Intervention and Causality

Volker Tresp
2015

Interventions

- In most of the lecture we talk about prediction
- When and how a can use predictive models for interventions, i.e., to estimate the results of interventions
- Let's assume that I learned a model

$$y = f(a, x)$$

where $y \in \{0, 1\}$ is the binary outcome (response) ($y = 1$ is good, healthy), $a \in \{0, 1\}$ is the action (treatment, exposure variable, risk factor, explanatory variable), x is some other contextual input, should I do

$$a_{opt} = \max_{a} f(a, x)$$

i.e., decide for the best action a given a learned the model (we don't consider modelling errors here)?

- Unfortunately, this is not always correct: the answer depends on the underlying causal structure!

- (By convention: $y = 1$ means healthy if $a = 1$ means that the treatment was applied to a person with a disease; in contrast, in exposure studies $y = 1$ means that the person gets the disease and $a = 1$ means that the person was exposed)
- This lecture tried to help you avoid embarrassment: "Data Miners found: Taking your breaks outside of the building causes lung cancer!"

Causality

- These slides are about dealing with intervention and causality and not learning causality from data
- It is also not about the best *predictive* model but about a model with which one can *evaluate and optimize* intervention
- This discussion also touches on collinearity (correlated inputs): if you detect collinearity you should think about which causal structure might have caused it and then use the results presented here to draw the right conclusions!

Data Set 1: Independent Inputs

- The first model (Model 1A) has a (action) and x as inputs and the learned regression is

$$\widehat{y} = -0.18a + 0.82x$$

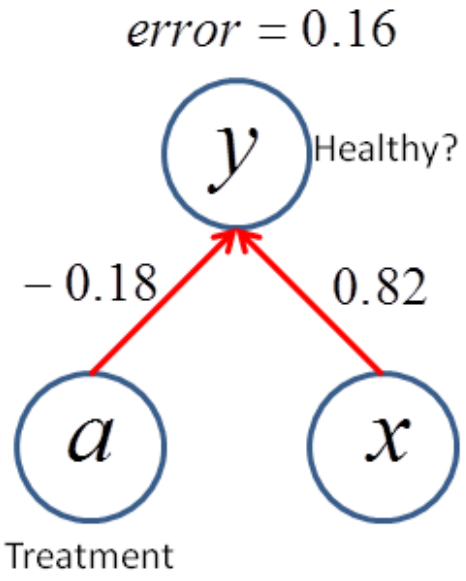
- In Model 1B I remove x and get

$$\widehat{y} = -0.18a$$

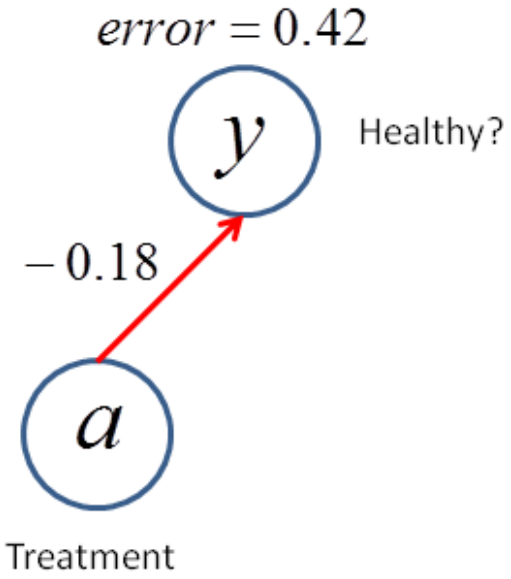
- Both models tell me that the medication is harmful and the recommendation would be not to use it

DataSet 1

Model 1A



Model 1B



Data Set 2: Dependent Inputs

- Now we use Data Set 2
- Model 2A has a (action) and x as inputs and the learned regression is, *as before*,

$$\widehat{y} = -0.18a + 0.82x$$

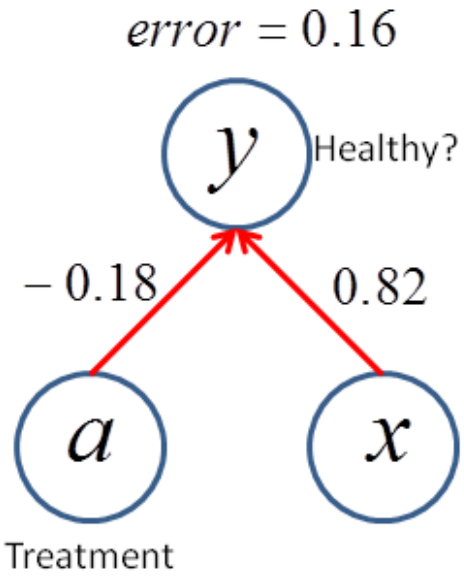
- In Model 2B I remove x and get

$$\widehat{y} = 0.42a$$

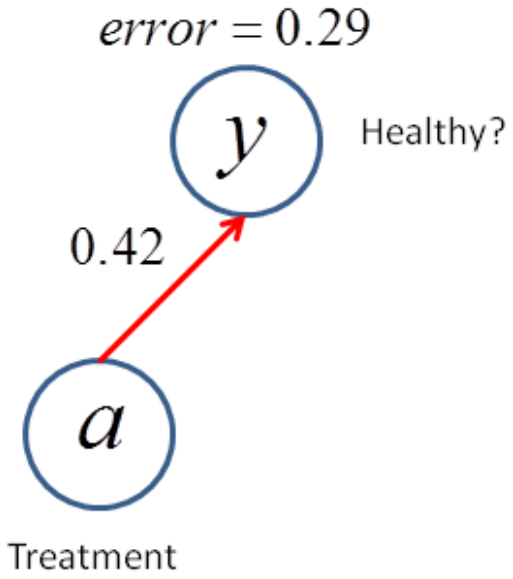
- Now the result is dramatically different: Now it seams that a has a positive influence on health!
- The difference is that a and x were independent in Data Set 1, but are dependent in Data Set 2.
- So what is the right answer with Data Set 2?

