Position: Data-driven Discovery with Large Generative Models

Bodhisattwa Prasad Majumder*, Harshit Surana*, Dhruv Agarwal*,
Sanchaita Hazra, Ashish Sabharwal, Peter Clark

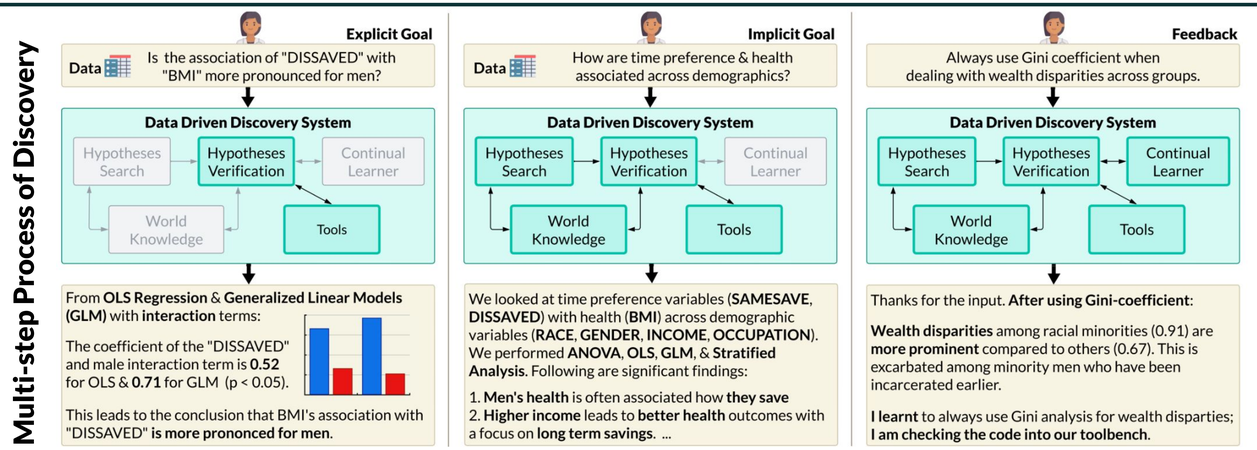
Data-driven Discovery: Following Newell & Simon (1976), we define a **heuristic search problem** that aims to describe a given set of observations by uncovering the laws that govern its data-generating process. E.g., "under context c, variables v have relationship r"

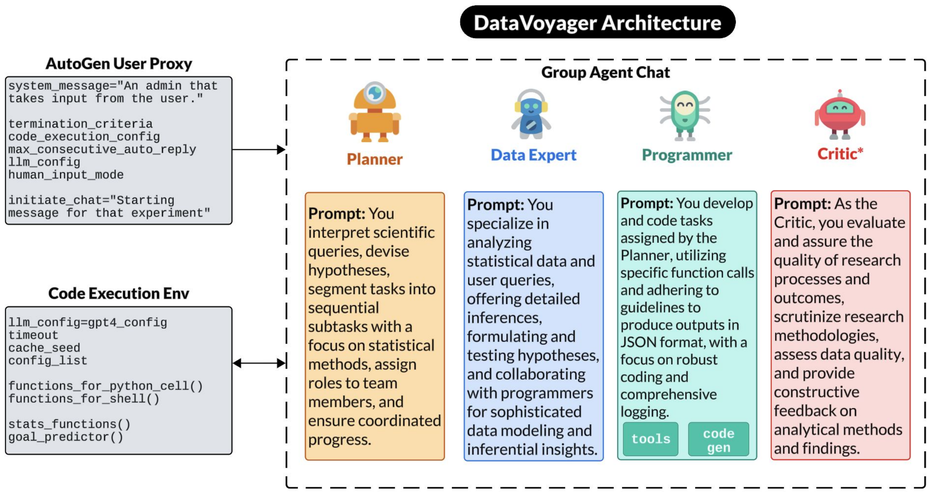
Why Data-driven Discovery:
1. Abundance of large-scale datasets that would benefit highly from automated discovery
2. Practicality of automated, inexpensive verification enabled by data without the need for additional data collection

