Outline

- Theory-based Bayesian framework for property induction
- Causal structure induction
  - Constraint-based (bottom-up) learning
  - Theory-based Bayesian learning

The origins of causal knowledge

- Question: how do people *reliably* come to *true* beliefs about the causal structure of their world?
- Answer must specify:
  - Prior causal knowledge
  - Causal inference procedure

Multiple goals

- Descriptive:
  – Prior knowledge must be psychologically realistic.
  – Inference procedure must generate the same beliefs that people do, given the same input.

- Explanatory:
  – Prior knowledge must be approximately correct.
  – Inference procedure (constrained by prior knowledge) must be reliable.

Analogy with vision

(Pearl, Cheng, Gopnik et al.)

External world structure

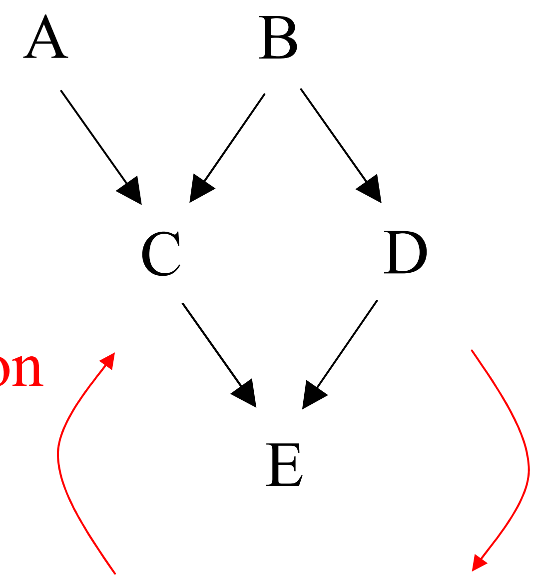
Vision (inverse graphics)

Graphics

Observed images

The fundamental problem

Hidden causal structure:



Observed data:

Case	A	B	C	D	E
1	0	1	1	1	1
2	1	0	1	0	1
3	0	0	0	1	0
4	0	1	1	0	1
. . . .					

Under-constrained problems

In both visual perception and causal induction, many world structures could have produced the same data.

Image removed due to copyright considerations. Please see:
Freeman, WT. "The Generic Viewpoint Assumption in a Framework for Visual Perception." *Nature* 368 (7 April 1994): 542-545.

Image

Possible world structures

Under-constrained problems

In both visual perception and causal induction, many world structures could have produced the same data.

$$P(A,B) \neq P(A)P(B)$$

Correlation

$$A \longrightarrow B \qquad\qquad A \longleftarrow B$$

$$\begin{tikzcd}[column sep=small] & X \arrow[dl] \arrow[dr] & \\ A & & B \end{tikzcd} \qquad \begin{tikzcd}[column sep=small] & X \arrow[dr] & \\ A \arrow[ur] & & B \end{tikzcd}$$

$$A \qquad\qquad B$$

Possible world structures

Questions in visual perception

- How is the external world represented?
  - 3-D models
  - 2-D views
  - Intermediate: 2 1/2-D sketch, layers, intrinsic images, etc.
- What kind of knowledge does the mind have about the world?
  - Structure of objects
  - Physics of surfaces
  - Statistics of scenes
- How does inference work?
  - Bottom-up, modular, context-free
  - Top-down, flexible, context-sensitive

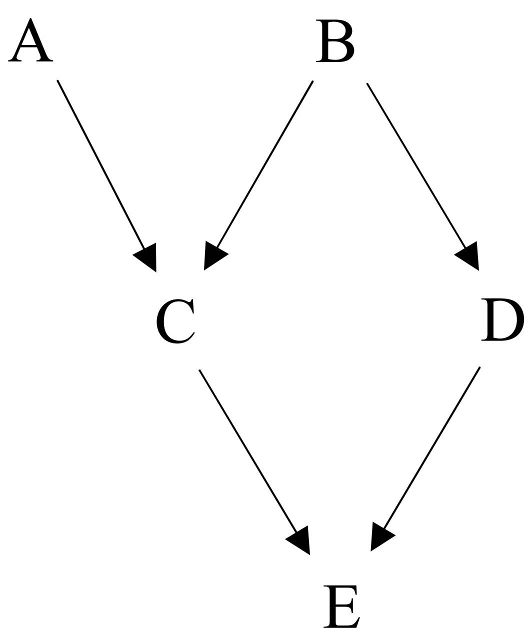
Questions in causal induction

- How is the external world represented?
  - Associations
  - Causal structures
  - Intermediate: Causal strength parameters
- What kind of knowledge does the mind have about the world?
  - Constraints on causal structure (e.g., causal order)
  - Faithfulness (observed independence relations are real)
  - Causal mechanisms
- How does inference work?
  - Bottom-up: constraint-based (data mining) approach
  - Top-down: theory-based Bayesian approach

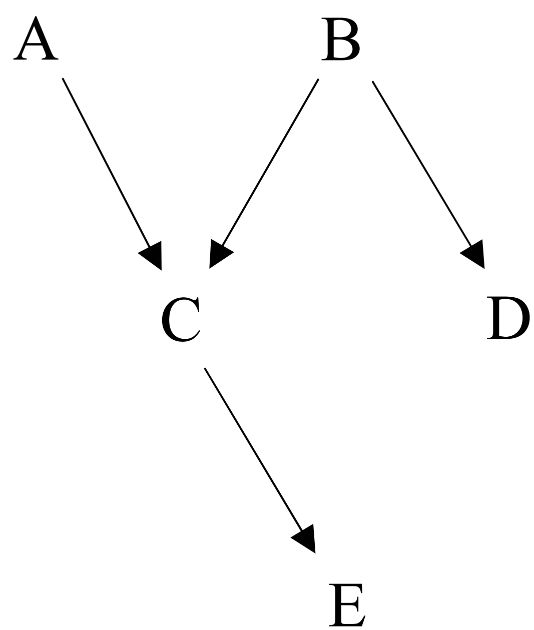
Some vocabulary

- Causal structure
  - What causes what.

Specifies *nothing* about causal mechanisms or parameterizations.

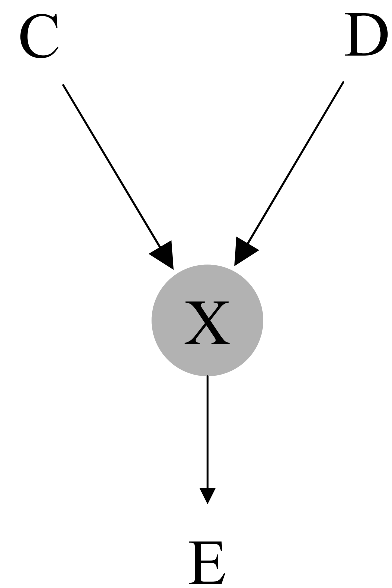


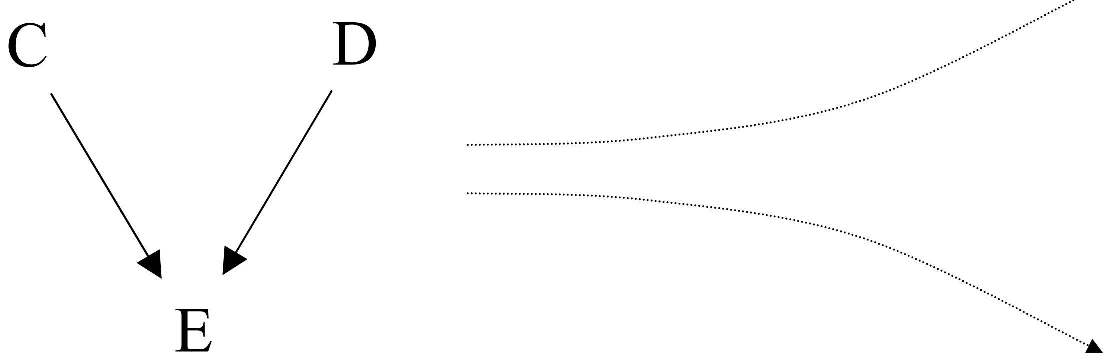
vs.