DataSet 2

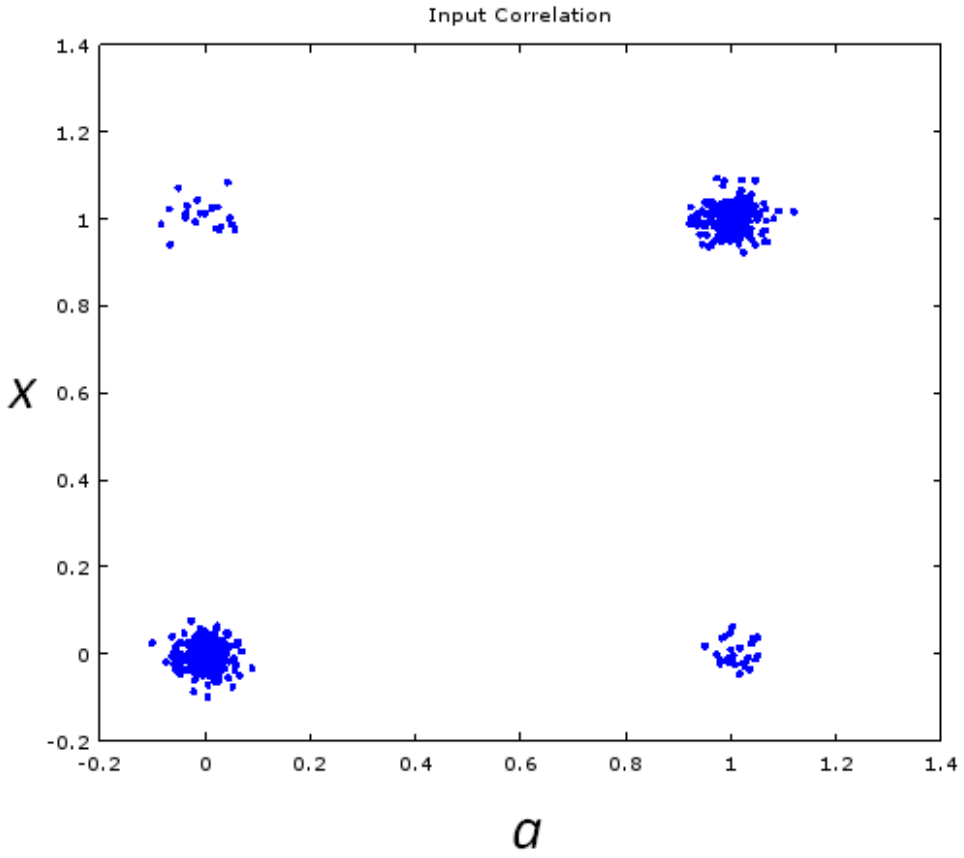
Model 2A



Model 2B

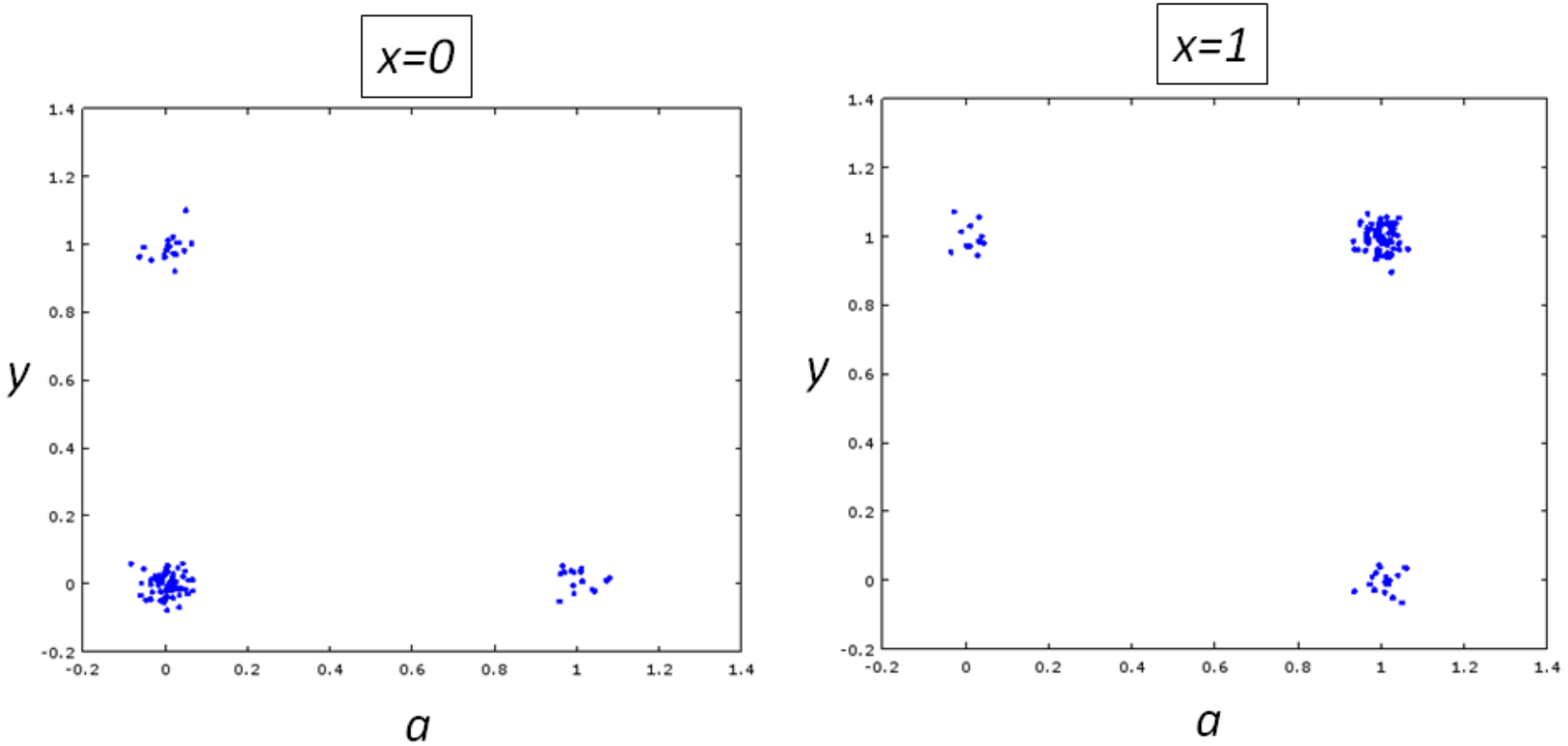


Data Set 2: *a* and *x*



We see the strong correlation between *a* and *x*

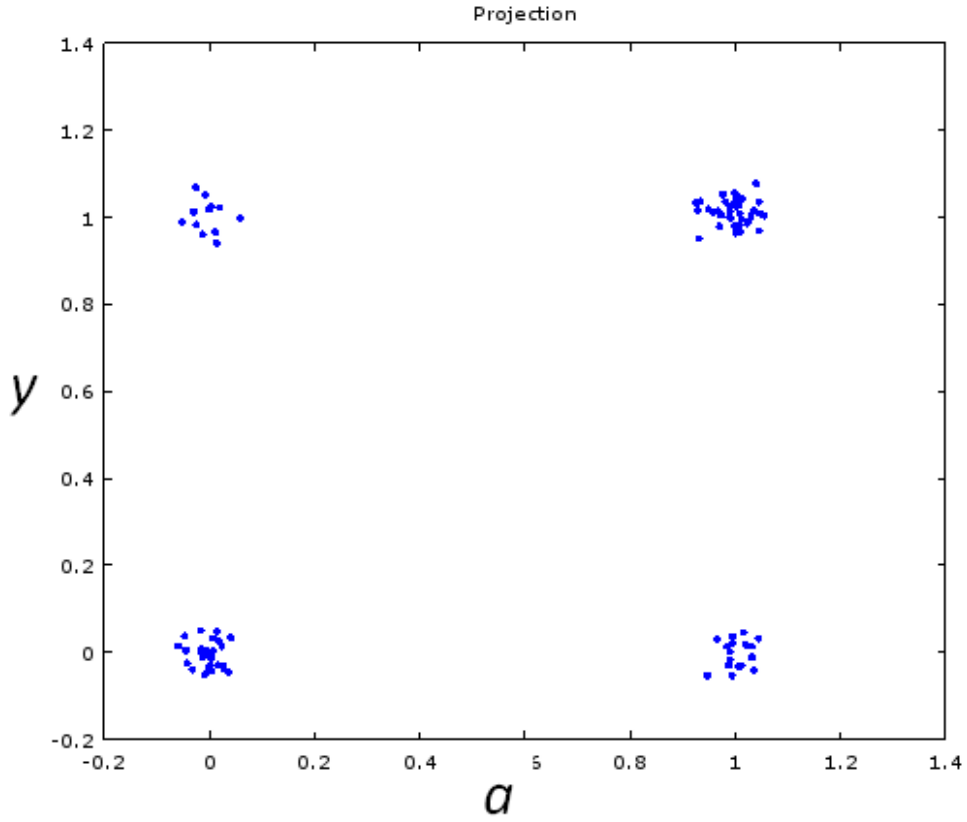
Data Set 2, Model 2A: inputs *a* and *x*



- Left: if *x=0*, with *a=0* there is a chance that the patient gets healthy
- Right: if *x=1*, all patients get healthy when *a=0*

So in both cases, the medication should NOT be given!
(here is the following the data are really 0/1, noise was added for illustration)

Data Set 2, Model 2B: input *a* only



If *a=0*, there is a greater chance that the patient gets healthy, so the medication should be given!

