Applying Causality toolkit to Real-world datasets

Arun Gaonkar

Mentored by : Roger Dev

Rube Goldberg Machine

Contents

What is Causality?

Importance of Causality

Causality Toolkits

1. HPCC Causality
2. Because

Causal Analysis – steps

1. Dataset
2. Pre-processing
3. Analyze & Interpret
4. Causal Model
5. Verify

Causal Analysis

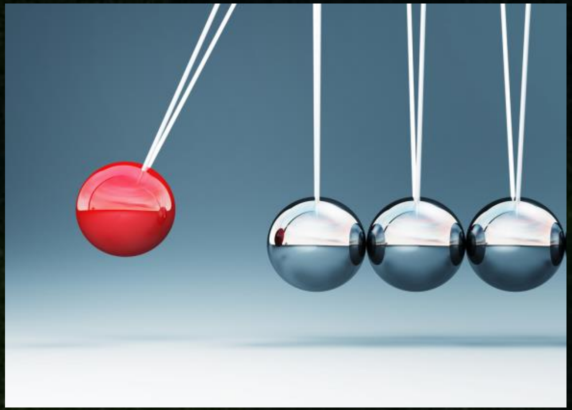
1. Causal Analysis on Synthetic Dataset
2. Causal Analysis on CDC-LLCP Dataset

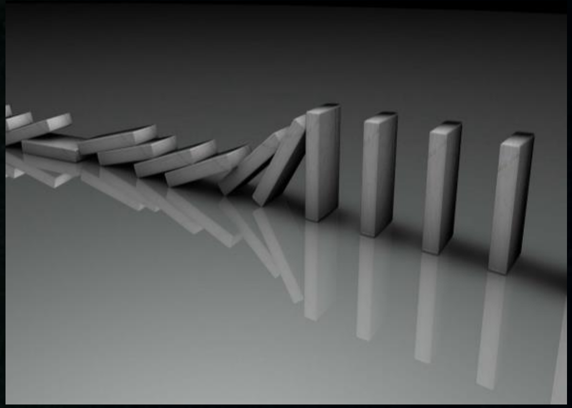
Verify

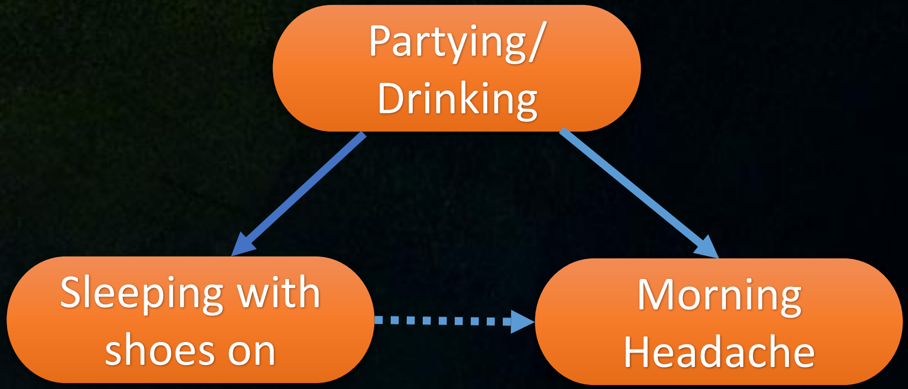
1. Verifying the Model
2. Verifying the Hypothesis

What is Causality?

- Cause – Responsible for effects
- Effect – Dependent on the Cause
- Is cause-effect relationship Universal?
- The Domino Effect
- Does Correlation imply Causation?

