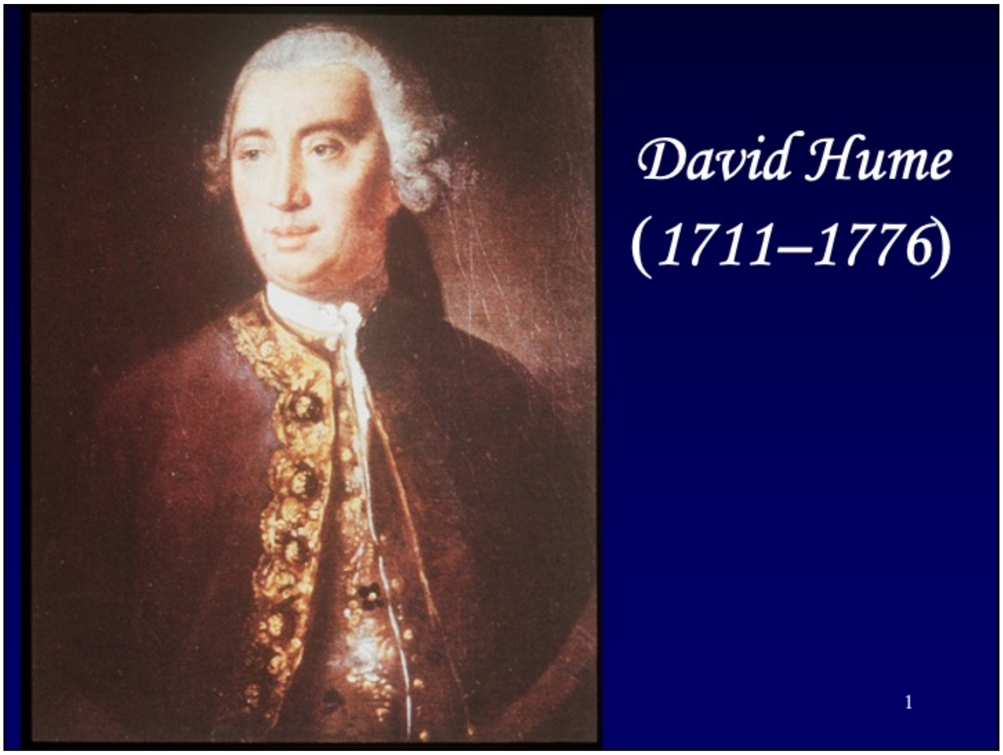
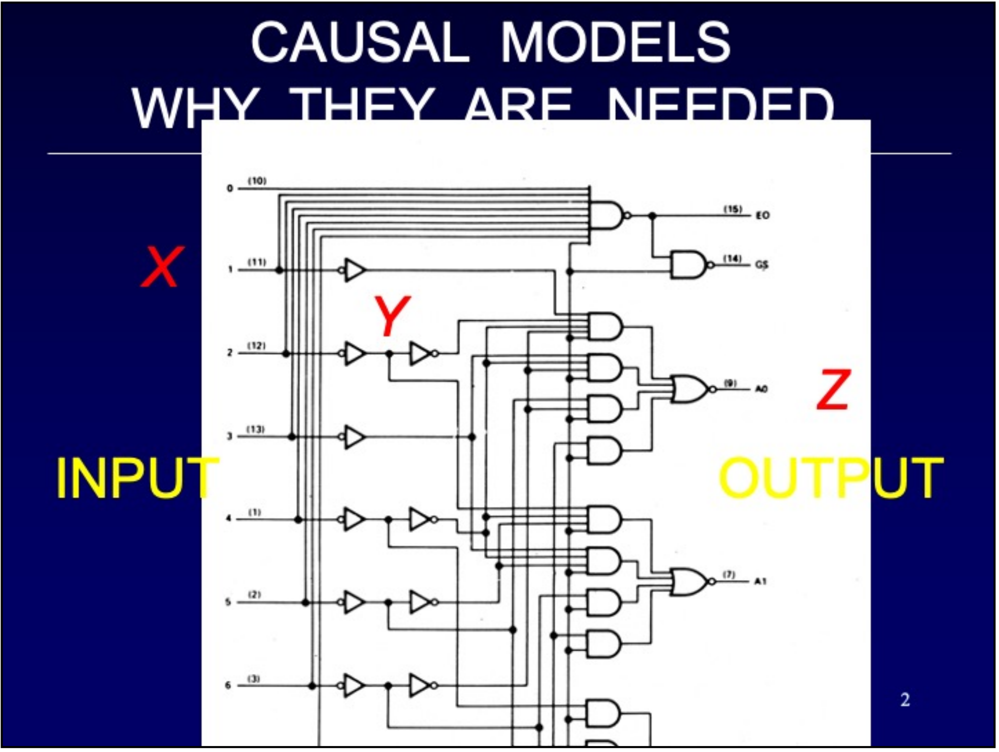


David Hume
(1711–1776)

1

The modern study of causation begins with the Scottish philosopher David Hume.

Hume has introduced to philosophy three revolutionary ideas that, today, are taken for granted by almost everybody, not only philosophers.



CAUSAL MODELS
WHY THEY ARE NEEDED

Slide2

Here is a causal model we all remember from high-school -- a circuit diagram.

There are 4 interesting points to notice in this example:

(1) It qualifies as a causal model -- because it contains the information to confirm or refute all action, counterfactual and explanatory sentences concerned with the operation of the circuit.

For example, anyone can figure out what the output would be like if we set Y to zero, or if we change this OR gate to a NOR gate or if we perform any of the billions combinations of such actions.

(2) Logical functions (Boolean input-output relation) is insufficient for answering such queries

(3)These actions were not specified in advance, they do not have special names and they do not show up in the diagram.

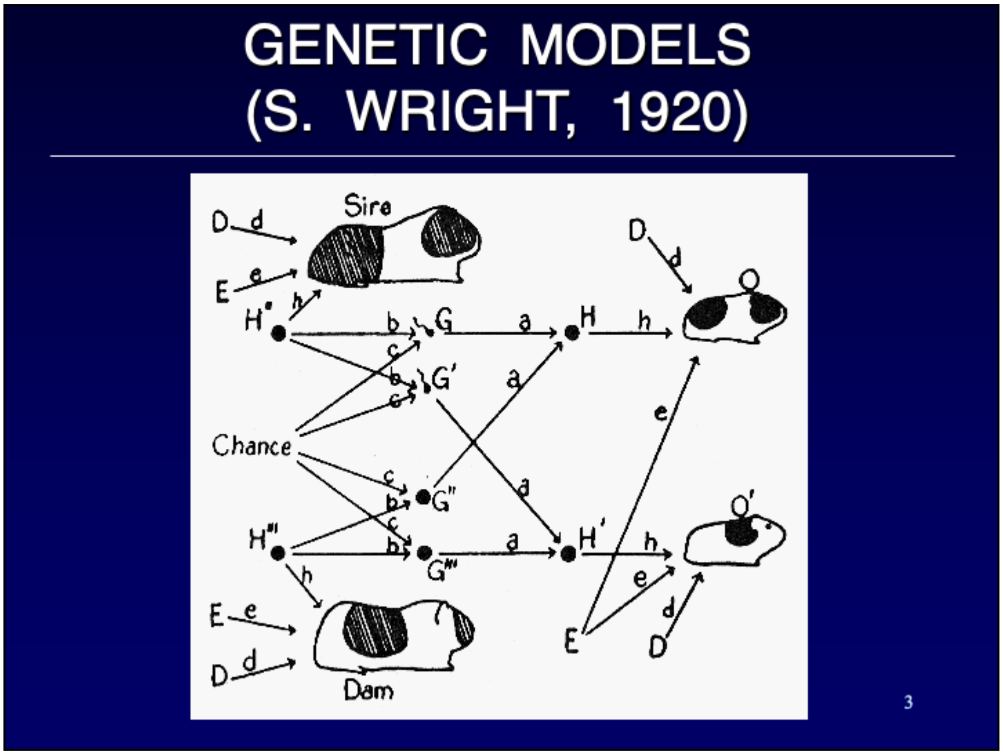
In fact, the great majority of the action queries that this circuit can answer have never been considered by the designer of this circuit.

(4) So how does the circuit encode this extra information?

