XAI: Correlation, Causality, and Feature Importance

Lyle Ungar

Different types of explanations
Shapley values
Correlation is not causality!

Why do people build models?

- **ML: prediction**
  - the y-hat culture $y = f(x; w)$
- **Statistics: hypothesis testing**
  - the beta-hat culture $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon$
- **The real world: often picking the best actions**
  - Requires causality

Why explain ML models?

- **Debugging/verification:** Explain a model
  - How much do each of these features contribute to the prediction?
  - Will this model work for other populations?
- **Science:** Explain the world
  - What protocols give the best surgery outcomes
  - What are the largest risk factors for suicide?
- **Decision support:** Explain a prediction
  - Why do you think this patient is at high risk for suicide?

Explanation is tricky

- **P(death) = f(age, height, weight, BMI, AVPU, GCS,…)**
  - How important is each of these features?
- **What does it mean to change "weight" as an input**
- **To explain the world, we often want causality**

Feature importance measures

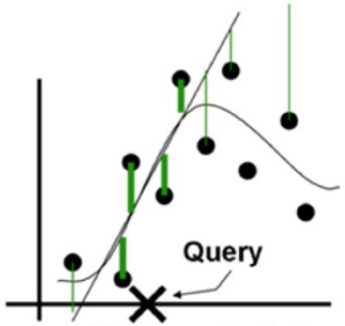
- **Univariate correlation**
- **Effect of 'removing' a feature**
  - Set feature to its average value (Partial Dependence Plot)
  - Permute it
  - Remove it and retrain
- **Shapley values**
  - Nice mathematical properties

LIME

Original Image
P(tree frog) = 0.54

Perturbed Instances	P(tree frog)
	0.85
	0.00001
	0.52

Generate a data set of perturbed instances by turning some of the interpretable components “off” (gray)



Locally weighted regression

Explanation
= pixels with high weights

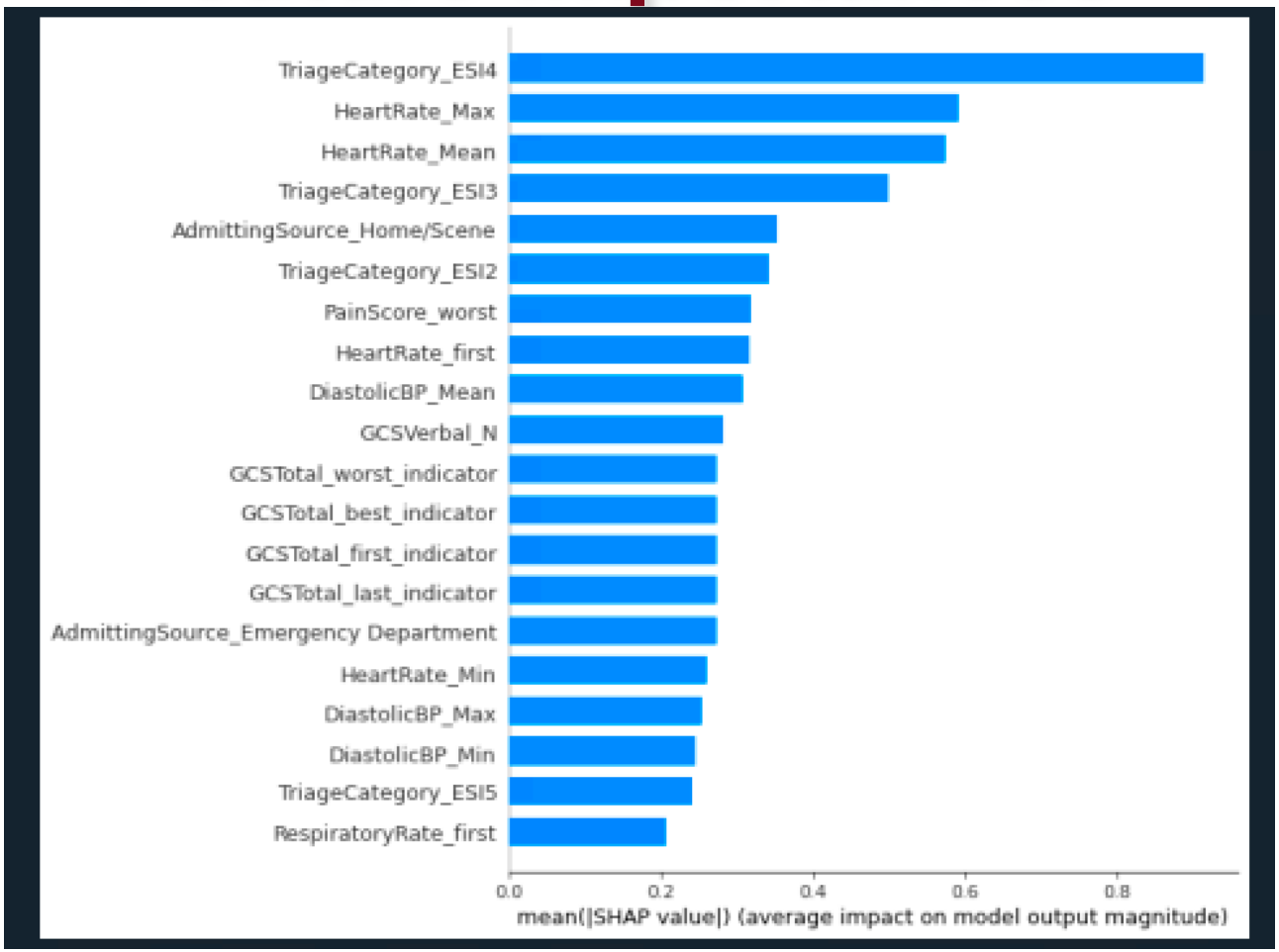
Types of explanations

- **Interventional vs. conditional**
  - *Interventional (aka marginal): :* Change one feature leaving the others fixed
    - Explain the model
  - *Conditional* change other features to respect correlations
- **Model-based vs. model-agnostic**
  - *Model-based:* Look inside the model (e.g. at coefficients)
  - *Model-agnostic:* explain the world
- **Local vs. global**
  - *Local:* feature importance for one patient
  - *Global:* average feature importance over observations

