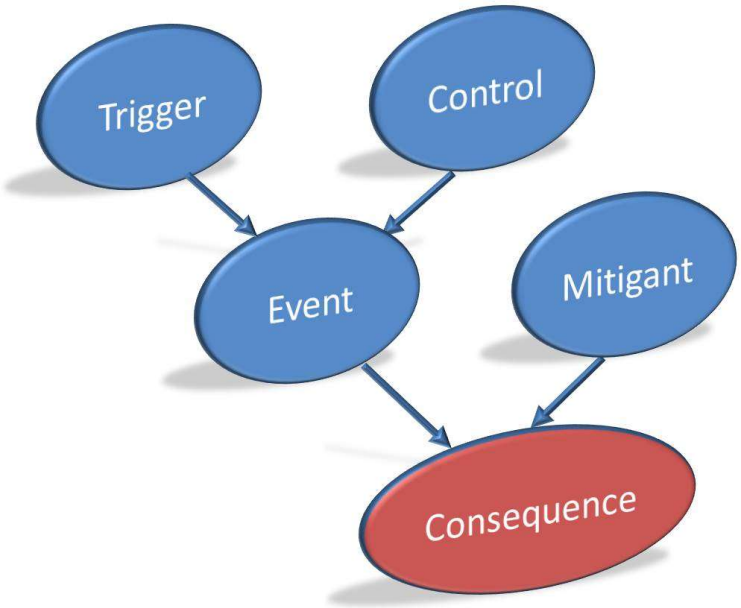


Risk and Decision Making for Data Science and AI

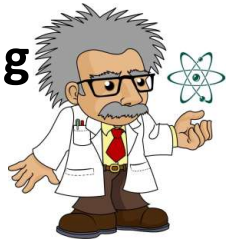
Lesson 7 Interventions and counterfactuals

Norman Fenton

@ProfNFenton

Pearl's ladder of causation

Imagining



Doing

Seeing

<u>Counterfactuals</u>: "What if I had done…"
If I hadn't taken this drug would my headache still have stopped?

<u>Intervention</u>: "What if I do…"
If I take this drug will it stop my headache?

<u>Association</u>: "What if I see…"
From trials data is this drug effective at stopping headaches?

'Standard' statistical methods and machine learning can ONLY really answer questions of association

In this module you will learn how to get to levels 2 and 3 using causal BNs.

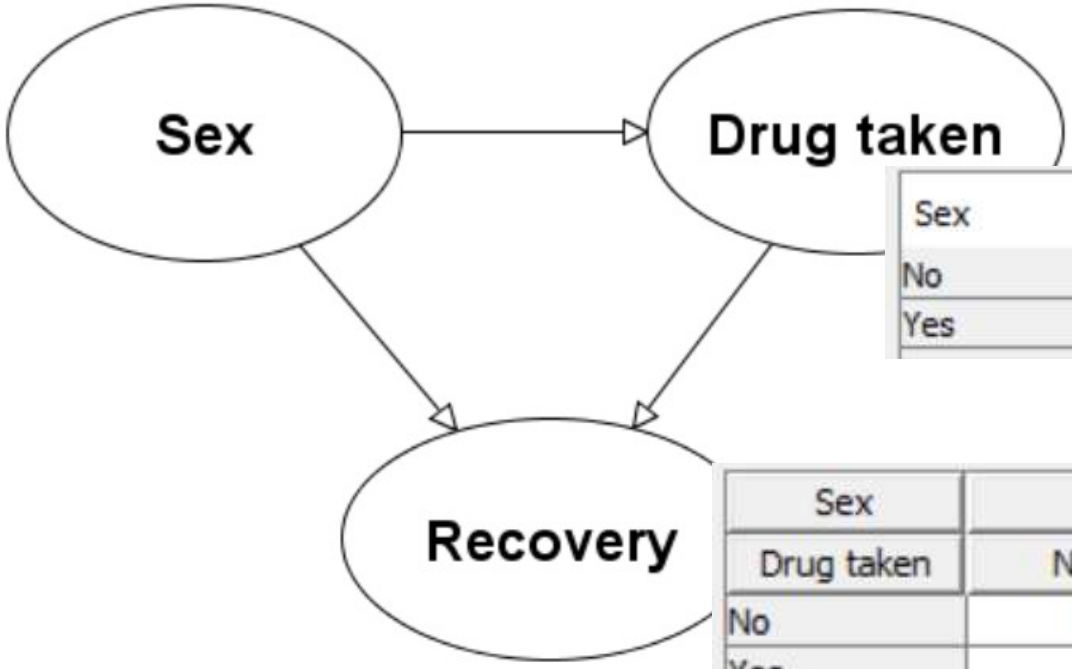
Drug example

Sex	**Female**		**Male**	
Drug taken	No	Yes	No	Yes
Recovered				
No	210	80	30	120
Yes	90	20	70	180
Recovery rate	***30%***	***20%***	***70%***	***60%***

