# ExDis: Causal Explanations for Disparate Trends

Tal Blau[1], Brit Youngmann[2], Anna Fariha[3], Yuval Moskovitch[1]

[1] Ben-Gurion University of the Negev
[2] Technion - Israel Institute of Technology
[3] University of Utah

## ① Motivation

- **Data Trends Drive High-Stakes Decisions**
  - Directly impacts policy design and societal equity.
- **The Core Challenges: Where and Why?**
  - **Where:** Disparities are often hidden within specific subpopulations.
  - **Why:** Surface correlations fail to reveal the true underlying causes.
- **Manual Search is Infeasible**
  - Finding localized gaps in high-dimensional data is like finding a needle in a haystack.
  - Crucial trends can easily be missed or completely reversed.

**The Goal**

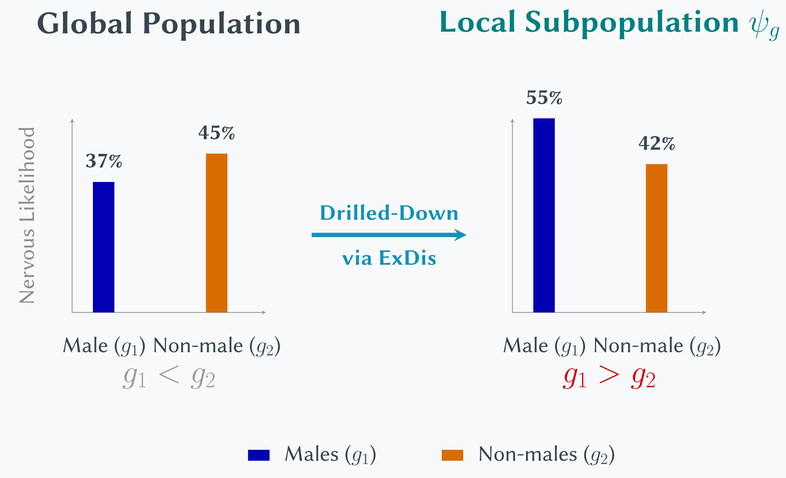
- Automate the discovery of **causal, actionable, and diverse** explanations across data regions.

## ② Example

- **MEPS Mental Health Study:** Looking at severe nervousness across different subpopulations.
- **Overall Averages:** Non-males ($g_2$) report higher rates ($45\%$) than Males ($g_1$) ($37\%$).
  - Basic policies might just target $g_2$ based on this broad trend. *But does this match the actual risk for specific people?*

**Key insight**

- **The Problem with Averages:** Looking only at the big picture hides real differences in people's lives by treating everyone the same.




**Simpson's Paradox Uncovered:** Non-males show higher nervousness globally ($g_1 < g_2$), but the trend completely reverses ($g_1 > g_2$) within the low-income, uninsured subpopulation isolated by ExDis.

- `ExDis` finds hidden groups where the overall trend flips completely.
- In this specific group, the numbers reverse: Males ($g_1$) jump to $55\%$, while Non-males ($g_2$) drop to $42\%$.
- To truly understand differences between groups, we must **find these hidden pockets** where the big trends break down.

## ③ Core Definitions & Problem Formulation

**1. Base Elements** :

- **Outcome Variable ($O \in A$):** The target aggregate metric monitored for disparities. Example: Total Compensation (TC).
- **Pattern ($\psi$):** A conjunction of equality predicates that isolates a specific subpopulation $\psi(D) \subseteq D$. Example: $\psi = (Ethnicity = Asian \wedge Role = Dataanalyst)$.
- **Groups of Interest ($g_1, g_2$):** disjoint patterns $(\psi_{g_1}, \psi_{g_2})$ whose aggregate outcomes are being contrasted. Example: $g_1 = DataAnalysts$ vs. $g_2 = Back-EndDevelopers$.

**2. Attribute Partitioning ($A \setminus \{O\} = I \cup M$)** two disjoint sets:

- **Immutable Attributes ($I$):** define **where** $\psi_g$. Example: $Age$.
- **Mutable Attributes ($M$):** define **why** $\psi_e$. Example: $EducationLevel$.

**3. Disparity Explanation ($\phi$)** A pair of patterns $(\psi_g, \psi_e)$ where $\psi_g$ over I — the subpopulation where g1 and g2 diverge. $\psi_e$ over M — the treatment that explains the gap inside.