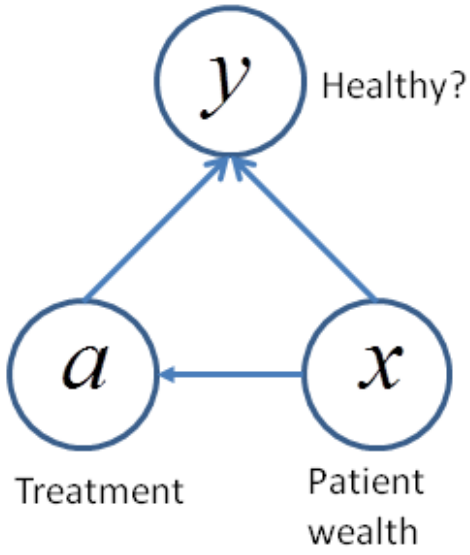
Recipe

- **Rule1:** You need to include all variables that are causes of a and influence y (confounders). In particular if a is decided by a person you need to include all information that the person has used!
- **Rule 2:** *Do not* include variables that are effects of a and influence y
- **Rule 3:** Recommended: Include variables that affect y but are independent of a, in particular if $f(\cdot)$ is nonlinear
- Let's consider a few examples

x is a Cause of a

- x might be the wealth of the patient. For example it might be that only wealthy people can afford the expensive treatment and that wealthy people have a healthier lifestyle and thus get healthier anyways and not because of the treatment.
- We need to apply Rule 1 and include x in the model. Model 2A is correct and the medication should not be given!
- A variable x that influences both treatment and outcome is called a **confounding variable**. Not taking the confounding variable into account results in confounding bias!

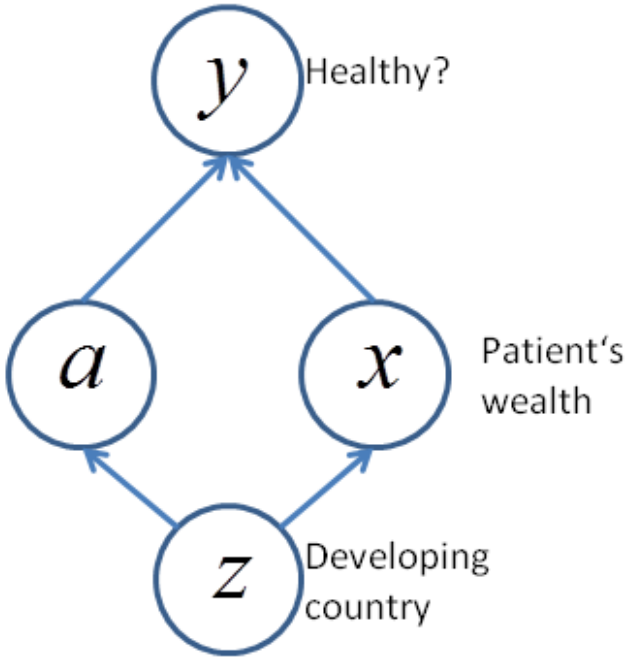
x is a Confounder



a and x have a common unknown cause z

- There is a common unknown cause z that explains the correlation and which does not have an effect on y
- If x is not included in the model, then z is a confounder that both influences a and y (via the unknown x) (back door criterion). When we include x, the influence of z on y is removed, thus x needs to be included (Model 2A). Again, the medication should not be given

z is a confounder if *x* is unknown (back door criterion)



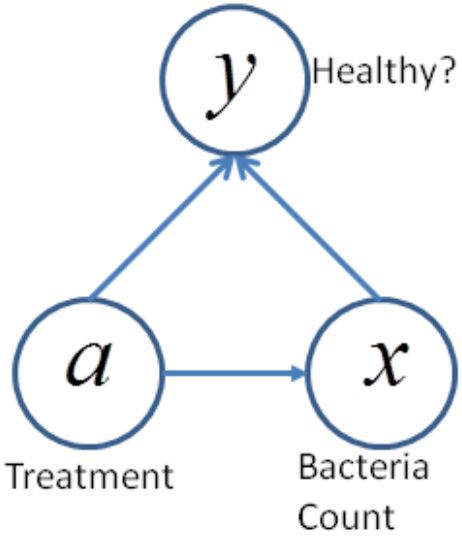
Dealing with Confounders

- When x is a confounder, then there are two options. Let's consider again that wealth is the confounder
- First option: As discussed, we include the confounder "wealth" as input in the model and use Model 2A
- Second option: We train two models without x as input. The first model only sees the data from the poor patients and the second model only sees the data from the rich patients. In both models we see that the medication is harmful. This strategy is called **stratification** and the two groups are called **strats**
- Dealing with confounders is called: "adjusting or controlling for confounders". The set of confounders to be included is called an admissible (or sufficient) set of variables for adjustment.

a is a Cause of x

- a is the cause of x. x might be the bacterial count measured shortly after the treatment is applied and which is much improved through the treatment
- We need to apply Rule 2. Now Model 2B is the correct model and the medication should be given!

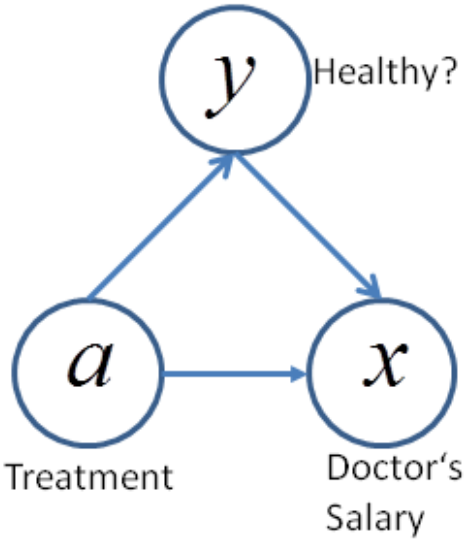
x is an effect and need to be removed from the model



a and y are both Causes of x

- Here, both a and y are causes of x. Here x might be the doctor's salary which depends on the outcome and the expensive treatment.
- Again we need to apply Rule 2 and remove it from the model and Model 2B gives the right answer: The medication should be given!
- Somewhat against intuition, considering less information is helpful in a causal analysis! The fact that Model 2A gives the wrong answer in this situation is called **Berkson's paradox**