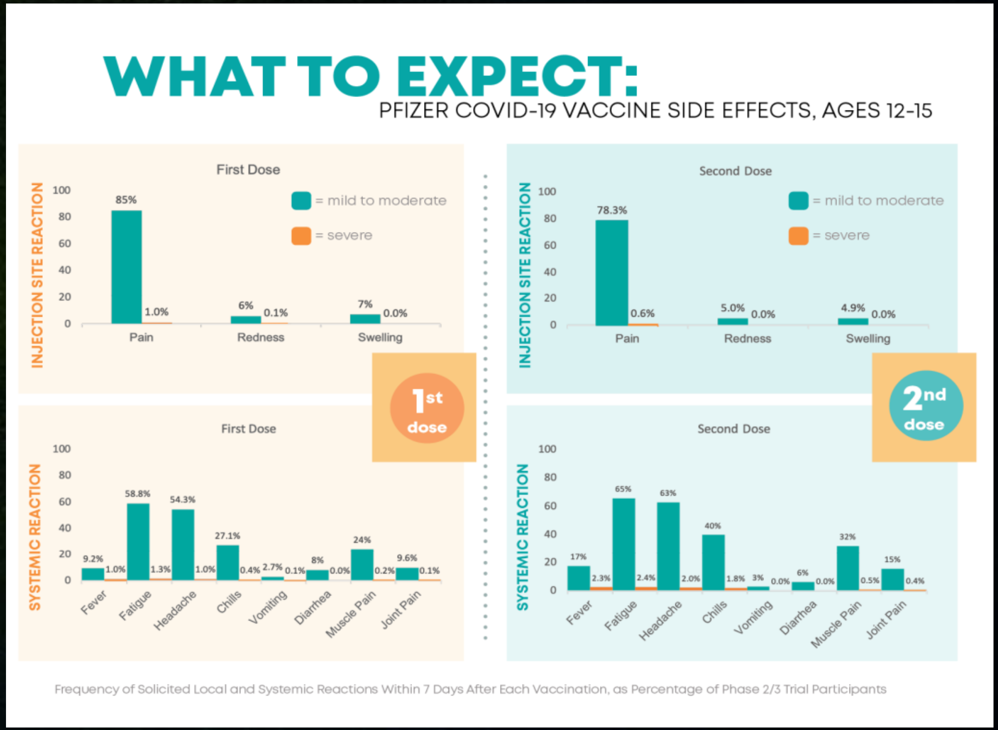






Importance of Causality

- Each person makes 35,000 conscious decisions everyday
- Helps in making Good decisions
- Time series analysis & Strategic Planning
- Medical Diagnostic analysis
- E.g.: COVID19 vaccination – Effects



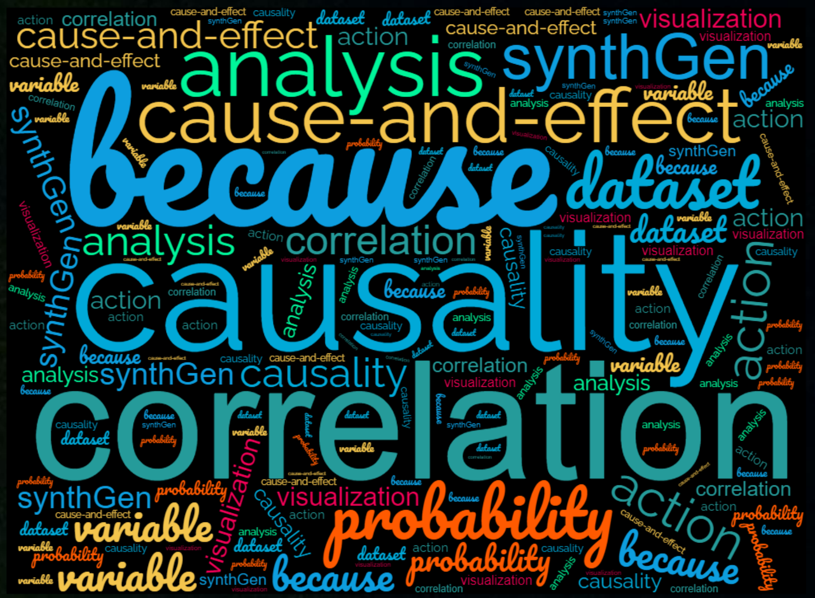
source : https://www.vdh.virginia.gov/richmond-city/covid-19-after-vaccination/

