Causal Inference
Making the right intervention

Paul Beaumont, PhD
Senior Data Scientist at QuantumBlack Singapore

QUANTUMBLACK
A MCKINSEY COMPANY

The world of machine learning – in a nutshell

Predictive modelling

- Estimate the target for new observations

$$\boxed{y} = f(x)$$

Explanatory modelling

- Describe the effect that a change of certain inputs has on the target

$$y = \boxed{f}(x)$$

Optimisation

- Find the inputs that give optimal performance
- f is known

$$y = f(\boxed{x})$$

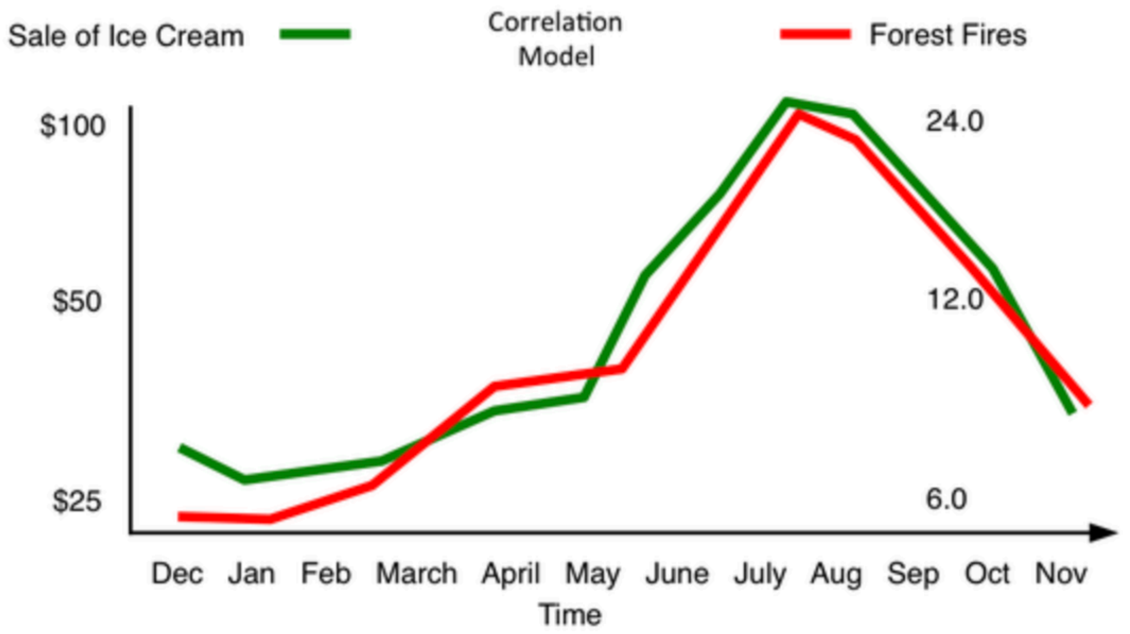
Galit Shmueli (2010) To Explain or to Predict? Statistical Science,

Key decisions require explanatory models

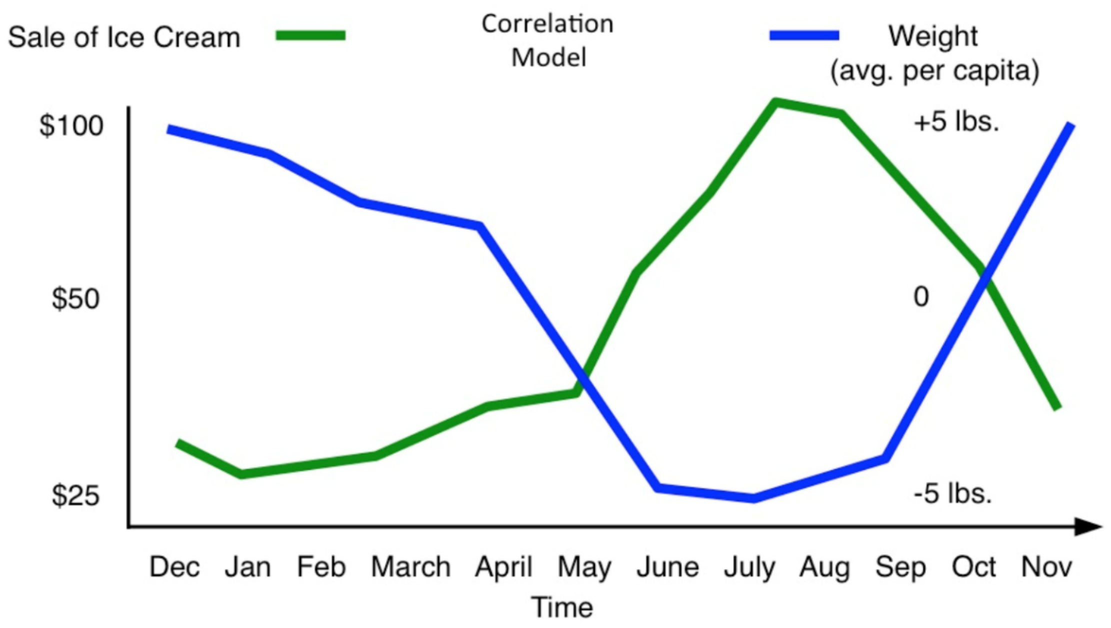
- Which medication will help a given patient?
- What marketing campaign will be most effective?
- How can a pharmaceutical company reduce non-conformities during their drug manufacturing process?
- What changes can a vehicle manufacturer make to their new product development process to reduce lead time?
- How can an company deploy resources to better serve customers?