Female	400.0
Male	400.0

210+90+80+20

30+70+120+180



210+90

30+70

Sex	Female	Male
No	300.0	100.0
Yes	100.0	300.0

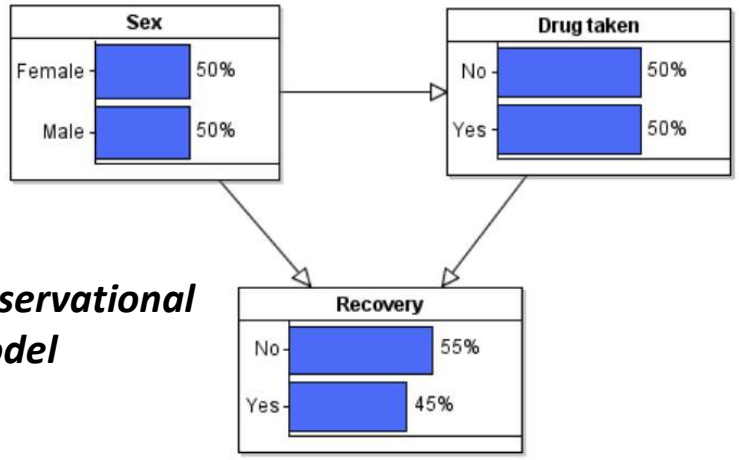
80+20

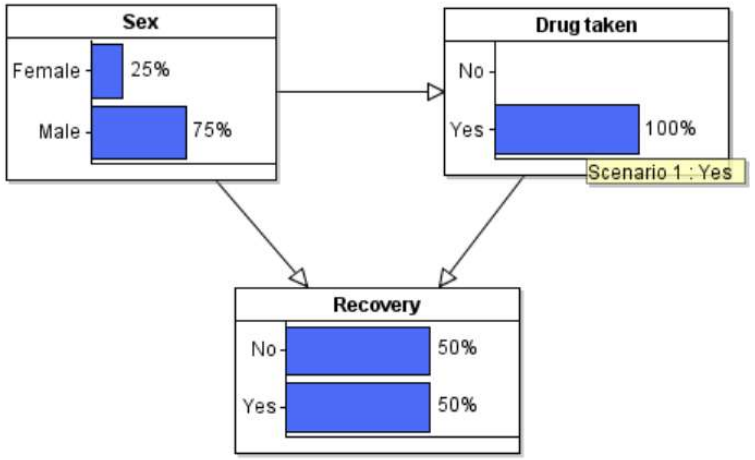
120+180

Sex	Female		Male	
Drug taken	No	Yes	No	Yes
No	210.0	80.0	30.0	120.0
Yes	90.0	20.0	70.0	180.0

