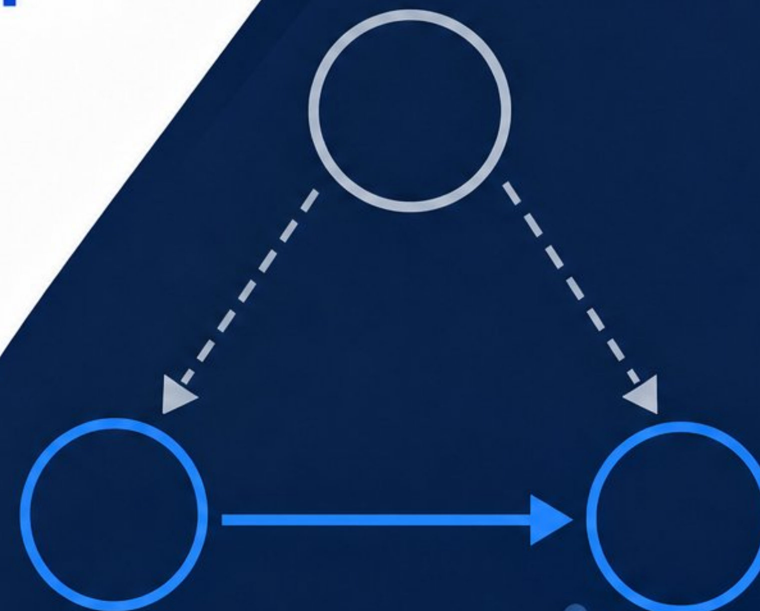
Correlation, Causation, and Confounding:

Making Sense of Relationships in Data

- Separating association from causation is critical for sound decision-making
- This training covers definitions, visual tools, pitfalls, and safeguards
- 2026 example: Unproven Tylenol-autism claim led to 10% drop in pregnant women's use
- Goal: Build 'causal literacy' to evaluate claims and ask better questions

