Week 5 — Bayesian Networks & Causal Bayes Nets

Friday, May 29, 2026

Prof. Joseph Austerweil

ChibaTech
School of
Design & Science

EN Press L

Agenda

Your assignment IS a Bayes net	0:00
Two complications + notation	0:10
Chibany Monty Hall + formal definition	0:20
d-separation + Bayes-Ball	0:35
Explaining away	1:00
Break	1:12
Observation → intervention	1:22
The do-operator	1:32
Blicket detector	1:47

Terminology: four overlapping names

Before we start — these terms get used loosely, sometimes interchangeably. Here's the relationship for today:

- **Graphical model** — *umbrella term.* Any probabilistic model drawn as a graph. Includes directed (Bayes nets) and undirected (Markov random fields) variants.
- **Bayesian network** *(= directed graphical model = belief network)* — a **DAG** where each node is a random variable and arrows encode conditional dependence. Today's main object.
- **Hierarchical Bayesian model** — Week 4's "priors over priors." Once drawn as a graph, it's a Bayes net with at least one **extra level of parameter nodes feeding into other parameter nodes.** So: **every hierarchical Bayesian model is a Bayes net.** The reverse isn't true — a Bayes net without that extra layer (e.g. Monty Hall, which we'll work through later today) is "flat," not hierarchical.
- **Causal Bayesian network** — same DAG, but arrows now mean **direct causation**. Adds the do-operator (Pearl). Every causal BN is a Bayes net; not every Bayes net is causal.

GMM = hierarchical Bayes ⊂ Bayes net ⊂ graphical model. Causal BN = Bayes net + causal reading.

Your assignment IS a Bayes net

Welcome

Week 5 of Human and Machine Learning.

Last week we wrestled with **hierarchical Bayes** — priors over priors. Right now most of you are mid-Clusters, working out **Gaussian mixtures**.

Today's claim: those are the same thing, drawn as a graph.

Chibany’s histogram, drawn differently

Remember Chibany’s mystery bentos? Two peaks at 350g and 500g.

In Clusters you’re writing a **two-Gaussian mixture** as a generative process and computing what each datapoint tells you about which cluster it came from.

Today’s claim: that generative process **is a Bayes net.** Same model — just drawn. You’ve been doing Bayesian-network inference all week without naming it.

We’ll leave the density and the “which cluster?” formula to your assignment — no spoilers today.

The generative process you wrote

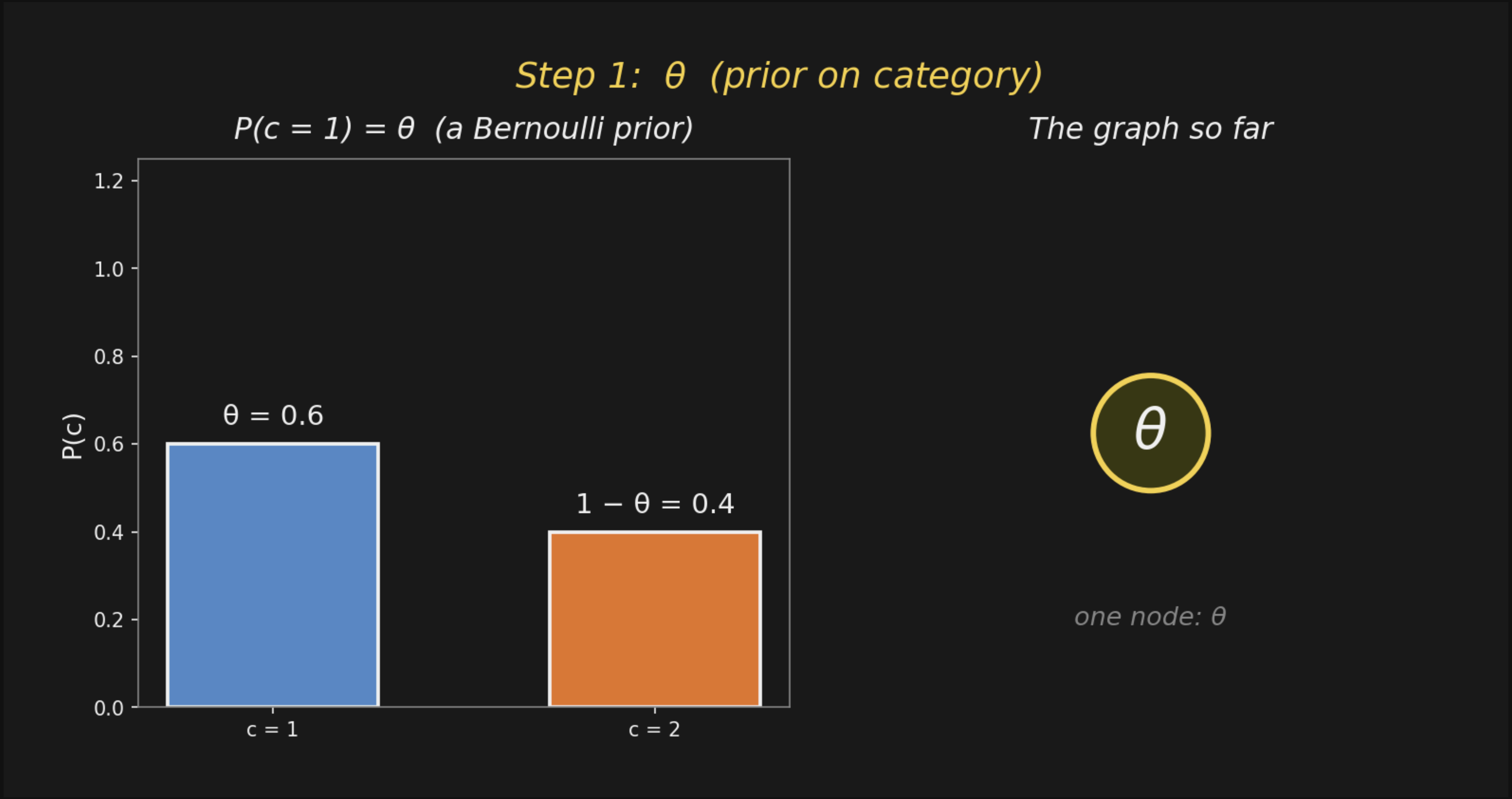
In Clusters Problem 2, you wrote the two-Gaussian mixture as:

$$c_n \mid \theta \;\sim\; \mathrm{Bernoulli}(\theta)$$

$$x_n \mid \mu_{c(n)},\, \sigma^2_{c(n)} \;\overset{iid}{\sim}\; \mathrm{N}(\mu_{c(n)},\, \sigma^2_{c(n)})$$

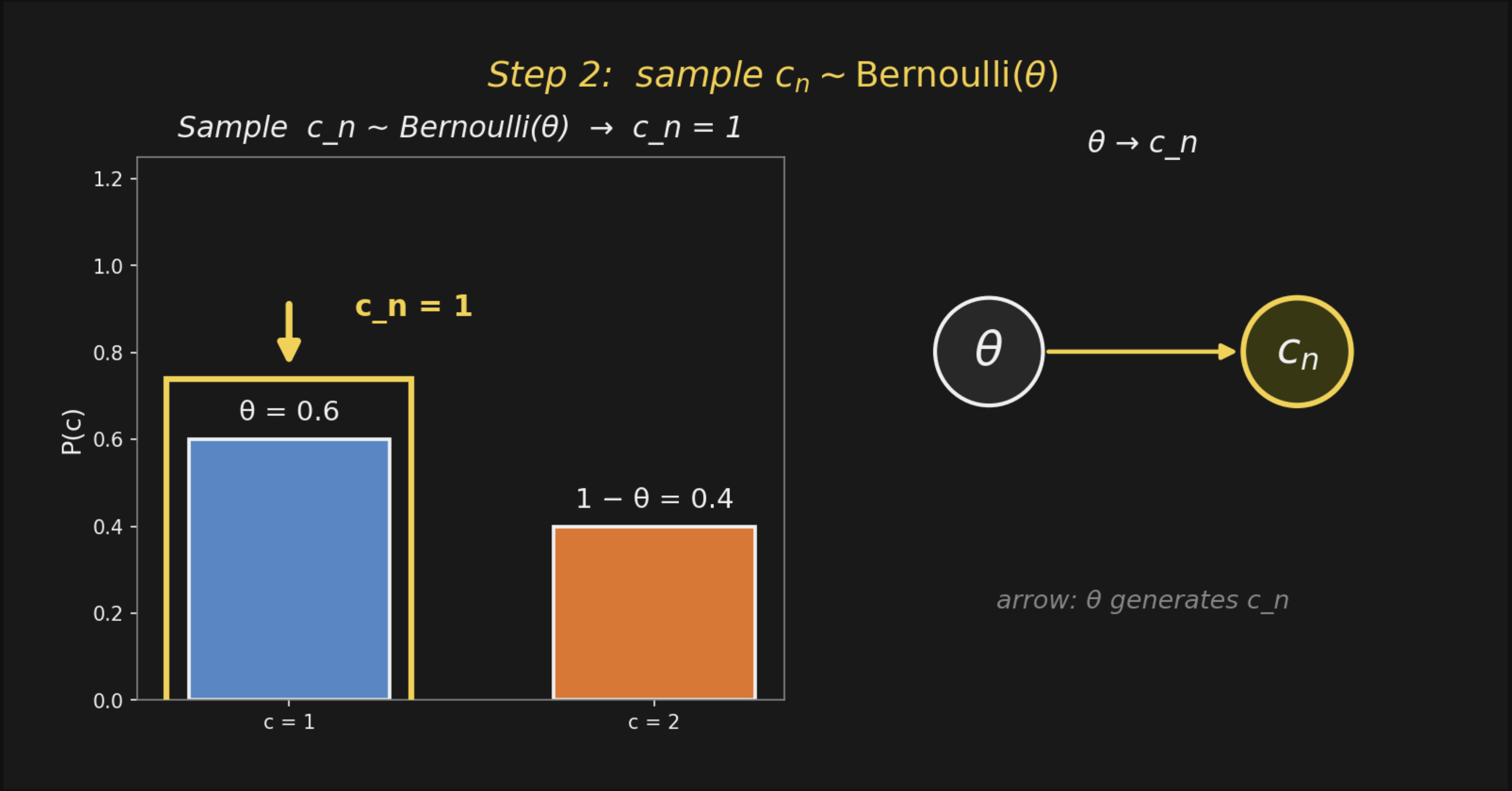
Two lines. Each line is **one sampling step.** Today we'll show that this generative process **is already a Bayes net** — we just haven't drawn it yet.

Step 1: introduce θ



The Bernoulli prior **on the left** has one parameter, θ. **On the right**, θ becomes one **node** of the graph. A node = a random variable (or parameter) in the model.

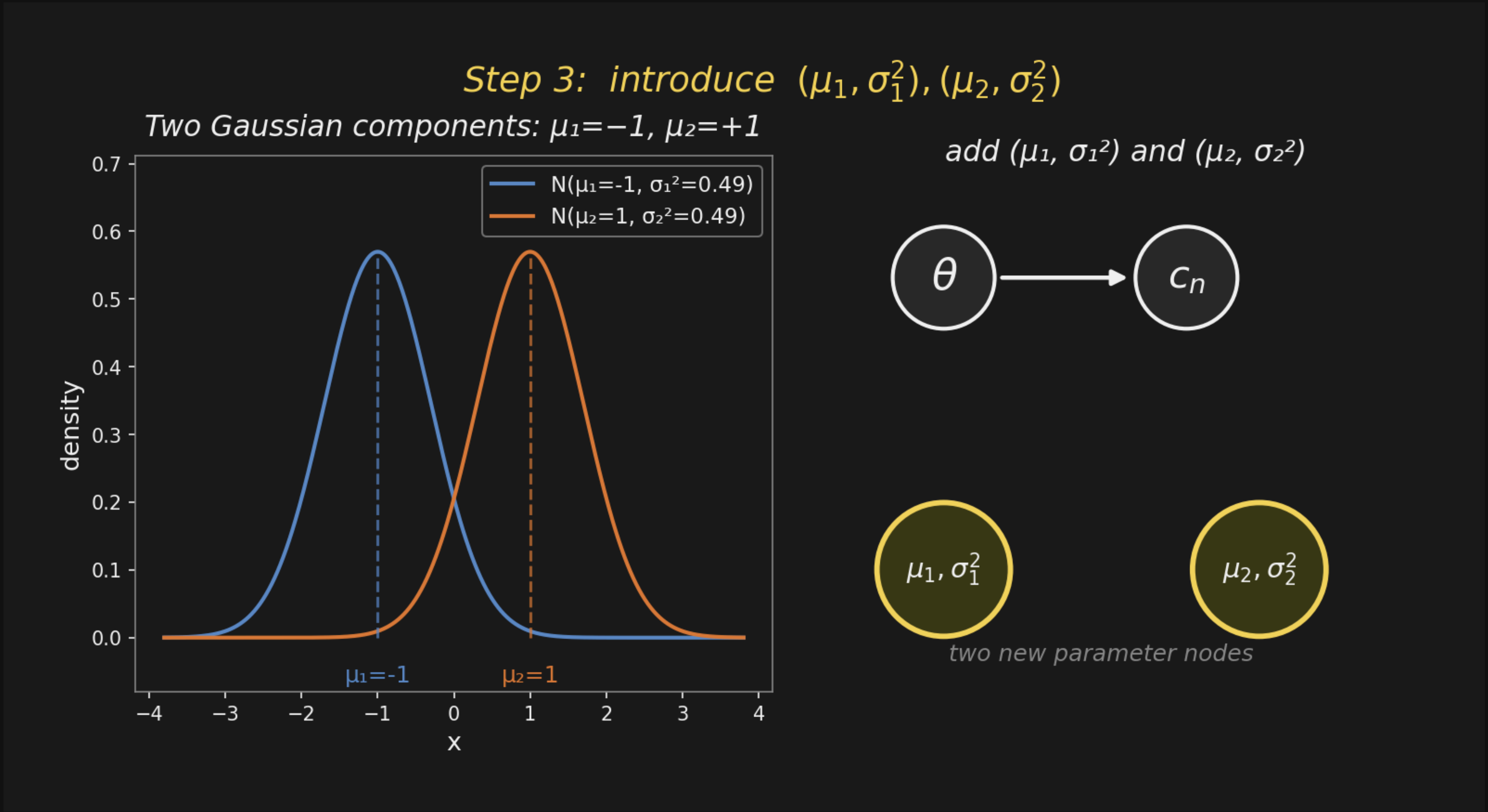
Step 2: sample $c_n \sim \mathrm{Bernoulli}(\theta)$



Sampling c_n flips a θ-weighted coin. Pick category 1 (highlighted bar). **On the graph:** c_n becomes a new node, and the line $c_n \mid \theta \sim \mathrm{Bernoulli}(\theta)$ becomes an **arrow** from θ to c_n.

Generative-process line ↔ arrow in the graph.

