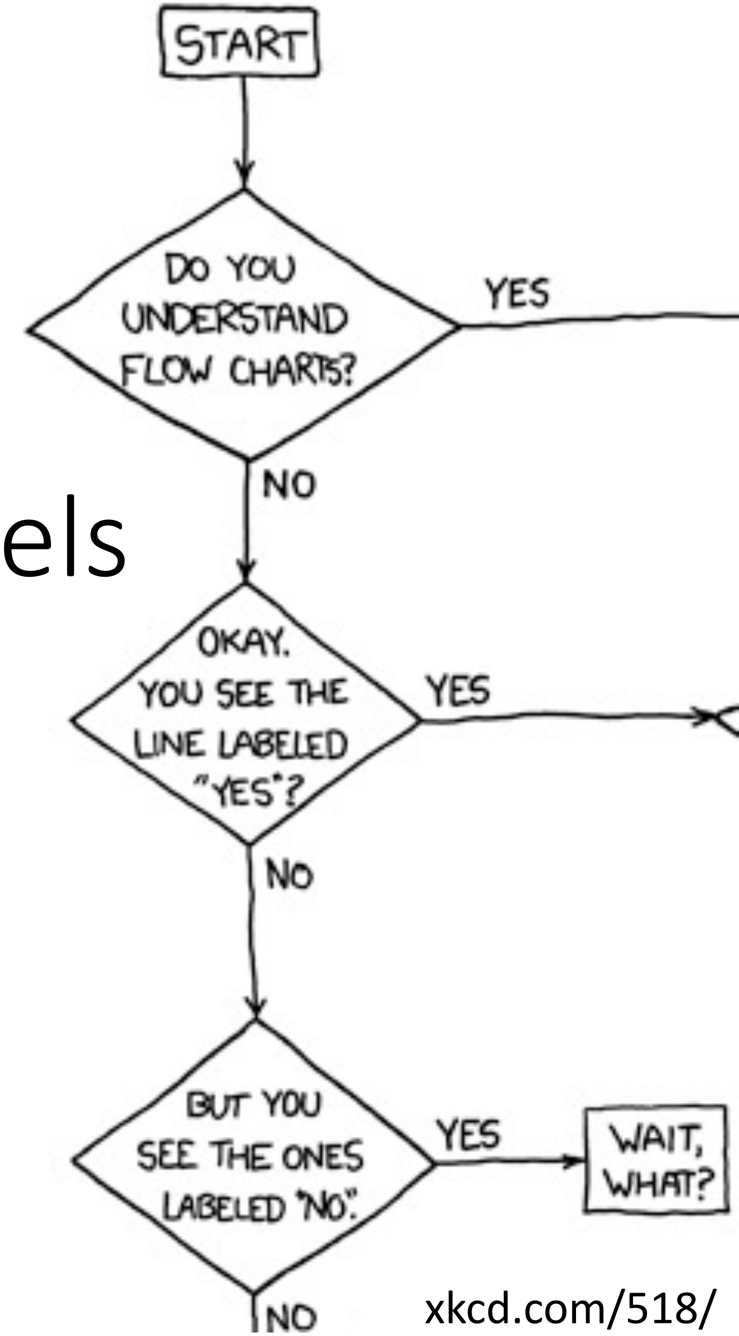
(Causal) Graphical Models



xkcd.com/518/

Zach Wood-Doughty

CS 396 Winter 2022

Some slides taken from Mark Dredze and Bryan Pardo
or inspired by Kevin Murphy

Last time: probabilities, independence, etc.

Are A and B Independent?

$P(A, B) = 0.25 + 0.05$

$P(A) = 0.3 + 0.2 + 0.05$

$P(B) = 0.3 + 0.1 + 0.05$

$P(A) \times P(B) = 0.55 \times 0.45$

$P(A, B) = 0.3 \neq 0.248$

A and B NOT independent

A	**B**	**C**	**Prob**
0	0	0	0.1
0	0	1	0.2
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.2
1	1	0	0.25
1	1	1	0.05

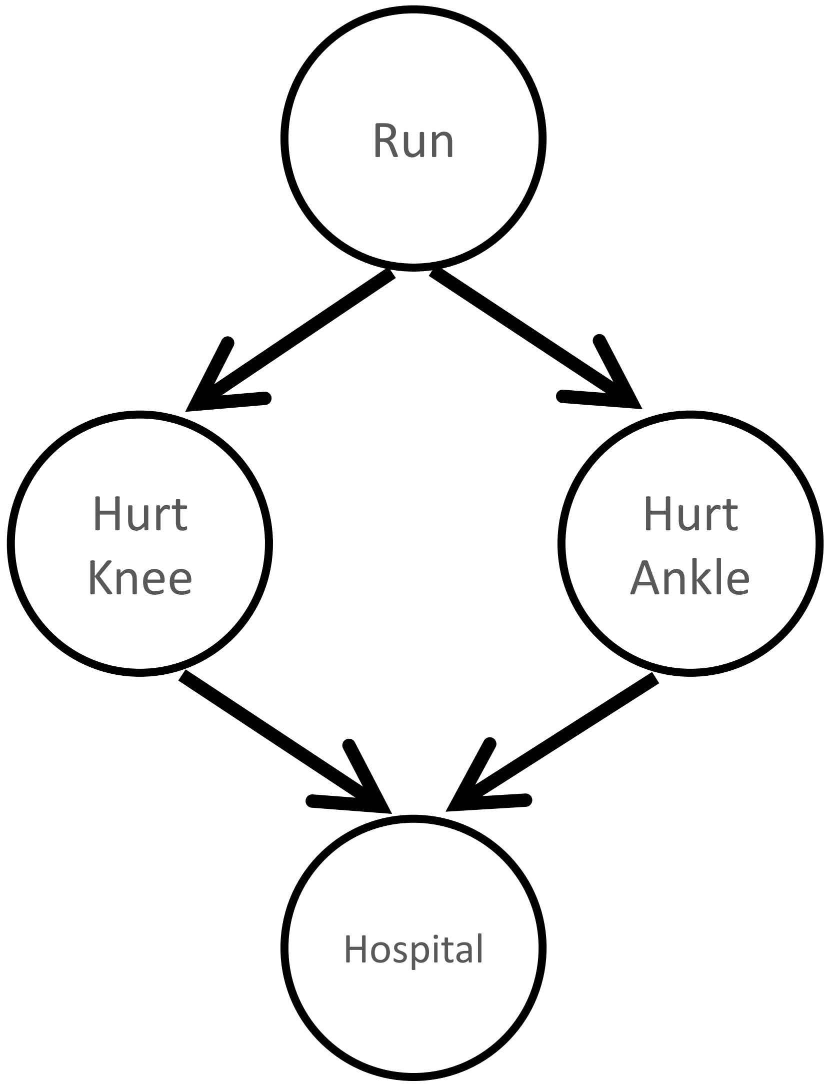
Example Probabilistic System

- A collection of related binary random variables
- Each day with some probability, a runner Avery:
  - Goes for a run (R)
  - Sprains an ankle (A)
  - Injuries their knee (K)
  - Goes to the hospital (H)
- Given a sprained ankle, what's the probability Avery goes to the hospital?
- What is the probability that Avery injuries their knee and goes to the hospital?