To explain the world, we often want causality

- **P(death) = f(age, height, weight, BMI, AVPU, GCS …)**
  - How important is each of these features?
- **Does age or does weight most affect p(death)**
- **How does AVPU or GCS affect p(death)**
  - AVPU scale: Alert, Voice, Pain, Unresponsive
  - GCS: Glasgow Coma Scale

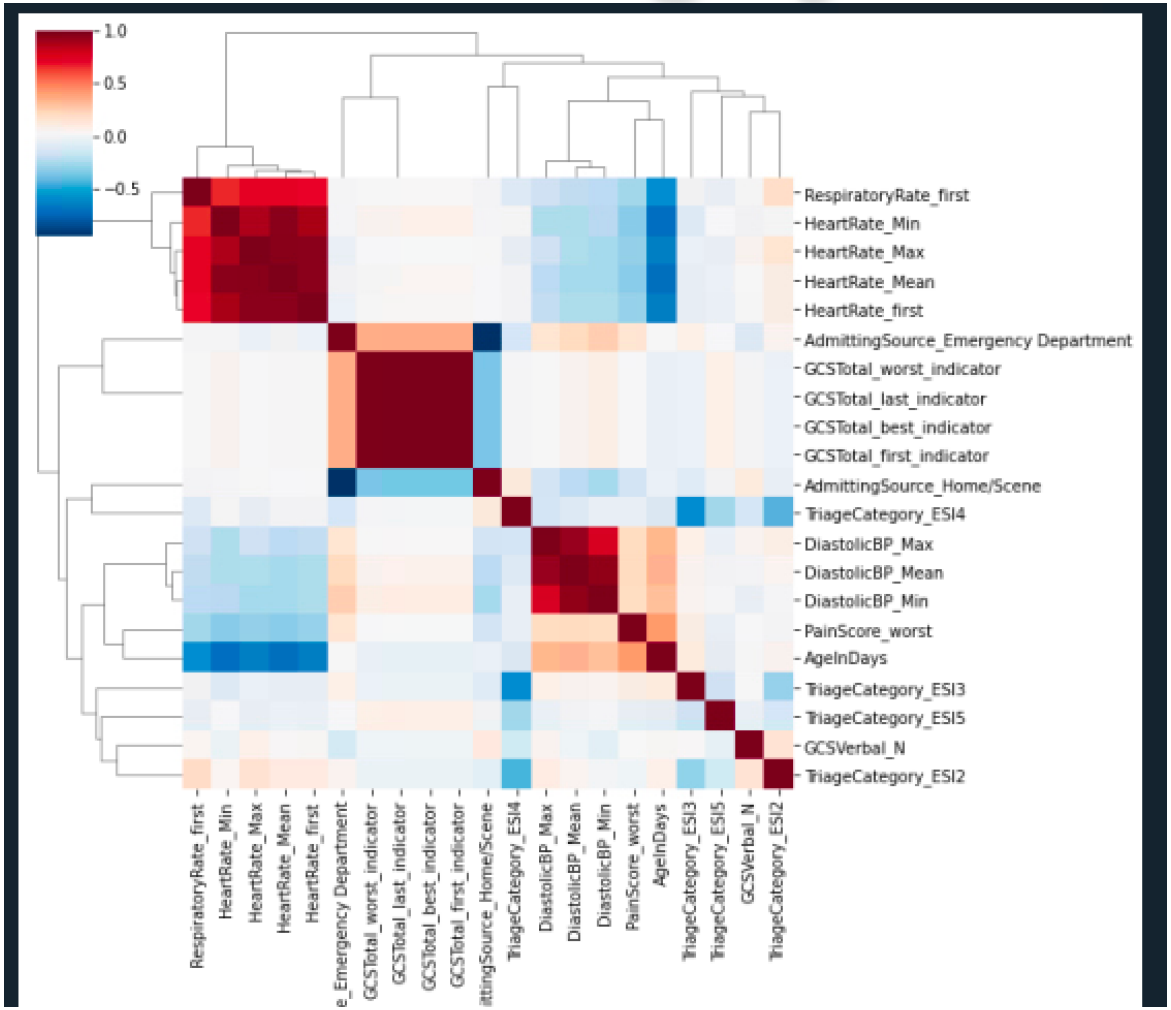
Feature importances



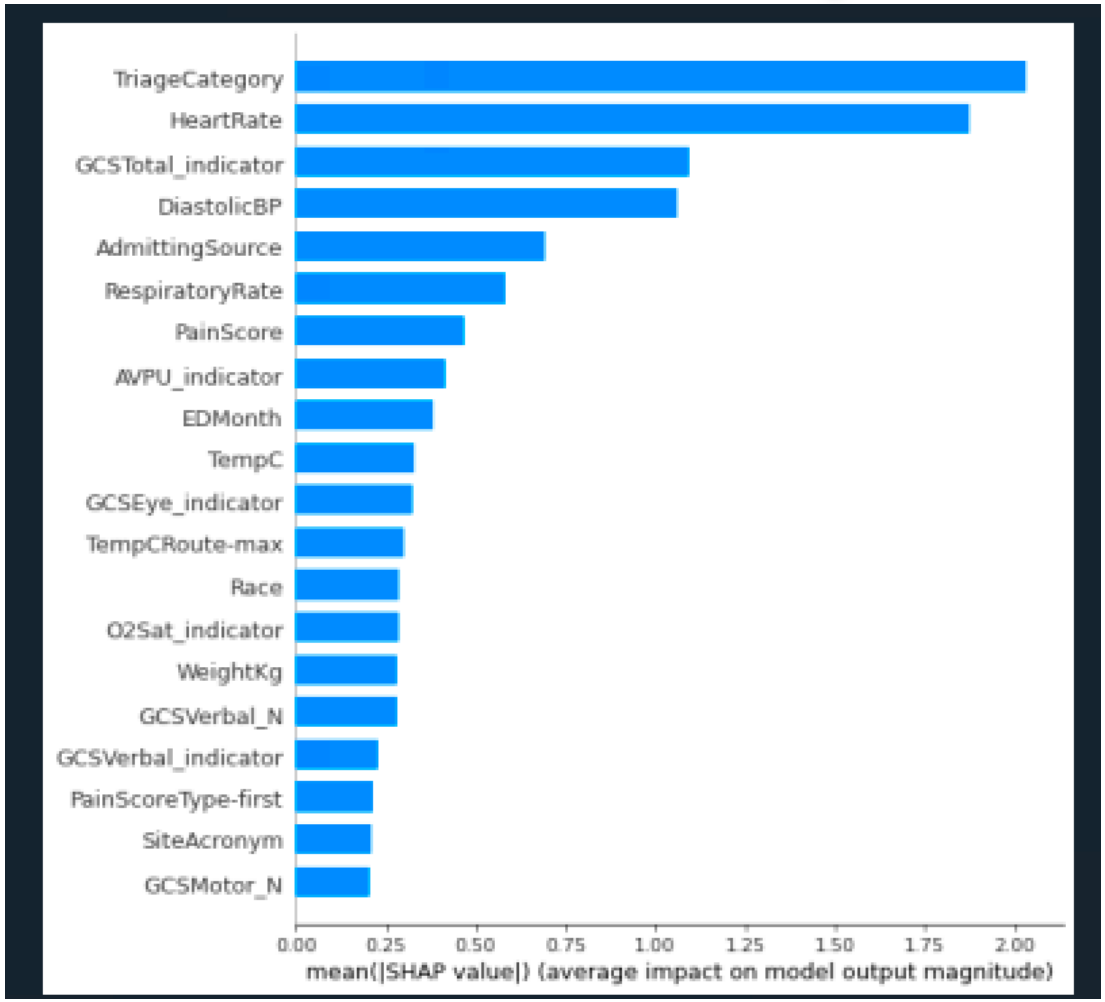
Pediatric mortality

Preliminary results of a large study

Features are highly correlated



Feature block importance



Pediatric mortality

For decision support

- **When does a clinician want an explanation for a risk score?**
- **What types of features are useful?**
  - **Demographic**: age, gender, …
  - **Previous physician actions**: triage category, tests ordered
  - **Physiologic measurements**: blood pressure, CO_2, ….
- **How to aggregate the features?**
  - Min, max, first, last, average measurements
  - Metabolic syndrome
  - Total number of hospital visits in the last year

Explanation is tricky

- **P(death) = f(age, height, weight, BMI, …)**
  - How important is each of these features?
- **Only in a model can you change one feature leaving the others fixed**
- **To explain the world, we want causality**

Measures of feature importance

- Correlation
- Replace with mean value
- Permutation importance
- Remove and retrain
- Partial Dependence Plot
- SHAP

Correlation

- $y = 0.5\, x_1 + 0.5\, x_2 + x_3$
  - where $x_1 = x_2$ generate data with $x_j \sim N(0,1)$
- **Correlation** (y, x_1) =
  - 1/sqrt(2) = 0.7
  - This is equivalent to
    - $y = x_1 + x_3$
    - where x_3 acts as noise and reduces the correlation

Zero out feature

- $y = 0.5\, x_1 + 0.5\, x_2 + x_3$
  - where $x_1 = x_2$ generate data with $x_j \sim N(0,1)$
- **Effect of replacing x_1 with its average**
  - $(0.5\, x_1 + 0.5\, x_2 + x_3) - (0.5*0 + 0.5\, x_2 + x_3)$
  - $= 0.5\, x_1$
  - *Makes an average difference of 0.5*