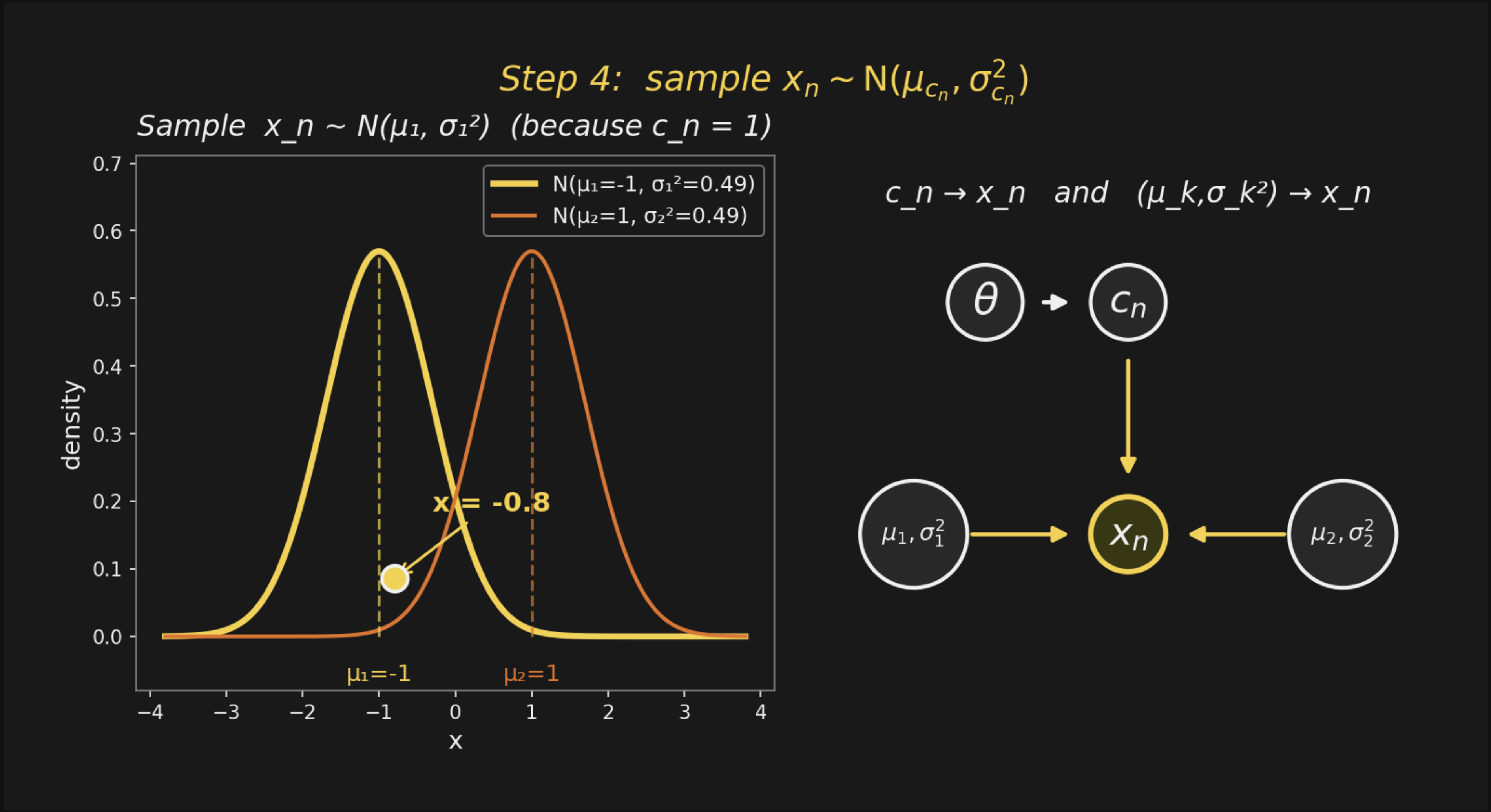
Step 3: introduce the components



The two Gaussians have parameters (μ_1, σ_1^2) and (μ_2, σ_2^2). **On the graph:** two new nodes for the parameters — one pair per component.

We're not yet drawing arrows from these into x_n. That happens in the next step.

Step 4: sample x_n, the observation



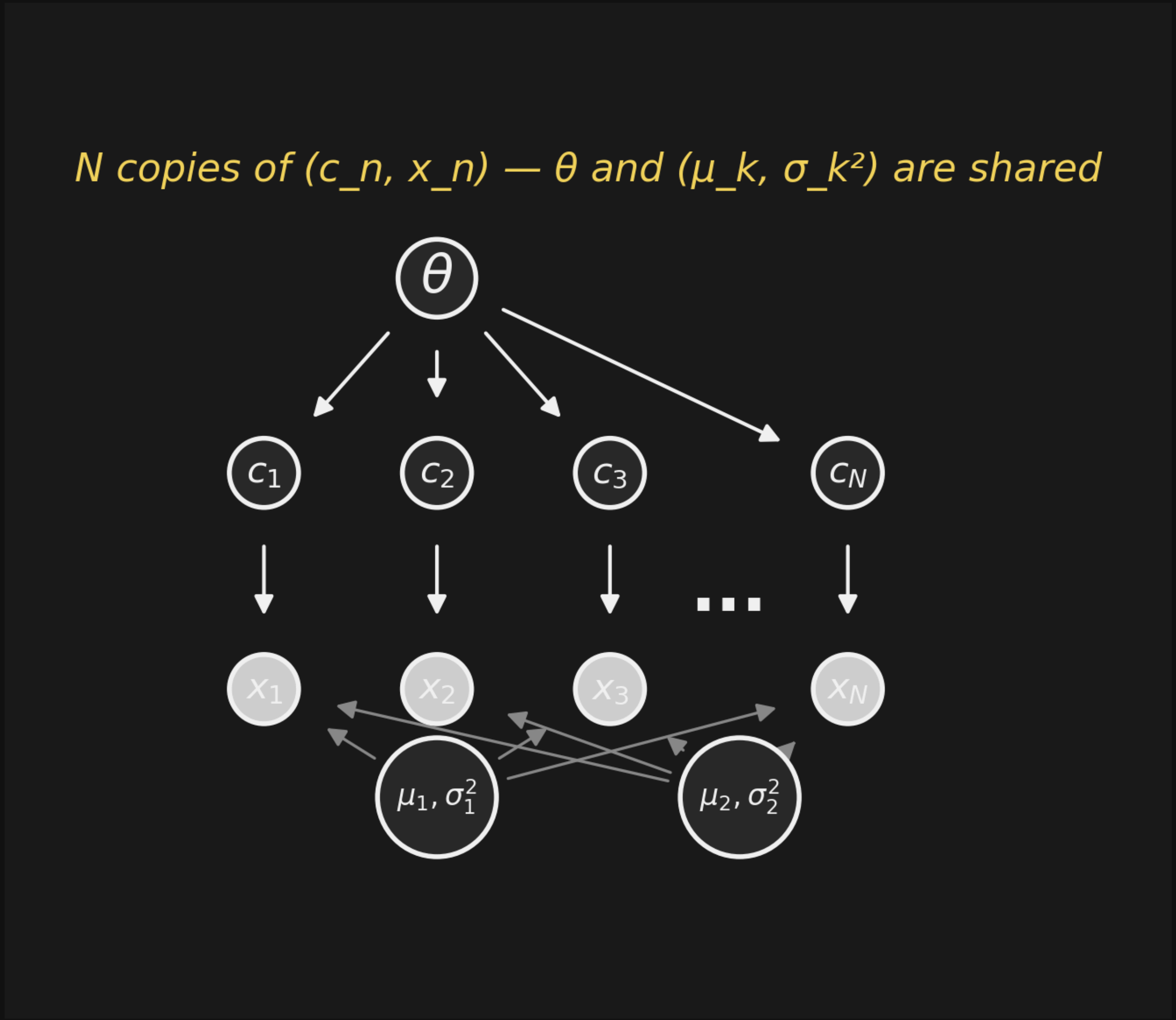
With $c_n = 1$, we sample $x_n \sim \mathrm{N}(\mu_1, \sigma_1^2)$ — a point under the highlighted Gaussian. **On the graph:** x_n is a new node (shaded — it's *observed*), with arrows from c_n and from each (μ_k, σ_k^2).

The second generative-process line gave us **three new arrows** at once.

ChibaTech School of Design & Science

EN Press L

Step 5: do this N times



The generative process has $n = 1, \ldots, N$. Each n gets its own (c_n, x_n) — but θ and the (μ_k, σ_k^2) are **shared**.

Drawn explicitly, this is an N-fold copy of the same little pattern.

ChibaTech School of Design & Science
EN Press L

Step 6: the plate — a shortcut

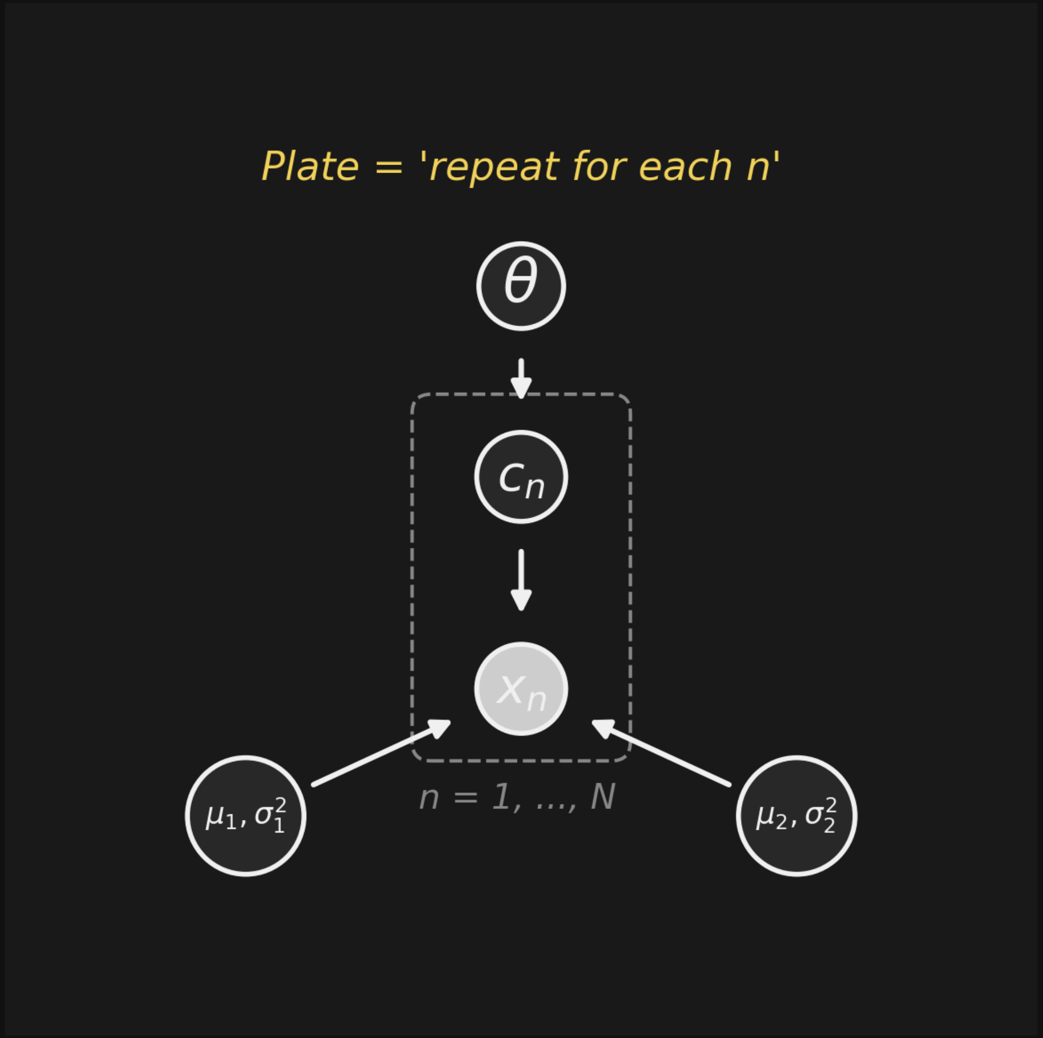
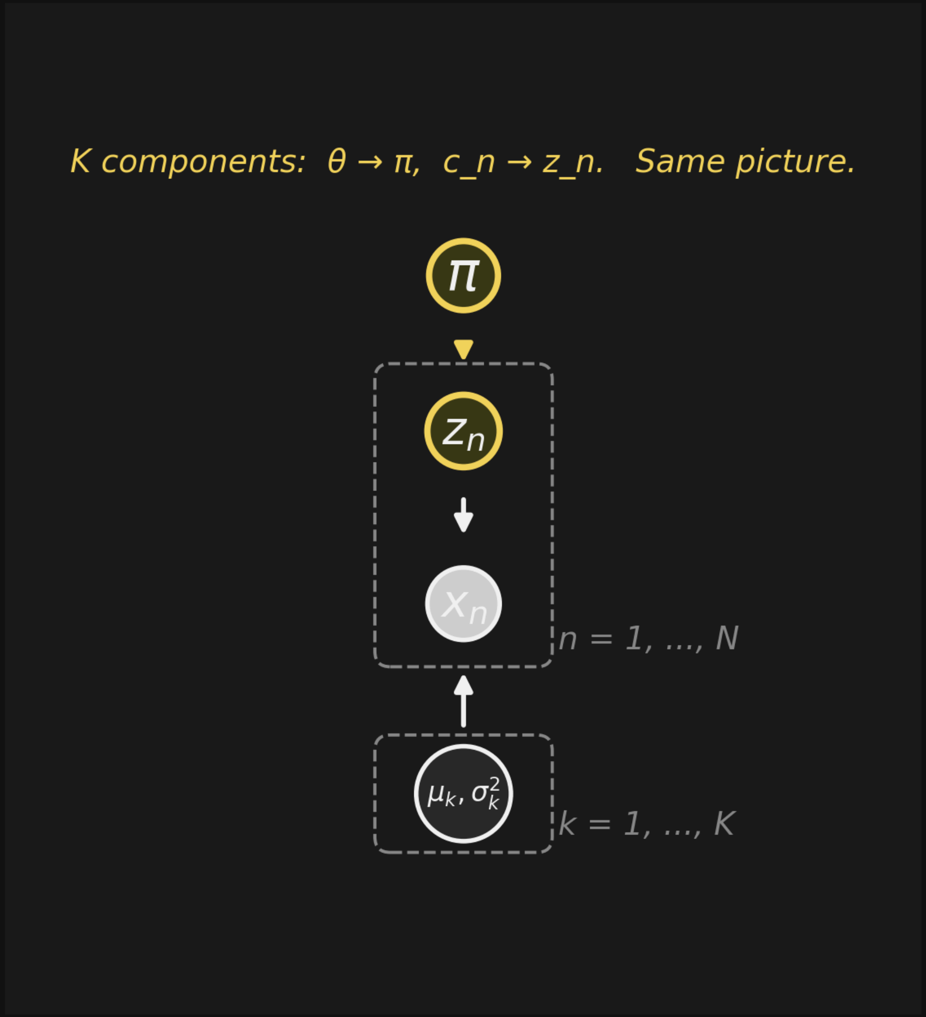


Plate notation: draw the repeating pattern *once* inside a dashed box, and label the box "$n = 1, \ldots, N$." The plate means **"copy everything inside, N times."**

Same graph, vastly more compact.

Step 7: generalize to K components



Bernoulli(θ) generalizes to Categorical(π) for K components. $c_n \to z_n$, $\theta \to \pi$, and (μ_k, σ_k^2) gets a K-plate.

Same picture. Two clusters or two hundred — the graph language stayed put.

The choice of K — DPMM (T3 Ch 6) avoids fixing it. Week 10 picks this up.

Reading inference off the graph

You've now seen the graph. Two things to take away — **without writing down the formula** (that's your assignment):

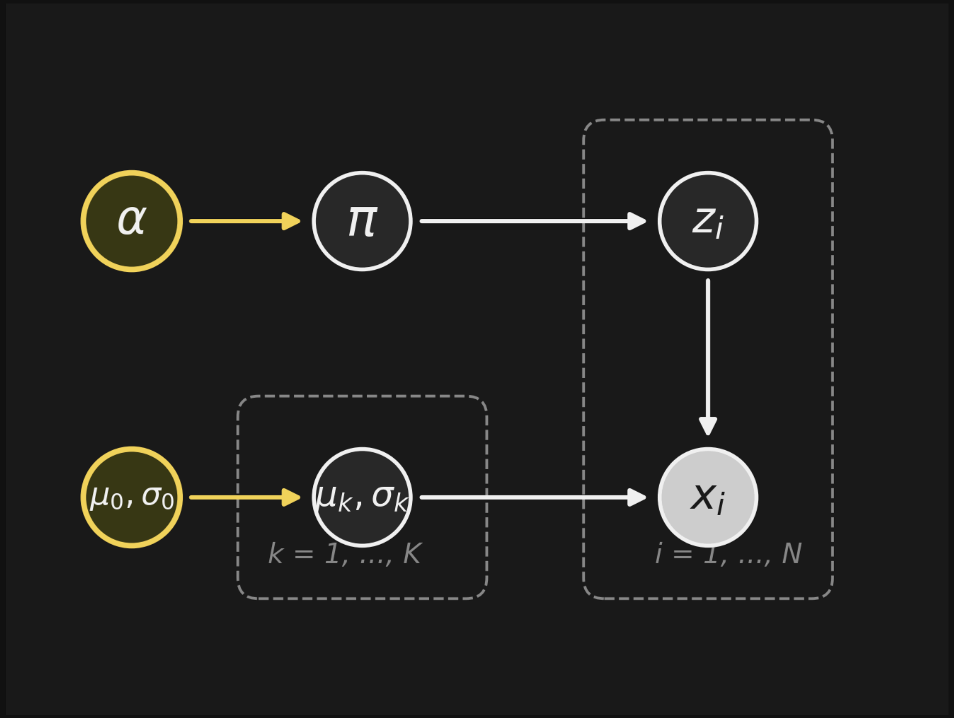
- **Down the arrow** = generative direction. Sample θ, then c_n, then x_n.
- **Up the arrow** = inference direction. Observe x_n, ask which c_n generated it.

The "which cluster?" calculation you're doing in Clusters Problem 2 is **inference going UP the arrow from x_n to c_n.** Prior $\rightarrow$ likelihood $\rightarrow$ posterior, read off the parent–child structure.

