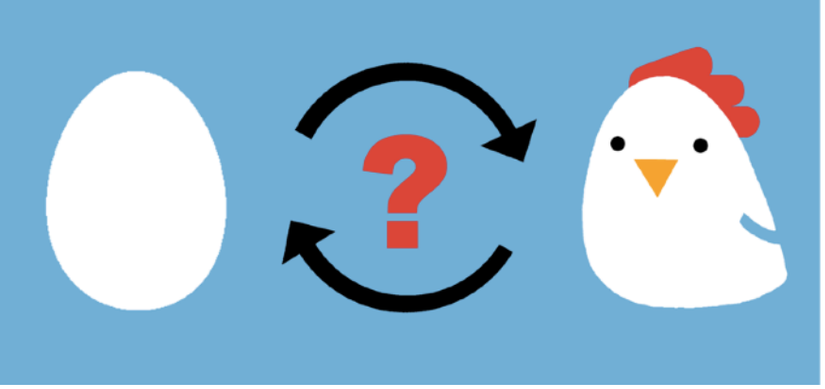


Causality and Machine Learning
(80-816/516)

Classes 7 (Feb 4, 2025)

Graphical Models and Causal Representations

Instructor:

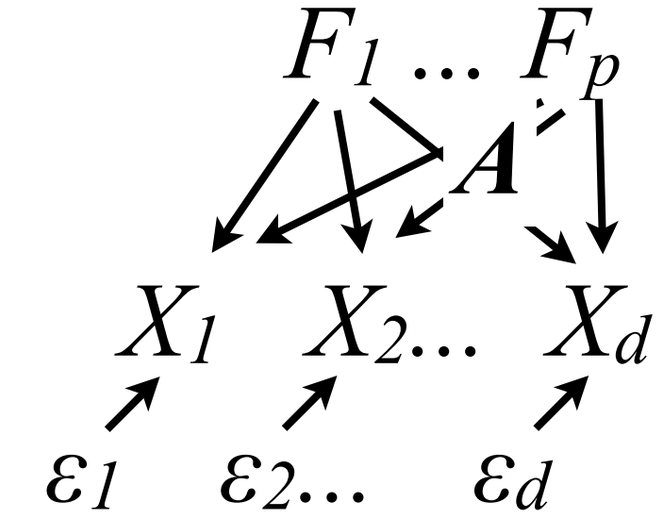
Kun Zhang (kunz1@cmu.edu)

Zoom link: https://cmu.zoom.us/j/8214572323)

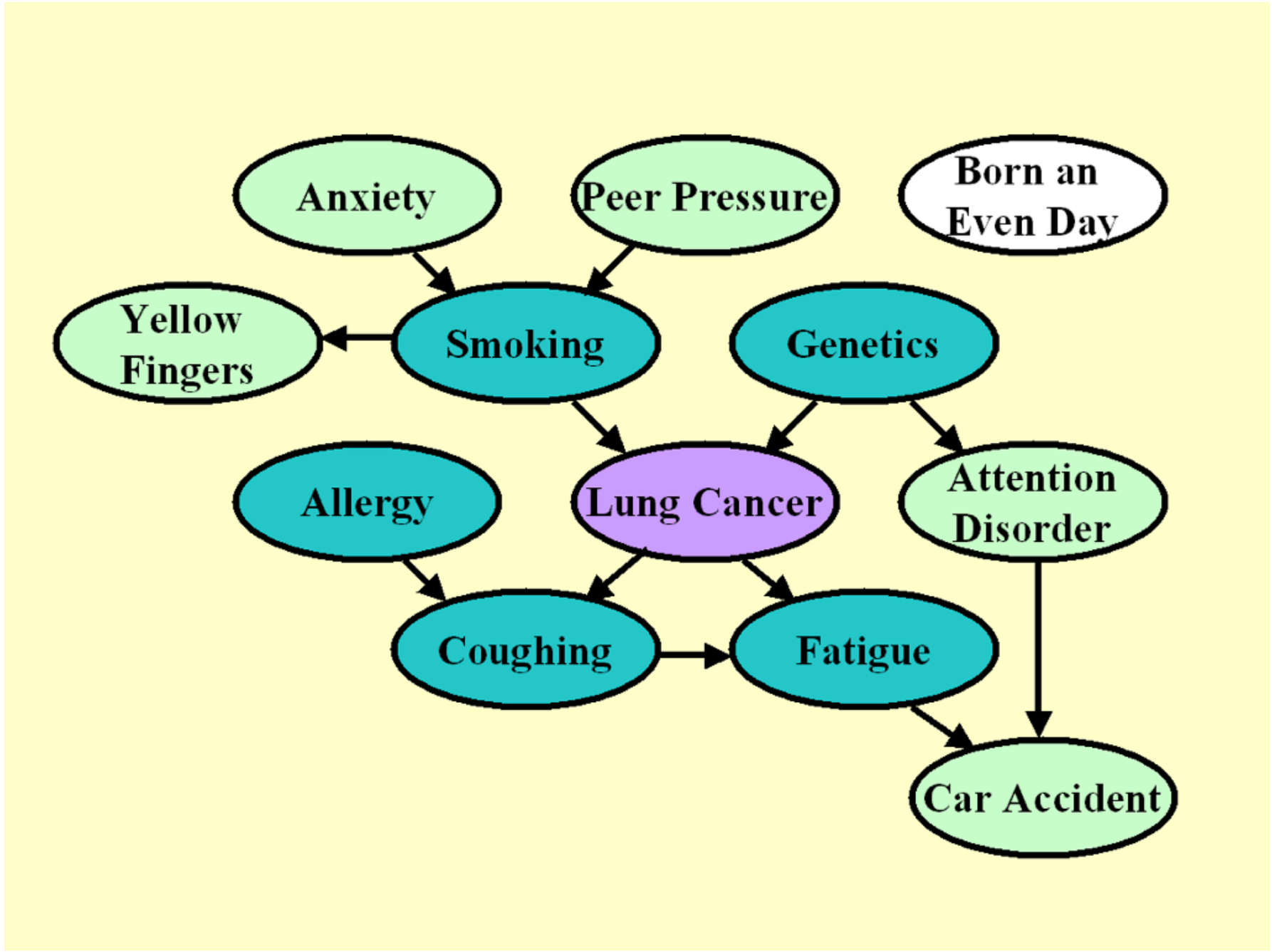
Office Hours: W 3:00–4:00PM (on Zoom or in person); other times by appointment

Outline

- Graphical models
- d-separation
- Connection between conditional independence in graphs and that in data?
- Causal interpretations?



A Graphical Representation



Probabilities & Graphical Models

- Why graphical models?
  - flexible, powerful and compact way to model relationships between random variables and do inference
- Why probabilities?

The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

James Clerk Maxwell (1850)

- Why causal discovery?
  - understanding, manipulation, prediction, fusion...

I would rather discover one true cause than gain the kingdom of Persia.

−Democritus (460 B.C. – 370 B. C.)

Graphical Models

- A **graph** comprises nodes (also called vertices) connected by links (also known as edges or arcs)
- Probabilistic **graphical** models: compactly encoding a complex distribution
  - Node: a random variable (or group of random variables)
  - Links: direct probabilistic interactions between them
- We mainly consider **directed acyclic** graphs (DAGs)

