CS372 Assignment 1: Relevant Lecture Content Summary

Assignment Context

Goal: Scale the T3 Causal Benchmark from 454 → 5,000+ vignettes for rigorous algorithm evaluation. Each group must analyze their assigned BenchmarkT3-BucketLarge file and expand it by approximately 10× while maintaining quality.

---

1. Pearl's Ladder of Causation (Core Framework)

The T3 Benchmark organizes cases by Pearl's three levels of causal reasoning:

Level 1: Association (L1)

- **Question Type:** P(Y|X) — "What if I see...?"
- **Requires:** Observational data only
- **Example Question:** "Do participants in Program A tend to get employed?"
- **LLM Capability:** ✓ LLMs can handle this level

Level 2: Intervention (L2)

- **Question Type:** P(Y|do(X)) — "What if I do...?"
- **Requires:** Causal graph + adjustment for confounders
- **Example Question:** "If we assign this person to Program A, what is their employment probability?"
- **LLM Capability:** Limited — requires causal graph and assumptions

Level 3: Counterfactual (L3)

- **Question Type:** P(y_x|x', y') — "What if I had...?"
- **Requires:** Full Structural Causal Model (SCM)
- **Example Question:** "This person did Program A and got employed. Would they have without it?"
- **LLM Capability:** × Not supported without SCM

Key Insight: Level-1 observational distributions cannot identify Level-2 interventional or Level-3 counterfactual quantities without additional assumptions or experimental information.

2. Key Causal Definitions (Essential for Case Classification)

Confounder

- A common cause of both X (treatment) and Y (outcome)
- Creates spurious association between X and Y
- **Action:** ADJUST for confounders to remove bias
- **DAG Structure:** Z → X and Z → Y (arrows pointing out from Z)

Mediator

- On the causal pathway between X and Y
- Explains HOW causes work
- **Action:** Do NOT adjust for mediators when estimating total effect (blocks the causal pathway)
- **DAG Structure:** X → Z → Y

Collider

- A common effect of both X and Y
- **Action:** Do NOT adjust — adjusting CREATES bias (opens a path that was closed)
- **DAG Structure:** X → Z and Y → Z (arrows pointing into Z)

Backdoor Path

- Any path from X to Y that begins with an arrow INTO X
- If open, induces non-causal (spurious) association
- Must be blocked for valid causal inference

---

3. Simpson's Paradox (Signature Trap for Group J)

The Core Problem

A trend that appears in aggregated data can reverse when the data is broken down by subgroups.

Job Training Example from Lecture:

	Program A	Program B
Overall Employment Rate	40%	50%

Naive conclusion: Program B is better

But stratified by experience:

	Program A	Program B
Experienced	80% employed (n=200)	70% employed (n=600)
Entry-level	30% employed (n=800)	20% employed (n=400)

Program A is better in EVERY subgroup, yet worse overall.

Why This Happens

- Program A was given mostly to entry-level participants (800 of 1000)
- Program B was given mostly to experienced participants (600 of 1000)
- Experience level acts as a **confounding variable**

The Lesson

The "correct" answer depends on the causal question:

Question	Correct Analysis
"Which program was *associated with* better outcomes?"	Aggregate (B looks better)
"Which *causes* better outcomes?"	Stratified (A is better)

4. The do-Operator and Backdoor Adjustment

Graph Surgery Concept

- **do(T = A)** = Assign everyone to Program A, regardless of confounders
- Cuts incoming arrows to the treatment variable
- Removes confounding by breaking the backdoor path

Backdoor Adjustment Formula

$$P(Y|do(T = A)) = \Sigma\_e\ P(Y|T = A, E = e) \cdot P(E = e)$$

Components:

- P(Y | T=A, E=e): Effect within each stratum (from stratified data)
- P(E = e): Target population distribution (from overall data)
- Σ_e: Weighted average (marginalization)

Intuition: Within each confounder level, there's no confounding. Compute effect per stratum, then average over the population.

Assumptions Required:

1. No unmeasured confounding
2. Positivity (each stratum has both treatment conditions)

---

5. Real-World Example: UC Berkeley Admissions (1973)

Aggregate data:

- Male Applicants: 44% admission rate
- Female Applicants: 35% admission rate
- Suggested gender discrimination

Department-level analysis revealed:

- Women had equal or higher admission rates in most departments
- Women disproportionately applied to highly competitive majors
- Men disproportionately applied to less competitive departments

The confounder: Department choice affects both acceptance probability and application distribution.