We won't write the formula here — finish the assignment first; we'll discuss after Friday.

Two complications + notation

Where do the priors come from?



Add a **hyperprior**: $\alpha \to \pi$ and $(\mu_0, \sigma_0) \to (\mu_k, \sigma_k)$.

This is **exactly Week 4's hierarchical Bayes**, drawn as a graph. The graph language already covered it; we just didn't have the words yet.

ChibaTech School of Design & Science
EN Press L

What if there are multiple parents?

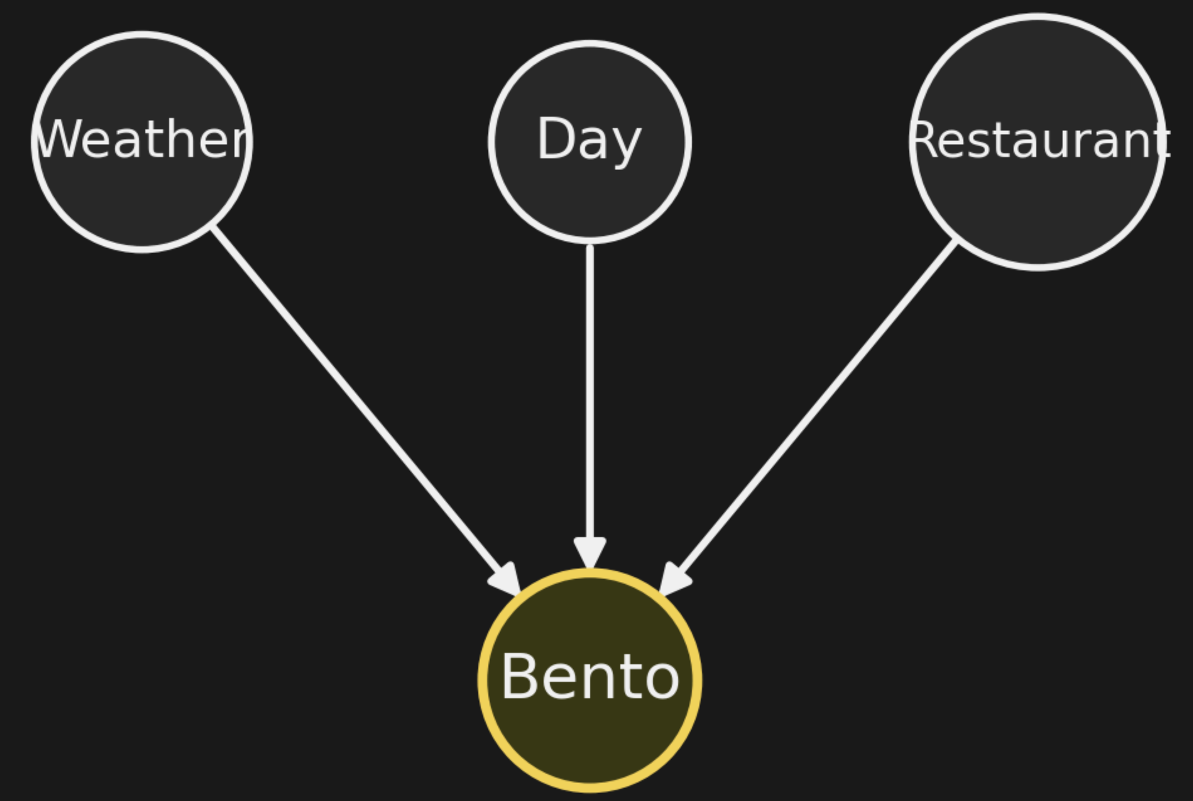
Chibany realizes: the bento (B) weight depends on more than just which cluster.

- **Weather** (W) — hot days mean lighter bentos
- **Day of the week** (D) — the cafeteria menu rotates
- **Restaurant** (R) — different students come from different places

Multiple causes, one observation. *How does the picture handle that?*

We'll use W, D, R, B as shorthand for these variables from here on.

Chibany’s full bento network



Three parents pointing into Bento. **What’s the joint distribution?**

The rule for multiple parents

Each parent contributes a factor. The joint factorizes as:

$$P(W, D, R, B) \;=\; P(W)\,P(D)\,P(R)\,P(B \mid W, D, R)$$

- W, D, R have no parents → marginals.
- B has three parents → conditional on **all three**.

One factor per node, conditioned on that node's parents. That's the rule.

Notation lock-in

Symbols we'll use:

- $G = (V, E)$ — a **directed acyclic graph** (DAG)
- $\mathrm{Pa}(X)$ — the **parents** of node X in G

Three examples so far:

- GMM: $\mathrm{Pa}(z_i) = \{\pi\}$, $\mathrm{Pa}(x_i) = \{z_i\}$
- GMM + hyperprior: add $\mathrm{Pa}(\pi) = \{\alpha\}$
- Bento: $\mathrm{Pa}(B) = \{W, D, R\}$

Joint factorizes as:

$$P(X_1, \ldots, X_n) = \prod_i P(X_i \mid \mathrm{Pa}(X_i))$$

Same picture, same rule, three problems.

Chibany Monty Hall + formal definition

A new puzzle for Chibany

The cafeteria scenario. Three bento boxes on the counter. Exactly one of them contains tonkatsu (the other two don’t). Chibany picks **box 1**. The cafeteria worker — who *knows* which box has the tonkatsu — opens **box 3** and reveals: not tonkatsu.

Should Chibany switch to box 2?

As a Bayes net

Three nodes:

- **Tonkatsu** (T): which box has it. Uniform: each box, 1/3.
- **Chibany Chooses** (C): which box Chibany chose. Independent of Tonkatsu.
- **Cafeteria Reveals** (R): which box the worker opens — depends on **both** parents.

A **collider**: two arrows pointing into the same node.

We'll use T, C, R as shorthand for these from here on. (C for Chibany's

Joint factorization

The graph gives us:

$$P(T, C, R) \;=\; P(T)\, P(C)\, P(R \mid T, C)$$

- $P(T) = 1/3$ for each box
- $P(C) = 1/3$ for each box (we'll assume Chibany chooses uniformly)
- $P(R \mid T, C)$ — the cafeteria worker's policy: open a box that's neither the tonkatsu nor Chibany's choice.

Before any observation, each box has probability 1/3 of being tonkatsu.

Now condition: $P(T \mid R = 3, C = 1)$

We observe $C = 1$ (Chibany chose box 1) and $R = 3$ (worker opened box 3). What's the posterior over T?

Two cases that matter:

- $T = 1$ (Chibany's choice is right). Worker opens box 2 OR box 3 — picks uniformly. So $P(R = 3 \mid T = 1, C = 1) = 1/2$.
- $T = 2$ (the other unopened box). Worker MUST open box 3 (only choice that's neither 1 nor 2). So $P(R = 3 \mid T = 2, C = 1) = 1$.
- $T = 3$ (impossible — worker just revealed it). $P(R = 3 \mid T = 3, C = 1) = 0$.

Bayes' rule gives the answer

$$P(T = k \mid R = 3, C = 1) \;\propto\; P(R = 3 \mid T = k, C = 1)\, P(T = k)$$

T	Prior	Likelihood	Unnormalized
1	1/3	1/2	1/6
2	1/3	1	1/3
3	1/3	0	0

Normalize: $P(T = 1 \mid \cdot) = 1/3$, $P(T = 2 \mid \cdot) = 2/3$.

Chibany should switch.

Now name it: the Markov factorization

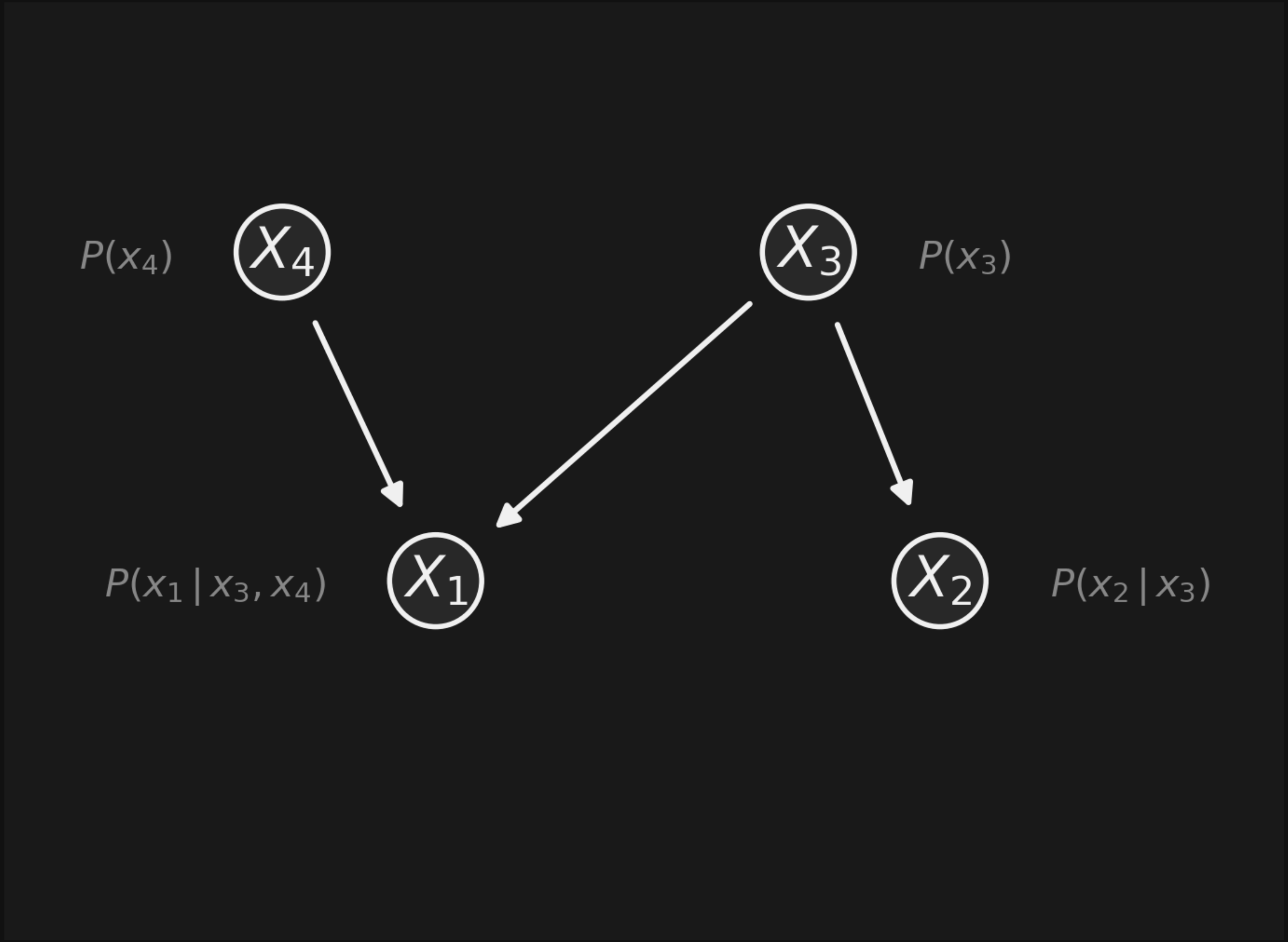
Definition. A **Bayesian network** for variables $X_1, \ldots, X_n$ is a **directed acyclic graph** G — one node per variable — together with a **conditional distribution** $P(X_i \mid \mathrm{Pa}_G(X_i))$ at each node.

Theorem (Markov factorization). The joint distribution factorizes as

$$P(X_1, \ldots, X_n) \;=\; \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}_G(X_i)).$$

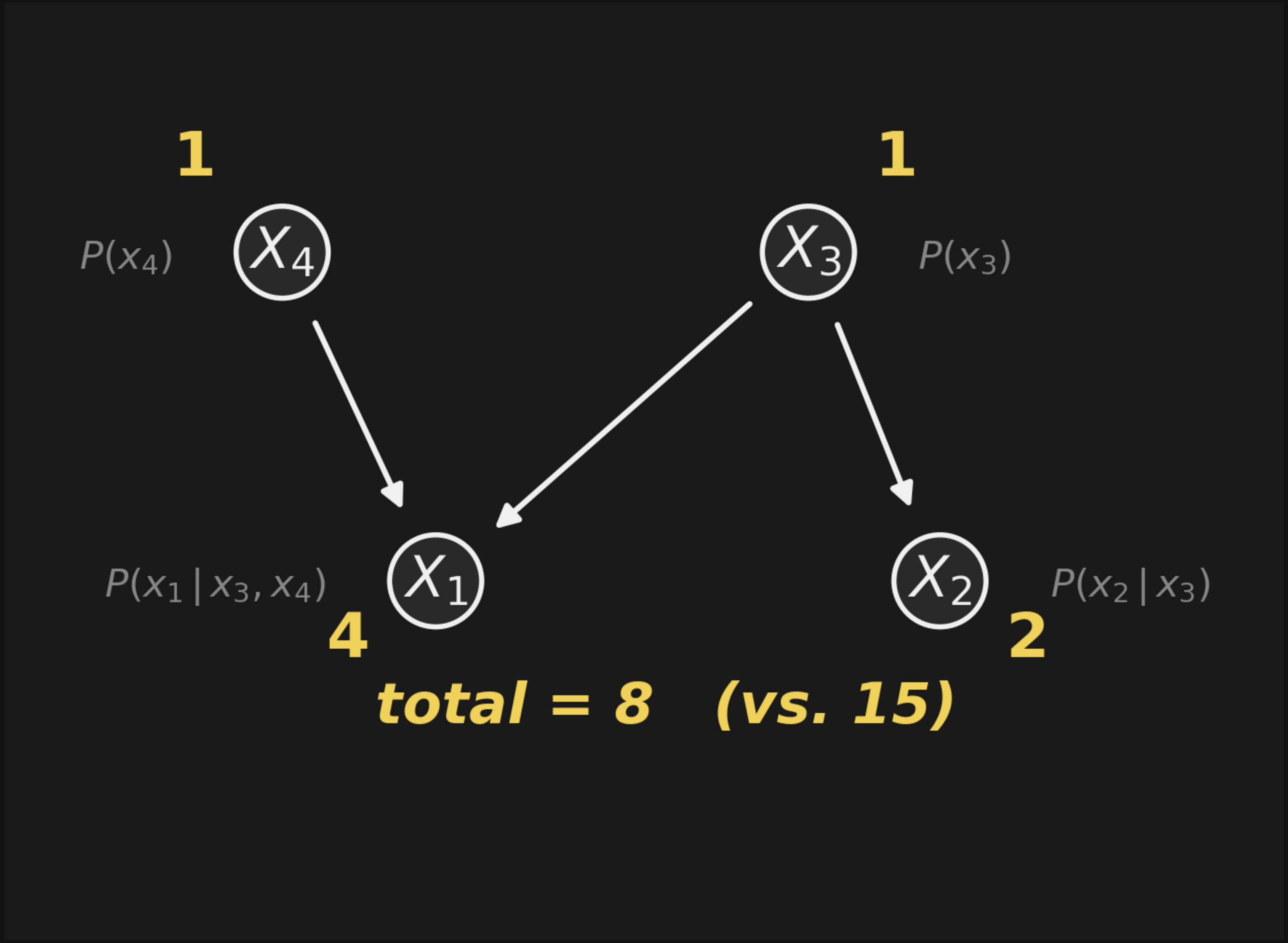
You’ve used this rule three times already — mixture, hyperprior, bento — and just used it for the Monty Hall posterior. Now it has a name.

How compact is this?