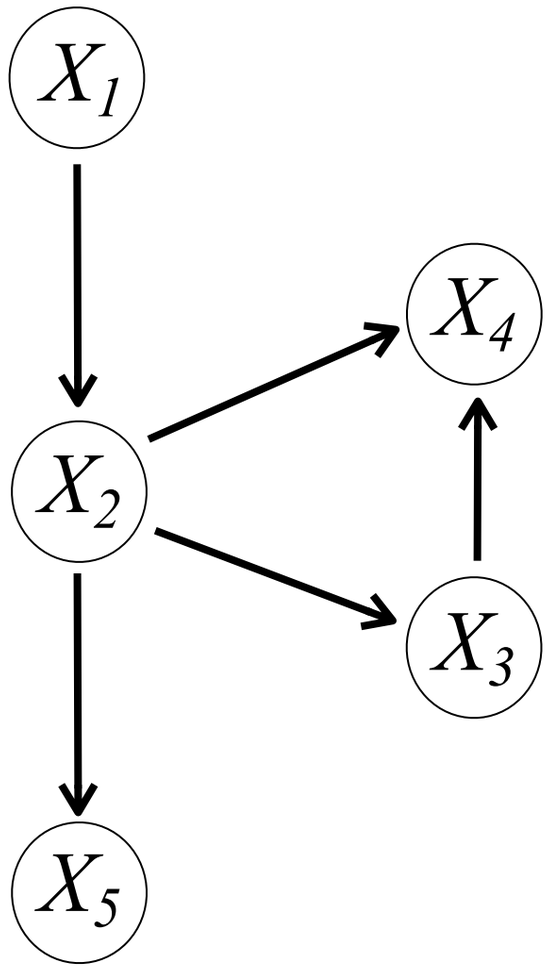
Probability theory + graph theory

represent uncertainty & interface models to data

intuitively appealing interface for humans

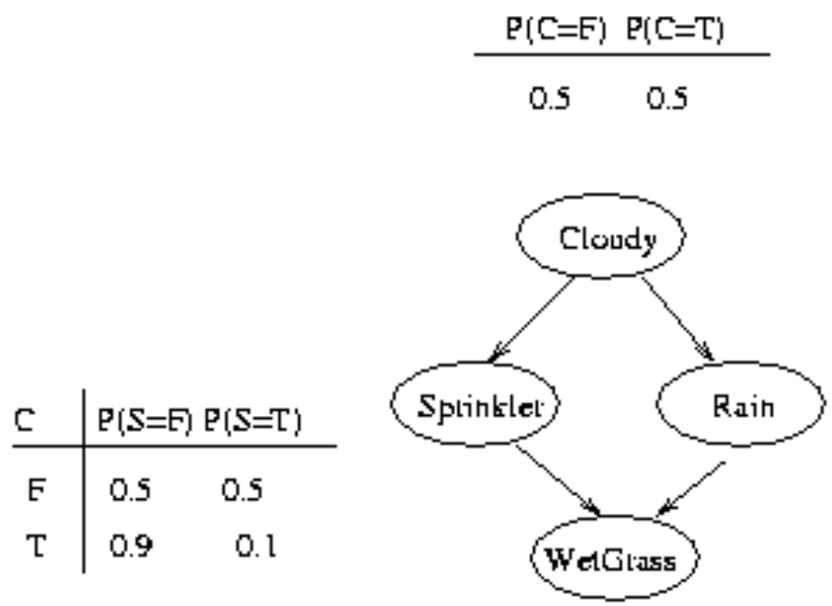
Terms

Terms:
nodes, edge, adjacent, path;
Parents $\mathrm{PA}(X_v)$, children,
spouses, ancestors,
descendants, Markov
blanket



Directed Graphical Models

- Also known as Bayesian networks or belief nets
- Two components
  - Graph structure (qualitative specification)
    - prior knowledge of causal/modular relationships
    - expert knowledge
    - learning from data
  - Conditional probability distributions (CPDs)
    - discrete variables : conditional distribution tables (CPTs)
    - continuous variables: SEMs



P(C=F)	P(C=T)
0.5	0.5

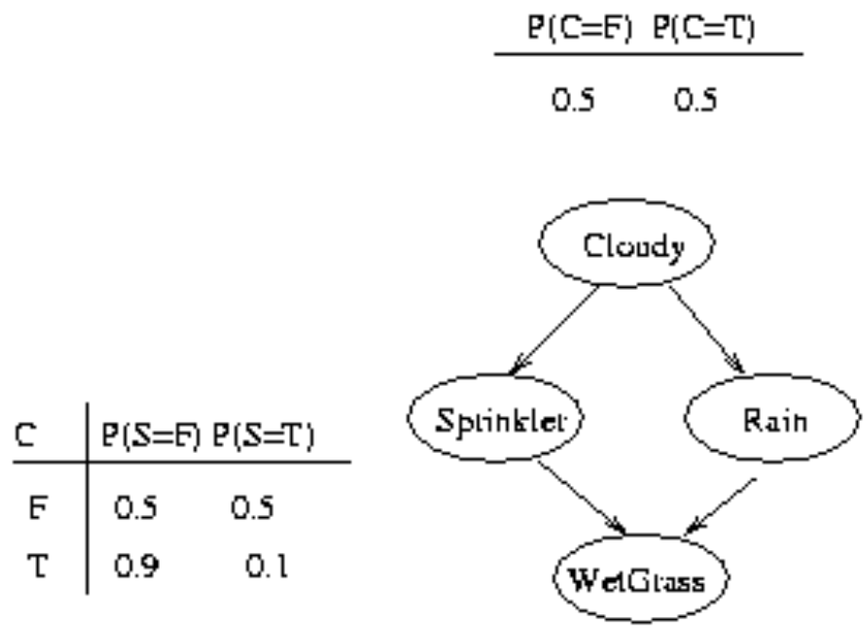
C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Tasks Related to Bayesian Networks

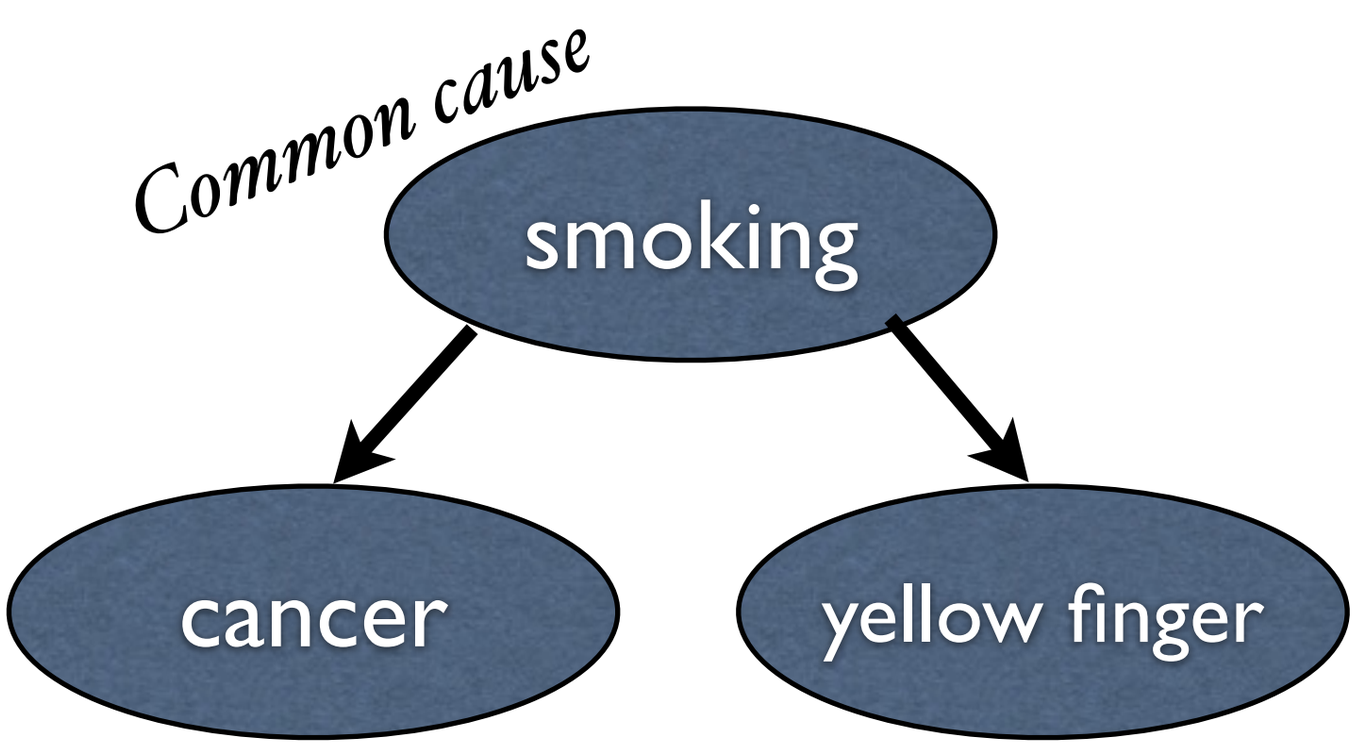
- **Probabilistic inference**: Calculate P(variables of interest | observed variables)
  - Most common task where we want to use Bayesian networks
    - How to find $P(S=1|W=1)$? $P(R=1|W=1)$?
- **Parameter learning**
- **Structure learning**: Learning the structure of the graphical model from observations