Permute feature

- $y = 0.5\, x_1 + 0.5\, x_2 + x_3$
  - where $x_1 = x_2$ generate data with $x_j \sim N(0,1)$
- **Effect of permuting x_1 on the prediction**
  - Same here as replacing with it's average

Retrain model

- $y = 0.5\,x_1 + 0.5\,x_2 + x_3$
  - where $x_1 = x_2$ generate data with $x_j \sim N(0,1)$
- **Importance as measured by removing and retraining**
  - 0

Partial Dependence Plot

- ***y = 0.5 x_1 + 0.5 x_2 + x_3***
  - where $x_1 = x_2$ generate data with $x_j \sim N(0,1)$
- **Partial Dependence Plot**
  - Look at effect of x_1, marginalizing over all the other values
  - $f_{PDP}(x_1) = (1/n)\ Sum_i\ (0.5\ x_1 + 0.5\ x_2^{(i)} + x_3^{(i)})$ $\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$
    $= 0.5\ x_1$
  - Effectively assumes features are independent

Shapley Values for feature importance

- **Local:** For one feature at one training point:

  Average over every subset of features ("coalition"):

  the change in prediction accuracy from replacing the removed features (including the target feature) with *baseline* values

- **Global:** Average the absolute values of these differences over all training points
- **Shapley Values** are a ***class*** of methods that
  - Have great axiomatic properties
  - Can be computed efficiently.

Shapley values, under linearity and independence

- **Local SV: For each feature *j*, for each observation *i*:**
  - Prediction accuracy difference between using the true value of x_j and *masking* the feature: using a random *baseline/background* value of it
    - $\varphi_{ij}(\boldsymbol{w}^T\boldsymbol{x}) = w_j(x_{ij} - E(x_j))$
- **Gobal SV: For each feature *j*:**
  - The average of absolute value of these accuracy differences over all training points
    - $\varphi_j(\boldsymbol{w}^T\boldsymbol{x}) = \Sigma_i |w_j(x_{ij} - E(x_j))|$

Shapley Values

- **Efficiency:** The feature contributions ϕ_j must add up to the difference of the prediction for x and the average.