Causality Toolkits – HPCC_Causality Bundle

- Toolkit developed by HPCC systems for Causal Analysis
- Parallelized on HPCC clusters
- Built on top of ‘Because’ module
- ECL programming language to implement

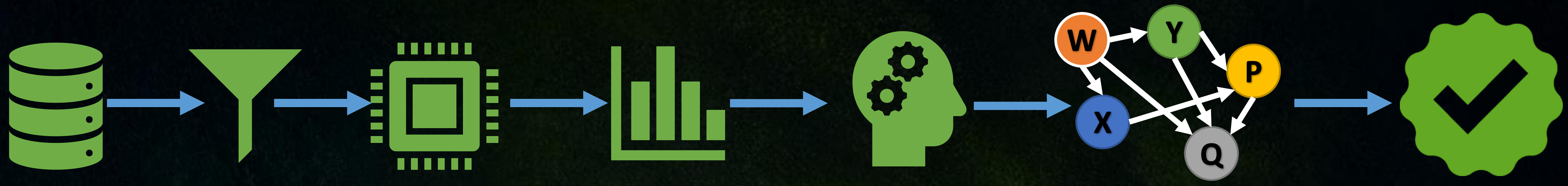
Causality Toolkits – Because

- HPCC Systems Causality Framework
- Python module for causal analysis
- Includes 4 sub-packages :
    1. Synthetic Data Generator – For synthetic data generation
    2. Probability – Statistical and Probabilistic analysis
    3. Visualization – For graphical representation and analysis
    4. Causality – For causal methods



Causal Analysis - Steps

Dataset	Pre-processing	Analyze	Interpret & Model	Verify
1. Find a dataset 2. Analyze the dataset 3. Cleaning preparing the dataset	1. Find interesting variables 2. Pre-processing the dataset	1. Find individual relationships between these variables 2. Analyze pairwise relations among these variables 3. Analyze these relations with conditionalization	1. Draw a causal model using from above observations 2. Draw a causal model with testDirection & Independence tests	1. Verify the causal model



Causal Analysis

on Synthetic Dataset

Dataset

- Synthetic Data Generation – Using HPCC_Causality Synth & Gen
- Based on Structural Equation Models (SEM)
- Generated Test Data for analysis

id	number	value
1	1	0.896642972931573
1	2	1.227285474143171
1	3	0.3716197319805742
1	4	-1.047257989481993
1	5	0.9470346507322929
1	6	0.534259202809647
1	7	-2.041073115309538
2	1	0.6408760999717116
2	2	-0.3660806211582063
2	3	0.6669649434036274
2	4	-2.924519924312789
2	5	0.6991835864092881
2	6	1.930282018205498
2	7	-4.925629162182375
3	1	-0.3548787600754421
3	2	1.395481996679391
3	3	-0.2872605014751051
3	4	1.170128088317976
3	5	-0.1271693932238411
3	6	-2.653658076468372
3	7	2.860091776444442
4	1	0.4120367667014634
4	2	-1.288768856928463
4	3	-0.6082476435350309
4	4	-0.4237578960614893
4	5	-0.7414302774501271