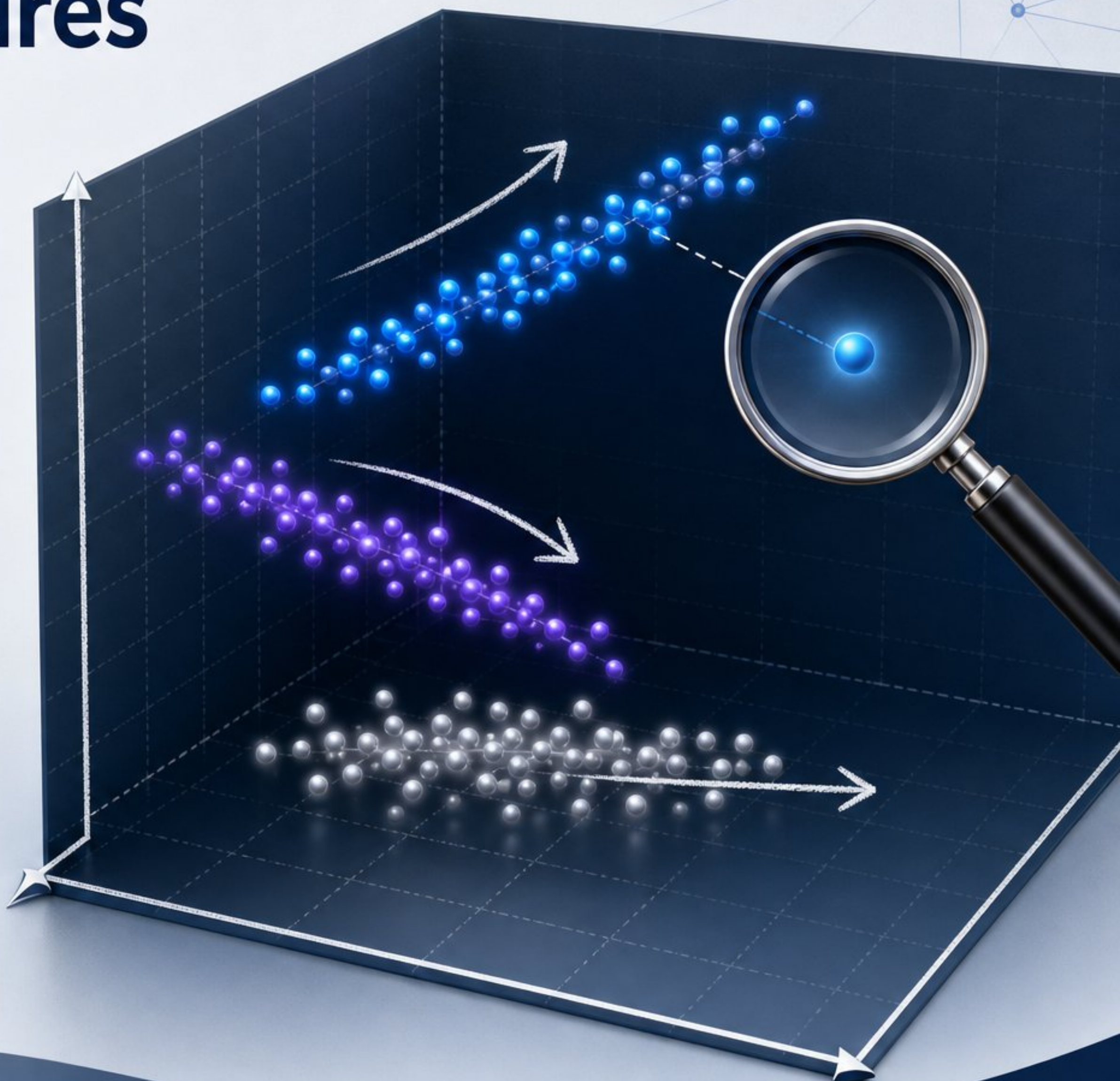


The High Cost of Getting It Wrong

- Trump's 2025 Tylenol-autism claim led to a 10% drop in use by pregnant women, risking untreated fevers.
- Ivermectin prescriptions for cancer nearly doubled after a 2025 celebrity endorsement, despite no causal evidence.
- A business may see engaged email users spend more and wrongly conclude emails cause spending.
- **Goal:** Shift from "this proves it" to "this question needs investigation."

