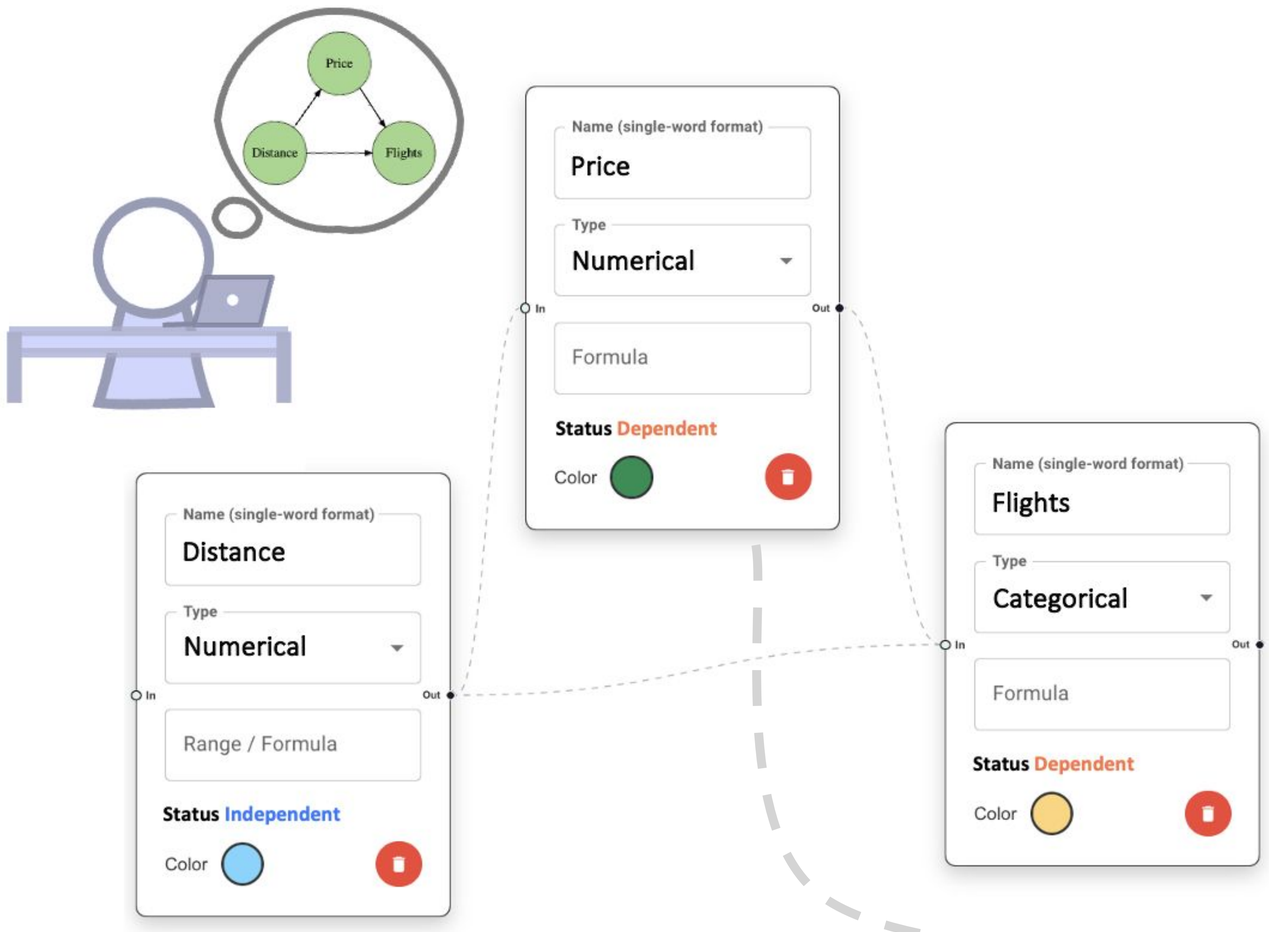
CausalSynth: An Interactive Web Application for Synthetic Dataset Generation and Visualization with User-Defined Causal Relationships

