REASONING WITH CAUSE AND EFFECT

Judea Pearl
University of California
Los Angeles

The subject of my lecture this evening is CAUSALITY. It is not an easy topic to speak about, but it is a fun topic to speak about. It is not easy because, like religion, sex and intelligence, causality was meant to be practiced, not analyzed. And it is fun, because, like religion, sex and intelligence, emotions run high, examples are plenty, there are plenty of interesting people to talk to, and above all, an exhilarating experience of watching our private thoughts magnified under the microscope of formal analysis.

David Hume
(*1711–1776*)

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.

HUME'S LEGACY

1. Analytical vs. empirical claims
2. Causal claims are empirical
3. All empirical claims originate from experience.

1. He made a sharp distinction between analytical and empirical claims --- the former are product of thoughts, the latter matter of fact.

2. He classified causal claims as empirical, rather than analytical.

3. He identified the source of all empirical claims with human experience, namely sensory input.

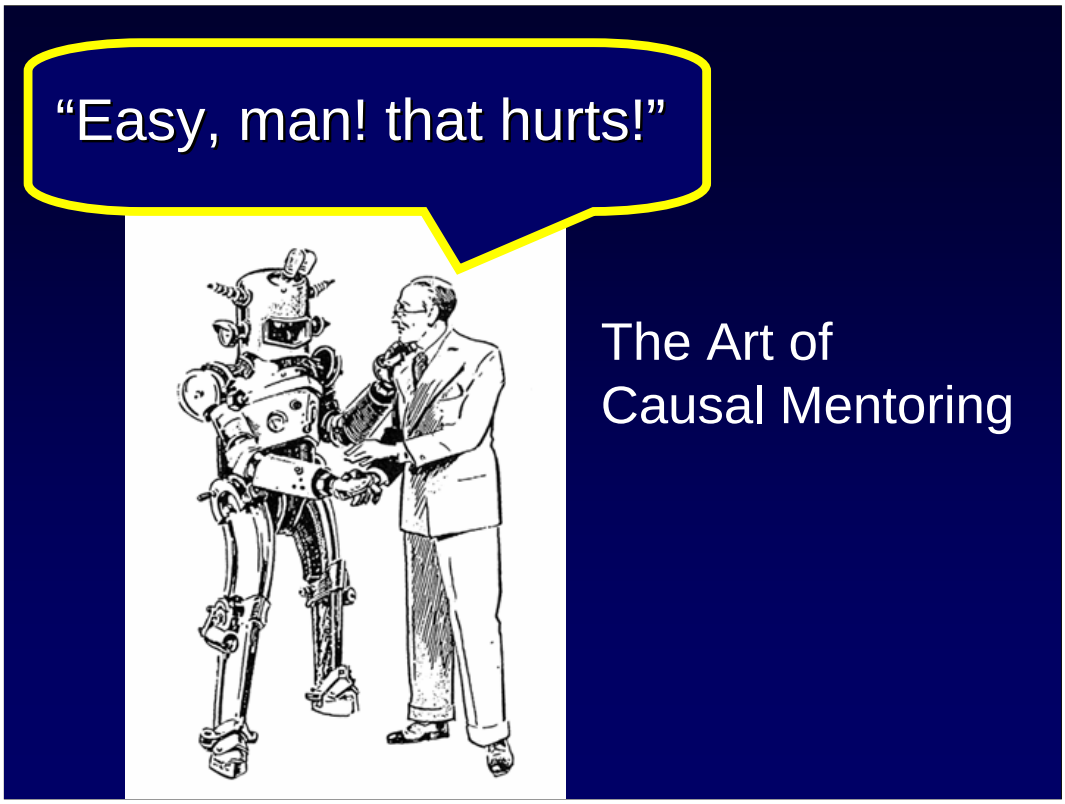
Putting (2) and (3) together have left philosophers baffled, for over two centuries, over two major riddles:

THE TWO RIDDLES OF CAUSATION

- What empirical evidence legitimizes a cause-effect connection?
- What inferences can be drawn from causal information? and how?

What gives us in AI the audacity to hope that today, after 2 centuries of philosophical debate, we can say something useful on this topic, is the fact that, for us, the question of causation is not purely academic.

We must build machines that make sense of what goes on in their environment, so they can recover when things do not turn out exactly as expected.



And we must build machines that understand causal talk, when we have the time to teach them what we know about the world. Because the way WE COMMUNICATE about the world is through this strange language called causation.

This pressure to build machines that both learn about and reason with cause and effect, something that David Hume did not experience, now casts new light on the riddles of causation, colored with engineering flavor.

OLD RIDDLES IN NEW DRESS

1. How should a robot acquire causal information from the environment?
2. How should a robot process causal information received from its creator-programmer?

I will not touch on the first riddle, because David Heckerman covered this topic on Tuesday evening, both eloquently and comprehensively.

I want to discuss primarily the second problem:

How we go from facts coupled with causal premises to conclusions that we could not obtain from either component alone.

On the surface, the second problem sounds trivial, take in the causal rules, apply them to the facts, and derive the conclusions by standard logical deduction.

But it is not as trivial as it sounds. The exercise of drawing the proper conclusions from causal inputs has met with traumatic experiences in AI.

CAUSATION AS A PROGRAMMER'S NIGHTMARE

Input:

1. “If the grass is wet, then it rained”
2. “if we break this bottle, the grass will get wet”

Output:

“If we break this bottle, then it rained”

One of my favorite example is the following:

(Wet grass example on slide).

CAUSATION AS A PROGRAMMER'S NIGHTMARE (Cont.) (Lin, 1995)

Input:

1. A suitcase will open iff both locks are open.
2. The right lock is open

Query:

What if we open the left lock?

Output:

The right lock might get closed.

Another troublesome example, which I first saw in Lin's paper of IJCAI-95 goes like that: (Suitcase Example on slide)

In these two examples, the strange output is derived from solid logical principles, chaining in the first, constraint-satisfaction in the second, yet, we feel that there is a missing ingredient there which the computer did not quite grasp, and that it has to do with causality.

Evidently there is some valuable information conveyed by causal vocabulary which is essential for correct understanding of the input. What is it that information? And what is that magic logic that should permit a computer to select the right information, and what is the semantics behind such logic ?

It is this sort of questions that I would like to address in my talk this evening, because I know that many people in this community are dealing with such questions, and have made promising proposals for answering them. Most notably are people working in qualitative physics, troubleshooting, planning under uncertainty, modeling behavior of physical systems, constructing theories of action and change, and perhaps even those working in natural language understanding, because our language is loaded with causal expressions.

Since 1990, I have examined many (though not all) of these proposals, together with others that have been suggested by philosophers and economists, and I have extracted from them a small set of basic principles which I would like to share with you tonight. I am now convinced, that the entire story of causality unfolds from just three basic principles:

THE BASIC PRINCIPLES

Causation = encoding of behavior under interventions

Interventions = surgeries on mechanisms

Mechanisms = stable functional relationships
= equations + graphs

1. The central theme is to view causality a computational scheme devised to facilitate prediction of the effects of actions.
2. I use the term "INTERVENTION" here, instead of ACTION, to emphasize that the role of causality can best be understood if we view actions as external entities, originating from outside our theory, not as a mode of behavior within the theory.

To understand the three principles it is better to start from the end and go backwards.

(3) The world is organized in the form of stable mechanisms, or physical laws, which are sufficient for determining all event that are of interest to the modeler. The mechanisms are autonomous – like mechanical linkages in a machine, or logic gates in electronic circuits -- we can change one without changing the others.

(2) Interventions ALWAYS involve the breakdown of mechanism. I will call this breakdown a "surgery" to emphasize its dual painful/remedial character.

(1) Causality tells us which mechanism is to be surgically modified by any given action.

These principles can be encapsulated neatly and organized in a mathematical object called a CAUSAL MODEL.

WHAT'S IN A CAUSAL MODEL?

Oracle that assigns truth value to causal sentences:

- Action sentences: *B* if we do *A*.
- Counterfactuals: $\neg B \Rightarrow B$ if it were *A*.
- Explanation: *B* occurred because of *A*.

Optional: with what probability?

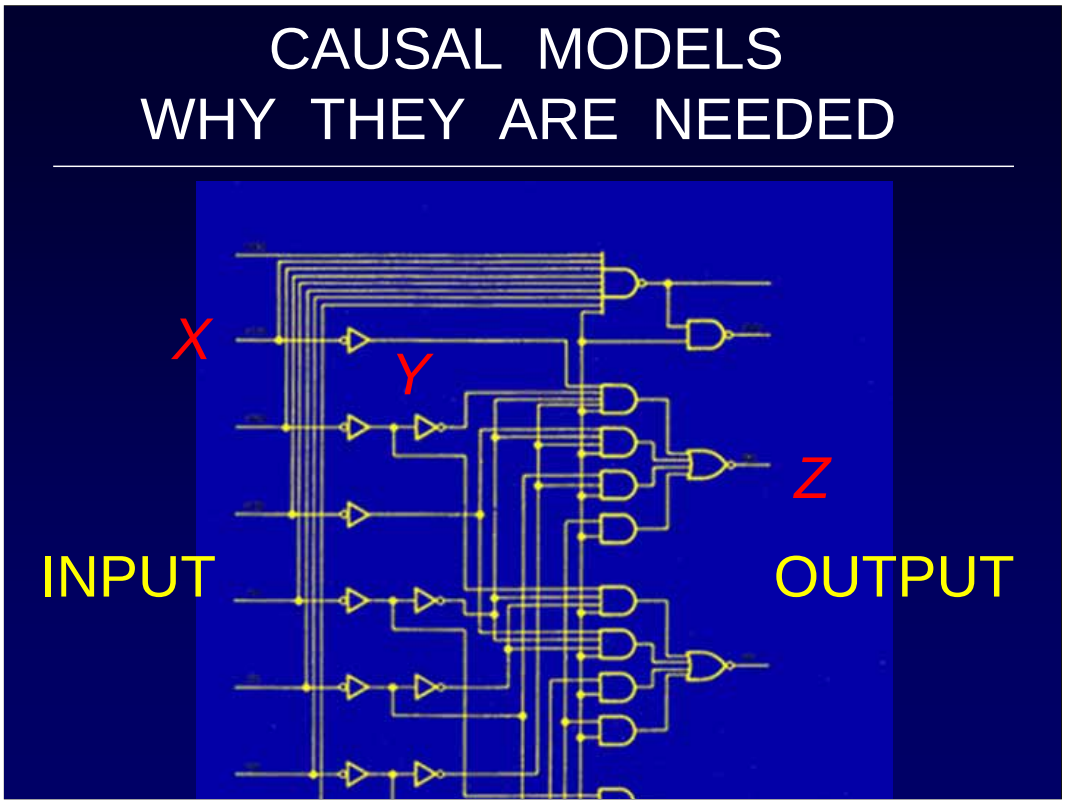
The purpose of a model is to assign truth values to sentences in a given language. If models in standard logic assign truth values to logical formulas, causal models embrace a wider class of sentences, including those that we normally classify as CAUSAL. What are those sentences?

Actions: *B* will be true if we do *A*.
Counterfactuals: *B* would be different if A were true
Explanation: *B* because of *A*

There could be more, but I will concentrate on these three, because they are commonly used, and because I believe that all other causal sentences can be reduced to these three.

The difference between action and counterfactuals is merely that the clash between the antecedant and the current state of affairs is explicit.

To allay any fear that a causal model is some complicated mathematical object, let me exemplify the beast with two familiar examples.



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?

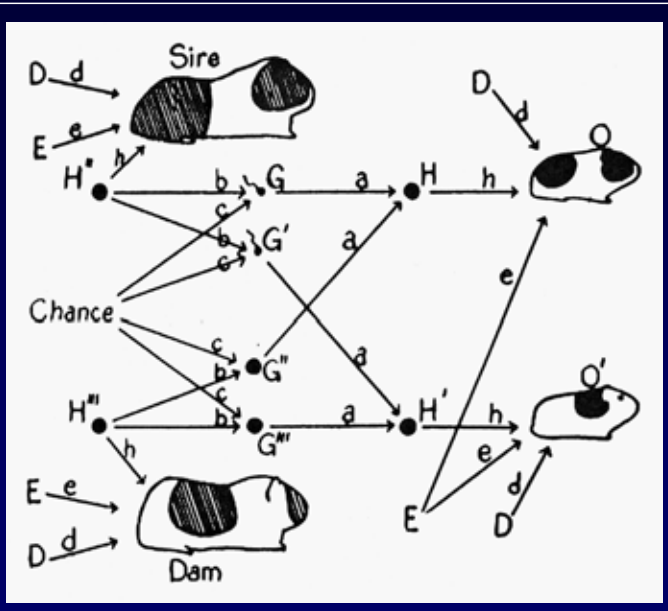
Through two encoding tricks:

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)

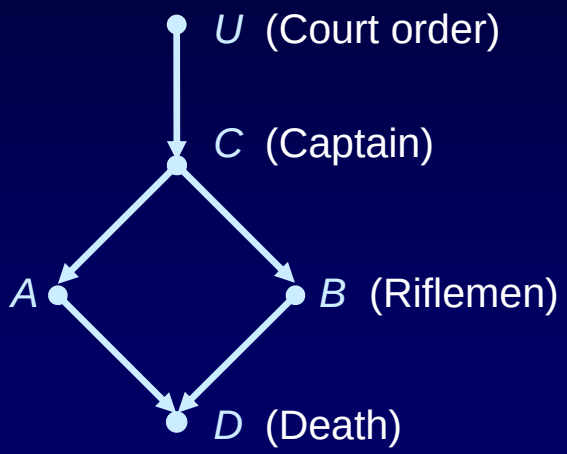


As another example, here is the first causal model that was put down on paper: Sewal Wright's path diagram, showing how the fur pattern of the litter guinea pigs is determined by various genetic and environmental factors. Again, (1) it qualifies as a causal model, (2) the algebraic equations in themselves 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 -

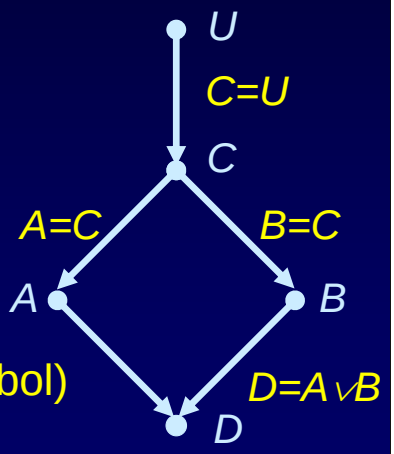
CAUSAL MODELS AT WORK
(The impatient firing-squad)



Though not many of us have had direct experience with this story, it is nevertheless familiar and vivid. It describes a tense moment in the life of a gentleman facing a firing squad.

CAUSAL MODELS AT WORK (Glossary)

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.

SENTENCES TO BE EVALUATED

S1. prediction: $\neg A \Rightarrow \neg D$

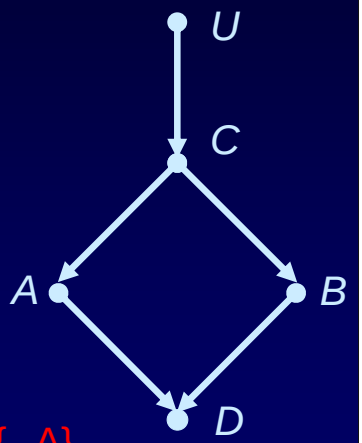
S2. abduction: $\neg D \Rightarrow \neg C$

S3. transduction: $A \Rightarrow B$

S4. action: $\neg C \Rightarrow D_A$

S5. counterfactual: $D \Rightarrow D_{\{\neg A\}}$

S6. explanation: Caused(A, D)



This slide lists the sentences we wish to evaluate. The simplest one are S1-S3 which are standard.

Next in difficulty are action sentences S4, -- requiring some causal information; next are counterfactuals S5 -- requiring more detailed causal information, and the hardest being explanation sentences (S6) whose semantics is still not completely settled -- to be discussed at the last part of the lecture.

STANDARD MODEL FOR STANDARD QUERIES

S1. (prediction): If rifleman-*A* shot, the prisoner is dead,

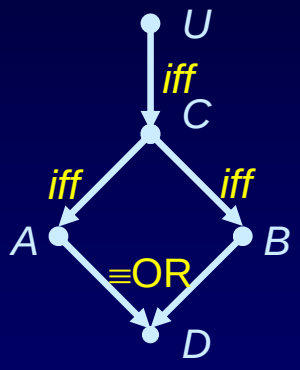
$$A \Rightarrow D$$

S2. (abduction): If the prisoner is alive, then the Captain did not signal,

$$\neg D \Rightarrow \neg C$$

S3. (transduction): If rifleman-*A* shot, then *B* shot as well,

$$A \Rightarrow B$$



Sentences S1-S3 involve standard logical connectives, because they deal with inferences from beliefs to beliefs about a static world.