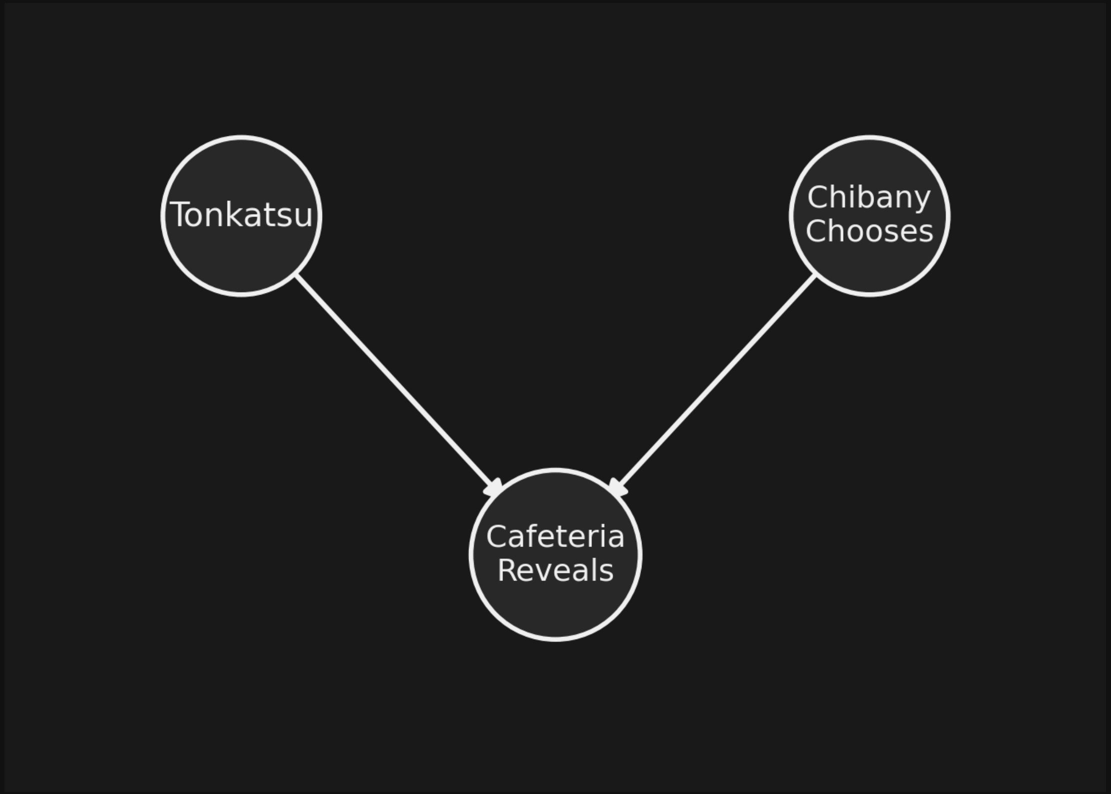
For 4 binary variables, the **full joint** needs $2^4 - 1 = 15$ numbers. This Bayes net needs how many?

How compact is this? — answer

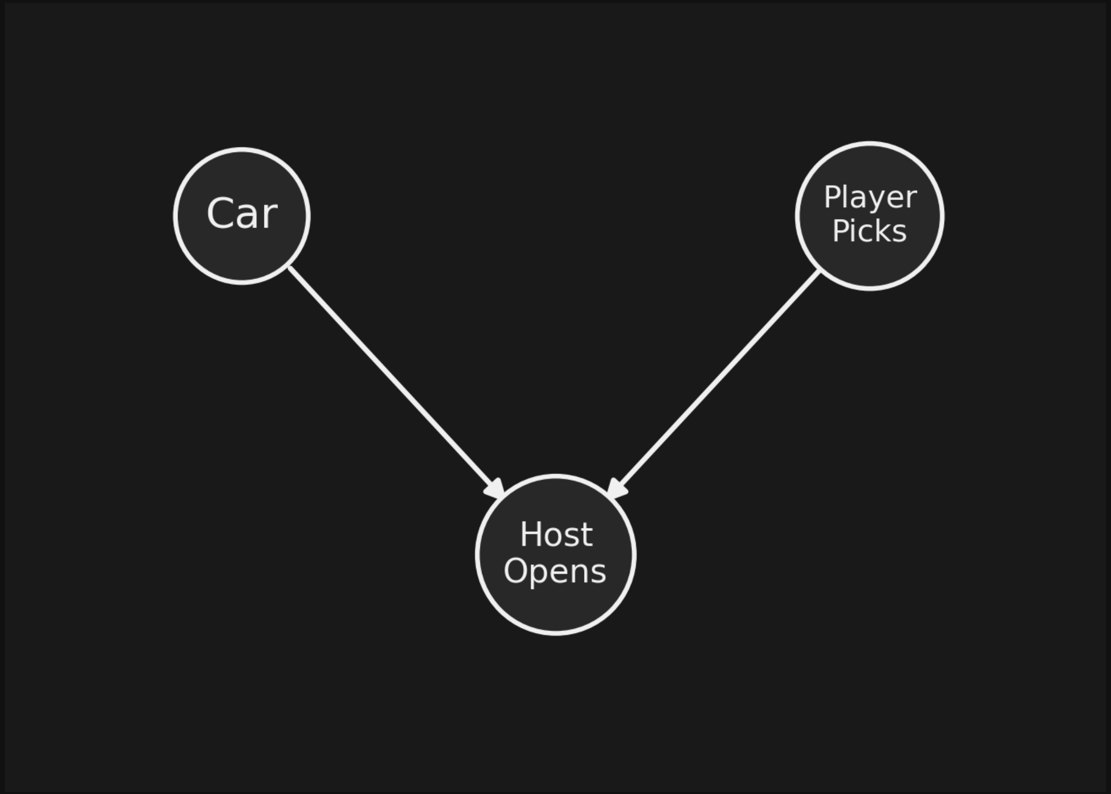


8 vs. 15. Modest gain here — exponential gain at scale.

Same network, different costume



Chibany version



Classic Monty Hall

Car / Host / Player and Tonkatsu / Cafeteria / Chibany — *literally* the same network. The 2/3 vs 1/3 answer is structural.

Poll — Bayes-net factorization

In the Chibany Monty Hall network (Tonkatsu → Reveals ← Chooses), the joint $P(T, R, C)$ factorizes as:

- **A.** $P(T)\,P(C)\,P(R \mid T, C)$
- **B.** $P(T \mid C)\,P(C)\,P(R)$
- **C.** $P(T)\,P(C \mid T)\,P(R \mid C)$
- **D.** Cannot factorize — all three are dependent.

Poll — answer

A. $P(T)\,P(C)\,P(R \mid T, C)$

Each node, given its parents, contributes one factor. T and C are root nodes (no parents) $\rightarrow$ marginals. R is a child of both $\rightarrow$ conditional on both.

d-separation + Bayes-Ball

Reading independence off the graph

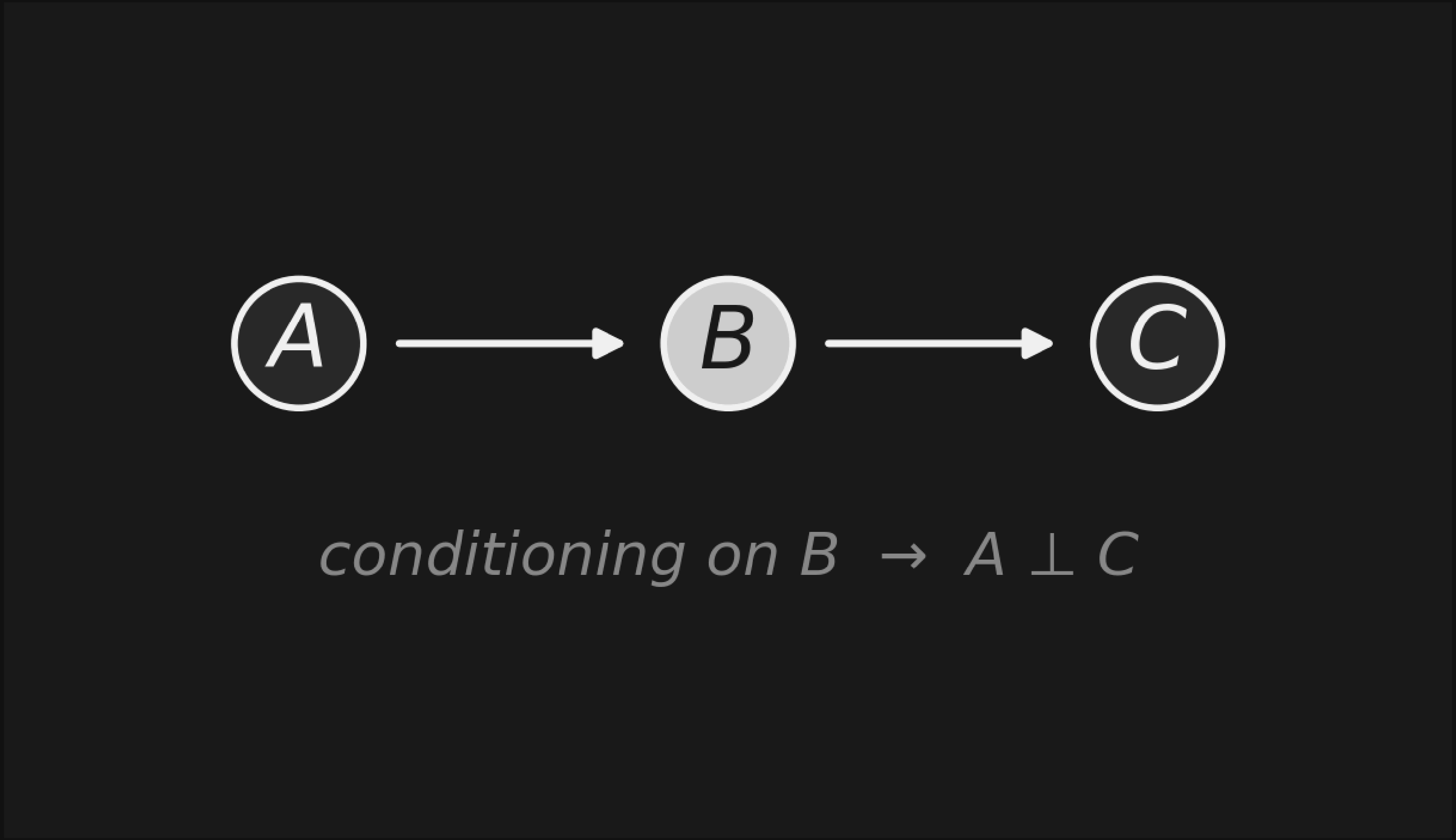
The Markov factorization tells us how the joint *factors*. But what about **conditional independence**?

Three patterns to know — every Bayes net is built from these:

- **Chain:** $A \to B \to C$
- **Fork:** $A \leftarrow B \to C$
- **Collider:** $A \to B \leftarrow C$

Conditioning on B behaves *differently* in each.

Chain: A → B → C

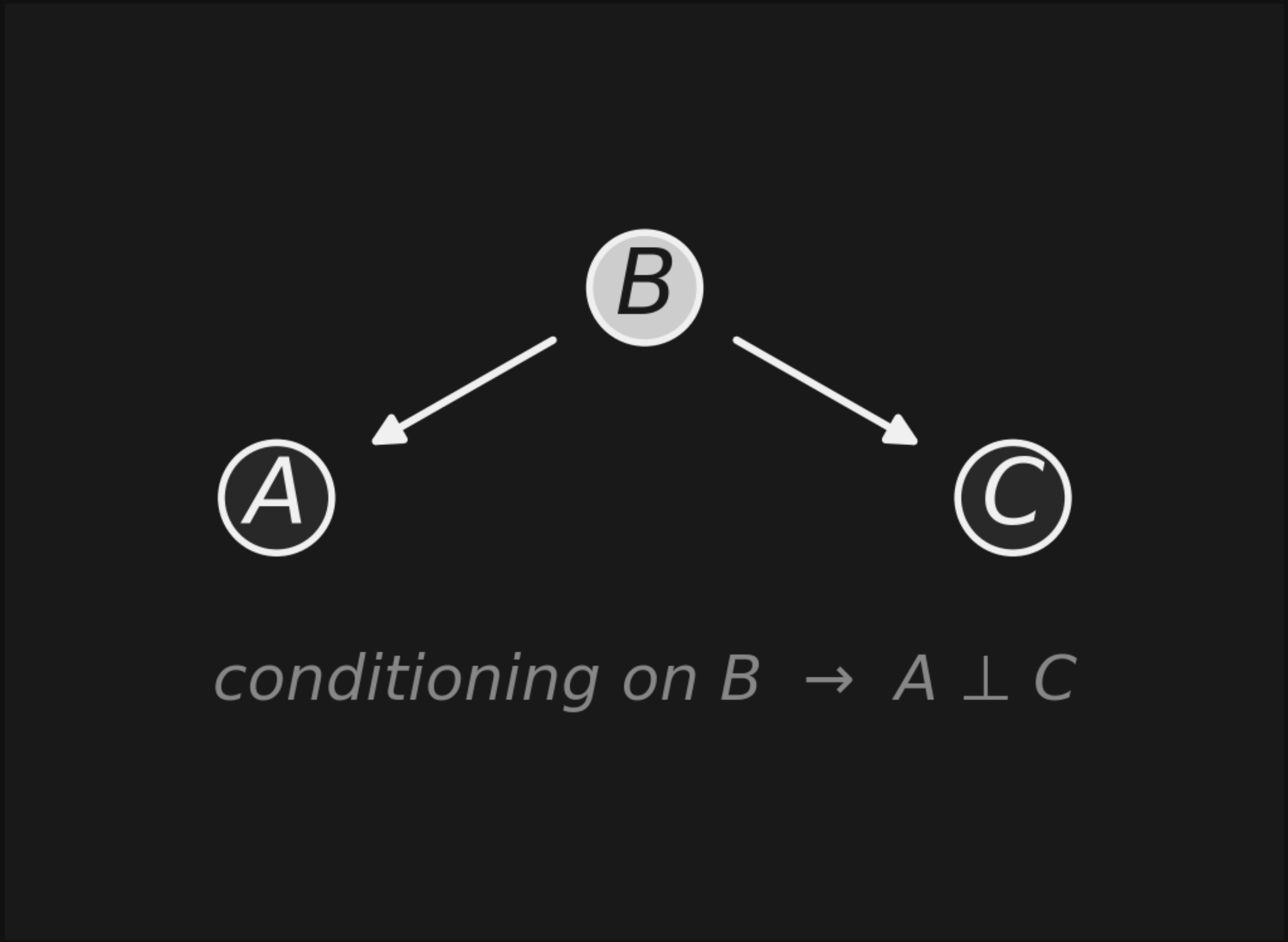


Conditioning on the middle node **blocks** information flow.

$$A \perp C \mid B$$

ChibaTech School of Design & Science

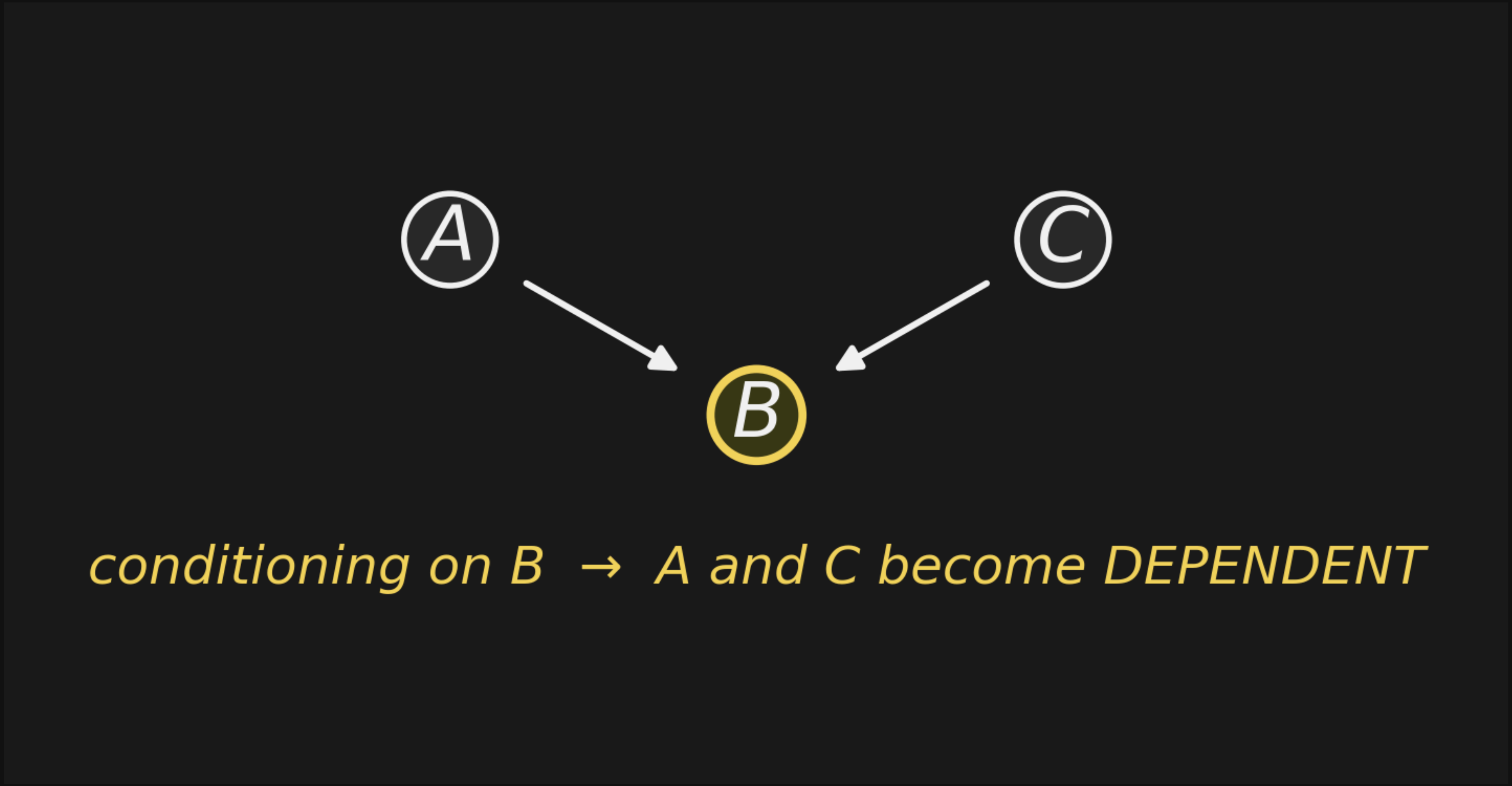
Fork: A ← B → C



A common cause. Conditioning on the cause **blocks** information flow.

$$A \perp C \mid B$$

Collider: A → B ← C



A common effect. Conditioning on the effect **induces** dependence between the otherwise-independent causes.

$$A \perp C \quad \text{but} \quad A \not\perp C \mid B$$

This is **backwards** from chain and fork. It's the surprise of the lecture.