Through two encoding tricks:

4.1 The symbolic units correspond to stable physical mechanisms (i.e., the logical gates)

4.2 Each variable has precisely one mechanism that determines its value.

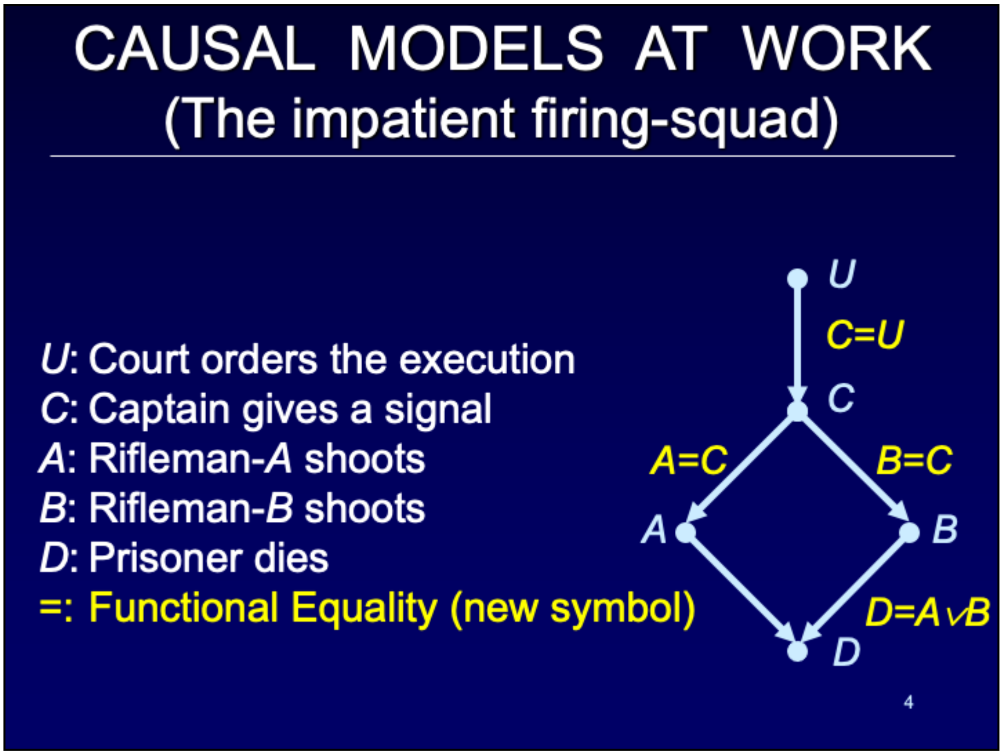


GENETIC MODELS
(S. WRIGHT, 1920)

As another example, here is the first causal model that was put down on paper: Sewal Wright's path diagram, showing how the fur pattern of the litter guinea pigs is determined by various genetic and environmental factors. Again, (1) it qualifies as a causal model, (2) the algebraic equations in themselves do not NOT qualify, and (3) the extra information comes from having each variable determined by a stable functional mechanism connecting it to its parents in the diagram.

Now that we are on familiar grounds, let us observe more closely the way a causal model encodes the information needed for answering causal queries.

Instead of a formal definition that you can find in the proceedings paper (Def. 1), I will illustrate the working of a causal model through another example, which can also be found in your proceedings -

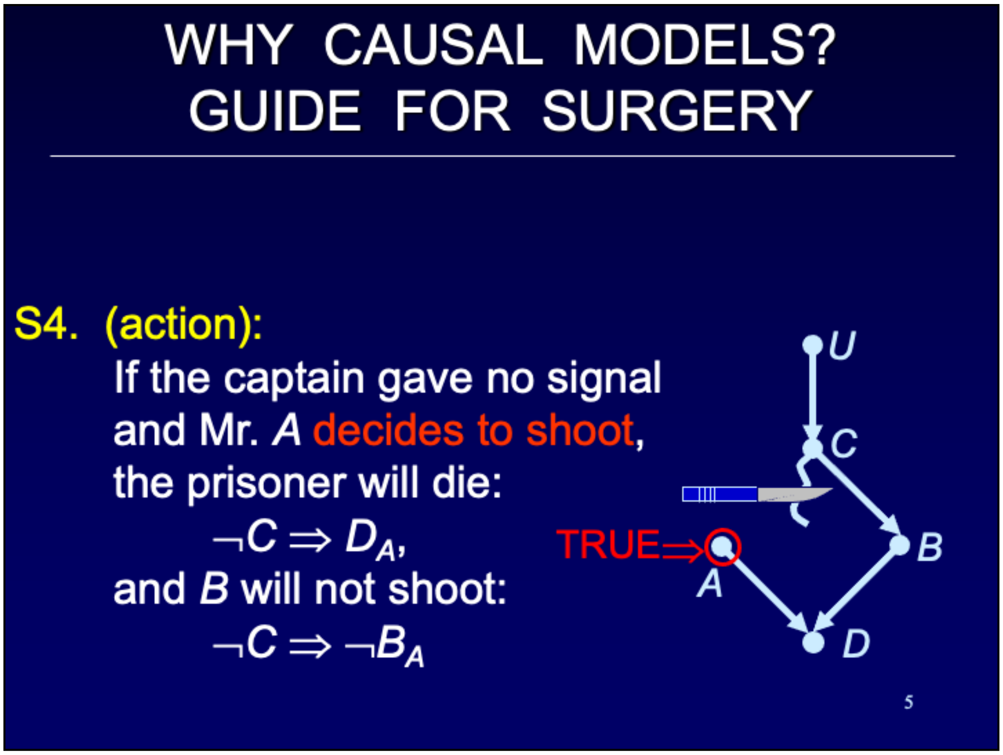


CAUSAL MODELS AT WORK
(The impatient firing-squad)

U: Court orders the execution
C: Captain gives a signal
A: Rifleman-*A* shoots
B: Rifleman-*B* shoots
D: Prisoner dies
=: Functional Equality (new symbol)

The meanings of the symbols is obvious from the story:

The only new symbol is the functional equality = which is borrowed here from Euler (around 1730's), meaning that the left hand side is determined by the right hand side and not the other way around.



WHY CAUSAL MODELS? GUIDE FOR SURGERY

S4. (action):

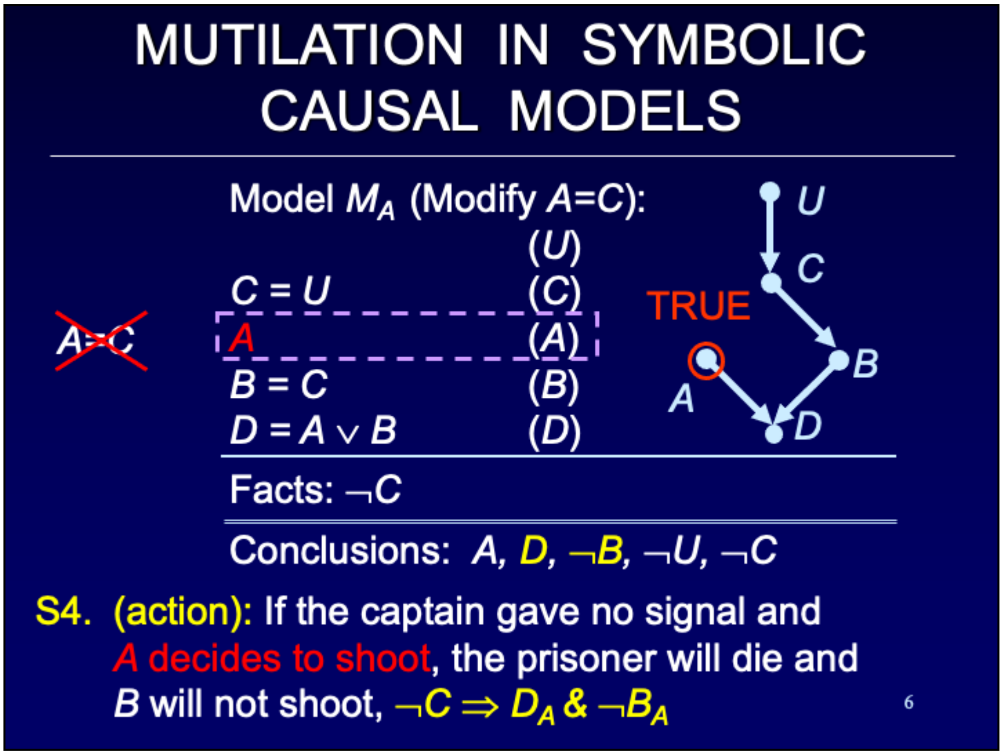
If the captain gave no signal and Mr. *A* decides to shoot, the prisoner will die:

$\neg C \Rightarrow D_A$,

and *B* will not shoot:

$\neg C \Rightarrow \neg B_A$

This surgery also suppresses abduction; from seeing *A* shoot we can infer that *B* shot as well (recall $A \Rightarrow B$), but from MAKING *A* shoot we can no longer infer what *B* does.



MUTILATION IN SYMBOLIC CAUSAL MODELS

Model M_A (Modify $A=C$):

~~$A=C$~~

(U)

$C = U \qquad (C)$

$A \qquad (A)$

$B = C \qquad (B)$

$D = A \vee B \qquad (D)$

Facts: $\neg C$

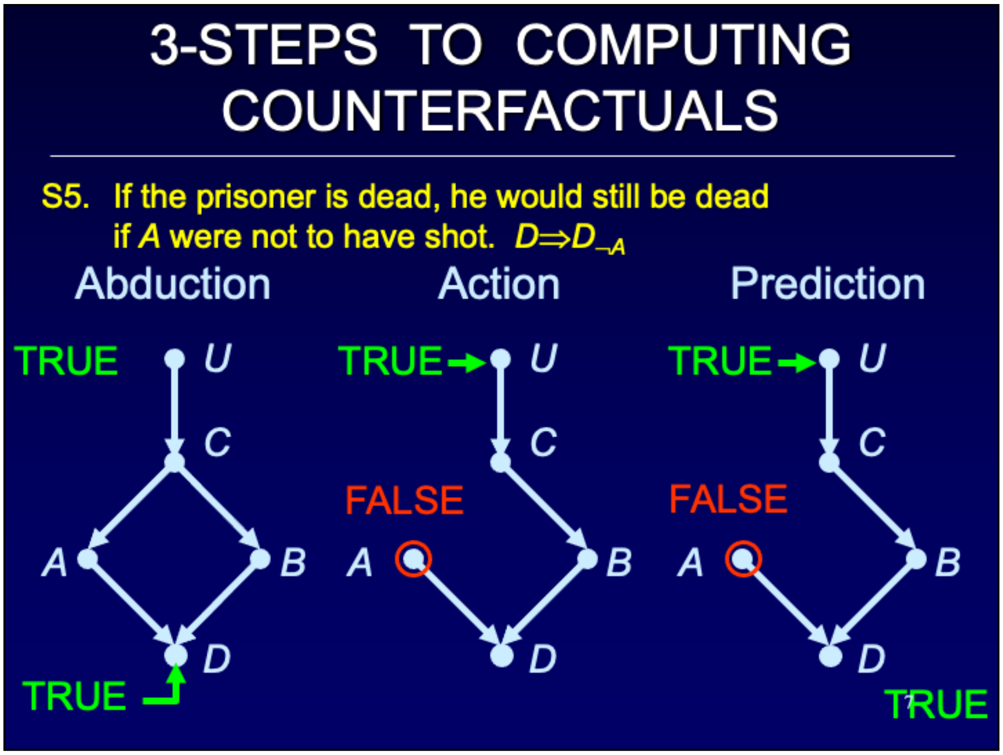
Conclusions: $A, D, \neg B, \neg U, \neg C$

S4. (action): If the captain gave no signal and *A* decides to shoot, the prisoner will die and *B* will not shoot, $\neg C \Rightarrow D_A \& \neg B_A$

6

Once we create the mutilated model M_A, we draw the conclusions by standard deduction and easily confirm:

S4: The prisoner will be dead -- D is true in M_A.