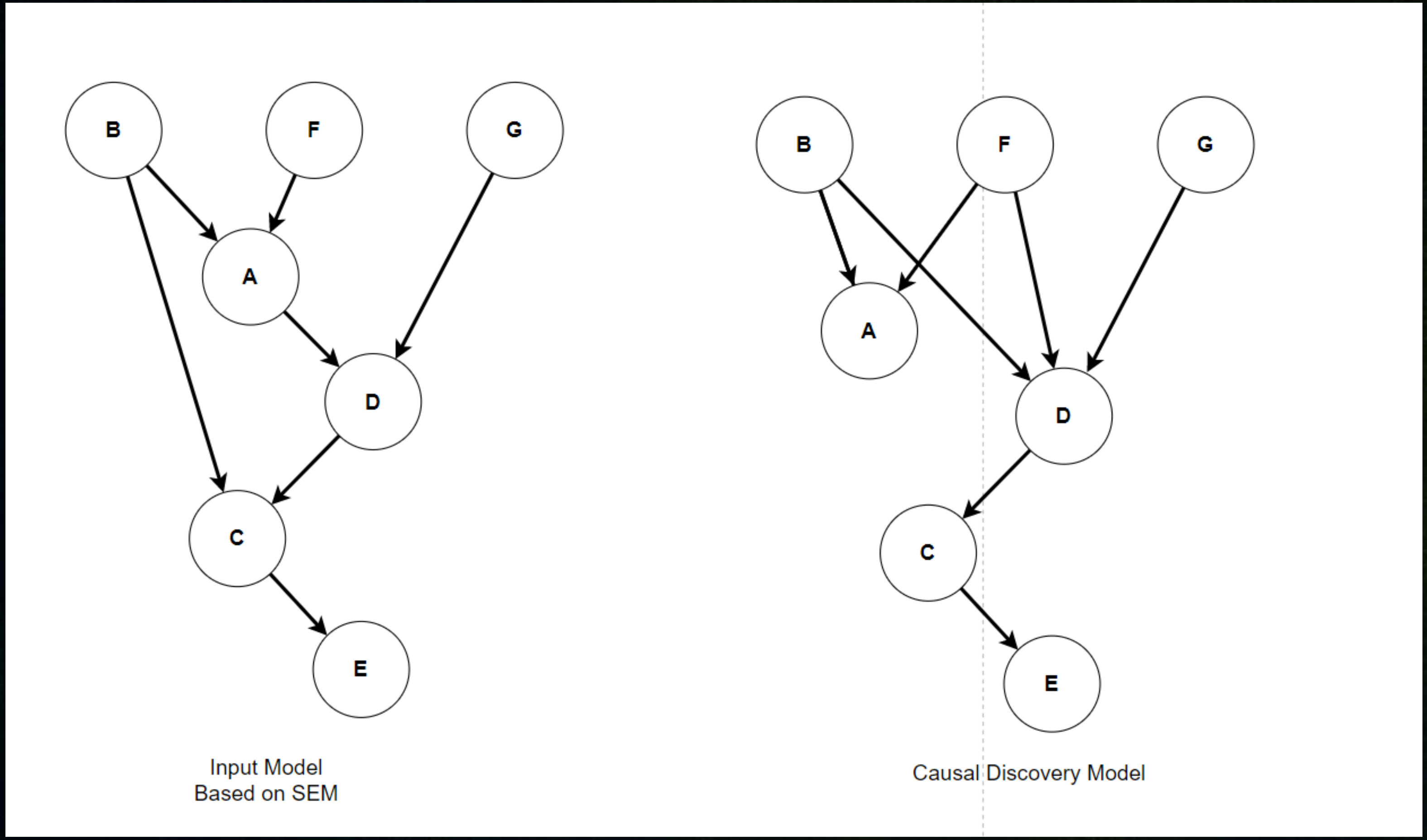
```
semRow := ROW({
    [],
    ['A', 'B', 'C', 'D', 'E', 'F', 'G'], // Variable names
    // Equations
    ['B = logistic(0,1)',
    'F = logistic(0,1)',
    'G = logistic(0,1)',
    'A = (B + F) / 2.0 + logistic(0,.4)',
    'D = (A + G) / 2.0 + logistic(0,.4)',
    'C = (B + A + D) / 3.0 + logistic(0,.4)',
    'E = C + logistic(0,.4)'
    ]}, SEM);

mySEM := DATASET([semRow], SEM);

testDat := HPCC_Causality.Synth(mySEM).Generate(nTestRecs);

// Note: The order of variables in the model much match the order of varNames in the SEM.
RVs := DATASET([
                {'A', ['B','F']},
                {'B', []},
                {'C', ['B', 'A', 'D']},
                {'D', ['A','G']},
                {'E', ['C']},
                {'F', []},
                {'G', []}
                ], Types.RV);
mod := DATASET([{'M8', RVs}], Types.cModel);
```