Some vocabulary

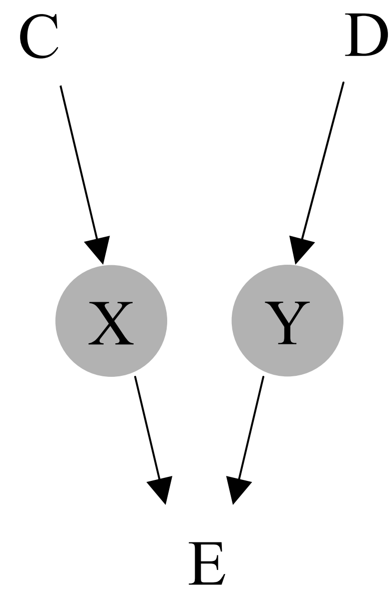
- Causal structure
  - What causes what.
- Causal mechanism
  - How causes influence effects.

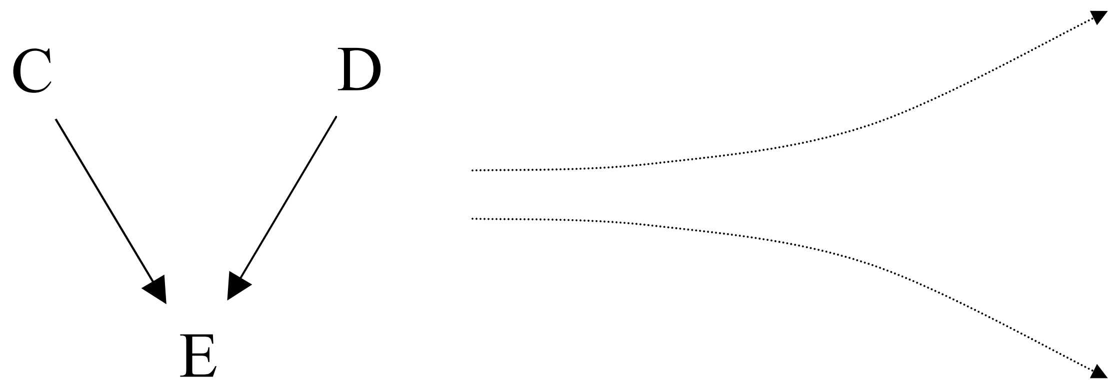




Some vocabulary

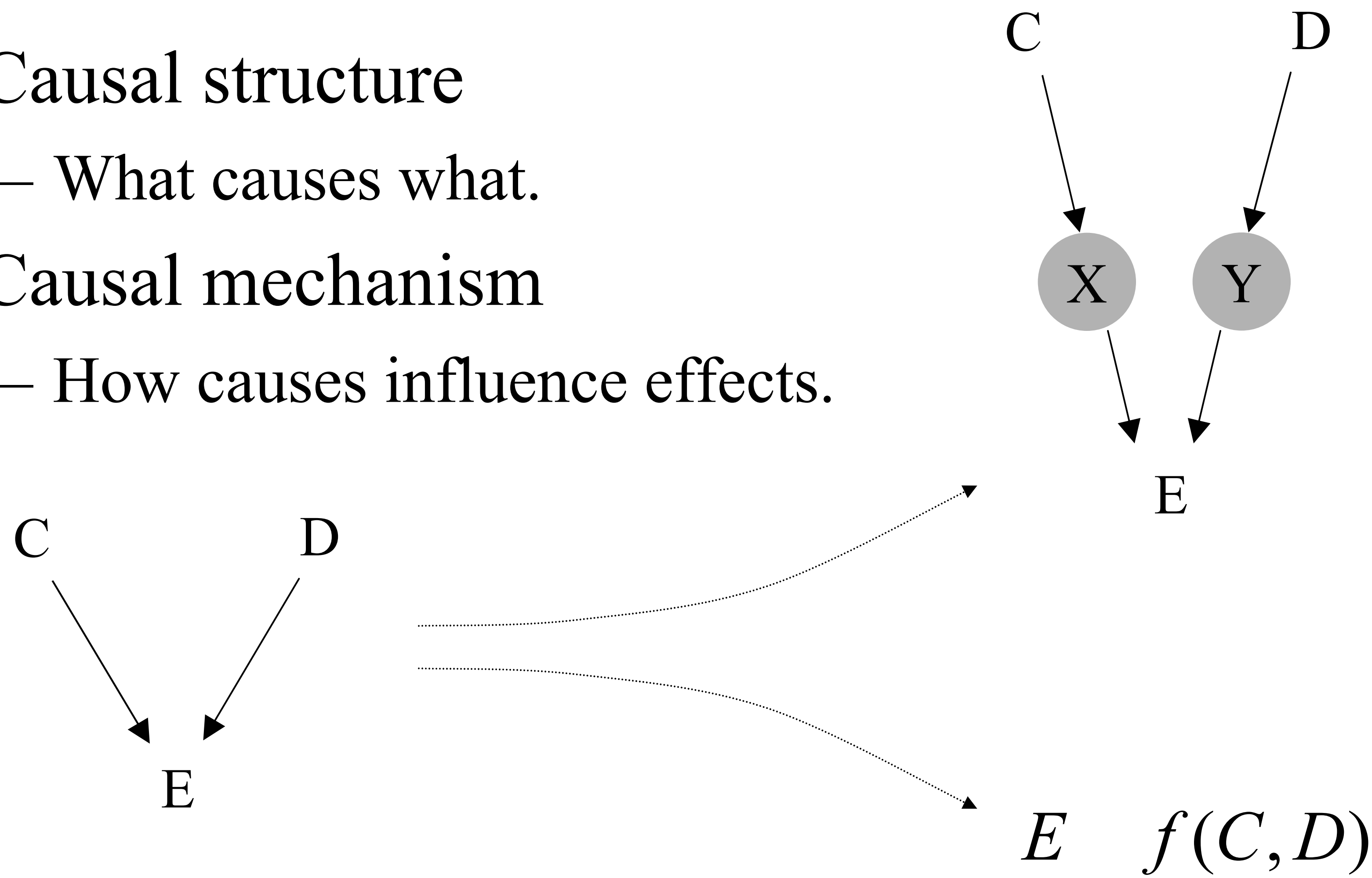
- Causal structure
  - What causes what.
- Causal mechanism
  - How causes influence effects.