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

- **Additivity:** For a game with combined payouts val+val$^+$ the respective Shapley values are as follows:
  - $\phi_j+\phi^+_j$
- **Dummy:** A feature j that has no effect has $\phi_j=0$
- **Symmetry:** For features i,j with identical effect, $\phi_i = \phi_j$

Shapley Values

- **Consistency:** if a model is altered so that the marginal contribution of a feature value increases (regardless of other features), the Shapley value also increases.

https://christophm.github.io/interpretable-ml-book/shap.html

Interventional SHAP

- $y = 0.5\, x_1 + 0.5\, x_2 + x_3$
  - where $x_1 = x_2$ generate data with $x_j \sim N(0,1)$
- **Same effect as replacing x_1 with its average**
  - $(0.5\, x_1 + 0.5\, x_2 + x_3) - (0.5*0 + 0.5\, x_2 + x_3)$
  - $= 0.5\, x_1$
  - *Makes an average difference of 0.5*

Conditional SHAP

- $y = 0.5\ x_1 + 0.5\ x_2 + x_3$
  - where $x_1 = x_2$ generate data with $x_j \sim N(0,1)$
- **Same effect as replacing x_1 and x_2 with their average (since they move together)**
  - $(0.5\ x_1 + 0.5\ x_2 + x_3) - (0.5*0 + 0.5\ 0 + x_3)$
  - $= 1.0\ x_1$
  - *Makes an average difference of 1.0*

Cotenability

It generally doesn't make sense to change one feature without others changing as well.

- *breaks the correlation structure of the features*

- What is the effect of changing the height in inches, but not the height in centimeters?
- What is the effect of changing rainfall, holding weather constant?

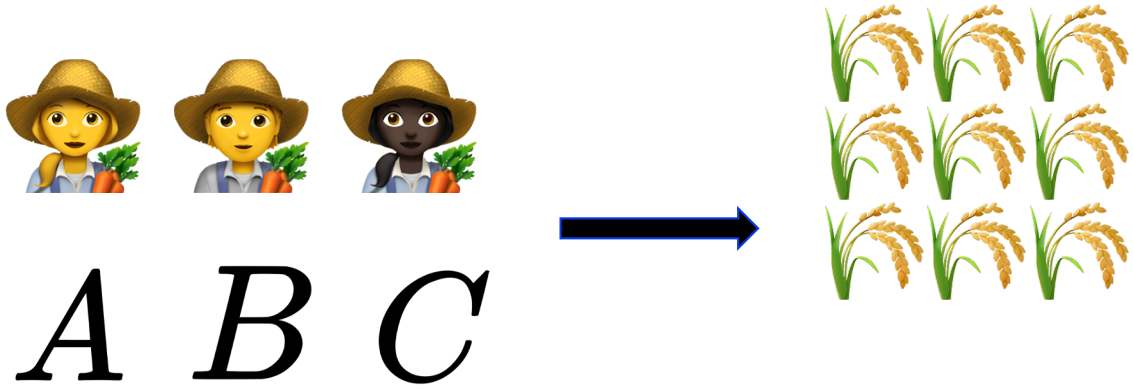
Shapley and SHAP

Lyle Ungar

Scott Lundberg and Su-In Lee

Shapley values example

- Alice, Bob, and Celine are farmers growing wheat



What is the marginal contribution of each farmer?
- given all possible coalitions of farmers

Tony Liu

Predict income > $50k

age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous. **capital-loss:** continuous.
hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-

Case study: predict income

https://github.com/slundberg/shap/blob/master/notebooks/tree_explainer/Census%20income%20classification%20with%20XGBoost.ipynb