Zhehao Wang, Arran Zeyu Wang, David Borland**, and **David Gotz

University of North Carolina at Chapel Hill



Motivation

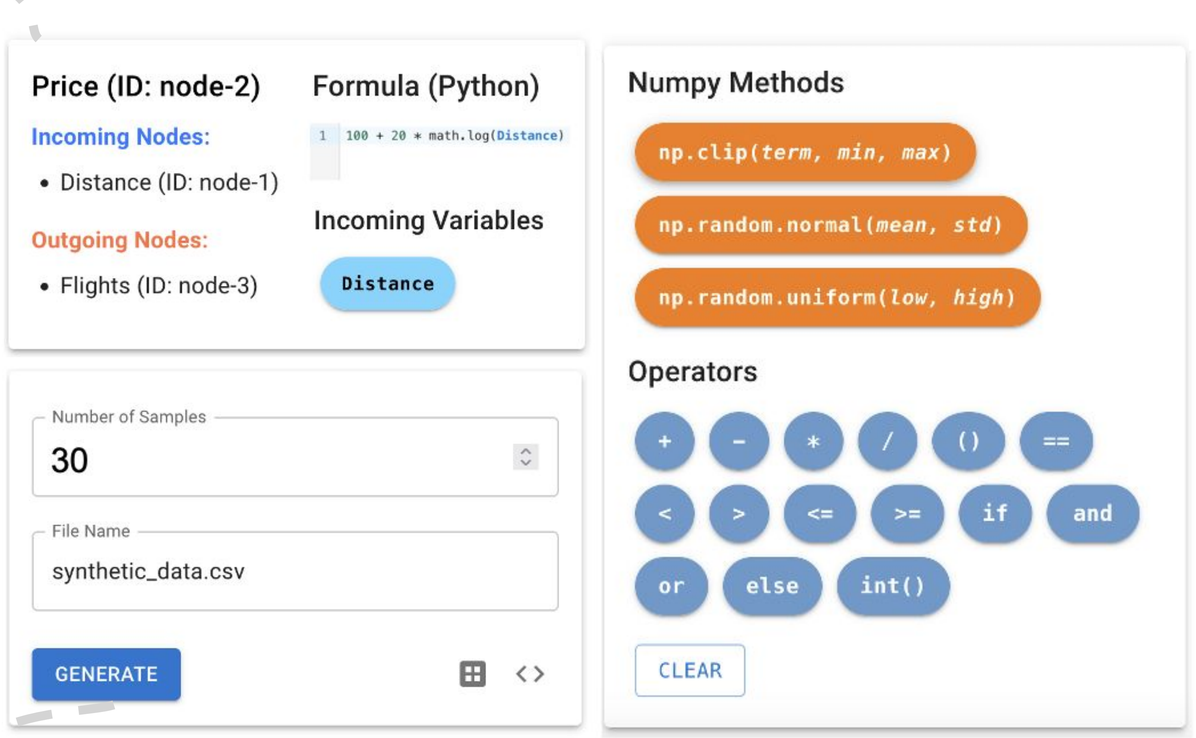
- **CausalSynth** generates and visualizes synthetic datasets based on *user-defined causal models* through a user-friendly web interface. It *addresses challenges in validating causal inferences* from observational data, which often *lacks clear ground truth*.