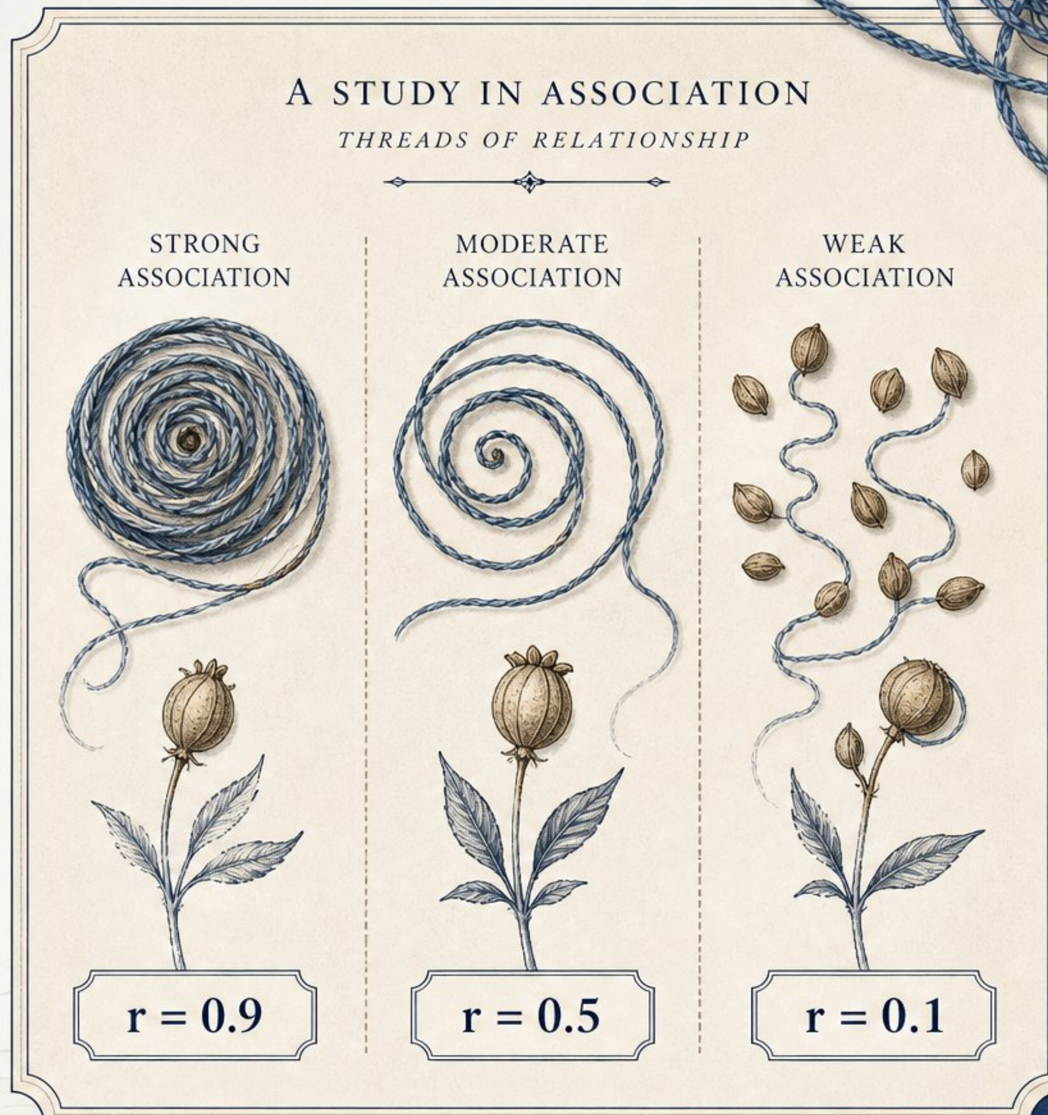
What Correlation Actually Measures

- Correlation measures strength and direction of linear relationships
- Positive: variables move together; Negative: move oppositely
- Zero correlation means no linear pattern exists
- It quantifies co-movement, but never explains why
- Visual tools: scatterplots, line charts, coefficients, heatmaps



Decoding Correlation Coefficients

- Pearson's r: strength and direction of linear relationships (-1 to +1)
- Spearman's rho: handles monotonic relationships using ranks
- Robust to outliers; does not require strict linearity
- Strong correlation (e.g., r=0.9) signals a pattern, not causation



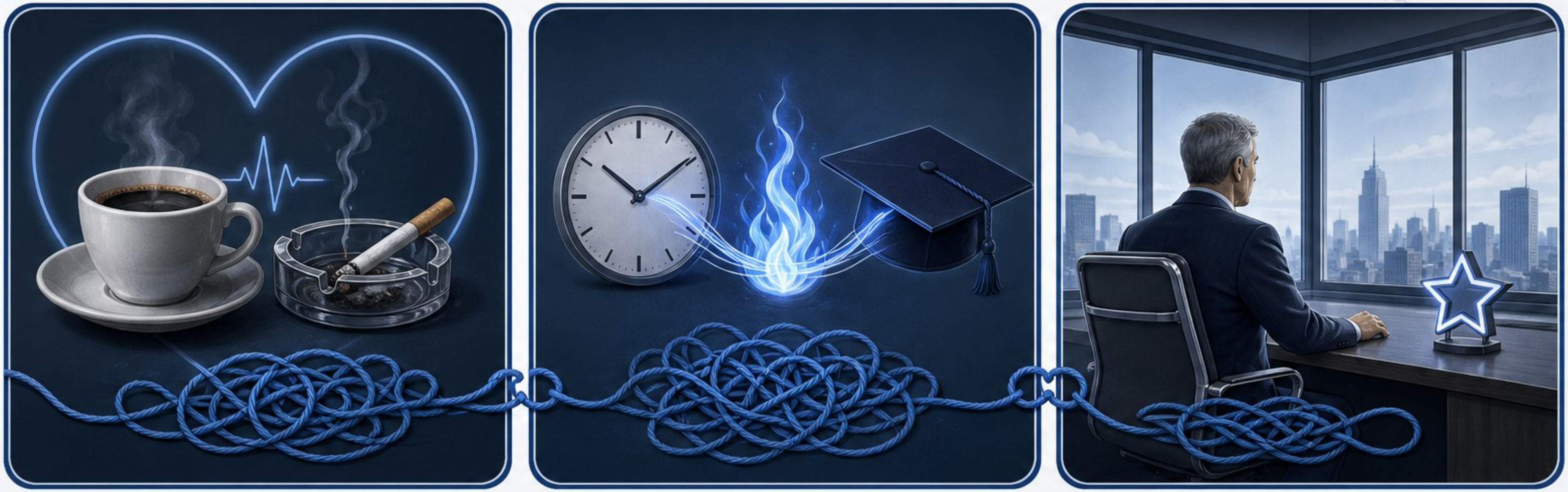
What Causation Really Means

- **Causal relationship:** Changing X produces a change in Y, other factors held constant
- **Three criteria:** Temporal order, plausible mechanism, elimination of alternatives
- **Gold standard:** RCT or A/B test breaks treatment-confounder links
- **Observational data** rarely proves causation; quasi-experiments can provide strong evidence

The Hidden Third Variable: Understanding Confounding

- A confounder influences both cause and effect, creating false links
- Ice cream sales and drownings: both rise due to hot weather
- Employee training vs. performance: leadership quality may drive both
- Never control for a collider—it opens bias, it doesn't remove it

Confounders in Action: From Health to HR



- Coffee & heart disease link confounded by **smoking**—heavier coffee drinkers also smoke more.
- Class attendance & grades confounded by **intrinsic motivation**—it drives both attendance and study time.
- Age & job satisfaction confounded by **job position**—older employees often hold more gratifying roles.
- **Domain expertise** is essential: you can't spot a confounder without understanding the system.