Example

- How do we answer these questions?
  - What is the structure of these variables?
  - What probabilities do I need to compute?
  - Are any of the variables independent of each other?
- How can we represent the variables in a way that answers these questions?

Graphical Models

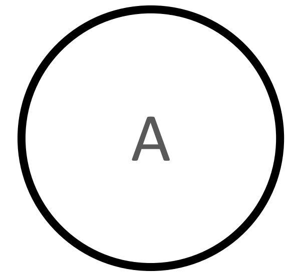


Graphical Models

- Combination of probability theory and graph theory
  - Combines uncertainty (probability) and complexity (graphs)
  - Represent a complex system as a (modular) graph
  - Standard algorithms for solving graph problems
- Many statistical models can be framed as graphical models
  - Logistic regression, linear regression, GMMs, etc.
  - Helps us reason about the underlying data generating process

Representation

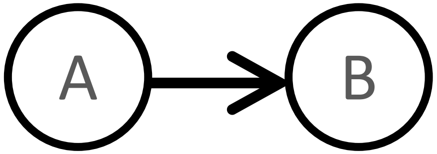
- A probabilistic system is encoded as a graph
- Nodes
  - Random variables
    - Could be discrete (this lecture) or continuous



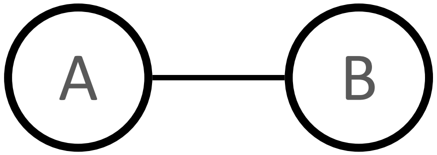
- Edges
  - Connections between two nodes
  - Indicates a direct relationship between two random variables
  - Note: the lack of an edge is very important
    - Implies no direct relationship

Graph Types

- Edge type determines graph type
- Directed (acyclic) graphs (DAG)
  - Edges have directions (A -> B)
  - Cycles are not allowed
  - Typically called Bayesian Networks
  - Directionality indicates (possible) causality

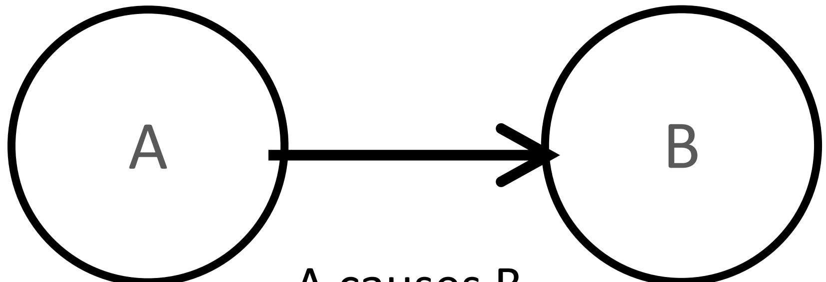


- Undirected graphs
  - Edges don’t have directions (A – B)
  - Typically called Markov Random Fields (MRFs)
    - Popular in physics and vision



Directed Graphs

- The direction of the edge indicates causation

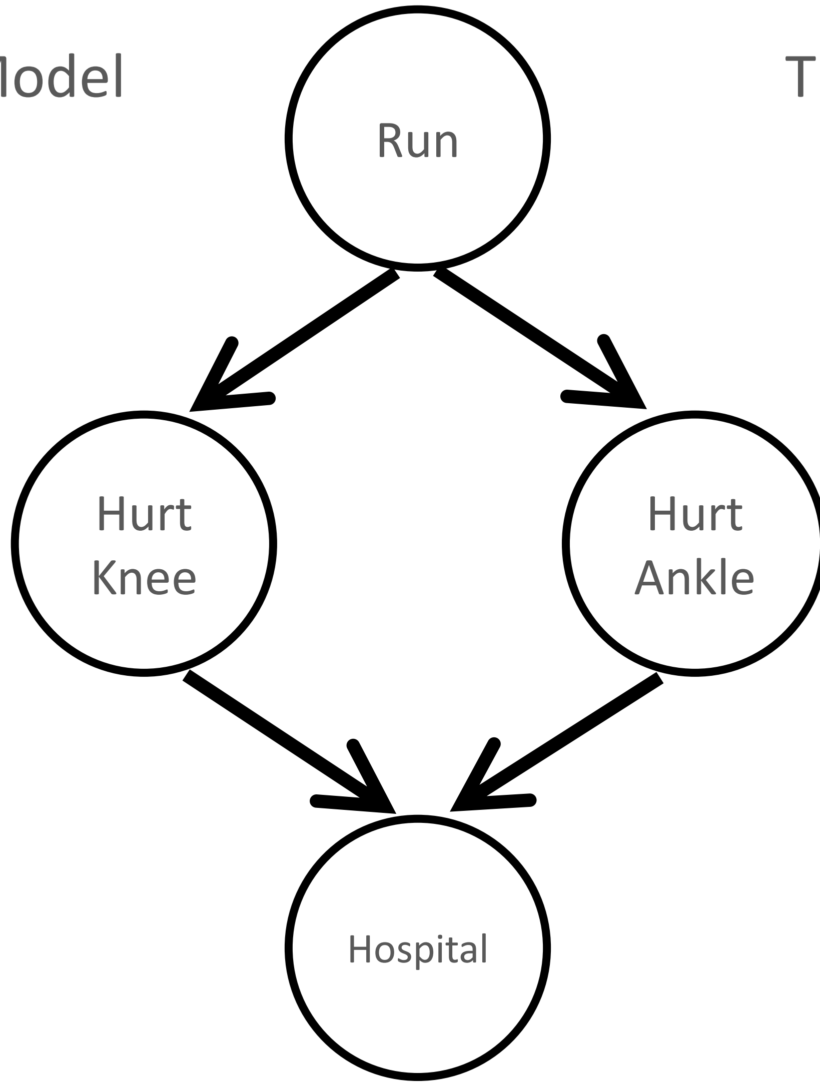


A causes B

- Causation can be very intuitive
  - We may know which random variable causes the other
  - Use this intuition to create a graph structure