0.34 Relationship

0.34 Education-Num

0.23 Age

0.23 Hours per week

0.22 Sex

0.22 Capital Gain

-0.20 Marital Status

0.15 Capital Loss

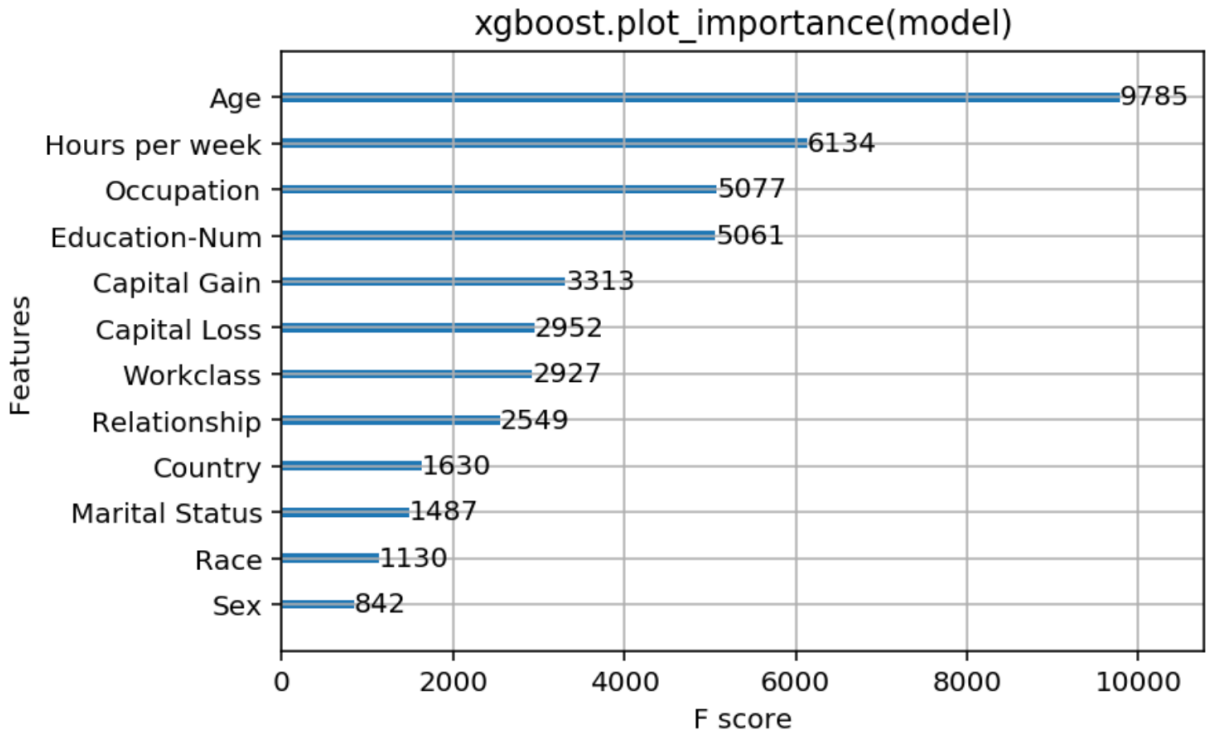
0.08 Occupation

0.07 Race

0.02 Country

0.05 Workclass

Feature importance - XGboost

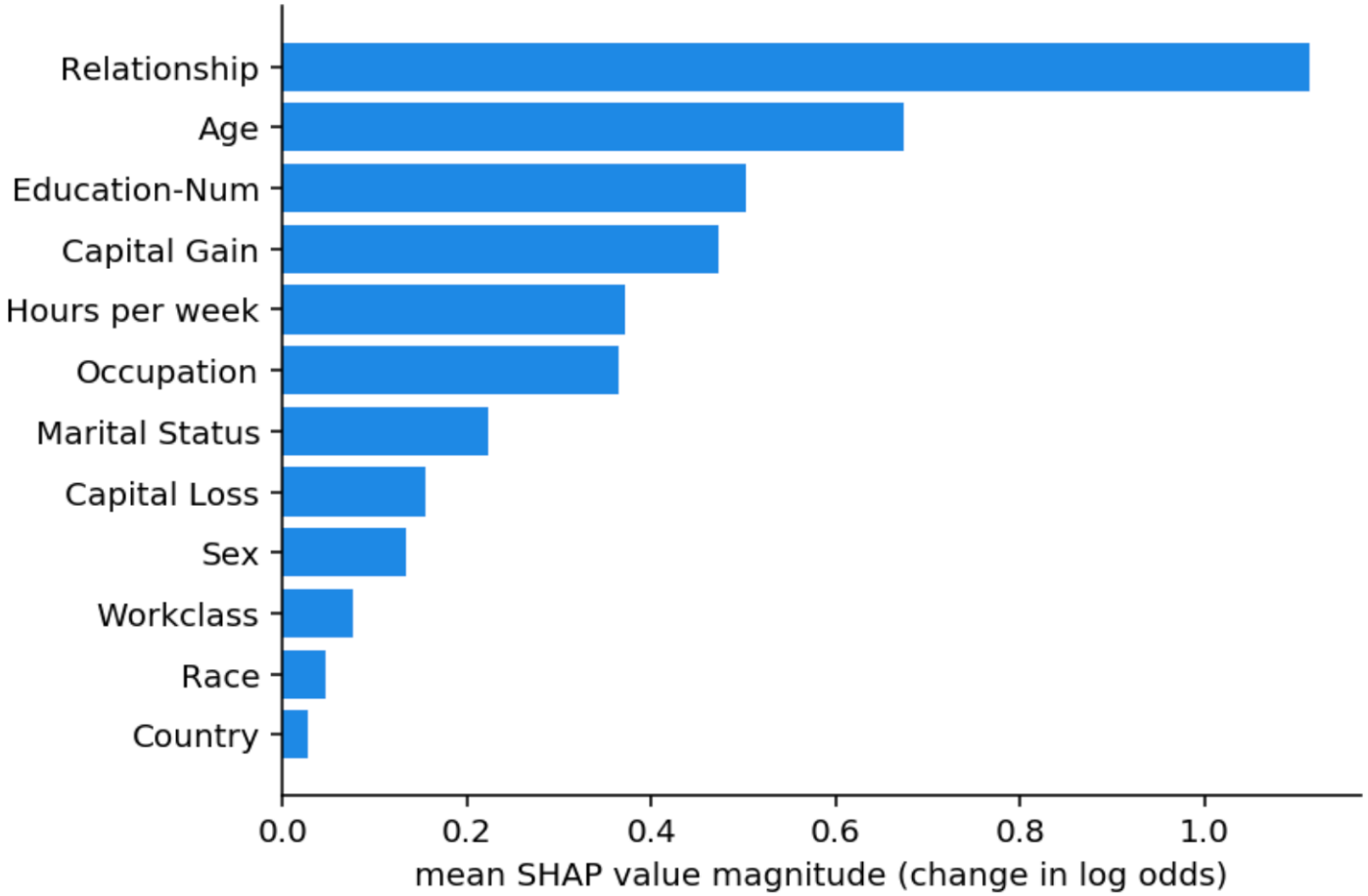


Results of running xgboost.plot_importance(model) for a model trained to predict if people will report over $50k of income from the classic "adult" census dataset (using a logistic loss).