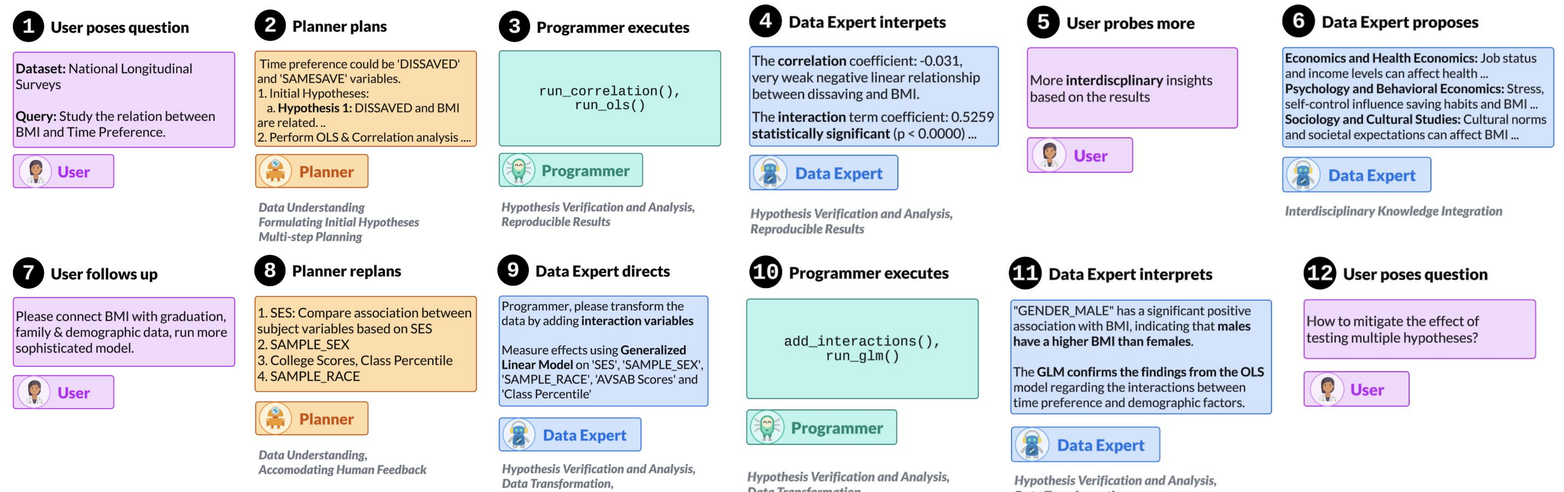
Why Large Generative Models (LGMs):
1. Previous works lacked the requisite computational power (Langley, 1984)
2. To harness pretrained domain and scientific knowledge for hypothesis search
3. Code generation and execution ability

LGMs present an incredible potential for automating discovery but LGMs are not *all* we need.

Multi-step Process of Discovery







Desiderata for Data-driven Discovery (+ anecdotal comparisons with existing frameworks and DataVoyager)

Comprehensive Data Understanding
Most frameworks (AutoML, WolframAlpha) have limited ability. LGMs can explore & understand context, if prompted explicitly.

Hypothesis Generation
Prev. work use heuristics, visualization, lit. retrieval for initial hypothesis search, though most fail to do iteratively where LGMs can do that in loop.

Planning Research Pathways
No frameworks including LGMs can consistently plan scientific workflows. LGMs may have scientific knowledge but cannot robustly apply it.

Hypothesis Verification
This is well-achieved by heuristics or free-form code generation. But explicit tool-calling w LGMS is required for long-tail domain analysis.

Accommodating Human Feedback
Systems must accommodate human feedback for better reasoning, update beliefs. Feedback sig. improves LGMs' exploration.

Reproducible & Robust Results
DV shows a POC for automated, reproducible experiments but opens up a novel case for explosion of false discoveries via *p*-hacking.

Limitations of Automatic Discovery

1. **Hallucinations,** memorization and superposition issues
2. **Costly** for high-throughput fields
3. **Propaganda-led dubious claims** created by bad actors, policy impacts
4. **Raises legal challenges** for intellectual property rights & authorship, liability in decision making

Fancy a benchmark?

discovery-bench