WHY CAUSAL MODELS? GUIDE FOR SURGERY

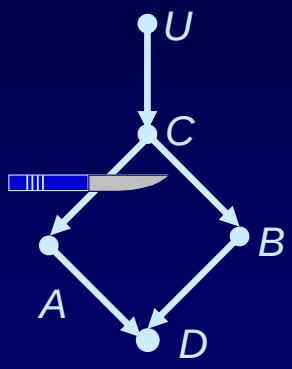
S4. (action):

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 is the first chance we have to witness what information a causal model provides on top of a logical model.

Shooting with no signal constitutes a blatant violation of one mechanism in the story: rifleman-*A*'s commitment to follow the Captain's signal. Violation renders this mechanism inactive, hence we must excise the corresponding equation from the model, using this knife, and replace it by a new mechanism: A = TRUE.

.

WHY CAUSAL MODELS? GUIDE FOR SURGERY

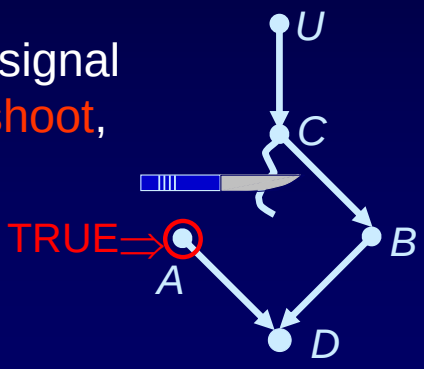
S4. (action):

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:

$\neg C \Rightarrow \neg B_A$

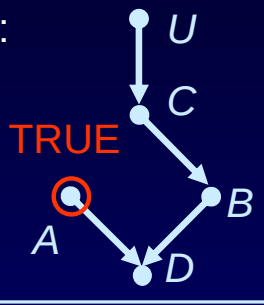


This surgery also suppresses abduction; from seeing *A* shoot we can infer that *B* shot as well (recall $A \Rightarrow B$), but from MAKING *A* shoot we can no longer infer what *B* does.

MUTILATION IN SYMBOLIC CAUSAL MODELS

Model M_A (Modify A=C):

	(U)
$C = U$	(C)
$A = C$	(A)
$B = C$	(B)
$D = A \vee B$	(D)



Facts: $\neg C$

Conclusions: ?

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

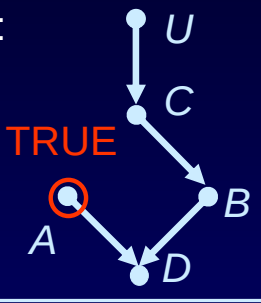
Everything we do with graphs we can, of course, do with symbols. We need however be careful to distinguish facts from rules (domain constraints), and to mark the privileged element in each rule (the left-hand-side).

Here we see for the first time the role of causal order: Which mechanism should be excised by the action *do*(*A*)? (note that *A* appears in two equations) The answer: Excise the equation in which *A* is the privileged variable.

MUTILATION IN SYMBOLIC CAUSAL MODELS

Model M_A (Modify A=C):

	(U)
$C = U$	(C)
~~$A = C$~~	(A)
$B = C$	(B)
$D = A \vee B$	(D)



Facts: $\neg C$

Conclusions: ?

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

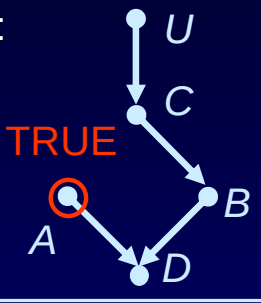
Once we create the mutilated model M_A, we draw the conclusions by standard deduction and easily confirm:

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

MUTILATION IN SYMBOLIC CAUSAL MODELS

Model M_A (Modify A=C):

	(U)
$C = U$	(C)
~~$A = C$~~	
A	(A)
$B = C$	(B)
$D = A \vee B$	(D)



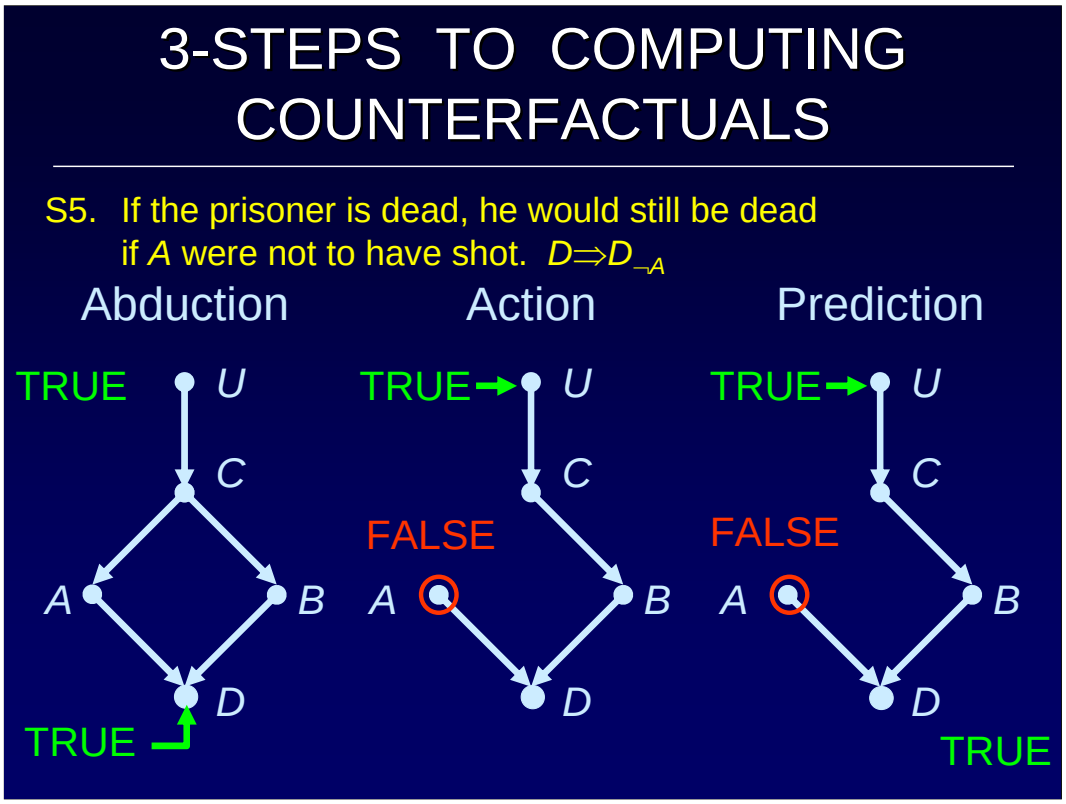
Facts: $\neg C$

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, $\neg C \Rightarrow D_A \,\&\, \neg B_A$

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.



Consider now our counterfactual sentence

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

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

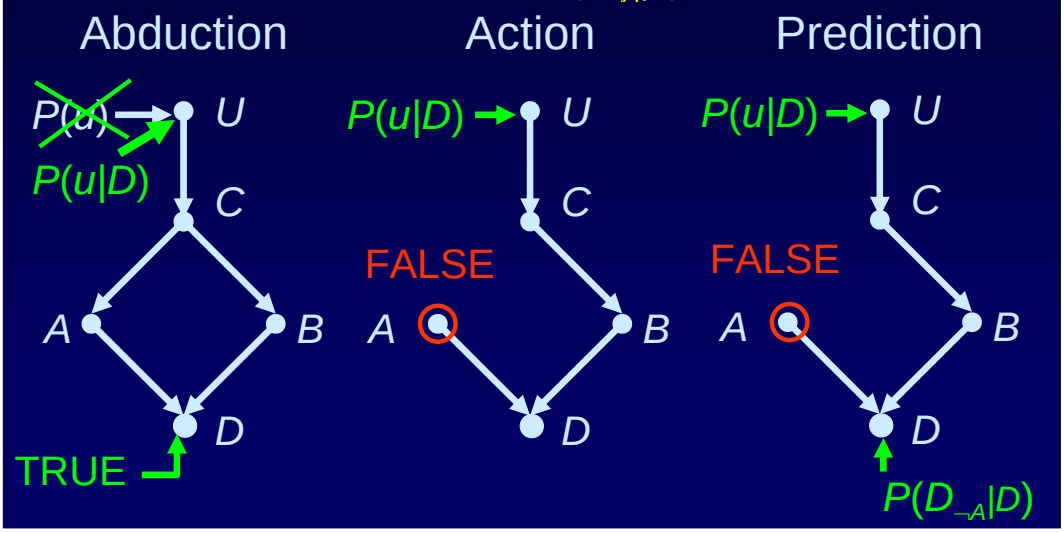
1 Abduction: Interpret the past in light of the evidence

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

3.Prediction: Project the consequences to the future.

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.

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)

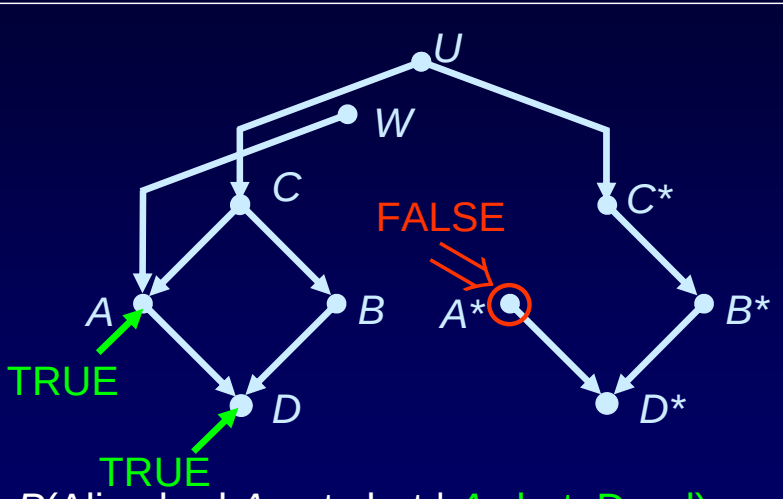
Facts: D

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)

PROBABILITY OF COUNTERFACTUALS
THE TWIN NETWORK



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

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.

CAUSAL MODEL (FORMAL)

$M = \langle U, V, F \rangle$ or $\langle U, V, F, P(u) \rangle$

U - Background variables

V - Endogenous variables

F - Set of functions $\{U \times V \setminus V_i \rightarrow V_i\}$

$$v_i = f_i(pa_i, u_i)$$

Submodel: $M_x = \langle U, V, F_x \rangle$, representing $do(x)$

F_x= Replaces equation for X with $X=x$

Actions and Counterfactuals:

$Y_x(u)$ = Solution of Y in M_x

$P(y \mid do(x)) \stackrel{\Delta}{=} P(Y_x=y)$

Let us now summarize the formal elements involved in this causal exercises.

(Reader: See proceedings paper for technical details)

WHY COUNTERFACTUALS?

Action queries are triggered by (modifiable) observations, demanding abductive step, i.e., counterfactual processing.

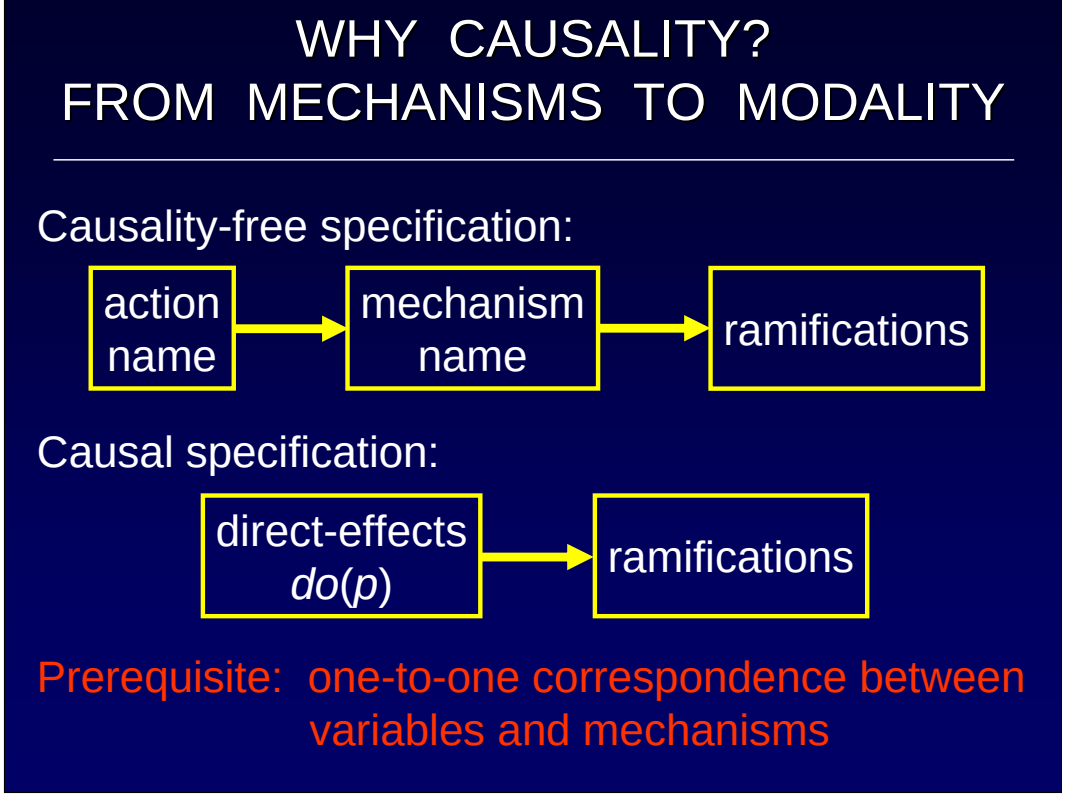
E.g., Troubleshooting

Observation:	The output is low
Action query:	Will the output get higher – if we replace the transistor?
Counterfactual query:	Would the output be higher – had the transistor been replaced?

We have seen that action queries can be answered in one step: Standard deduction on a mutilated submodel. Counterfactual queries, on the other hand, required a preparatory stage of abduction. The question naturally arises: who needs counterfactuals? and why spend time on computing such convoluted sentences? It turns out that counterfactuals are commonplace and pure action sentences are a fiction. Action queries are brought into focus by certain undesired observations, potentially modifiable by the actions. The step of abduction, which is characteristic of counterfactual queries, cannot be disposed of, and must therefore precede the surgery step. This makes most action queries semantically identical to counterfactual queries.

The two sentences in this example from troubleshooting are equivalent: Both demand abductive step to account for the observation.

And this unfortunately complicates things a bit. In probabilistic analysis, functional specification is needed, conditional-probabilities alone are not sufficient for answering observation-triggered action queries. In symbolic analysis: abnormalities must be explicated in functional details; the catch-all phrase "AND NOT ABNORMAL *p*" is not sufficient.



This brings us to the million dollar question: WHY CAUSALITY

So far we have discussed actions, counterfactuals, surgeries, mechanism, abduction etc, but is causality really necessary?

Indeed, if we know which mechanisms each action modifies, and the nature of the modification, we can avoid all talk of causation -- the ramification of each action can be obtained by simply MUTILATING then SIMULATING. The price we pay is that we need to specify an action, not by its direct effects but, rather, by the mechanisms which the action modifies.

For example, instead of saying "this action moves the coffee cup to location *X*" I would need to say "this action neutralizes the static friction of the coffee cup, and replaces it with a forward acceleration a for a period of 1 second, followed by deceleration for a period of 2 seconds ...".

This is awfully clumsy: Most mechanisms do not have names in non-technical languages, and when they do, the names do not match the granularity of ordinary language. Causality enables us to reason correctly about actions while keeping the mechanism IMPLICIT. All we need to specify is the action's direct effects, the rest follows by mutilation-simulation.

But to figure out which mechanism deserves mutilation, there must be one-to-one correspondence between variables and mechanisms. Is that a realistic requirement?

In general, NO. A random collection of n equations on n variables would not enjoy this property. Even a resistive network (e.g., voltage divider) does not enjoy it. But from the fact that causal thinking is so pervasive in our language we may conclude that our understanding of the world is more structured, and that it does enjoy the 1-1 correspondence . We say: "raise taxes", "clean your face", "make him laugh“ and in general, *do*(*p*) and, miraculously, people understand us without asking for mechanism name. (H. Simon devised a test for deciding when 1-1 correspondence exists, see proceedings paper)

SURGERY IN STRIPS STYLE

Action: $do(V_i = v^*)$
Current state: $V_i(u) = v$

DELETE-LIST	ADD-LIST
$V_i = v$ + ramifications	$V_i = v^*$ + ramifications
MECHANISM DELETE-LIST	**MECHANISM ADD-LIST**
$v_i = f_i(pa_i, u_i)$	$f_i(\cdot) = v^*$

Perhaps the best "AI proof" of the ubiquity of the modality do(p) is the existence of the language STRIPS, in which actions are specified via direct effects -- the ADD-LIST.

