp I_Z|X, W||X$$

[Hyttinen et al. 2015]

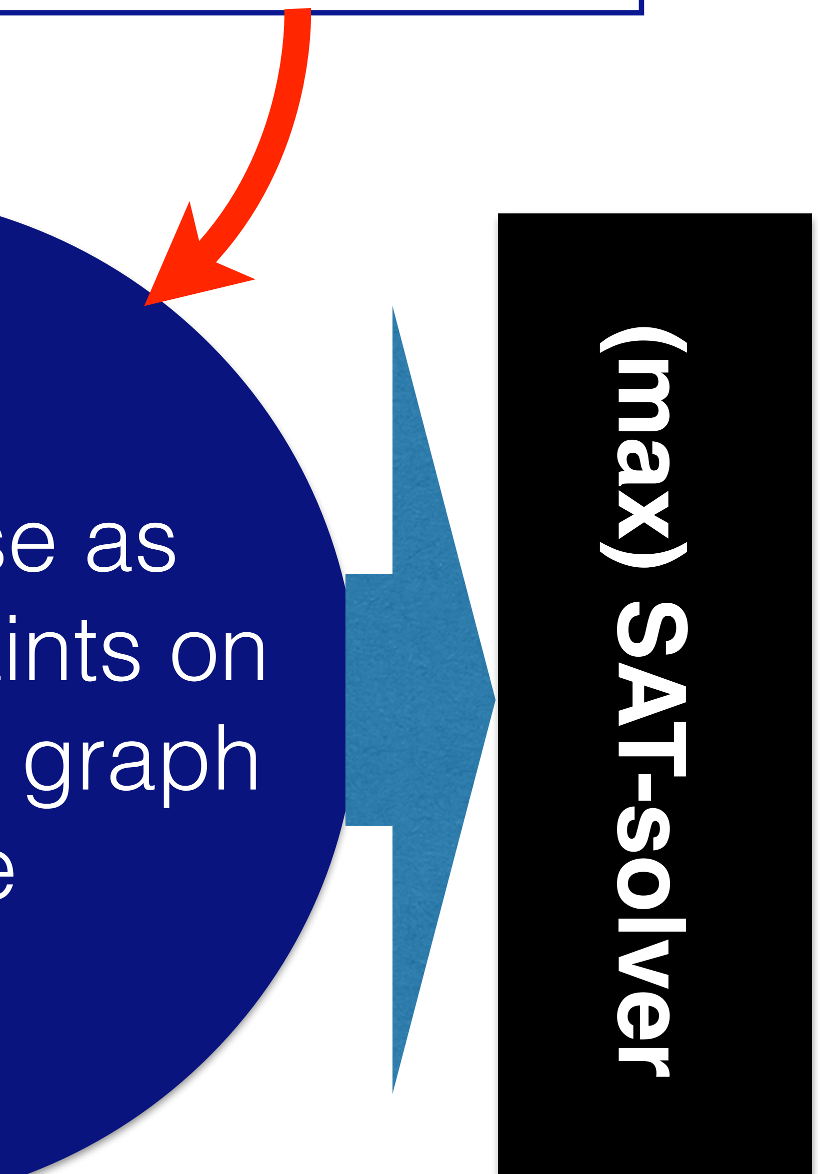
High-Level



data sample
assumptions, e.g.
• causal Markov
• causal faithfulness
• etc.
background knowledge, e.g.
• pathways
• tier ordering
• “priors”
• etc.
setting
• time series
• internal latent structures
• etc.
w x y z
samples
w x y
samples
(in)dependence constraints
$x \not\perp\!\!\!\perp y \,|\, \mathbf{C} \,||\, \mathbf{J}$
Encode these as logical constraints on the underlying graph structure
(max) SAT-solver

High-Level

tting
ime series
nternal latent structures
tc.



High-Level

tting

ime series

nternal latent structures

tc.

QUERY?

se as
ints on
graph
e

(max) SAT-solver

High-Level

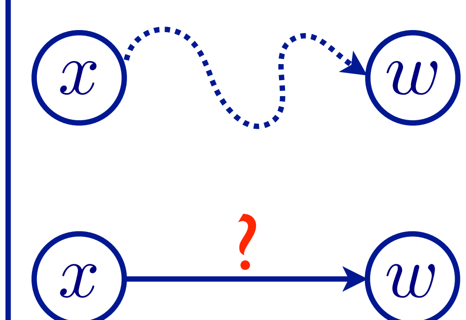
tting
ime series
nternal latent structures
tc.