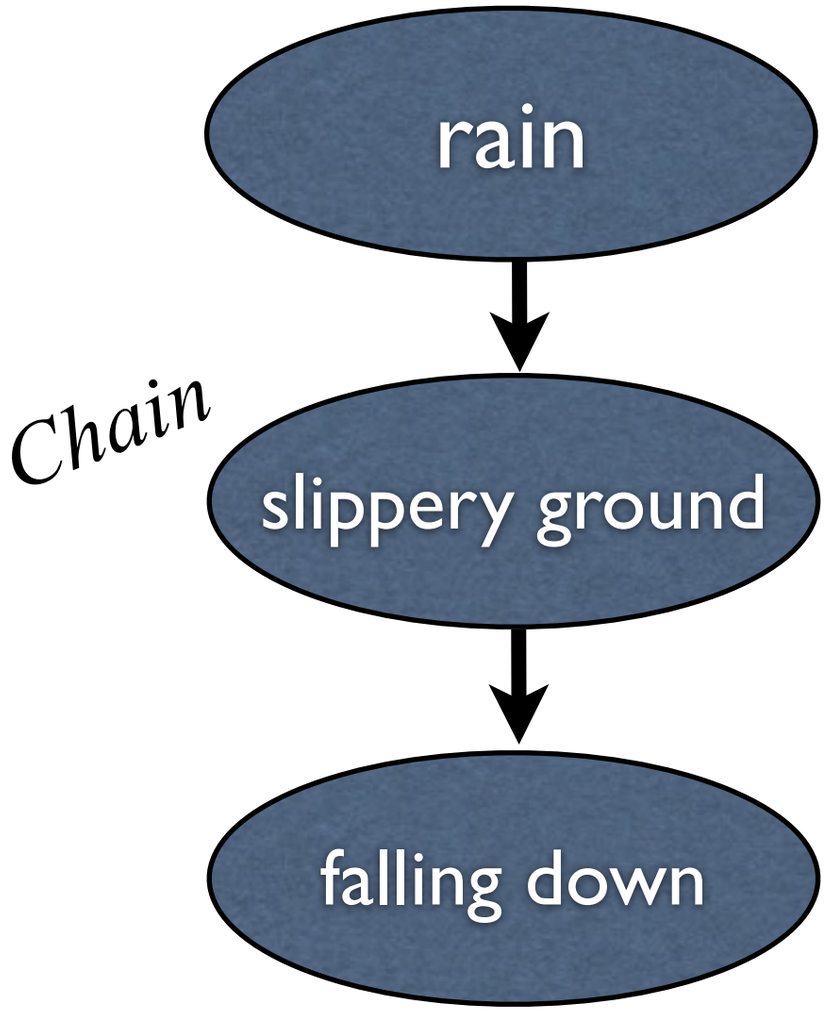


Bayesian Networks: Story

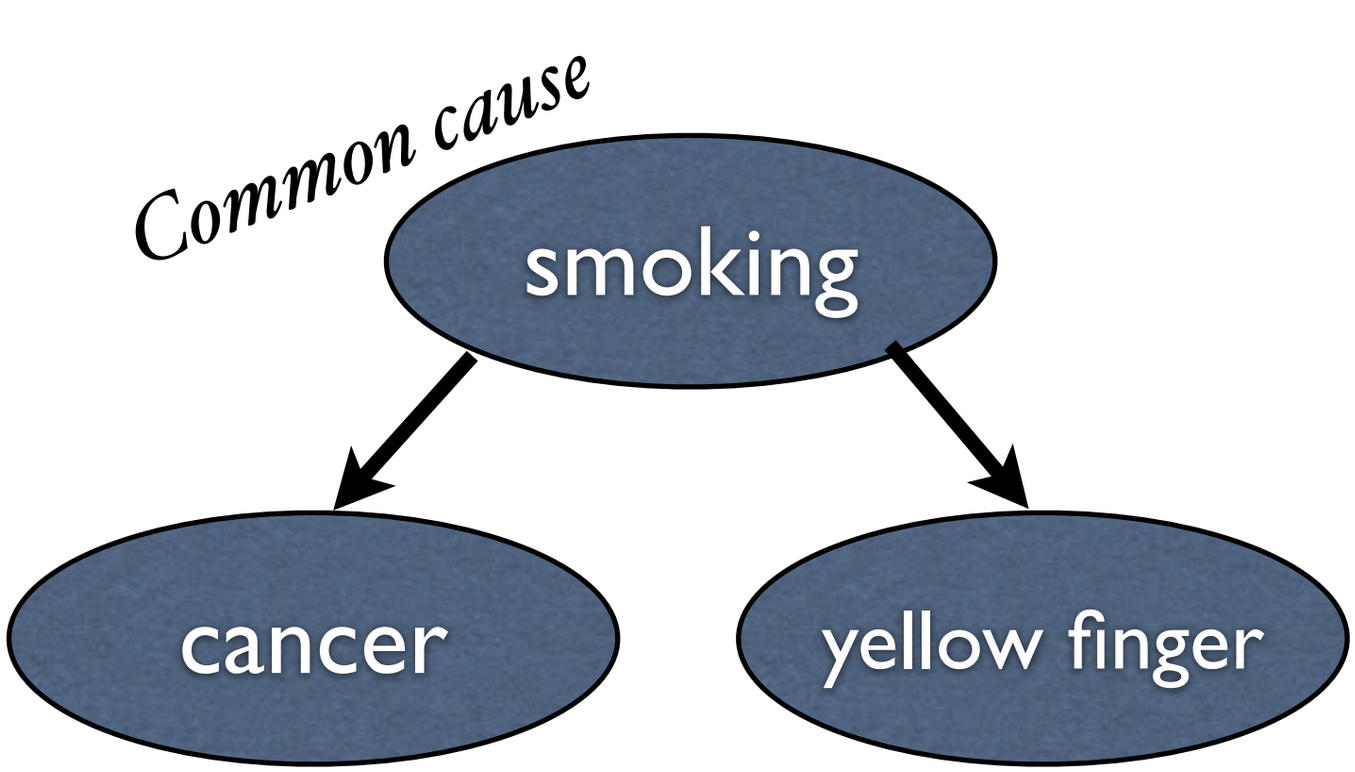
- Breakthrough in early 1980s (by Pearl et al.)
- In a joint probability distribution, every variable is, in general, related to all other variables.
- Pearl and others realized:
  - It is often reasonable to make the assumption that each variable is directly related to only a few other variables
  - This leads to **modularity**: Allowing decomposing a complex model into small manageable pieces
  - Giving rise to **Bayesian networks**

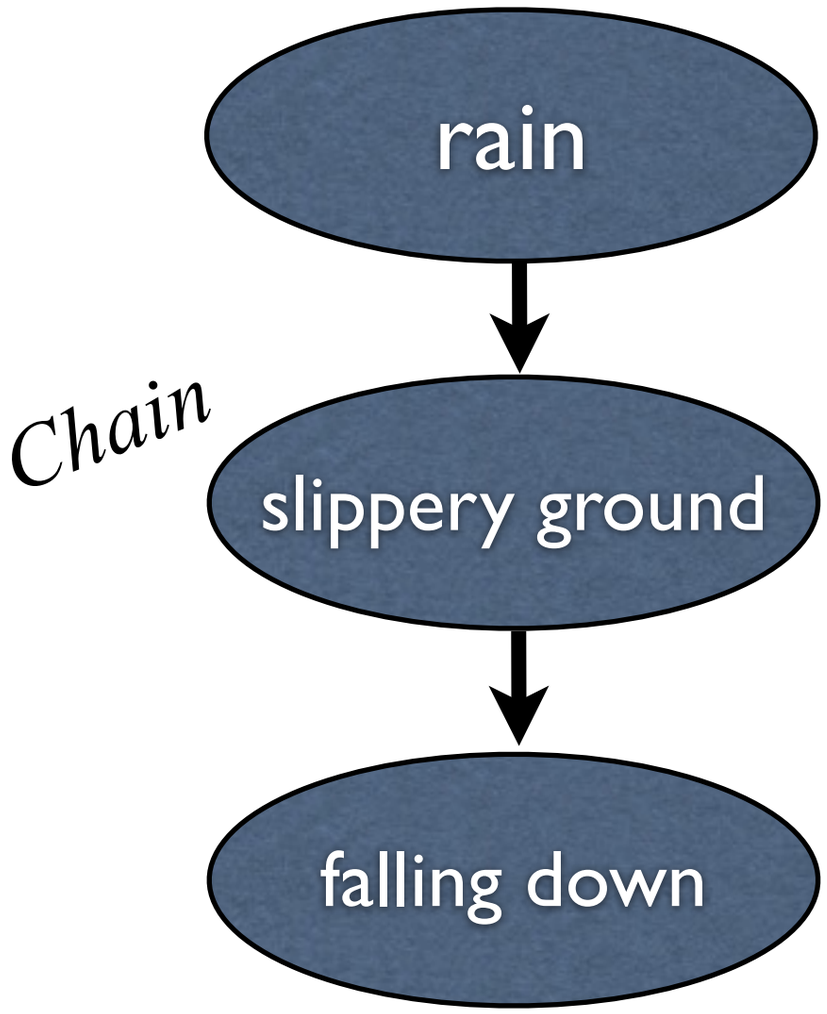
What Independence Relationships Can You See?