Simpson's Paradox: When Trends Reverse

- A trend in groups can reverse when data is combined
- Unequal group sizes and a lurking variable distort results
- Women admitted at higher rates per department, lower overall
- Due to applying disproportionately to more competitive programs
- Always ask: Could a hidden grouping drive this pattern?

Spotting Non-Causal Patterns: A Practical Checklist

Red flag:
relationship vanishes when splitting data by a third variable

Trap:
both variables grow over time, but neither causes the other

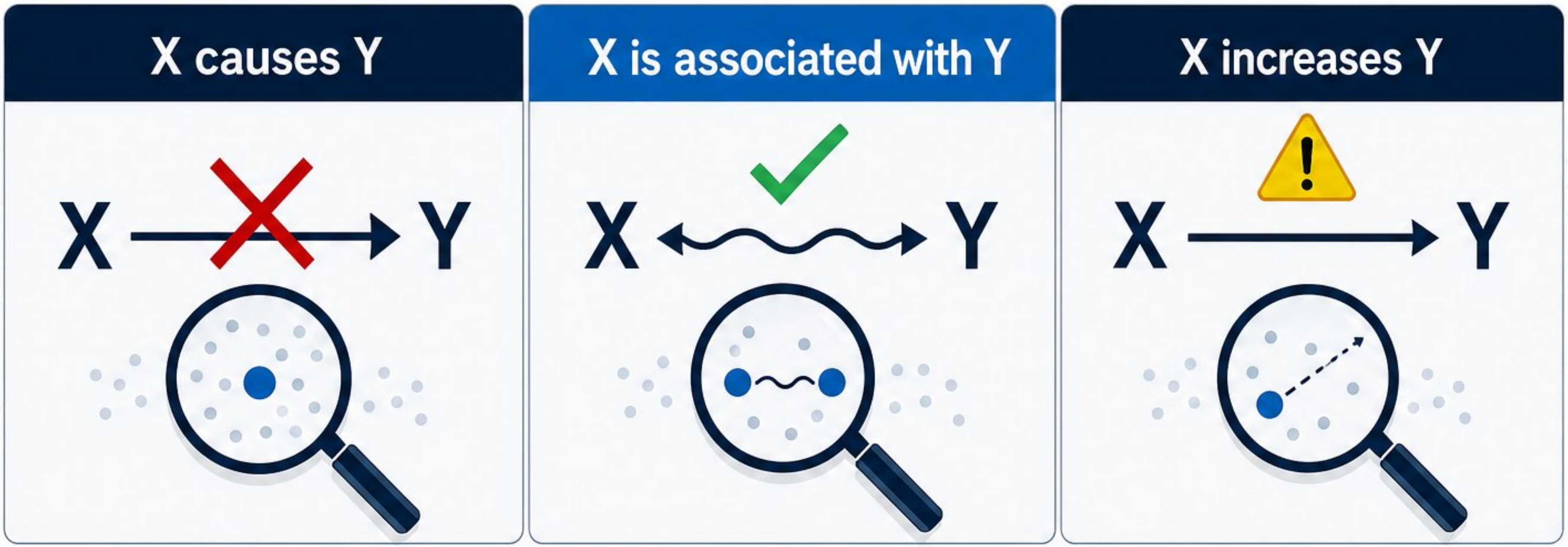
Check:
temporal order, plausible mechanism, alternative explanations

Ask:
“Could a hidden group or confounder drive this pattern?”

Mapping Causal Assumptions with DAGs

- **DAG:** nodes are variables, arrows are direct causal effects, no cycles allowed
- Makes implicit assumptions **explicit** and **auditable** before analysis
- **Backdoor Criterion** reveals which variables to control for and which to avoid
- **Practical workflow:** whiteboard the DAG with your team, then decide what to measure

Practical Strategies to Avoid the Correlation Trap



- Replace 'X leads to Y' with 'X is associated with Y' when evidence is only correlational
- Phrases like 'is linked with' are often interpreted by the public as 'causes'
- Strengthen causal claims with natural experiments, diff-in-diffs, or instrumental variables
- **Pre-decision checklist:** Is the claim causal? What alternatives exist? Who is affected if wrong?

Applying the Framework in Your Work

- **Map domain variables:** marketing, HR, product, ops KPIs
- **Watch common confounders:** seasonality, engagement, tenure
- **Challenge claims:** 'What else could explain this pattern?'
- **Commit to one** behavior change when you see a data claim

Key Takeaways and Extended Learning

- Correlation measures co-movement; causation needs direct, isolated influence
- Confounding creates spurious correlation through a hidden third variable
- Mindset shift: "This correlation raises a question I should investigate"
- Metaphor: umbrellas correlate with accidents—rain causes both
- Resources: The Causal Mindset Handbook (Gallea, 2026) and Cause and Effect Business Analytics (Haughton et al., 2026)