Let us compare causal surgeries to STRIPS surgeries. Both accept actions as modalities, both perform surgeries BUT: STRIPS perform the surgery on propositions (the DELETE-LIST) while causal theories, by exploiting their 1-1 correspondence, can infer the mechanism to be excised and performs the surgery on mechanisms, not on propositions. The result is that ramifications need not be specified, they can be inferred from the MUTILATE-SIMULATE cycle.

MID-STORY OUTLINE

Background:
- From Hume to robotics

Semantics and principles:
- Causal models, Surgeries, Actions and Counterfactuals

Applications I:
- Evaluating Actions and Plans from Data and Theories

Applications II:
- Finding Explanations and Single-event Causation

This brings us to our mid-story outline. We have talked about the story of causation from Hume to robotics, we have discussed the semantics of causal utterances and the principles behind the interpretation of action and counterfactual sentences, and now it is time to ask about the applications of these principles.

I will talk about two types of applications, the first relates to the evaluation of actions and the second to finding explanations.

The next slides provides a somewhat more elaborate list of these applications. with slide 48 (Applications-II)

APPLICATIONS

1. Predicting effects of actions and policies
2. Learning causal relationships from assumptions and data
3. Troubleshooting physical systems and plans
4. Finding explanations for reported events
5. Generating verbal explanations
6. Understanding causal talk
7. Formulating theories of causal thinking

Let us talk about item 1 for a minute. We saw that if we have a causal model M, then predicting the ramifications of an action is trivial -- mutilate and solve.

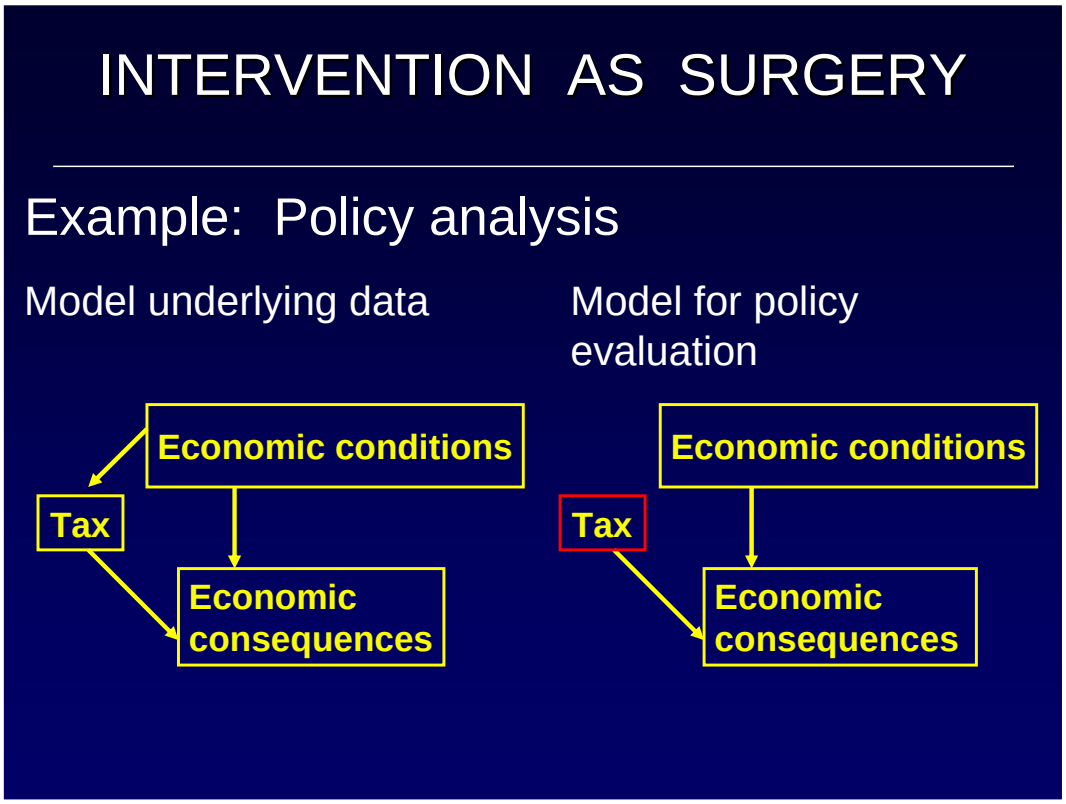
If instead of a complete model we only have a probabilistic model, it is again trivial: we mutilate and propagate probabilities in the resultant causal network.

The important point is that we can specify knowledge using causal vocabulary, and can handle actions that are specified as modalities.

But what if we do not have even a probabilistic model? This is where item 2 comes in.

In certain applications we are lucky to have data that may supplement missing fragments of the model, and the question is whether the data available is sufficient for computing the effect of actions.

Let us illustrate this possibility in a simple example taken from economics:



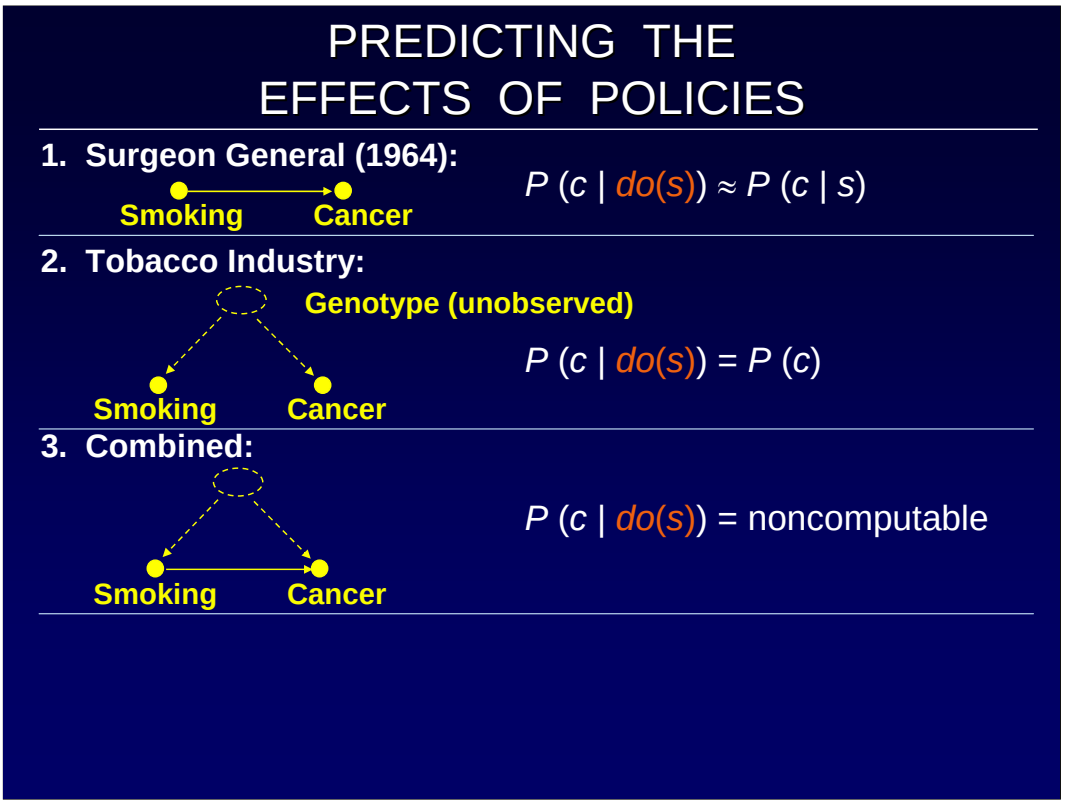
Economic policies are made in a manner similar to the way actions were taken in the firing squad story: Viewed from the outside, they are taken in response to economic indicators or political pressure, while viewed from the policy maker perspective, the next decision is chosen under the pretense of free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



In 1964, the Surgeon General issued a report linking cigarette smoking to death, cancer and most particularly, lung cancer.

The report was based on non-experimental studies, in which a strong correlation was found between smoking and lung cancer, and the claim was that the correlation found is causal, namely: If we ban smoking, the rate of cancer cases will be roughly the same as the one we find today among non-smokers in the population.

These studies came under severe attacks from the tobacco industry, backed by some very prominent statisticians, among them Sir Ronald Fisher.

The claim was that the observed correlations can also be explained by a model in which there is no causal connection between smoking and lung cancer. Instead, an unobserved genotype might exist which simultaneously causes cancer and produces an inborn craving for nicotine.

Formally, this claim would be written in our notation as: P(cancer | do(smoke)) = P(cancer) stating that making the population smoke or stop smoking would have no effect on the rate of cancer cases.

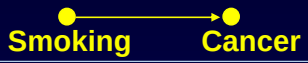
Controlled experiment could decide between the two models, but these are impossible, and now also illegal to conduct.

This is all history. Now we enter a hypothetical era where representatives of both sides decide to meet and iron out their differences.

The tobacco industry concedes that there might be some weak causal link between smoking, and cancer and representatives of the health group concede that there might be some weak links to genetic factors. Accordingly, they draw this combined model (no. 3 in the slide), and the question boils down to assessing, from the data, the strengths of the various links.

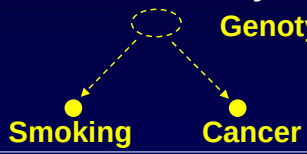
PREDICTING THE EFFECTS OF POLICIES

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



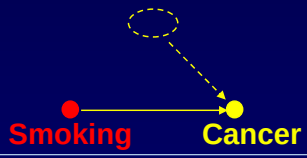
$P(c \mid do(s)) \approx P(c \mid s)$

2. **Tobacco Industry:**



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

3. **Combined:**

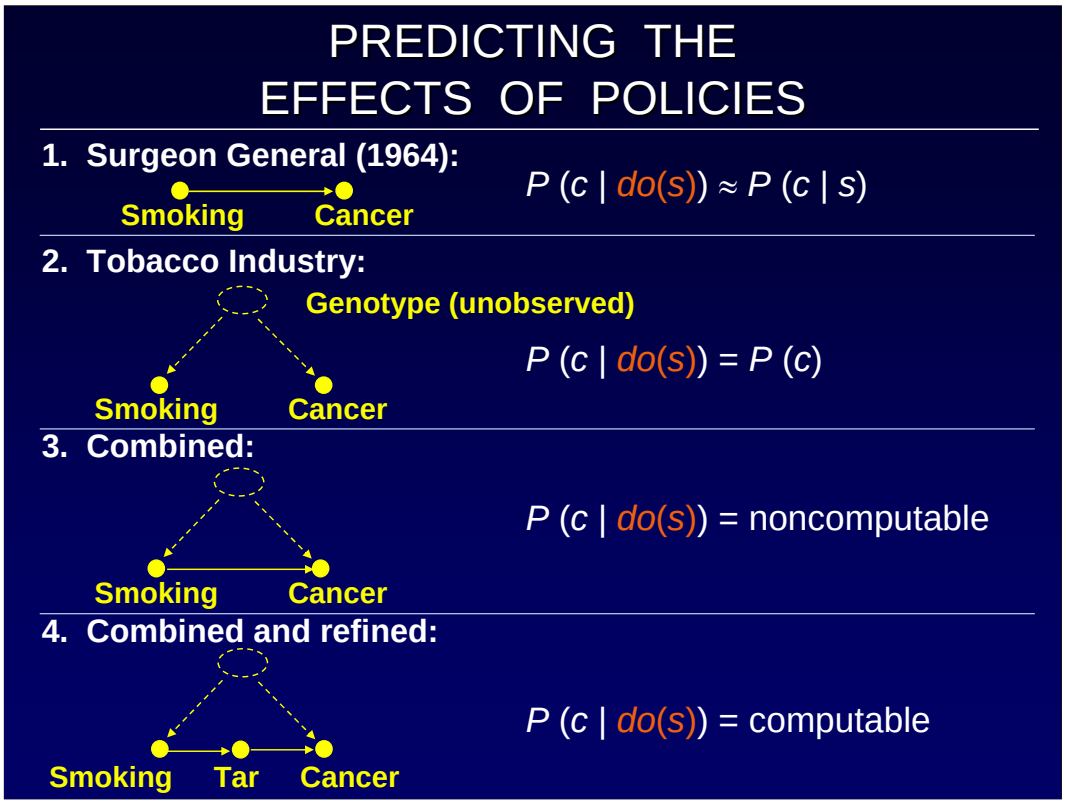


$P(c \mid do(s))$ = noncomputable

Or, speaking in mutilation language, the question boils down to assessing the effect of smoking in the mutilated model shown here, from data taken under the wholesome model shown before.

They submit the query to a statistician and the answer comes back immediately: IMPOSSIBLE. Meaning: there is no way to estimate the strength for the causal links from the data, because any data whatsoever can perfectly fit either one of the extreme models shown in (1) and (2).

So they give up, and decide to continue the political battle as usual.



Before parting, a suggestion comes up: perhaps we can resolve our differences if we measure some auxiliary factors,

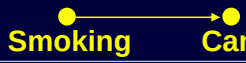
For example, since the causal link model is based on the understanding that smoking affects lung cancer through the accumulation of tar deposits in the lungs, perhaps we can measure the amount of tar deposits in the lungs of sampled individuals, and this might provide the necessary information for quantifying the links?

Both sides agree that this is a reasonable suggestion, so they submit a new query to the statistician: Can we find the effect of smoking on cancer assuming that an intermediate measurement of tar deposits is available???

The statistician comes back with good news: IT IS COMPUTABLE

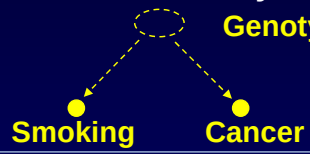
PREDICTING THE EFFECTS OF POLICIES

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



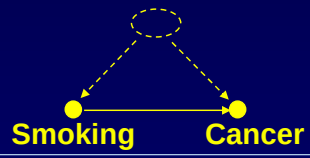
$P(c \mid do(s)) \approx P(c \mid s)$

2. **Tobacco Industry:**



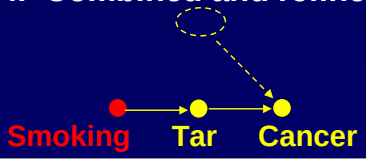
$P(c \mid do(s)) = P(c)$

3. **Combined:**



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

4. **Combined and refined:**

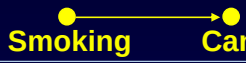


$P(c \mid do(s))$ = computable

In other words, it is possible now to infer the effect of smoking in the mutilated model shown here (No. 4), from data taken under the original wholesome model:

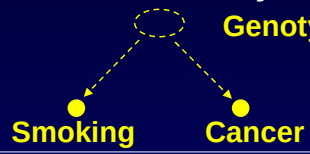
PREDICTING THE EFFECTS OF POLICIES

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



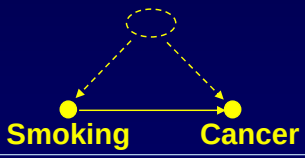
$P(c \mid do(s)) \approx P(c \mid s)$

2. **Tobacco Industry:**



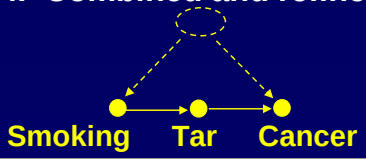
$P(c \mid do(s)) = P(c)$

3. **Combined:**



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

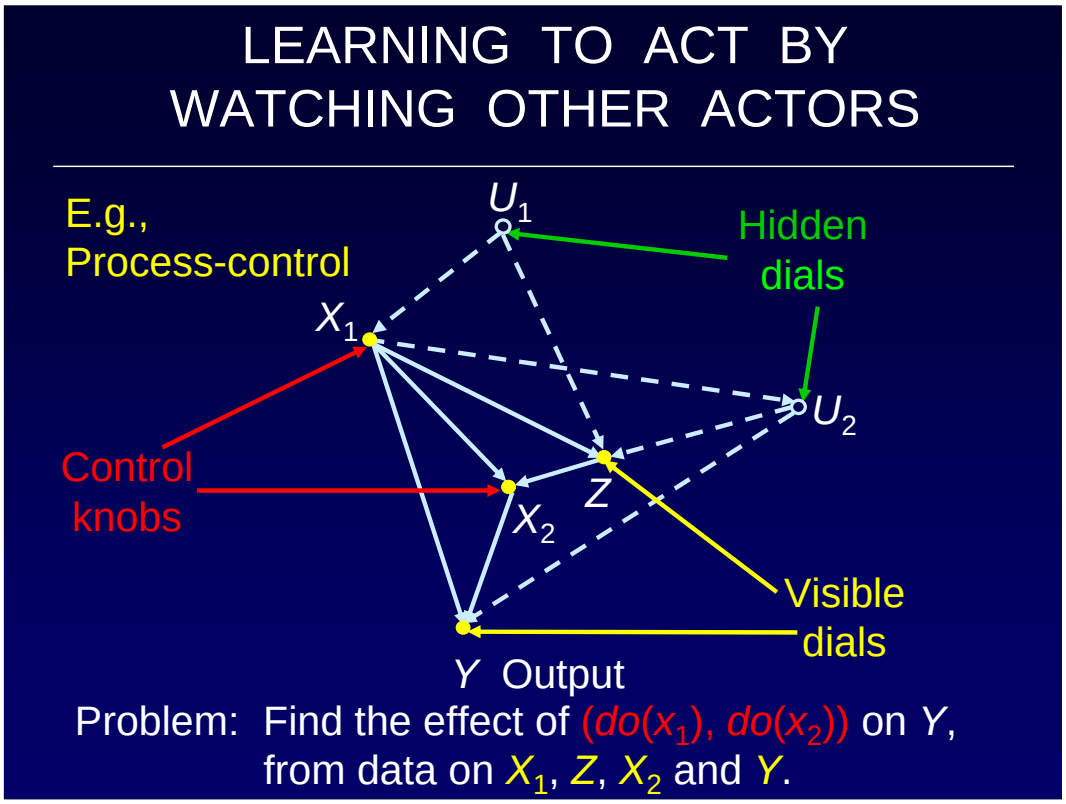
4. **Combined and refined:**



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

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

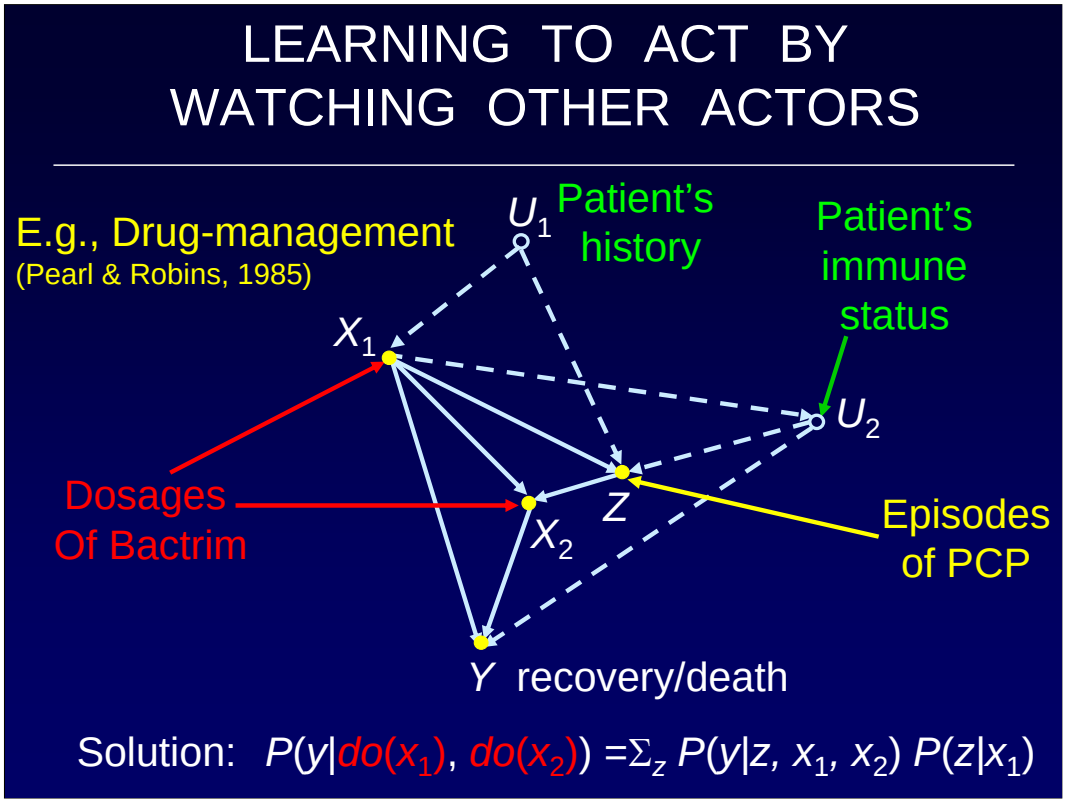
Moreover, the solution can be obtained in close mathematical form, using symbolic manipulations that mimic the surgery semantics.



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.



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.

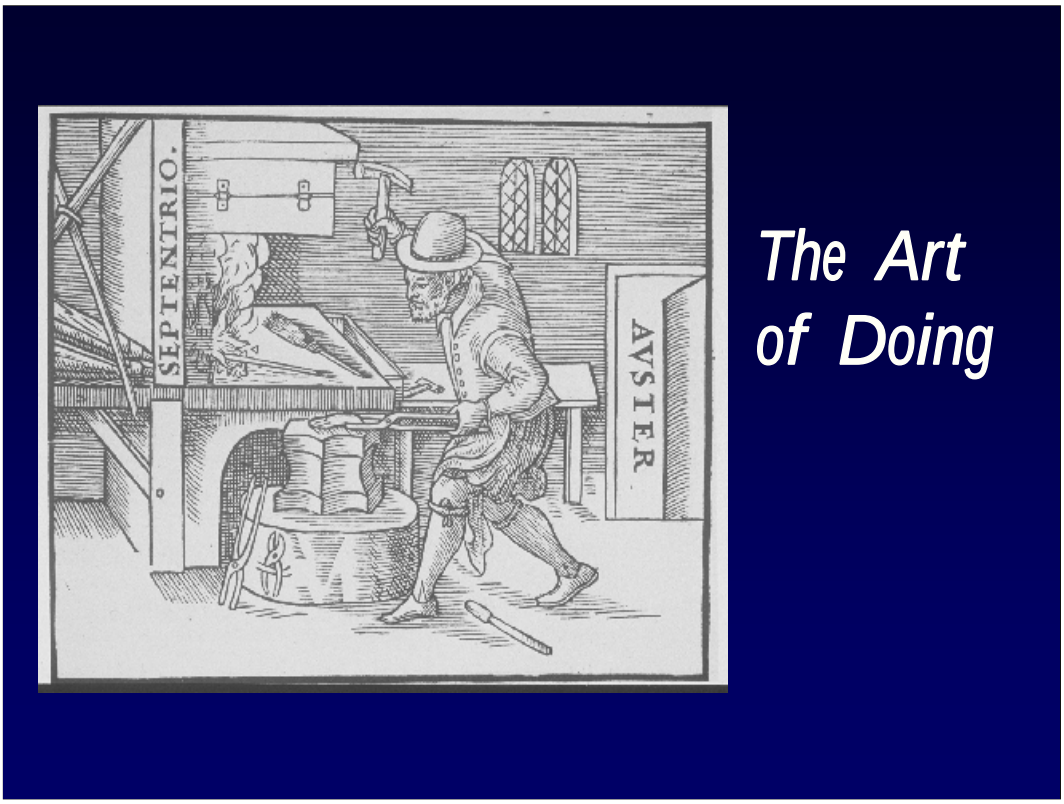
The problem we face is as follows. Assume we have collected a large amount of data on the behavior of many patients and physicians, which is summarized in the form of (an estimated) joint distribution P of the observed four variables (X1, Z, X2, Y). A new patient comes in and we wish to determine the 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?

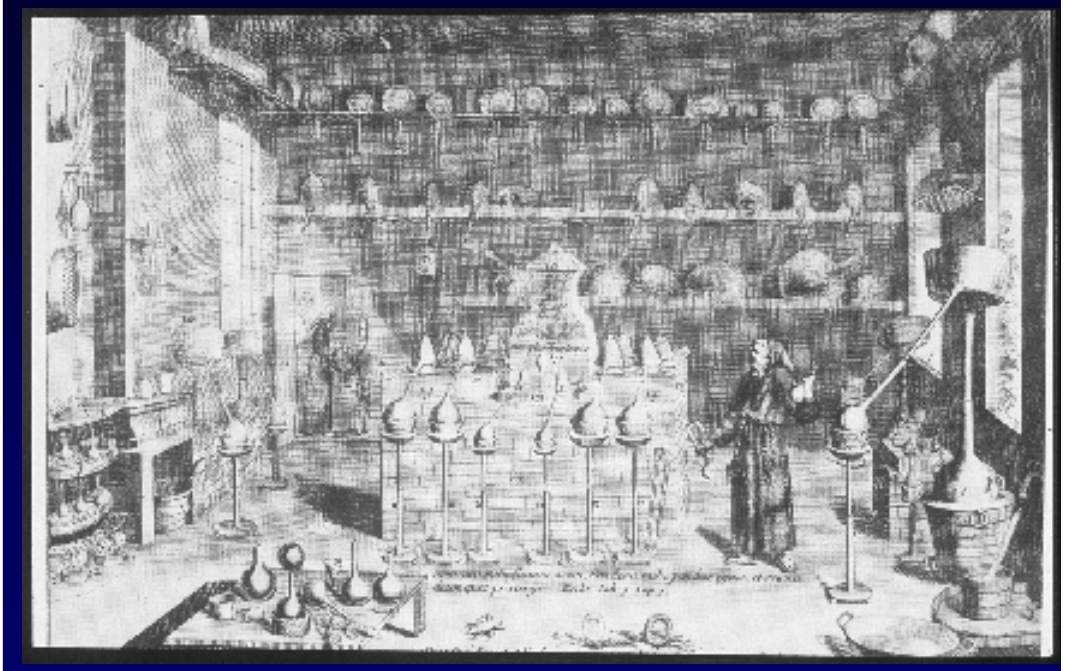
If we look at the history of science, we find to our astonishment that such a calculus does not in fact exist. It is true that Science rests on two components: One consisting of passive observations (epitomized by astronomy), and the other consisting of voluntary intervention,



Represented here by the black smith from Gilbert's De Magnet (1600)

But algebra was not equally fair to these two components. Mathematical techniques were developed exclusively to support the former (seeing) not the latter (doing) -- no calculus was developed to help this artisan make a better magnet.

Combining Seeing and Doing



Even in the laboratory, a place where the two components combine, the "seeing" part enjoys the benefits of algebra, whereas the "doing" part is at the mercy of the scientist's judgment. When actions change chemical from one test tube to another, a new set of equations becomes applicable, and algebraic techniques are useful for solving such equations. But there is no algebraic operation to represent the transfer from one test tube to another, and no algebra for selecting the correct set of equations when laboratory conditions change. Such selection has thus far relied on unaided scientific judgment.

NEEDED: ALGEBRA OF DOING

Available: algebra of seeing

e.g., What is the chance it rained if we see the grass wet?

$P(rain \mid wet) = ?$ $\{= P(wet|rain)\frac{P(rain)}{P(wet)}\}$

Needed: algebra of doing

e.g., What is the chance it rained if we make the grass wet?

$P(rain \mid do(wet)) = ?$ $\{= P(rain)\}$

Let me convince you of this misbalance using a very simple example.

If we wish to find the chance it rained, given that we "see" the grass wet, we can express our question in a formal sentence, and use the machinery of probability theory to transform the sentence into other expressions that are more convenient or informative.

But suppose we ask a different question: "What is the chance it rained if we MAKE the grass wet?"

We cannot even express our query in the syntax of probability, because the vertical bar is already taken to mean "given that we see".

We know intuitively what the answer should be: P(rain), because making the grass wet does not change the chance of rain. But can this intuitive answer, and others like it, be derived mechanically, so as to comfort our thoughts when intuition fails?

The answer is YES, and it takes a new algebra, using the do(x) operator, for which we have a simple semantics in terms of surgeries. To make it into a genuine calculus, we also need to translate the surgery semantics into rules of inference. These are described in the next slide.

RULES OF CAUSAL CALCULUS

Rule 1: Ignoring observations

$P(y \mid do\{x\}, z, w) = P(y \mid do\{x\}, w)$

if $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}}$

Rule 2: Action/observation exchange

$P(y \mid do\{x\}, do\{z\}, w) = P(y \mid do\{x\}, z, w)$

if $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}}$

Rule 3: Ignoring actions

$P(y \mid do\{x\}, do\{z\}, w) = P(y \mid do\{x\}, w)$

if $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\overline{Z(W)}}}$

The calculus consists of 3 rules that permit us to transform expressions involving actions and observations, into other expressions of this type.

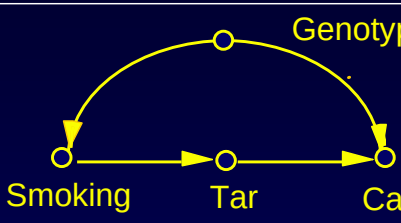
The first allows us to ignore an irrelevant observation, the third to ignore an irrelevant action, the second allows us to exchange an action with an observation of the same fact.

What are those green symbols on the right?

These are d-separation conditions in various subgraphs of the diagram that indicate when the transformation is legal.

We will see them in action in the smoking-cancer example that was discussed earlier.

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

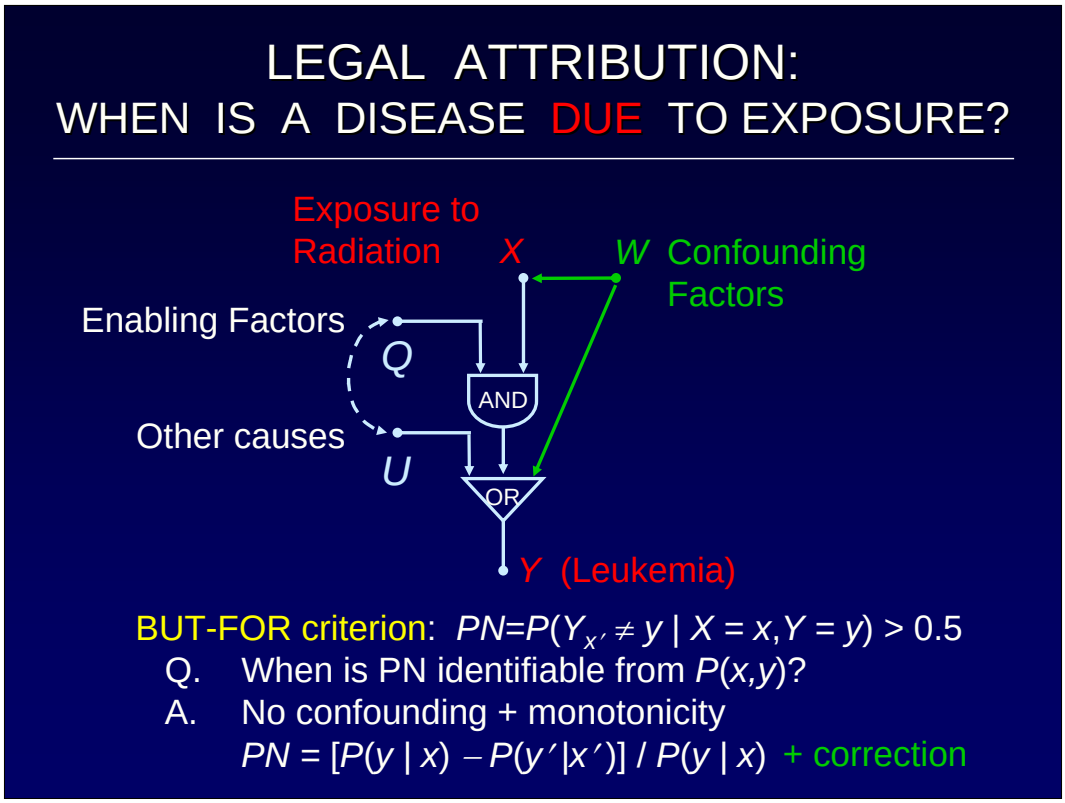
Here we see how one can prove that the effect of smoking on cancer can be determined from data on three variables: Smoking, Tar and Cancer.

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.



We now demonstrate how causal calculus can answer questions of attribution, namely finding causes of effects, rather than effects of causes.

The US army has conducted many nuclear experiments in Nevada in the period 1940-1955. Data taken over a period of 12 years indicate that fallout radiation apparently has resulted in high number of deaths from leukemia in children residing in South Utah. A law suit was filed. The question is: is the Army liable for THOSE DEATHS?

According to a fairly common judicial standard, damage will be paid iff it is more probable than not that death would not have occurred but for the action. Can we calculate this probability PN?

The answer is Yes; PN is given by the formula on the bottom of this slide. But we must assume two conditions: 1. no confounding, and 2. monotonicity (radiation cannot prevent leukemia).