3-STEPS TO COMPUTING COUNTERFACTUALS

S5. If the prisoner is dead, he would still be dead if *A* were not to have shot. $D \Rightarrow D_{\neg A}$

Consider now our counterfactual sentence

S5: If the prisoner is Dead, he would still be dead if *A* were not to have shot. D ==> $D_{\neg A}$

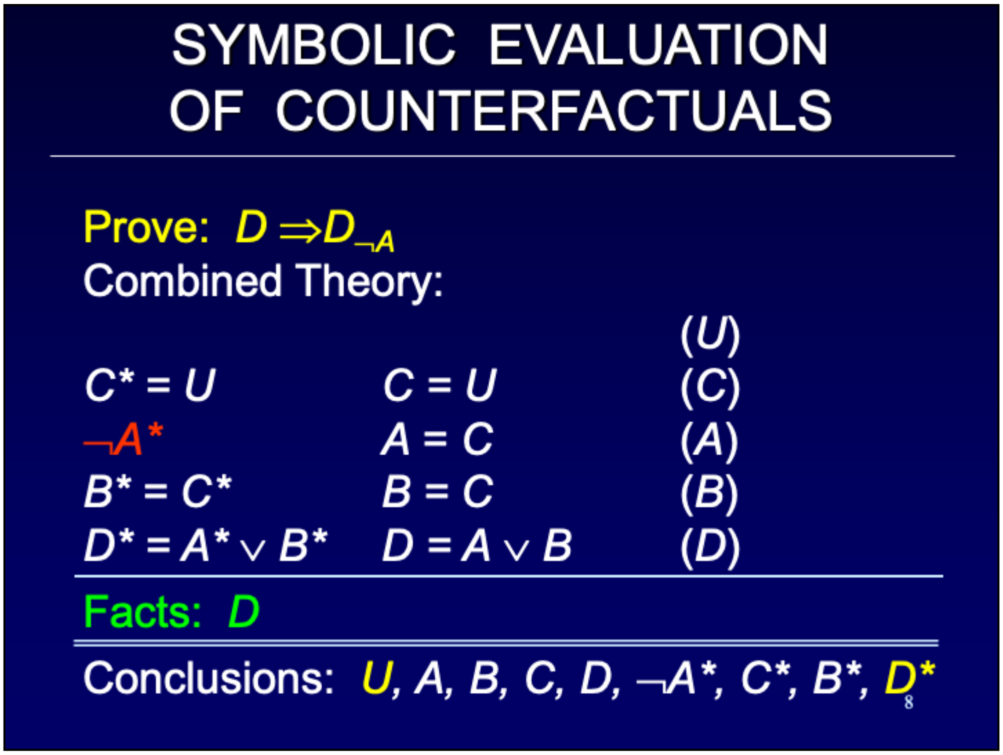
The antecedant $\{\neg A\}$ should still be treated as interventional surgery, but only after we fully account for the evidence given: *D*.

This calls for three steps

1 Abduction: Interpret the past in light of the evidence

2. Action: Bend the course of history (minimally) to account for the hypothetical antecedant ($\neg A$).

3.Prediction: Project the consequences to the future.



SYMBOLIC EVALUATION OF COUNTERFACTUALS

Prove: $D \Rightarrow D_{\neg A}$

Combined Theory:

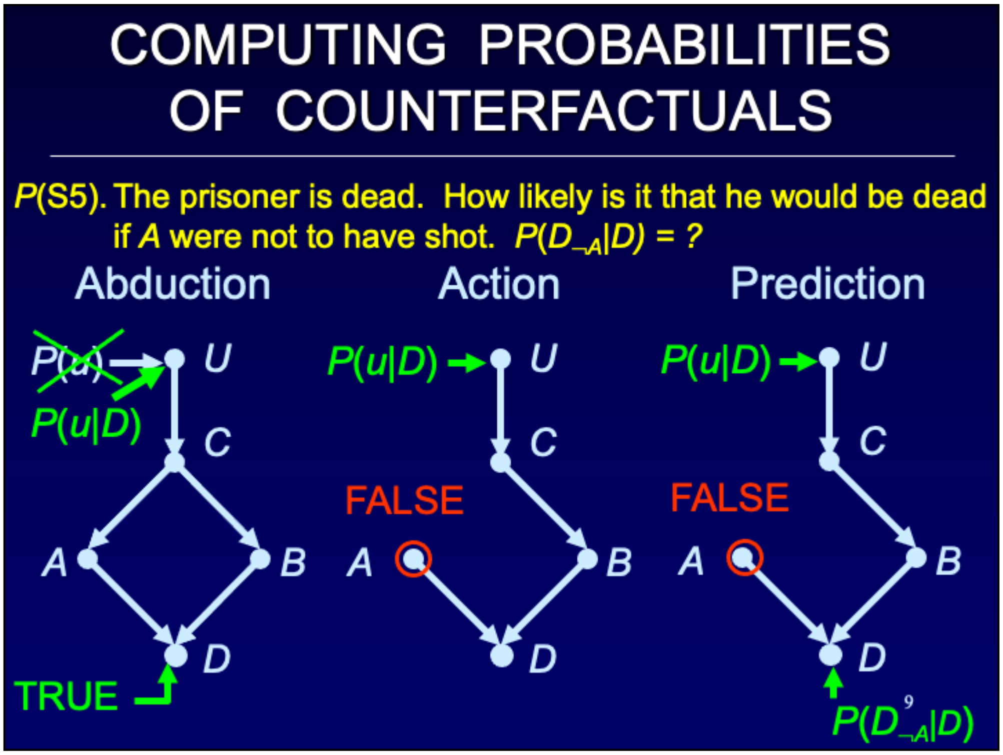
		(U)
$C^* = U$	$C = U$	(C)
$\neg A^*$	$A = C$	(A)
$B^* = C^*$	$B = C$	(B)
$D^* = A^* \vee B^*$	$D = A \vee B$	(D)

Facts: D

Conclusions: $U, A, B, C, D, \neg A^*, C^*, B^*, D^*$

We can combine the first two steps into one, if we use two models, M and M_A, to represent the actual and hypothetical worlds, respectively.

(Reader: See proceeding paper for technical details)



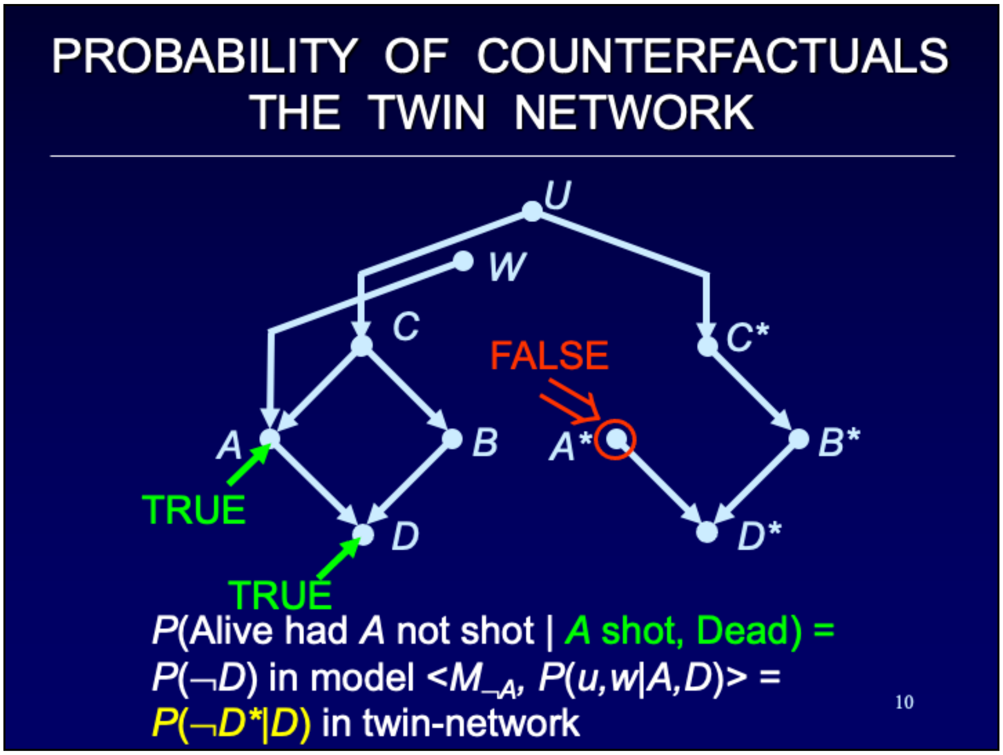
COMPUTING PROBABILITIES OF COUNTERFACTUALS

P(S5). The prisoner is dead. How likely is it that he would be dead if *A* were not to have shot. $P(D_{\neg A}|D)$ = ?

