Announcements

- **We will "grade" quizzes on Saturday**
  - So please finish by Friday, midnight
- **We will keep office hours until the final**
  - As best we can
- **Extra review session Monday**
  - Usual class time and location
- **Final exam next Wednesday**
  - Same structure as midterm

Correlation, Causality and Variable Importance

Lyle Ungar

Different types of explanations
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**
  - Reinforcement learning
  - Predictive modeling
  - Causal modeling

Model Interpretation

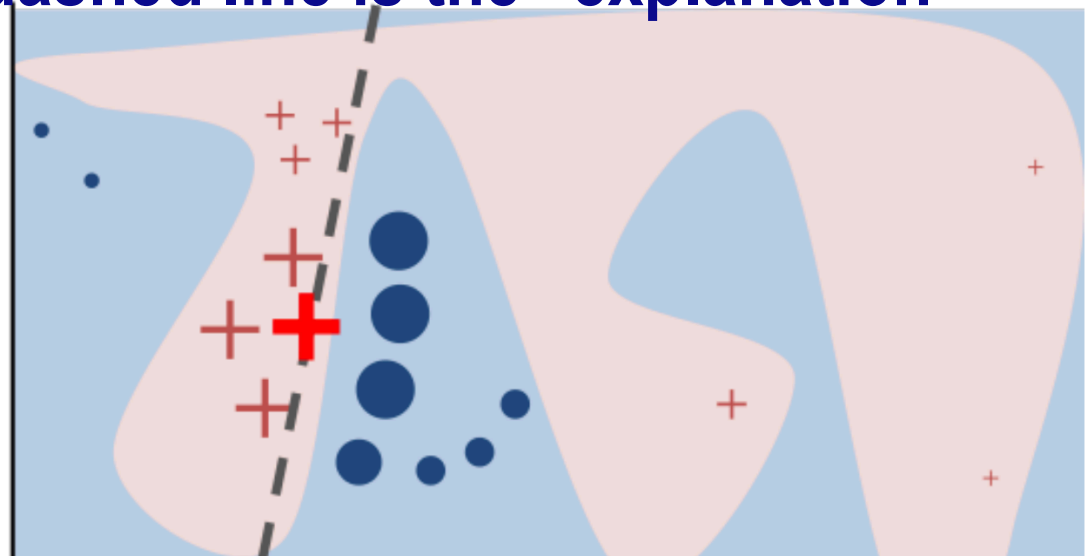
- **Explain the world**
  - Effect of changing an input on the output
- **Explain the model**
  - What input most affects this latent variable or prediction?
  - What predictions does this feature or latent variable most affect?
- **Explain a prediction**
  - Decision tree: path taken
  - Regression: largest values of $w_j(x_{ij}-\mu_j)$
  - **LIME** (Local Interpretable Model-Agnostic Explanations)

Explain a prediction

- **Why did you predict this value or label for x_i?**
- **Linear regression**
  - E.g. : $y = c_1 x_1 + c_2 x_2 + c_3 x_3 + c_4 x_4 + c_5 x_5 + \ldots$
  - **Influence** of x_{ij} is $c_j(x_{ij} - avg(x_j))$
- **Clustering**
  - Show nearby ***x*** and its label

LIME algorithm

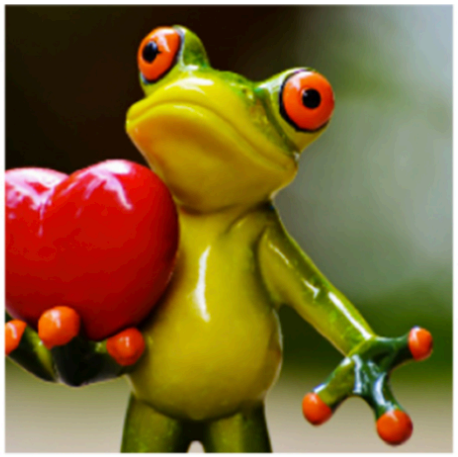
- Sample instances near the target instance +
- Predict labels using full model
- Fit a sparse locally weighted regression
- The dashed line is the “explanation”



LIME

Why Should I Trust You?":
Explaining the Predictions of Any Classifier
Ribeiro Singh & Guestrin