This result, although it is not mentioned explicitly in any textbooks on epidemiology, statistics or Law, is not as startling as some of its corollaries: 1. There is a simple correction term to this formula that accounts for confounding. 2. There is a test for monotonicity. 3. In the absence of monotonicity, the corrected formula still provides a lower bound on the probability of causation.

Before I go to the topic of explanation, I would like to say a few words on the role of AI in such applications as statistics, public health, and social science.

One of the reasons I find these areas to be fertile grounds to try out new ideas is that, unlike AI, tangible rewards can be reaped from solving relative small problems. Problems involving barely 4 to 5 variables, which we in AI regard as toy-problems, carry tremendous payoffs in public health and social science.

Billions of dollars are invested each year on various public-health studies; is chocolate ice-cream good for you or bad for you, would red wine increase or decrease your heart rate?, etc. etc..

The same applies to the social sciences. Would increasing police budget decrease or increase crime rates? Is the Colorado school incident due to TV violence or failure of public education? The Inter-university Consortium for Political and Social Research has distributed about 800 gigabytes worth of such studies in 1993 alone.

Unfortunately the causal-analytical methodology currently available to researchers in these fields is rather primitive, and every innovation can make a tremendous difference. Moreover, the major stumbling block has not been statistical, but rather: CONCEPTUAL -- lack of semantics, and lack of formal machinery for handling causal knowledge and causal queries -- perfect for AI involvement. This has been changing recently as new techniques are beginning to emerge from AI laboratories. I predict that a quiet revolution will take place in the next decade in the way causality is handled in statistics, epidemiology, social science, economics, and business. While news of this revolution will never make it to DARPA's newsletter, and even NSF is not equipped to appreciate or support it, it will nevertheless have enormous intellectual and technological impact on our society. I spent many pages on these applications in my new book on causality (Cambridge University Press, 2000) and I hope this lecture gives you some of its flavor.

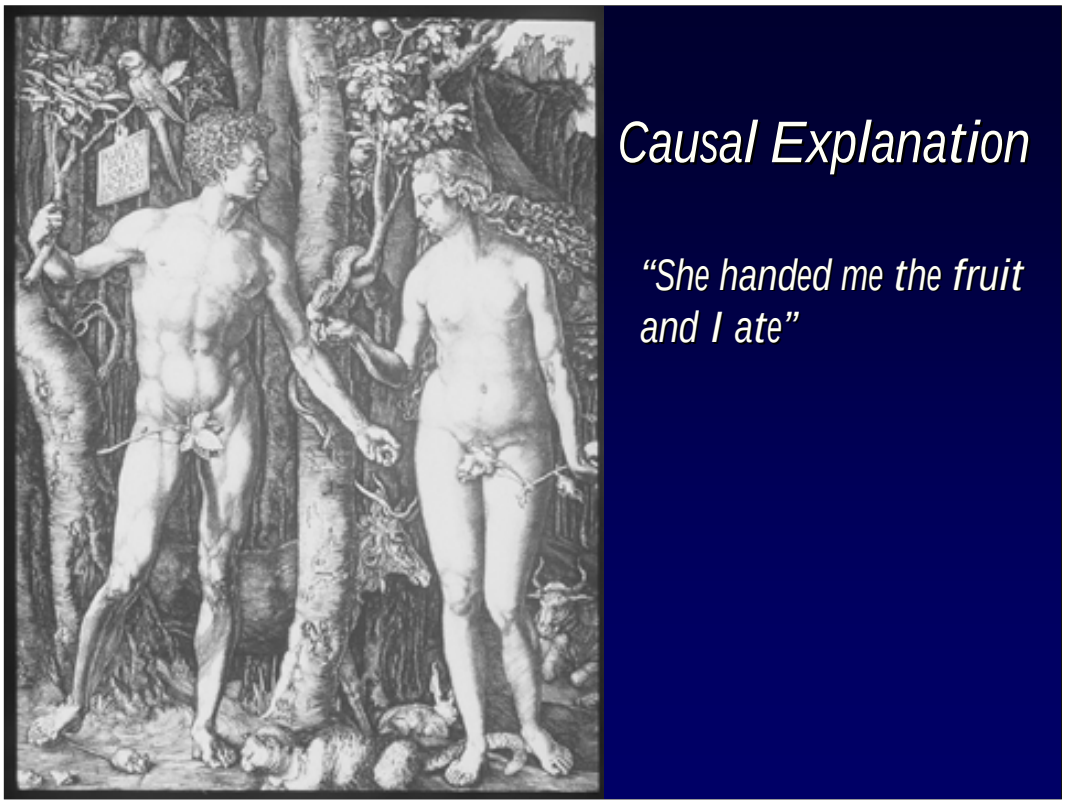
APPLICATIONS-II

4. Finding explanations for reported events
5. Generating verbal explanations
6. Understanding causal talk
7. Formulating theories of causal thinking

We now come to one of the grand problems in AI:

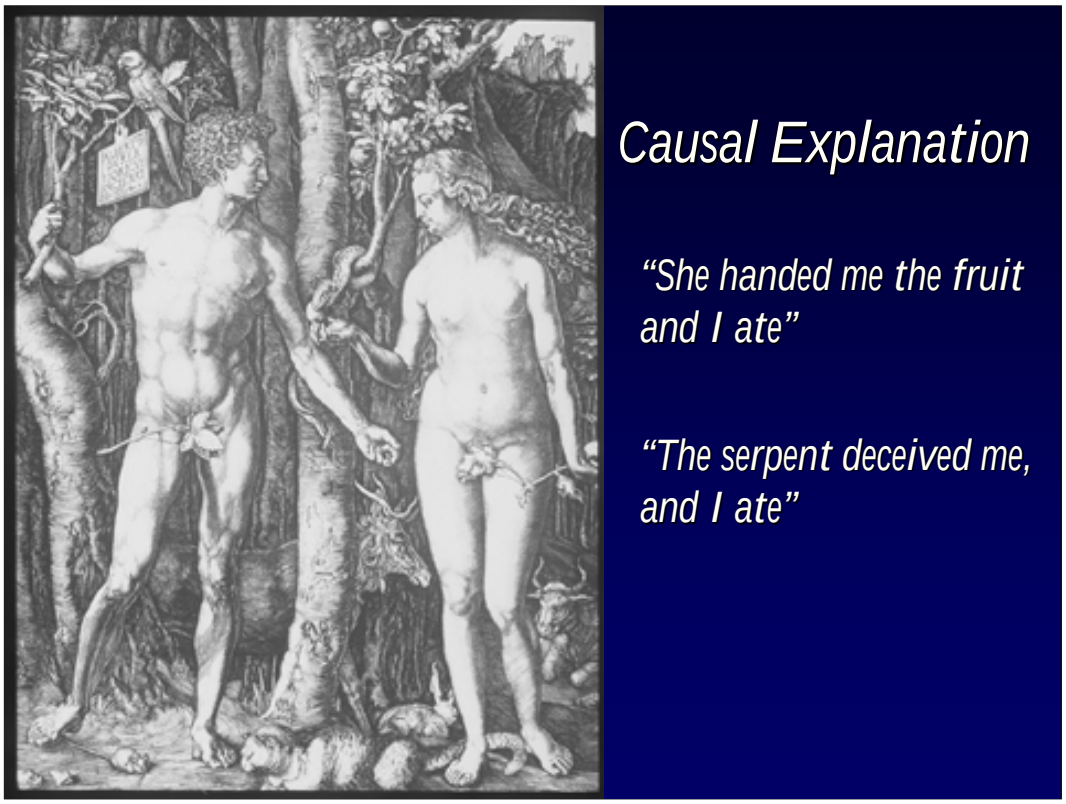
Generating meaningful explanations

It is the hardest of all causal tasks considered thus far, because the semantics of explanation is still debatable. I barely touched on this issue in the proceedings paper, but some promising solutions are currently in the making, and are described in greater detail in Chapter 10 of my forthcoming book.



The art of generating explanations is as old as mankind. According to the bible, it was Adam who first discovered the ubiquitous nature of causal explanation when he answered God's question with:

"She handed me the fruit and I ate"



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."

ACTUAL CAUSATION AND THE COUNTERFACTUAL TEST

"We may define a cause to be an object followed by another,..., where, if the first object had not been, the second never had existed."

Hume, Enquiry, 1748

Lewis (1973): "*x* CAUSED *y*" if *x* and *y* are true, and *y* is false in the closest non-*x*-world.

Structural interpretation:

(i) $X(u)=x$

(ii) $Y(u)=y$

(iii) $Y_{x'}(u) \neq y$ for $x' \neq x$

The modern formulation of this concept start again with David Hume. It was given a possible-world semantics by David Lewis, and even simpler semantics using our structural-interpretation of counterfactuals.

Notice how we write, in surgery language, the sentence:

"If the first object (*x*) had not been, the second (*y*) never had existed."

$$Y_{x'}(u) \neq y \text{ for } x' \neq x$$

Meaning: The solution for Y in a model mutilated by the operator $do(X=x')$ is not equal to *y*.

But this definition of "cause" is known to be ridden with problems.

PROBLEMS WITH THE COUNTERFACTUAL TEST

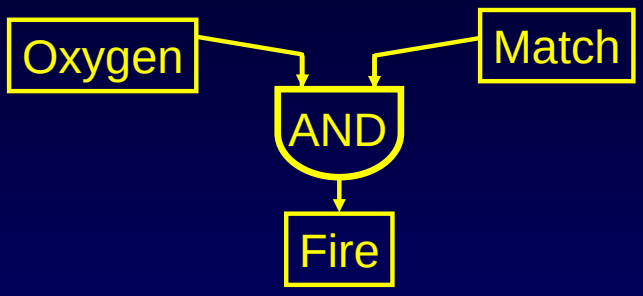
1. NECESSITY –
   Ignores aspects of sufficiency (Production)
   Fails in presence of other causes (Overdetermination)

2. COARSENESS –
   Ignores structure of intervening mechanisms.
   Fails when other causes are preempted (Preemption)

SOLUTION:
Supplement counterfactual test with Sustenance

I will first demonstrate these two problems by examples, and then provide a general solution using a notion called "Sustenance" which is easy to formulate in our structural-model semantics.

THE IMPORTANCE OF SUFFICIENCY (PRODUCTION)



Observation: Fire broke out.
Question: Why is oxygen an awkward explanation?
Answer: Because Oxygen is (usually) not sufficient

P(Oxygen is sufficient) = *P*(Match is lighted) = low
P(Match is sufficient) = *P*(Oxygen present) = high

Let us first look at the aspects of sufficiency (or production), namely, the capacity of a cause to produce the effect in situations where the effect is absence.

In our example, both Match and Oxygen are necessary, and none is sufficient alone. So, why is Match considered an adequate explanation and Oxygen an awkward explanation? The asymmetry surfaces when we compute the probability of sufficiency:

P(Oxygen is sufficient to produce Fire) = low

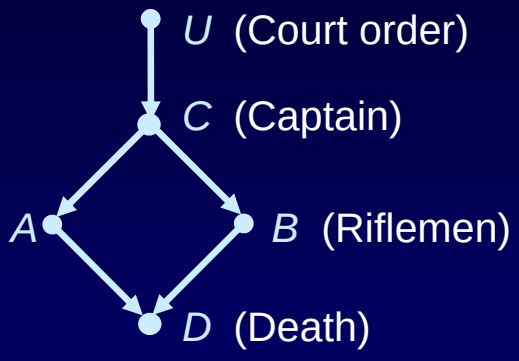
P(Match is sufficient to produce Fire) = high

Recall: $P(x \text{ is sufficient for } y) = P(Y_x = y | X \neq x, Y \neq y)$ which is well defined in our language.

Thus, we see that human judgment of explanation adequacy takes into account not merely how necessary a factor was for the effect but also how sufficient it was.

Another manifestation of sufficiency occurs in a phenomenon known as over-determination.

OVERDETERMINATION:
HOW THE COUNTERFACTUAL TEST FAILS?



Observation: Dead prisoner with two bullets.
Query: Was *A* a cause of death?
Answer: Yes, *A* sustains *D* against *B*.

Here, we consider each rifleman to be a cause of death. Why?

The prisoner would have died without *A.*

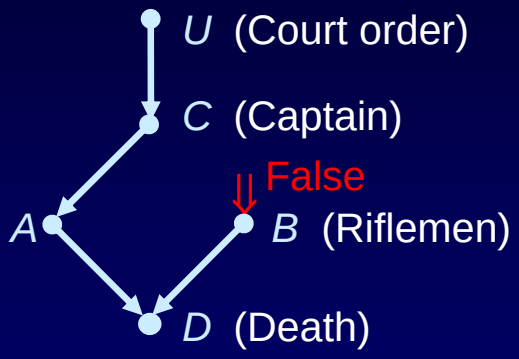
The answer lies in the concept of SUSTENANCE:

Death would still occur even if for some reason

B's rifle gets stuck, but only it *A* occurs.

Sustenance is a fundamental concept that helps dissolve many (if not all) of the problems associated with actual causation. So let us see what it entails.

OVERDETERMINATION:
HOW THE SUSTENANCE TEST SUCCEEDS?



Observation: Dead prisoner with two bullets.
Query: Was *A* a cause of death?
Answer: Yes, *A* sustains *D* against *B*.

Sustenance instructs us to imagine a new world, contrary to the scenario at hand, in which some structural contingencies are introduced. And in that contingency-inflicted world, we are to perform Hume's counterfactual test.

(Lewis talked about a related concept of "quasi dependence“ namely, counterfactual dependence if only "the surroundings were different". Sustenance offers a formal explication of this idea.)

Let us see formally how sustenance stands relative to necessity and sufficiency.

NUANCES IN CAUSAL TALK

y depends on *x* (in *u*)
$X(u)=x,\ Y(u)=y,\ Y_{x'}(u)=y'$

x can produce *y* (in *u*)
$X(u)=x',\ Y(u)=y',\ Y_{x}(u)=y$

x sustains *y* relative to *W=w'*
$X(u)=x,\ Y(u)=y,\ Y_{xw'}(u)=y,\ Y_{x'w'}(u)=y'$

Here we see the formal definitions of “dependence” (or necessity), “production” (or sufficiency) and “sustenance”.

The last condition $Y_{x'w'}(u)=y'$ weakens necessity by requiring that *Y* differ from *y* (under $x' \neq x$) only under one special condition, when W is set to some *w'*.

But the third condition, $Y_{xw}(u)=y'$ substantially strengthens sufficiency, insisting that *Y* retain its value *y* (under *x*) for every setting of *W = w*.

NUANCES IN CAUSAL TALK

y depends on *x* (in *u*)
$X(u)=x, Y(u)=y, Y_{x'}(u)=y'$

x can produce *y* (in *u*)
$X(u)=x', Y(u)=y', Y_{x}(u)=y$

x *x* sustains *y* relative to *W*=*w*′
$X(u)=x, Y(u)=y, Y_{xw'}(u)=y, Y_{x'w'}(u)=y'$

x caused *y*,
necessary for,
responsible for,
y due to *x*,
y attributed to *x*.

These three aspects of causation have several manifestations in causal talk.

The expressions on the right are associated with dependence (or necessity)

NUANCES IN CAUSAL TALK

y depends on *x* (in *u*)

$X(u)=x, Y(u)=y, Y_{x'}(u)=y'$

x can produce *y* (in *u*)

$X(u)=x', Y(u)=y', Y_{x}(u)=y$

x sustains *y* relative to *W*=*w*′

$X(u)=x, Y(u)=y, Y_{xw'}(u)=y, Y_{x'w'}(u)=y'$

x causes *y*, sufficient for, enables, triggers, brings about, activates, responds to, susceptible to.

Notice that when production is invoked, the present tense is used: "*x* causes *y*", instead of "*x* caused *y*"

NUANCES IN CAUSAL TALK

y depends on *x* (in *u*)
$X(u)=x, Y(u)=y, Y_{x'}(u)=y'$

x can produce *y* (in *u*)
$X(u)=x', Y(u)=y', Y_{x}(u)=y$

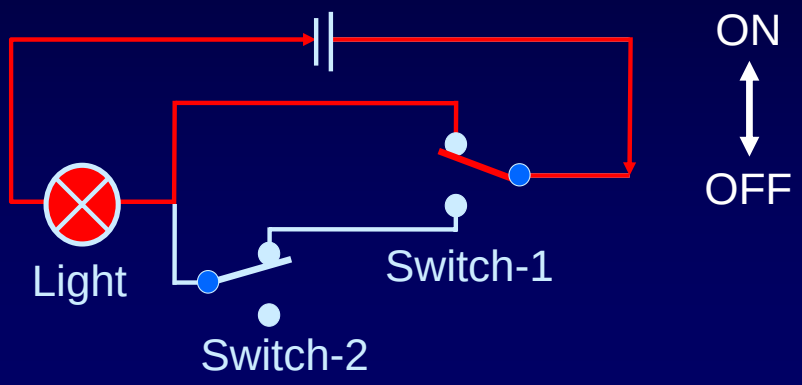
x sustains *y* relative to *W*=*w*′
$X(u)=x, Y(u)=y, Y_{xw'}(u)=y, Y_{x'w'}(u)=y'$

maintain, protect, uphold, keep up, back up, prolong, support, rests on.