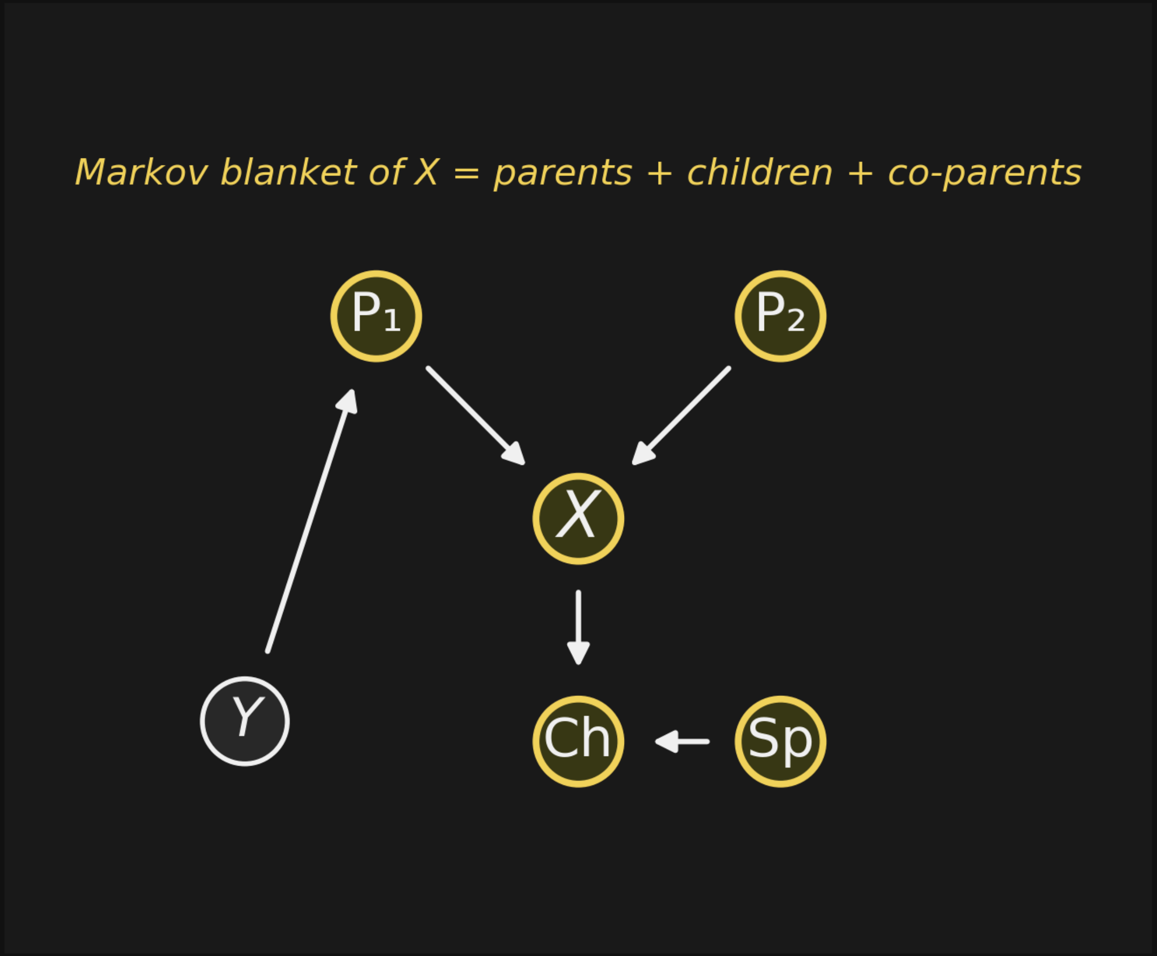
Markov blanket

Definition. The Markov blanket of node X:

- **parents** of X
- **children** of X
- **other parents** of X's children (the "spouses")

Conditioning on the Markov blanket makes X **independent of everything else** in the network.

The spouses are the non-obvious part — they're there because of *collider conditioning* flowing through X's children.



Bayes-Ball: interactive demo

Poll — collider conditioning

You see the cafeteria worker reveal a bento (not Chibany's pick, not tonkatsu). You haven't been told which bento Chibany picked. Does the worker's choice tell you anything about Chibany's pick?

- **A.** Yes — the worker's choice and Chibany's pick become dependent.
- **B.** No — they were independent before, so they're still independent.
- **C.** Only if you also know which bento has the tonkatsu.
- **D.** Chibany's pick was random, so nothing tells you about it.

Poll — answer

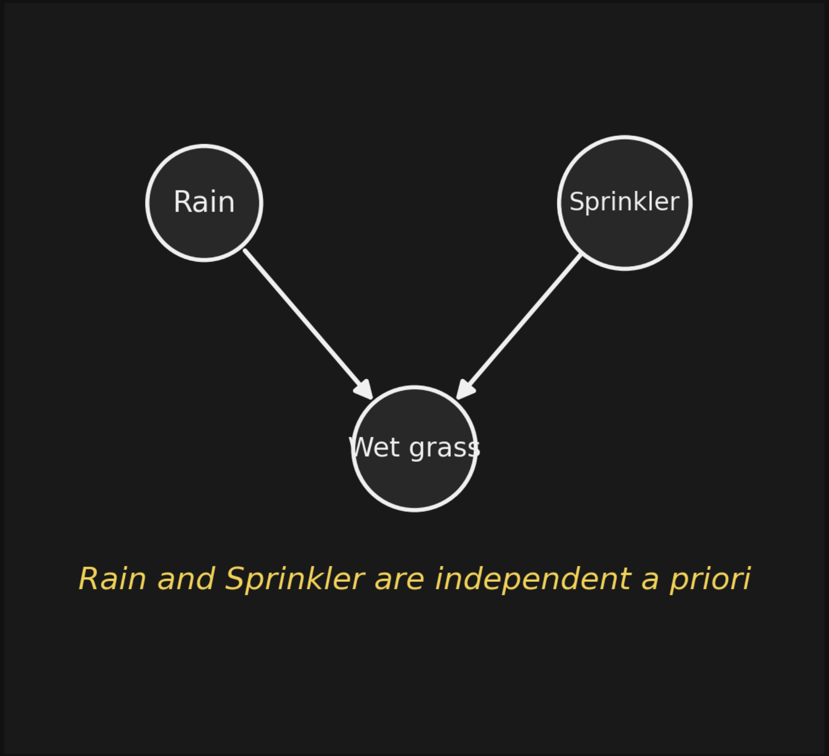
A. Yes — collider conditioning induces dependence.

The worker's choice constrains Chibany's pick: if the worker opened box 3, Chibany must have picked box 1 or 2 (otherwise the worker would have opened something else). The choices are now correlated *through* the constraint.

This is **explaining away** — Block 5.

Explaining away

Sprinkler / Rain / Wet grass



Three nodes: **Rain** (R), **Sprinkler** (S), **Wet grass** (W). Two independent causes of wet grass. A priori,

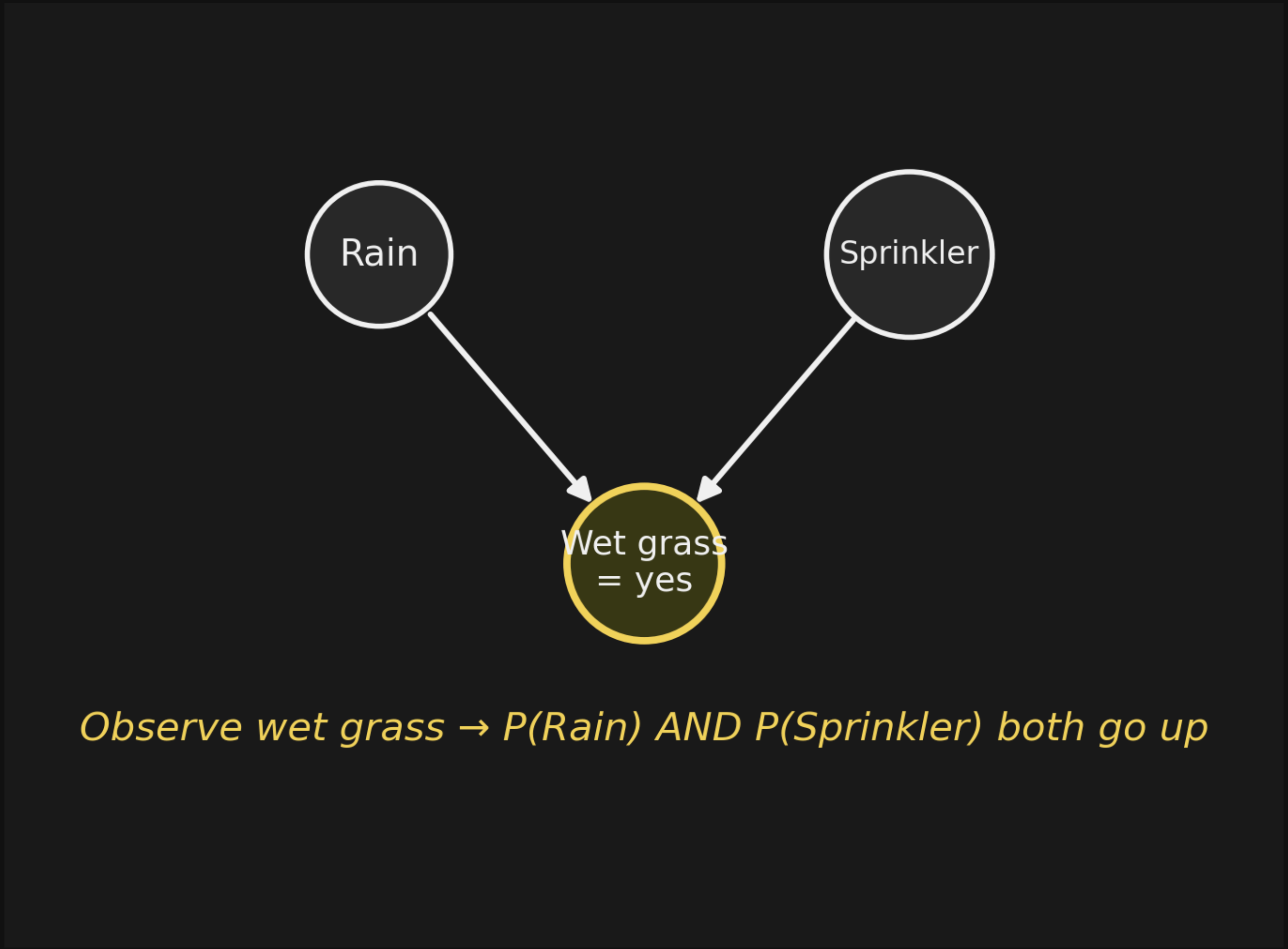
$$P(R) = 0.3, \quad P(S) = 0.3, \quad P(R, S) = P(R)\,P(S).$$

Rain and Sprinkler are **independent**. The grass is wet if **either** cause fires:

$$W = 1 \iff (R = 1) \text{ or } (S = 1). \quad \text{(deterministic OR)}$$

R, S, W from here on.

Observe wet grass

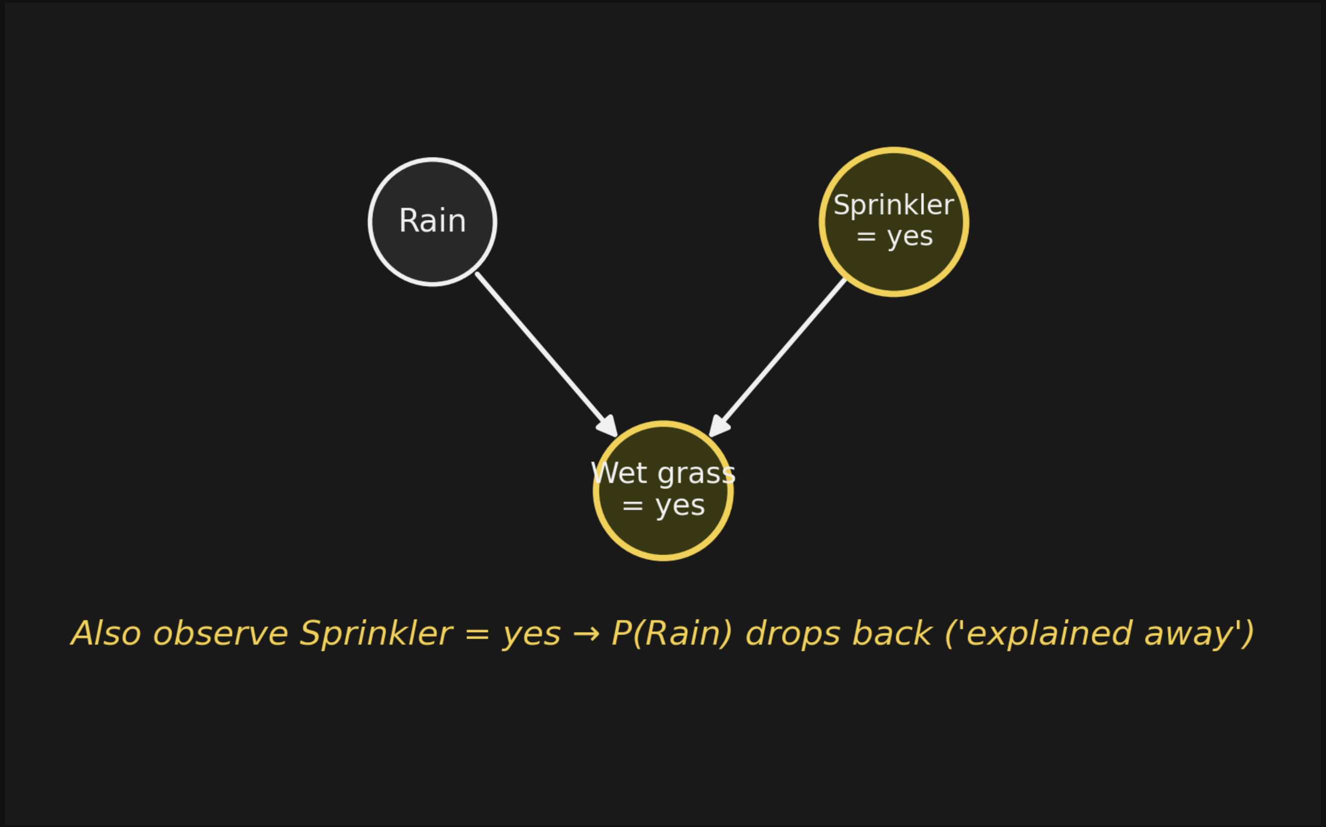


Now we walk outside: the grass is wet ($W = 1$). Both probabilities **rise**:

$$P(R \mid W) \approx 0.59, \quad P(S \mid W) \approx 0.59.$$

Either cause is now more plausible — the observation supports both.

Also learn sprinkler was on



A neighbor mentions: “I saw your sprinkler running this morning.” So $S = 1$. Now:

$$P(R \mid W, S = 1) = 0.3 = P(R).$$

Rain’s probability **drops all the way back to its prior** — sprinkler already “explains” the wetness, so wet grass tells us **nothing more** about rain.

Conditioning on the collider’s parents *and* the collider itself induces this competition. That’s explaining away.

EN Press L

What just happened

The numbers:

Conditioning set	$P(R=1)$
nothing	0.30
$W=1$	0.59
$W=1,\ S=1$	0.30

Rain rose, then fell back to its prior. The fall is **explaining away**.