Suppose we are not entirely ignorant of *U*, but can assess the degree of belief *P*(*u*).

The same 3-steps apply to the computation of the counterfactual probability (that the prisoner be dead if *A* were not to have shot)

The only difference is that we now use the evidence to update *P*(*u*) into *P*(*u*|*e*), and draw probabilistic instead of logical conclusions.



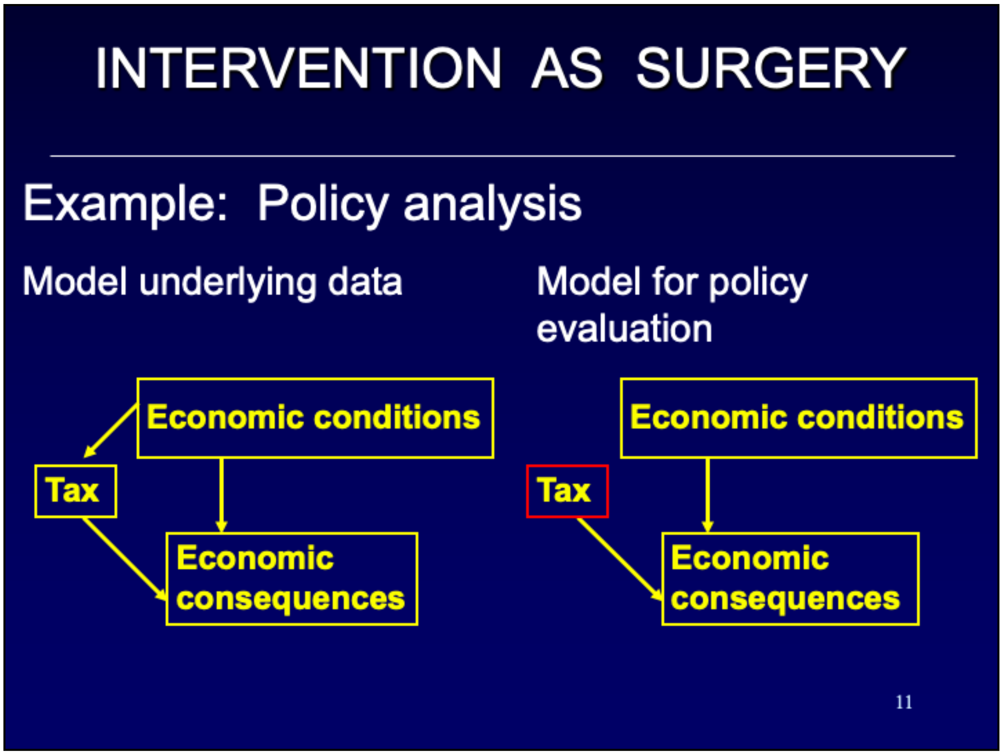
PROBABILITY OF COUNTERFACTUALS THE TWIN NETWORK

P(Alive had A not shot | A shot, Dead) =
$P(\neg D)$ in model $<M_{\neg A}, P(u,w|A,D)>$ =
$P(\neg D^*|D)$ in twin-network

10

Graphically, the two models can be represented by two graphs sharing the *U* variables (called TWIN-NETWORK).

The Twin-model is particularly useful in probabilistic calculations, because we can simply propagate evidence (using Bayes-network techniques) from the actual to the hypothetical network.



INTERVENTION AS SURGERY

Example: Policy analysis

Model underlying data

Economic conditions

Tax

Economic consequences

Model for policy evaluation

Economic conditions

Tax

Economic consequences

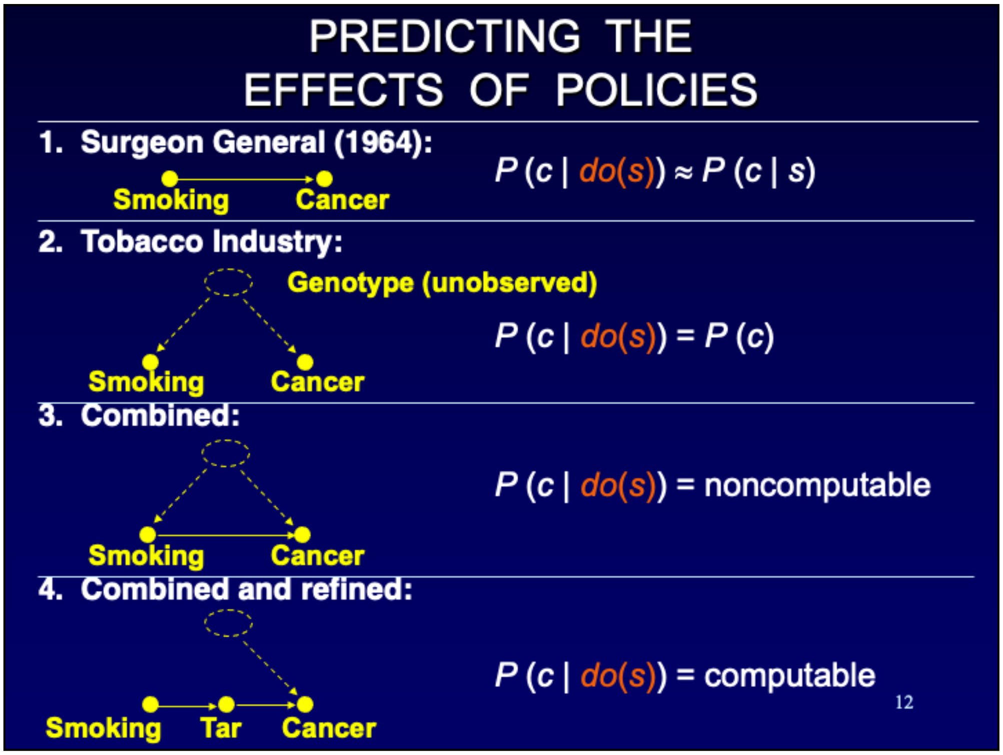
Economic policies are made in a manner similar to the way actions were taken in the firing squad story: Viewed from the outside, they are taken in response to economic indicators or political pressure, while viewed from the policy maker perspective, the next decision is chosen under the pretense of free will ....

Like rifleman-A, the policy maker should and does consider the ramification of non-routine actions that do not conform to the dictates of the model.

If we knew the model, there would be no problem calculating the ramifications of each pending decision -- mutilate and predict -- but being ignorant of the functional relationships and the probability of u, and having only the skeleton of the causal graph in our hands, we hope to supplement this information with what we can learn from economical data.

Unfortunately, economical data are taken under a wholesome graph, and we need to predict ramifications under a mutilated graph. Can we still extract useful information from such data?

The answer is YES. As long as we can measure every variable that is a common cause of two or more other measured variables, it is possible to infer the probabilities of the mutilated model directly from those of the nonmutilated model REGARDLESS of the underlying functions. The transformation is given by the manipulation theorem described in the book by Spirtes Glymour and Schienes (1993).



PREDICTING THE EFFECTS OF POLICIES

1. **Surgeon General (1964):**

Smoking → Cancer

$P(c \mid do(s)) \approx P(c \mid s)$

2. **Tobacco Industry:**

Genotype (unobserved)

Smoking Cancer

$P(c \mid do(s)) = P(c)$

3. **Combined:**

Smoking → Cancer

$P(c \mid do(s))$ = noncomputable

4. **Combined and refined:**

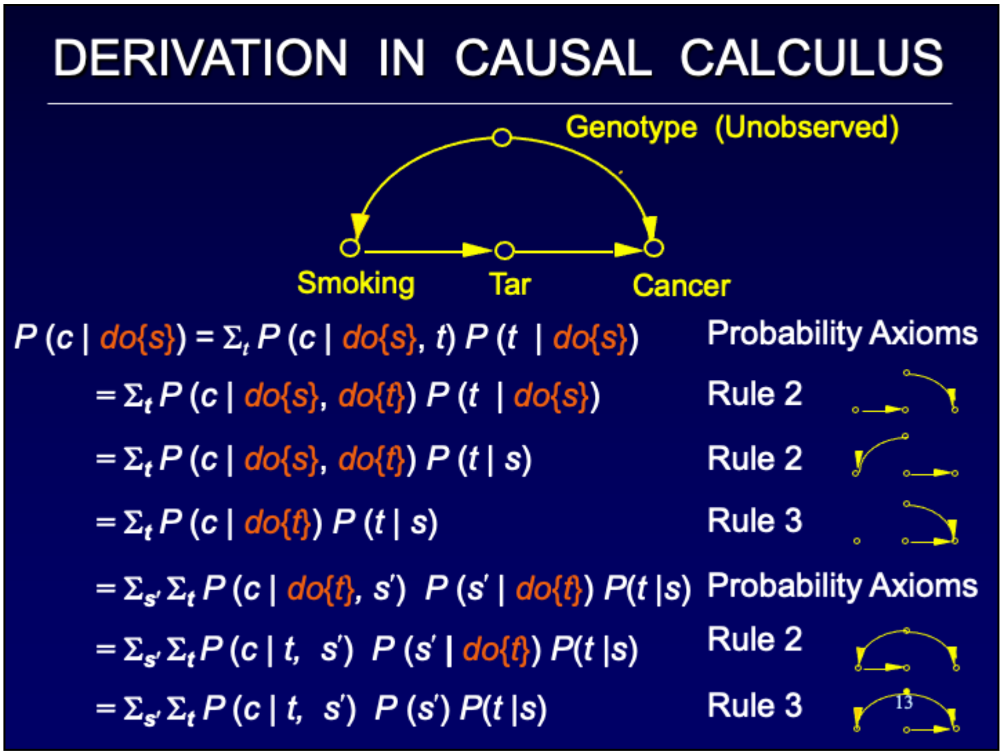
Smoking → Tar → Cancer

$P(c \mid do(s))$ = computable

12

This inference is valid as long as the data contains measurements of all three variables: Smoking, Tar and Cancer.