Finally, here are some expressions connected with the notion of sustenance.

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.

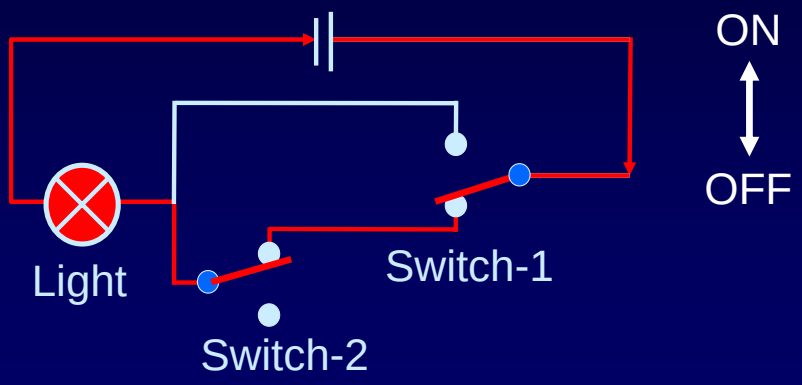
If someone were to ask us what caused the light to be on, we would point to Switch-1. After all, S1 causes the current to flow through this wire, while S2 is totally out of the game.

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

Light = S1 $\vee$ S2

PREEMPTION: HOW THE COUNTERFACTUAL TEST FAILS

Which switch is the actual cause of light? S_1!

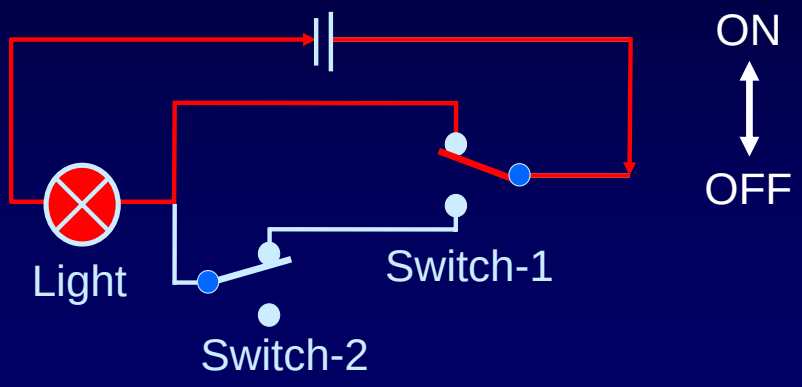


Deceiving symmetry: *Light* = $S_1 \vee S_2$

Turning Switch-1 off merely re-directs the current, but keeps the light on.

PREEMPTION: HOW THE COUNTERFACTUAL TEST FAILS

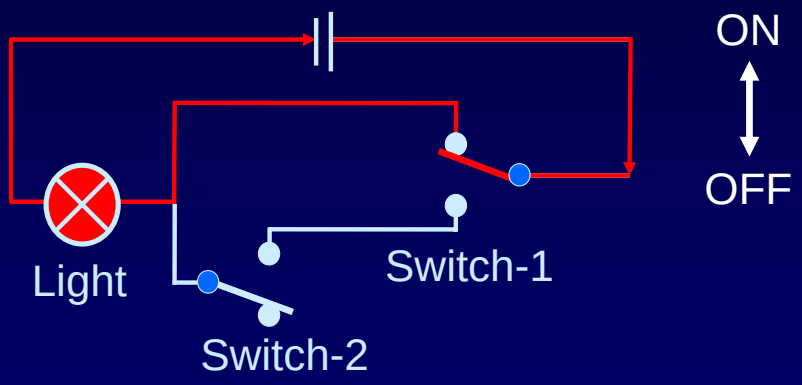
Which switch is the actual cause of light? S_1!



Deceiving symmetry: *Light* = $S_1 \vee S_2$

PREEMPTION: HOW THE COUNTERFACTUAL TEST FAILS

Which switch is the actual cause of light? S_1!

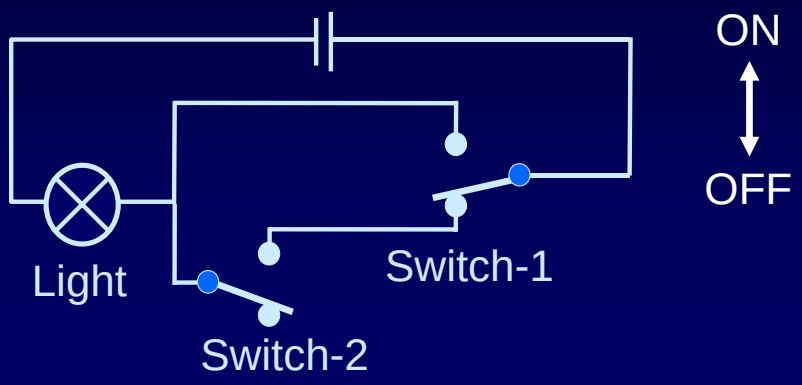


Deceiving symmetry: *Light* = $S_1 \vee S_2$

Turning Switch-2 off has no effect whatsoever.

PREEMPTION: HOW THE COUNTERFACTUAL TEST FAILS

Which switch is the actual cause of light? S_1!



Deceiving symmetry: *Light* = $S_1 \vee S_2$

The light turns off if and only if both switches are off.

This example is interesting because it is for the first time that we witness the effect of structure on our perception of actual causation.

Evidently, our mind takes into consideration, not merely input-output relationships, but also the inner structure of the process leading from causes to effects. How?

CAUSAL BEAM
Locally sustaining sub-process

ACTUAL CAUSATION

"*x* is an actual cause of *y*" in scenario *u*,
if *x* passes the following test:

1. Construct a new model *Beam*(*u*, *w′*)
   - 1.1 In each family, retain a subset of parents that minimally sustains the child
   - 1.2 Set the other parents to some value *w′*
2. Test if *x* is necessary for *y* in *Beam*(*u*, *w′*) for some *w′*

The solution I would like to propose here (this is not in your proceedings but is explained in my book) is based on local sustenance relationships.

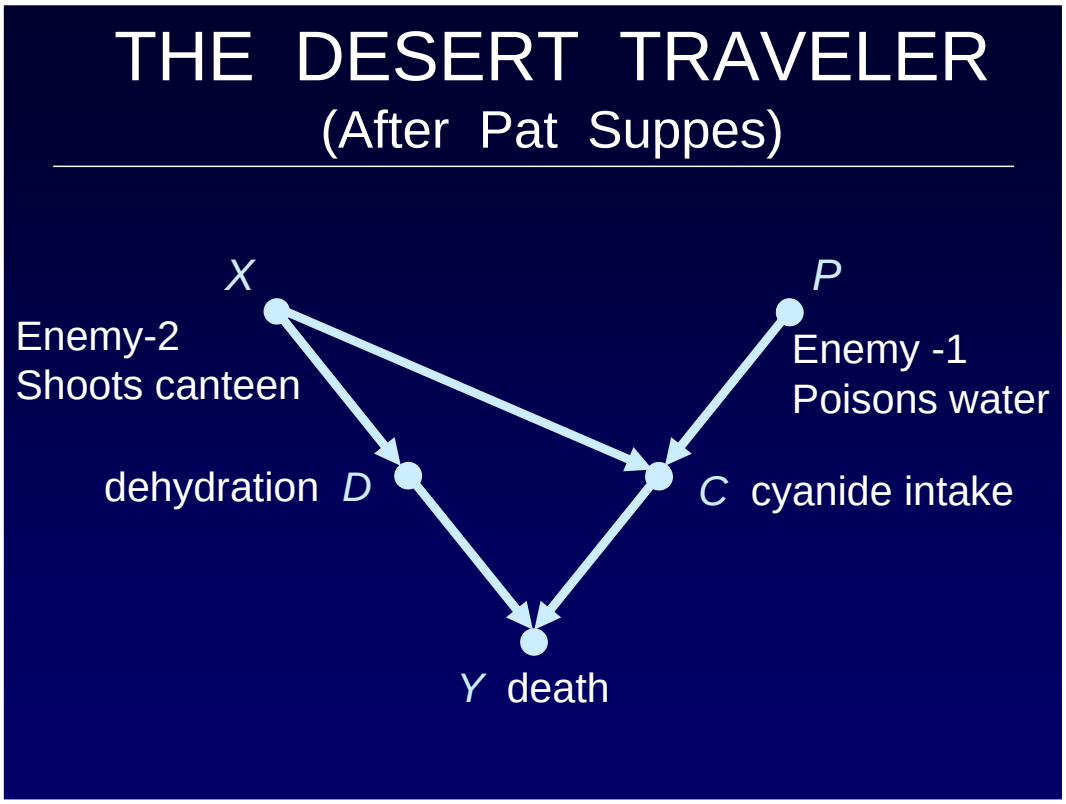
Given a causal model, and a specific scenario in this model, we construct a new model by pruning away, from every family, all parents except those that minimally sustain the value of the child.

I call such a model a causal beam.

In this new model we conduct the counterfactual test, and we proclaim an event $X=x$ the actual cause of $Y=y$ if y depends on x.

I will next demonstrate this construction using a classical example due to P. Suppes.

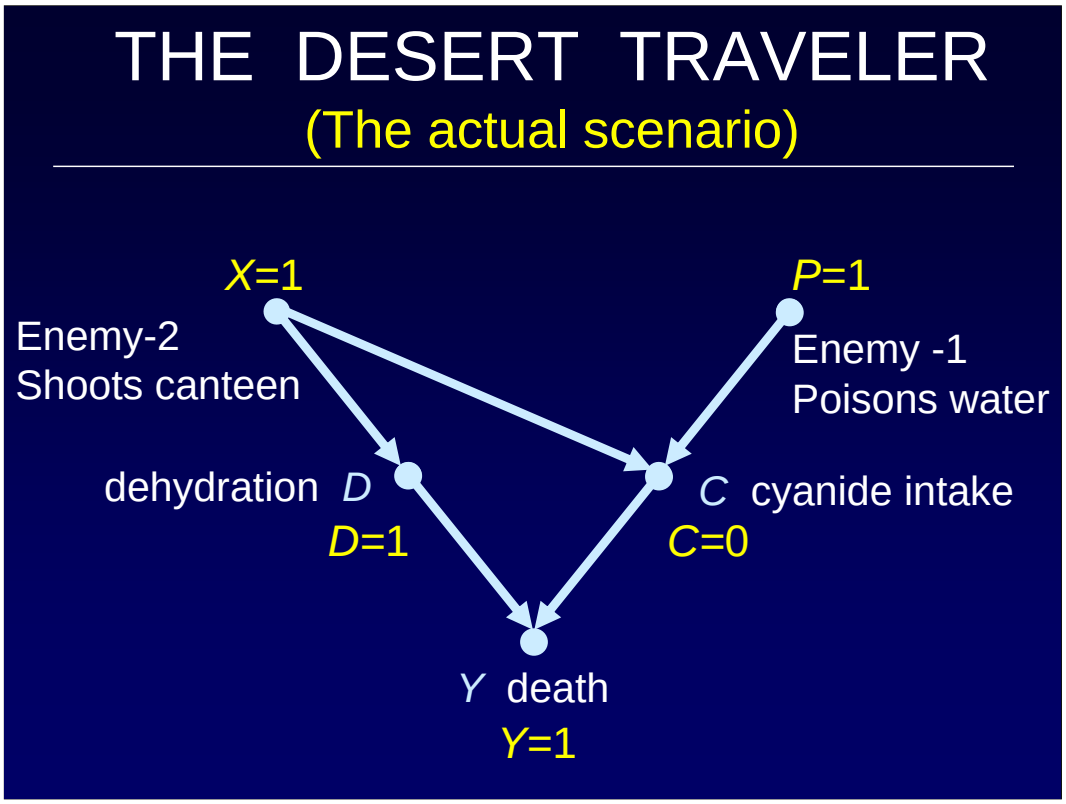
It is isomorphic to the two-switch problem, but more blood-thirsty.



A desert traveler *T* has two enemies. Enemy-1 poisons *T*'s canteen, and Enemy-2, unaware of Enemy-1's action, shoots and empties the canteen. A week later, *T* is found dead and the two enemies confess to action and intention. A jury must decide whose action was the *actual cause* of *T*'s death.

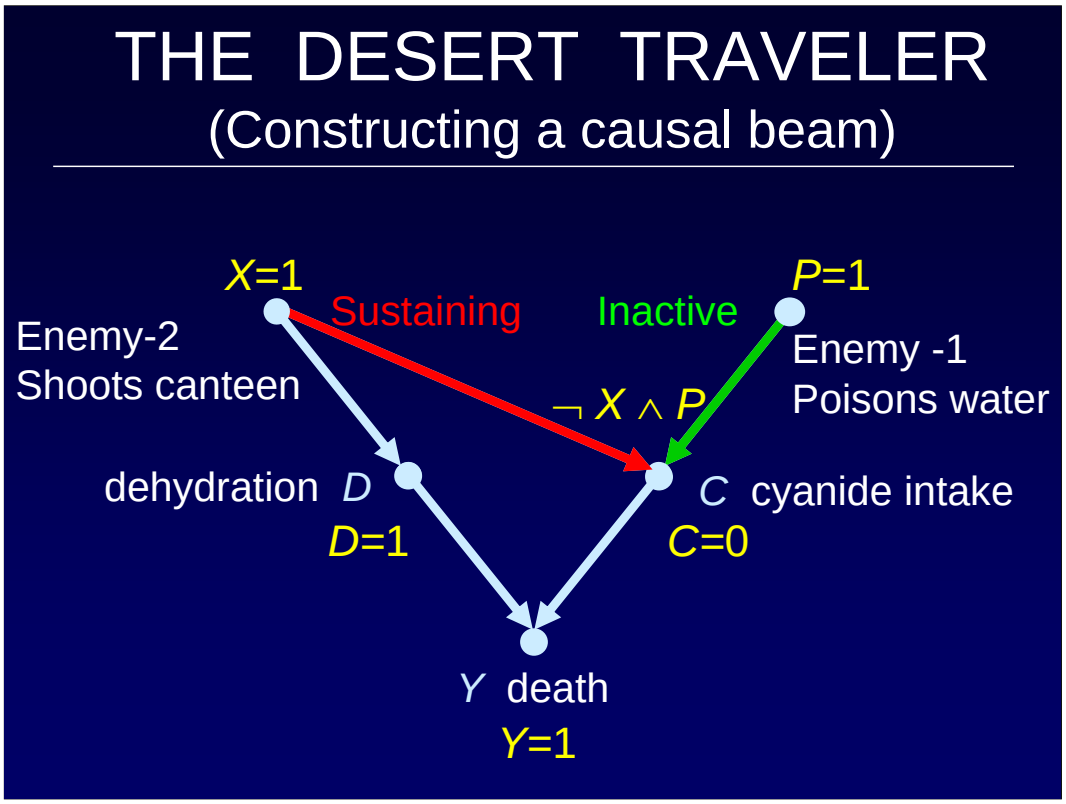
Enemy-1 claims: *T* died of thirst

Enemy-2 claims: I have only prolonged *T*'s life.

