What makes real datasets so hard to analyze?

1. Variations in data quality, format, provenance–trying to solve this!
2. Many datasets are not causal
3. Missing data
4. Irrelevant or deterministic variables
5. Mixtures of continuous and discrete variables
6. Mixtures of structures
7. Latent variables
8. Continuous variables usually not linear, Gaussian or even additive
9. Sample sizes can be too large or too small
10. Not always sparse
11. Not always i.i.d.
12. Faithfulness often fails
13. Selection bias
14. Feedback
15. Proxy variables
16. Figuring out ground truth can be painful

Algorithmic problems

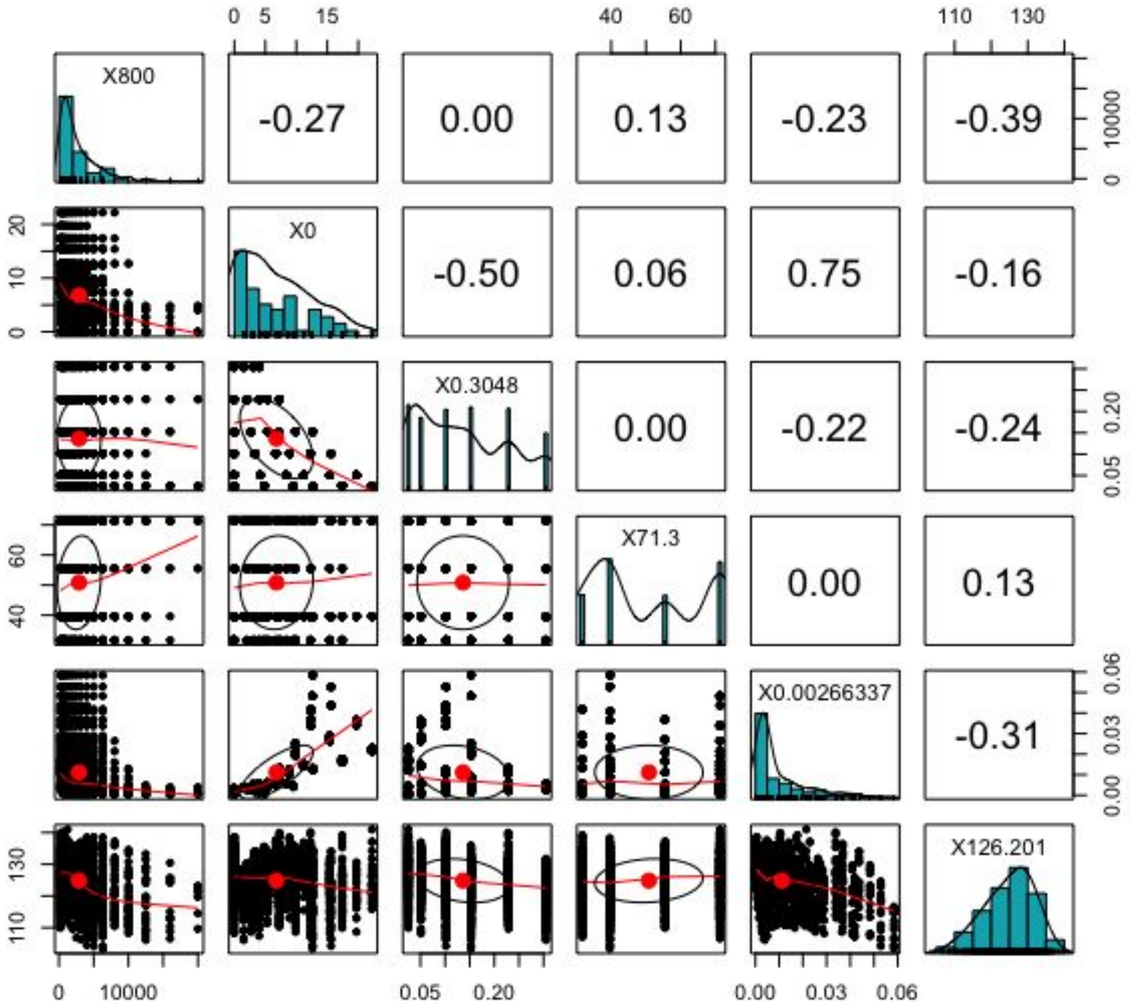
- There are many algorithms to try
- There are many parameter choices one could make
- In most cases, the “go to” algorithms don’t work very well
- Different algorithms can give radically different answers
- Sometimes the same algorithm can give different answers on different runs, due to order dependence, different bootstrap samples, algorithmic randomness, or simply different parameter settings

Airfoil Self-Noise (UCI)

Continuous, N = 1503

Attribute Information:

1. Frequency, in Hertzs.
2. Angle of attack, in degrees.
3. Chord length, in meters.
4. Free-stream velocity, in meters per second.
5. Suction side displacement thickness, in meters.
6. Scaled sound pressure level, in decibels.



Airfoil Self-Noise (UCI) - Some Ground Truth

This is a NASA wind tunnel experiment. We get ground truth from the description of the experiment

- **Tier 1 (forbid within): Chord, Attack, Velocity**
- **Tier 2: Frequency, Displacement**
- **Tier 3: Pressure**

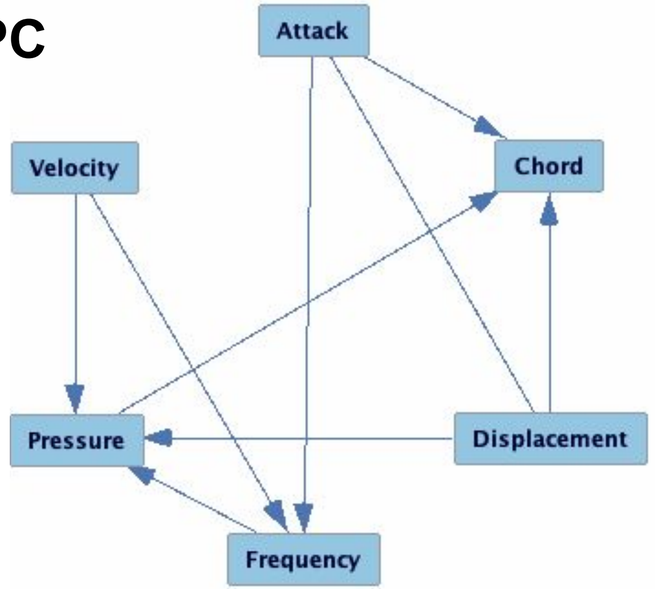
= Satisfies ground truth

- **Violates linear/Gaussian/additive**
- **Latent variables definitely**

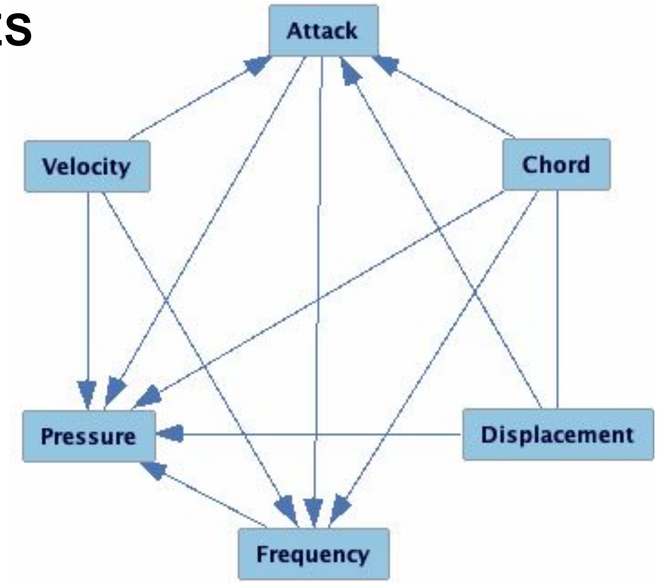
Note that for FCI:

- If an edge is green that means there is no latent confounder. Otherwise, there is possibly latent confounder.
- If an edge is **bold** (thickened) that means it is definitely direct. Otherwise, it is possibly direct.