We'd expect these explanations to make causal sense before trusting the model

Does ice cream cause forest fires?



Is ice cream the new diet food?

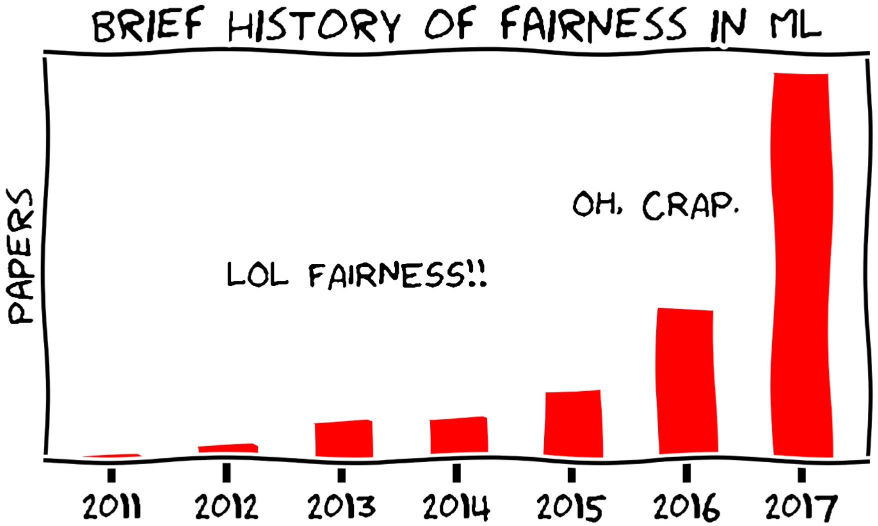


Actionable insights?

Causal inference is a hot topic in data science…

- Desire for **causal** methods given the prevalence of machine learning algorithms in all parts of society.
- **Counterfactual fairness**[1]: A decision is fair towards an individual if it is the same in
  - (a) the actual world and
  - (b) a counterfactual world where the individual belonged to a different demographic group.
- Close relationship to **Reinforcement learning**

Pearl: "Systems that operate in purely statistical mode of inference ... cannot reason about interventions … and, therefore, cannot serve as the basis **for strong AI**."[2]



1. Kusner, Loftus, Russell, Silva (2017) Counterfactual Fairness. NeurIPS
2: Pearl (2018) The seven tools of causal inference with reflections on machine learning, CACM
Image: https://fairmlclass.github.io/

… but machine learning unfortunately often doesn’t care about causality

```
model = sm.OLS(train.y, train[['t']])
results = model.fit()
print(results.summary())
```