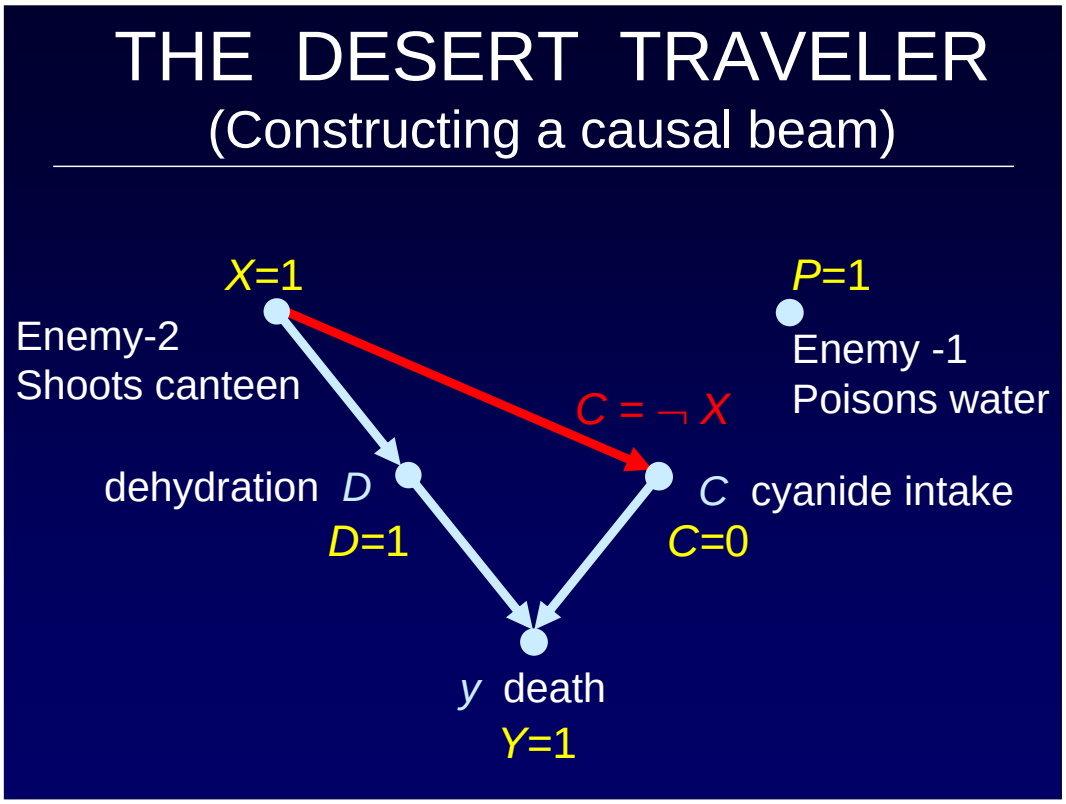


Now let us construct the causal beam associated with the natural scenario, in which we have:

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

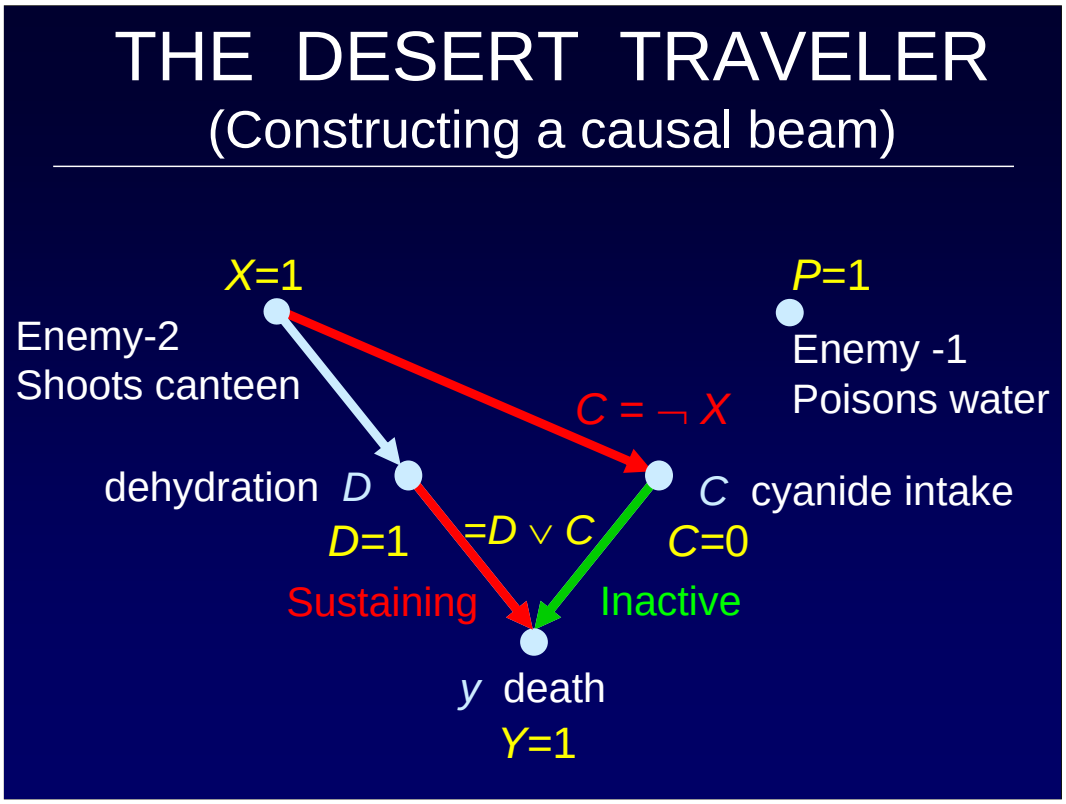


Consider the Cyanide family. Since emptying the canteen is sufficient for sustaining no Cyanide intake, regardless of poisoning, we label the link $P \rightarrow C$ "inactive", and the link $X \rightarrow C$ "sustaining".



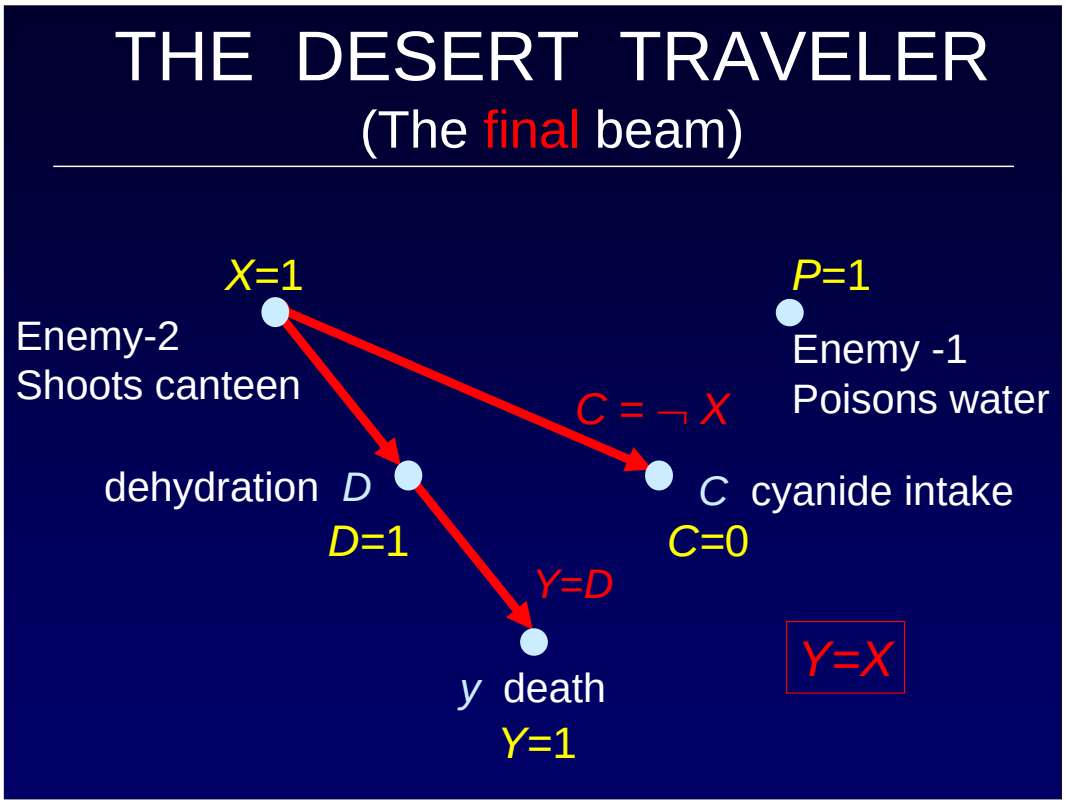
The link $P \rightarrow C$ is inactive in the current scenario, which allows us to retain just one parent of C, with the functional relationship $C = \neg X$.

We repeat this process on other parent-child families.



Next consider the *Y*-family (in the situation *D=1, C=0)*.

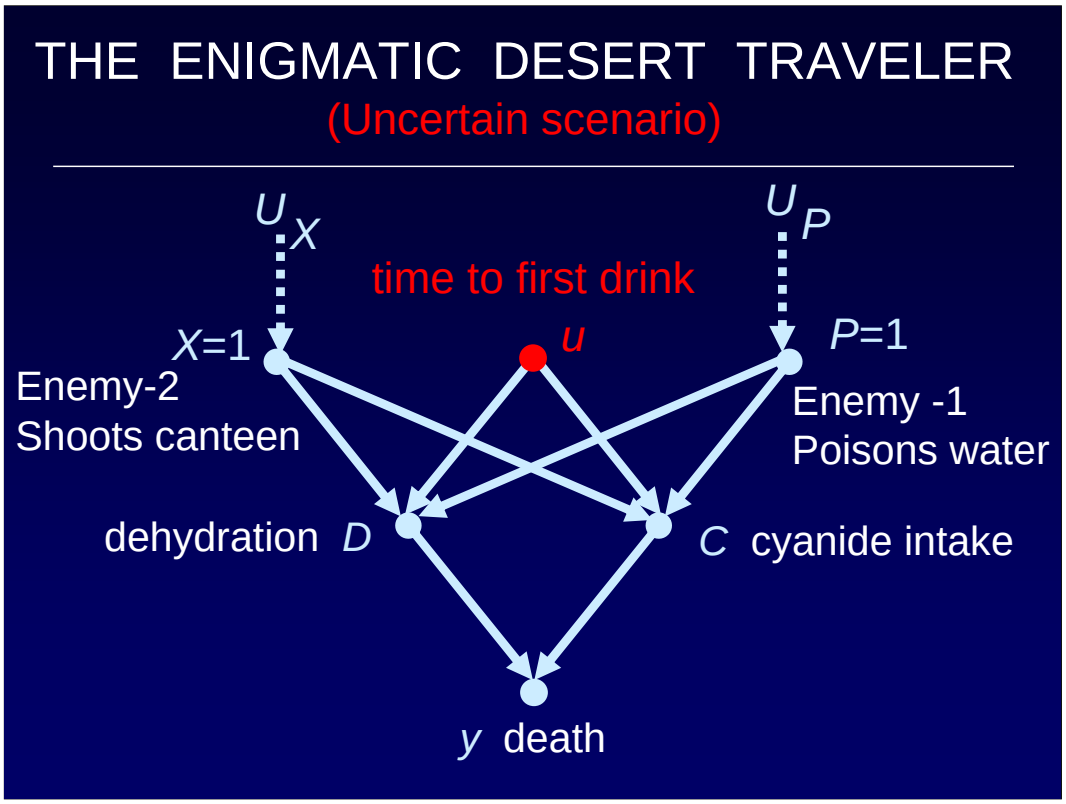
Since dehydration would sustain death regardless of cyanide intake, we label the link $C \rightarrow Y$ "inactive" and the link $D \rightarrow Y$ "sustaining".



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

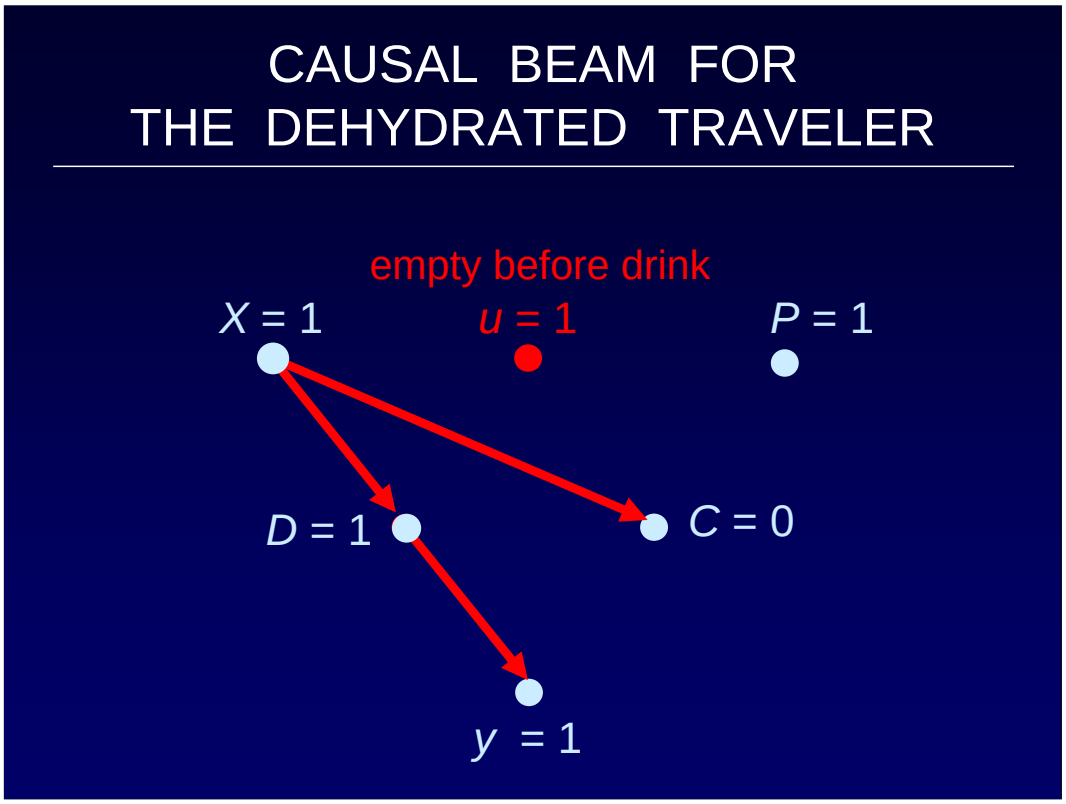
In this final model we conduct the counterfactual test and find that the test is satisfied since $Y = X$.

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.

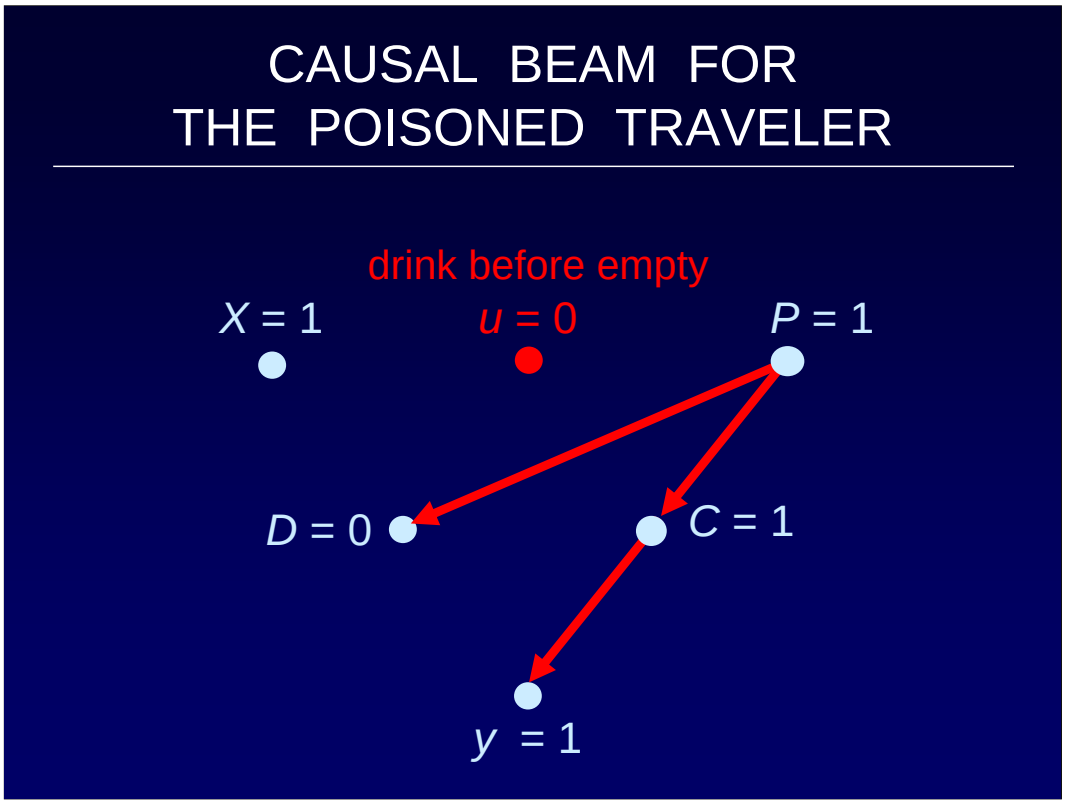


Things will change of course, if the we do not know whether the traveler craved for water before the shot.

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



If the canteen was emptied before T drank, we have the dehydration scenario, as before.