- **Do local perturbations to x**



Original Image

Interpretable Components

- **Fit locally weighted model**

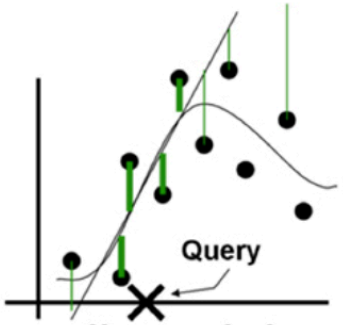
https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

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

LIME explains alternate predictions

P() = 0.54

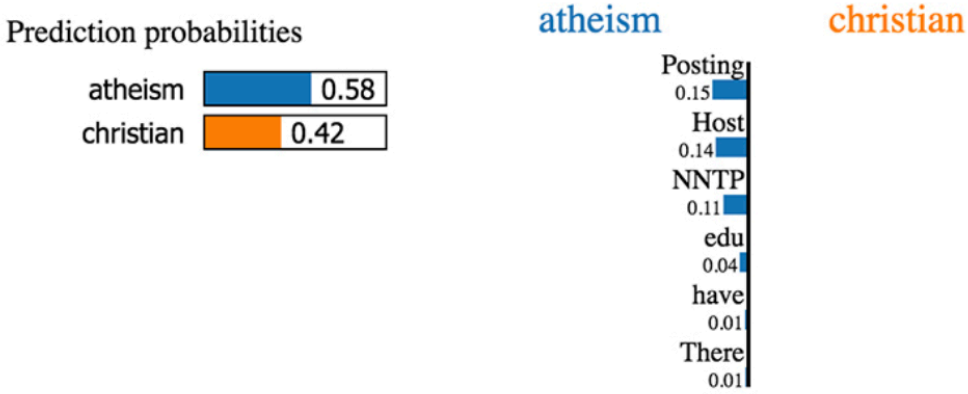
P() = 0.07

P() = 0.05

https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

LIME

- Works on SVMs, Random Forests, Nnets …
- Works on text or images



https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

Explain a model

- **Interpretation**
  - Find items closest to the cluster center
  - Find words closest to a predicted vector embedding
  - Variable importance
- **Probing**
  - What labels can I predict with this latent variable?
- ***argmax*$_{\boldsymbol{x}}$ *f(**x**)* for hidden nodes or outputs**
  - Which input (image, document …) maximizes *p(Y=y)?*

Regression Feature Importance

- **The accuracy loss from leaving out a variable when building a model**
  - What is the importance of x_1 in

    $y = c_1 x_1 + c_2 x_2 + c_3 x_3 + c_4 x_4 + c_5 x_5$

    with $x_1 = x_2 = x_3 = x_4$
- **The accuracy loss from pegging a variable to its average value in a trained model**

Random Forest Feature Importance

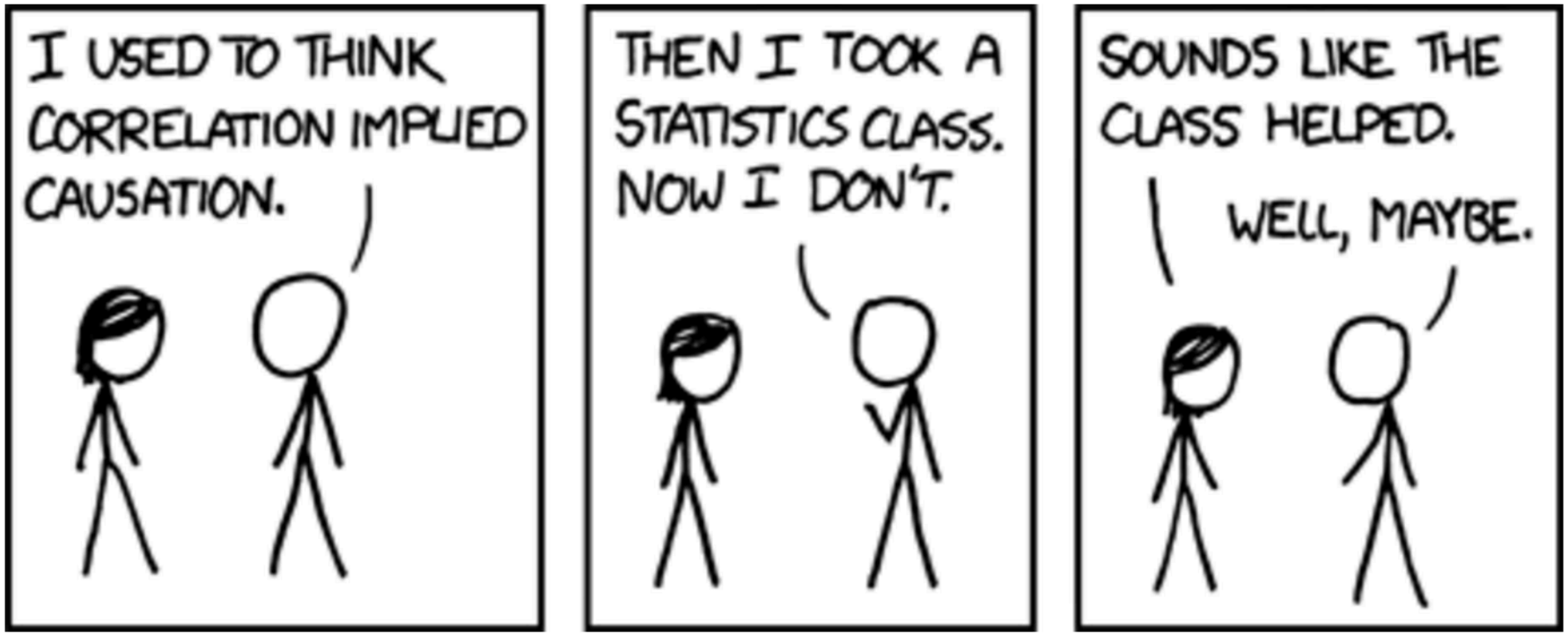
- **Find test set error, *Err***
- **Permute a variable x_j, find new test set error, Err_t**
- **Variable importance is the difference, (Err – Err_t) divided by the standard error**