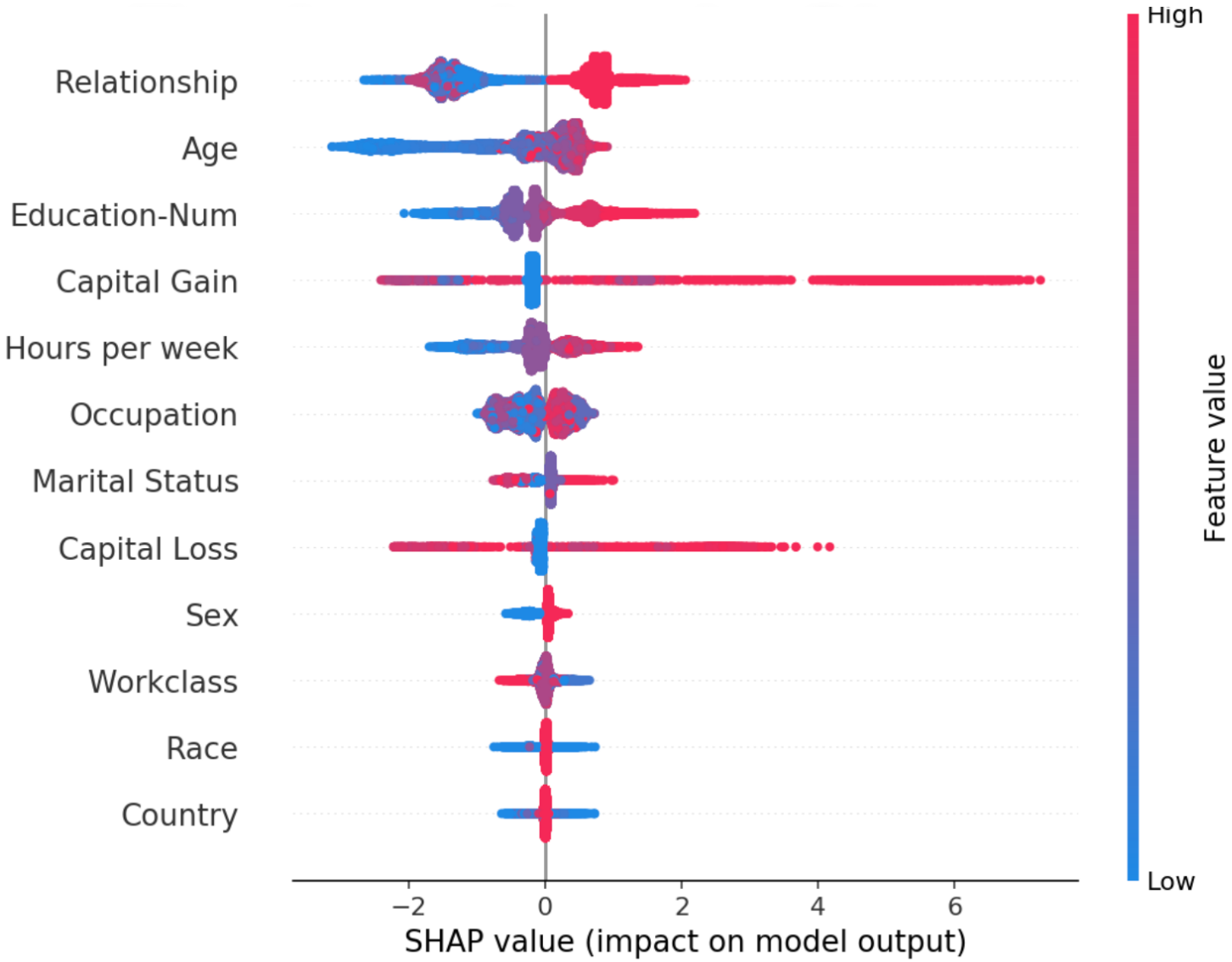
Different methods give very different results

https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27

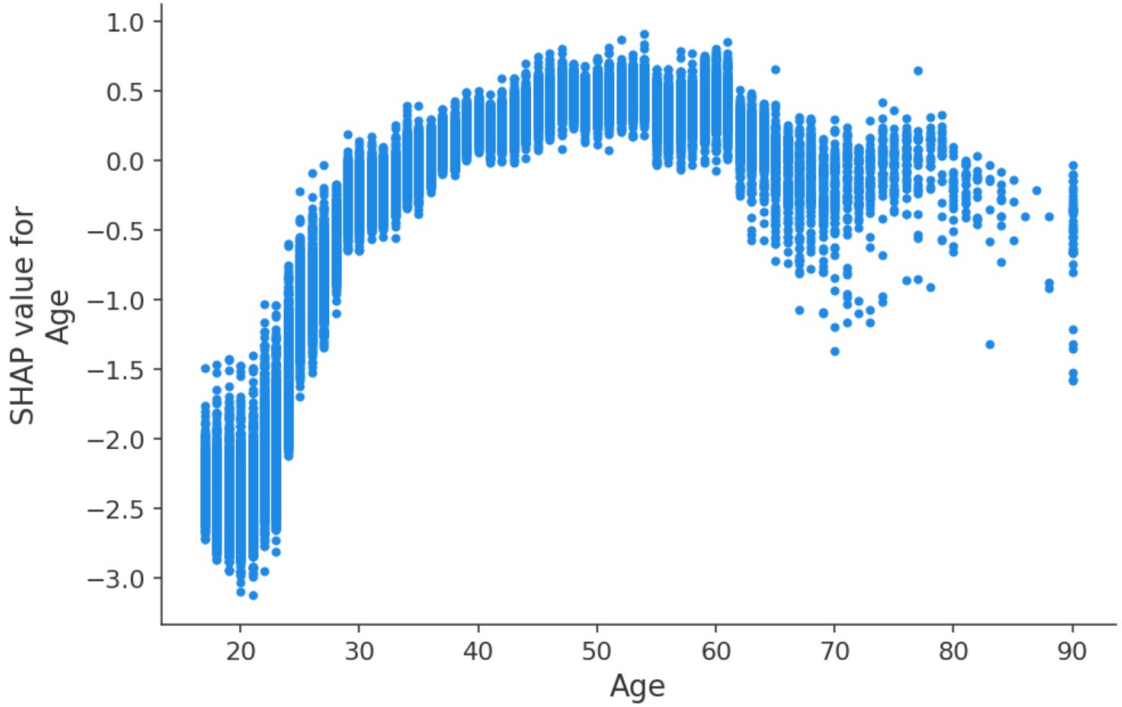
Feature importance - SHAP



https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27


Relationship
Age
Education-Num
Capital Gain
Hours per week
Occupation
Marital Status
Capital Loss
Sex
Workclass
Race
Country
−2
0
2
4
6
SHAP value (impact on model output)
High
Feature value
Low