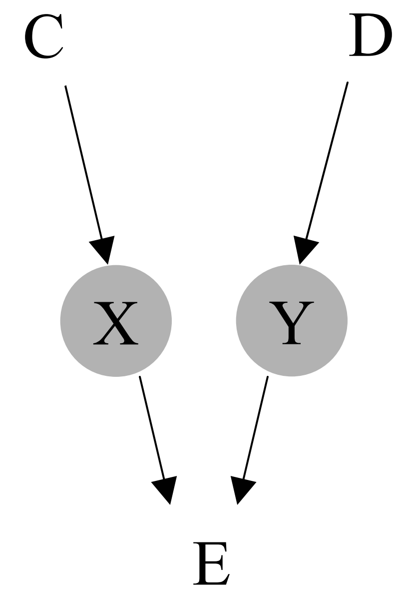
Some vocabulary

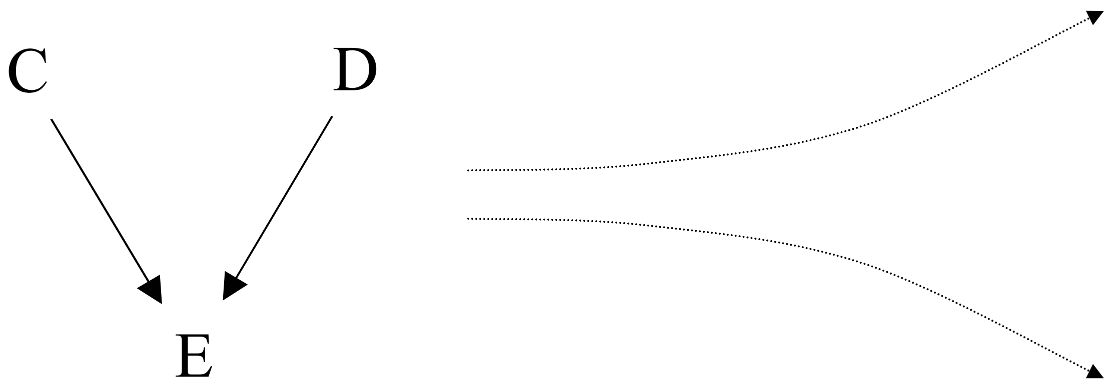
- Causal structure
  - What causes what.
- Causal mechanism
  - How causes influence effects.



Some vocabulary

- Causal structure
  - What causes what.
- Causal mechanism
  - How causes influence effects.





$$E \quad f(C, D, \varepsilon)$$
$$\varepsilon \sim \text{Gaussian}(\mu, \sigma)$$

Some vocabulary

- Causal structure
  – What causes what.
- Causal mechanism
  – How causes influence effects.

Knowledge about causal structures and mechanisms can be represented at different scales of detail.

Abstract ("light") mechanism knowledge will be particularly important: e.g.,

- deterministic, quasi-deterministic, semi-deterministic or stochastic?
- strong or weak?
- generative or preventive influence?
- independent of or interactive with other causes?

Some vocabulary

- Causal structure
  - What causes what.
- Causal mechanism
  - How causes influence effects.
- Parameterization
  - Form of $P(\text{effect}|\text{causes})$, e.g. "noisy-OR"
- Causal strengths (parameters)
  - Relative contributions of different causes given a particular mechanism or parameterization.

Approaches to structure learning

- Constraint-based learning (Pearl, Glymour, Gopnik):
  - Assume structure is unknown, no knowledge of parameterization or parameters
- Bayesian learning (Heckerman, Friedman/Koller):
  - Assume structure is unknown, arbitrary parameterization.
- Theory-based Bayesian inference (T & G):
  - Assume structure is partially unknown, parameterization is known but parameters may not be. *Prior knowledge about structure and parameterization depends on domain theories (derived from ontology and mechanisms).*

Approaches to structure learning

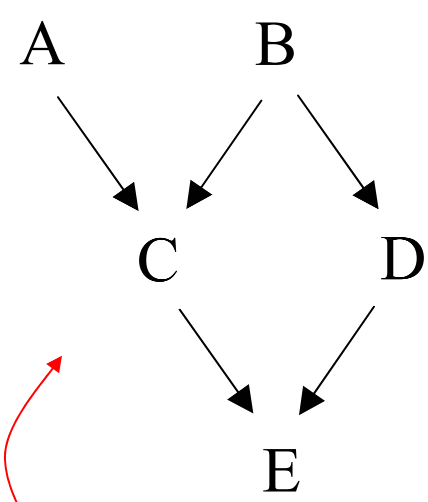
- Constraint-based learning (Pearl, Glymour, Gopnik):
  - Assume structure is unknown, no knowledge of parameterization or parameters
- Bayesian learning (Heckerman, Friedman/Koller):
  - Assume structure is unknown, arbitrary parameterization.
- Theory-based Bayesian inference (T & G):
  - Assume structure is partially unknown, parameterization is known but parameters may not be. *Prior knowledge about structure and parameterization depends on domain theories (derived from ontology and mechanisms).*

Causal inference in science

- Standard question: is X a direct cause of Y?
- Standard empirical methodologies in many domains:
  - Psychology
  - Medicine
  - Epidemiology
  - Economics
  - Biology
- Constraint-based inference attempts to formalize this methodology.

Constraint-based learning

Causal graph:



Faithfulness assumption

Causal Markov assumption

Probability distribution:

$$P(A,B,C,D,E) \qquad \prod_{V \in \{A,B,C,D,E\}} P(V \mid \text{parents}[V])$$

$$P(A,B,C,D,E) \qquad P(A)P(B)P(C \mid A,B)P(D \mid B)P(E \mid C,D)$$

Definition of "cause"

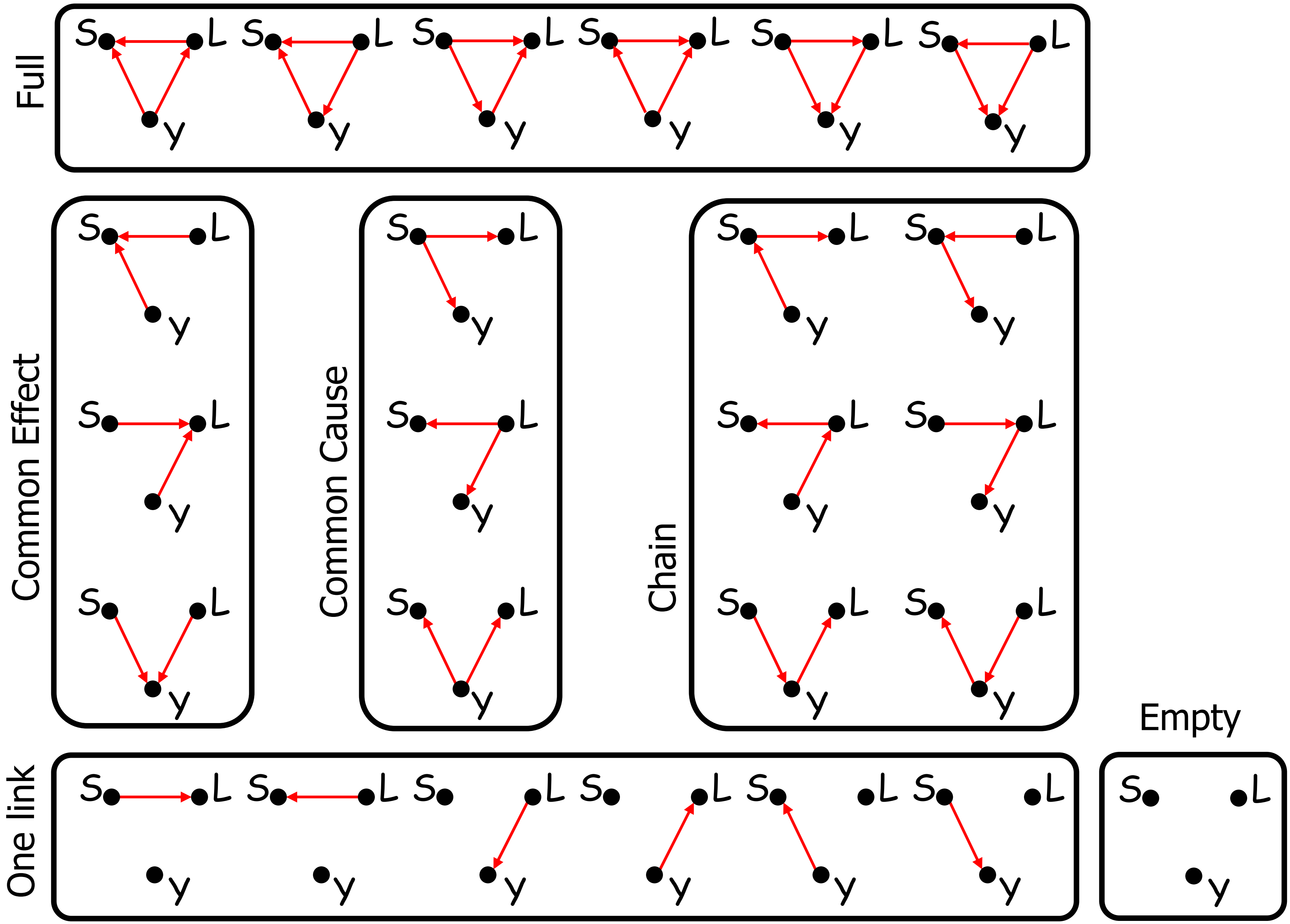
- Under the *causal Markov* principle, A is a *direct cause* of B implies that when all other potentially relevant variables are held constant, the probability of B depends upon the presence or absence of A.