Berkson's paradox



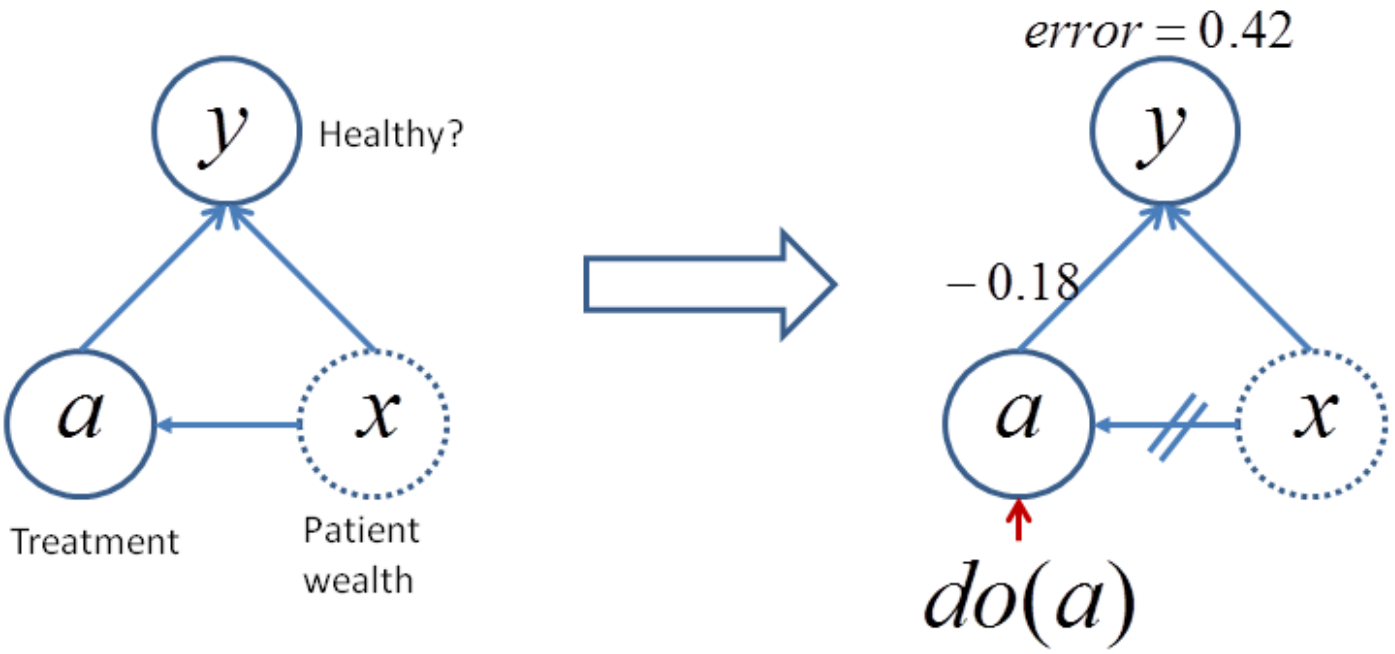
Recipe: Not Always Obvious

- Consider smoking, coughing, lungCancer, cancerGeneMutation. We want to do a causal analysis, if smoking causes lung cancer
- We observe patients with these properties and it would not be clear what came first and what came last
- Our prior knowledge tells us, that the cancerGeneMutation potentially made us smoke and is a cause for lung cancer, so we need to include it as inputs in the model
- Our prior knowledge tells us, that the coughing was caused by smoking and we should NOT include it as inputs in the model (we should also not include coughing if there are other *additional* causes of coughing, not related to smoking)
- (Of course, if we want to *predict* lung cancer, then we can include coughing in the model!)
- A causal evaluation is greatly simplified if indeed, temporal information is available! If you are interested in time series: The Granger causality test is a statistical hypothesis test for determining whether one time series is useful in forecasting another

x is a Hidden Confounder

- Let's assume that x is a confounder, but I do not know or cannot measure x. Then x is called a hidden confounder
- But if I use Model 2B (without x) I obviously get the wrong answer! This is a serious problem!
- The solution is to obtain data from experiments, where I manipulate a independent of x with $do(a)$! If a is chosen randomly, then this is called a randomized study!

The Intervention makes *a* and *x* independent



Probabilistic Model for a Randomized Study

- In a randomized study, the causal link from x to a can be removed and the link from $do(a)$ to a is introduced
- Probabilistically, the observed probabilistic model without intervention is

$$P(x)P(a|x)P(y|a,x)$$

and conditioning on a gives

$$P(y|a) = \frac{1}{P(a)} \sum_x P(x)P(a|x)P(y|a,x)$$

- In contrast the model where we manipulate a is $P(x)P(y|a,x)$

$$P(y|do(a)) = \sum_x P(x)P(y|a,x)$$

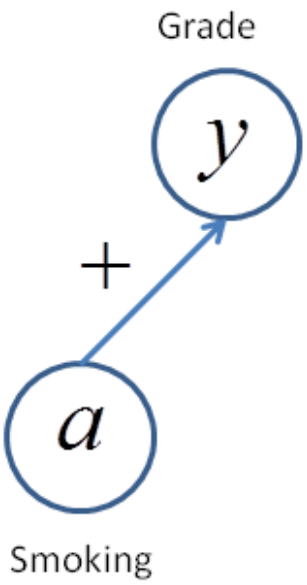
Note that $P(a|x)$ drops from the equation in the model when I manipulate a

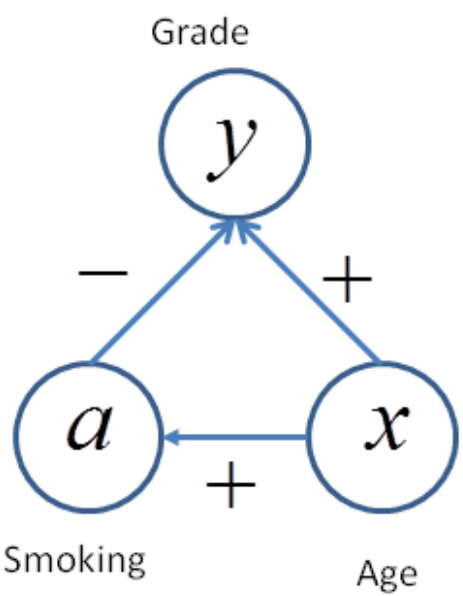
If I cannot Do Randomized Studies

- Again, try to include all possible confounders in the model!

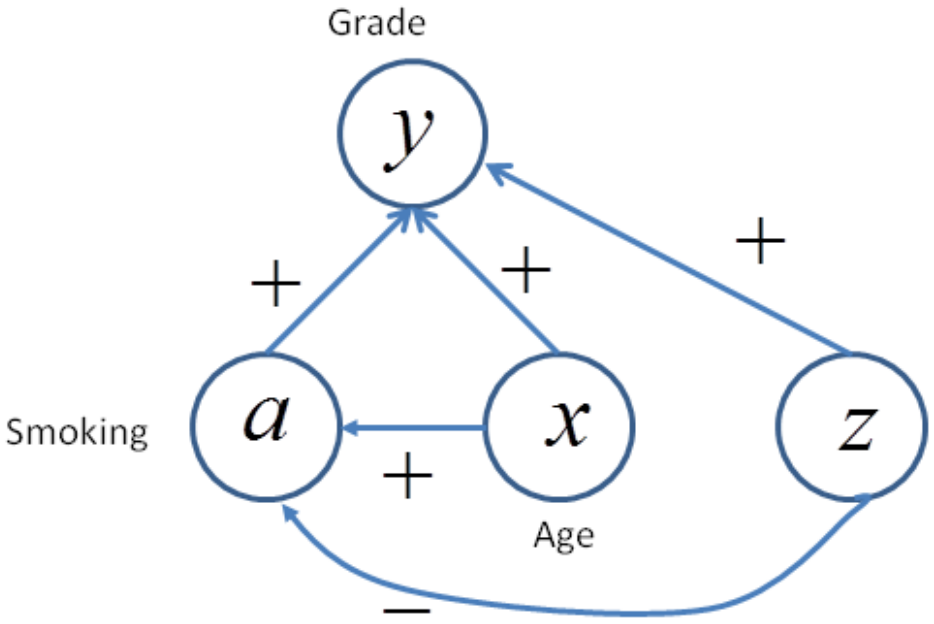
Simpson's Paradox

- When confounders are included sequentially, the intermediate results might look quite confusing





SIMPSON'S PARADOX



Which factors to include and which not?