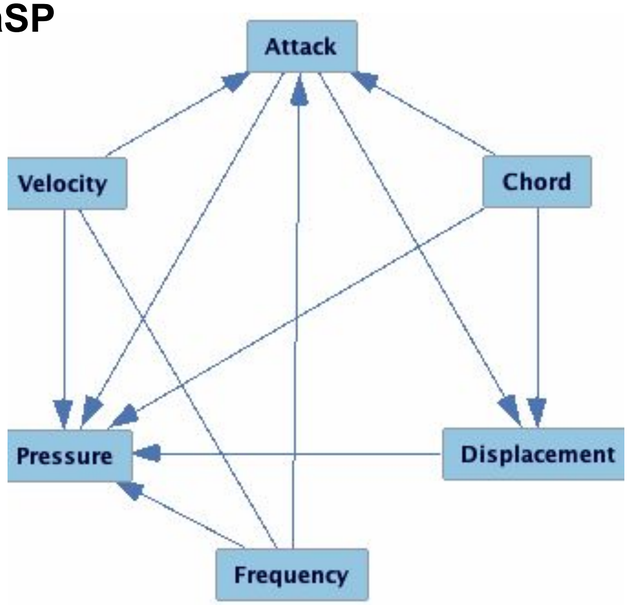
PC



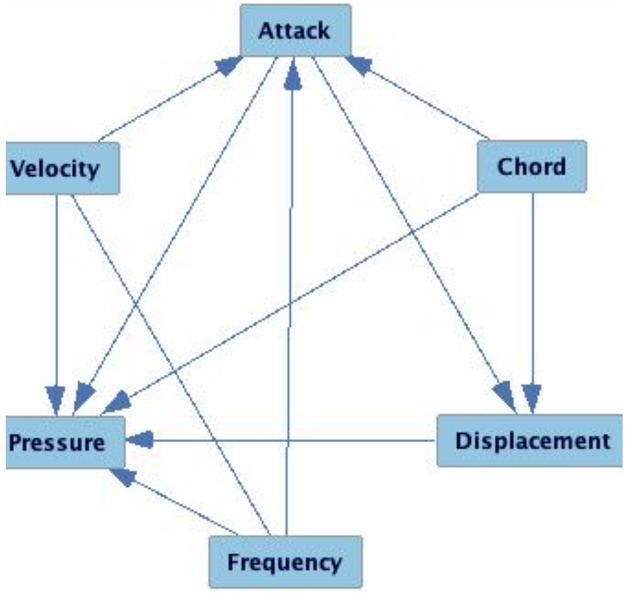
FGES



GRaSP



SP



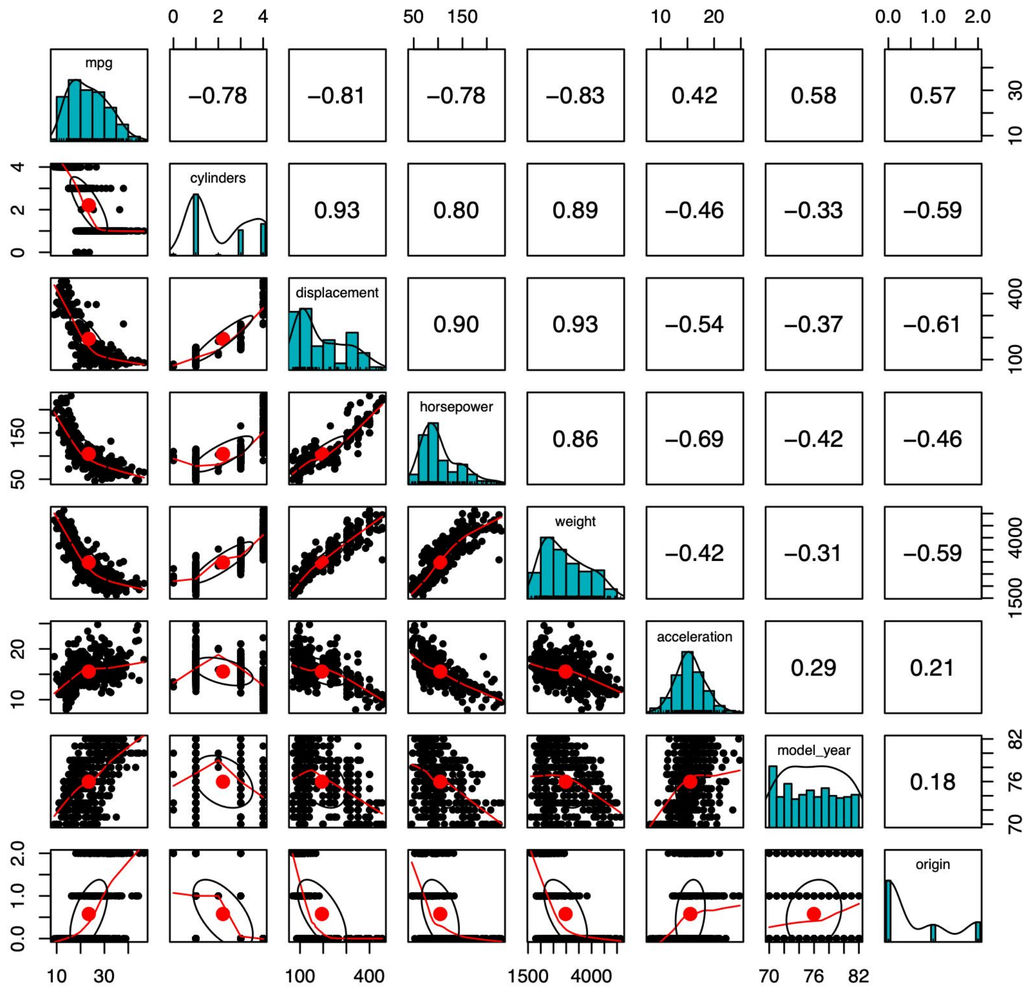
Auto MPG (UCI)

Mixed, N = 396
4 missing value rows removed

Attribute Information:

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete

-. car name: string (unique for each instance) —— REMOVED

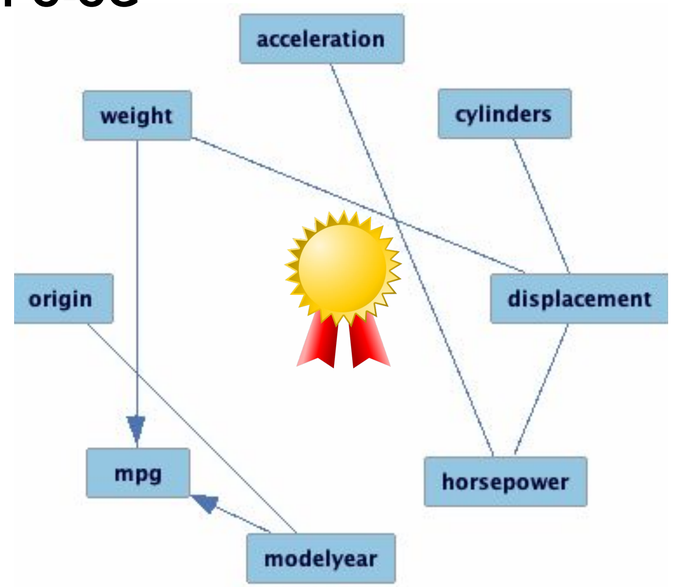


Auto MPG (UCI) - Some Ground Truth

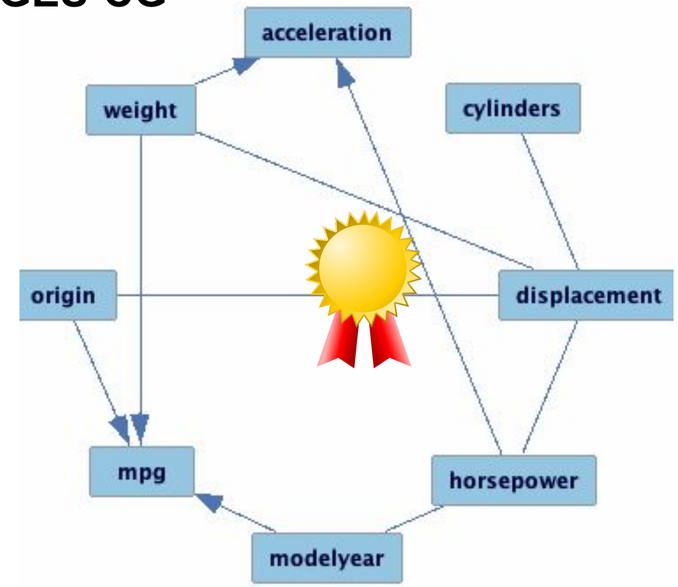
- **Tier 1 model_year, origin, weight**
- **Tier 2 cylinders**
- **Tier 3 displacement**
- **Tier 4 horsepower**
- **Tier 5 acceleration, mpg**

- **Violates linear/Gaussian/additive**
- **Latent variables**
- **Mixed continuous/discrete**

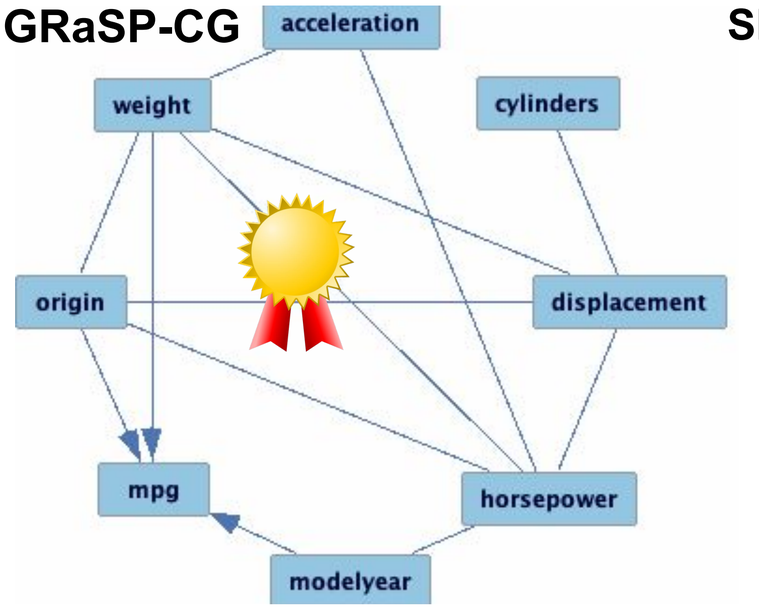
PC-CG



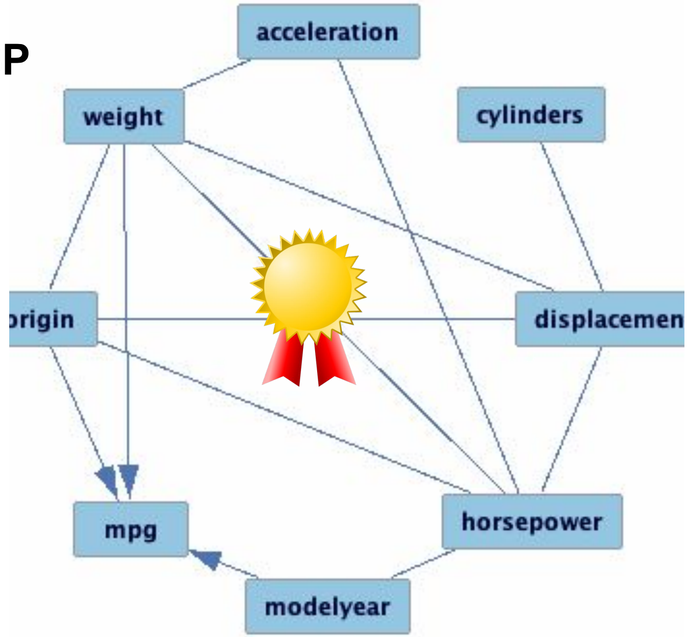
FGES-CG



GRaSP-CG



SP



Abalone (UCI)