- Under the *faithfulness* assumption, (in)dependence and conditional (in)dependence relations in the observed data imply constraints on the hidden causal structure *(see picture)*.

Example

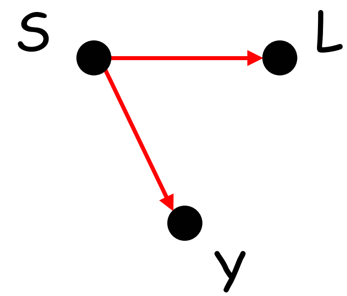
- What is the causal structure relating smoking (*S*), yellow teeth (*Y*), and lung cancer (*L*)?
- Epidemiological Data:

Patient	Smoking?	Yellow teeth?	Lung Cancer?
1	yes	yes	yes
2	yes	yes	no
3	yes	no	yes
4	no	no	no
5	yes	yes	yes
6	yes	no	no
7	yes	no	yes
8	no	no	no
. . . .			


Full
Common Effect
Common Cause
Chain
Empty
One link
S
L
y

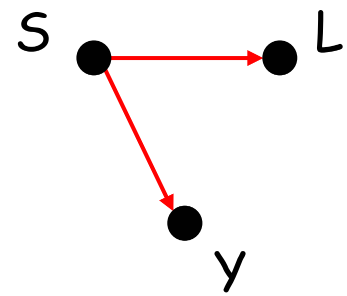
Inference process

- A hypothesis:



Inference process

- A hypothesis:



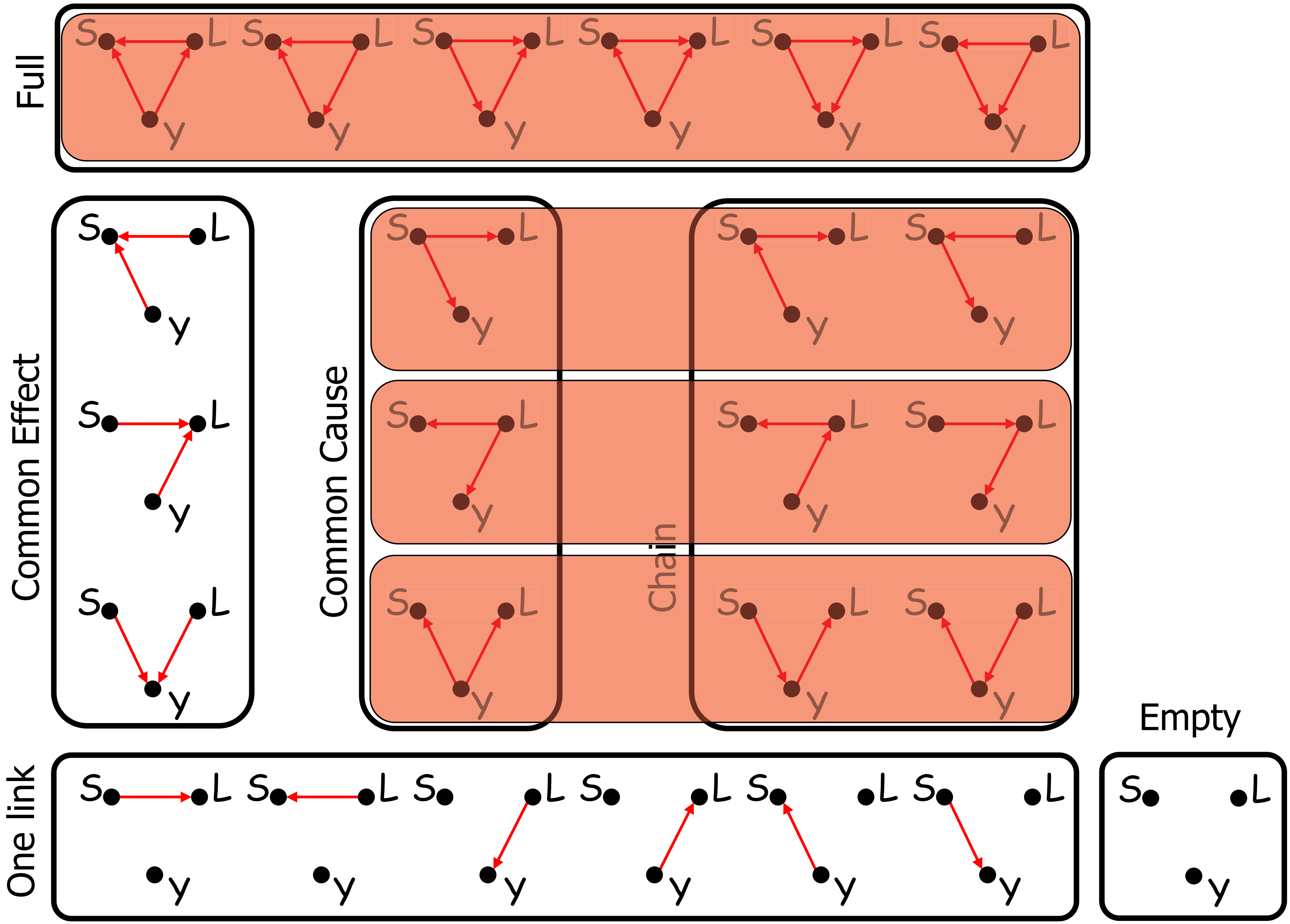
- What evidence would support this hypothesis?
- Would that evidence be consistent with any other hypothesis?

Example

- What is the causal structure relating smoking (*S*), yellow teeth (*Y*), and lung cancer (*L*)?
- Expected simple correlations:
  – smoking, yellow teeth: yes
  – smoking, lung cancer: yes
  – yellow teeth, lung cancer: yes
- Expected partial (conditional) correlations:
  – smoking, yellow teeth | lung cancer: yes
  – smoking, lung cancer | yellow teeth: yes
  – yellow teeth, lung cancer | smoking: no

Example

- What is the causal structure relating smoking (S), yellow teeth (Y), and lung cancer (L)?
- Expected simple correlations:
  - smoking, yellow teeth: yes
  - smoking, lung cancer: yes
  - yellow teeth, lung cancer: yes
- Under faithfulness, two variables that are correlated must share a common ancestor.
  - In this example, each pair of nodes must share a common ancestor.