The take-home in one sentence:

Conditioning on a common effect (the collider) makes the independent causes *dependent*. Then learning one cause *reduces* the support for the other.

Same mechanism, two stages.

Why Chibany should switch — explained

Remember Monty Hall? Tonkatsu and Chibany’s pick were independent a priori. We then conditioned on the **collider** (the worker’s reveal).

That’s explaining away.

The worker had to open *some* box. If Chibany picked correctly ($T = 1$), the worker had a free choice. If Chibany picked wrong ($T \neq 1$), the worker was *forced* to open the specific other box. The “forcing” is what makes the posterior favor switching.

Break

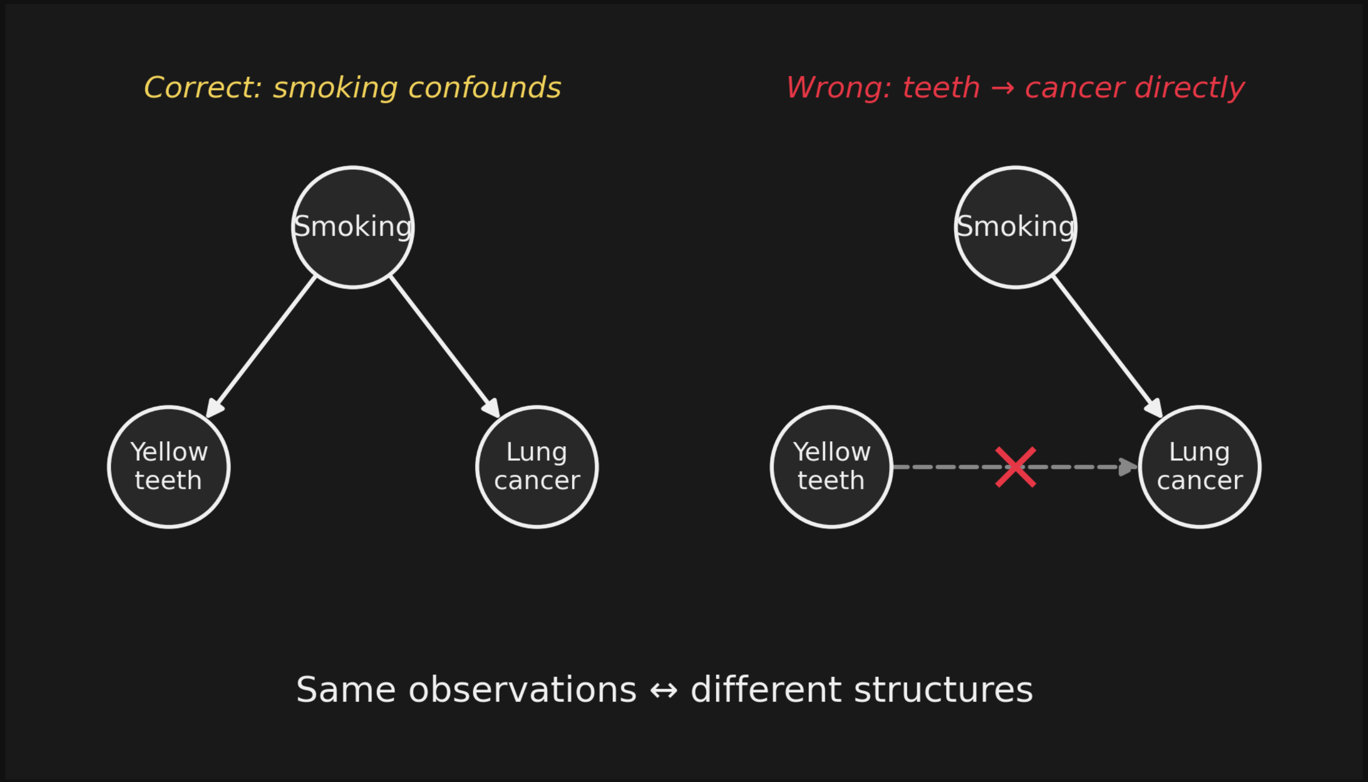
Observation → intervention

A claim from the dentist

Chibany is at the dentist. “People with yellow teeth,” the dentist says, “tend to have higher rates of lung cancer.”

Chibany asks: “Should I whiten my teeth?”

The confound



Three variables: **Smoking** (S), **Yellow teeth** (T), **Lung cancer** (L). Two structures give *exactly the same* statistical correlation:

- S causes both T and L (correct).
- T directly causes L (wrong, but observationally indistinguishable).

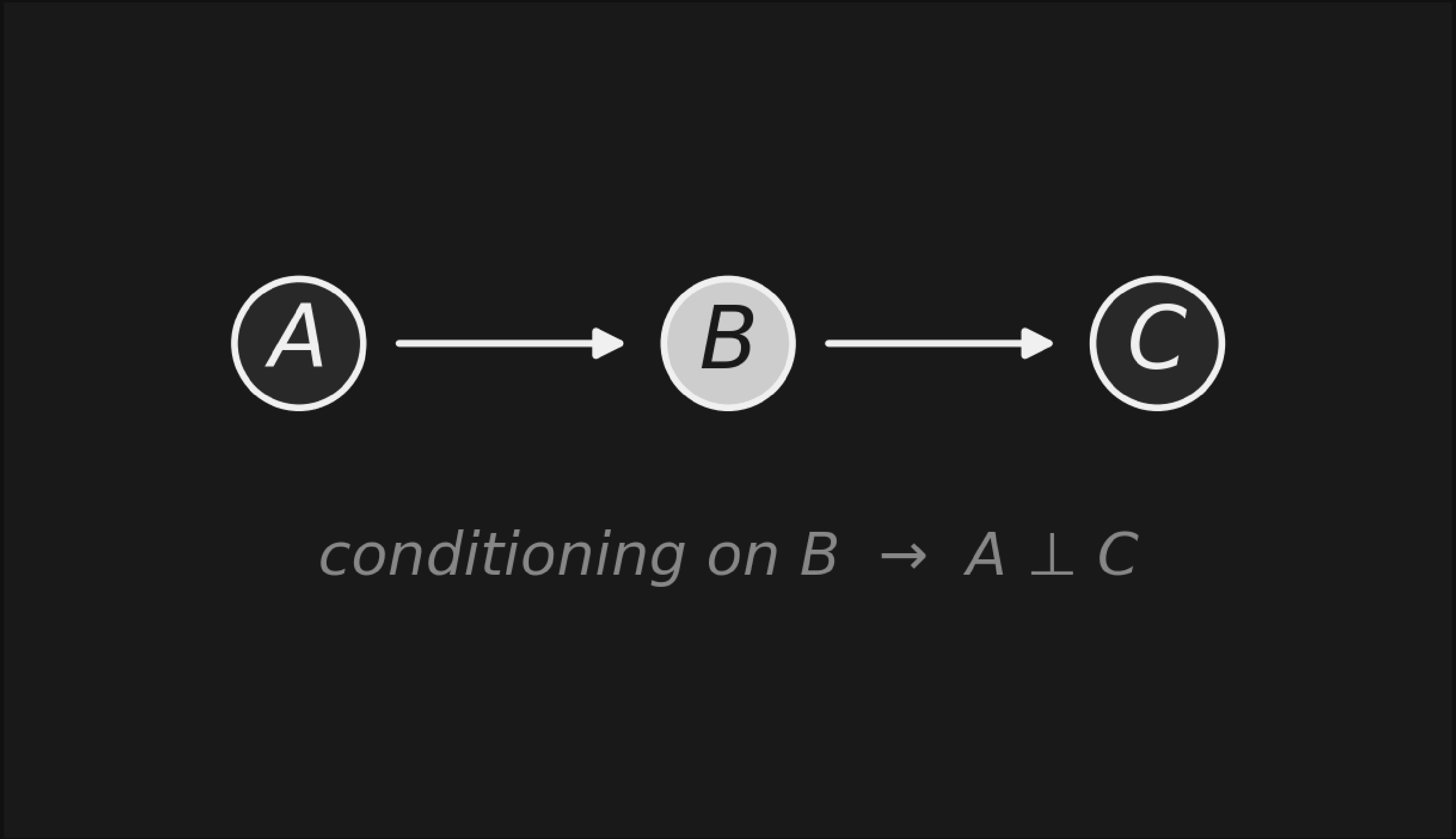
Observation alone cannot tell them apart. You'd see the same correlation either way.

S, T, L from here on.

ChibaTech School of Design & Science

EN Press L

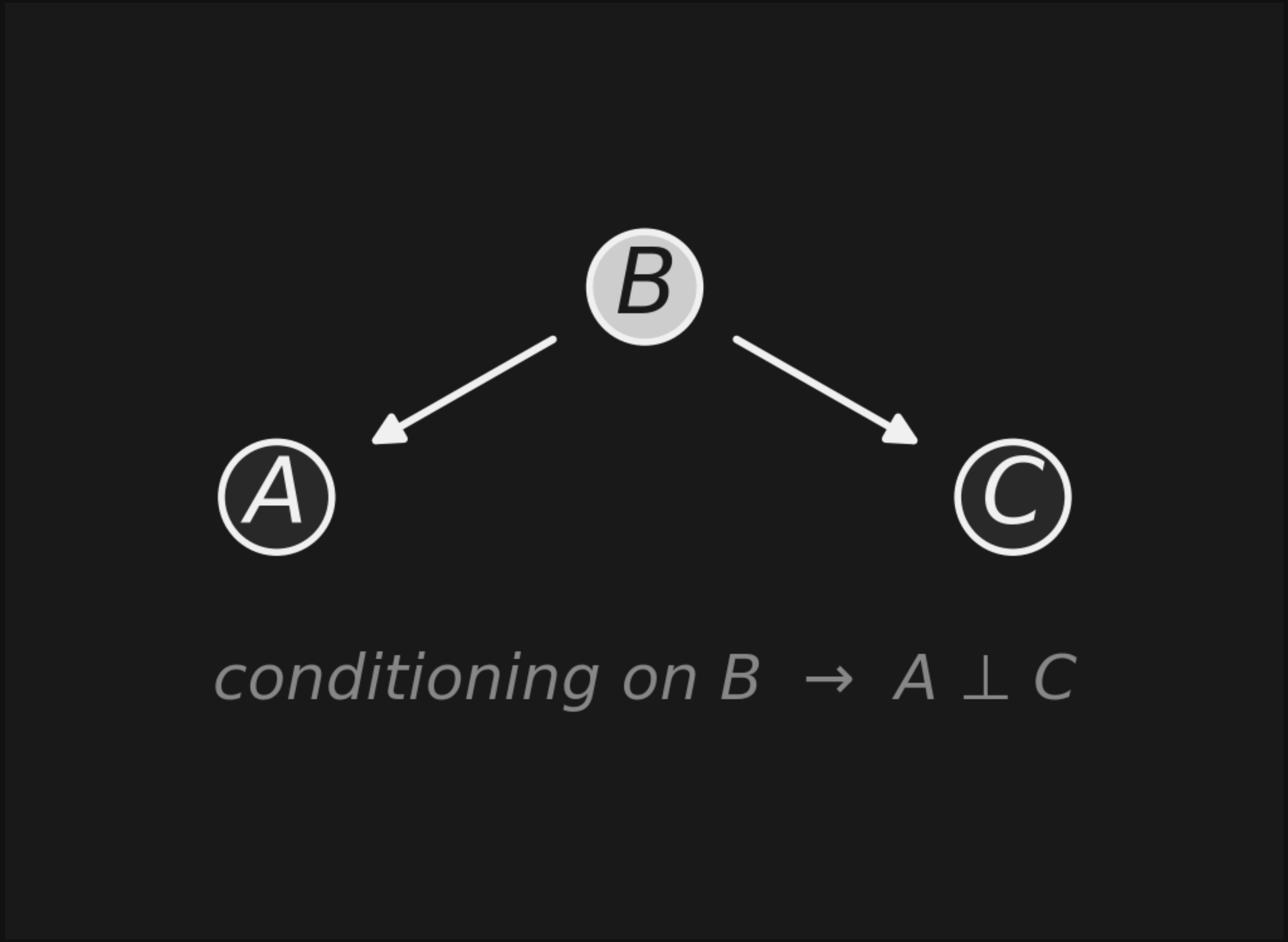
Three causal stories, one correlation: chain



Story 1 (chain): $T \to S \to L$ — yellow teeth cause smoking, smoking causes lung cancer. Implausible, but observationally consistent.

Conditional independence: $T \perp L \mid S$.

Three causal stories, one correlation: fork



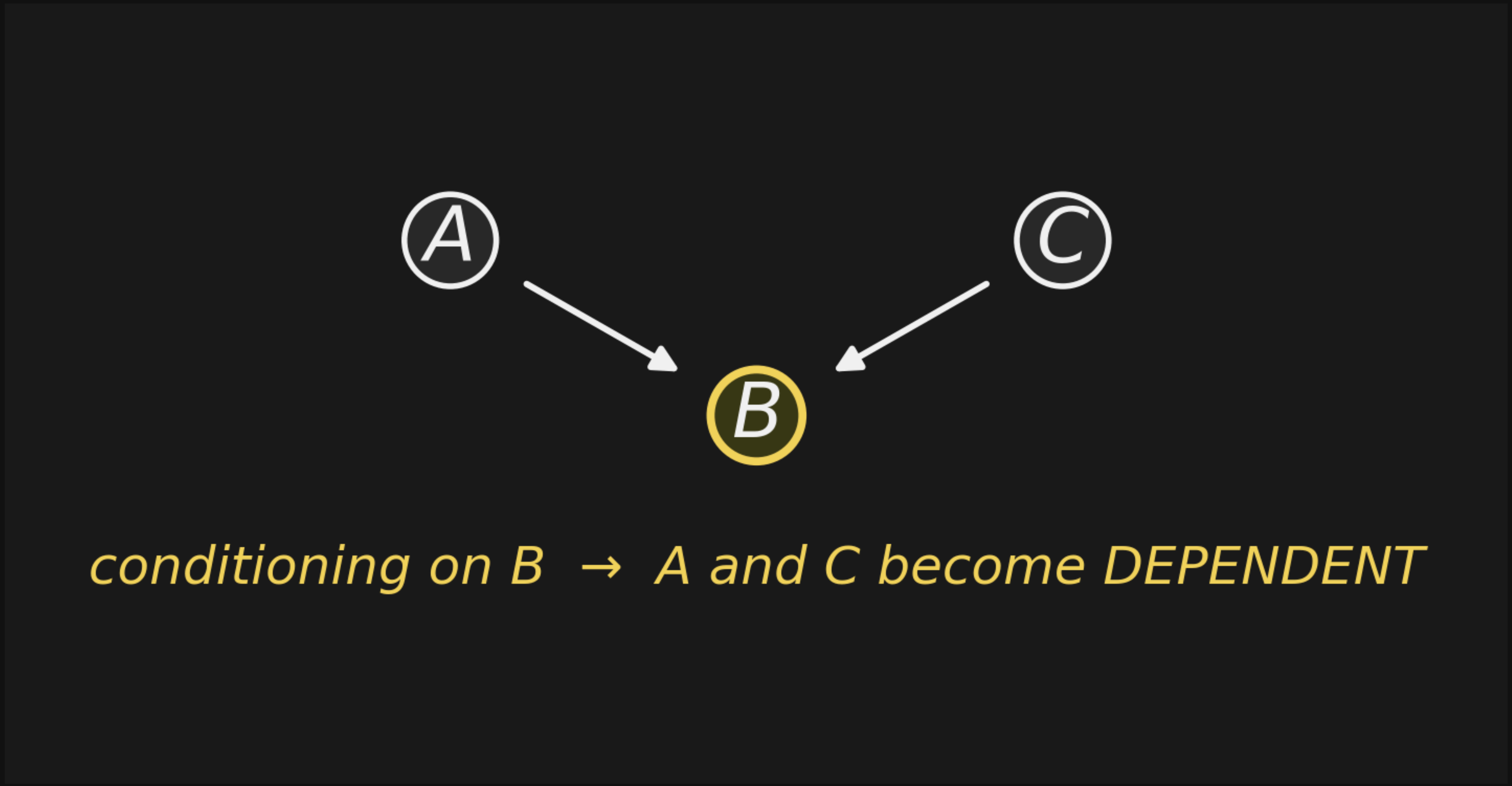
Story 2 (fork): $T \leftarrow S \rightarrow L$ — smoking causes both. The *real* story for the smoking-yellow-teeth-cancer triangle.

Conditional independence: $T \perp L \mid S$.

Identical to chain! Observation cannot tell these apart.