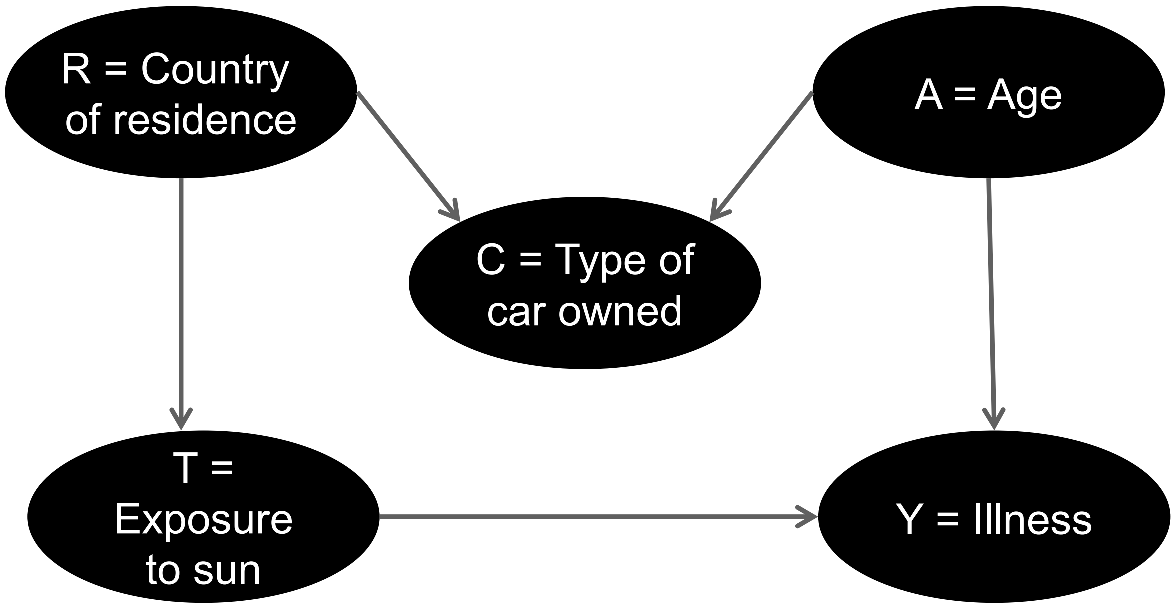
Test RMSE = 10

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.019
Model:                            OLS   Adj. R-squared:                  0.019
Method:                 Least Squares   F-statistic:                 2.766e+04
Date:                Wed, 13 Feb 2019   Prob (F-statistic):               0.00
Time:                        07:25:25   Log-Likelihood:            -5.2235e+06
No. Observations:             1401801   AIC:                         1.045e+07
Df Residuals:                 1401800   BIC:                         1.045e+07
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
t              0.9984      0.006    166.317      0.000       0.987       1.010
==============================================================================
```

```
model = sm.OLS(train.y, train[['t', 'car']])
results = model.fit()
print(results.summary())
```

Test RMSE = 7.7

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.408
Model:                            OLS   Adj. R-squared:                  0.408
Method:                 Least Squares   F-statistic:                 4.835e+05
Date:                Wed, 13 Feb 2019   Prob (F-statistic):               0.00
Time:                        07:27:29   Log-Likelihood:            -4.8695e+06
No. Observations:             1401801   AIC:                         9.739e+06
Df Residuals:                 1401799   BIC:                         9.739e+06
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
t             -1.0036      0.005   -196.455      0.000      -1.014      -0.994
car            4.0024      0.004    959.740      0.000       3.994       4.011
==============================================================================
```



$R = \varepsilon_1$

$A = \varepsilon_2$

$C = R + A + \varepsilon_3$

$T = R + \varepsilon_4$

$Y = T + 10\,A + \varepsilon_5$

$\varepsilon_1, \dots, \varepsilon_5$ independent normal(0,1)

Goal:
Find the effect of sun exposure on the illness

Pearl (2014) Comment: Understanding Simpson’s Paradox, The American Statistician.
Simpson machine generator: http://dagitty.net/learn/simpson/

… but machine learning unfortunately often doesn't care about causality

```
model = sm.OLS(train.y, train[['t']])
results = model.fit()
print(results.summary())
```

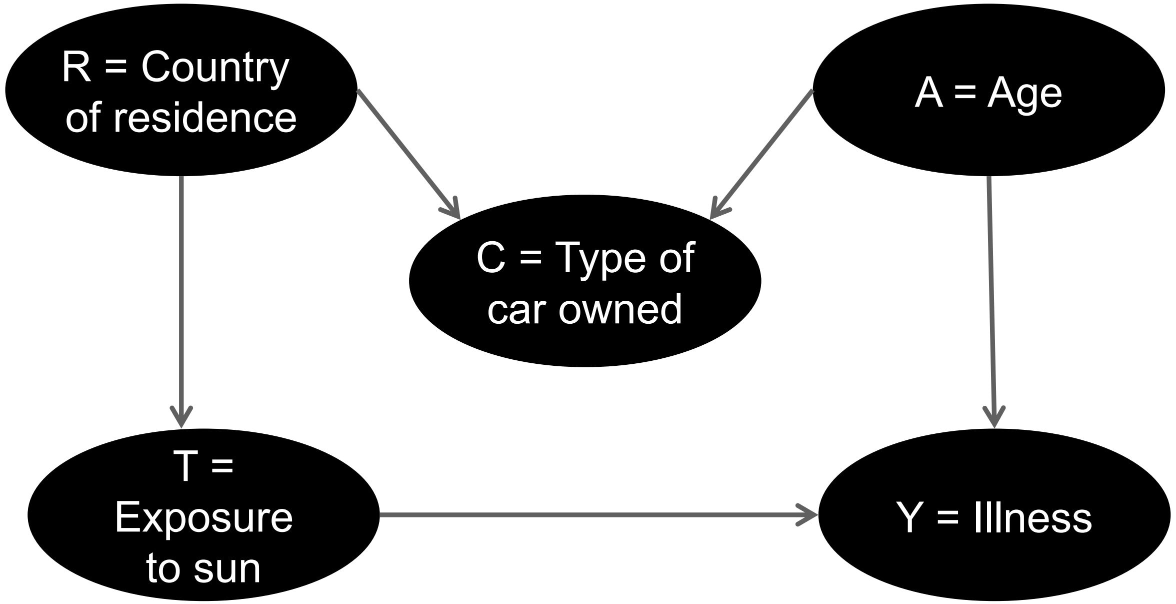
Test RMSE = 10

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.019
Model:                            OLS   Adj. R-squared:                  0.019
Method:                 Least Squares   F-statistic:                 2.766e+04
Date:                Wed, 13 Feb 2019   Prob (F-statistic):               0.00
Time:                        07:25:25   Log-Likelihood:            -5.2235e+06
No. Observations:             1401801   AIC:                         1.045e+07
Df Residuals:                 1401800   BIC:                         1.045e+07
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
t              0.9984      0.006    166.317      0.000       0.987       1.010
==============================================================================
```