Partial Dependence Plot

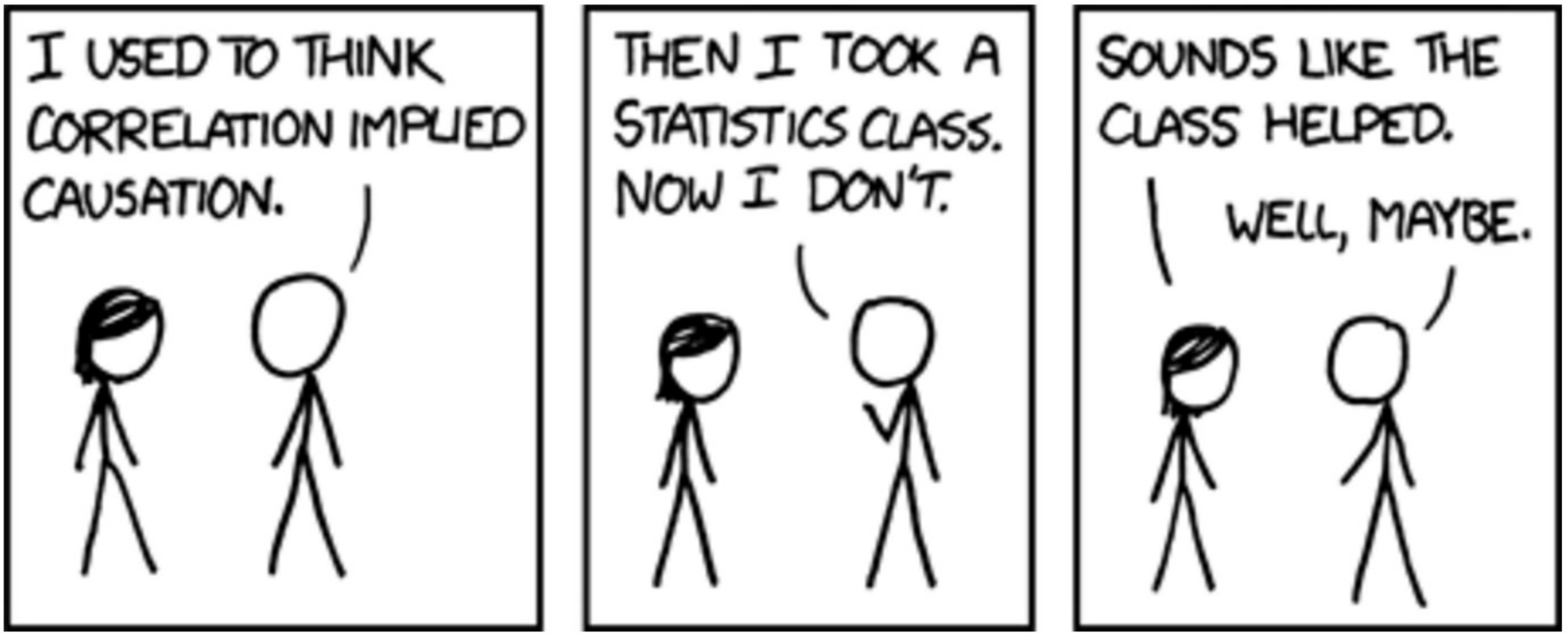


https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27

What you should know

- **Feature importance often measured as**
  - Effect on y_i of changing feature x_{ij} from average to its value *holding other features fixed*
    - Or zeroing it out (LIME)
  - But this ignores the correlation between features
    - E.g., height, weight, BMI
- **Explaining the model vs. explaining the world**

What is Causality?



https://xkcd.com/552/

High correlation between…

- **Radio ownership and population in insane asylums**
  - England, 20th century
- **Daily ice cream consumption and rape incidents**
  - US, 21st century
- **Stork population and babies born**
  - Germany, 20th century

Storks and Babies

New evidence for the theory of the stork.

- **Höfer T, Przyrembel H, Verleger S.**
- Paediatr Perinat Epidemiol. 2004 Jan;18(1):88-92.

Data from Berlin (Germany) show a significant correlation between the increase in the stork population around the city and the increase in [baby] deliveries outside city hospitals (out-of-hospital deliveries). However, there is no correlation between deliveries in hospital buildings (clinical deliveries) and the stork population. The decline in the number of pairs of storks in the German state of Lower Saxony between 1970 and 1985 correlated with the decrease of deliveries in that area.

Causality and Regression

- $y = c_1 x_1 + c_2 x_2$
  - y : crop yield
  - x_1: temperature
  - x_2: rainfall

Do higher temperatures cause higher crop yields?

- **Increased temperature decreases yield?**
  - $y = -0.1 x_1$
- **Increased temperature increases yield?**
  - $y = 0.2 x_1 + 0.4 x_2$

Causality and feature selection

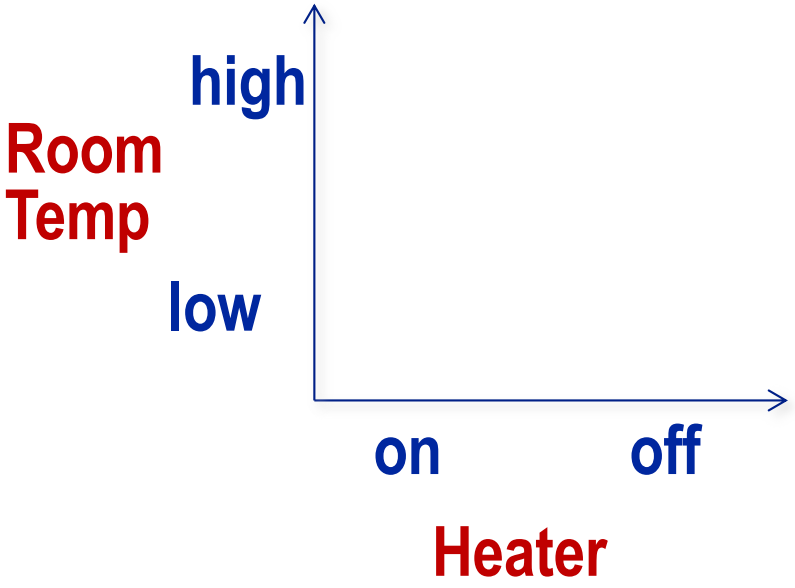
- $y = c_1 x_1 + c_2 x_2 + c_3 x_3$
  - y : customer lifetime value
  - x_1: customer car value
  - x_2: customer house value
  - x_3: customer mortgage payment
- **Stepwise regression selects only x_3**
  - What does this mean?
  - Is this a problem?