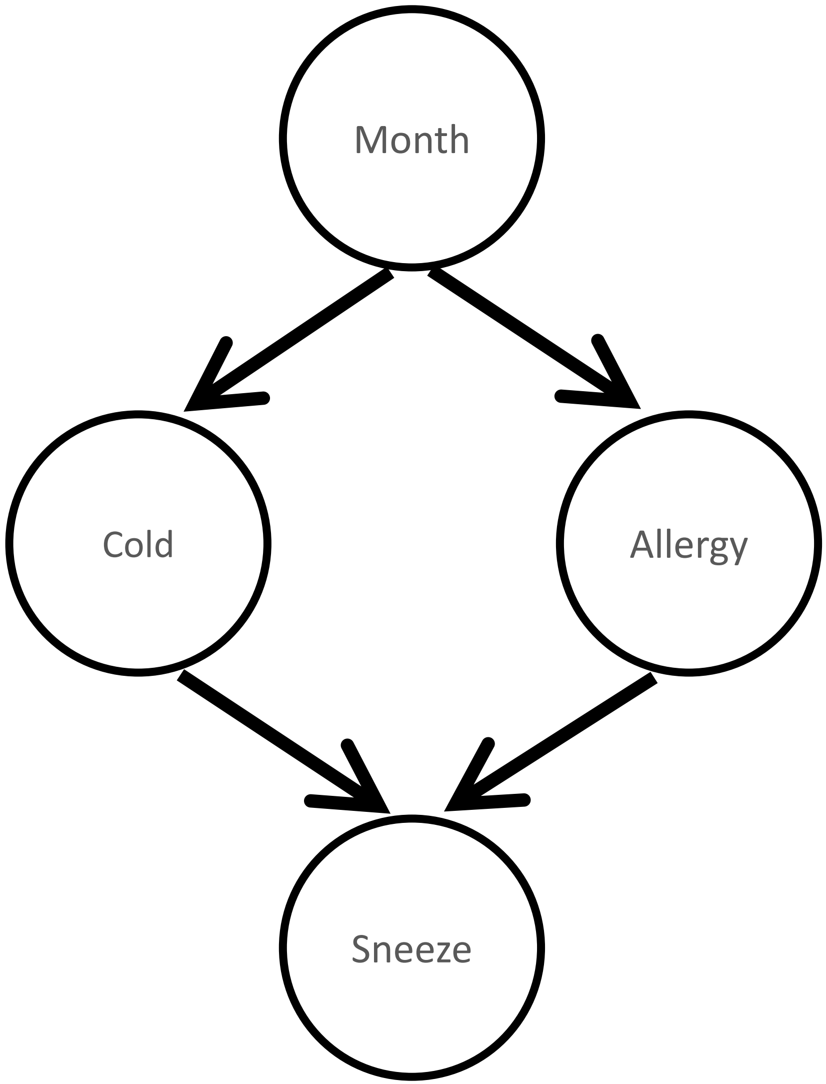
Example

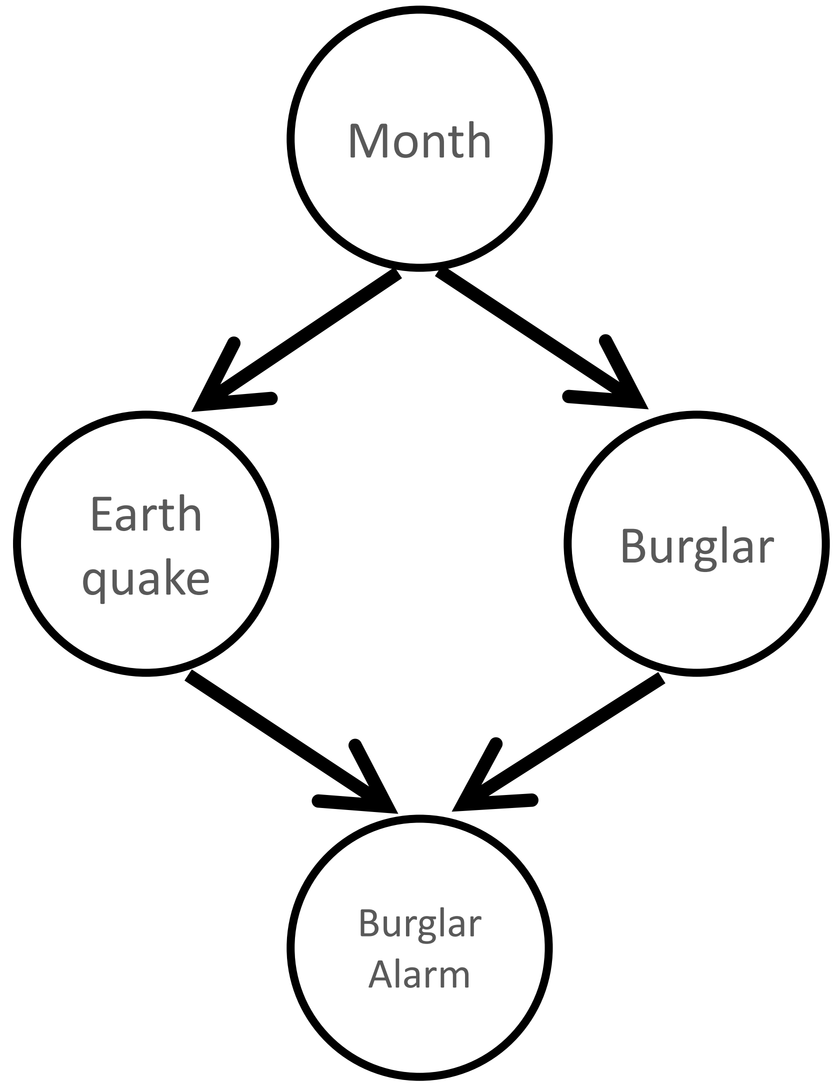
Generative Model



The Generative Story

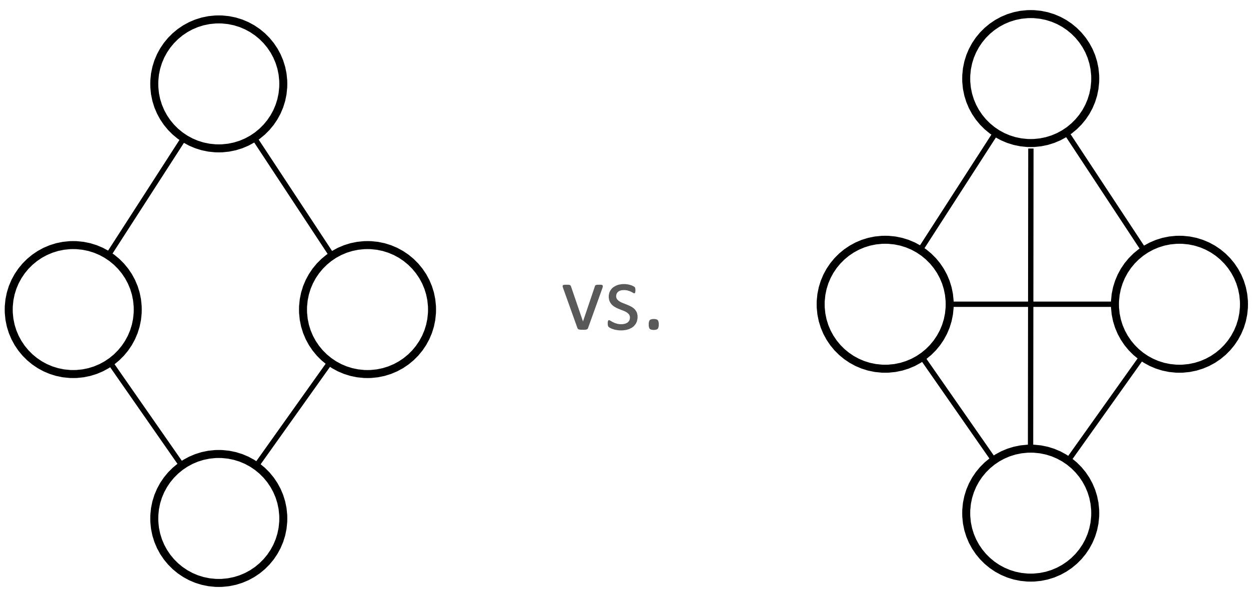
More examples





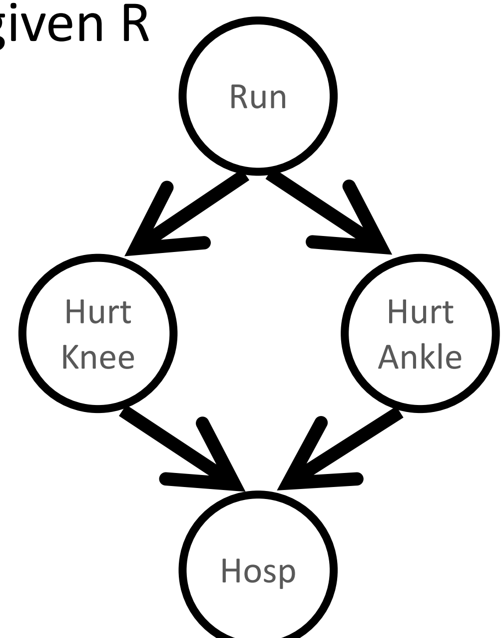
Advantages?

- What have we gained with this representation?
  - We could just draw a graph where everything is connected



Factorization

- Consider the joint probability of our example
  - What is the size of the conditional probability table for the p(R, A, K, H) distribution?
  - What can we do to simplify?
  - Notice that A and K are independent given R



Product Rule

- Can use the product rule to decompose joint probabilities
  - p(a,b,c) = p(c|a,b) p(a,b)
  - p(a,b,c) = p(c|a,b) p(b|a) p(a)
- This is true for any distribution
- Same for K variables

$$p(x_1 \ldots x_K) = p(x_K \mid x_1 \ldots x_{K-1}) \ldots p(x_2 \mid x_1) p(x_1)$$

Recall: independence

- My height is independent of my favorite basketball team
- This is **domain** knowledge, typically supplied by the problem designer
- Independence implies:

$$A \perp B \Rightarrow p(A \mid B) = p(A)$$

$$A \perp B \mid C \Rightarrow p(A, B \mid C) = p(A \mid C)p(B \mid C)$$

How does independence help?

$$A \perp B \Rightarrow p(A \mid B) = p(A)$$