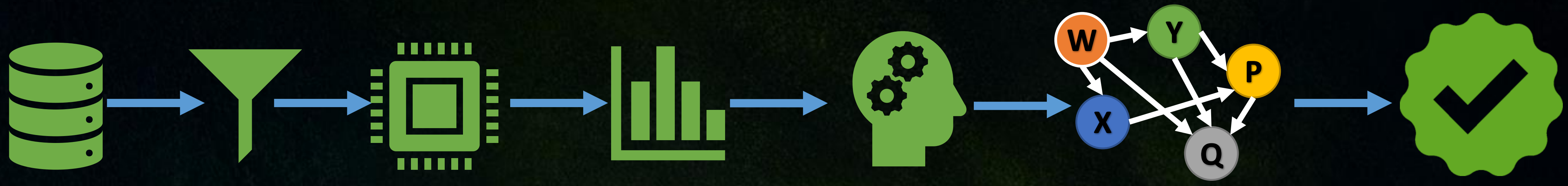
Causal Model & Discovery



Model Validation & Result

Error (Type 3 -- Incorrect Causal Direction) between A and C. Direction appears to be reversed.. Rho = -0.0005730759083272894
Error (Type 3 -- Incorrect Causal Direction) between D and C. Direction appears to be reversed.. Rho = -0.0013527315938702844
Warning (Type 3 -- Incorrect Causal Direction) between C and E. True direction could not be verified.. Rho = -3.0427271323660705e-05