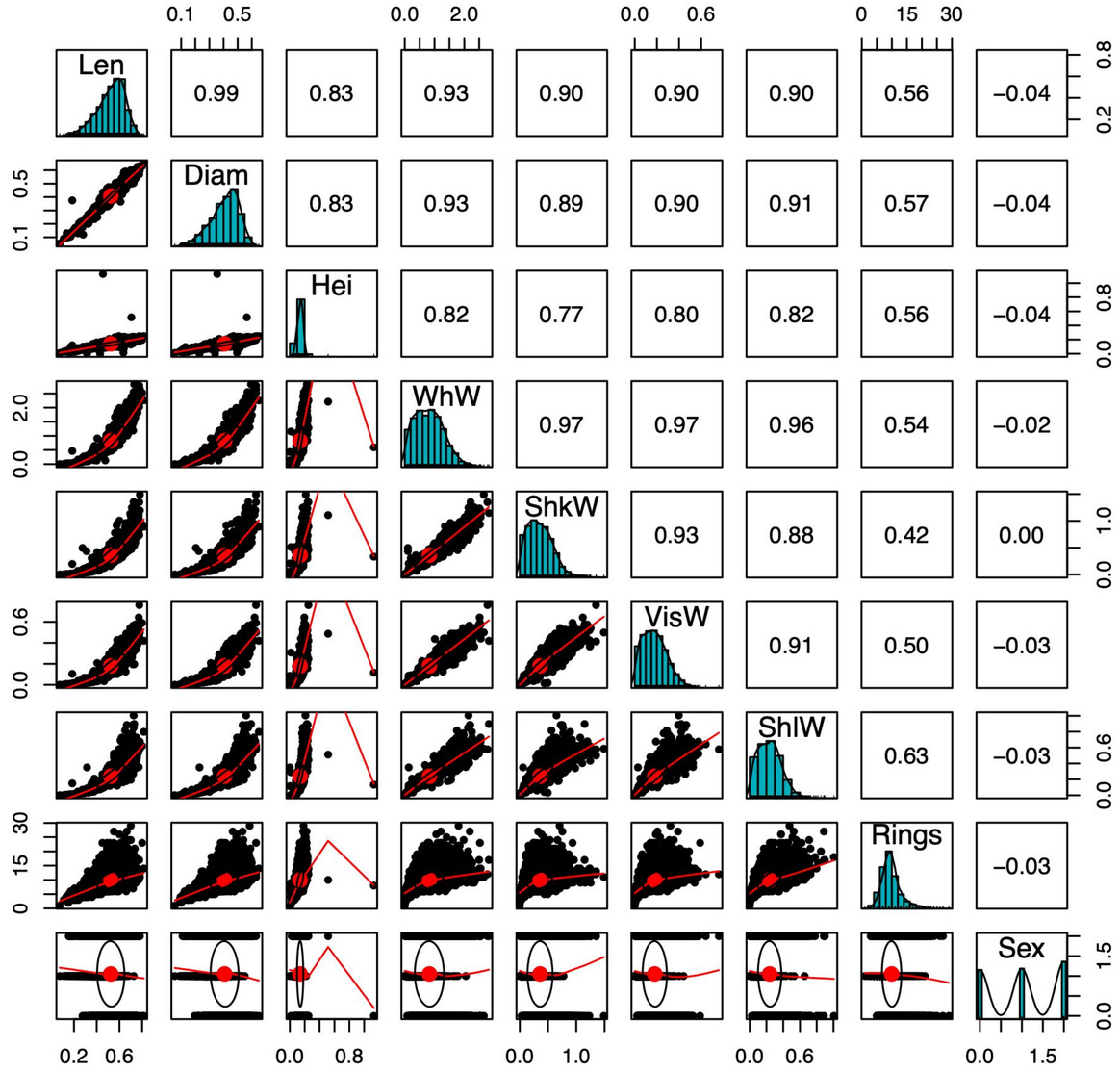
Mixed, N = 4177

Attribute Information:

Given is the attribute name, attribute type, the measurement unit and a brief description.

Name / Data Type / Measurement Unit / Description

1. **Sex / nominal / -- / M, F, and I (infant)**
2. **Length / continuous / mm / Longest shell measurement**
3. **Diameter / continuous / mm / perpendicular to length**
4. **Height / continuous / mm / with meat in shell**
5. **Whole weight / continuous / grams / whole abalone**
6. **Shucked weight / continuous / grams / weight of meat**
7. **Viscera weight / continuous / grams / gut weight (after bleeding)**
8. **Shell weight / continuous / grams / after being dried**
9. **Rings / integer / -- / +1.5 gives the age in years**



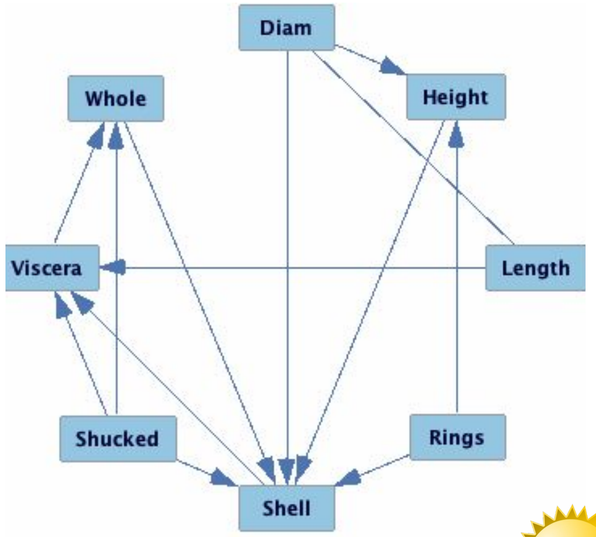
Abalone (UCI) - Some Ground Truth

- **Tier 1 Rings, Sex**
- **Tier 2 ShkW, ShlW, VisW**
- **Tier 3 WhW**
- **Tier 4 Diam, Hei, Len**

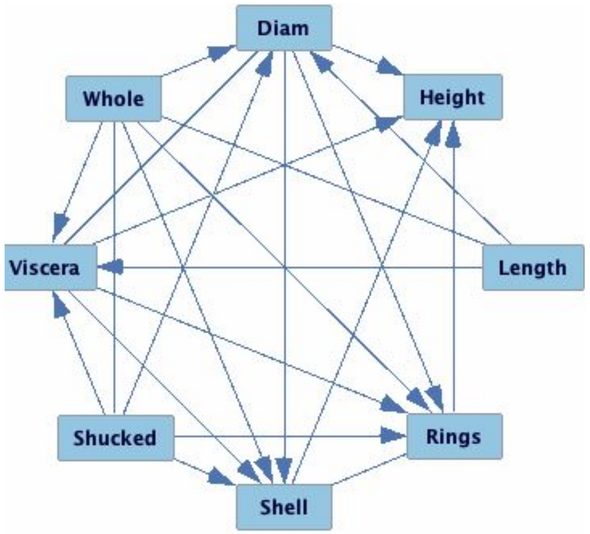
Note that here *Sex* is usually omitted from the dataset because it is discrete; the remaining variables are all continuous. But sex is clearly a causal variable, so for greater completeness, we analyze the data as *mixed*.

- **Violates linear/Gaussaian/additive**
- **Rings is *multiplicative***
- **I'm leaving out a number of *pairwise analyses of Rings with other variables, which are relevant***

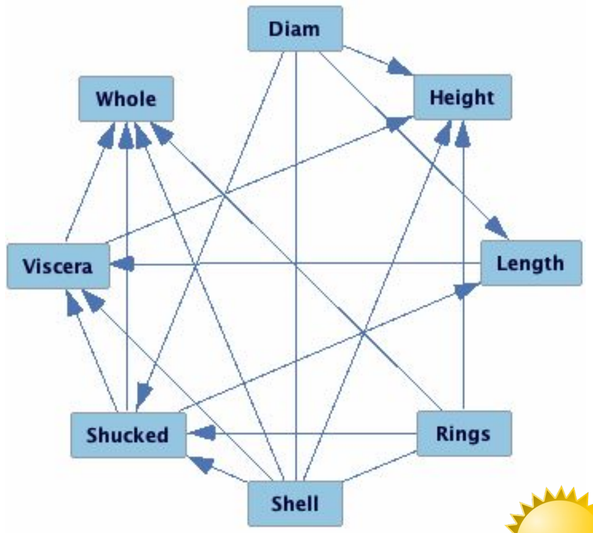
PC-CG



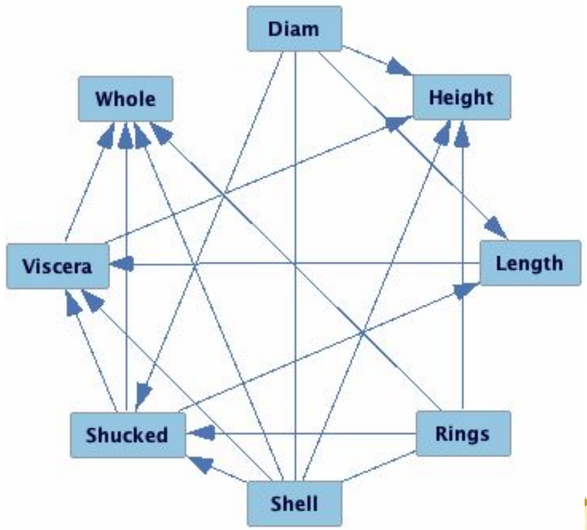
FGES-CG



GRaSP



SP



Wine Quality, Red (UCI)