**4. Disparity Score ($\Delta$)** :

$$\Delta(\phi) = \frac{|CATE_{G_D}(\psi_e, O \mid \psi_g \wedge \psi_{g_1}) - CATE_{G_D}(\psi_e, O \mid \psi_g \wedge \psi_{g_2})|}{\max|O|}$$

**5. Support** :

$$support(\phi) = \frac{|\psi_{g \wedge g_1}(D) \cup \psi_{g \wedge g_2}(D)|}{|D|}$$

**6. Diversity (Jaccard Similarity)** :

$$SIM(\phi, \phi') = \frac{|\psi_g(D_{g_1 \cup g_2}) \cap \psi_{g'}(D_{g_1 \cup g_2})|}{|\psi_g(D_{g_1 \cup g_2}) \cup \psi_{g'}(D_{g_1 \cup g_2})|}$$

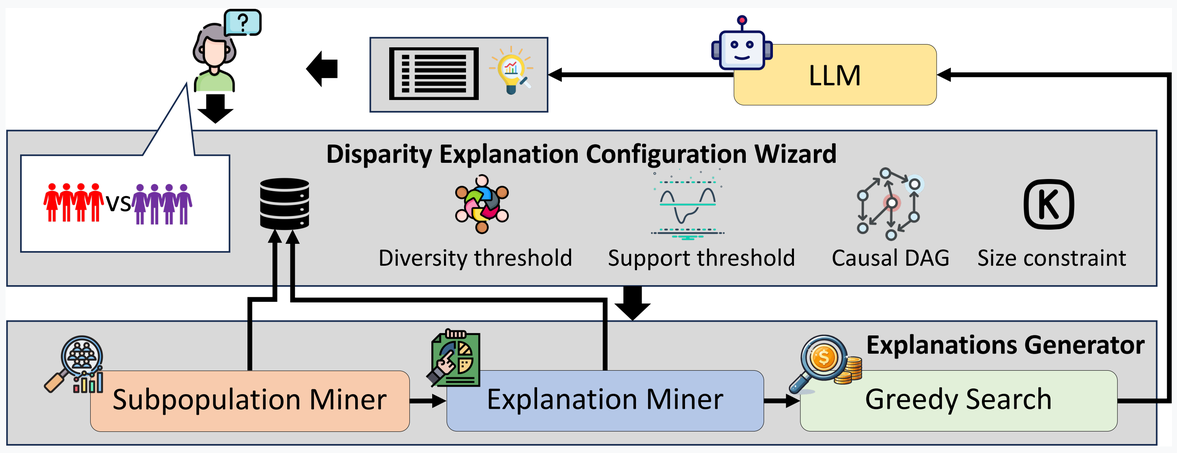
**Problem: Disparity Explanation Selection**

Given a database instance $D$, a causal model $G_D$, an outcome $O$, two groups of interest $g_1$ and $g_2$, a fixed budget $k \in N^+$, a support threshold $\sigma$, and a similarity threshold $\tau$, select an optimal explanation set $\Phi \subseteq \{\phi_1, \ldots, \phi_l\}$ such that:

1. **Size Constraint:** The size of the explanation set is bounded by the budget ($|\Phi| \leq k$).
2. **Support Constraint:** Each explanation covers a representative fraction of data ($\forall \phi_i \in \Phi : support(\phi_i) \geq \sigma$).
3. **Diversity Constraint:** Pairwise subpopulation overlap is strictly bounded to eliminate redundancy ($\forall \phi_i, \phi_j \in \Phi : SIM(\phi_i, \phi_j) \leq \tau$).
4. **Objective Function:** Maximize the collective causal explanation strength:

$$Maximize \Delta(\Phi) = \sum_{\phi \in \Phi} \Delta(\phi)$$

## ④ System Architecture & Overview




The *ExDis* framework operates via a unified pipeline splitting user configuration from localized heuristic execution:

**1. Configuration Wizard (Problem Formulation)** :

- **Cohorts:** Users select an outcome and target groups (can be overlapping).
- **Schema Split:** Explicitly flags features as mutable or immutable.
- **Causal DAG:** Built manually, uploaded via DOT files, or discovered algorithmically.

**2. Explanations Generator (Heuristic Engine)**
Executes a three-stage scalable generation process to avoid exponential search complexity :

- **Subpopulation Miner:** One Apriori pass over I attributes; keep only patterns with $support \geq \sigma$.
- **Explanation Miner:** For each candidate, lattice-search treatment patterns over M attributes.
- **Greedy Search Selector:** Hierarchical-cluster on subpopulation overlap; iteratively pick the highest-$\Delta$ candidate that respects $\tau$.

**3. Natural Language Translation** : Converts logic to plain English via an integrated LLM.

Feel free to reach out with any questions!
tbl@post.bgu.ac.il, brity@technion.ac.il,
afariha@cs.utah.edu, yuvalmos@bgu.ac.il

*See the demo video here:*