... and so on ... until all confounders are included

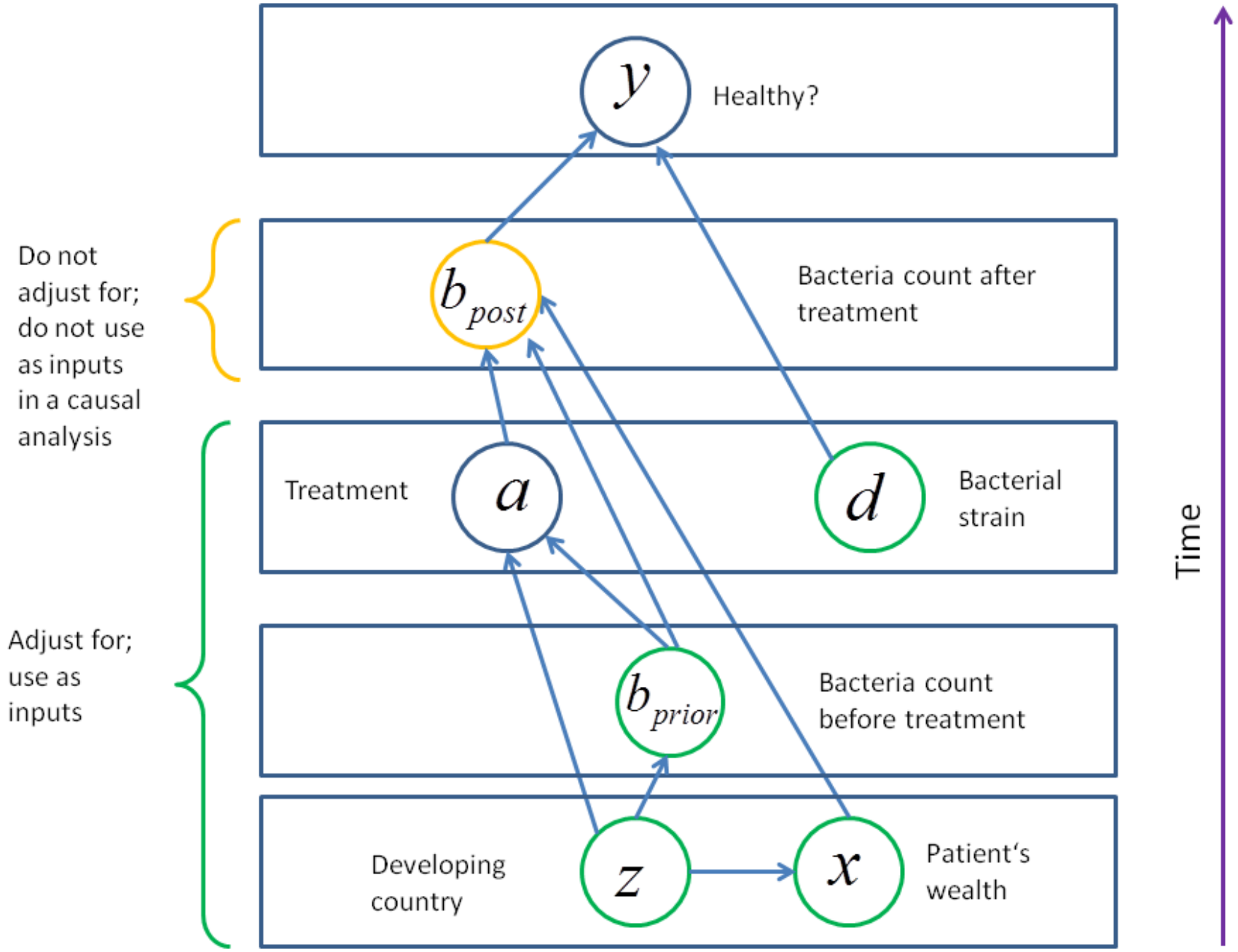
Effect Modification

- So far we only considered linear models
- In nonlinear models, a might have a positive effect for some states of x and a negative effect for others. This issue is called: effect modification, interaction, or heterogeneity between strata. Example: when the DAX is high, sell DAY, if the DAY is low, buy DAX
- Thus in general, we want

$$a_{opt}(x) = \arg\max_{a} f(a, x)$$

More Complex Example

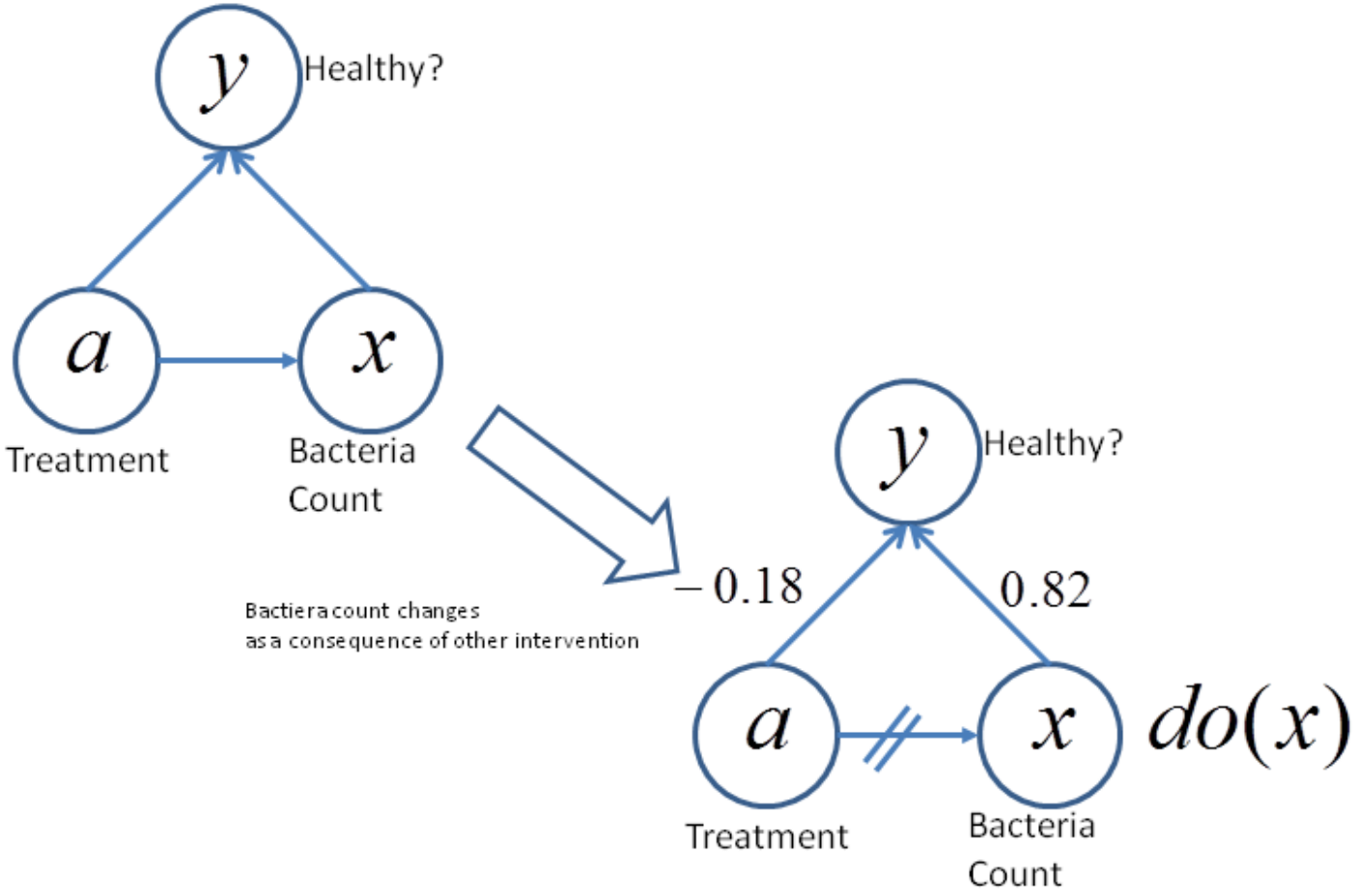
- The next example is slightly more complex and illustrates the recipe


y
Healthy?
Do not adjust for; do not use as inputs in a causal analysis
b post
Bacteria count after treatment
Treatment
a
d
Bacterial strain
Adjust for; use as inputs
b prior
Bacteria count before treatment
Developing country
z
x
Patient's wealth
Time

Manipulating Effects

- Consider again that a causes x
- If I can manipulate the effect x with $do(x)$, of course the model also changes

If we can Manipulate Bacteria independently?