From the R package for Random forests

What you should know

- **Feature importance often measured as**
  - Effect on y_i of changing feature x_{ij} from average to its value
    - Or zeroing it out (LIME)
    - Holding other features fixed
  - Effect on total error of randomly permuting a feature, $\boldsymbol{x}_j$
- **Many ways to explain models**
  - Many of them do not reflect causality
  - E.g. change only one of a set of correlated features
    - height, weight, BMI

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?

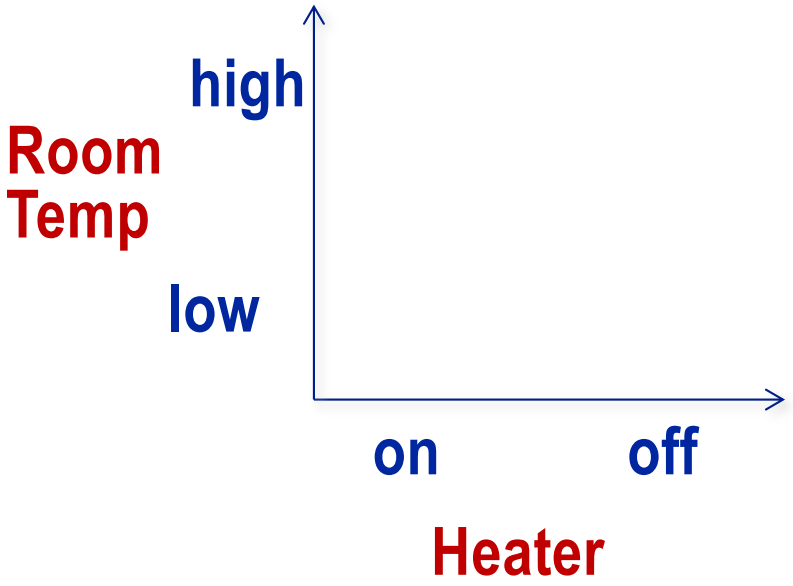
- **Increased temperature decreases yield?**
  - $y = -0.1 x_1$
- **Increased temperature increases yield?**
  - $y = 0.2 x_1 + 0.4 x_2$

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?

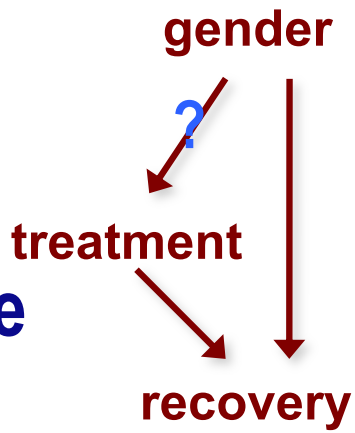
Feedback complicates causality

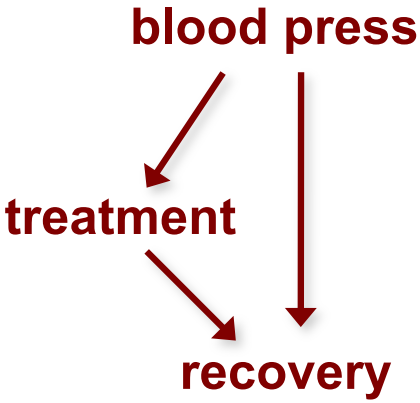
- Room temperature as a function of whether the heat is on