```
model = sm.OLS(train.y, train[['t', 'car']])
results = model.fit()
print(results.summary())
```

Test RMSE = 7.7

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.408
Model:                            OLS   Adj. R-squared:                  0.408
Method:                 Least Squares   F-statistic:                 4.835e+05
Date:                Wed, 13 Feb 2019   Prob (F-statistic):               0.00
Time:                        07:27:29   Log-Likelihood:            -4.8695e+06
No. Observations:             1401801   AIC:                         9.739e+06
Df Residuals:                 1401799   BIC:                         9.739e+06
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
t             -1.0036      0.005   -196.455      0.000      -1.014      -0.994
car            4.0024      0.004    959.740      0.000       3.994       4.011
==============================================================================
```



$$R = \varepsilon_1$$
$$A = \varepsilon_2$$
$$C = R + A + \varepsilon_3$$
$$T = R + \varepsilon_4$$
$$Y = T + 10\,A + \varepsilon_5$$

$\varepsilon_1, \dots, \varepsilon_5$ independent normal(0,1)

Goal:
Find the effect of sun exposure on the illness

Pearl (2014) Comment: Understanding Simpson's Paradox, The American Statistician.
Simpson machine generator: http://dagitty.net/learn/simpson/

Randomised Control Trials test for causality but have limitations; most data is observational

Randomised Control Trials (RCT)

- Randomly assign treatment to individuals:

$$Y \perp T$$

- Often small data set
- Limited generalizability, risk if participants are not representative of population
- Unethical in many cases
- Unconfounded by design

Observational Studies

- Data is generated without the causal question in mind
- Often large and rich data set
- Most common case because:
  - Did not think of the question when data was created
  - Financial and reputational risk
  - Budget and time constraints
- Potential problem with hidden confounding

There are 2 key challenges we need to solve when working with observational data:

1. Finding the causal direction
2. Confounding

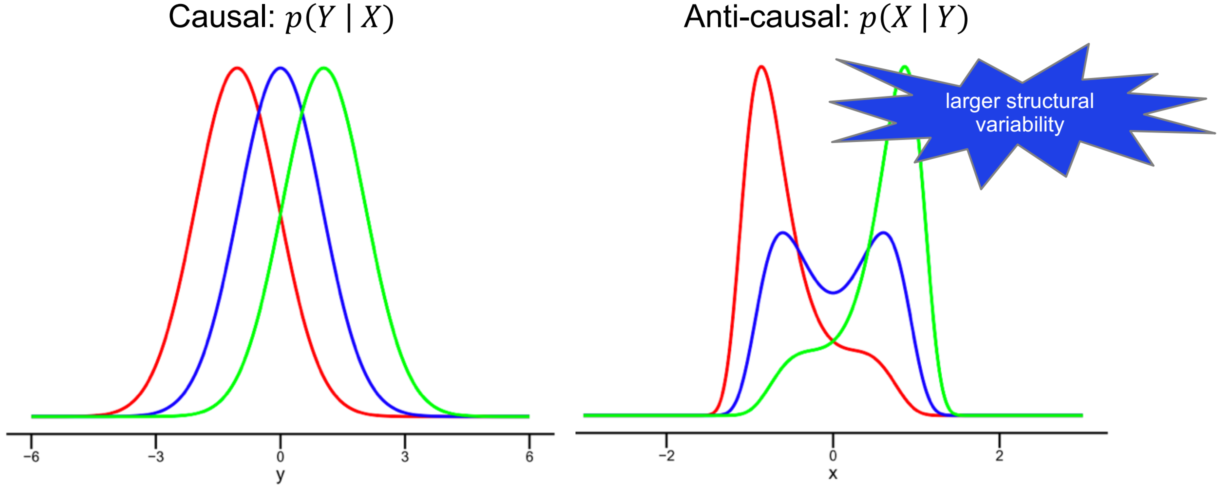
…let’s discuss some potential solutions!

Identifying the causal direction of a relationship purely from data is something the research community is working on

Mitrovic, Sejdinovic & Teh’s NeurIPS 2018 paper “Causal Inference via Kernel Deviance Measures (KCDC)” postulates that sometimes a causal direction can be determined from distributions of the data.