Moreover, the solution can be obtained in close mathematical form, using symbolic manipulations that mimic the surgery semantics.



DERIVATION IN CAUSAL CALCULUS

$P(c \mid do\{s\}) = \Sigma_t P(c \mid do\{s\}, t)\, P(t \mid do\{s\})$ Probability Axioms

$= \Sigma_t P(c \mid do\{s\}, do\{t\})\, P(t \mid do\{s\})$ Rule 2

$= \Sigma_t P(c \mid do\{s\}, do\{t\})\, P(t \mid s)$ Rule 2

$= \Sigma_t P(c \mid do\{t\})\, P(t \mid s)$ Rule 3

$= \Sigma_{s'} \Sigma_t P(c \mid do\{t\}, s')\; P(s' \mid do\{t\})\, P(t \mid s)$ Probability Axioms

$= \Sigma_{s'} \Sigma_t P(c \mid t, s')\; P(s' \mid do\{t\})\, P(t \mid s)$ Rule 2

$= \Sigma_{s'} \Sigma_t P(c \mid t, s')\; P(s')\, P(t \mid s)$ Rule 3

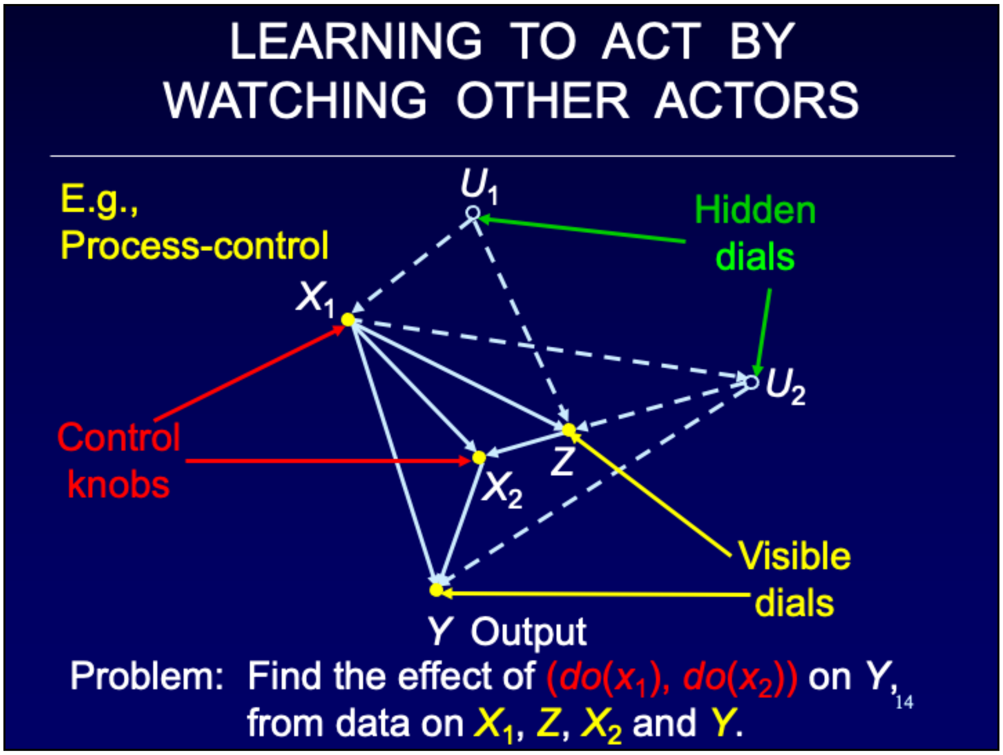
Here we see how one can prove that the effect of smoking on cancer can be determined from data on three variables: Smoking, Tar and Cancer.

The question boils down to computing P(cancer) under the hypothetical action do(smoking), from non-experimental data, namely, from expressions involving NO ACTIONS. Or: we need to eliminate the "do" symbol from the initial expression.

The elimination proceeds like ordinary solution of algebraic equation -- in each stage, a new rule is applied, licensed by some subgraph of the diagram, until eventually leading to a formula involving only WHITE SYMBOLS, meaning an expression computable from non-experimental data.

Now, if I were not a modest person, I would say that this is an amazing result. Watch what is going on here: we are not given any information whatsoever on the hidden genotype, it may be continuous or discrete, unidimensional or multidimensional. Yet, measuring an auxiliary variable TAR someplace else in the system, enables us to predict what the world would be like in the hypothetical situation where people were free of the influence of this hidden genotype. Data on the visible allows us to infer the effects of the invisible. Moreover, a person can also figure out the answer to the question: "I am about to smoke -- should I"?

I think it is amazing, because I cannot do this calculation in my head. It demonstrates the immense power of having a formal language in an area that many respectable scientists prefer to see handled by unaided judgment.



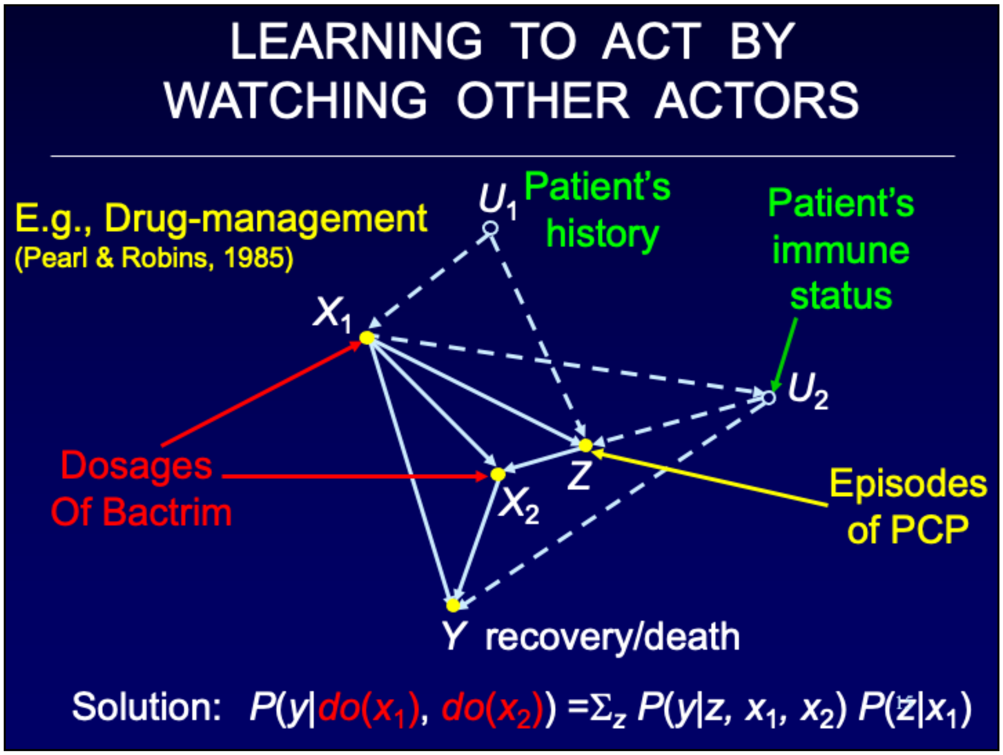
LEARNING TO ACT BY WATCHING OTHER ACTORS

Problem: Find the effect of $(do(x_1), do(x_2))$ on Y, from data on X_1, Z, X_2 and Y.

The common theme in the past two examples was the need to predict the effect of our actions by watching the behavior of other actors (past policy makers in the case of economic decisions, and past smokers-nonsmokers in the smoking-cancer example).

This is a recurring problem in many applications, and here are a couple of additional examples:

In this example, we need to predict the effect of a plan (sequence of actions) after watching an expert control a production process. The expert observes dials which we cannot observe, though we know what quantities those dials indicate.



LEARNING TO ACT BY WATCHING OTHER ACTORS

E.g., Drug-management
(Pearl & Robins, 1985)

U_1 Patient's history

Patient's immune status

X_1

U_2

Dosages Of Bactrim

Z

X_2

Episodes of PCP

Y recovery/death

Solution: $P(y|do(x_1), do(x_2)) = \Sigma_z P(y|z, x_1, x_2) P(z|x_1)$

The second example (due to J Robins) comes from sequential treatment of AIDS patients.