EN Press L

Three causal stories, one correlation: collider



Story 3 (collider): $T \to S \leftarrow L$ — yellow teeth and lung cancer each independently cause smoking (also implausible, but bear with it).

Conditional dependence: $T \not\perp L \mid S$.

Observationally distinct — but the wrong structure for a smoking confound. Most real-world confounds are forks, indistinguishable from chains.

What we need

What observation tells you:

- statistical dependence
- joint distribution
- can’t disambiguate chain from fork

This is **Level 1** of Pearl’s ladder.

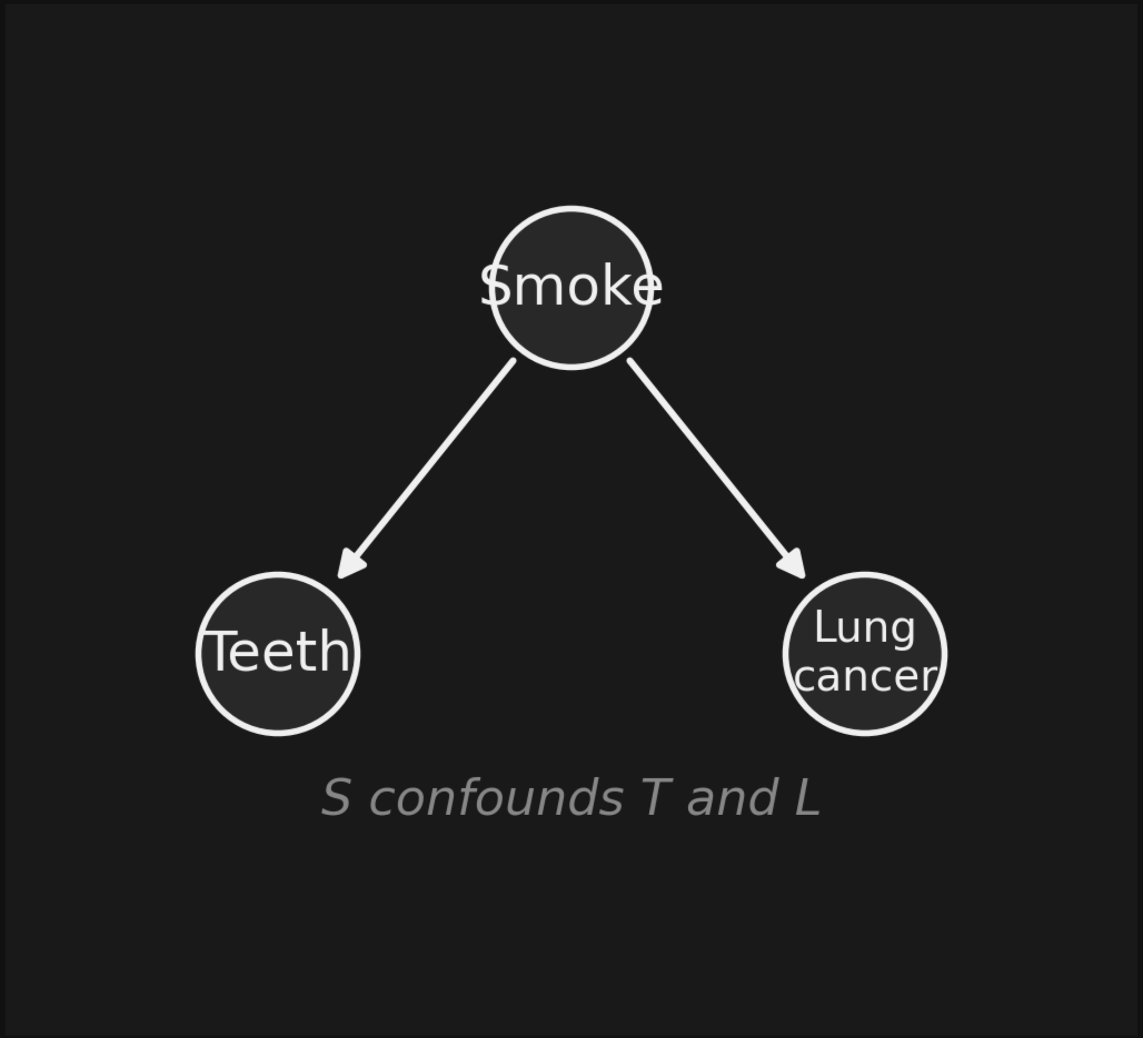
What intervention would do:

- cut incoming arrows
- reveal the cause
- break Markov equivalence

This is **Level 2** — coming up next.

The do-operator

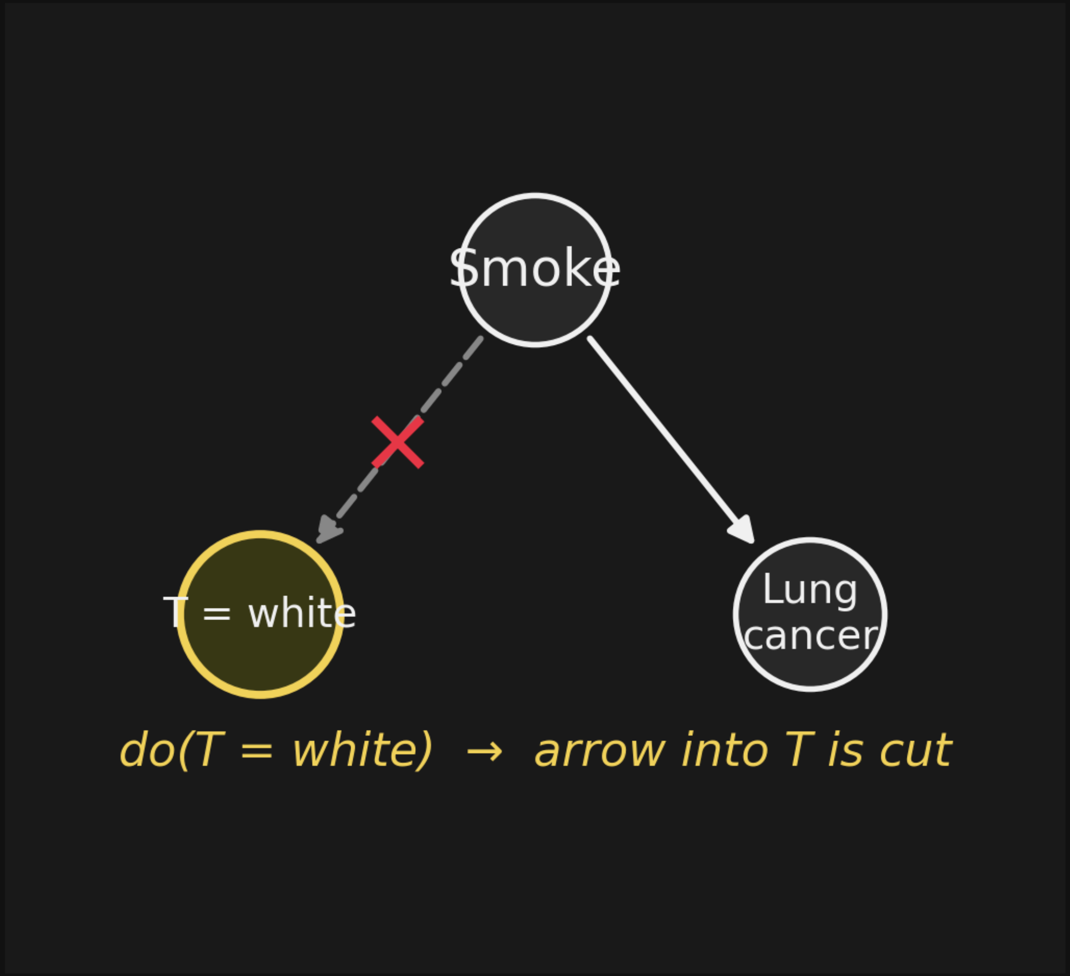
The original network



$S \to T$ (smoking causes yellow teeth), $S \to L$ (smoking causes lung cancer). T and L are correlated *because of* the common cause S.

If we **observe** $T = \mathrm{yellow}$, we update our belief about S, and *therefore* about L. Hence $P(L \mid T = \mathrm{yellow}) > P(L)$.

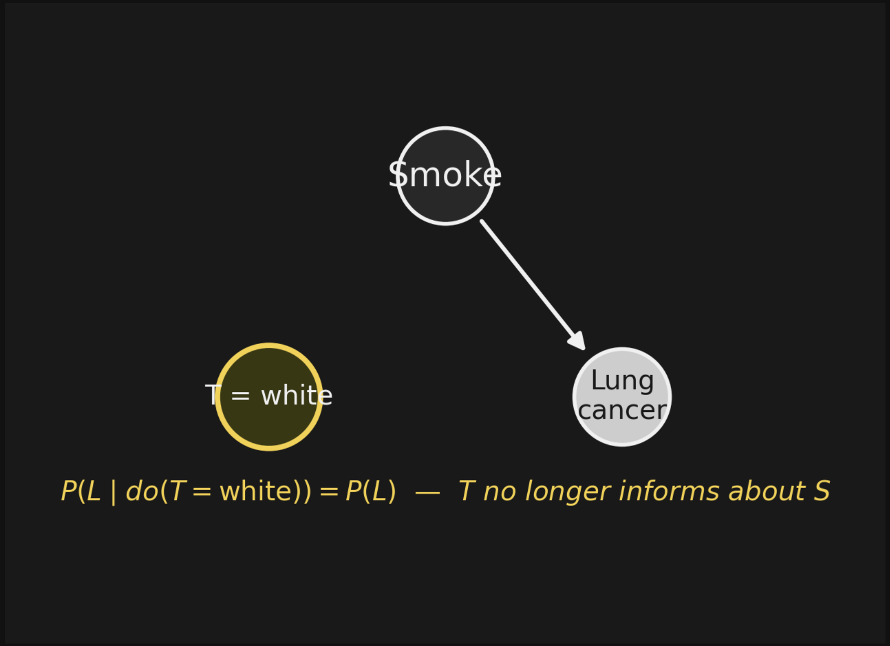
do(T = white): cut the incoming arrow



Graph surgery. We set $T = \mathrm{white}$ by intervention — we paid for whitening. The intervention **cuts the incoming arrow** $S \to T$.

Why cut? Because T is no longer being *generated* by smoking. T is being generated by Chibany's wallet. The mechanism changed.

Compute P(L | do(T = white))



After surgery, T has no parents. The factorization becomes:

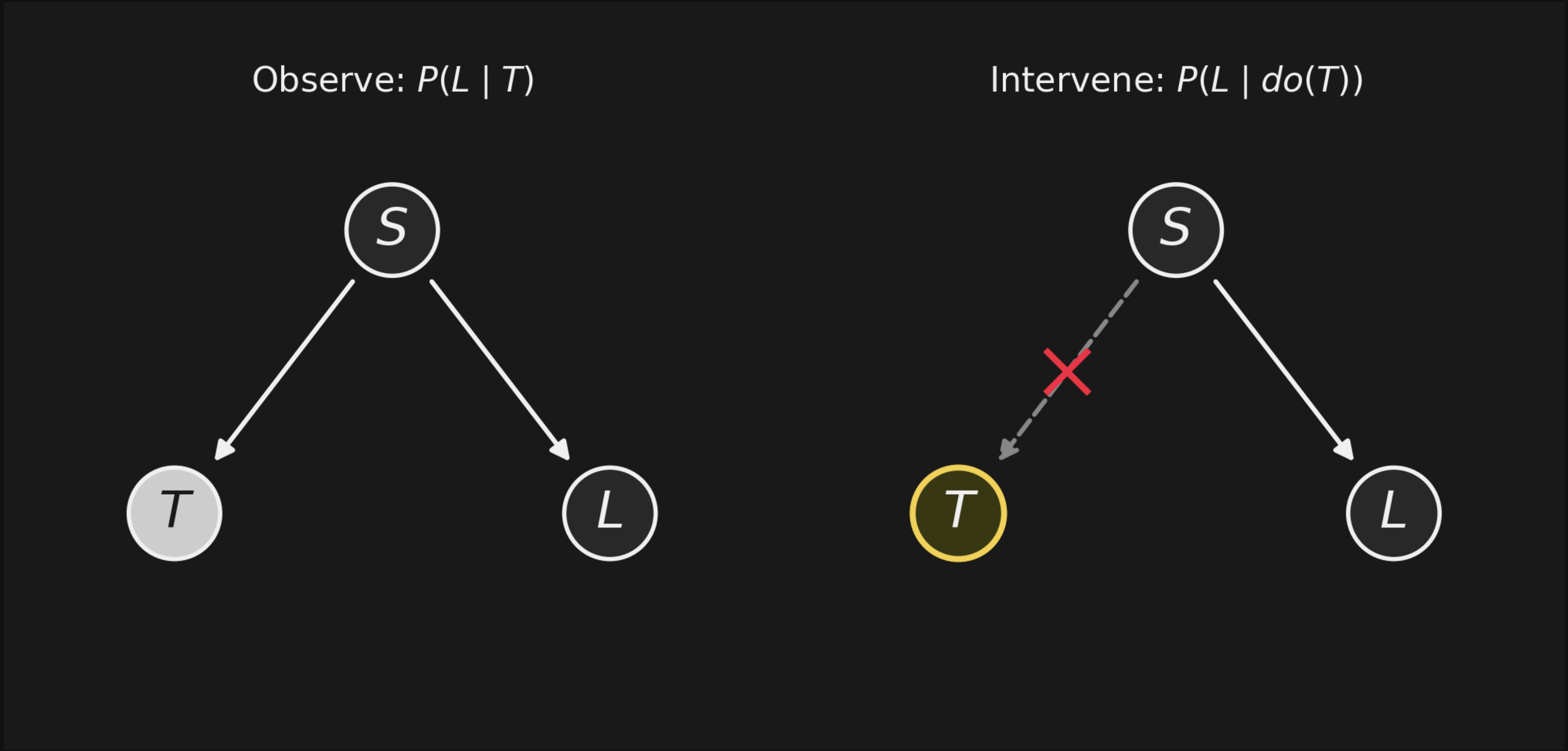
$$P(S, L, T = \text{white}) = P(S)\, P(L \mid S).$$

Marginalize over S to get $P(L \mid do(T = \text{white}))$:

$$P(L \mid do(T = \text{white})) = \sum_S P(L \mid S)\, P(S) = P(L).$$

Intervention severed the path from T back to L through S. T no longer carries any information about S, so knowing T tells you nothing new about L.

$P(L \mid T)$ vs. $P(L \mid do(T))$



Same notation, different operation, different answer.

- $P(L \mid T)$: passive observation. Yellow teeth tell you about smoking.
- $P(L \mid do(T))$: active intervention. Yellow teeth tell you about the dentist's bill.

Poll — do vs. condition

In the smoking network $S \to T, S \to L$, Chibany pays to whiten their teeth (intervention: $T = \text{white}$).

What is $P(\text{lung cancer} \mid do(T = \text{white}))$?

- **A.** Lower than $P(\text{lung cancer})$ — white teeth predict less smoking.
- **B.** Same as $P(\text{lung cancer})$ — intervention cuts the $S \to T$ edge.
- **C.** Higher than $P(\text{lung cancer})$ — paint chemicals are toxic.
- **D.** Unknown without more data.