A	B	P(A, B)
F	F	0.56
T	F	0.24
F	T	0.14
T	T	0.06

$$\begin{aligned} p(A) &= \sum_B p(A, B) \\ &= p(A, B) + p(A, \neg B) \\ &= 0.24 + 0.06 = 0.3 \end{aligned}$$

$$\begin{aligned} p(A|B) &= \frac{p(A, B)}{p(B)} \\ &= \frac{p(A, B)}{\sum_A p(A, B)} \\ &= \frac{p(A, B)}{p(A, B) + p(\neg A, B)} \\ &= \frac{0.06}{0.06 + 0.14} \\ &= 0.06/0.2 = 0.3 \end{aligned}$$

$$A \perp B \mid C \Rightarrow p(A, B \mid C) = p(A \mid C)p(B \mid C)$$

Conditional Independence

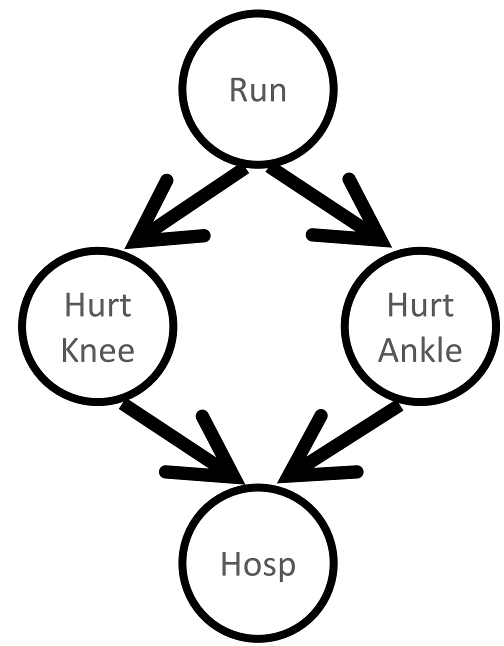
- Random variable **X** is conditionally independent of **Y** given **Z** if their conditional probabilities (given Z) are independent
- $p(x,y|z) = p(x|z)p(y|z)$
  - or $p(x|z, y) = p(x \mid z)$
- **X**: I need an umbrella and **Y**: the ground is wet
- Not independent! If ground is wet, it's probably raining (**Z**)
- I am told it is raining, now what?
- Knowing it's raining, needing an umbrella becomes independent of the ground being wet
- I gain no new information knowing that the ground is wet

  $P(x \mid z, y) = p(x \mid z)$

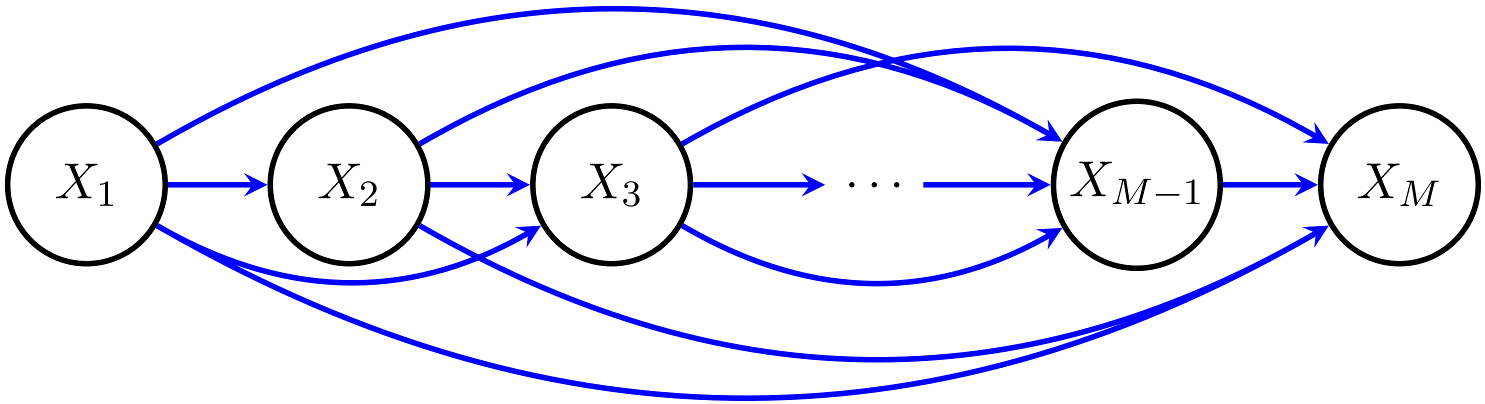
Factorization

- For any graphical model we can write the joint distribution using conditional probabilities
  - We just need conditional probabilities for a node given its parents