Modelling Financial Decisions

- My friend is a big investor (big enough that his decisions influence the stock market). In the morning he buys DAX $a = 1$ or he sells DAX $a = 0$. His decision is only based on the morning mood of his cat. When his cat is in good mood $m = 1$ he buys DAX, otherwise he sells DAX, so we have $m \to a$. His competitor is also a big investor and always looks at my friend's decisions and also tends to buys or sells DAX using this information and we have $a \to c$. y is the gain or loss in the next morning.
- He asks me to train a model to predict his gain. Based on historic data, I model

$$\widehat{y} = f(a, m, c)$$

Improving Financial Decisions

- The model works quite well. Now he asks me: can I use your model to help me improve my decisions?
- First we need to remove c from the model —since it is only known after the decision is made— to get the correct causal effect and we should model

$$\widehat{y} = f(a, m)$$

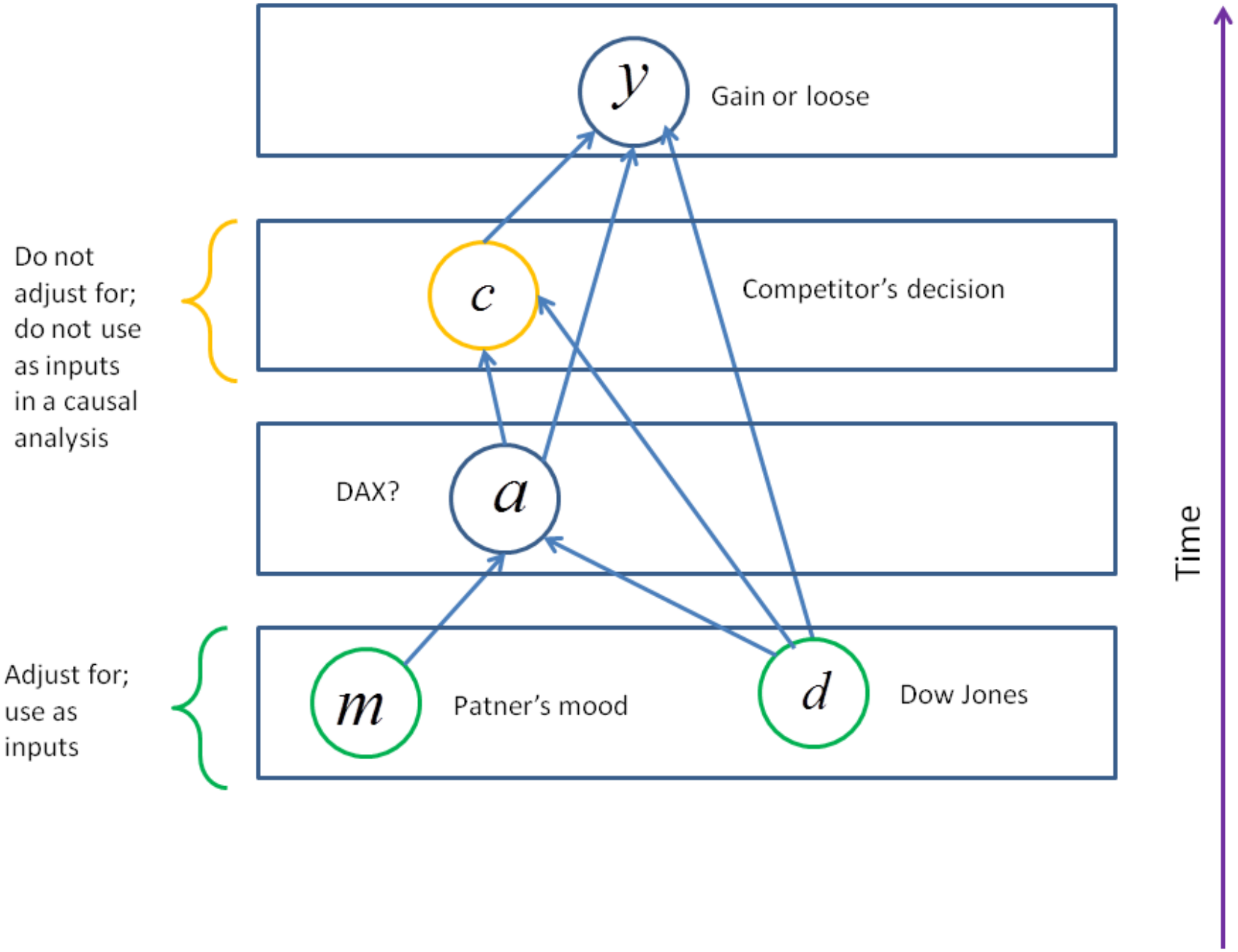
  Since we know about the mood, which is a cause for the action, we leave it in the model (otherwise it would be a confounder)

- Unfortunately, the advice from this model is not very good. The reason is that my friend also looks at the Dow Jones index d but did not tell me; thus d is a confounder
- After interrogation, he confesses that he also looks at the Dow Jones and I include it in the model

$$\widehat{y} = f(a, m, d)$$

- If I learn this model and optimize decisions based on this model, I should get good results with

$$a_{opt}(m, d) = \arg\max_{a} f(a, m, d)$$

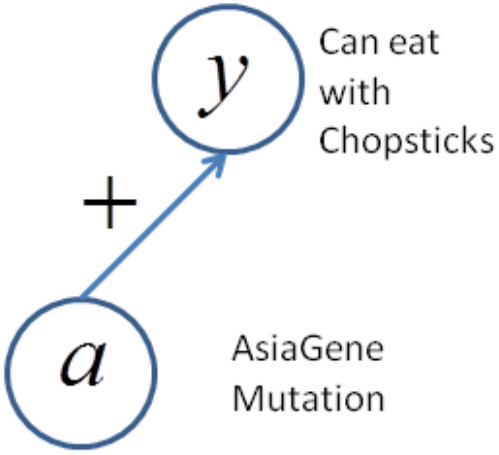

y
Gain or loose
Do not adjust for; do not use as inputs in a causal analysis
c
Competitor's decision
DAX?
a
Adjust for; use as inputs
m
Patner's mood
d
Dow Jones
Time