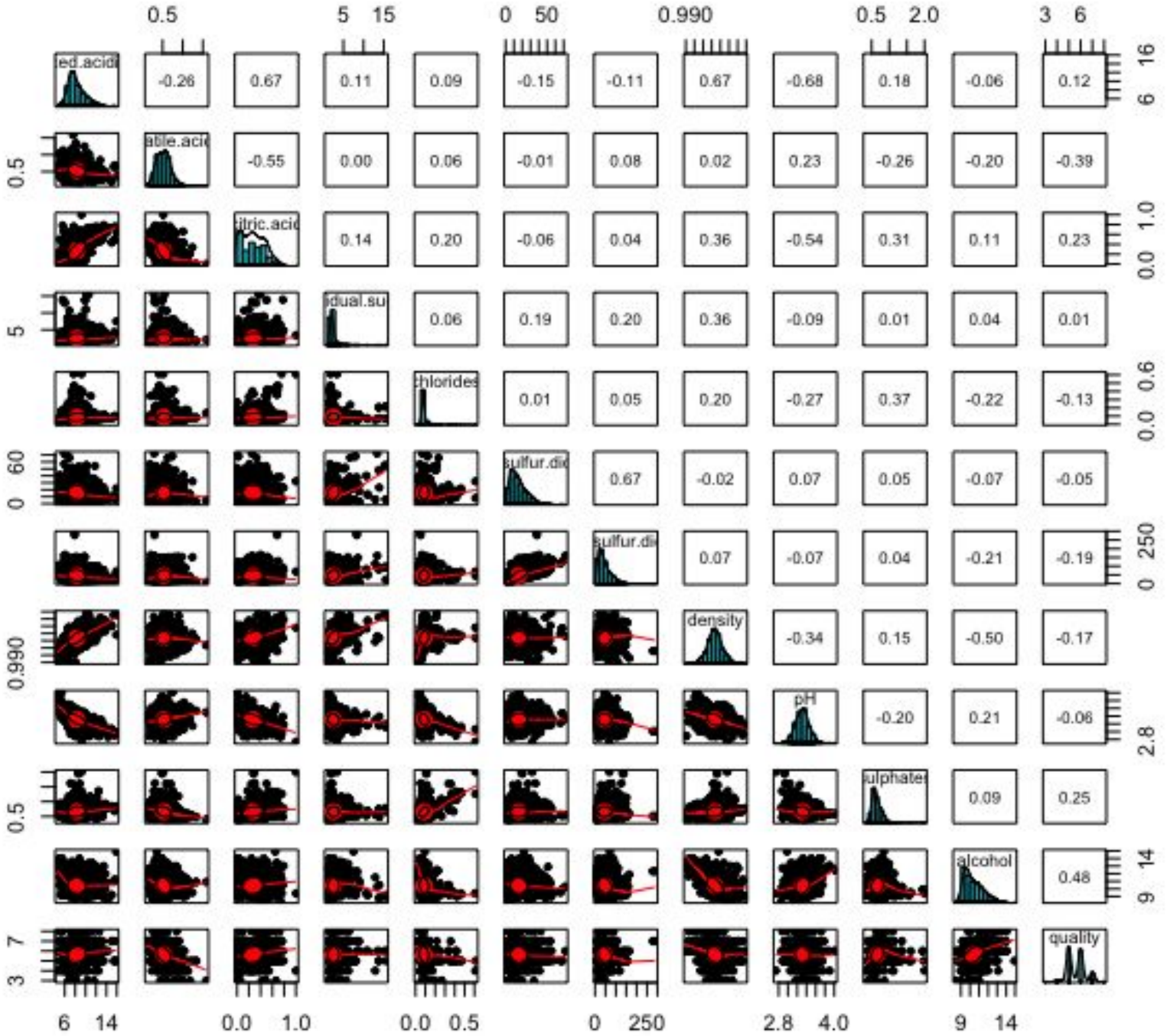
Continuous, N = 4177

Input variables (based on physicochemical tests):

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol

Finally from a taste test:

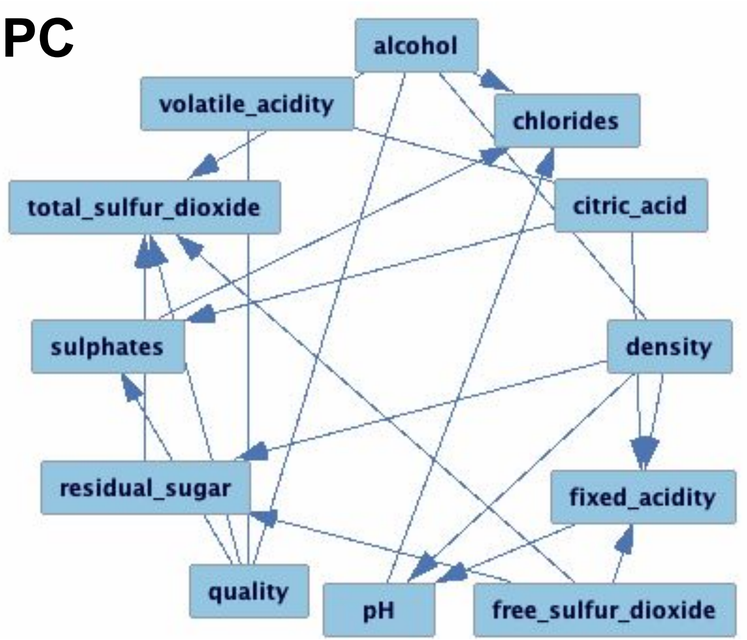
12. quality (score between 0 and 10)



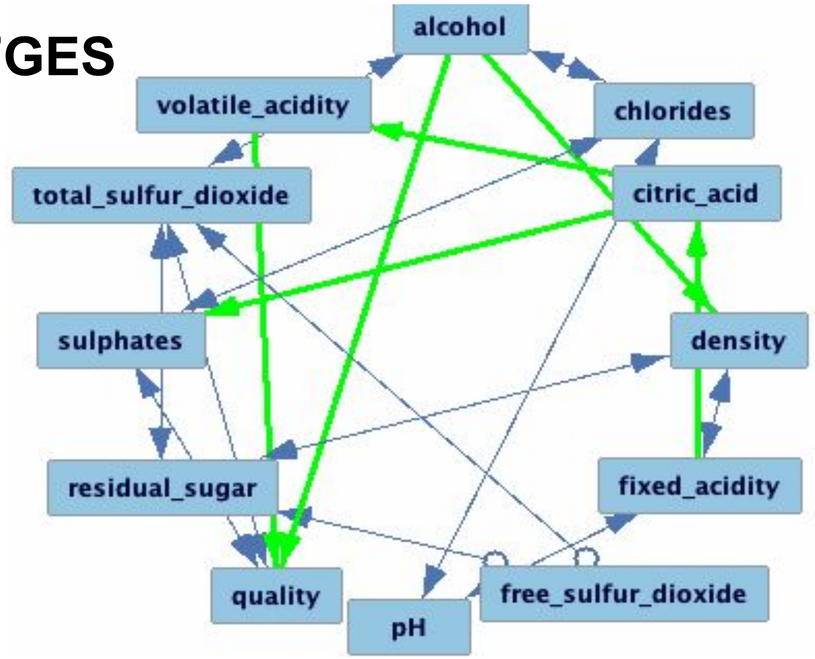
Wine Quality, Red (UCI) - Some Ground Truth