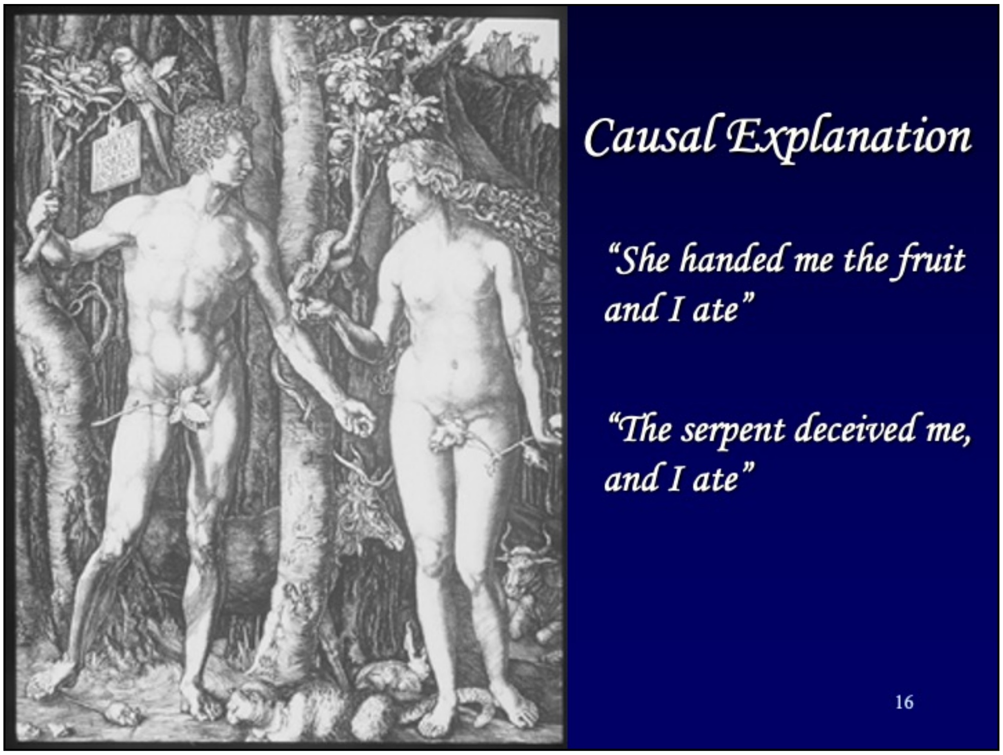
The variables X1 and X2 stand for treatments that physicians prescribe to a patient at two different times, Z represents observations that the second physician consults to determine X2, and Y represents the patient's survival. The hidden variables U1 and U2 represent, respectively, part of the patient history and the patient disposition to recover. Doctors used the patient's earlier PCP history (U1) to prescribe X1, but its value was not recorded for data analysis.

The problem we face is as follows. Assume we have collected a large amount of data on the behavior of many patients and physicians, which is summarized in the form of (an estimated) joint distribution P of the observed four variables (X1, Z, X2, Y). A new patient comes in and we wish to determine the impact of the (unconditional) plan (do(x1), do(x2)) on survival (Y), where x1 and x2 are two predetermined dosages of bactrim, to be administered at two prespecified times.

Many of you have probably noticed the similarity of this problem to Markov Decision processes, where it is required to find an optimal sequence of action to bring about a certain response. The problem here is both simpler and harder. Simpler, because we are only required to evaluate a given strategy, and harder, because we are not given the transition probabilities associated with the elementary actions -- those need to be learned from data. As you can see on the bottom line, this task is feasible - the answer is expressible as a probabilistic quantity that is estimable for the data.

How can this be accomplished? To reduce an expression involving do(x) to those involving ordinary probabilities we need a calculus for doing. A calculus that enables us to deduce behavior under intervention from behavior under passive observations.

Do we have such a calculus?



Causal Explanation

"She handed me the fruit and I ate"

"The serpent deceived me, and I ate"

16

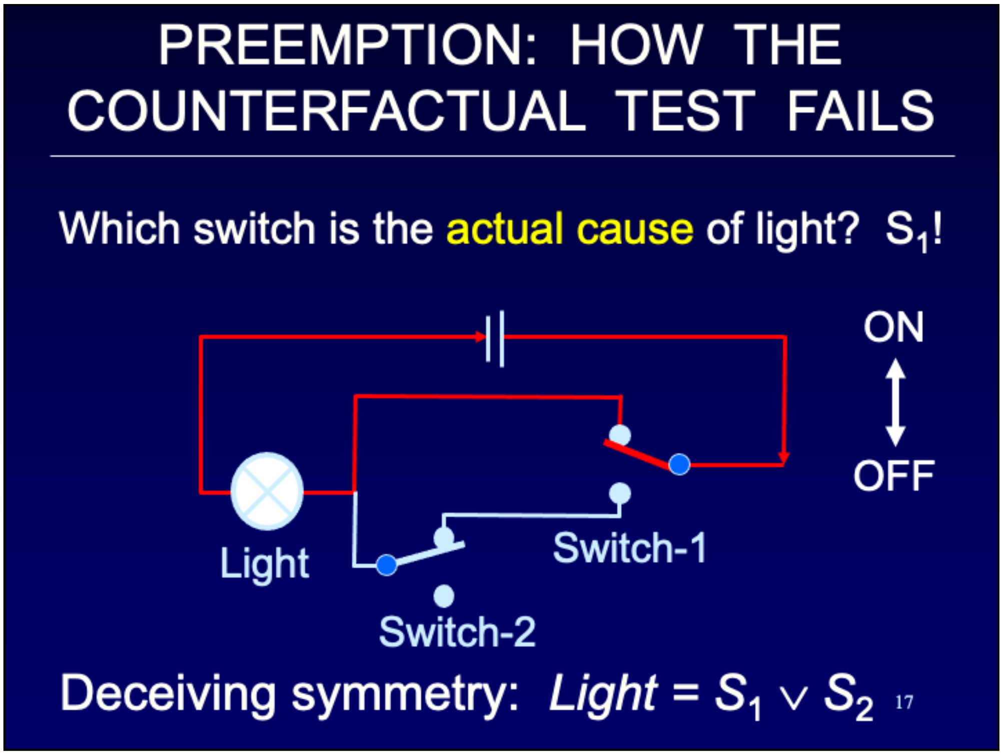
Eve is quick to catch on:

"The serpent deceived me, and I ate"

Explanations here are used for exonerating one from blame, passing on the responsibility to others:

The interpretation therefore is counterfactual:

"Had she not given me the fruit, I would not have eaten."



PREEMPTION: HOW THE COUNTERFACTUAL TEST FAILS

Which switch is the actual cause of light? S_1!

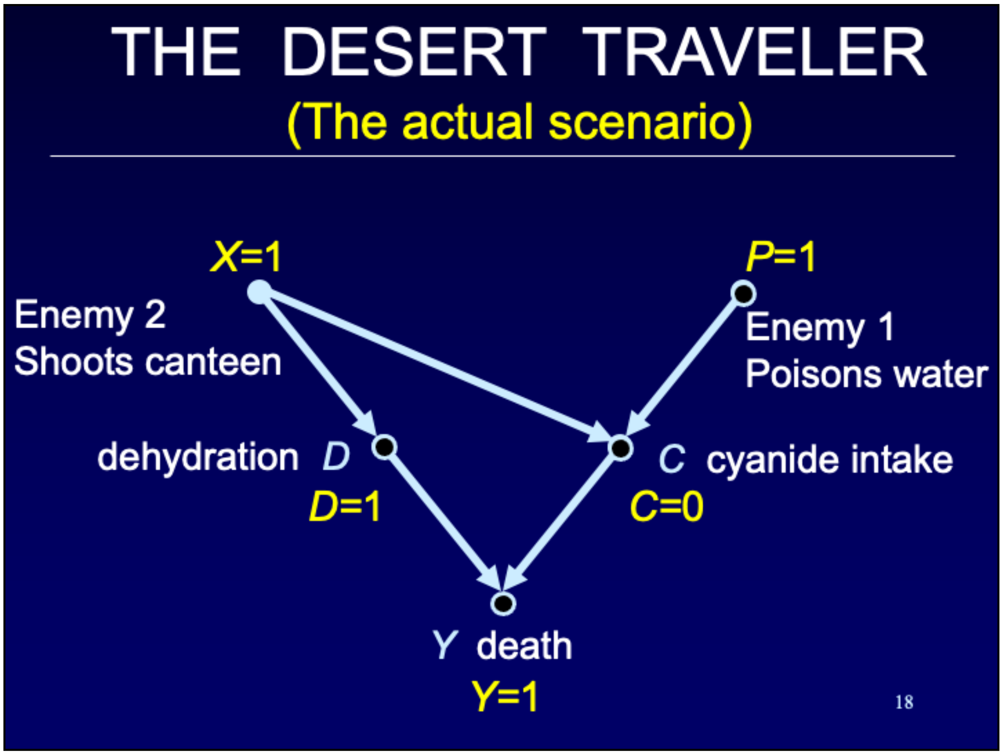
Deceiving symmetry: $Light = S_1 \vee S_2$

We now come to the 2nd difficulty with the counterfactual test, its failure to incorporate structural information.

If someone were to ask us what caused the light to be on, we would point to Switch-1. After all, S1 causes the current to flow through this wire, while S2 is totally out of the game.