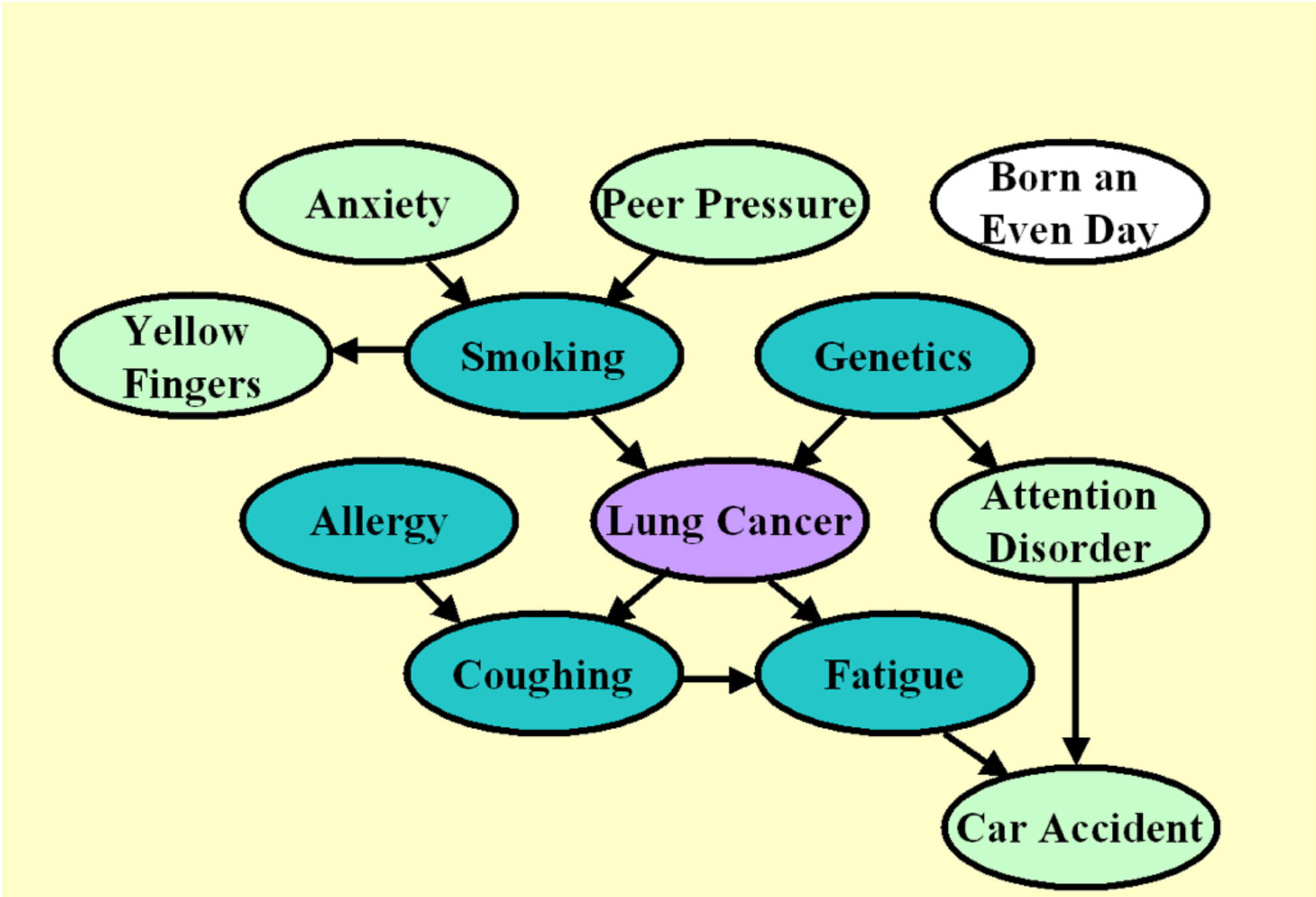
(Local) Markov Condition





- Each variable is independent from its non-descendants given its parents

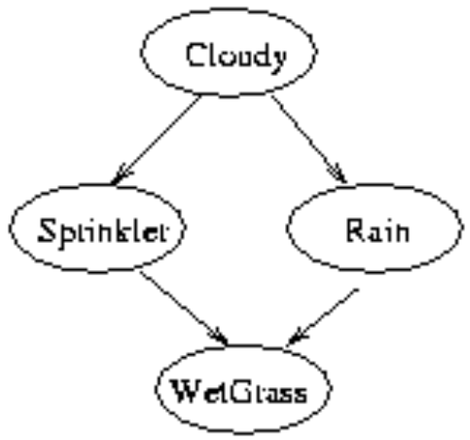
For Instance, What Independence Relations can You See?



Factorization According to Directed Graphs

- Chain rule of probability gives

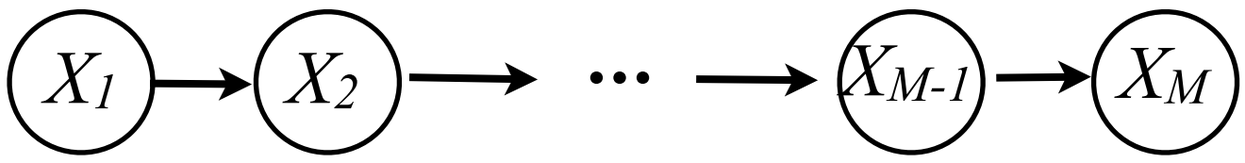
  P(C,S,R,W) = P(C) P(S|C) P(R|C,S) P(W|C,S,R)



- According to the CI relationships:

  P(C,S,R,W) = P(C) P(S|C) P(R|C) P(W|S,R)

- The graph structure allows us to represent the joint distribution more compactly (*Markov factorization* or *Markov decomposition* of the joint distribution):
  - $P(X_1,\ldots,X_n) = \prod_i P(X_i \mid PA_i)$
- Remember this example?

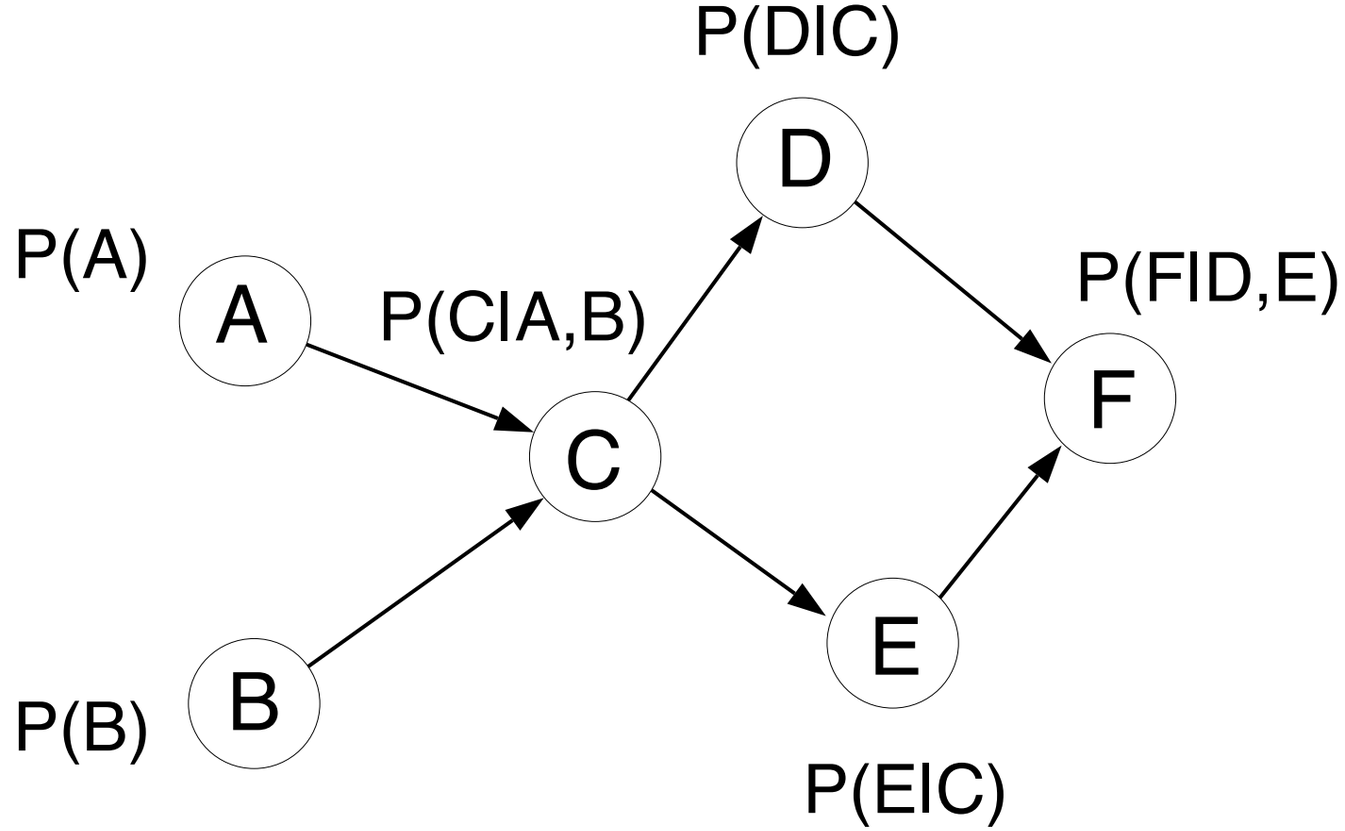


If we aim to represent causal info, is CI info enough?

X→Y or X←Y?