- **Tier 1 alcohol chlorides, citric_acid, density, fixed_acidity, free_sulfur_dioxide, pH, residual_sugar, sulphates, total_sulfur_dioxide, volatile_acidity**
- **Tier 2 quality**
- **Required edges (from the expert knowledge I read):**
  - **alcohol→quality**
  - **total_sulfur_dioxide→quality**

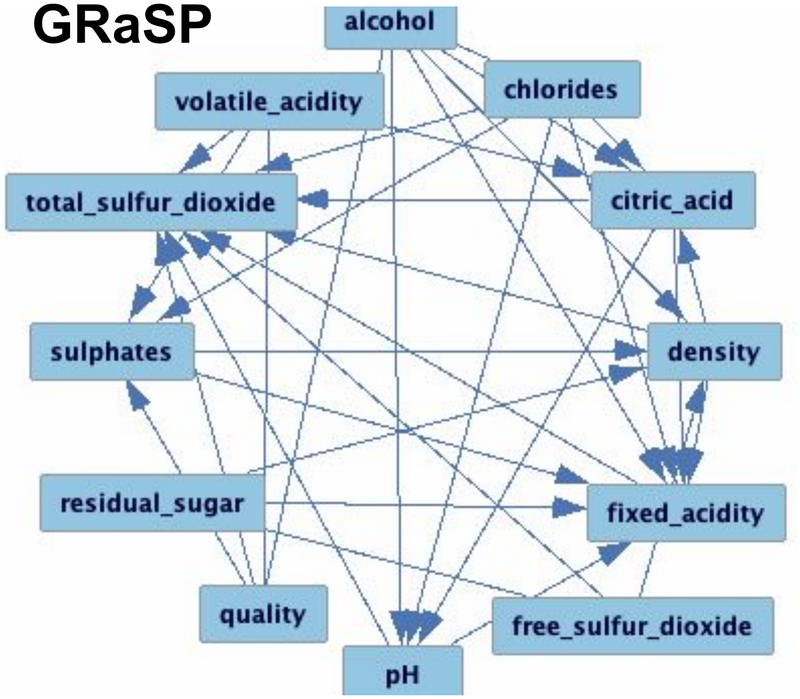
PC



FGES



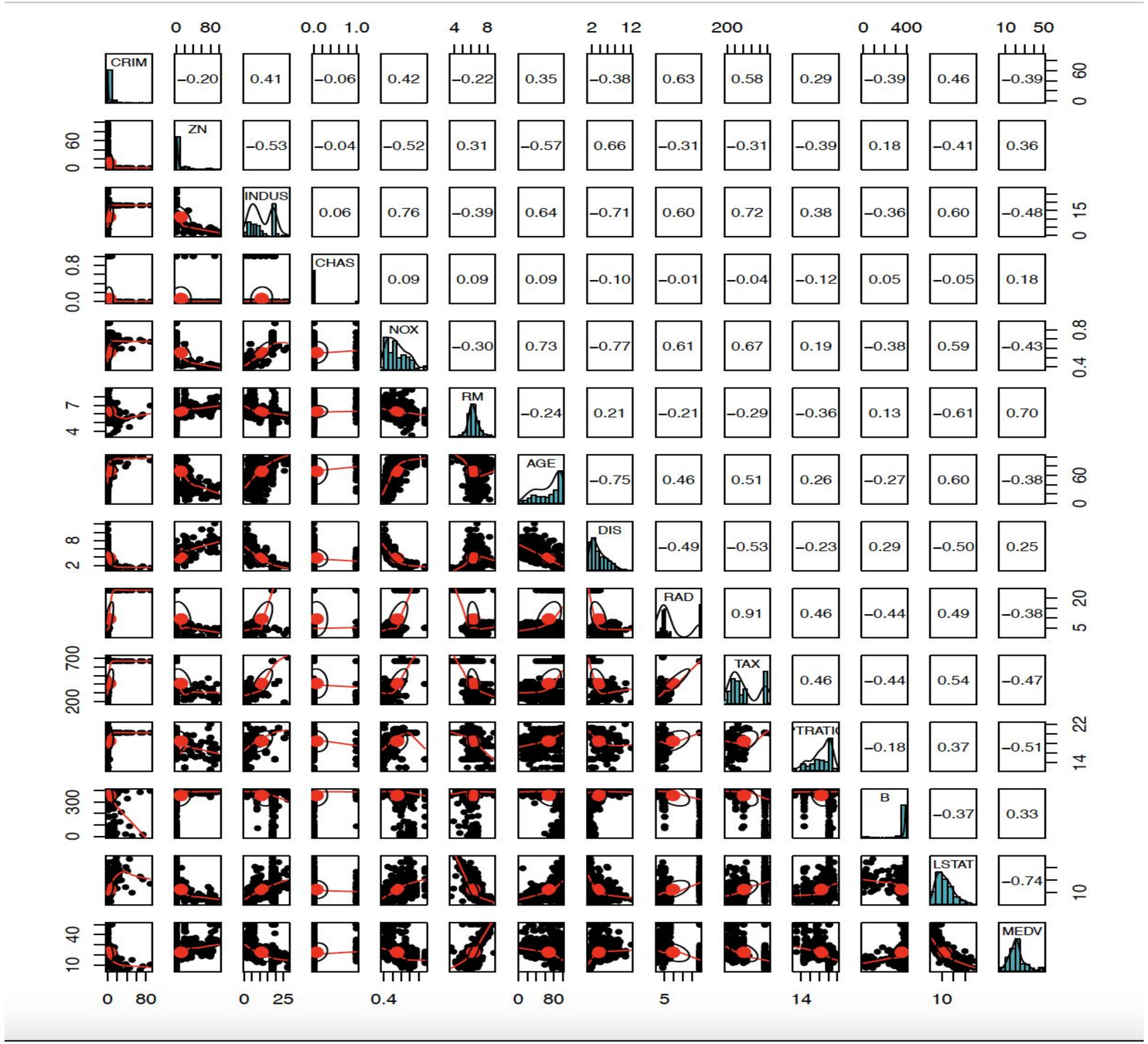
GRaSP



Boston Housing

Mixed, N = 506

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town
4. CHAS - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in $1000's