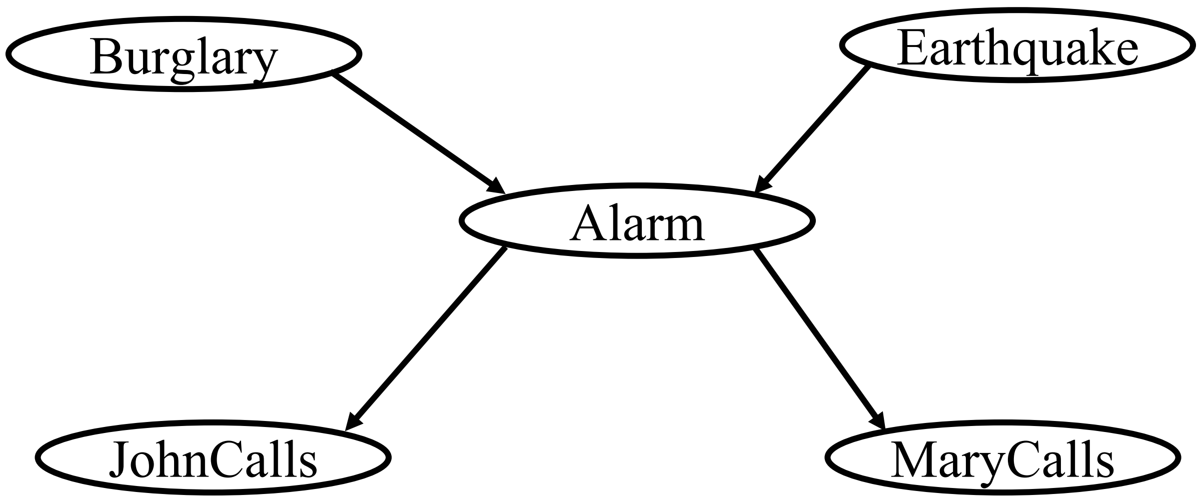


Full
Common Effect
Common Cause
Chain
Empty
One link
S
L
y

Global semantics

Joint probability distribution factorizes into product of local conditional probabilities:

$$P(V_1,\ldots,V_n) \quad \prod_{i\ \ 1}^{n} P(V_i \mid \text{parents}[V_i])$$



$$P(B, E, A, J, M)$$

$$P(B)\ P(E)\ P(A \mid B, E)\ P(J \mid A)\ P(M \mid A)$$

Local semantics

Global factorization is equivalent to a set of constraints on pairwise relationships between variables.

"Markov property": Each node is conditionally independent of its non-descendants given its parents.

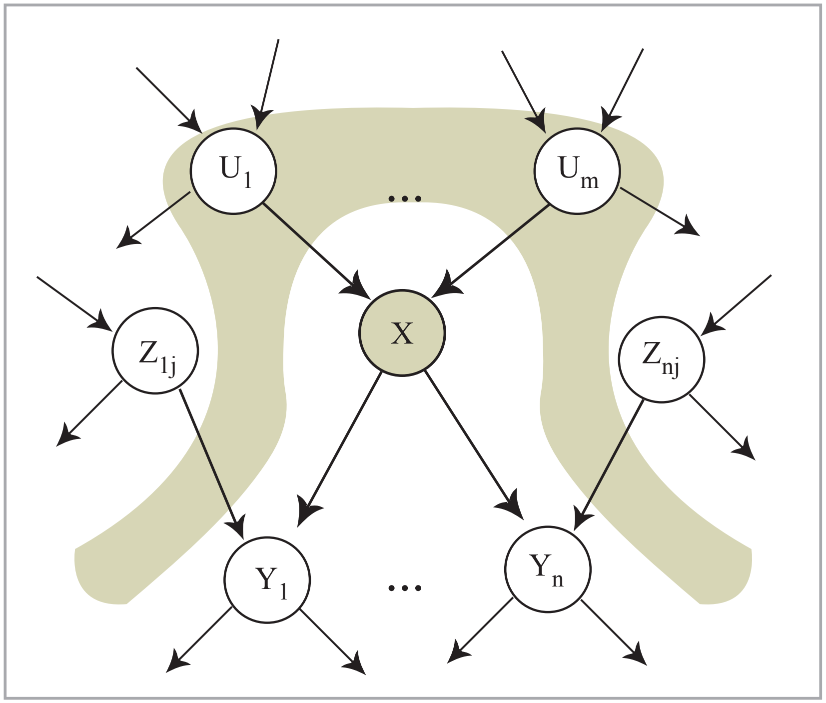


Image by MIT OCW.

Local semantics

Global factorization is equivalent to a set of constraints on pairwise relationships between variables.

Each node is conditionally independent of all others given its "Markov blanket": parents, children, children's parents.