Model Confidence is 93.75%



Causal Analysis

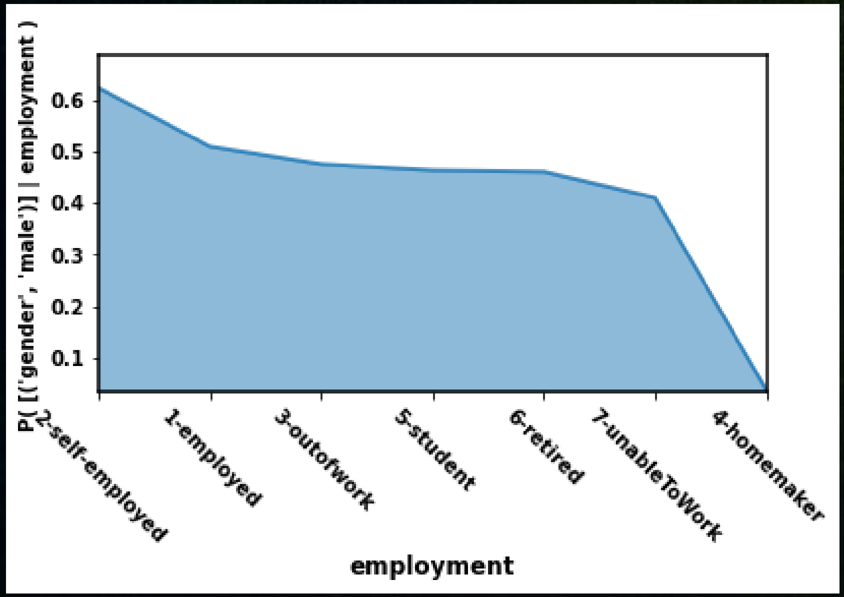
on CDC-LLCP Dataset

Dataset & Preprocessing

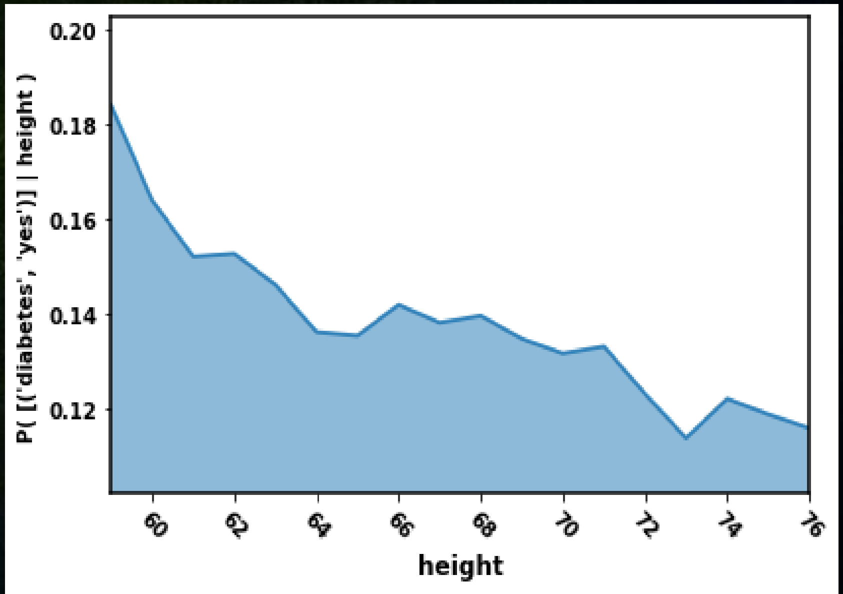
- Survey from CDC’s Population Health Surveillance Branch
- Questionnaire to collect uniform state-specific data on health, disease, disability, care & cause in US
- Dataset contains 401,958 records for 279 different types of questions
- After filtering and preprocessing, dataset now has 279,922 records and 31 different variables
- Using Python Because module