Example:

If $X \rightarrow Y$, $y|\, x \sim N(x^3 + \mathrm{x}, \sigma^2)$ then



“… asymmetry is realized by the Kolmogorov complexity of the mechanism in the causal direction being ***independent*** *of the input value of the cause.”*

…but sometimes expert help will be necessary

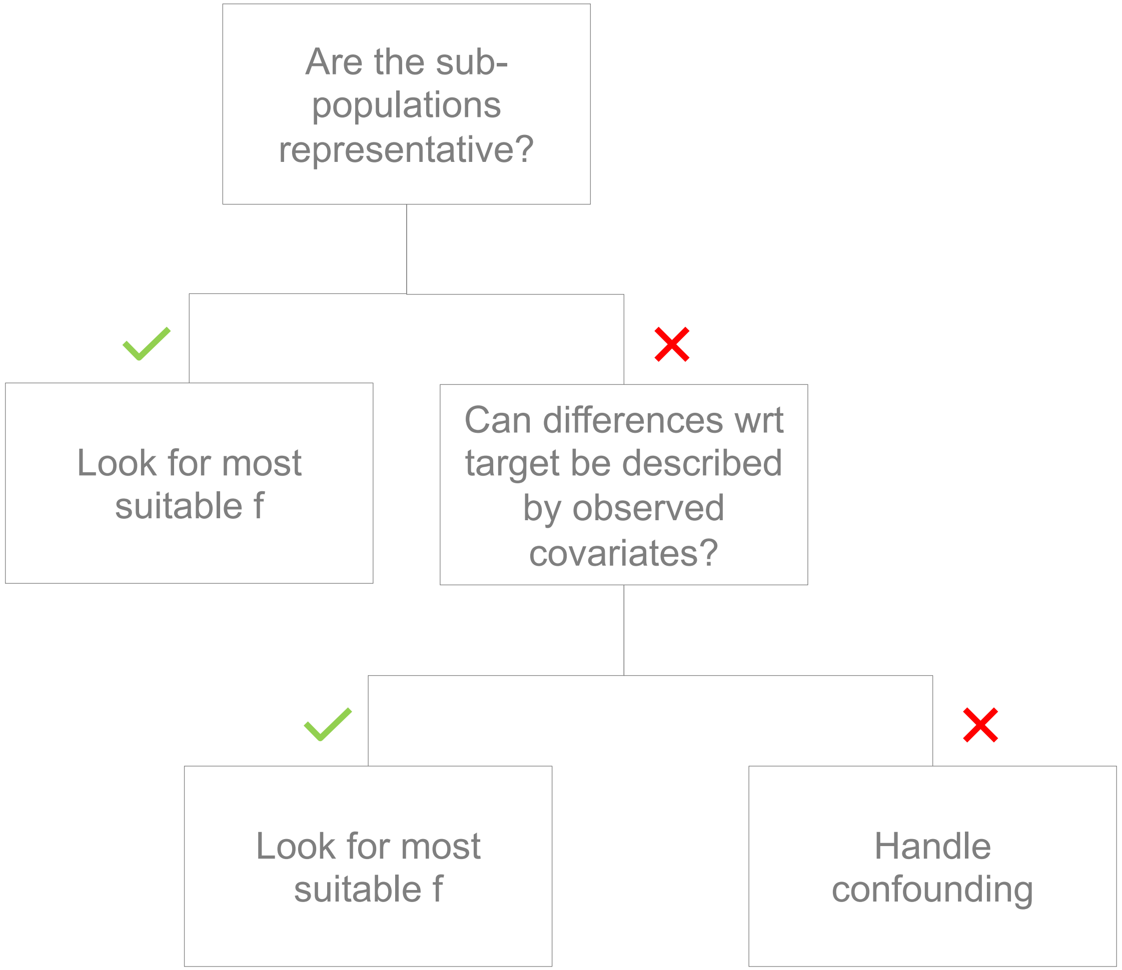
arXiv: [1804.04622]

Confounding poses a risk to causal inference on observational data

- A **confounder** is a variable that influences both the treatment and the target
- Confounding can **limit identifiability** of the causal effect

Suitable fixes

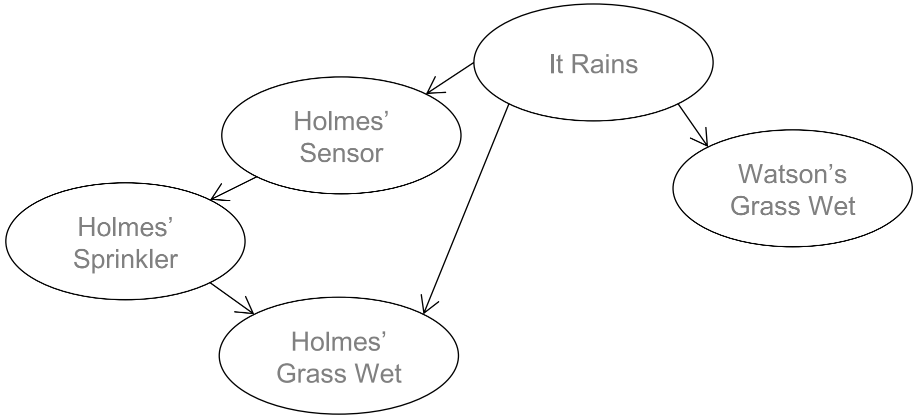
- Match populations using **propensity score matching**
- Capture non-linear relationships between Y, X and highly-varied X across treatment groups using **ML**
- Obtain confidence intervals using **Causal Forests, BART**
- Model relationships between all variables and encode subject matter expertise using **Bayesian Networks** or **structural equations**