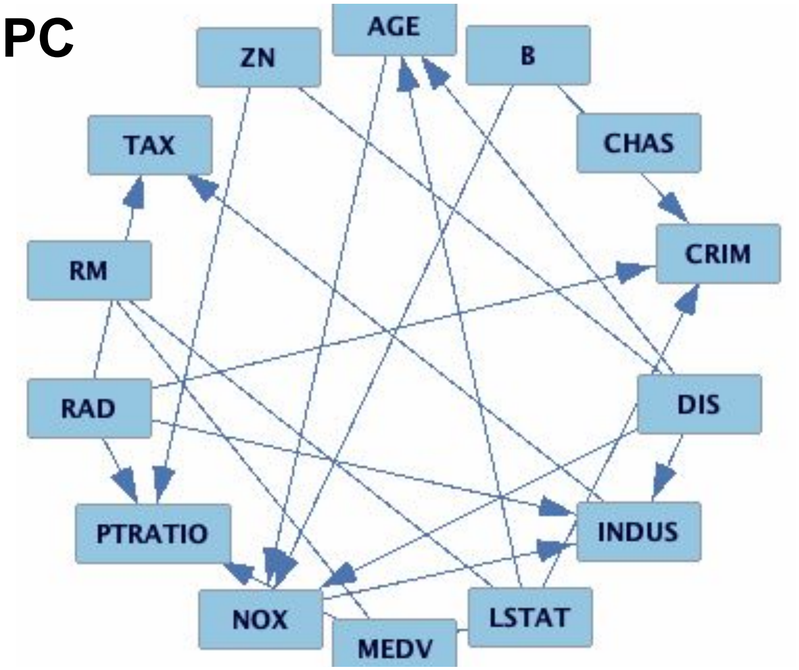
Boston Housing (UCI archive) - Some Ground Truth

Used this knowledge for Tetrad to guide search (but not CL because I don't know how yet). Recommended by Zhao and Hastie.

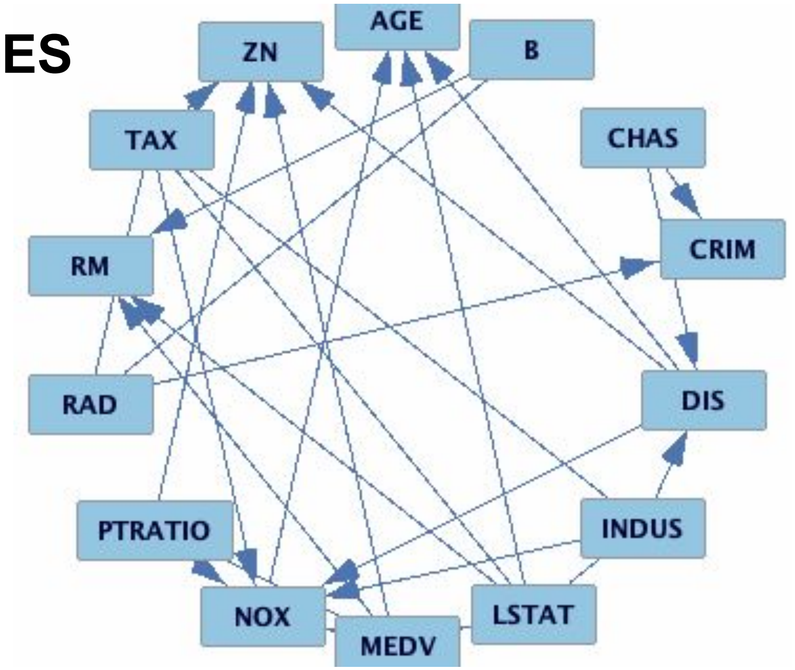
- **Tier 1: Other vars**
- **Tier 2: NOX**

- **Violates linear/Gaussian/additive**
- **Mixed**

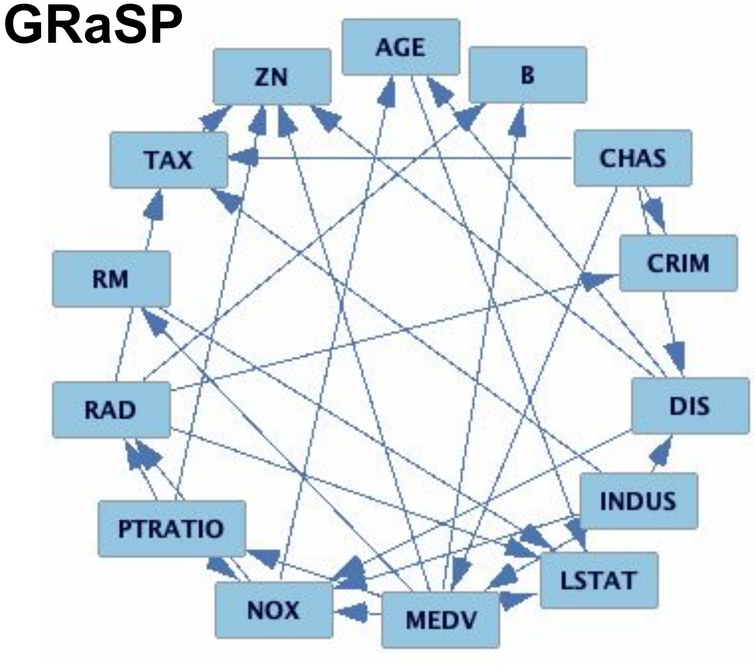
PC



FGES



GRaSP



Apple Fitbit (Kaggle)

Mixed, N = 6264

Variables cleaned up a bit…

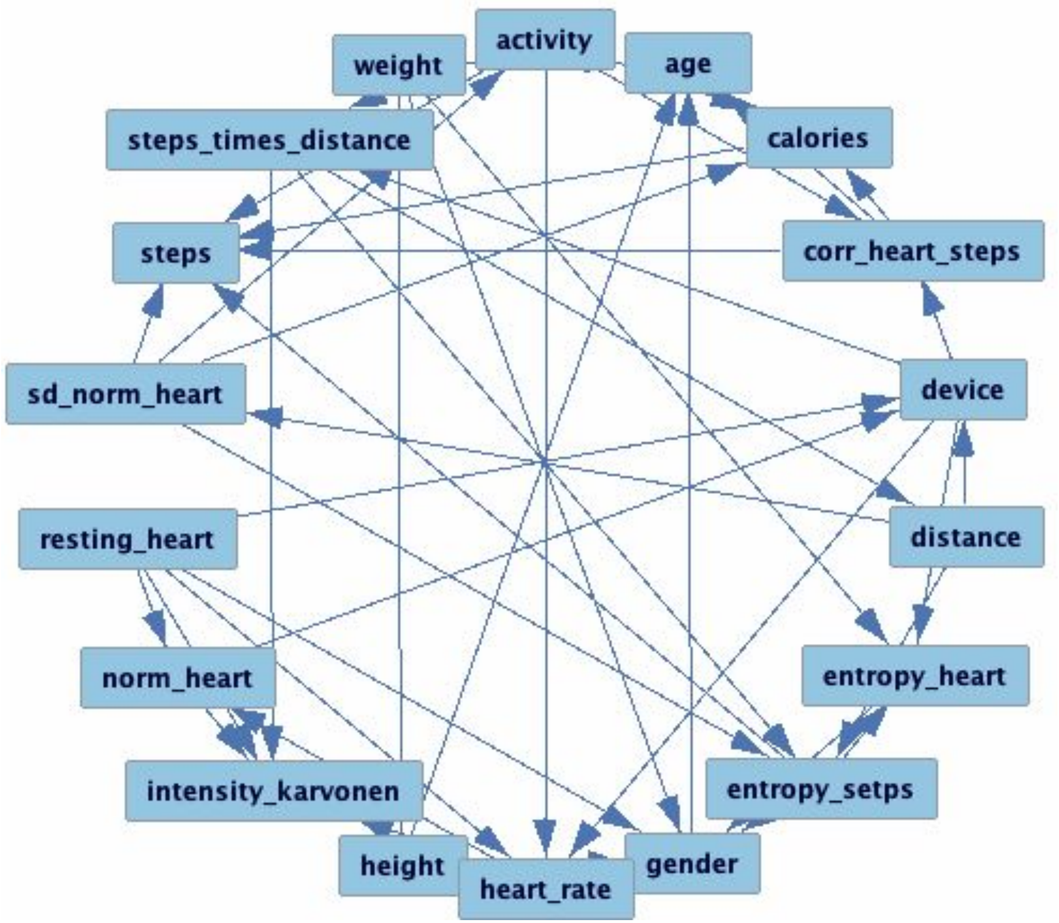
1. age
2. gender
3. height
4. weight
5. steps
6. heart_rate
7. calories
8. distance
9. entropy_heart
10. entropy_steps
11. resting_heart
12. corr_heart_steps
13. norm_heart
14. sd_norm_heart
15. steps_times_distance
16. device
17. activity