Analysis & Interpretation

- The correlation between Gender-Male and Employment

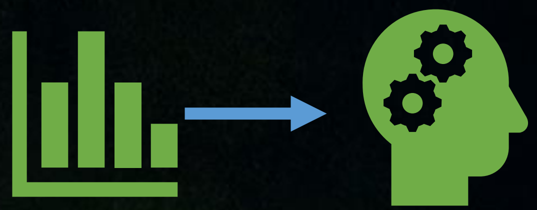


- The correlation between Height and Diabetes



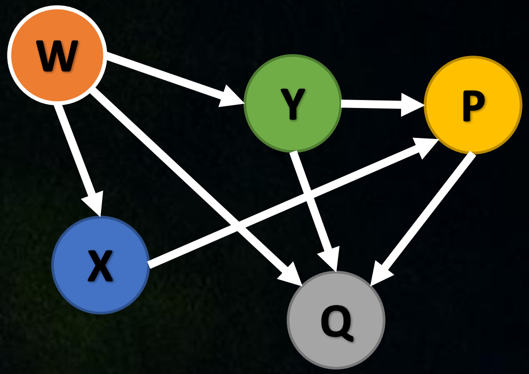
More analysis in Jupyter Notebook

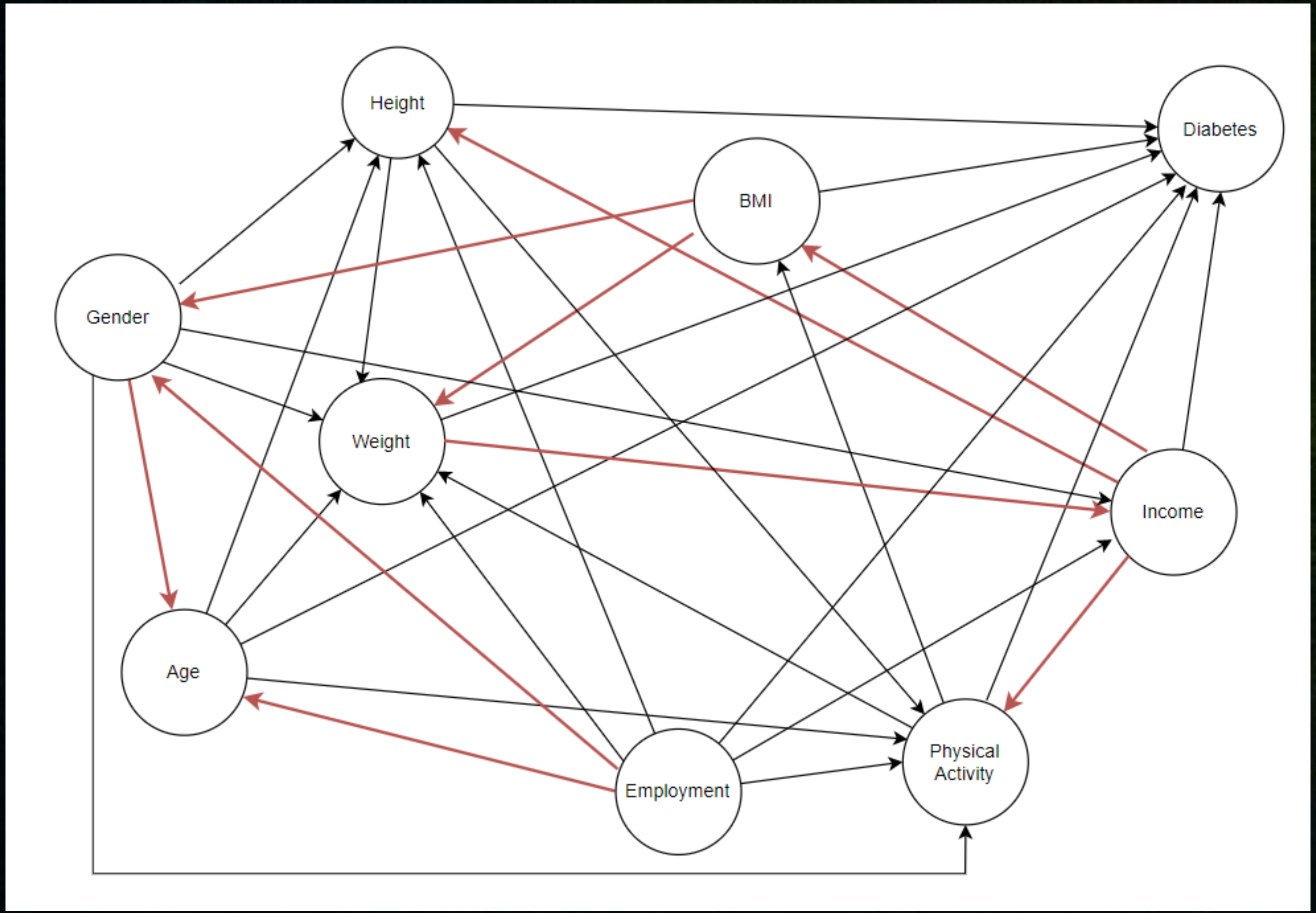
Analysis & Interpretation



More analysis in Jupyter Notebook

Causal Model





Verifying the Model

- Most of the relations are practically and analytically correct

- Some relations are unexpected. like,

  1. Gender causing Age
  2. Weight causing Income
  3. Income causing BMI
  4. Income causing Physical Activity

  Even though they are unexpected, a probable valid proof can be generated.

- There are some other relations, such as,

  1. BMI causing Gender
  2. Employment causing Gender
  3. Employment causing Age
  4. Income causing Height

  For which, any explanation is rationally invalid.

Verifying the Hypothesis

- What are all the factors that can influence the likelihood of a person having Diabetes?

1. Age – Direct
2. Gender – Indirect
3. Weight – Indirect
4. Height – Direct
5. BMI – Direct
6. Income – Direct
7. Employment – Direct
8. Physical Activity – Direct

Conclusion

- Causality toolkit can be applied to analyze real-world datasets
- Hidden variable and their cause-effects can be observed
- In Real-world, variables are inter-related, causing complexities
- Not 100% effective yet for complex datasets

Thank You

Special Thanks to:

Roger Dev

Lorraine Chapman

Zheyu Shen