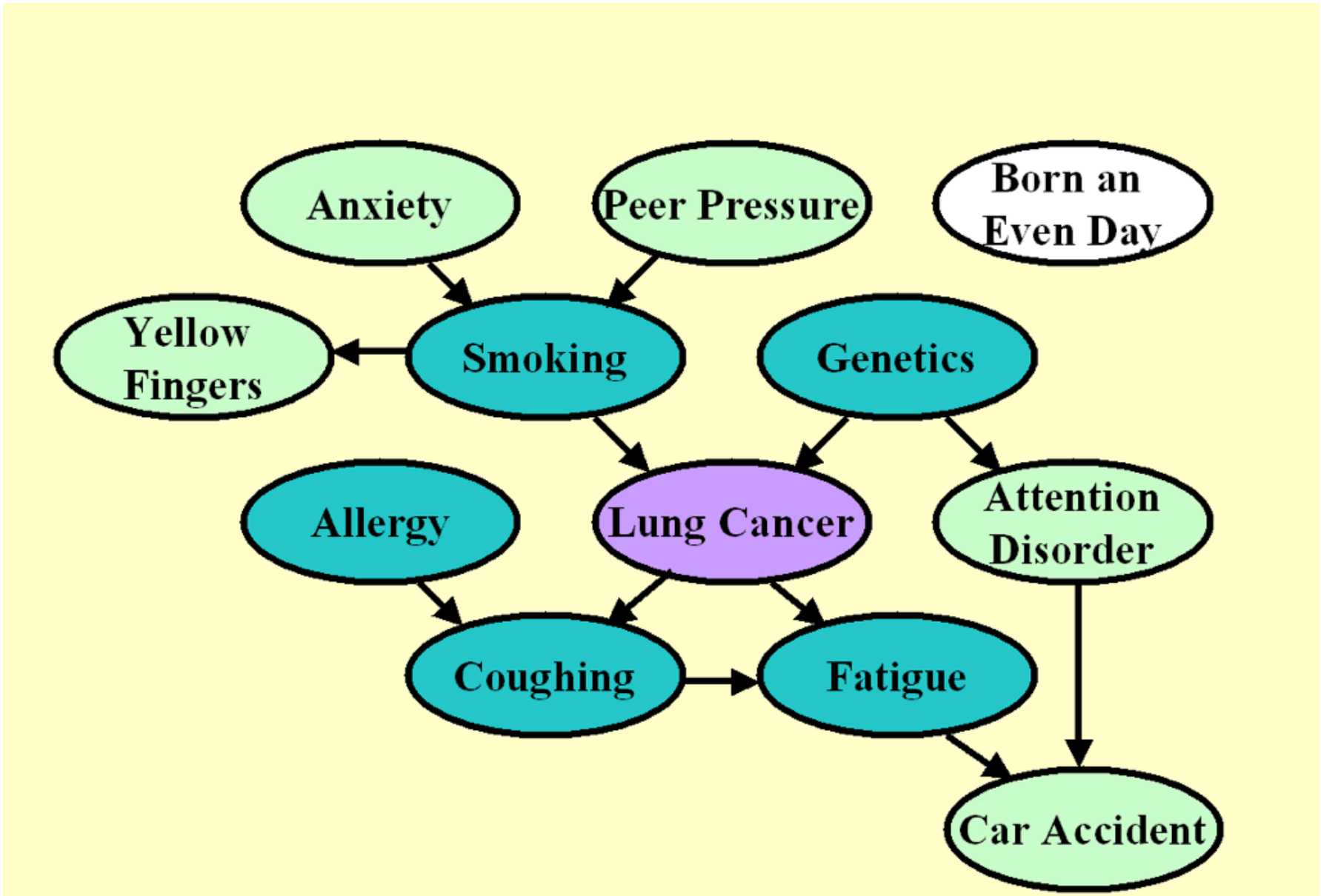
Factorization According to Directed Graphs: Procedure

- Associate a conditional probability with each node
- Then take the product of the local probabilities to yield the global probabilities



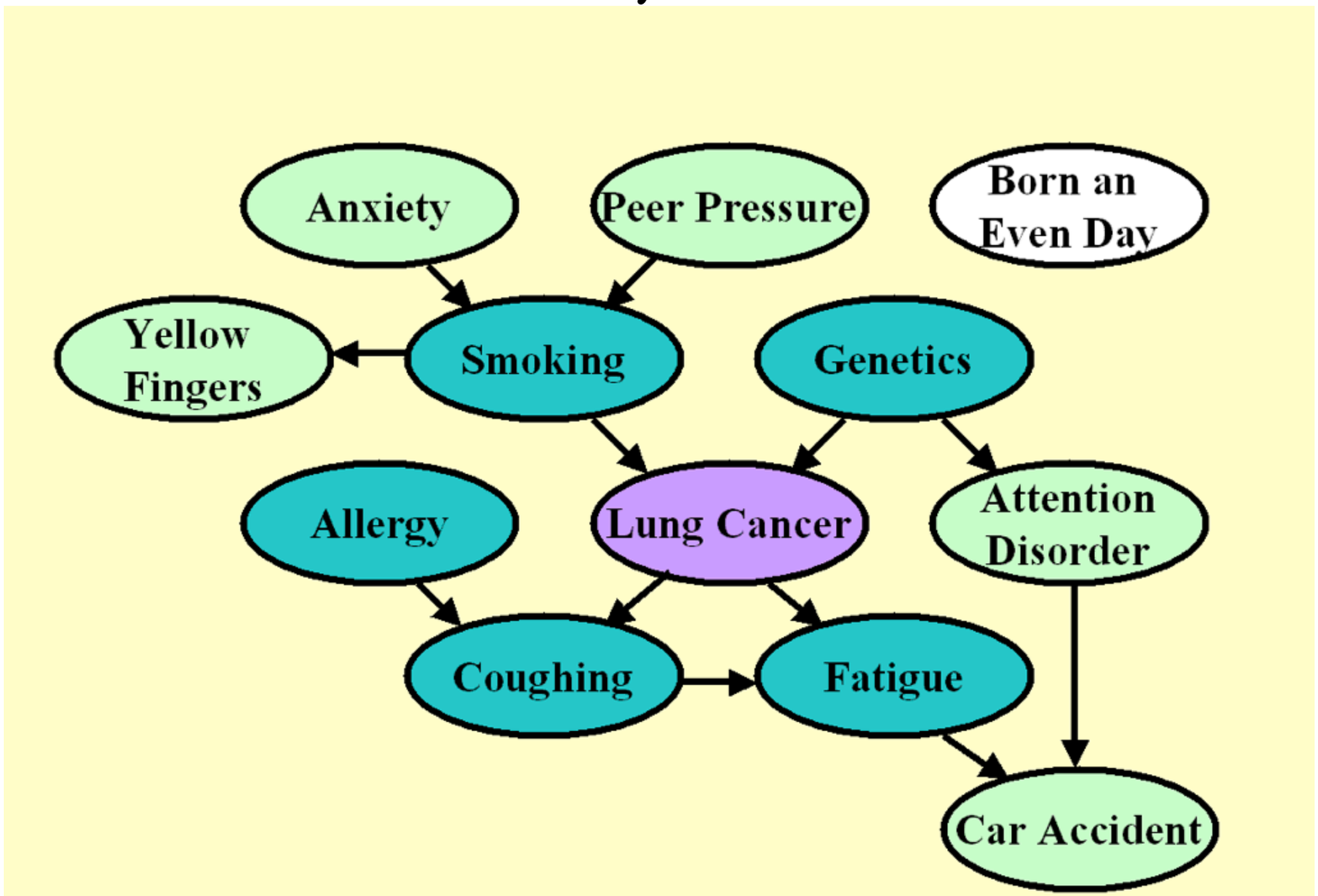
Is Local Markov Condition Enough?

- Can we see whether two arbitrary variables, X and Y, are conditionally independent given an arbitrary set of variables, $\mathbf{Z}$?



D-Separation Tells Conditional Independence

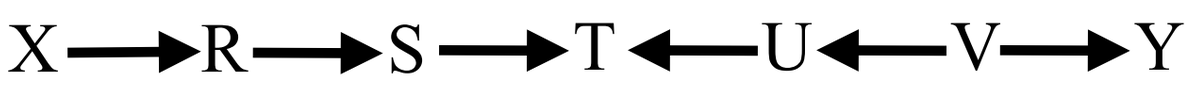
- If every path from a node in **X** to a node in **Y** is d-separated by **Z**, then **X** and **Y** are always conditionally independent given **Z**
- d: directional… You will see why



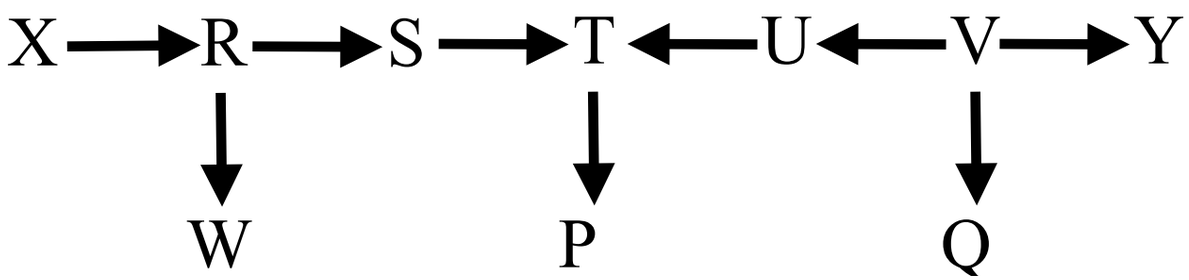
D-Separation

- A set of nodes **Z** d-separates two sets of nodes **X** and **Y** if every **path** from a node in **X** to a node in **Y** is blocked given **Z**.

- A path p is blocked by a set of nodes **Z** if
  - p contains a chain $i \rightarrow m \rightarrow j$ or a common cause $i \leftarrow m \rightarrow j$ such that the middle node m is in **Z**, **or**
  - p contains a collider $i \rightarrow m \leftarrow j$ such that the middle node m is in not **Z** and no descendant of m is in **Z**



X and Y d-separated by {R,V}?
S and U d-separated by {R,V}?



X and Y d-separated by {R, P}?