Poll — answer

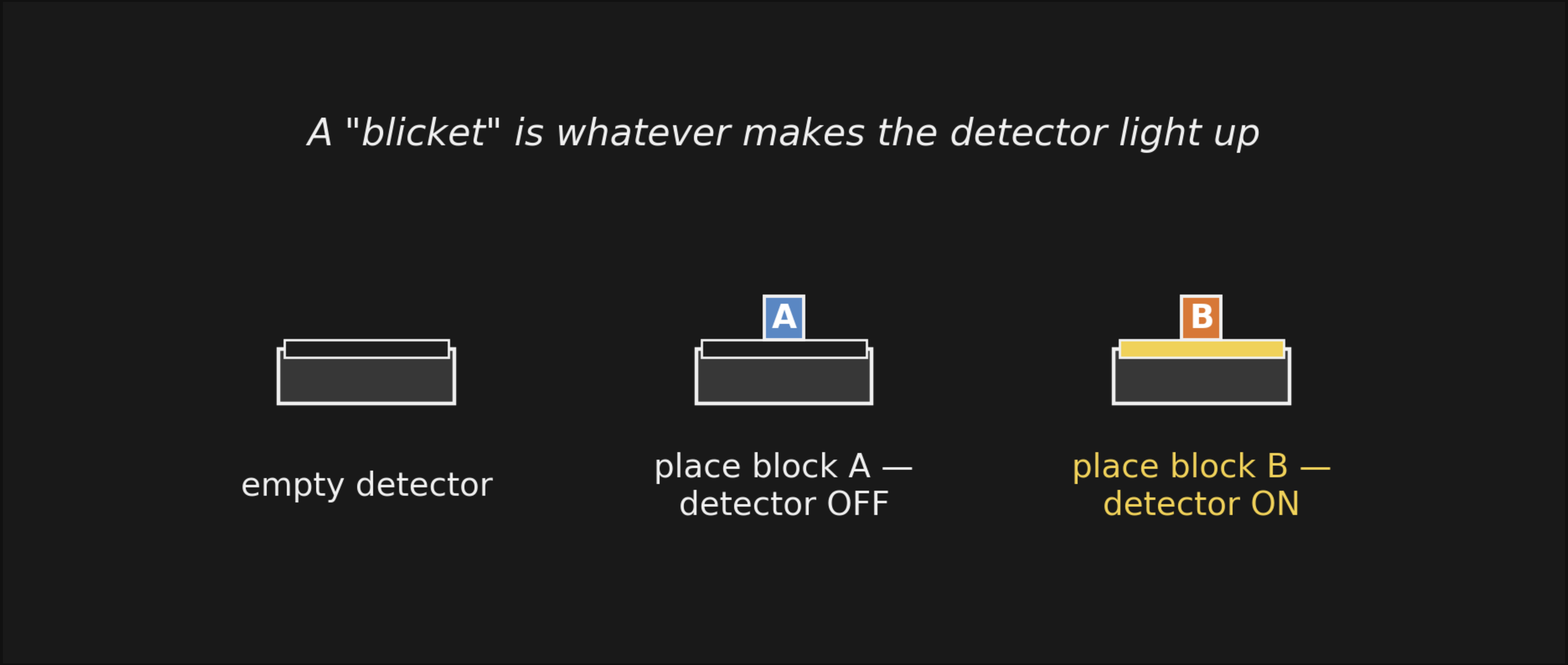
B. Same as $P(\text{lung cancer})$.

The intervention cuts the $S \to T$ edge. T no longer informs about S, and the path from T to L (which went through S) is severed.

(A) is the trap — it's the right answer if you *observe* T, but the wrong answer if you *intervene* on T.

Causal cognition — the blicket detector

The blicket detector



A children’s experiment (Gopnik & Sobel, 2000). A “blicket” is whatever makes the machine light up. Children watch blocks placed on the detector and infer which blocks are blickets.

The experiment as a Bayes net

The physical setup:

- A machine ("detector") that lights up.
- A collection of blocks. Each block is or isn't a blicket.
- Place blocks on the machine; observe whether it lights up.

The child's job: figure out which blocks are blickets.

The Bayes net:

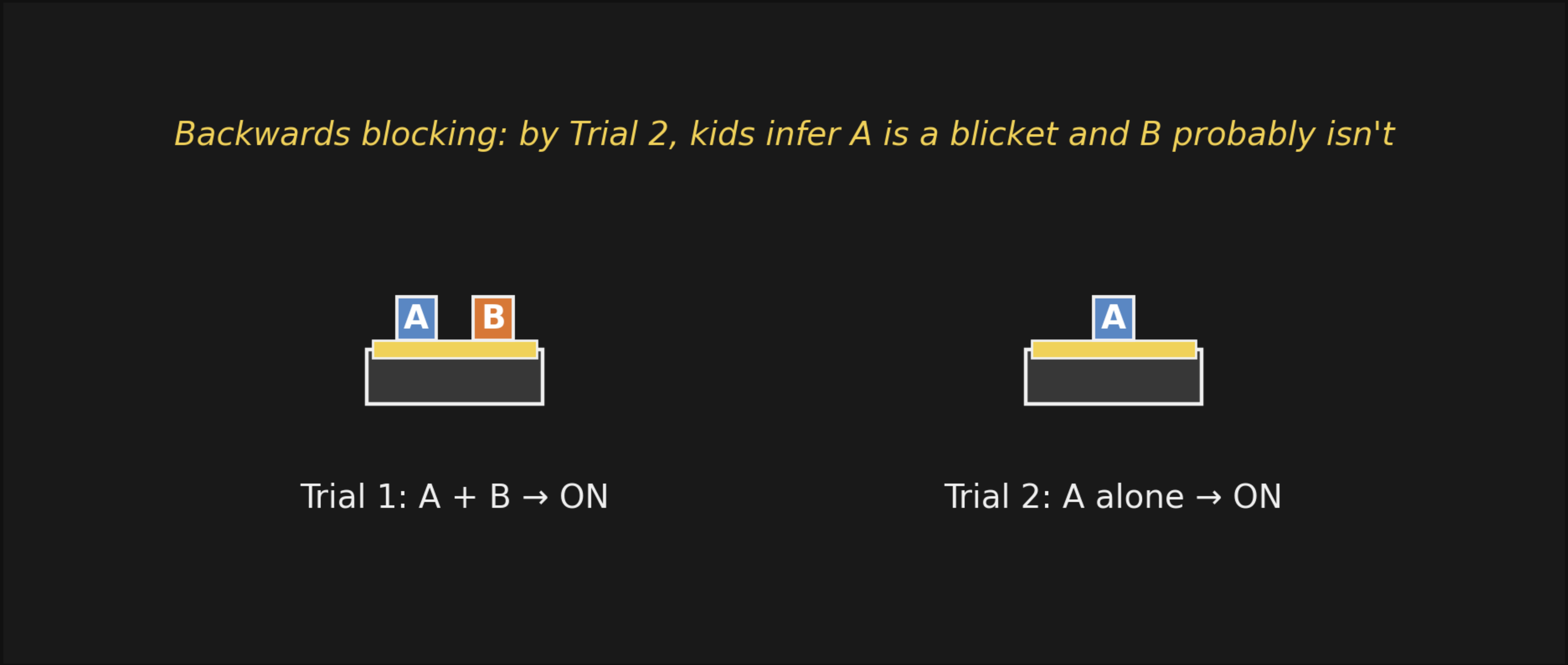
For two blocks A, B:

- A-blicket? → Detector lights
- B-blicket? → Detector lights

A **collider**: detector is the common effect.

Each trial **conditions on the collider**. Inference is *explaining away*.

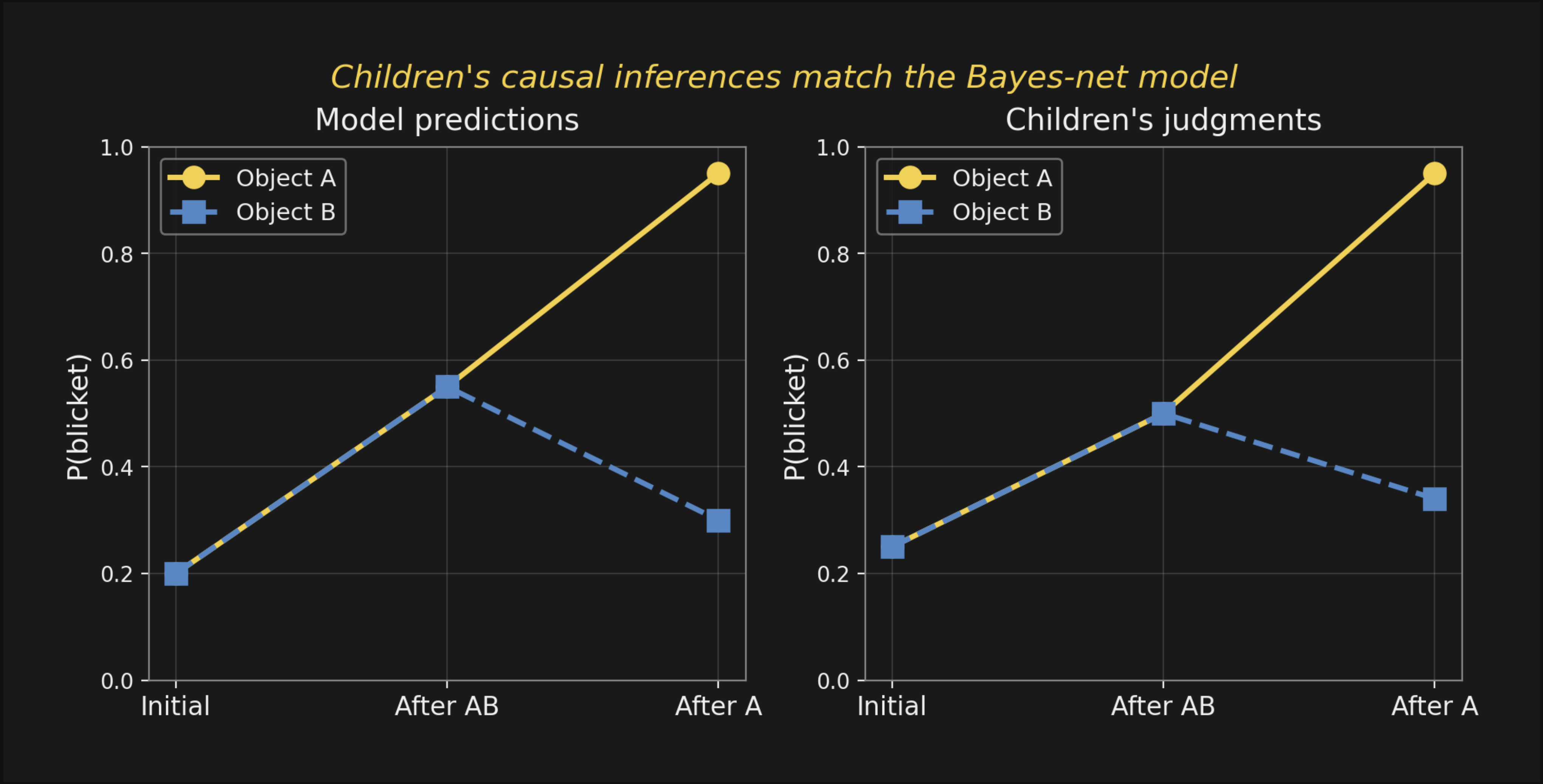
Backwards blocking



- **Trial 1:** A and B together → detector ON.
- **Trial 2:** A alone → detector ON.

After Trial 2, what do kids think about B?

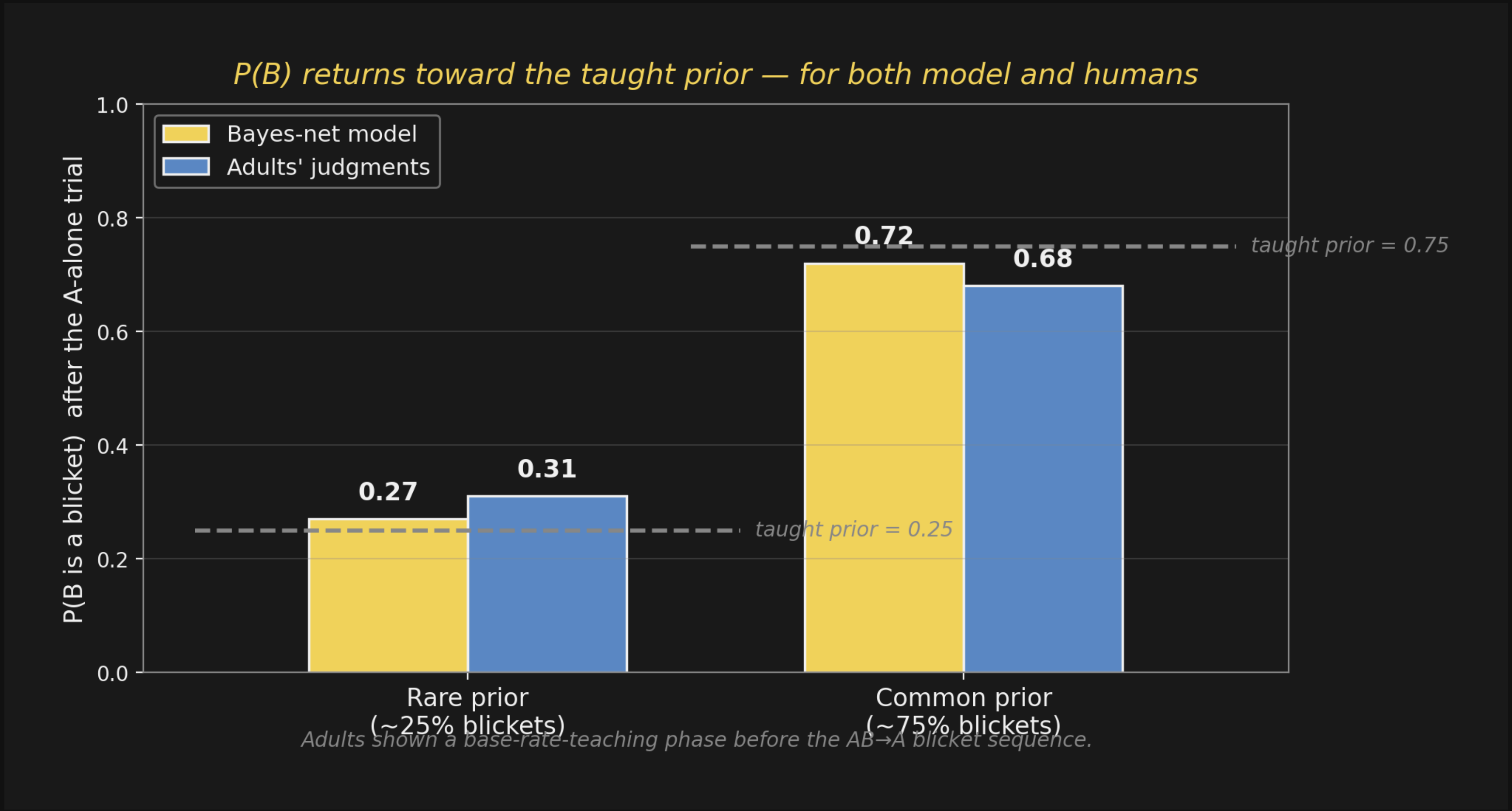
Children match the Bayes net



Children’s judgments **drop** for object B after the A-alone trial — even though nothing changed about B’s data. **Children do graph surgery.** They reason about the causal structure, not just statistical patterns.

The formal machinery isn’t just for engineers.

Did the prior come back?



Design. Before the AB → A blicket sequence, adults watch a base-rate-teaching phase showing **~25%** ("rare prior") or **~75%** ("common prior") of arbitrary blocks are blickets. Then they run the classic backwards-blocking trials and rate $P(B)$.

Bayes-net prediction. After A alone explains the AB result, the posterior on B should **relax back toward the taught prior**. Associative learning predicts a uniformly low rating regardless.

Result. Both the model AND adults snap $P(B)$ back to the taught base rate — strong evidence that adults are reasoning *Bayesianly* over a causal graph, not just learning associations.

ChibaTech School of Design & Science

EN Press L

Information theory + close

Entropy as expected surprise

Define the **surprise** of seeing event $X = x$ as

$$\text{surprise}(x) \;=\; \log \frac{1}{P(x)} \;=\; -\log P(x).$$

Smaller probability → larger surprise. The minus sign is there because $\log P(x) \le 0$ when $P(x) \le 1$, so $-\log P(x) \ge 0$ — surprise is non-negative.

Event probability	Surprise (bits)
$P(x) = 1$ (certain)	0
$P(x) = 1/2$ (fair coin)	1
$P(x) = 0.01$ (rare)	≈ 6.6

Mutual information

How much does knowing Y reduce uncertainty about X?

$$I(X;Y) \;=\; H(X) - H(X \mid Y)$$

Properties:

- $I(X;Y) \geq 0$ always.
- $I(X;Y) = 0 \iff X \perp Y$.
- $I(X;Y) = I(Y;X)$ — symmetric.

Conditional version:

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z).$$

$X \perp Y \mid Z$ iff $I(X;Y \mid Z) = 0$.

The collider, in info-theoretic clothing

Take the Sprinkler / Rain / Wet-Grass collider:

- $I(\mathrm{Rain}; \mathrm{Sprinkler}) = 0$ — independent a priori.
- $I(\mathrm{Rain}; \mathrm{Sprinkler} \mid \mathrm{Wet}) > 0$ — dependent after conditioning.

Conditioning on a collider creates mutual information from nothing.

That’s explaining away — viewed through a different lens. Info theory and Bayes nets are two languages describing the same structural facts.

Next week — Week 6

Markov chains + networks. Random walks. Memory search.

Reading: Abbott, Austerweil & Griffiths (2012) — *Human memory search as a random walk in a semantic network.*

We’ve spent today on the **structure** of multi-variable distributions. Next week we **walk** on a network instead of conditioning on one.

Thanks

Questions?

Office hours: by appointment. Slack me.

Reflections due before next Friday — pick any required or optional reading from this week. 5 of 11 across the semester.