“True” intervention effect is ‘wrong’ in the observational model

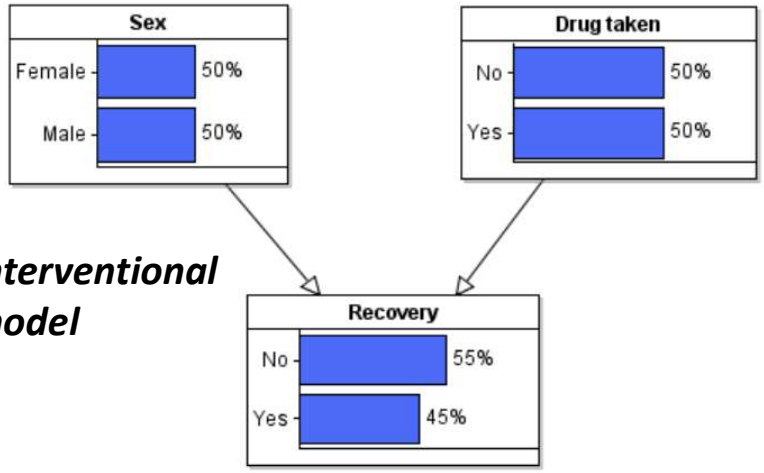
Observational model

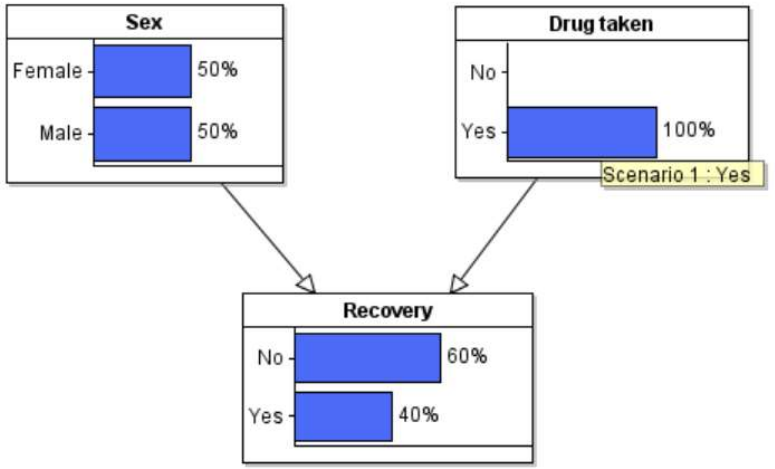




Sex influences whether Drug is taken. If we observe drug taken the probability the subject is a Male increases

Interventional model

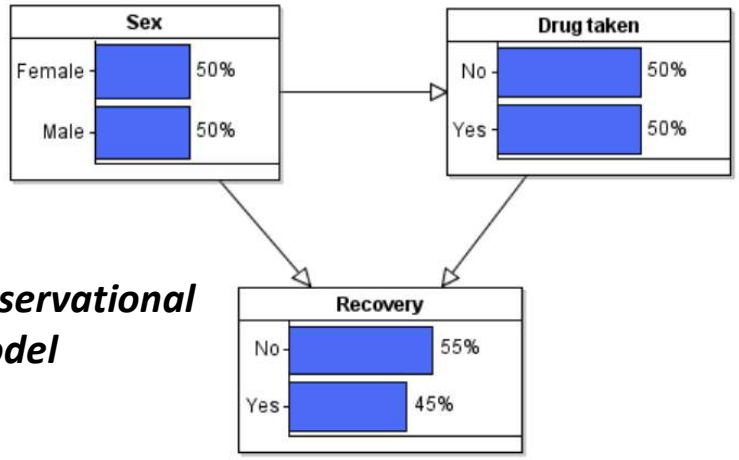


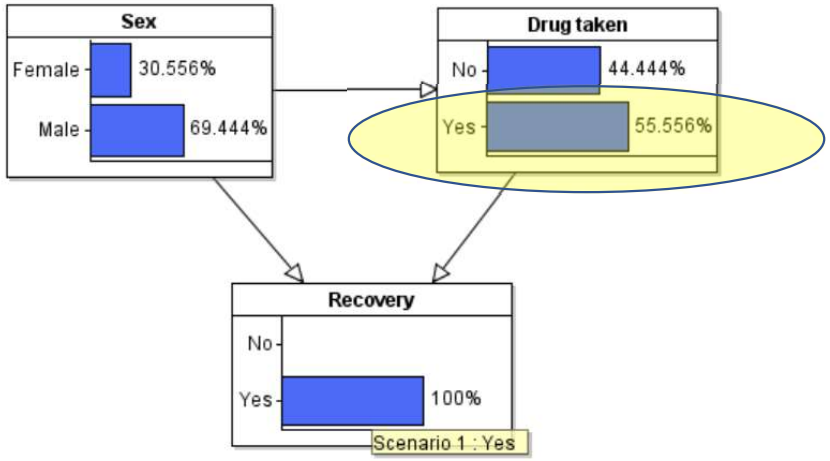


By removing links into the ‘intervention’ variable, if we observe drug taken then this is independent of Sex