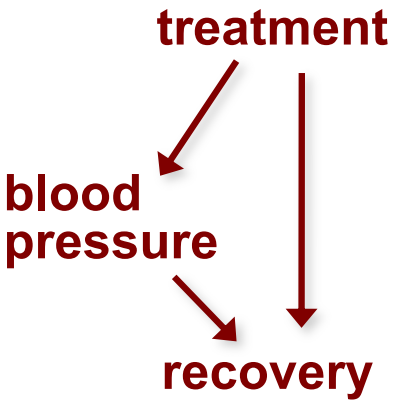


Causality Matters

- 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

Questions

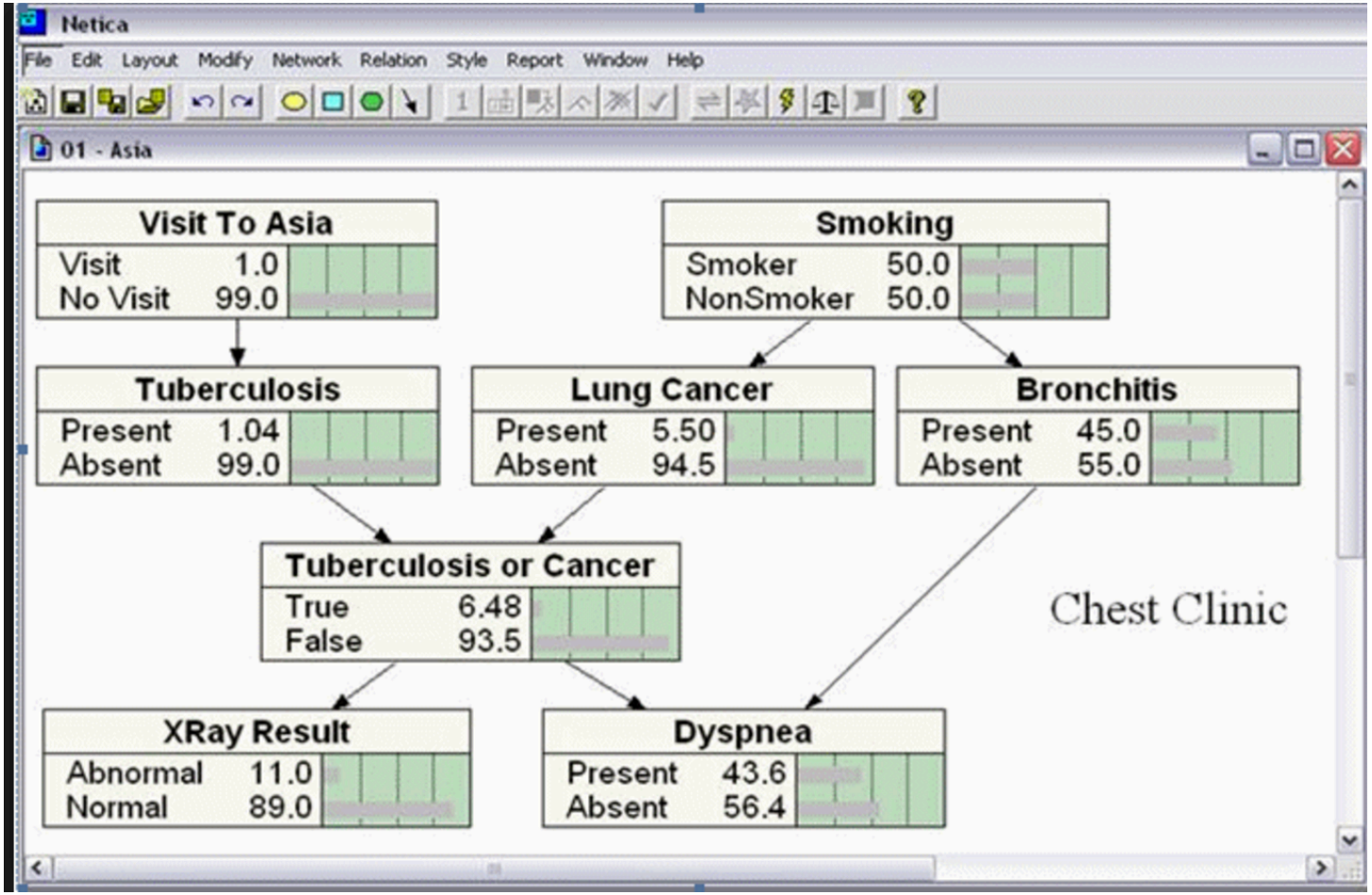
- **Elastic net selects zip code, house price, and owning a second house, but not income, as predictors of buying a boat.**
  - What might be going on?
  - Is this a problem?
- **Predicted polymer quality does not depend on the temperature of the reactor**
  - What might be going on?