---

6. Why Target Population Matters

The backdoor adjustment formula uses P(E = e) — but which population's distribution?

Same per-stratum effect, different target populations: | Population | % Experienced | % Entry | P(Y|do(T=A)) | |---|---|---|---| | Region 1 (tech hub) | 90% | 10% | 0.75 | | Region 2 (rural) | 10% | 90% | 0.35 | | Overall population | 40% | 60% | 0.50 |

Key insight: You must specify "Effect for whom?"

---

7. Counterfactuals: The Three-Step Process (Level 3)

For questions like: "Person did Program A and got employed. Would they have been employed without the program?"

Step 1: Abduction

Infer latent factors U from evidence (person did Program A, got employed)

Step 2: Action

Apply do(T = None) — surgery on the graph

Step 3: Prediction

With their specific U, compute outcome in the new world

Requires: Full Structural Causal Model (SCM) with functional equations, not just DAG

---

8. LLM Failure Patterns (Critical for Vignette Design)

Failure Pattern 1: Sensitivity to Wording

- Small wording changes can change outcomes
- Redacting key causal trigger words (e.g., "changing", "causes") strongly affects accuracy
- Even minor word changes can hurt accuracy

Implication for vignette creation: Test robustness to paraphrase

Failure Pattern 2: Semantic Cues Override Data

- When labels carry strong connotations, models may follow semantics rather than evidence
- LLMs can pick answers aligned with label meaning even when data supports opposite

Implication: Design cases where semantic intuition conflicts with data

Failure Pattern 3: No Grounded Intervention Mechanism

- Correct answers on famous examples don't imply reliable intervention computation
- Models rely on non-causal textual signals and can ignore actual data

Implication: Include novel scenarios not in training data

Failure Pattern 4: Simple, Unpredictable Mistakes

- Even with high average accuracy, LLMs make simple mistakes on specific inputs
- Inconsistency in applying causal criteria (which principle is relevant?)

Implication: Include cases that test consistent application of principles

---

9. Domain-Specific Signature Traps

Based on assignment groups, relevant traps include:

Indication Bias (Medicine - Groups A1, A2)

- Treatment assigned based on condition severity

- Sicker patients get more aggressive treatment → appear to have worse outcomes

Equilibrium Effects (Economics - Groups B1, B2)

- Interventions change the system equilibrium
- Initial effect differs from long-term effect after market adjusts

Attribution & Preemption (Law/Ethics - Groups C1, C2)

- Multiple sufficient causes
- Determining which cause is responsible when any could have produced the outcome

Outcome Bias (Sports - Groups D1, D2)

- Judging decisions by their outcomes rather than the information available at decision time

Regression to Mean (Daily Life - Groups E1, E2)

- Extreme observations tend to be followed by less extreme ones
- Often mistaken for treatment effects

Survivorship Bias (History - Groups F1, F2)

- Only observing successful cases that "survived" a selection process
- Missing data on failures

Self-Fulfilling Prophecies (Markets - Groups G1, G2)

- Predictions that cause themselves to become true
- Feedback from prediction to outcome

Feedback Loops (Environment - Groups H1, H2)

- Bidirectional causation over time
- Effect becomes cause in subsequent periods

Goodhart's Law (AI & Tech - Groups I1, I2)

- "When a measure becomes a target, it ceases to be a good measure"
- Optimization pressure corrupts the metric

Simpson's Paradox (Social Science - Groups J1, J2)

- Aggregate trends reverse when stratified by confounders
- Detailed in Section 3 above

---

10. What LLMs CAN and CANNOT Do for Causal Reasoning

What LLMs CAN Do:

- Suggest candidate variables from domain knowledge
- Propose edges based on known relationships in training data
- Generate hypotheses (e.g., "Age might confound the relationship")
- Help draft graphs and causal context from natural language

What LLMs CANNOT Do:

- Validate causal direction from data alone (requires assumptions or interventions)
- Guarantee complete confounders (may propose plausible ones, but not exhaustive)
- Distinguish correlation from causation (fundamentally observational training)