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k \mid \text{parents}_k)$$



Counting parameters in CPTs



X_1	X_2	…	X_M	P(X)
F	F	F	F	0.001
T	F	F	F	0.014
F	T	F	F	0.004
T	T	F	F	0.002
				…

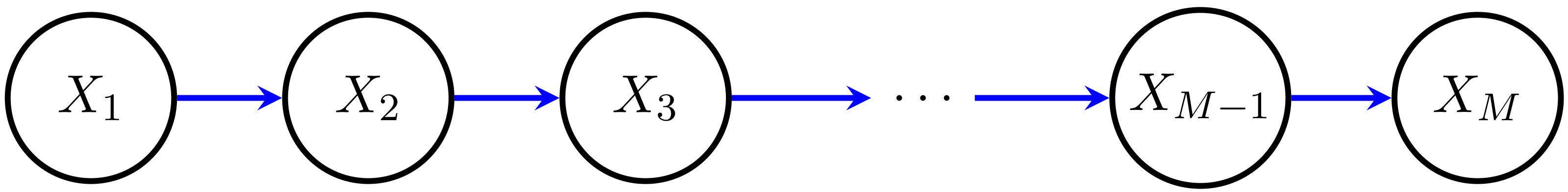
$P(X_1)$
.5

X_1	$P(X_2 \mid X_1)$
F	0.5
T	0.3

X_1	X_2	$P(X_3 \mid X_2, X_1)$
F	F	0.4
T	F	0.3
F	T	0.2
T	T	0.7

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k \mid \text{parents}_k)$$

Counting parameters in CPTs



X_1	X_2	…	X_M	P(X)
F	F	F	F	0.001
T	F	F	F	0.014
F	T	F	F	0.004
T	T	F	F	0.002
				…

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k \mid \text{parents}_k)$$

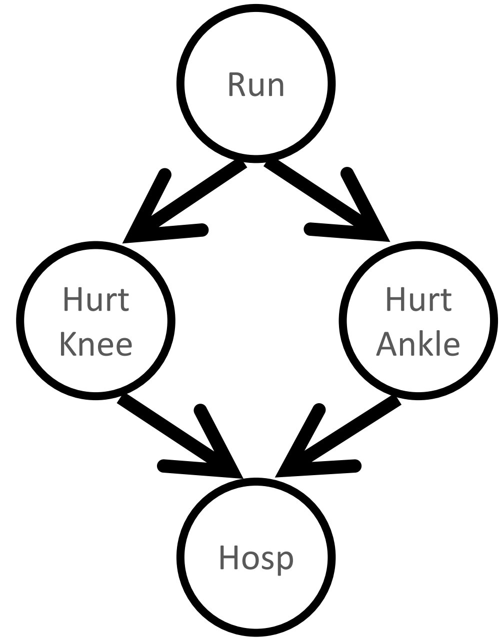
$P(X_1)$
.5

X_1	$P(X_2 \mid X_1)$
F	0.5
T	0.3

X_1	X_2	$P(X_3 \mid X_2, X_1)$
F	F	0.4
T	F	0.4
F	T	0.2
T	T	0.2

X_2	$P(X_3 \mid X_2)$
F	0.4
T	0.2

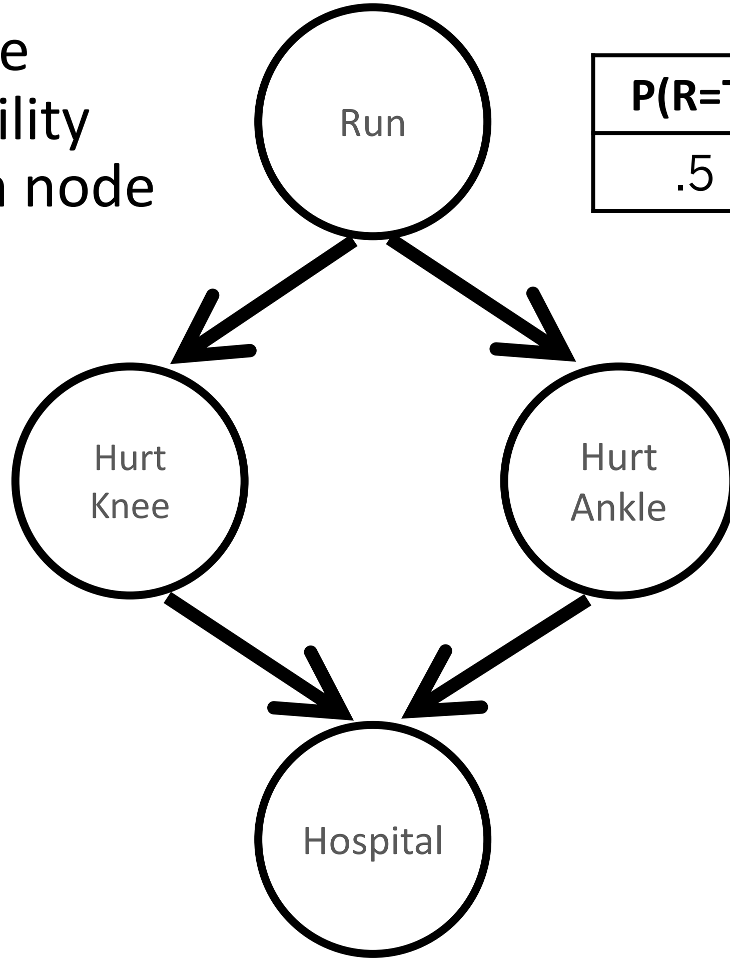
Factorization



- Consider the joint probability of our example
  - The full p(R, A, K, H) is complex
  - What can we do to simplify?
  - Notice that A and K are independent given R
- Factor the joint probability according to the graph
  - $p(R, A, K, H) = p(H \mid A,K)\, p(A \mid R)\, p(K \mid R)\, p(R)$
  - This is simpler to compute, with fewer conditional probabilities track.