Causality and Correlation

- $y = c_1 x_1 + c_2 x_2$
  - y : satisfaction with life
  - x_1: income
  - x_2: county income
- **Regression**
  - $y = .587\ x_1$ being richer makes you happier
  - $y = .047\ x_2$ having richer neighbors is good?
  - $y = .587\ x_1 - .013\ x_2$ having richer neighbors makes you less happy

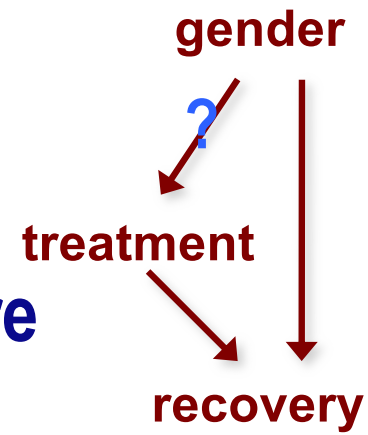
Feedback complicates causality

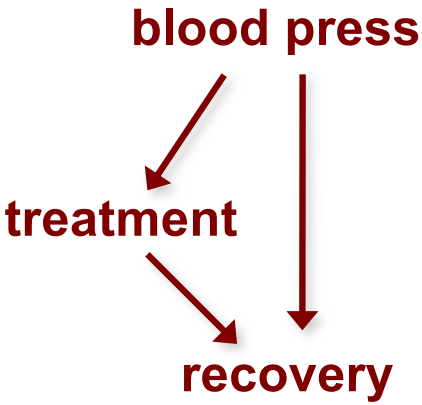
- Room temperature as a function of whether the heat is on

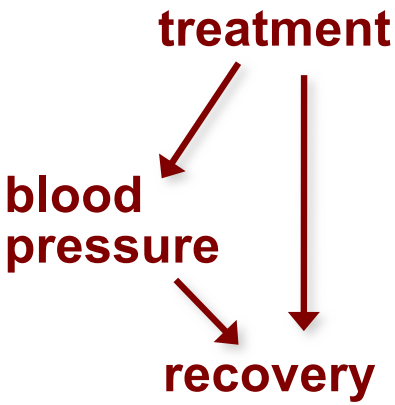


Causality Matters

- Is the treatment more effective for men than women?
- Does treatment cause high blood pressure or high blood pressure cause treatment?







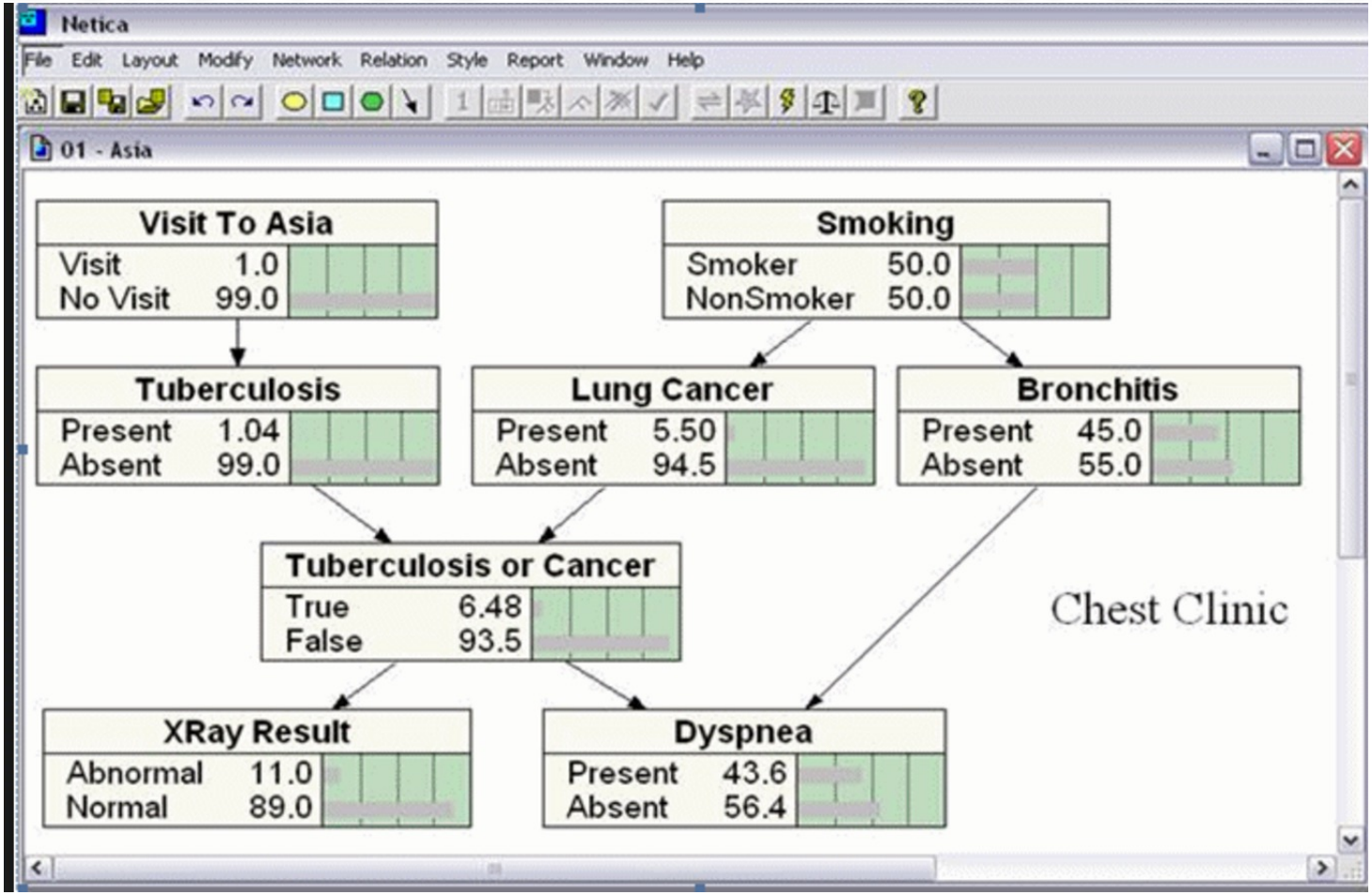
Causality is usually impossible to infer from data

Questions

- **Among patients with pneumonia admitted to a hospital, those with asthma had a lower chance of dying**
  - What might be going on?
  - Is this a problem?

Rich Caruana

Belief Nets often model causality



Can add decision ('do') nodes to a Belief Network

What you should know

- **Many explanations of feature importance**
  - Replace feature with average or permute it
  - be careful when interpreting correlated features
- **Machine learning finds correlation – not causality**
  - Finding causality requires experiments
    - Or talking to experts
- **Actions can be added to most of our models**
  - Then need to both *learn the model* and *select the optimal action*
  - Exploration in RL is experimentation