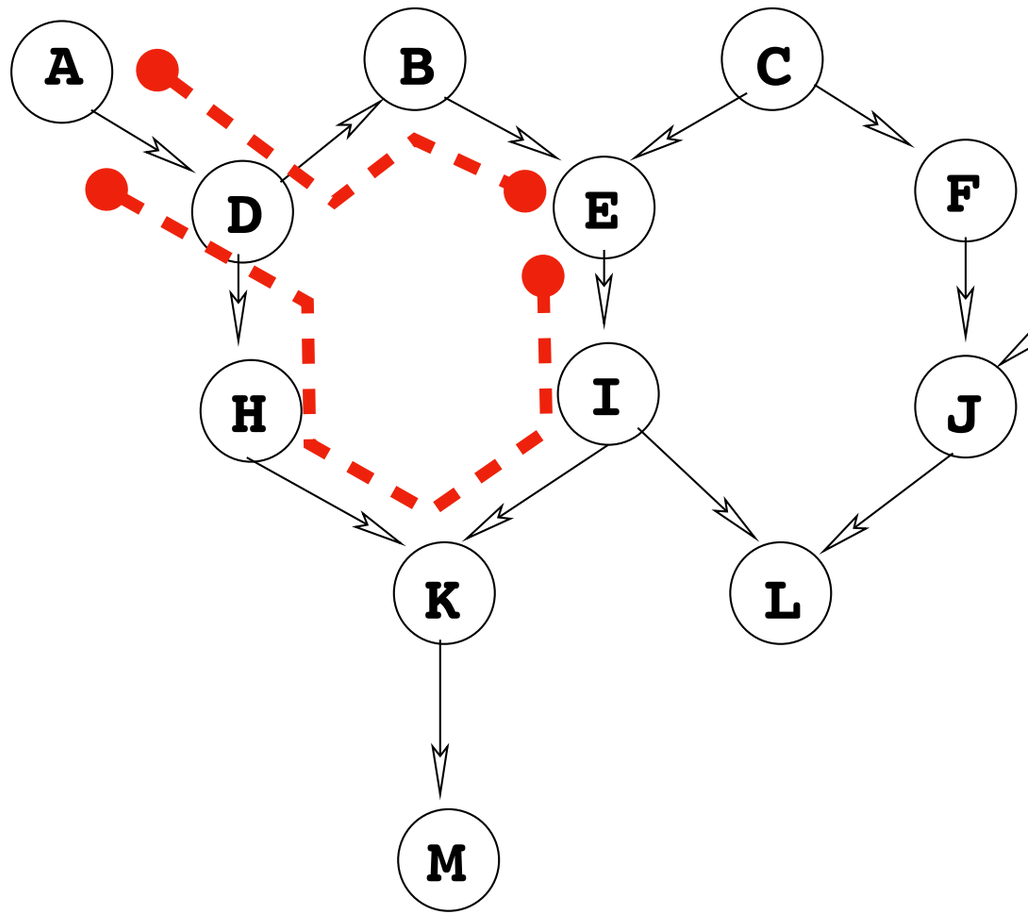
D-Separation

- A set of nodes **Z** d-separates two sets of nodes **X** and **Y** if every **path** from a node in **X** to a node in **Y** is blocked given **Z**.

- A path p is blocked by a set of nodes **Z** if

  - p contains a chain $i \rightarrow m \rightarrow j$ or a common cause $i \leftarrow m \rightarrow j$ such that the middle node m is in **Z**, **or**

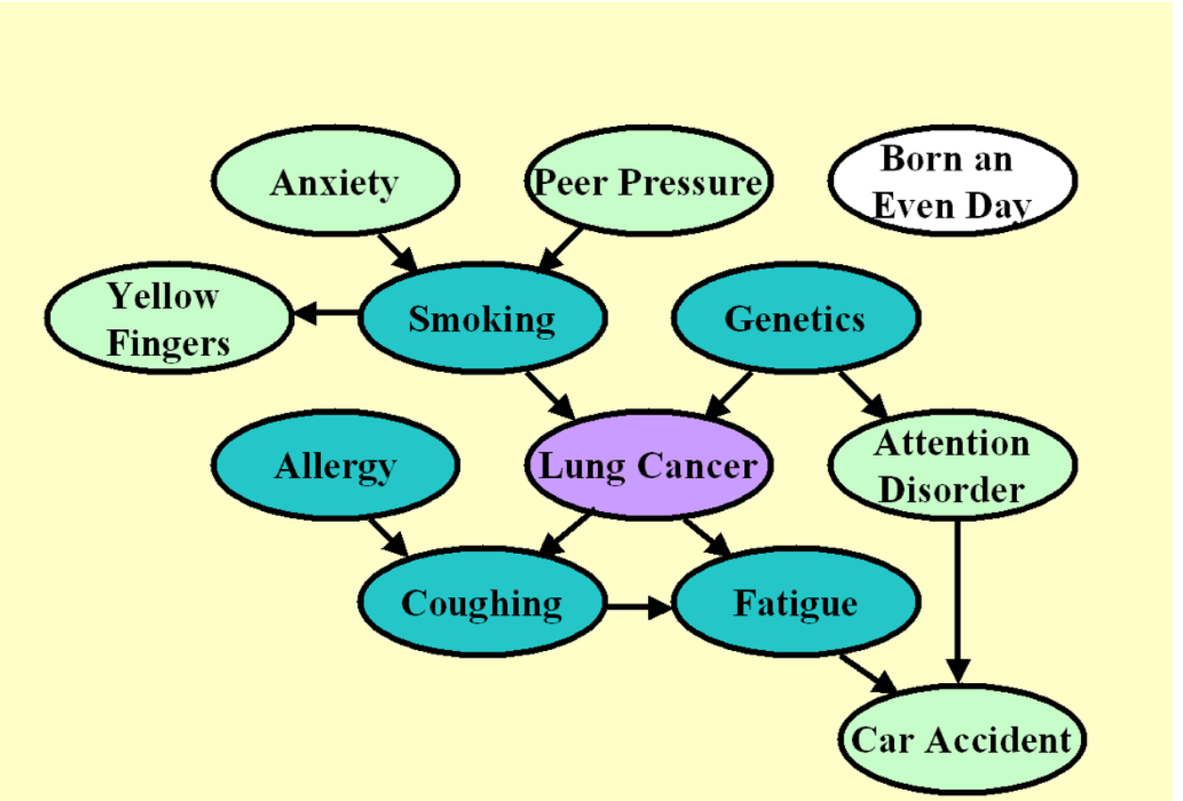
  - p contains a collider $i \rightarrow m \leftarrow j$ such that the middle node m is not in **Z** and no descendant of m is in **Z**

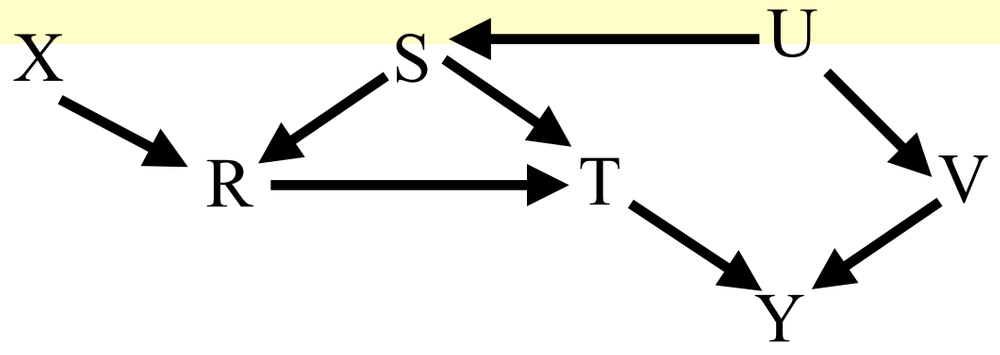


A and E d-separated by B ?
A and E d-separated by {B, M} ?

D-Separation: Intuition

- Suppose **X** and **Y** are d-separated by **Z**
- Then if you fix **Z**, **X** and **Y**
  - do not cause each other and
  - do not share a common cause
- **X** and **Y** are independent (conditional on **Z**)!