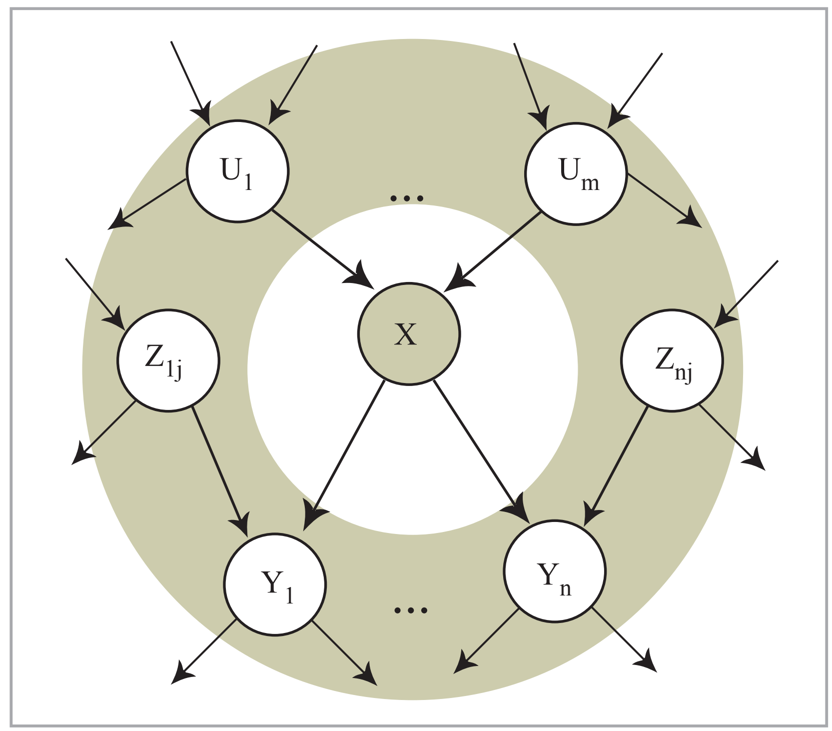
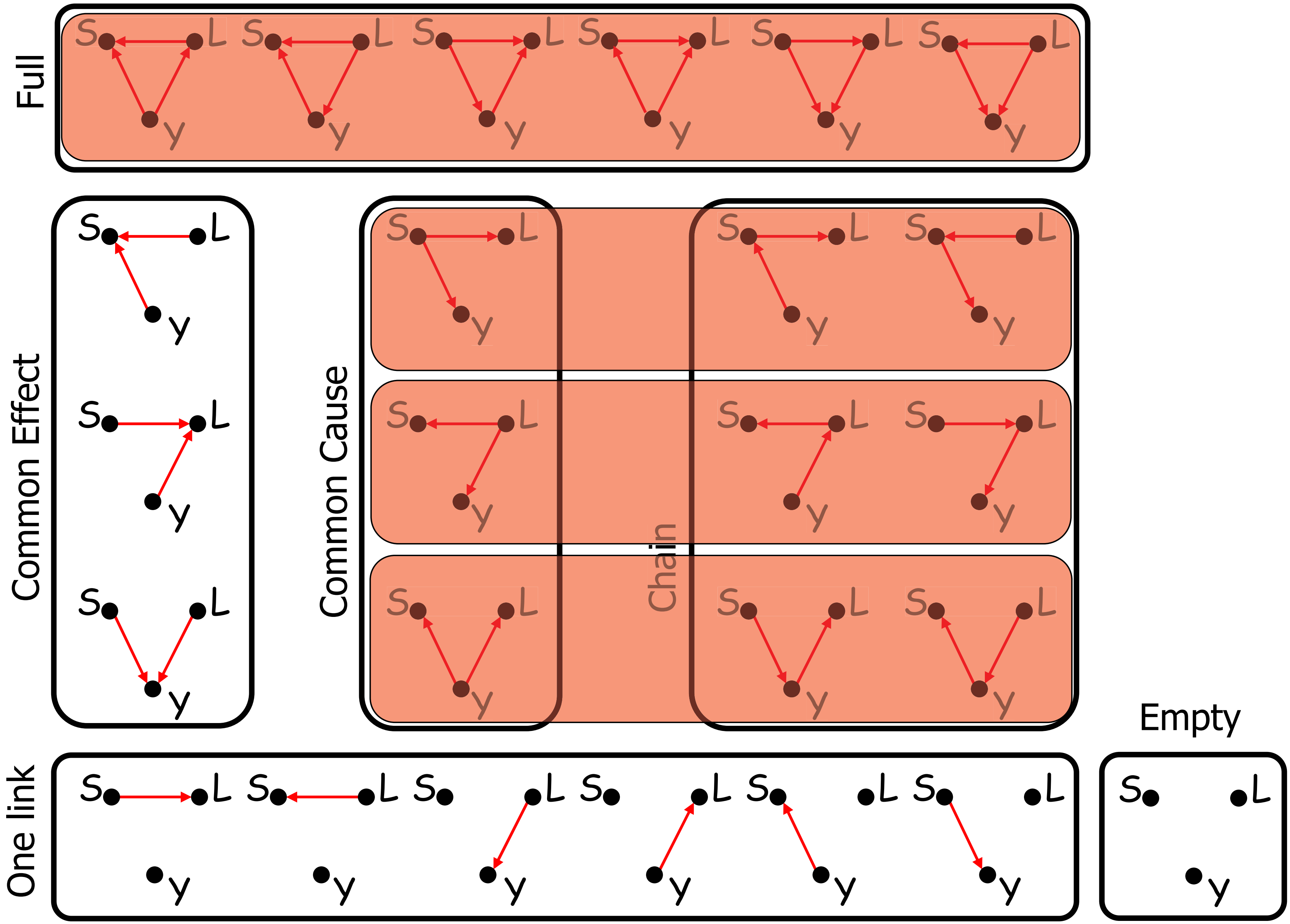


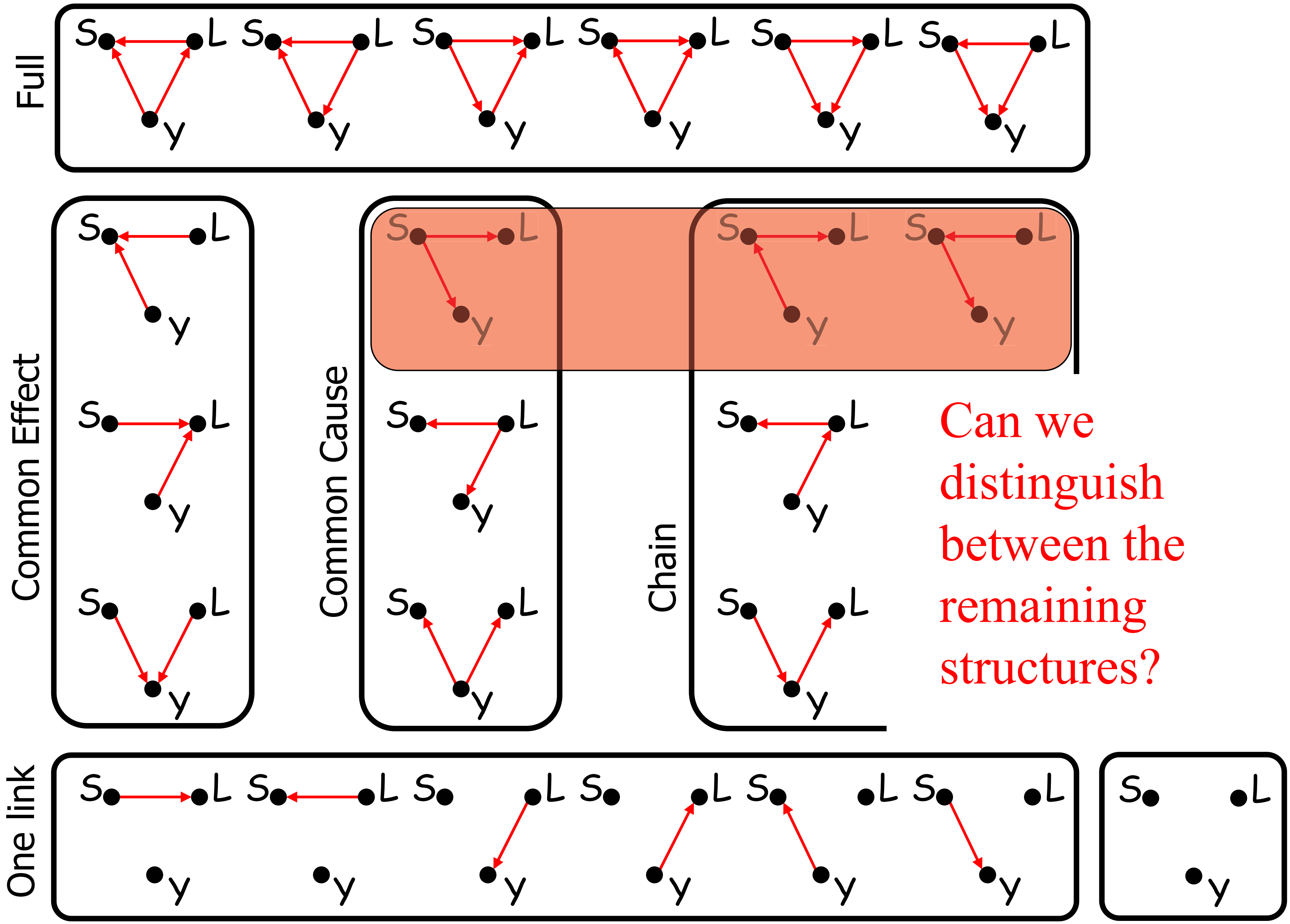
Image by MIT OCW.

Example

- What is the causal structure relating smoking, yellow teeth, and lung cancer?
- Expected partial (conditional) correlations:
  - smoking, yellow teeth | lung cancer: yes
  - smoking, lung cancer | yellow teeth: yes
  - yellow teeth, lung cancer | smoking: no
- Under faithfulness:
  - If two variables L and Y are conditionally independent given S, then L and Y must not be in each other's Markov blanket, and S must be in the Markov blanket of both.