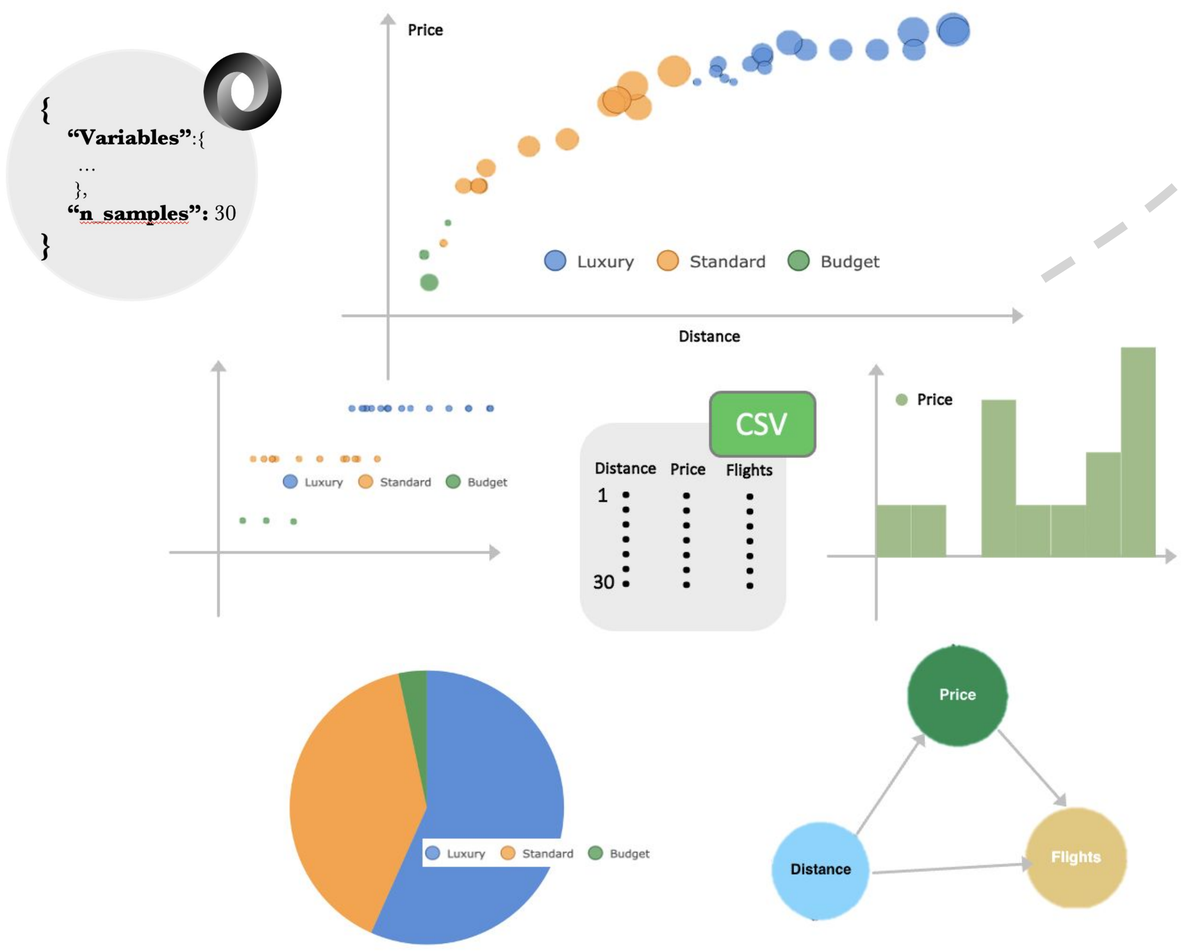
Getting Started

- Conceptualize a causal model for air flights
  - **Distance** affects **Price**
  - **Distance** and **Price** affect **Flights**
  - **Flights** has three categories:
    - *Budget*, *Standard*, and *Luxury*.
- Create an acyclic graph in **CausalSynth**
  - Connect the nodes based on the causal model
  - Optionally color-code variables

Model Definition & Data Generation

- **Specify Variable Interrelations**
  - On the right, **Price** is selected
    - Dependent on **Distance**
  - Input Python formula for **Price**
    - Incoming variables for **Price**, in this case **Distance**, shown as color-coded buttons
    - Virtual keyboard contains common NumPy methods and Python operators to facilitate sanitized input
  - Repeat similar process for remaining variables
- **Generate Samples**
  - Specify number of samples and filename
  - Click “Generate” button to create the dataset





Inspection / Visualization of Model & Data

- **Visualize the Data & Causal Model**
  - Review generated dataset with built-in CSV previewer
  - Various visualizations provided:
    - Bubble plots, scatter plots, histograms, pie charts, and acyclic graphs
  - In the example bubble plot:
    - Color represent **Flights** categories.
    - Size represents **Price** relative to counterparts of the same category.
    - Note a logarithmic relationship between **Price** and **Distance**, and that as **Price** increases, **Flights** tend to be more luxurious.
- **Export, Modify, and Re-import the Causal Model**
  - JSON file can be modified and re-imported to restore the model, ensuring traceability for users.