Minimal ground truth:

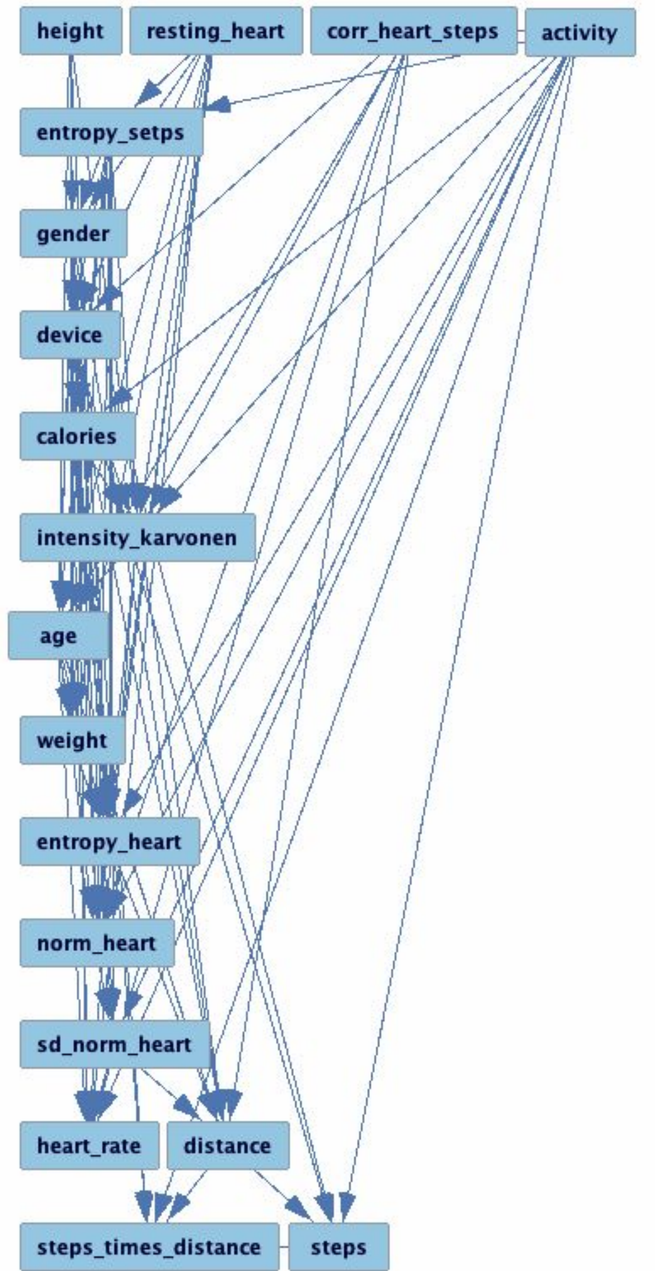
- **Tier 1:** age, height, gender, weight, device, activity
- **Tier 2:** everything else

- **Violates i.i.d.,** sample of ~50 people and each instance is their stats over a 1 minute period during a 65 minute session
- **Mixed continuous/discrete**
- **Violates linear/Gaussian**

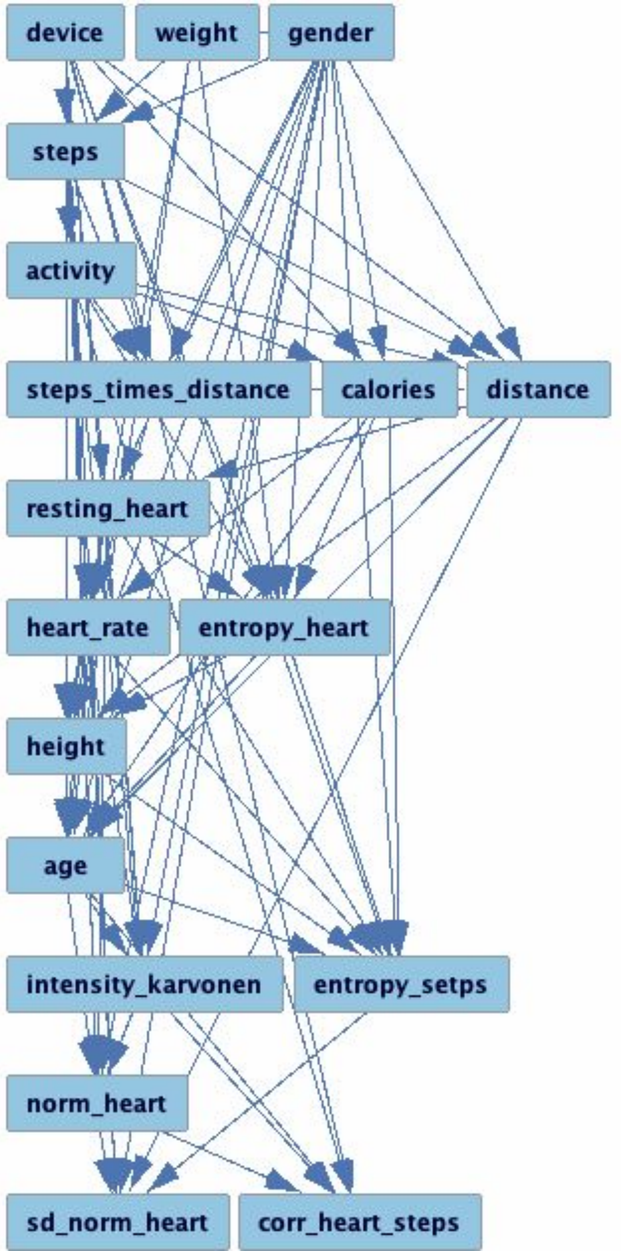
PC


activity
weight
age
steps_times_distance
calories
steps
corr_heart_steps
sd_norm_heart
device
resting_heart
distance
norm_heart
entropy_heart
intensity_karvonen
entropy_setps
height
heart_rate
gender

FGES


height
resting_heart
corr_heart_steps
activity
entropy_setps
gender
device
calories
intensity_karvonen
age
weight
entropy_heart
norm_heart
sd_norm_heart
heart_rate
distance
steps_times_distance
steps

GRaSP


device
weight
gender
steps
activity
steps_times_distance
calories
distance
resting_heart
heart_rate
entropy_heart
height
age
intensity_karvonen
entropy_setps
norm_heart
sd_norm_heart
corr_heart_steps