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

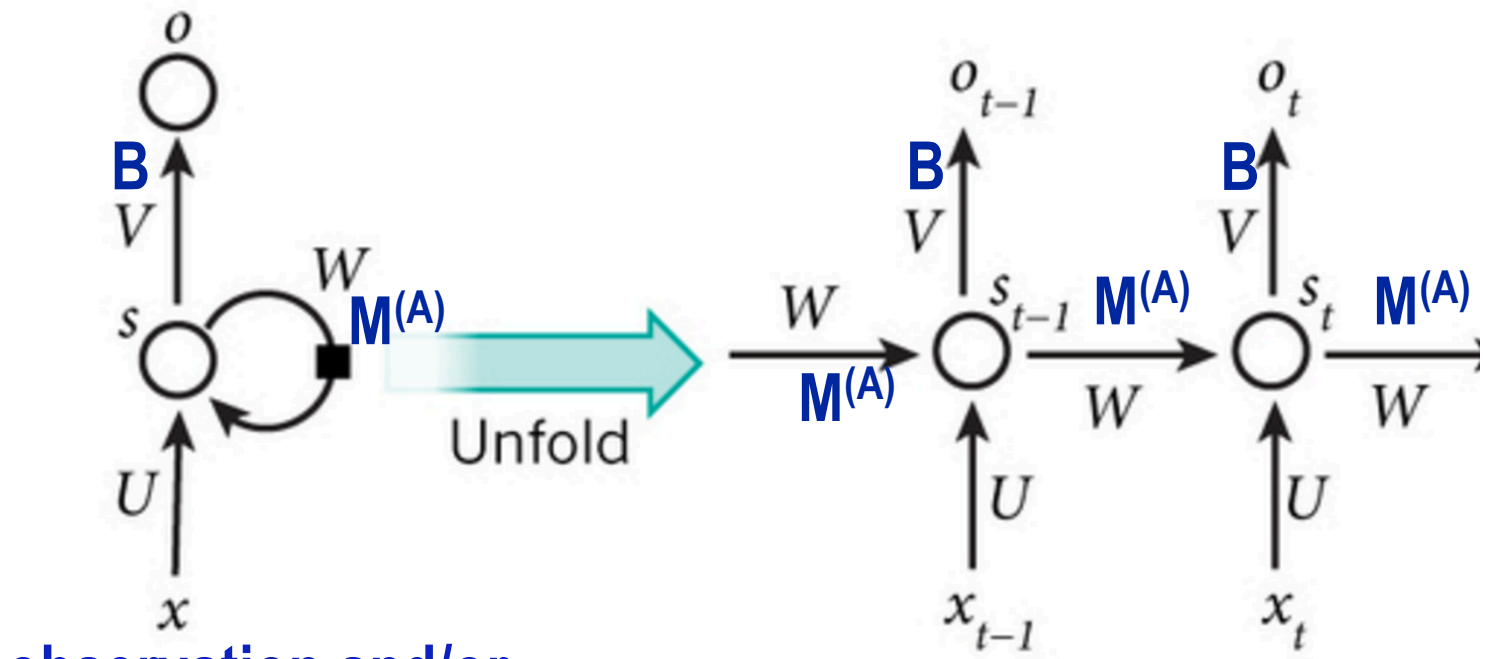


Can add decision ('do') nodes to a Belief Network

Can add actions to models

- Markov Process → Markov Decision Process (MDP)
- HMM → Partially Observable Markov Decision Process (POMDP)
- Neural Network → Neural Network

MDP/POMDP or Recurrent Net



What you should know

- **Machine learning finds correlation – not causality**
  - Finding causality requires experiments
    - Or talking to experts
  - But correlations suggest possible causality
  - Interpretable models help identify incorrect causality
- **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