Conditional Probability Tables

- The CPTs specify the conditional probability distribution at each node
- CPTs reflect local information only



P(R=T)
.5

R	P(K=T)
F	0.01
T	0.1

A	K	P(H=T)
F	F	0
T	F	0.3
F	T	.9
T	T	.99

R	P(A=T)
F	0.05
T	0.2

Independence and d-separation in DAGs

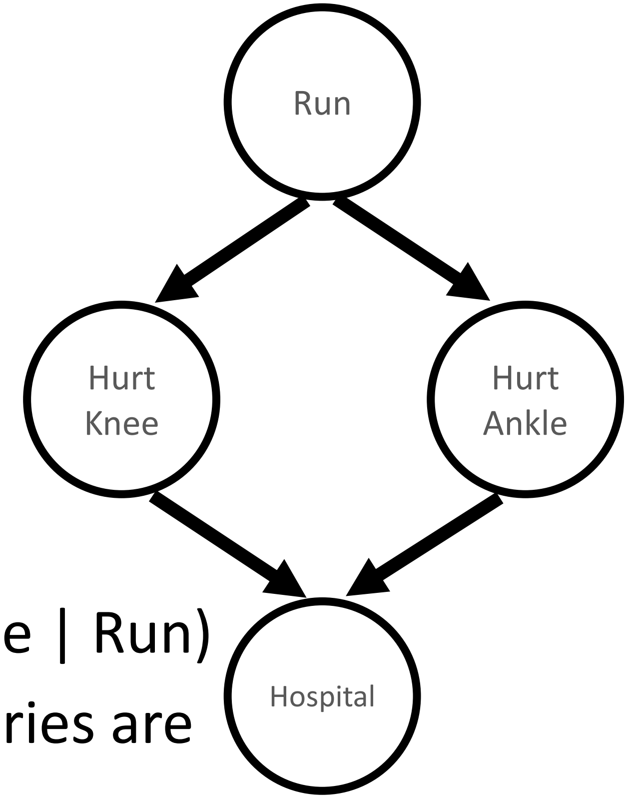
- DAGs are helpful because they make it much easier to visualize a complex probabilistic system

Independence

- The best part of graphical models is what they do not show
- Consider the network

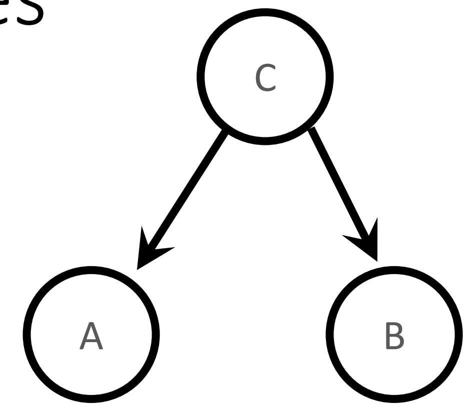
- A and B are independent
  - $P(A,B) = P(A)\,P(B)$
  - Variable independence allows us to build efficient models

Conditional Independence

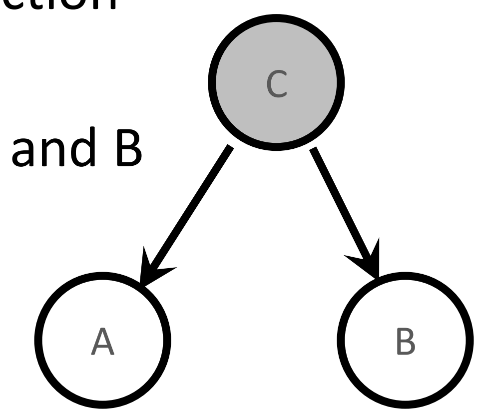


- Are Knee and Ankle injuries independent?
- Not marginally independent
- but conditionally independent given Run
- P(Knee, Ankle | Run) = p(Knee | Run) p(Ankle | Run)
- Once we know whether Avery ran, both injuries are independent of each other
- How do we know if something is independent?
  - We can read it from the paths of the graph!

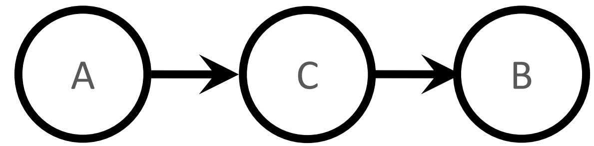
Example 1: Fork or "tail-to-tail" nodes



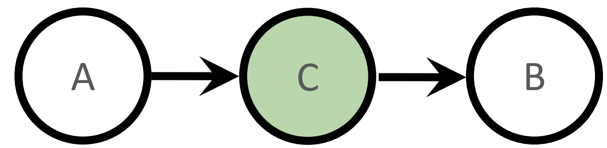
- Are A and B independent?
  - No, because both depend on C
- Are A and B conditionally independent?
  - Yes, because common cause is their only connection
  - We'll call C's connection with A and B a "fork"
  - Conditioning on C "blocks the path" between A and B



Example 2: Chain or “head-to-tail” nodes



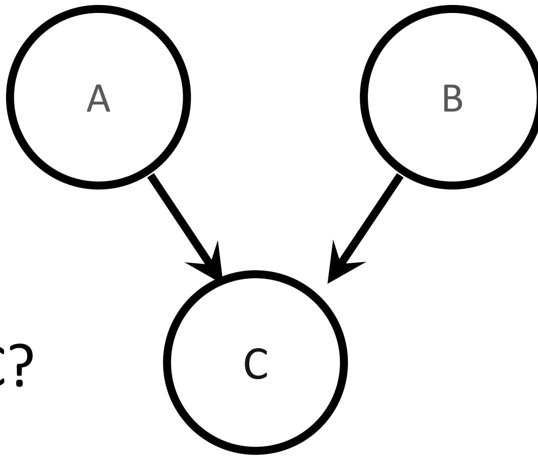
- Are A and B independent?
  - No. A causes C which causes B
- Are A and B conditionally independent?
  - Yes, because if we know C, knowing A tells us nothing about B



  - The connection of A and B to C is a “chain” or “head-to-tail”
  - Conditioning on C “blocks the path” between A and B

Example 3: Collider or “head-to-head” nodes

- Are A and B independent?
  - Yes, because A and B are generated without common parents
- Are A and B conditionally independent given C?
  - Counter-intuitively, no!
  - When C is unobserved, the path is **blocked**
  - When C is observed, the path becomes **unblocked**

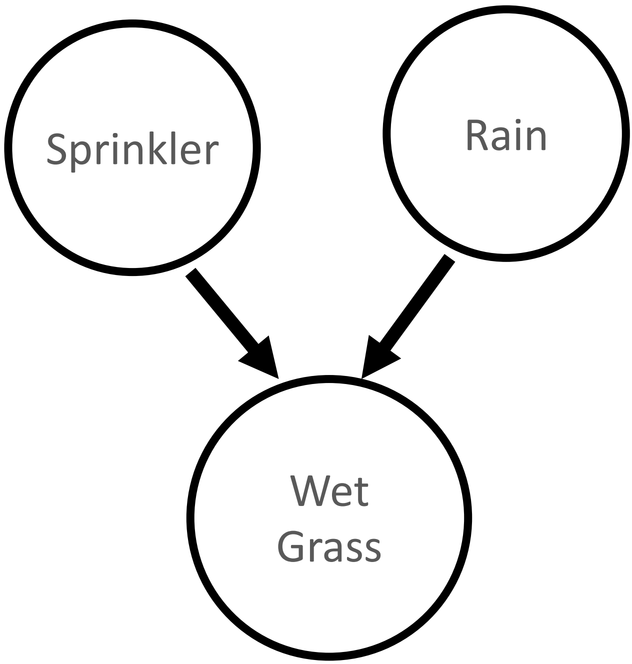


Blocked vs. Unblocked?