Backward inference reveals the fundamental difference

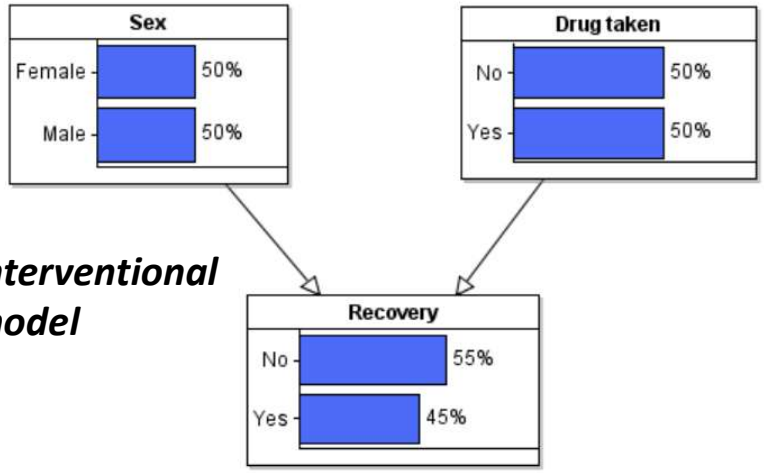
Observational model

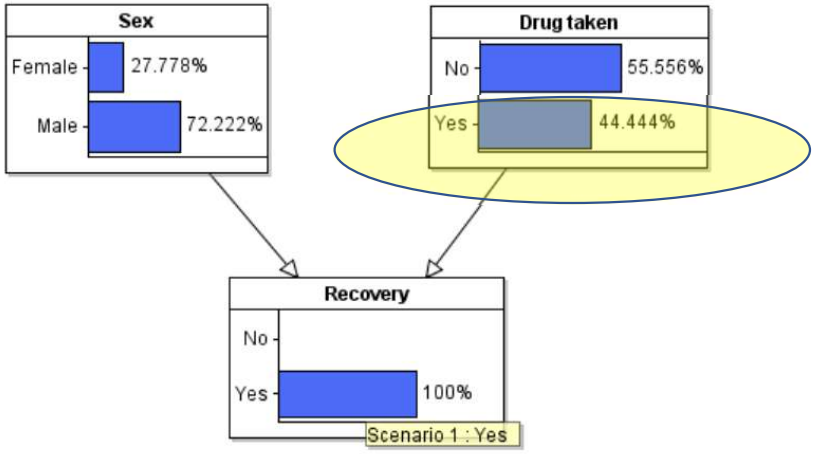




If we know patient recovered then probability drug taken ***increases***

Interventional model





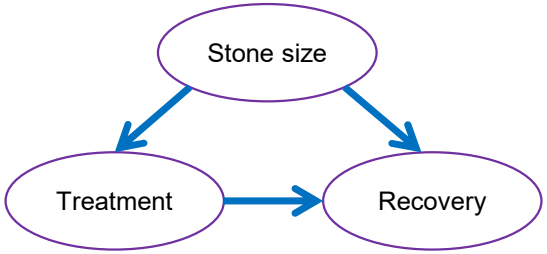
If we know patient recovered then probability drug taken ***decreases***

A real medical study

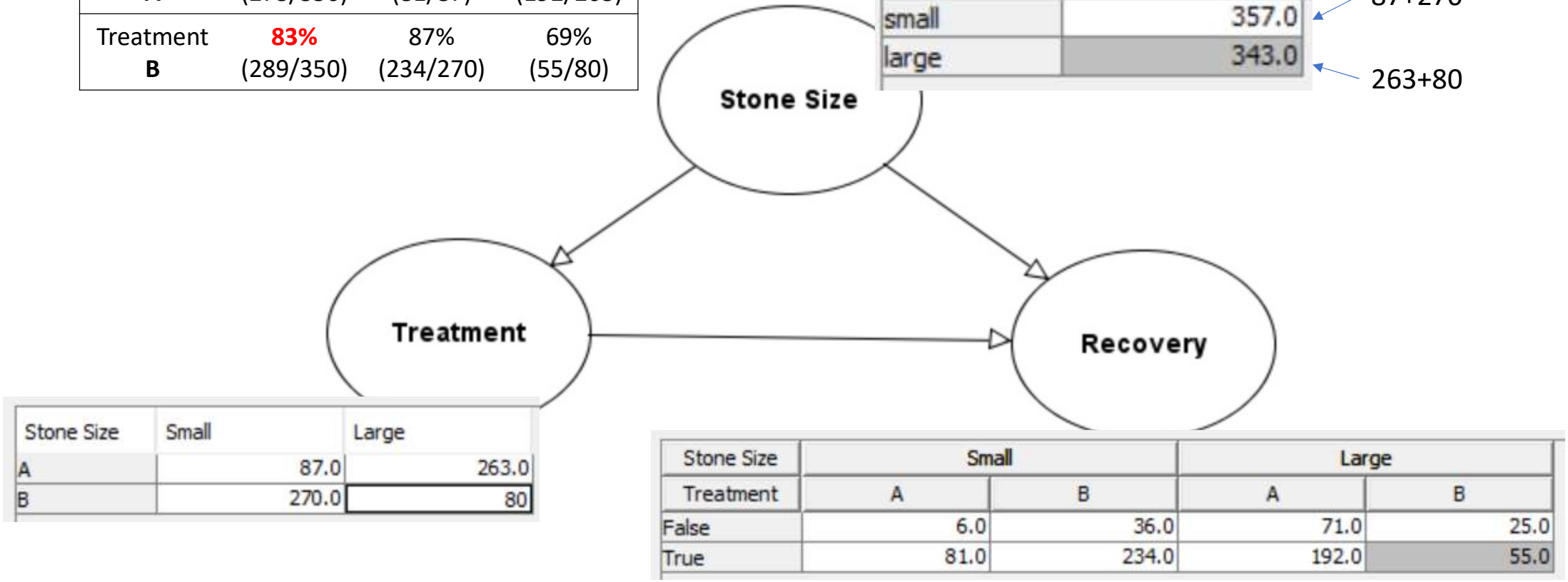
	Overall	Patients with small stones	Patients with large stones
Treatment **A**	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment **B**	**83%** (289/350)	87% (234/270)	69% (55/80)