On the other hand, the overall functional relationship between the switches and the light is deceptively symmetric:

$$Light = S1 \vee S2$$

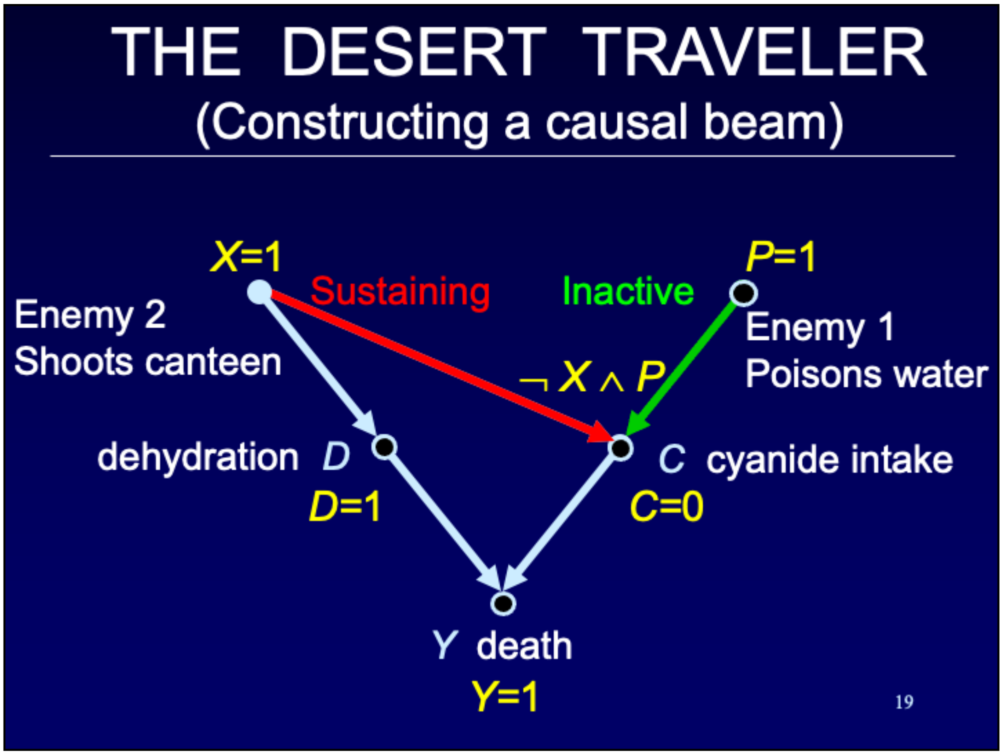


THE DESERT TRAVELER

(The actual scenario)

Now let us construct the causal beam associated with the natural scenario, in which we have:

Death (*Y=1*), Dehydration (*D=1*) and no poisoning (*C=0*).

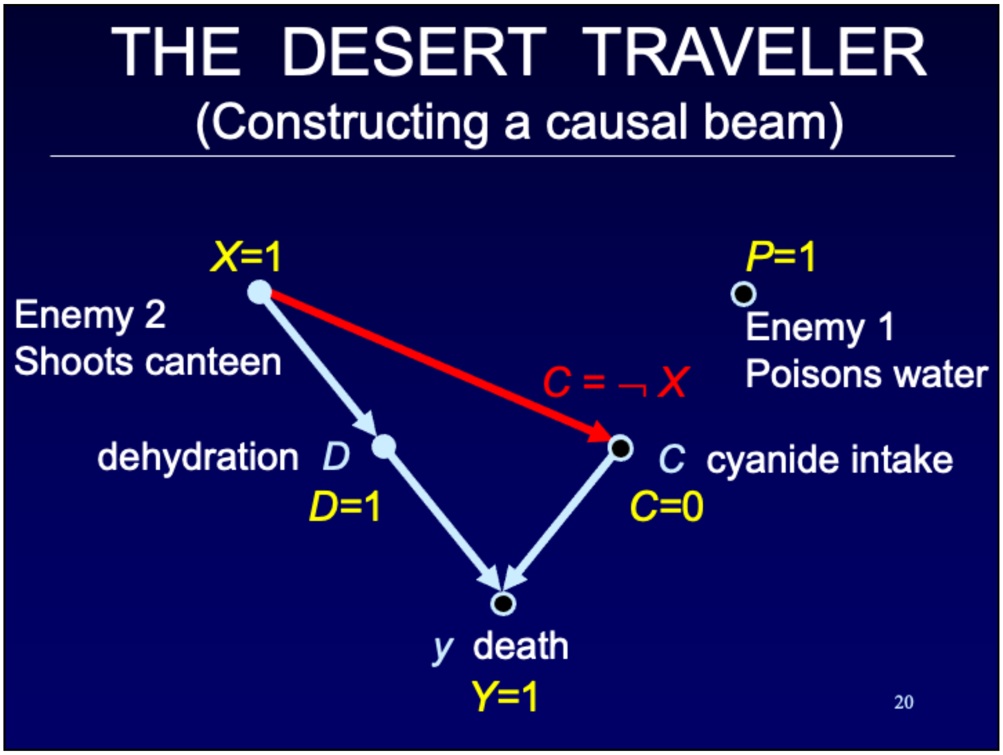


THE DESERT TRAVELER

(Constructing a causal beam)

19

Consider the Cyanide family. Since emptying the canteen is sufficient for sustaining no Cyanide intake, regardless of poisoning, we label the link *P*→*C* “inactive”, and the link *X*→*C* “sustaining”.



THE DESERT TRAVELER
(Constructing a causal beam)

X=1

Enemy 2
Shoots canteen

P=1

Enemy 1
Poisons water

$C = \neg X$

dehydration D

D=1

C cyanide intake

C=0

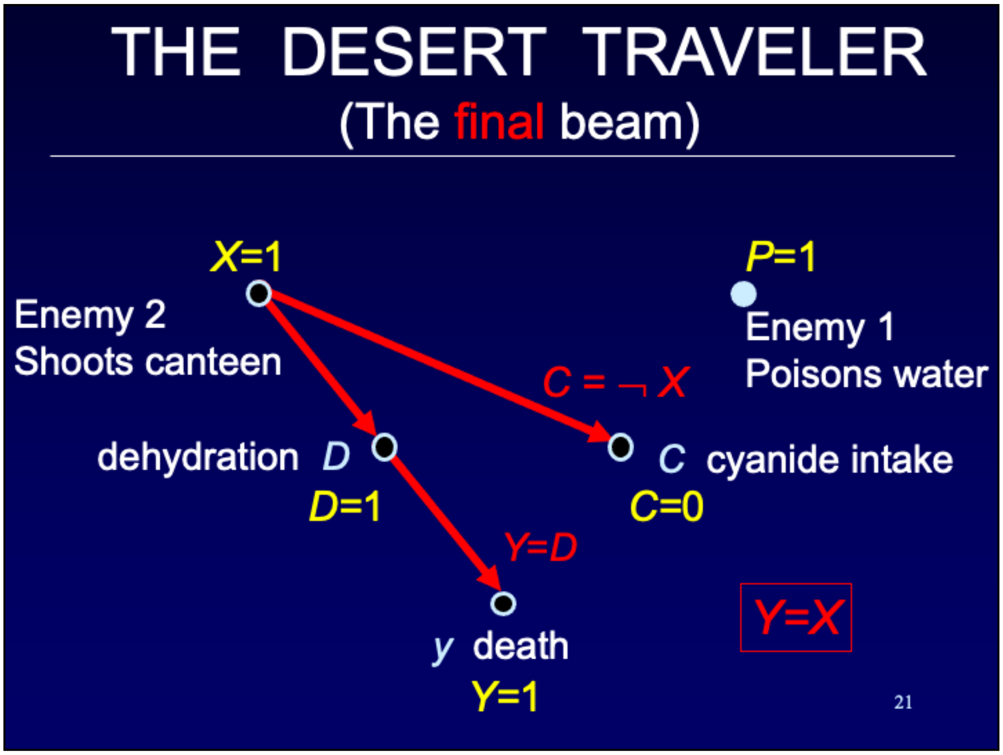
y death

Y=1

20

The link $P \rightarrow C$ is inactive in the current scenario, which allows us to retain just one parent of C, with the functional relationship $C = \neg X$.

We repeat this process on other parent-child families.

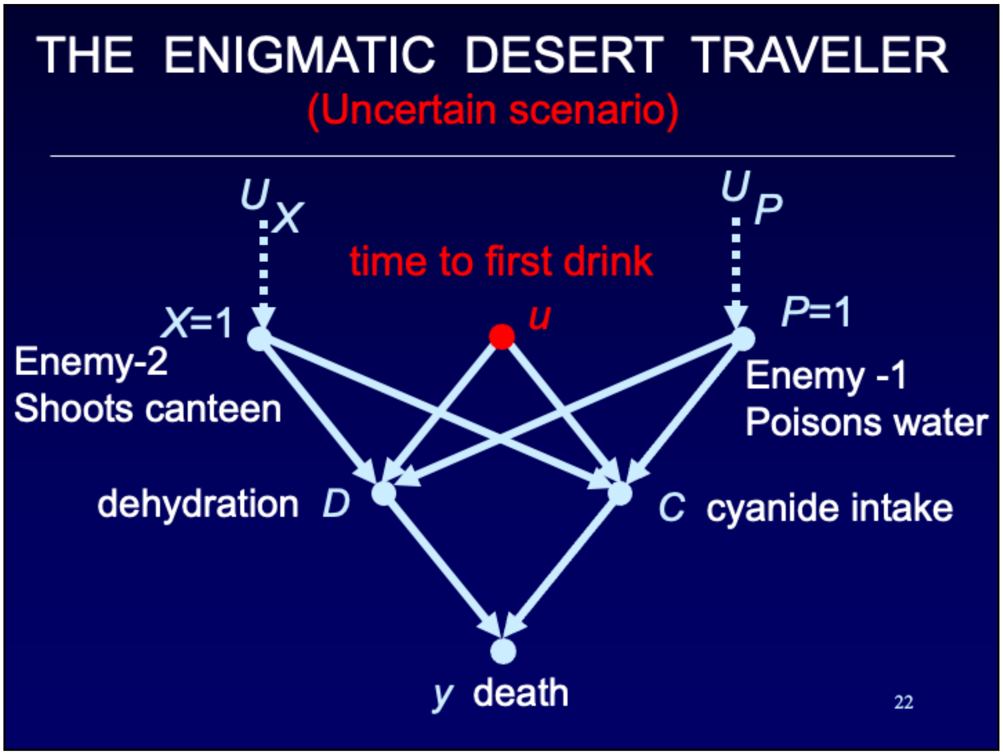


THE DESERT TRAVELER
(The final beam)

We drop the link $C \rightarrow Y$ and we end up with a causal beam leading form shooting to death through dehydration.