QUERY?

se as
ints on
graph
e

(max) SAT-solver

$$P(y \,|\, do(x))$$



Just getting started…

Just getting started…

- application

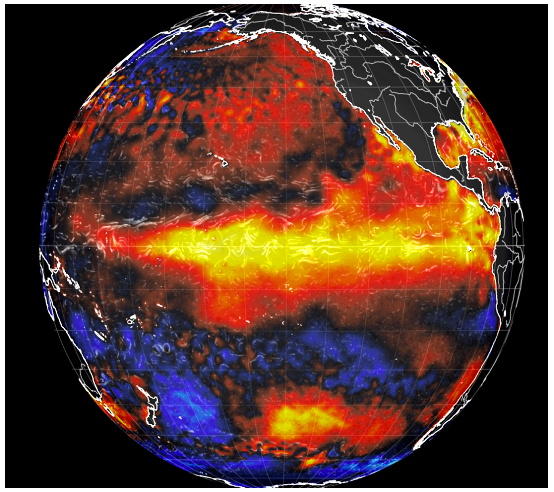
[Stekhoven et al. 2012]

Just getting started…

- application

[Stekhoven et al. 2012]

- multi-scale causal analysis: micro- to macro-variables



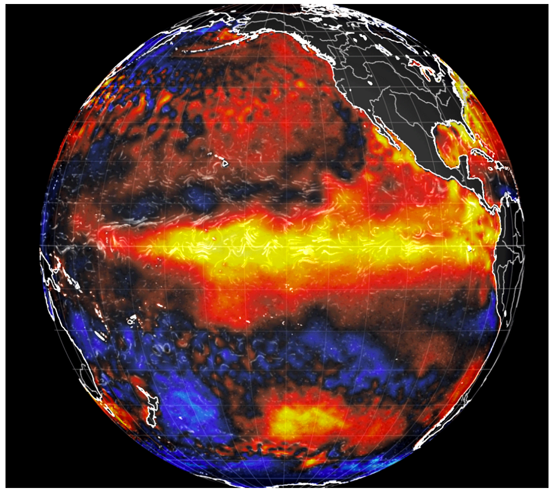
[Chalupka et al. 2016]

Just getting started…

- application

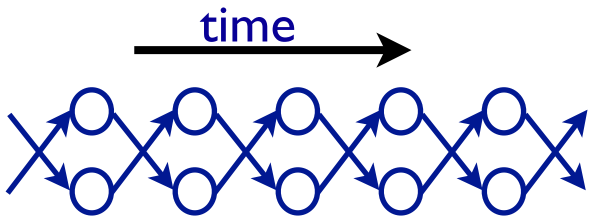
[Stekhoven et al. 2012]

- multi-scale causal analysis: micro- to macro-variables



[Chalupka et al. 2016]

- time-series and dynamics

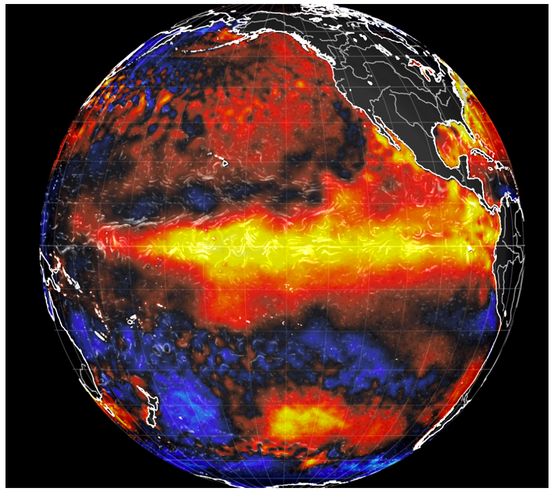


Just getting started…

- application

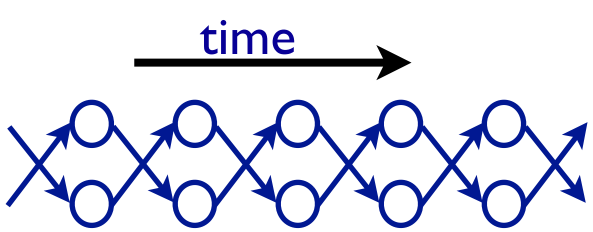
[Stekhoven et al. 2012]

- multi-scale causal analysis: micro- to macro-variables

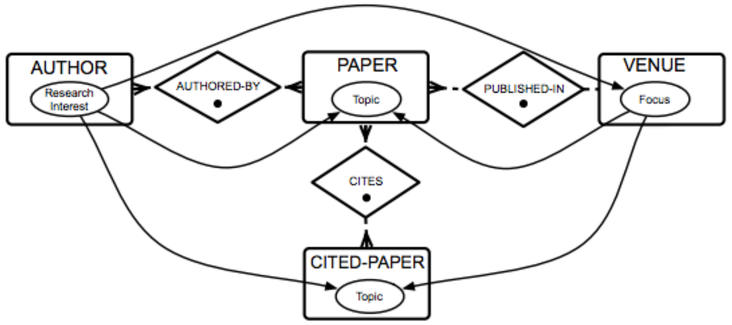


[Chalupka et al. 2016]