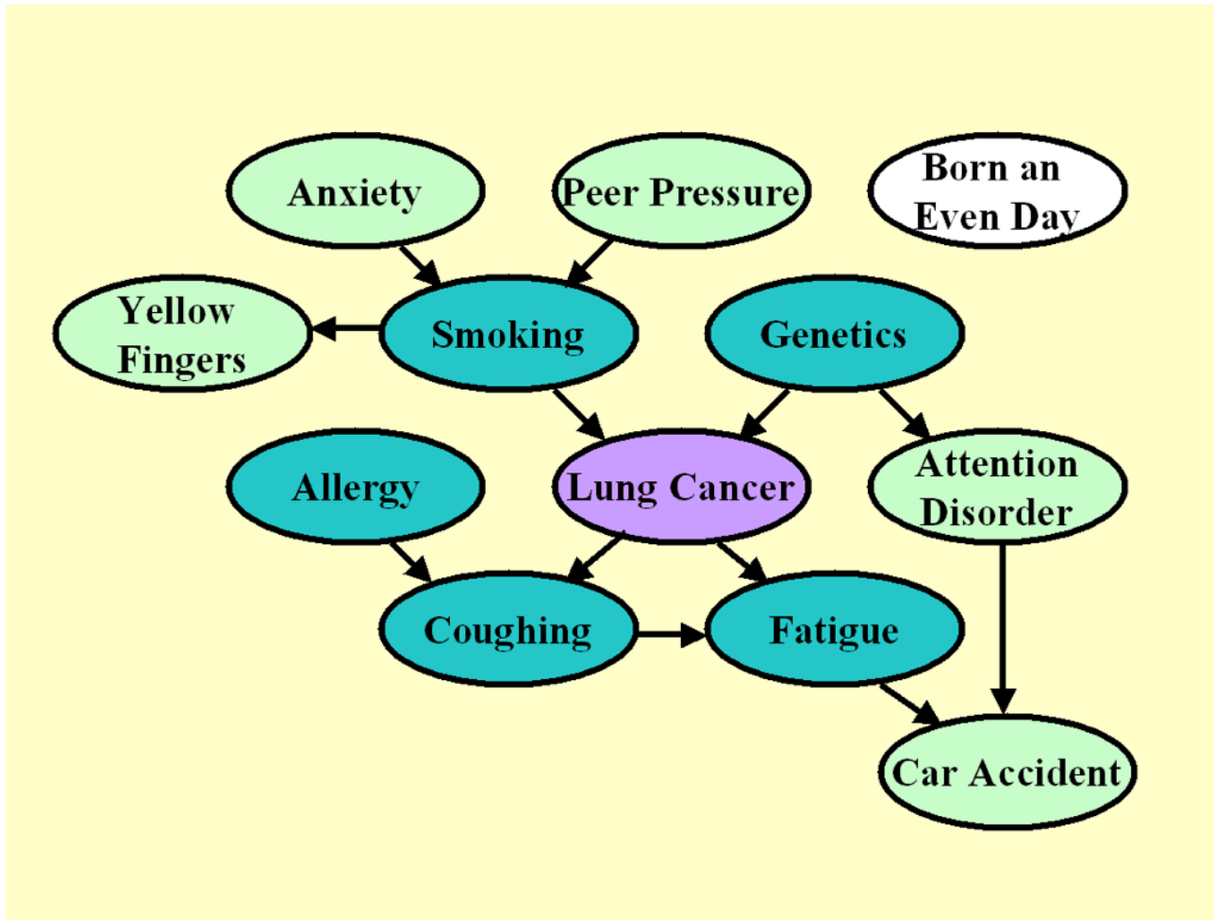




1. X and Y d-separated by {R}?
2. X and Y d-separated by {R, T}?
3. X and Y d-separated by {T, V}?
4. X and V d-separated by ∅ ?

Local & Global Markov Conditions

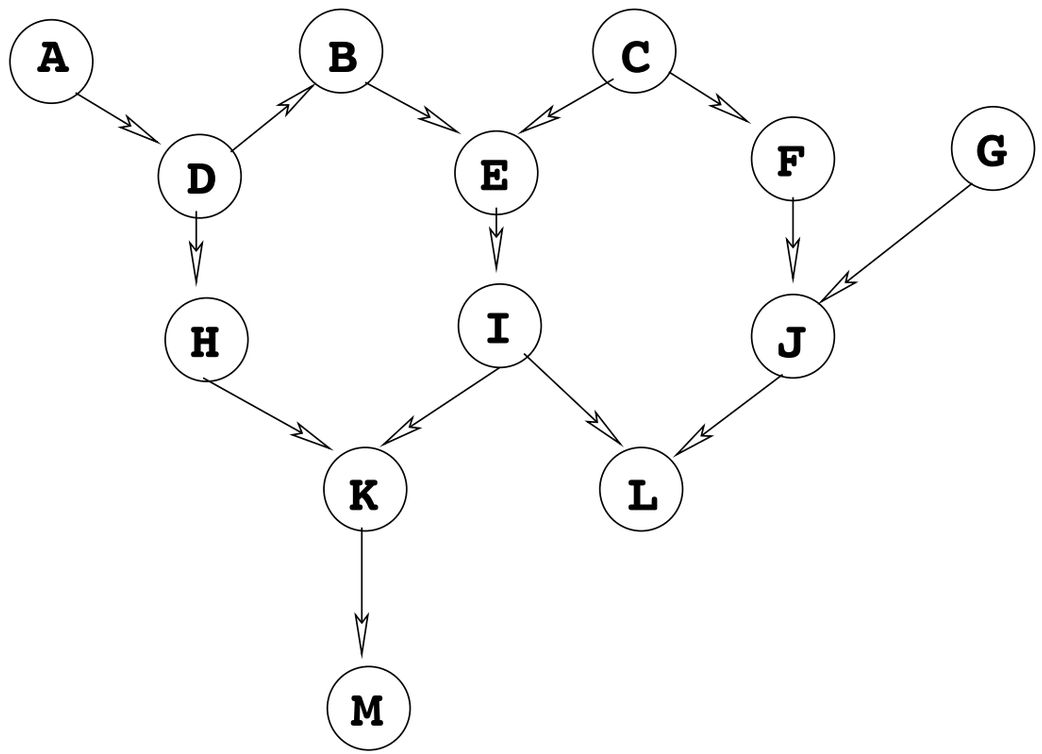
- **Local** Markov condition:
  - In a DAG, a variable X is independent of all its non-descendants given its parents



- **Global** Markov condition:
  - Given a DAG, let X and Y be two variables and $\mathbf{Z}$ be a set of variables that does not contain X or Y. If $\mathbf{Z}$ d-separates X and Y, then $X \perp\!\!\!\perp Y \mid \mathbf{Z}$.
- Actually equivalent on DAGs!

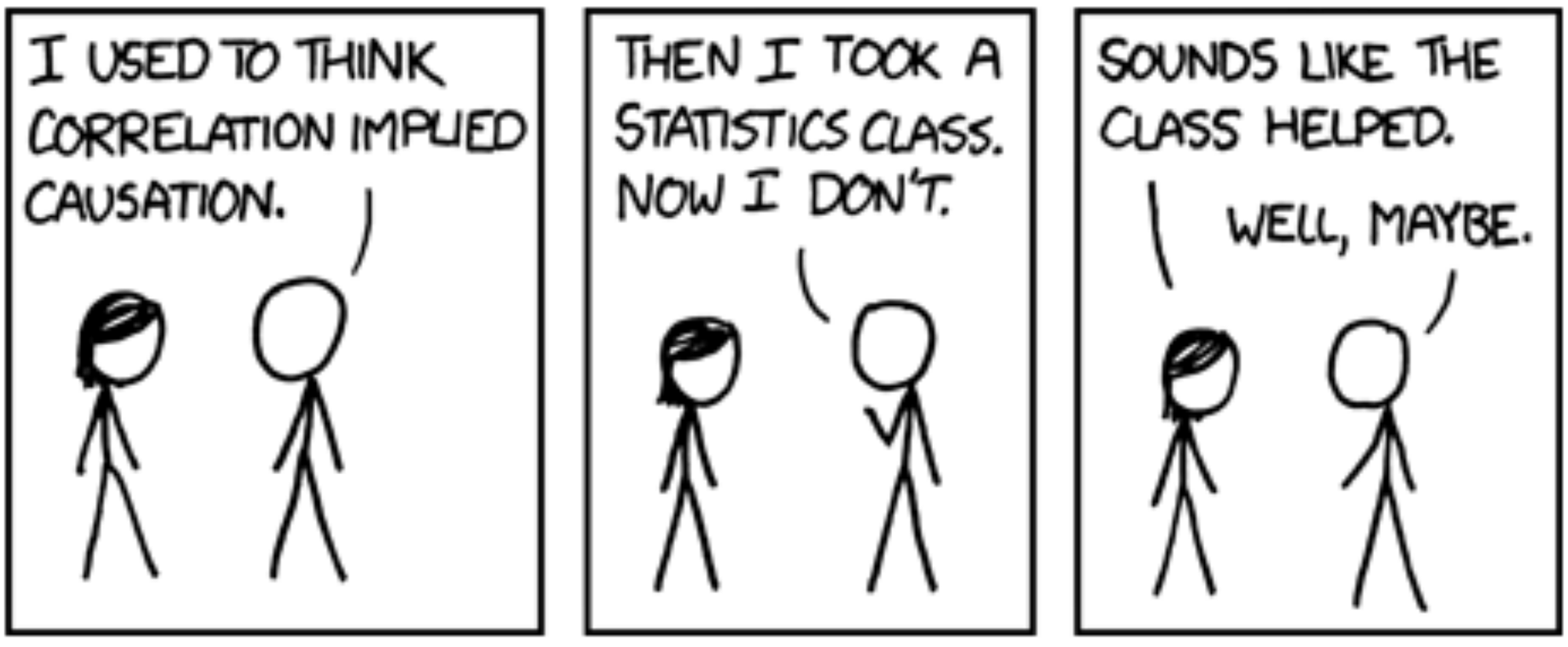
Markov Blanket

- In a DAG, the Markov Blanket of a node X is the set consisting of
  - Parents of X
  - Children of X
  - Parents of children (i.e., spouses) of X
- In a DAG, a variable X is conditionally independent from all other variables given its Markov Blanket
  - Implied by d-separation...
- The Markov blanket of I ?