Augmenting modelling using expert knowledge can help

Graphical models encode domain expertise

- **Bayesian Networks** are graphical models where the graph is a DAG
- Assumptions are marked by the (lack of) edges
- Incorporates the human understandable part of the model
- Facilitates discussion with subject matter experts



Structural equations facilitate counterfactuals

- Mathematical encoding of the transformations of parent nodes into child nodes
- Each function is autonomous to possible changes in the form of the other functions

$$D = f_1(L, \varepsilon_1)$$

$$W = f_2(M, L, \varepsilon_2)$$

$$H = f_3(W, D, \varepsilon_3)$$

Pearl (1995) Causal diagrams for empirical research. Biometrika 82
Pearl (2000) Causality: Models, Reasoning, and Inference. Cambridge University Press (2nd edition 2009)

Graphical models, notably Bayesian Networks, are an intuitive way to encode context knowledge

Structure learning

- **Computationally demanding**
- Constraint-based methods
- Score-based methods
  - Continuous optimization: DAGs with NO TEARS

- Hybrid learning where domain expertise edits the network structure:
  - Ensure causal direction
  - Add missing (but weak) associations
  - Handle spurious data relationships

Model performance

- Provided sufficient data, BNs should outperform simpler interpretable models
  - Allow for modelling of interdependencies between variables, rather than additive

Perform inference

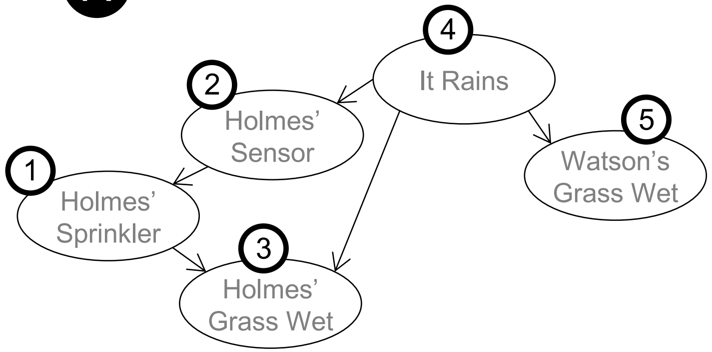
- **Maximum likelihood estimation** for one-step probabilities
- Conditional distributions as product of one-step probabilities along the route
- **Junction tree algorithm** for efficient execution of inference

Zheng et al. (2018). DAGs with NO TEARS: Continuous Optimization for Structure Learning, NeurIPS

Bayesian Networks have historically struggled to get traction as they were difficult to learn; new methods drastically change this

Previous techniques suffered because they needed to "check acyclicity holds" and this is a combinatorial optimization problem. The authors of *DAGs with NO TEARS (Zheng et al.)* convert this to a continuous test (that is faster and easier to incorporate into search algorithms), leveraging the properties of the adjacency matrix

A



$$\begin{vmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{vmatrix} = W$$

1 3
2 1
4 2
4 5

B

$$\begin{vmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{vmatrix} \begin{vmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{vmatrix} = \begin{vmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{vmatrix} = W^2$$

4 1
2 3

A The leading diagonal (or trace) of a DAG's adjacency matrix, *W*, is all zeros.

B Raising *W* to a power, *k* will produce all possible paths *k* steps away. In a DAG, trace(W^k) = 0 for all *k*.

- trace(W^k) = 0 for all *k* is true iff:

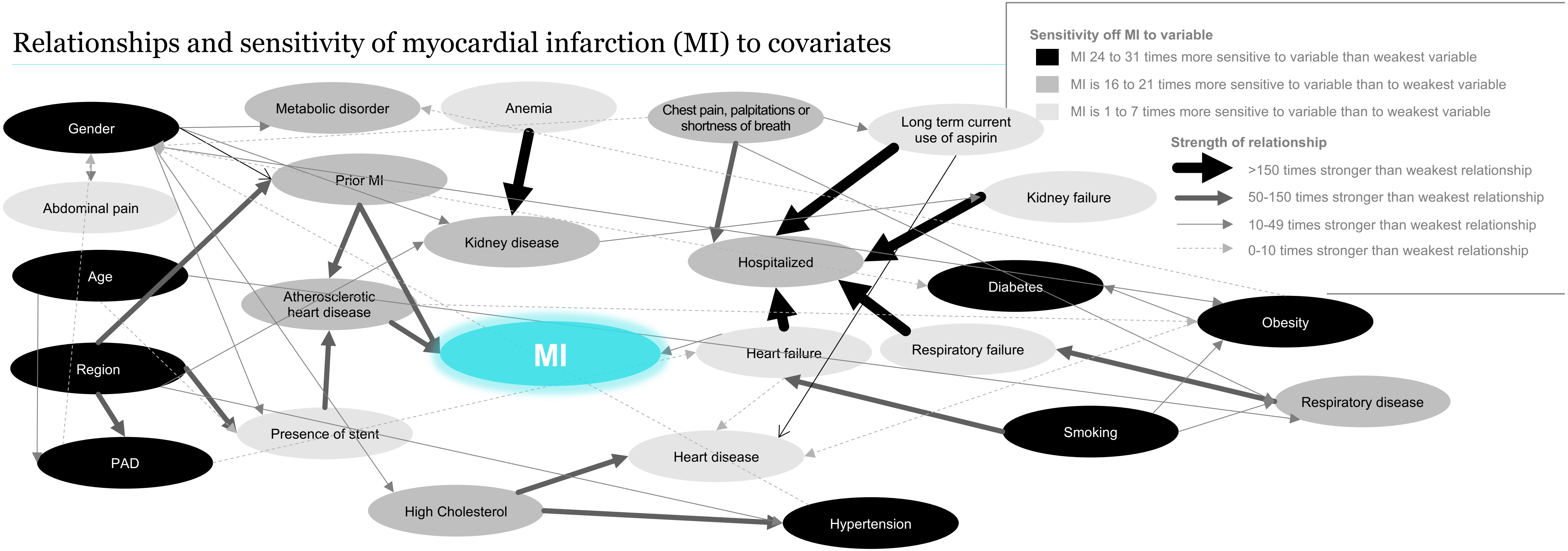
$$\sum_{k=1}\sum_{i}^{d}\frac{(W^{2k})_{ii}}{k!} = trace\left(e^{(W\odot W)}\right) - d = 0 \ (< \epsilon)$$

https://arxiv.org/pdf/1803.01422.pdf

ILLUSTRATIVE

Causal models can be used to support decision making in important domains such as healthcare

Relationships and sensitivity of myocardial infarction (MI) to covariates



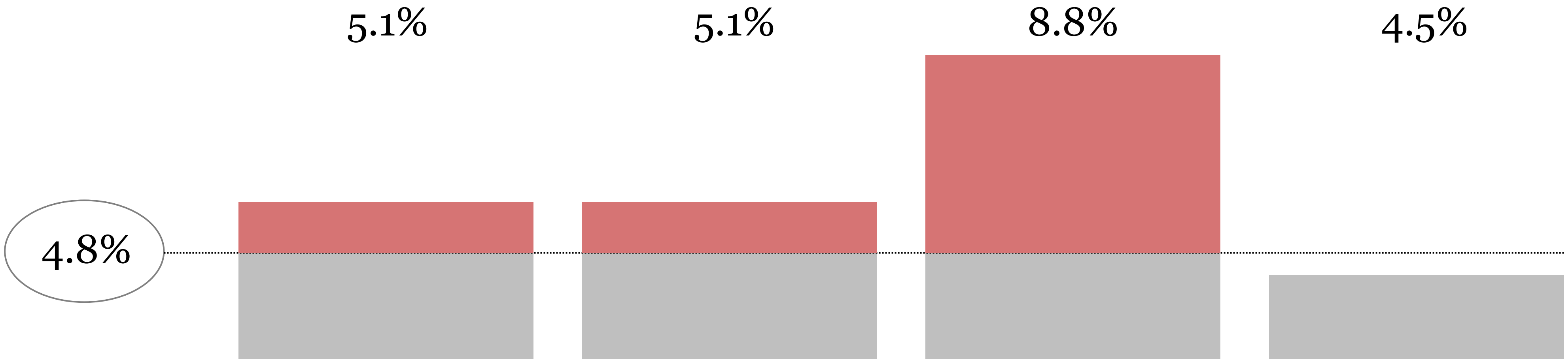
- The network structure is generated from both data and domain knowledge.
- Incorporating domain expertise ensures the model represents a domain expert's view of causal relationships
- Quantifying the relationship between patient demographics, comorbidities, and cardiovascular events can be used to identify key drivers of patient risk

Trained on Truven Claims data (2015 – 2017)
Structure learning uses DAGs with NO TEARS: Continuous Optimization for Structure Learning; Zheng et al.; NeurIPS 2018

With this approach we could better understand patient journeys

Risk of MI within 12 months[1]

Baseline risk of CAD/PAD patient having a MI event



Mrs. Smith, smoker, BMI 32, age 50, PAD

Mrs. Smith's lifestyle drives an increase in MI risk, as well as an increased risk of diabetes (2.1% change to 2.6%)

Mrs. Smith, now aged 55, has diabetes

Mrs. Smith's risk of MI is not much increased – but she is now at increased risk of heart failure (4.5% to 11.7%)