The outcome of two treatments (success rate) for patients with kidney stones were compared based on 700 patients, half of whom had treatment A and half of whom had treatment B

Treatment B is **better overall** (recovery rate: 83%), but **worse in every subcategory** (recovery rate 87% and 69% compared to 93% and 73% respectively) when we split data by the size of kidney stones



	Overall	Patients with small stones	Patients with large stones
Treatment **A**	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment **B**	**83%** (289/350)	87% (234/270)	69% (55/80)



small	357.0
large	343.0

87+270

263+80

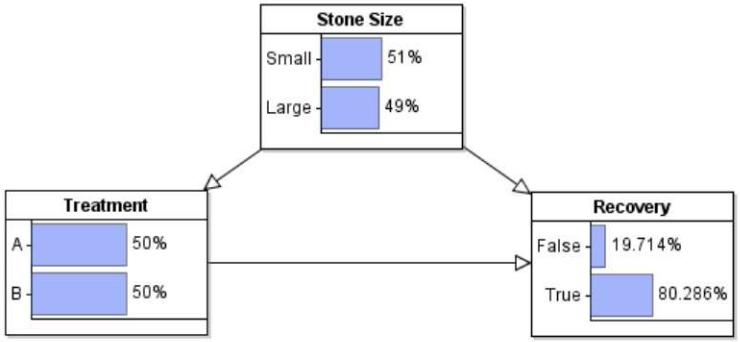
Stone Size	Small	Large
A	87.0	263.0
B	270.0	80

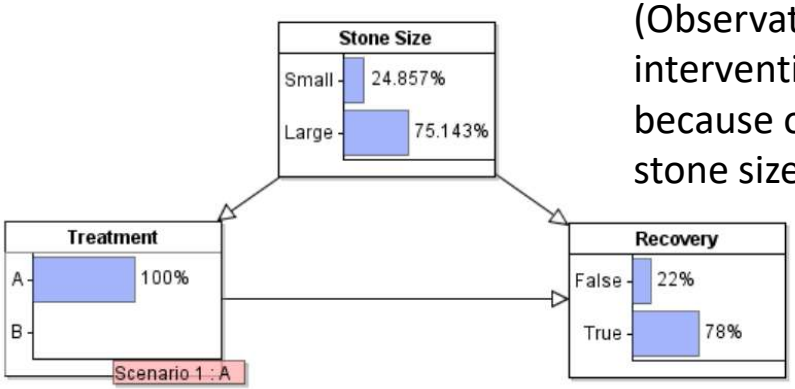
Stone Size	Small		Large	
Treatment	A	B	A	B
False	6.0	36.0	71.0	25.0
True	81.0	234.0	192.0	55.0

If I do not know patient stone size which treatment should we give?