The Chopsticks Gene

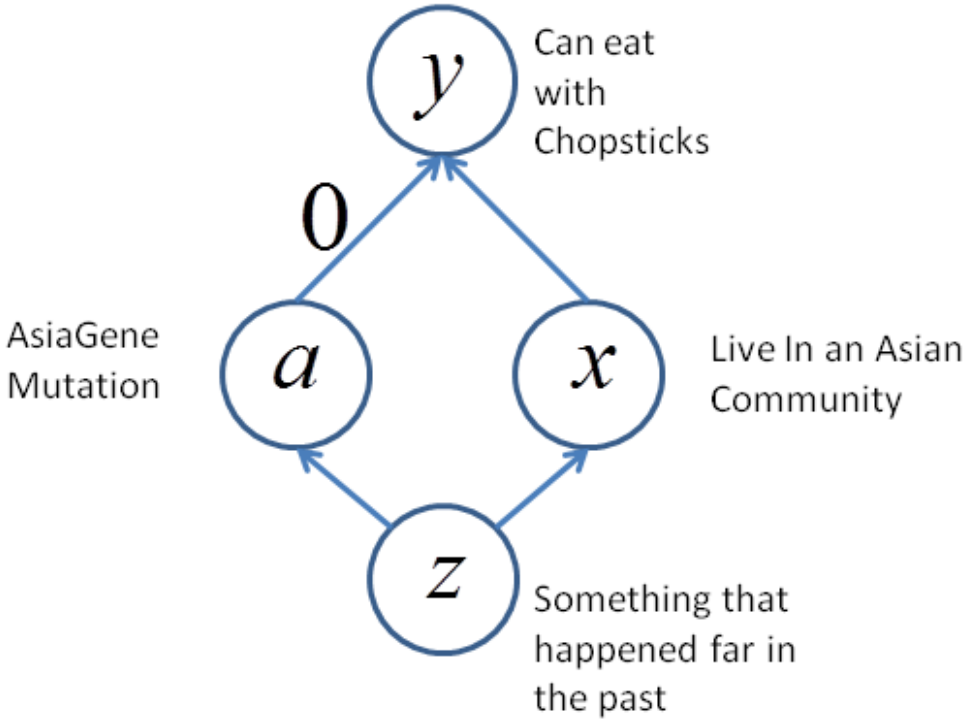
- Since we do not manipulate genes in humans, we cannot do randomized experiments to study the effects of local gene mutations (variants) (SNPs)
- A big issue are correlated SNPs when their locations on the DNA are close by. Asymptoticaly one might find the correct model, but with finite data, correlations are a problem. State of the art: The SNPs influence is modeled for individual SNPs (thus we train 1 Mio models). Correlated SNPs are modeled, e.g., by leading PCA components calculated on all (1Mio) SNPs
- Still, there are surprises: Researchers have found the gene which makes you manage to eat with chopsticks
- "Living in an Asian community" is the confounder which causes a person to have the Asia Gene and being good at eating with chopsticks (the real cause is probably something that happened far in the past but "Living in an Asian community" is a good substitute)

- Thus “Living in an Asian community” should be included in the model; in reality it is hoped that the PCA components already contain information about a general asian background

Asia Gene



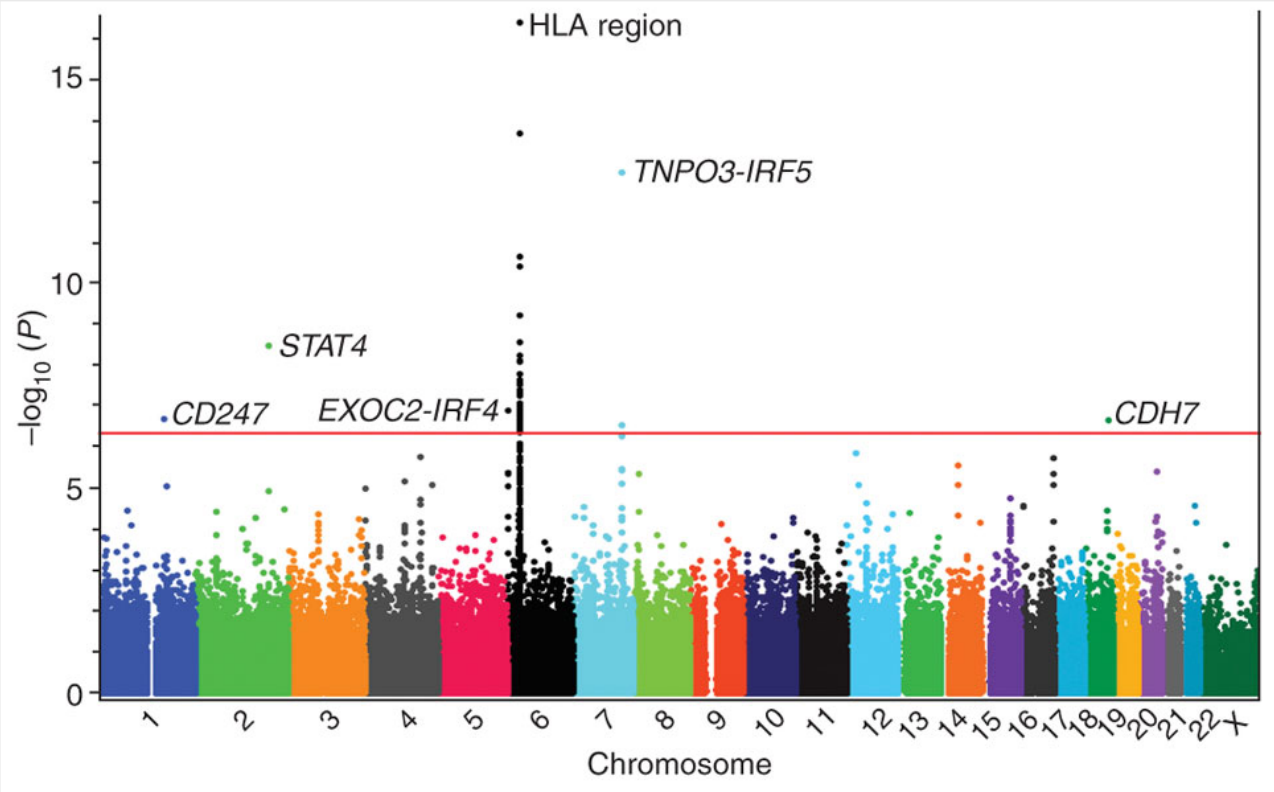
Incorrect Causal Model



Correct Causal Model

GWAS Study

Correlation with disease (systemic sclerosis) versus location of SNPs on the gene. The regression weight of a single SNP as an input is calculated with other inputs representing general personal traits and are possible confounders (male/female, Caucasian, Asian, PCA features derived from all SNPs, ...). Repeated for all SNPs (maybe 1 Mio).



Secondary Use of Clinical Data

- The situation is similar, if I need to optimize medical decisions observed in practice. If the decision law used by the doctor is known, I need to include all variables which were used by the doctor
- This is typically a problem since not all information that is used by the doctor is typically documented. Also, information, that is only available at a later instance in time (e.g., a blood count done later) should not be included!

Controlled Plants

- The situation is similar, if I need to optimize a plant which is controlled. I need to include in the system model all variables that are used in the control law, but not variables that are measured after the control action! This task is called system identification

$$\widehat{y} = f(a, x)$$

- If y_{des} is the desired output,

$$a_{opt}(x) = \arg\min_{a}(f(a, x) - y_{des})^2$$

- The control law then is

$$a_{opt}(x) = g(y_{des}, x)$$

Learning the Inverse with Optimal Actions

- Note that if $\widehat{y} = f(a, x)$ is a forward model, then $a_{opt}(x) = g(y_{des}, x)$ is an inverse model
- Such an inverse model was learned in a Siemens application for the control of a rolling mill: x is the state of the steel (thickness, temperature, ...), y_{des} is the desired thickness of the steal and $a_{opt}(x)$ is the correct rolling force to achieve the desired thickness. $a_{opt}(y_{des}, x)$ needs to be estimated when a new piece of steel enters the mill. Then the feedback control sets in and within a second the true $a_{opt}(y_{des}, x)$ is known which can then be used for training
- We had perfect training data: $x, y_{des}, a_{opt}(y_{des}, x)$. The project was a great success and made Siemens world leader in this application