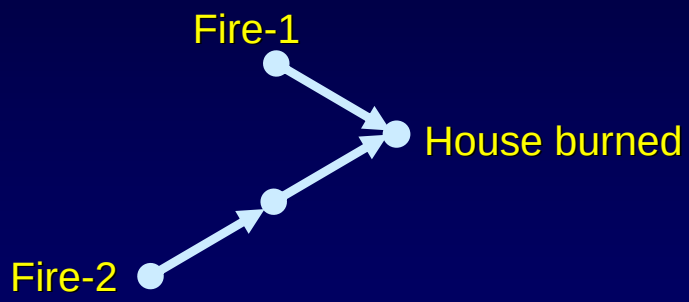
On the other hand, if *T* drank before the canteen was emptied we have a new causal beam, in which Enemy-1 is classified as the cause of death.

If *U* is uncertain we can use *P*(*u*) to compute the probability *P*(*x* caused *y*), because the sentence "*x* was the actual cause of *y*" receives definite truth-value in every *u*.

Thus, $P(x \text{ caused } y) = \sum_{u \mid x \text{ caused } y \text{ in } u} P(u)$

TEMPORAL PREEMPTION

Fire-1 is the actual cause of damage



Yet, Fire-1 fails the counterfactual test

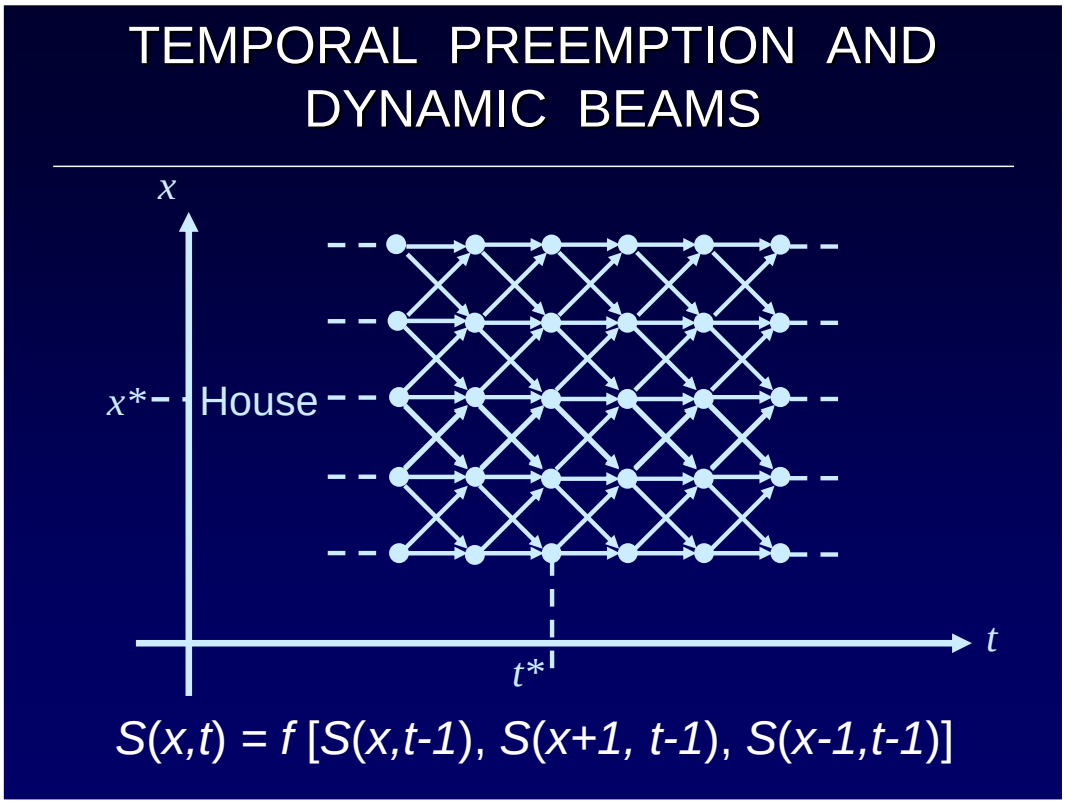
We come now to resolve the third objection against the counterfactual test -- temporal preemption.

Consider two fires advancing toward a house. If Fire-1 burned the house before Fire-2 we (and many juries nationwide) would consider Fire-1 "the actual cause" for the damage, though Fire-2 would have done the same if it were not for Fire-1. If we simply write the structural model as

$$H = F_1 \vee F_2,$$

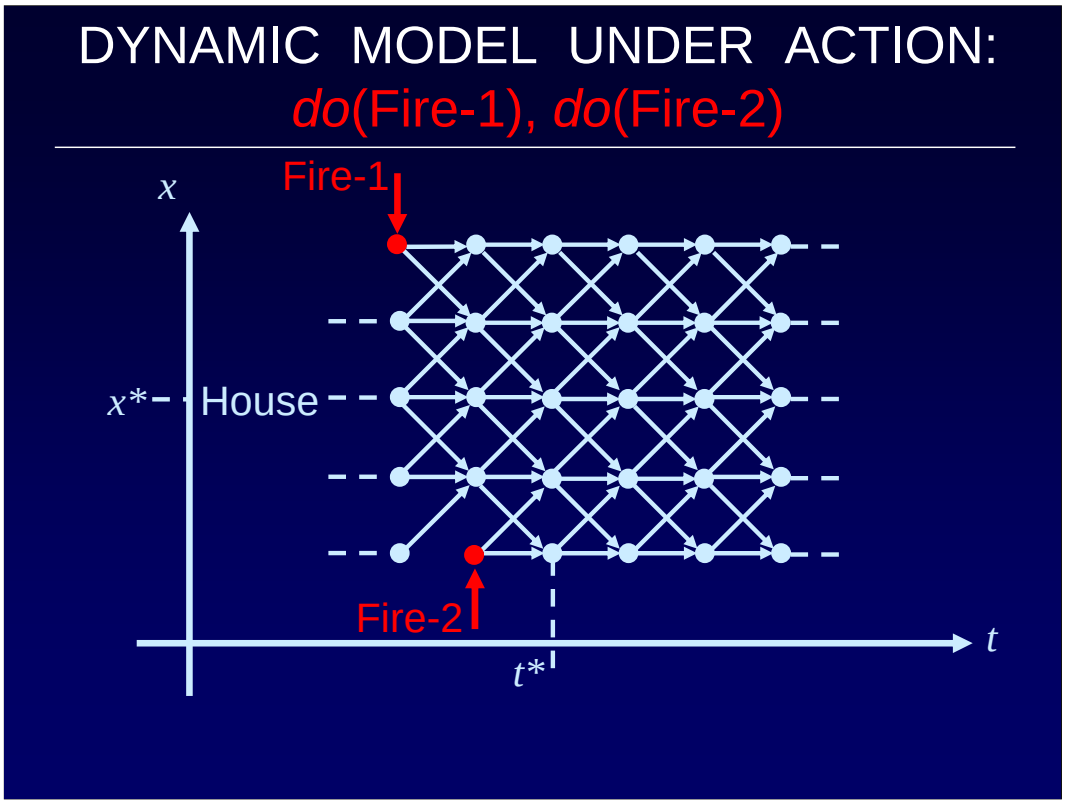
where *H* stands for "house burns down," the beam method would classify each fire equally as a contributory cause, which is counterintuitive.

Here the second cause becomes ineffective only because the effect "has already happened" -- a temporal notion that cannot be expressed in the static causal model we have used thus far. Remarkably, the idea of a causal bean still gives us the correct result if we use a dynamic model of the story.

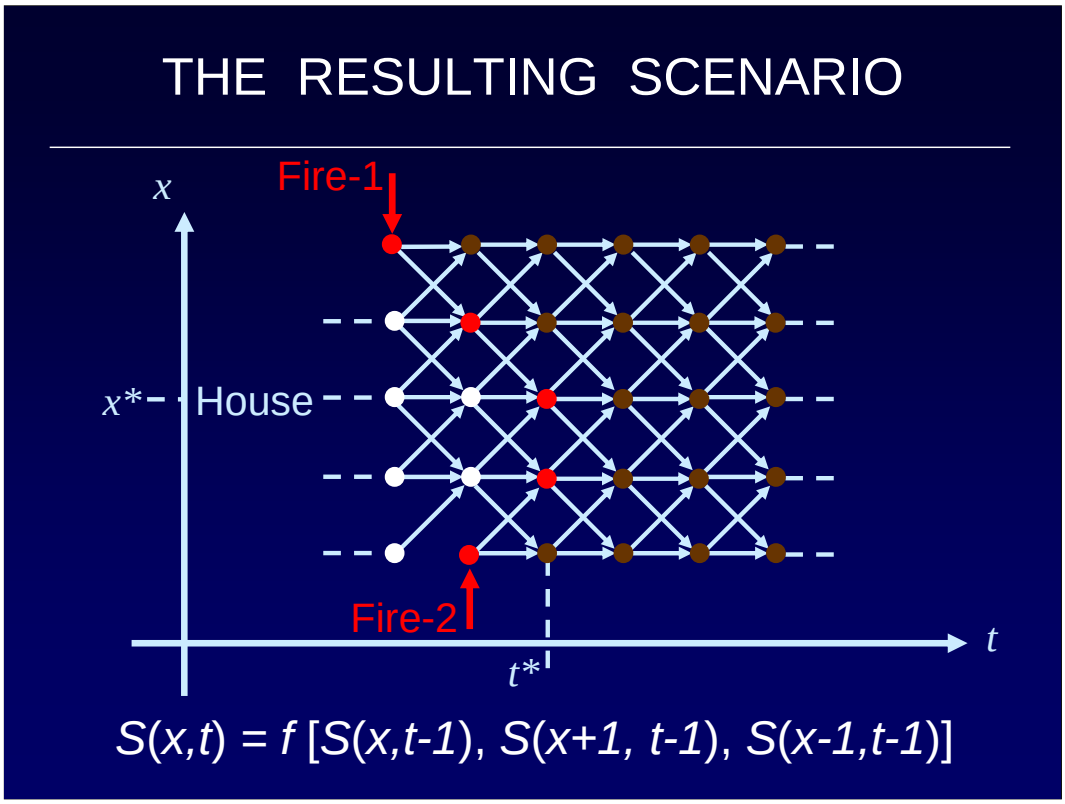


Dynamic structural equation are obtained when we index variables by time and ask for the mechanisms that determine their values.

For example, we may designate by $S(x,t)$ the state of the field in location x and time t, and describe each variable $S(x,t)$ as dependent on three other variables: the state of the adjacent region to the north, the state of the adjacent region to the south and the previous state at the same location.



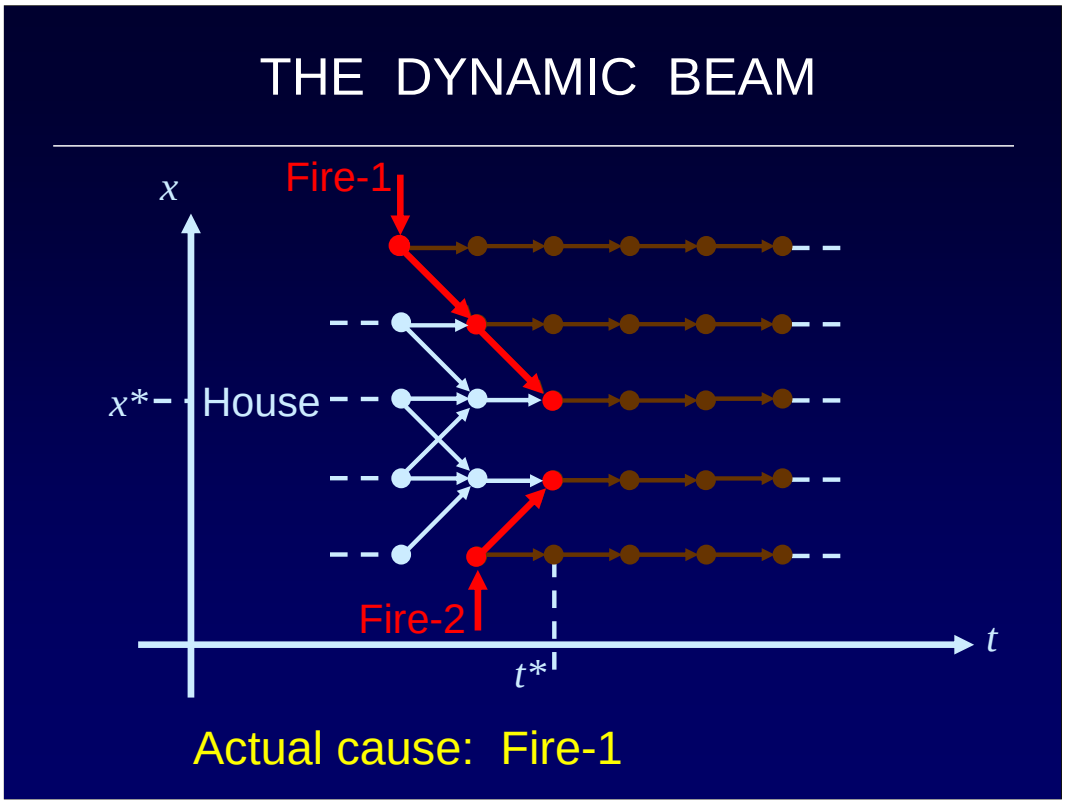
To test which action was the cause of the damage, we first simulate the two actions at their corresponding times and locations, as shown in the slide.



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.



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.

CONCLUSIONS

"I would rather discover one causal relation than be King of Persia"

Democritus (430-380 BC)

Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance).

A. Einstein, April 23, 1953

I would like now to conclude this lecture by quoting two great scientists.

The first is Democritus, the father of the atomic theory of matter, who said: I would rather discover ONE causal relation than be King of Persia.

Admittedly, the political situation in Persia has changed somewhat from the time he made this statement, but I believe Democritus has a valid point in reminding us of the many application areas that could benefit from the discovery of even ONE causal relation, namely from the solution of ONE toy problem, on AI scale.

I have discussed these applications earlier, which include medicine, biology, economics, and social science, and I believe AI is in a unique position to help those areas, because only AI enjoys the combined strength of model-searching, learning and the logic of causation.

CONCLUSIONS

"I would rather discover one causal relation than be King of Persia"

Democritus (430-380 BC)

Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance).

A. Einstein, April 23, 1953

The second quote is from Albert Einstein who, a year before his death, attributes the progress of Western Science to two fortunate events: The invention of formal logic by the Greek geometers, and the Galilean idea that causes could be discovered by experiments.

As I have demonstrated earlier, experimental science have not fully benefited from the power of formal methods --- formal mathematics was used primarily for analyzing passive observations, under fixed boundary conditions, while the design of new experiments, and the transitions between boundary conditions, have been managed entirely by the unaided human intellect.

The development of autonomous agents and intelligent robots requires a new type of analysis, in which the DOING component of science enjoys the benefit of formal mathematics, side by side with its observational component, a tiny glimpse of such analysis I have labored to uncover in this lecture. I am convinced that the meeting of these two components will eventually bring about another scientific revolution, perhaps equal in impact and profoundness to the one that took place during the renaissance. AI will be the major player in this revolution, and I hope each of you take part in seeing it off the ground.

ACKNOWLEDGEMENT-I

Collaborators in Causality:

Alex Balke
David Chickering
Adnan Darwiche
Rina Dechter
Hector Geffner
David Galles
Moisés Goldszmidt
Sander Greenland
David Heckerman
Jin Kim
Jamie Robins
Tom Verma

My next to final slide lists the cast of this show,--- a wonderful team of colleagues and students with whom I was fortunate to collaborate. Most of these names should be familiar to you from other stages and other shows, except perhaps Greenland and Robins, two epidemiologists who are currently carrying the banner of causal analysis in epidemiology.

ACKNOWLEDGEMENT-II

Influential ideas:

S. Wright (1920)
T. Haavelmo (1943)
H. Simon (1953)
I.J. Good (1961)
R. Strotz & H. Wold (1963)
D. Lewis (1973)
R. Reiter (1987)
Y. Shoham (1988)
M. Druzdzel & H. Simon (1993)
P. Spirtes, C. Glymour & R. Scheines (1993)
P. Nayak (1994)
F. Lin (1995)
D. Heckerman & R. Shachter (1995)
N. Hall (1998)
J. Halpern (1998)
D. Michie (1998)

I have borrowed many of these ideas from other authors, the most influential ones are listed here, while others are cited in the proceedings paper. I will only mention that the fundamental idea that actions be conceived as modifiers of mechanisms, goes back to Jacob Marschak and Herbert Simon. Strotz and Wold were the first to represent actions by "wiping out" equations, and I would never have taken seriously the writings of these "early" economists if it were not for Peter Spirtes' lecture, 100 Miles from here, Uppsala, 1991, where I first learned about manipulations and manipulated graphs.

Thanks you all.