In this final model we conduct the counterfactual test and find that the test is satisfied since $Y = X$.

This gives us the asymmetry we need to classify the shooter as the cause of death, not the poisoner, though none meets the counterfactual test for necessity on a global scale -- the asymmetry emanates from structural information.

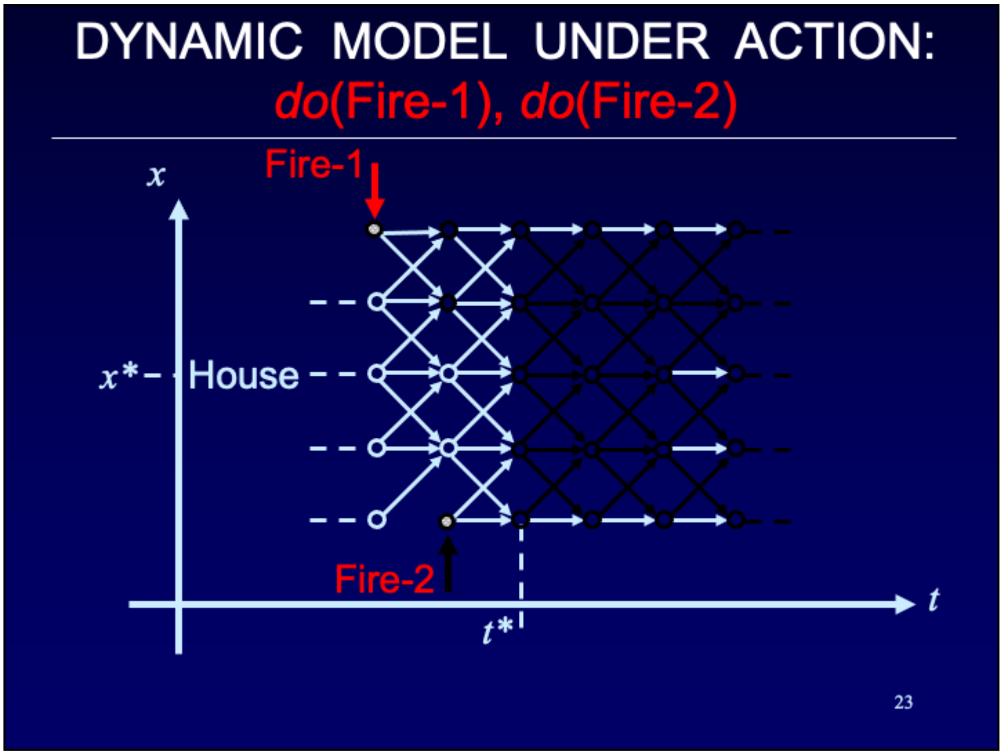


THE ENIGMATIC DESERT TRAVELER

(Uncertain scenario)

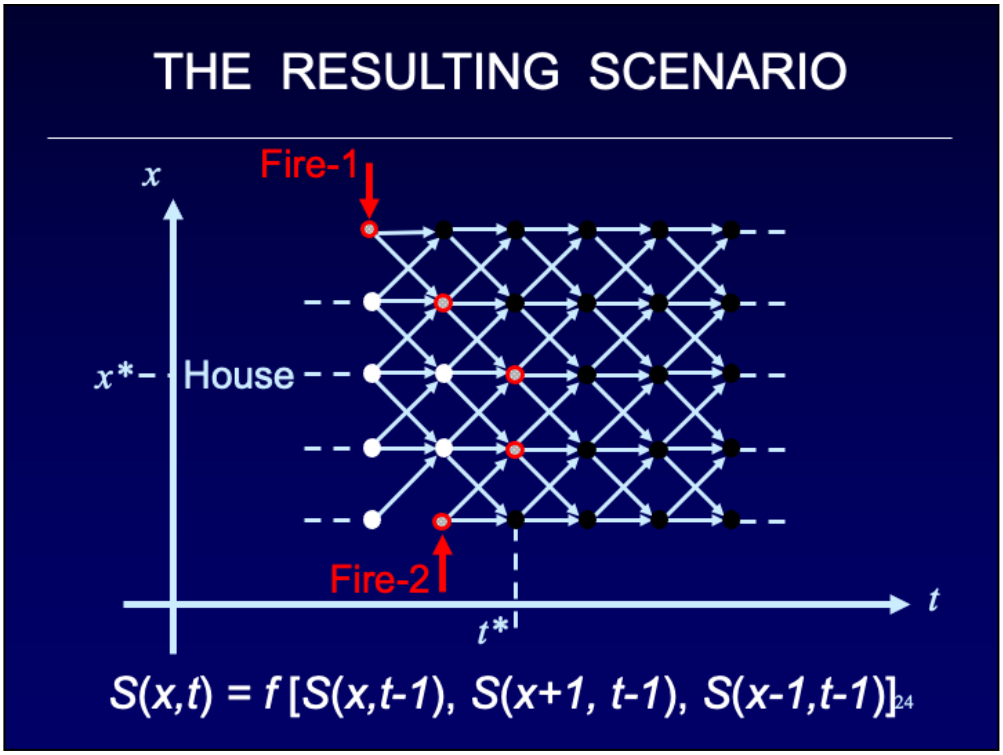
Things will change of course, if the we do not know whether the traveler craved for water before the shot.

Our uncertainty can be model by introducing a background variable, *U*, to represent the time when the traveler first reached for drink.



DYNAMIC MODEL UNDER ACTION: *do*(Fire-1), *do*(Fire-2)

To test which action was the cause of the damage, we first simulate the two actions at their corresponding times and locations, as shown in the slide.



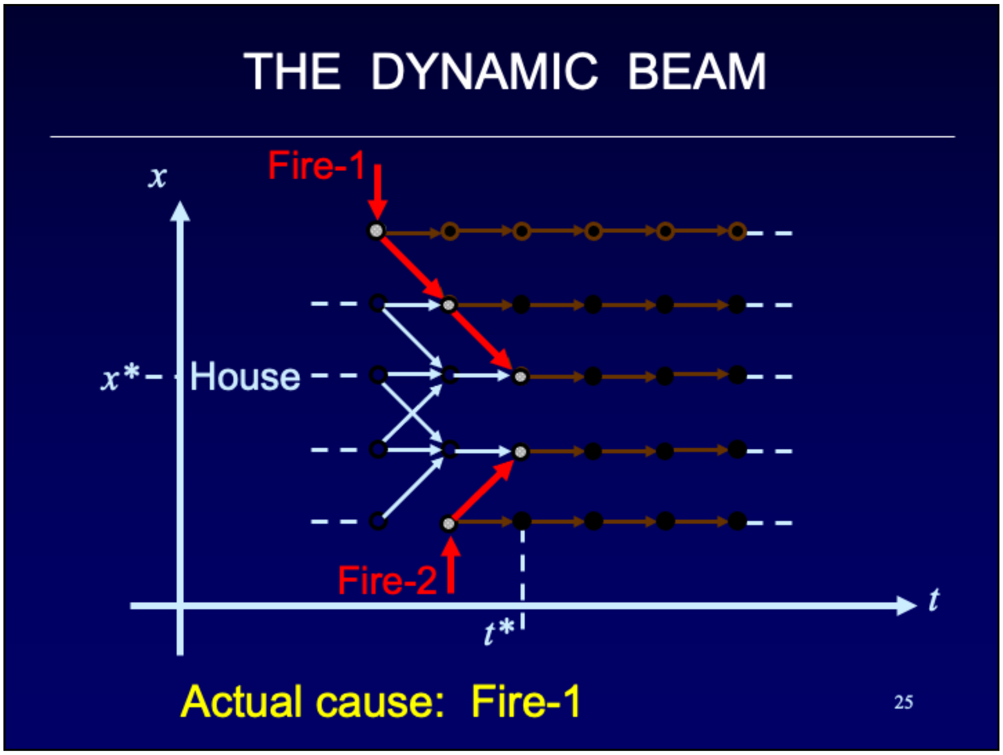
THE RESULTING SCENARIO

$S(x,t) = f\,[S(x,t\text{-}1),\ S(x+1,\ t\text{-}1),\ S(x\text{-}1,t\text{-}1)]$

To apply the beam test to this dynamic model, we first need to compute the scenario that unfolds from these actions.

Applying the process-equations recursively, from left to right, simulates the propagation of the two fires, and gives us the actual value for each variable in this spatio-temporal domain.

Here, white represents unconsumed regions, red represents regions on fire, and brown represent burned regions.



THE DYNAMIC BEAM

Actual cause: Fire-1

25

We are now ready to construct the beam and conduct the test for causation.

The resulting beam is unique and is shown in the slide above.

The symmetry is clearly broken -- there is a dependence between Fire-1 and the conditions of the house x^* at all times $t \geq t^*$; no such dependence exists for Fire-2.

Thus, the earlier fire is proclaimed the actual cause of the house burning.