Causality vs. Dependence

- Causality ➜ dependence ! Dependence ➜ causality



(http://imgs.xkcd.com/comics/correlation.png)

X and Y are associated iff

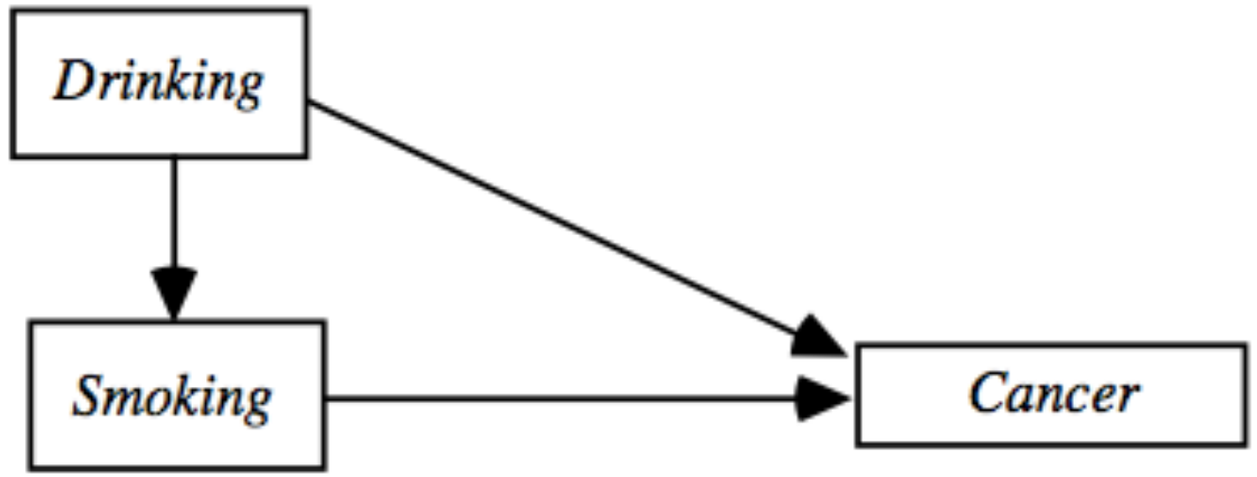
$\exists x_1 \neq x_2 \; P(Y|X=x_1) \neq P(Y|X=x_2)$

X is a cause of Y iff

$\exists x_1 \neq x_2 \; P(Y|do\ X=x_1) \neq P(Y|do\ X=x_2)$

Representing Causal Relations with Directed Graphs

- A directed graph represents a causally sufficient causal structure

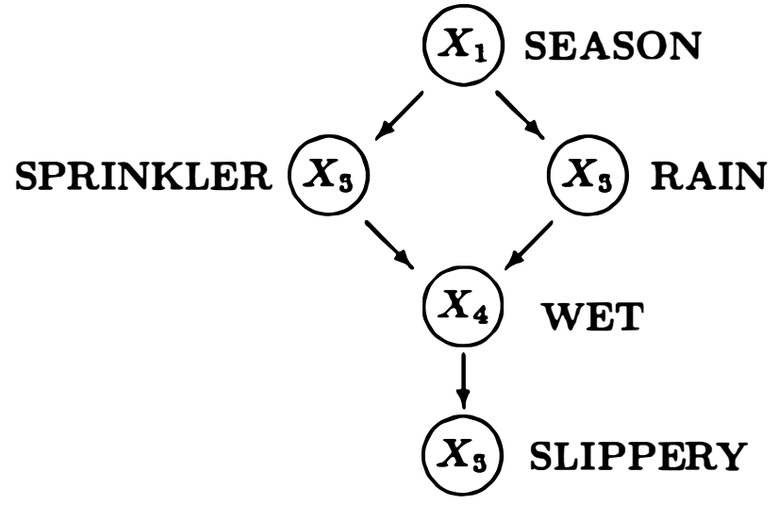


(adapted from "Causation, Prediction, and Search" by SGS, 1995)

- Directed edge from A to B means A is a direct cause of B relative to the given variable set $\boldsymbol{V}$

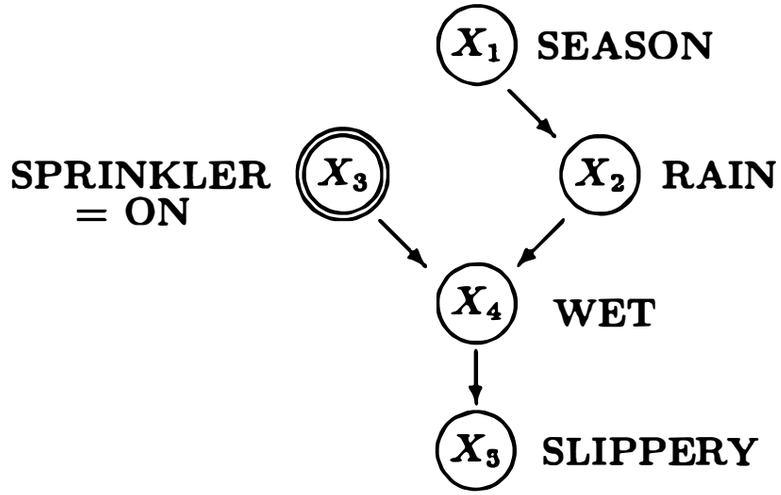
Causal Bayesian Networks (CBNs)

- Bayesian networks: DAGs
- Causal Bayesian networks
  - More meaningful & able to **represent and respond to external or spontaneous changes**



Let $P_x(V)$ be the distribution of V resulting from intervention $do(X=x)$. A DAG G is a CBN if

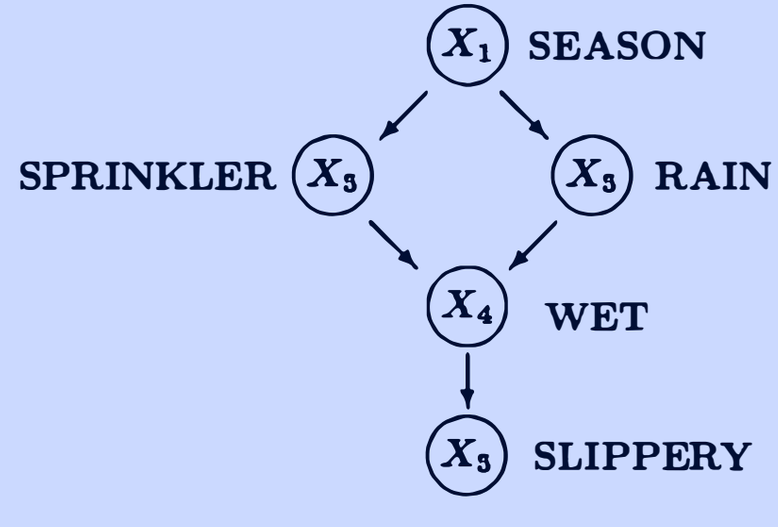
1. $P_x(V)$ is Markov relative to G;
2. $P_x(V_i=v_i)=1$ for all $V_i \in$X and v_i consistent with $X=x$;
3. $P_x(V_i \mid PA_i) = P(V_i \mid PA_i)$ for all $V_i \notin$X, i.e., $P(V_i \mid PA_i)$ remains invariant to interventions not involving V_i.



What is $P_{X3=ON}(X_1,X_2,X_4,X_5)$?

Structural Causal Models

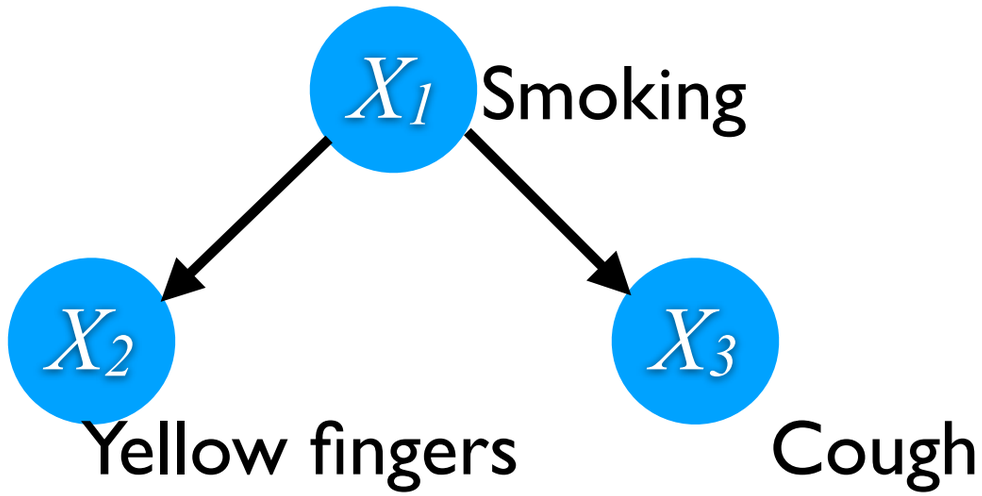
- $X_i = f_i(PA_i, E_i)$, $i=1,\ldots,n$
  - E_i: exogenous variables / errors / disturbances
  - Each equation represents an *autonomous* mechanism
    - Describes how nature assigns values to variables of interest
- Distinction between structural equations & algebraic equations
- Associated with graphical causal models



$$PA_i \longrightarrow X_i$$

$$\begin{aligned}
X_1 &= E_1, \\
X_2 &= f_2(X_1, E_2), \\
X_3 &= f_3(X_1, E_3), \\
X_4 &= f_2(X_3, X_2, E_4), \\
X_5 &= f_5(X_4, E_5)
\end{aligned}$$

Three Types of Problems in Current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
…	…	…

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

$$P(X3 \mid X2=1)$$

- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

$$P(X3 \mid \text{do}\ (X2=1))$$

- **Counterfactual**: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

$$P(X3_{\,X2=1} \mid X2 = 0, X3 = 1)$$

Summary: Graphical Models

- Directed acyclic graph
- d-separation
- Local and global Markov condition
- Causal graphical representation