- time-series and dynamics



- violations of the Markov property: non-causal relations



[Maier et al. 2013]

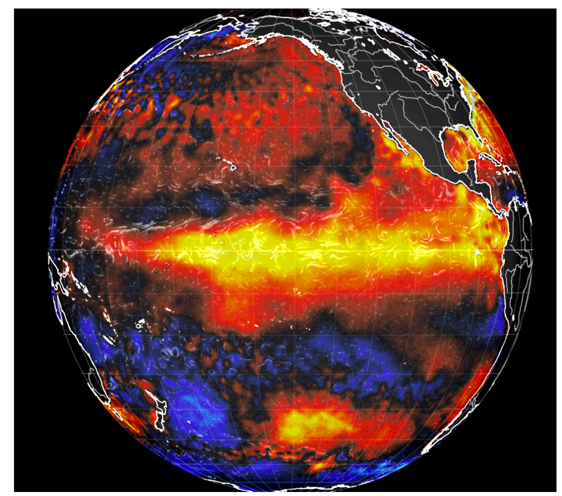
Just getting started…

Sokolova talk

- application

[Stekhoven et al. 2012]

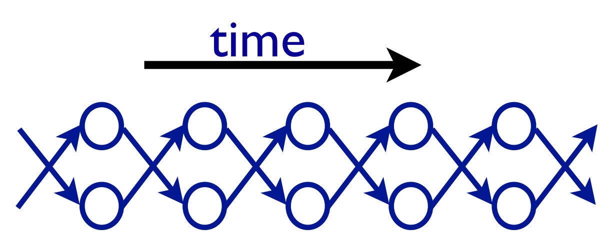
- multi-scale causal analysis: micro- to macro-variables



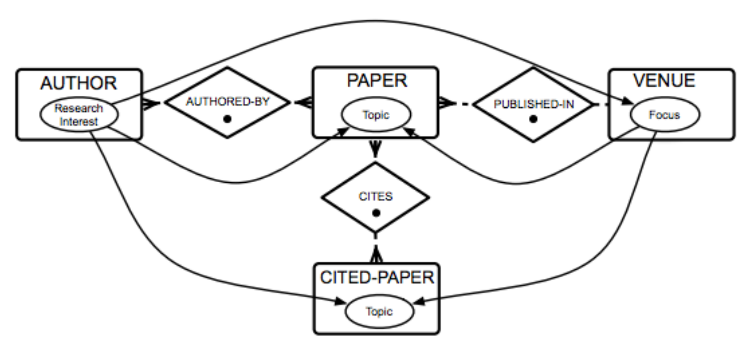
[Chalupka et al. 2016]

Blondel talk

- time-series and dynamics



- violations of the Markov property: non-causal relations



[Maier et al. 2013]

References

Limitations

- Verma & Pearl, **Equivalence and synthesis of causal models,** UAI 1990.
- Frydenberg, **The chain graph Markov property,** Scandinavian Journal of Statistics 1990.
- Geiger & Pearl, **On the logic of influence diagrams**, UAI 1988.
- Meek, **Strong completeness and faithfulness in Bayesian networks**, UAI 1995.

LiNGaM

- Shimizu et al, **A linear non-Gaussian acyclic model for causal discovery**, JMLR, 2006.
- Hoyer et al., **Estimation of causal effects using linear non-Gaussian causal models with hidden variables**, IJAR 2008.
- Lacerda et al., **Discoverying cyclic causal models by Independent Component Analysis**, UAI 2008.

Additive noise models

- Hoyer et al., **Nonlinear causal discovery with additive noise models**, NIPS 2009.
- Mooij et al., **Regression by dependence minimization and its application to causal inference**, ICML 2009.
- Peters et al., **Causal inference on discrete data using additive noise models**, IEEE..., 2011.
- Peters et al., **Identifiability of causal graphs using functional models**, UAI 2011.

SAT-based approaches

- Triantafillou et al., **Learning causal structure from overlapping variable sets,** AISTATS 2010.
- Claassen & Heskes, **A logical characterization of constraint-based causal discovery**, UAI 2011.
- Hyttinen et al., **Discovering cyclic causal models with latent variables: A SAT-based approach**, UAI 2013.
- Hyttinen et al., **Constraint-based Causal Discovery: Conflict Resolution with Answer Set Programming,** UAI 2014.
- Hyttinen et al., **Do-calculus when the true graph is unknown,** UAI 2015.
- Triantafillou & Tsamardinos, **Constraint-based Causal Discovery from Multiple Interventions Over Overlapping Variable Sets,** JMLR 2015.

Other references

- Maier et al., **A sound and complete algorithm for learning causal models from relational data,** UAI 2013.
- Chalupka et al., **Unsupervised discovery of El Niño using causal feature learning on microlevel climate data,** UAI 2016.
- Stekhoven et al., **Causal stability ranking,** Bioinformatics 2012.

Thank you!