Hybrid Approach (Roadmap):

1. **LLM:** Generate candidate variables and initial graph structure
2. **Causal Discovery Algorithms:** Test conditional independencies in data
3. **Human Expert:** Validate, add domain constraints, resolve conflicts
4. **Do-Calculus:** Compute P(Y|do(X)) from final graph

---

11. DAG Constraints

Forbidden: Cycles

- Violates "Acyclic" in DAG
- Real feedback loops exist (e.g., poverty ↔ health)

Solution: Time-Unrolling

- Model $X_t \rightarrow Y_t \rightarrow X_{t+1} \rightarrow Y_{t+1}$
- Feedback unrolled over time becomes acyclic

Other Approaches for Cycles:

- Equilibrium / steady-state models
- Dynamic Bayesian networks
- Structural equation models (SEM)

---

12. Benchmark Quality Criteria

When expanding the benchmark, ensure vignettes:

1. **Have clear causal structure** — identifiable confounders, mediators, colliders
2. **Test specific Pearl levels** — L1, L2, or L3 reasoning
3. **Include signature traps** — domain-specific pitfalls
4. **Have unambiguous correct answers** — based on causal reasoning, not intuition
5. **Vary in difficulty** — easy to challenging within each level
6. **Are novel** — not likely in LLM training data
7. **Test robustness** — similar scenarios with different surface features

---

13. Case Structure Requirements (from Assignment)

Each case must include:

- **Scenario:** Clear description of situation
- **Variables:** Key variables with roles (Treatment, Outcome, Confounder, etc.)
- **Annotations:**
  - Case ID
  - Pearl Level (L1, L2, L3)
  - Domain
  - Trap Type
  - Trap Subtype (if applicable)
  - Difficulty level
  - Subdomain
  - Causal Structure
  - Key Insight
- **Hidden Timestamp:** Question revealing temporal/causal ordering
- **Conditional Answers:** "Answer if..." sections for different scenarios
- **Wise Refusal:** Response identifying missing information or potential biases

---

14. Key Formulas and Notation

Probability Notation:

- P(Y|X) — Conditional probability (observational, Level 1)
- P(Y|do(X)) — Interventional probability (Level 2)
- P(y_x|x', y') — Counterfactual probability (Level 3)

Backdoor Criterion:

Adjust for variables that block ALL backdoor paths from treatment to outcome.

Backdoor Adjustment:

$$P(Y|do(T)) = \Sigma_z P(Y|T, Z=z) \cdot P(Z=z)$$

---

15. Summary: The do-Calculus Recipe

1. **Draw the causal graph** — From domain knowledge (what causes what?)
2. **Identify backdoor paths** — Paths from T to Y starting with arrow into T
3. **Find adjustment set** — Variables that block ALL backdoor paths
4. **Apply the formula** — $P(Y|do(T)) = \Sigma_z P(Y|T, Z=z) \cdot P(Z=z)$
5. **Interpret** — This is the causal effect (what happens if we intervene)

No causal graph → Don't know what to adjust for → Cannot resolve paradoxes **With causal graph** → Backdoor criterion → Principled causal inference

---

16. Discussion Questions from Lecture (Good for Vignette Ideas)

1. Can you tell a participant "You have 40% chance of employment if you enroll"?
   - **Answer:** No — this assumes no confounding (C is correct: depends on study design)
2. If correlation is 100%, can we be certain the program works?
   - **Answer:** No — could still be perfect confounding
3. What if there's an unknown confounder?
   - **Answer:** Causal estimate could be completely wrong
4. How can we EVER be confident about causation?
   - **Answer:** All of: RCTs, explicit assumptions + sensitivity analysis, multiple converging evidence

---

Quick Reference: Confounder vs Mediator vs Collider

Type	Structure	Adjustment Rule	Example
Confounder	Z → X, Z → Y	ADJUST (removes bias)	Experience → Program choice, Experience → Employment
Mediator	X → Z → Y	DON'T ADJUST (blocks effect)	Program → Skills → Employment

Type	Structure	Adjustment Rule	Example
Collider	X → Z ← Y	DON'T ADJUST (creates bias)	Program → Interviewed ← Ability