Learning the Inverse with Observed Actions

- Another issue is if I can learn an inverse model

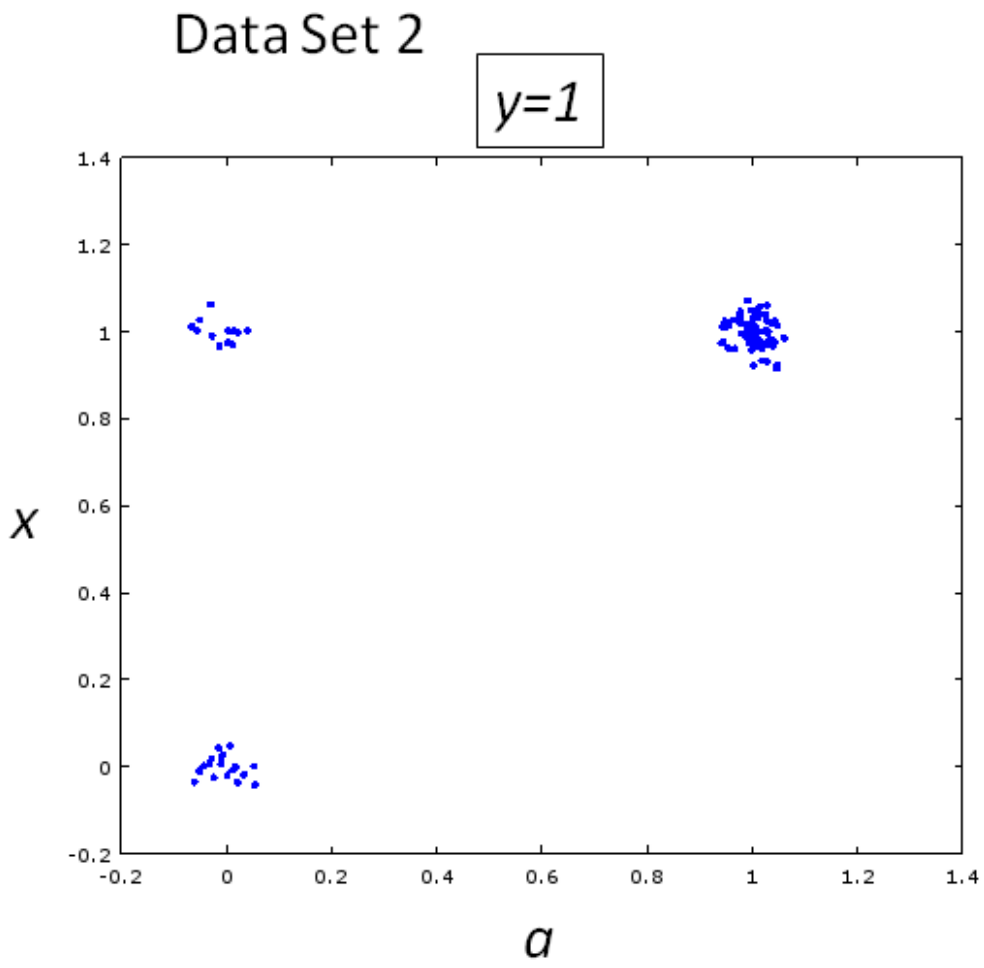
$$\widehat{a} = g'(y, x)$$

  by simply observing the plant under some control and then define

$$a'_{opt}(x) := g'(x, y_{des}) = E(a|y_{des})$$

- With no or little uncertainty in the model and if the forward model is invertible, then this can work

- On the other hand, if there is strong correlation between a and x in the observed data and if the model is noisy, then this can give the wrong answers, as in our standard example (see next figure)!



We look at the data where *y=1* (desired output)

- If *x=0*, we see that *g'(x, y=1) = E(a|x, y=1)=0*, so the medication should not be given, which is right
- If x=1, we see that *g'(x, y=1) = E(a|x, y=1)=0.84*, so the medication should be given, which is wrong!

Exploration

- Note that another issue is that a learned model might not be valid in regions of state space never explored under control: For example, if $a = g(x)$, then $f(a, x) = f(g(x), x)$ might not be a good model for uncommon a
- Thus, some form of exploration might be necessary

Machine Learning versus Statistics

- This lecture tried to help you avoid embarrassment: "Data Miners found: Taking your breaks outside of the building causes lung cancer!". Now the know that there might be a confounding factor: smoking!
- Machine Learning is primarily concerned with prediction accuracy. What is the best prediction given the available information? Regularization is always applied, in particular with few data points and when inputs are correlated.
- In statistics one is traditionally concerned with the interpretability of parameters and the elimination of non-causal correlated inputs and the search for the minimum "true" set of inputs is of great concern. If relevant inputs are missing, then the parameters are more difficult to interpret
- Unfortunately, a causal analysis is typically not done in statistics. For more, Judea Pearl, Causal inference in statistics: An overview (tech report). Pearl argues that a causal analysis is essential for drawing conclusions from observed data (see also: http://singapore.cs.ucla.edu/LECTURE/)

- Note that we made causal assumptions. Discovering (learning) causal structure from observed data is another area of research (see lecture on Bayes nets). Pearl argues that causality cannot completely be learned from observed data and needs to be use in form of prior knowledge

But What is Causality

- http://singapore.cs.ucla.edu/LECTURE/
- "All philosophers," says Bertrand Russell, "imagine that causation is one of the fundamental axioms of science, yet oddly enough, in advanced sciences, the word 'cause' never occurs ... The law of causality, I believe, is a relic of bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm ..."
- "Likewise, we say that the ratio f/a helps us DETERMINE the mass, not that it CAUSES the mass. Such distinctions are not supported by the equations of physics, and this leads us to ask whether the whole causal vocabulary is purely metaphysical. "surviving, like the monarchy...etc."
- "I felt like a buccaneer of Drake's days -... I interpreted that sentence of Galton to mean that there was a category broader than causation, namely correlation, of which causation was only the limit, and that this new conception of correlation brought psychology, anthropology, medicine, and sociology in large parts into the field of mathematical treatment." Karl Pearson (1934)

- Laplace's demon was the first published articulation of causal or scientific determinism by Pierre-Simon Laplace in 1814. According to determinism, if someone (the Demon) knows the precise location and momentum of every atom in the universe, their past and future values for any given time are entailed; they can be calculated from the laws of classical mechanics. (Possible arguments against it: entropy, quantum mechanics (Copenhagen interpretation), relativity, chaos theory, computational complexity, ...)

- J. Pearl's argument: The answer is: YES. If you wish to include the entire universe in the model, causality disappears because interventions disappear - the manipulator and the manipulated loose their distinction. However, scientists rarely consider the entirety of the universe as an object of investigation. In most cases the scientist carves a piece from the universe and proclaims that piece: IN namely, the FOCUS of investigation. The rest of the universe is then considered OUT or BACKGROUND, and is summarized by what we call BOUNDARY CONDITIONS. This choice of INs and OUTs creates asymmetry in the way we look at things, and it is this asymmetry that permits us to talk about "outside intervention", hence, causality and cause-effect directionality.

Concrete Example

- Assume that *lungCancer* is a deterministic function of a *geneMutation* and *smoking*
- We can predict any of the three quantities from the other. All three are "IN". There is now way to determine causality from observation. The world is inherently causal but I cannot identify causality
- When I manipulate a variable using an external intervention $do()$, then the links to that variables are removed in "IN" and I can decide what is the cause and what is the effect, i.e., $\mathit{lungCancer} = f(\mathit{smoking}, \mathit{geneMutation})$
- If I can distinguish IN from OUT, i.e., if I allow for manipulations from outside of "IN" I can introduce the concept if *time*: effects are after the cause
- Meaning of probability: If I observe data for many persons in population, but *geneMutation* is OUT and unknown, I can treat the influence of *geneMutation* as being random