- Terminology: y is a descendent of x if there is a path from x to y (following the arrows)
- A fork or chain node only blocks a path when it is **observed**
- A collider node blocks a path when it is **unobserved**
    - A collider will become unblocked if either it, **or any of its descendants**, is observed

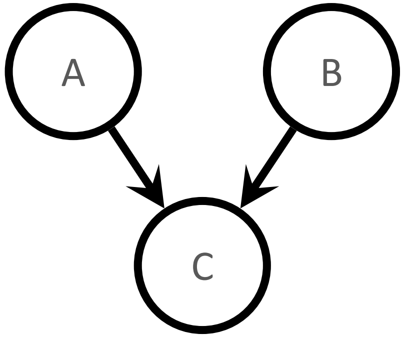
Collider dependence

- Suppose you see the grass outside is wet
- The two causes (sprinkler/rain) compete to explain the grass



Explaining Away

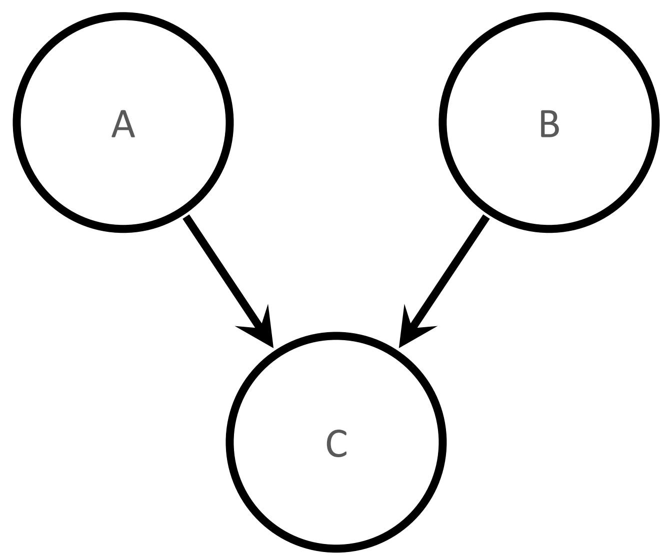
- Competing explanations for the same event
  - The rain explained the grass, so sprinkler is now less likely
  - The rain explained away the state of the grass
  - Don’t "need" to use sprinkler to explain it
- Thus, the observed head-to-head is unblocked
  - Once we know the value of C,
    we learn something about A and B



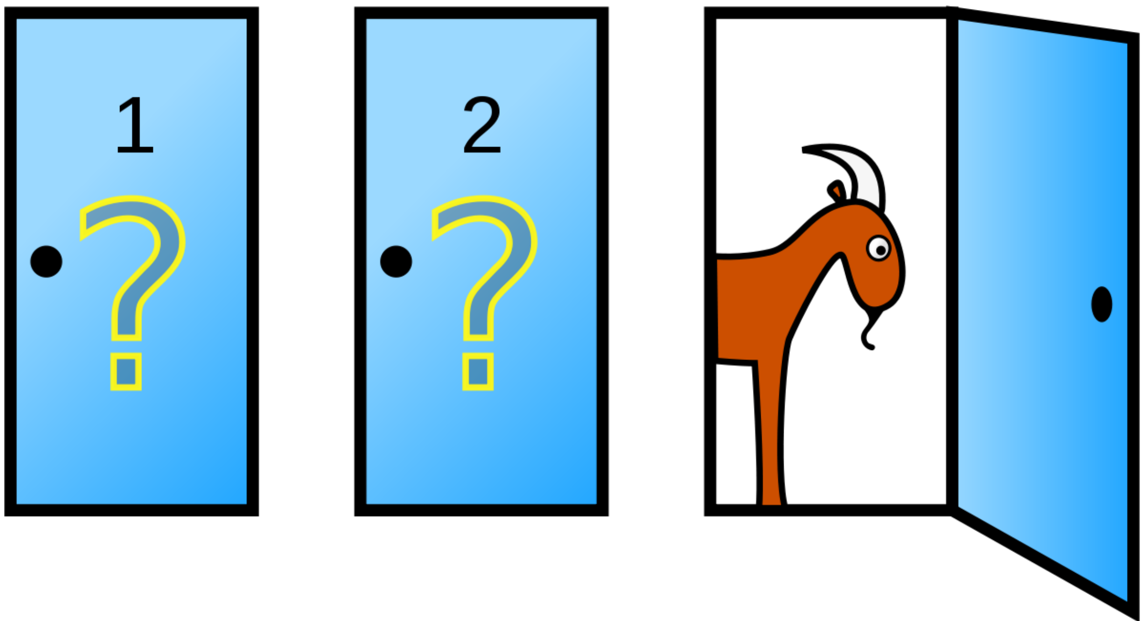
Explaining away

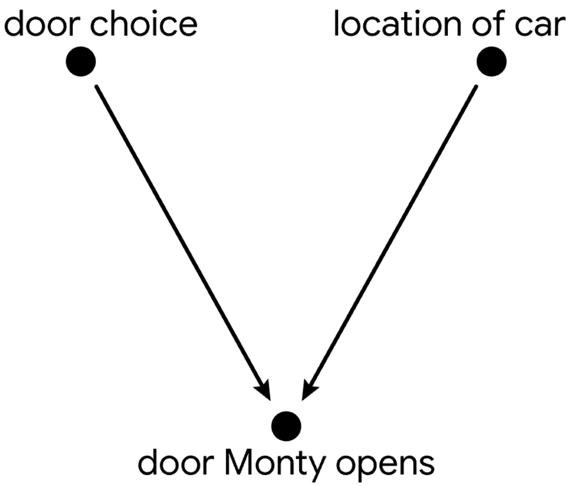
- Suppose A and B are coin flips (heads = 1)
- C is 1 if either A or B is 1
- A and B are independent: $p(A=0)p(B=0) = p(A=0, B=0)$

- But, if we condition on C, this changes!
- $p(A=0 \mid C=1) > 0$ and $p(B=0 \mid C=1) > 0$
- But $p(A=0, B=0 \mid C=1) = 0$