Mrs. Smith, age 58, has developed symptoms of heart failure

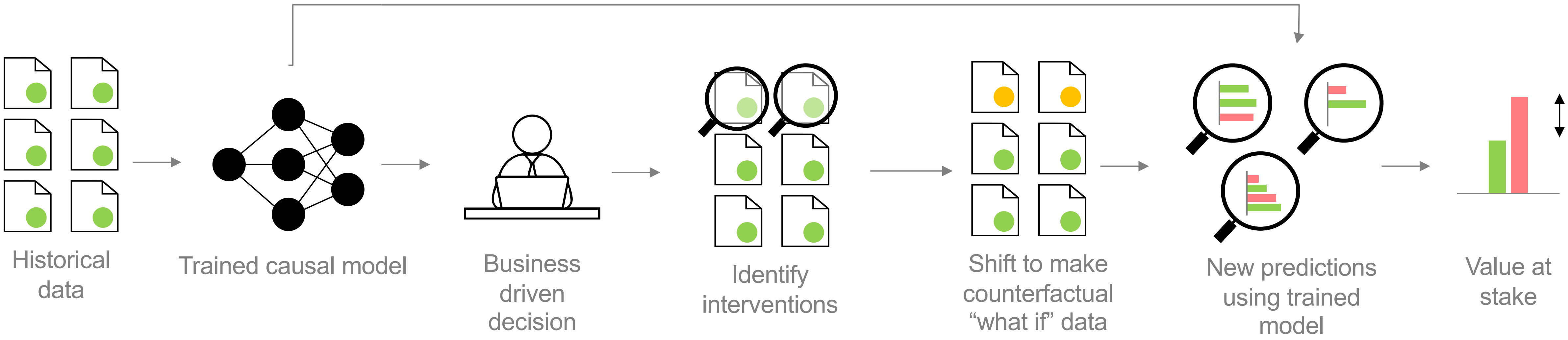
Based on the symptoms of heart failure, Mrs Smith is now at much higher risk of MI

Mrs Smith, age 62, has made lifestyle changes: no longer smokes and has a BMI of 29

Her risk of an MI is now below that of the average CAD/PAD patient

SOURCE: Truven Claims Data (2105-2017)

More generally, if we model causally we can apply data science to business problems and perform counterfactual analysis to ask “what if?”

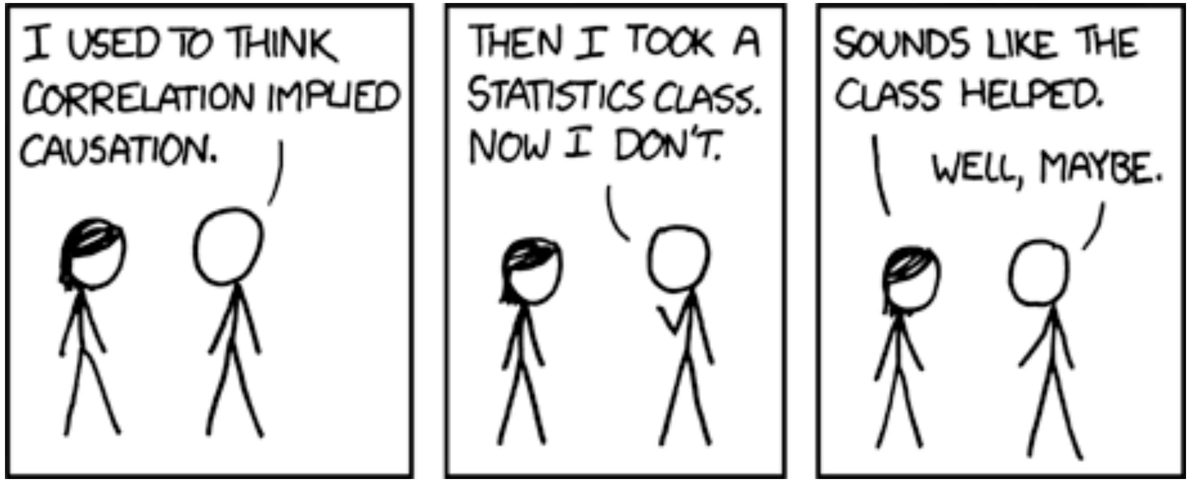


- Once we have trained a causal model, we identify counterfactuals that we would like to test and “intervene” on.
- These are generated by changing the historical data to reflect the actions of the intervention, and new predictions (of a target) are generated.
- Comparing these to the target from the “real” data allows us to calculate the value at stake of implementing the counterfactual change.

- **If our models aren’t causal, our “what if’s” could be very inaccurate**

Takeaways

- If we want to trust models for decisions, then we should expect them to make **causal sense**
- Training on observational data is common, and the causal direction of relationships is not always clear
- Methods exist to **help us identify possible causal relationships**, but domain experts can also help
- **Models that respect causality** also exist and thanks to recent advances are now easier to learn and deploy



Comic: https://xkcd.com/552/