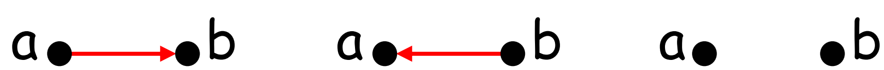
Full
Common Effect
Common Cause
Chain
Empty
One link
S
L
y



Can we distinguish between the remaining structures?

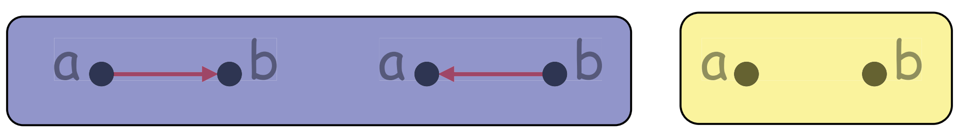
The limits of constraint-based inference

- *Markov equivalence class*: A set of causal graphs that cannot be distinguished based on (in)dependence relations.
- With two variables, there are three possible causal graphs and two equivalence classes:



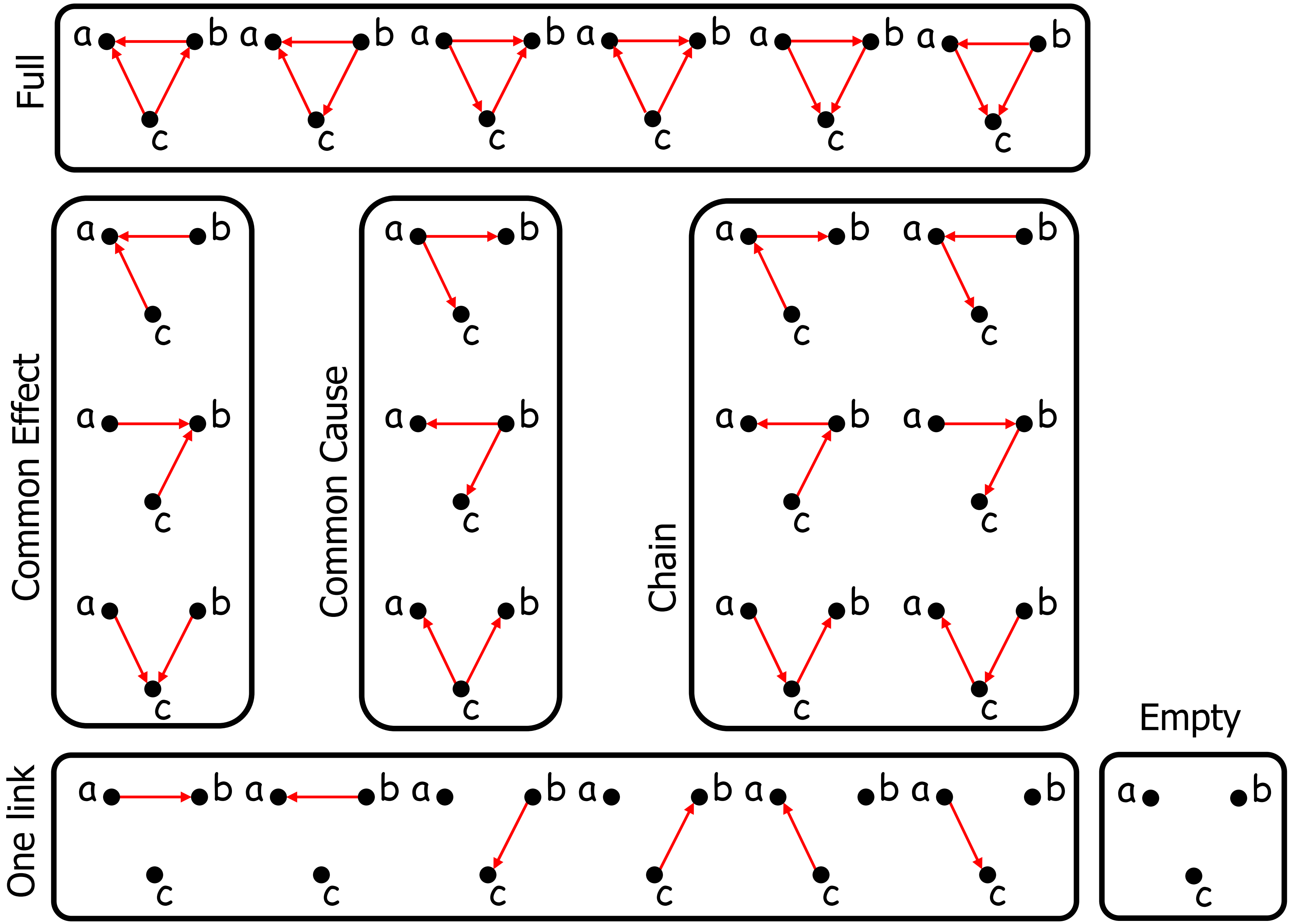
The limits of constraint-based inference

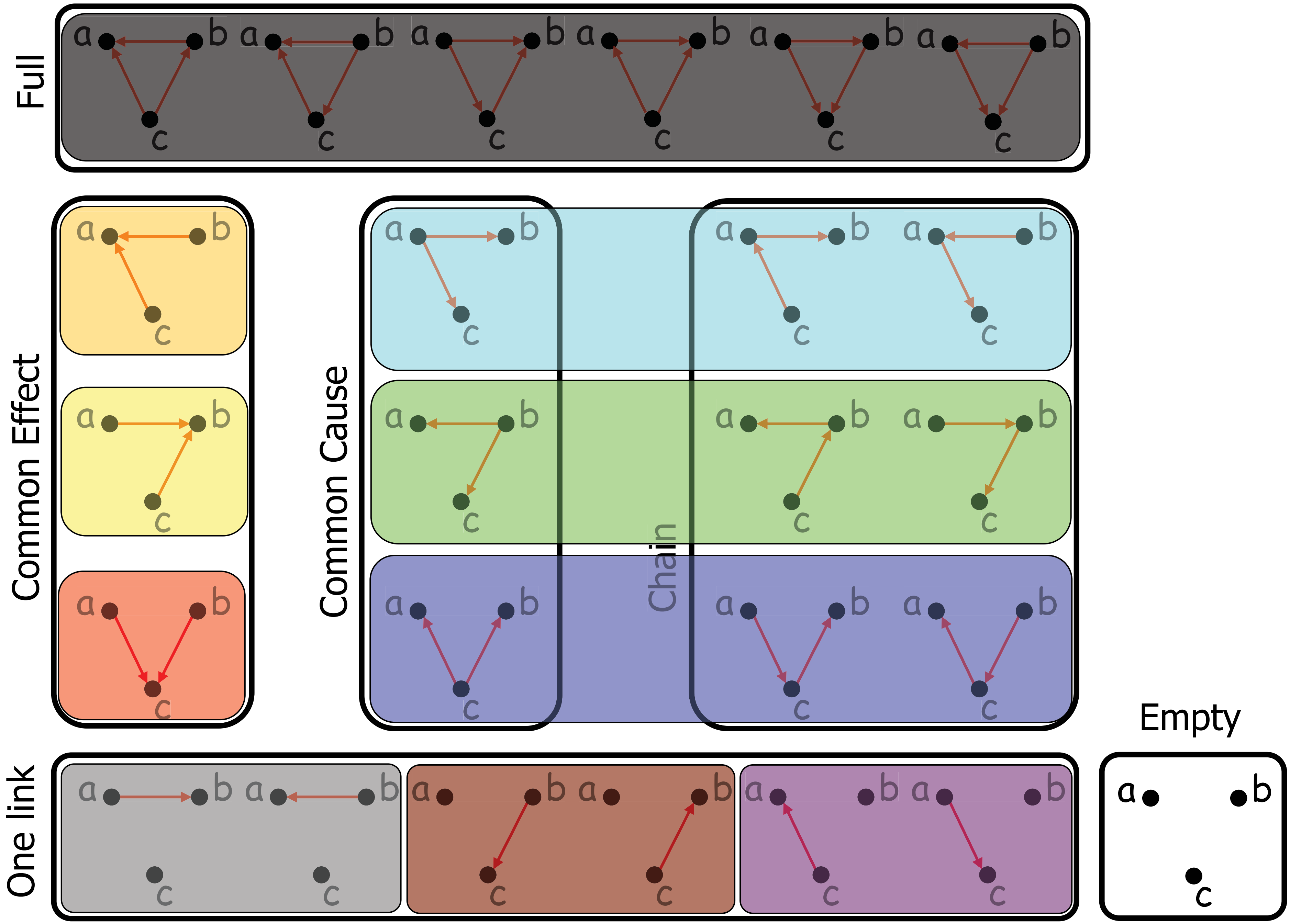
- *Markov equivalence class*: A set of causal graphs that cannot be distinguished based on (in)dependence relations.
- With two variables, there are three possible causal graphs and two equivalence classes:



A and *B* not independent.

A and *B* independent.


Full
Common Effect
Common Cause
Chain
Empty
One link
a
b
c



Full
Common Effect
Common Cause
Chain
One link
Empty
a
b
c

Additional sources of constraint

- Prior knowledge about causal structure
  - Temporal order
  - Domain-specific constraints
- Interventions
  - Exogenously clamp one or more variables to some known value, and observe other variables over a series of cases.

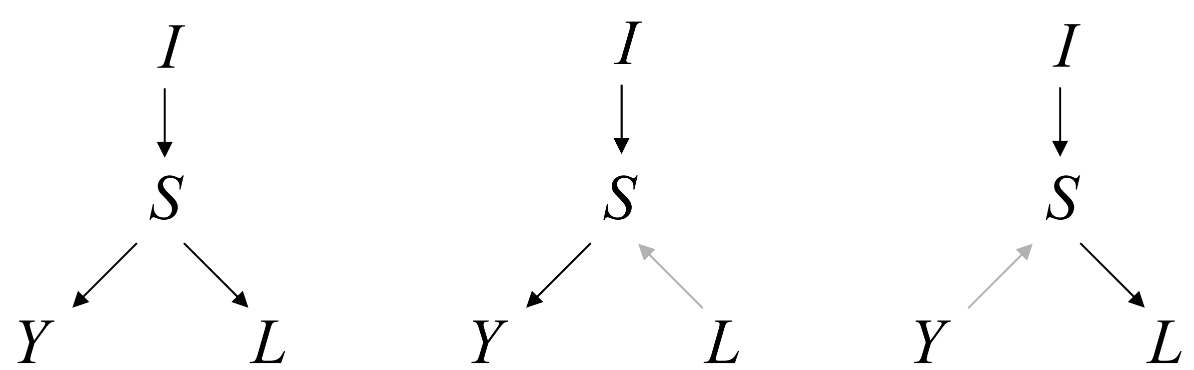
Interventions

- Example: Force a sample of subjects to smoke.
- Ideal interventions block all other direct causes of the manipulated variable:

$$
\begin{tikzcd}
& S \arrow[dl] \arrow[dr] & \\
Y & & L
\end{tikzcd}
\qquad
\begin{tikzcd}
& S \arrow[dl] & \\
Y & & L \arrow[ul]
\end{tikzcd}
\qquad
\begin{tikzcd}
& S \arrow[dr] & \\
Y \arrow[ur] & & L
\end{tikzcd}
$$

Interventions

- Example: Force a sample of subjects to smoke, and another sample to not smoke.
- Ideal interventions block all other direct causes of the manipulated variable:



Interventions

- Example: Force a sample of subjects to smoke, and another sample to not smoke.
- *Non-ideal* interventions simply add an extra cause that is under the learner's control:

$$\begin{tikzcd} & I \arrow[d] & \\ & S \arrow[dl] \arrow[dr] & \\ Y & & L \end{tikzcd} \qquad \begin{tikzcd} & I \arrow[d] & \\ & S \arrow[dl] & \\ Y & & L \arrow[ul] \end{tikzcd} \qquad \begin{tikzcd} & I \arrow[d] & \\ & S \arrow[dr] & \\ Y \arrow[ur] & & L \end{tikzcd}$$

Advantages of the constraint-based approach

- Deductive
- Domain-general
- No essential role for domain knowledge:
  - Knowledge of possible causal structures not needed.
  - Knowledge of possible causal mechanisms not used.

Disadvantages of the constraint-based approach

- Deductive
- Domain-general
- No essential role for domain knowledge:
  - Knowledge of possible causal structures not needed.
  - Knowledge of possible causal mechanisms not used.
- Requires large sample sizes to make reliable inferences.

Example

- What is the causal structure relating smoking, yellow teeth, and lung cancer?
- Epidemiological Data:

Patient	Smoking?	Yellow teeth?	Lung Cancer?
1	yes	yes	yes
2	yes	yes	no
3	yes	no	yes
4	no	no	no
5	yes	yes	yes
6	yes	no	no
7	yes	no	yes
8	no	no	no

. . . .

Computing (in)dependence

- Standard methods based on χ^2 test:

	$V=0$	$V=1$
$U=0$	a	c
$U=1$	b	d

$$\chi^2 \quad \frac{(a+b+c+d)(a\times d-b\times c)^2}{(a+b)(c+d)(a+c)(b+d)}$$

significantly > 0: not independent

not significantly > 0: independent

Computing (in)dependence

- Are smoking and yellow teeth independent?

	Y=0	Y=1
S=0	2	0
S=1	3	3

$$\chi^2 = 1.6,\ p = 0.21$$

Computing (in)dependence

- Are smoking and lung cancer independent?

	L=0	L=1
S=0	2	0
S=1	2	4

$$\chi^2 = 2.67, p = 0.10$$

Computing (in)dependence

- Are lung cancer and yellow teeth conditionally independent given smoking?

$\boxed{S=1}$

	$L=0$	$L=1$
$Y=0$	1	2
$Y=1$	1	2

$\chi^2 = 0,\ p = 1.0$

$\boxed{S=0}$

	$L=0$	$L=1$
$Y=0$	2	0
$Y=1$	0	0

$\chi^2 = \text{undefined}$

Disadvantages of the constraint-based approach

- Deductive
- Domain-general
- No essential role for domain knowledge:
  - Knowledge of possible causal structures not needed.
  - Knowledge of possible causal mechanisms not used.
- Requires large sample sizes to make reliable inferences.

The Blicket detector

Image removed due to copyright considerations. Please see:
Gopnick, A., and D. M. Sobel. "Detecting Blickets: How Young Children use Information about Novel Causal Powers in Categorization and Induction." *Child Development* 71 (2000): 1205-1222.

Image removed due to copyright considerations. Please see:
Gopnick, A., and D. M. Sobel. "Detecting Blickets: How Young Children use Information about Novel Causal Powers in Categorization and Induction." *Child Development* 71 (2000): 1205-1222.

The Blicket detector

- Can we explain these inferences using constraint-based learning?
- What other explanations can we come up with?