Monty Hall



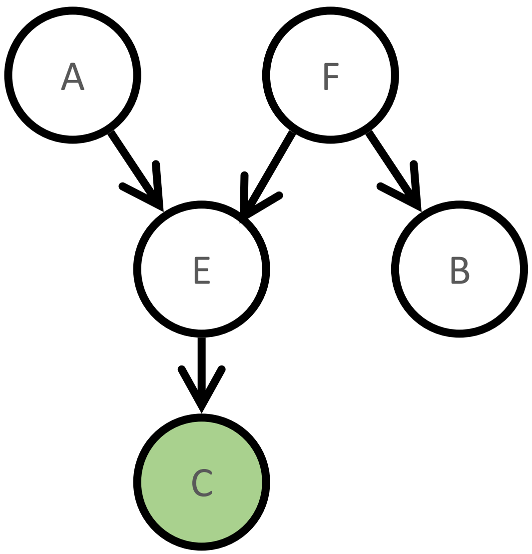


https://guzintamath.com/textsavvy/2019/09/22/monty-hall-and-the-colliders/

D-Separation

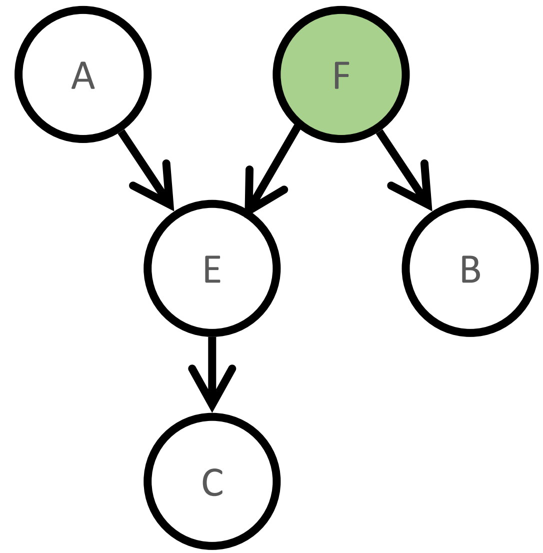
- Two nodes A and B are **d-separated** given observed node(s) C if all paths between A and B are blocked
  - Blocked paths: two arrows on the path meet head-to-tail or tail-to-tail at a node in set C
  - Or, the arrows meet head-to-head at a node which isn't in C
    - And none of its descendants are either
- If two (sets of) nodes are d-separated they are conditionally-independent!

Are A and B d-separated?



No

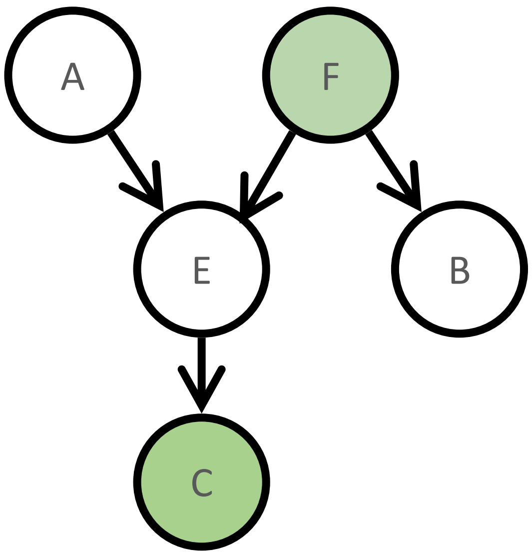
C is a descendent of head to head E



Yes

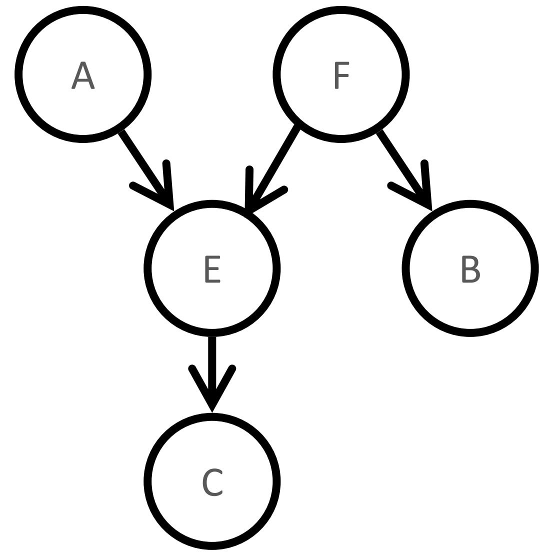
F is a tail to tail node

Are A and B d-separated?



Yes

F is a tail-to-tail node



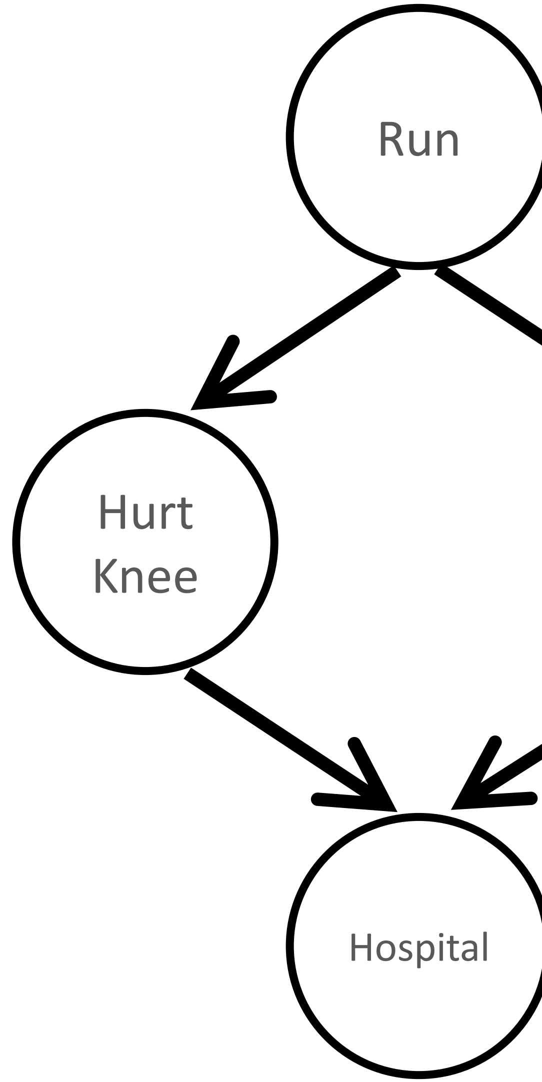
Yes

E is head-to-head

Causal DAGs

- DAGs are not necessarily causally-interpretable

- Need to assume
  (a) interventions are possible
  (b) no missing variables



P(R=T)
.5

A	K	P(H=T)
F	F	0
T	F	0.3
F	T	.9
T	T	.99