Observational reasoning with causal model (rung 1)

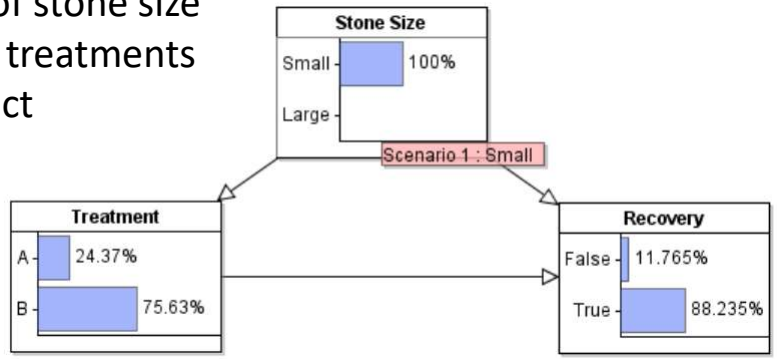
Marginals



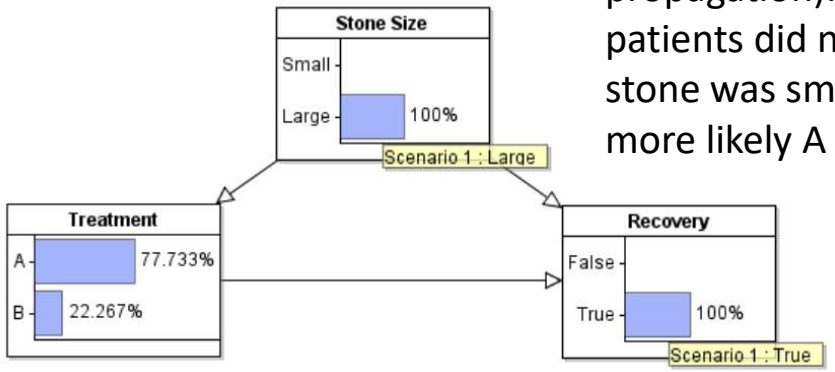


(Observational) impact of intervention (not real effect because of dependency on stone size)

(Observational) impact of stone size on both treatments and effect

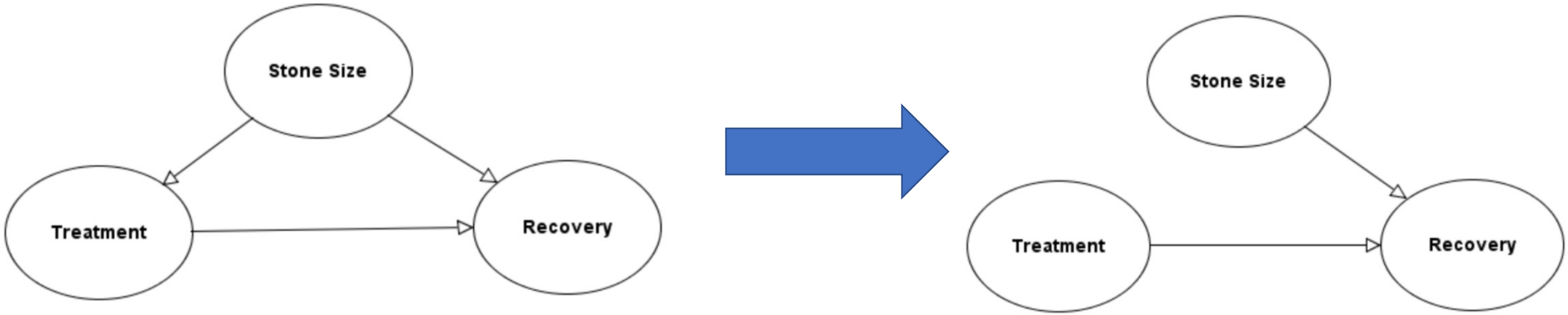


Explaining away (back propagation). If we know patients did not recover and stone was small it is much more likely A was used



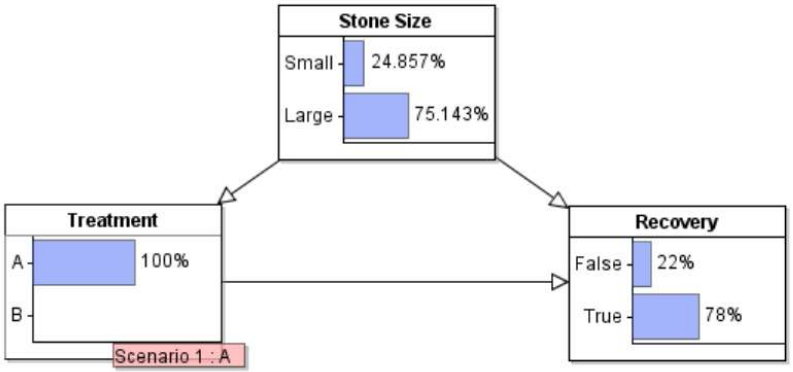
To calculate the unbiased result…

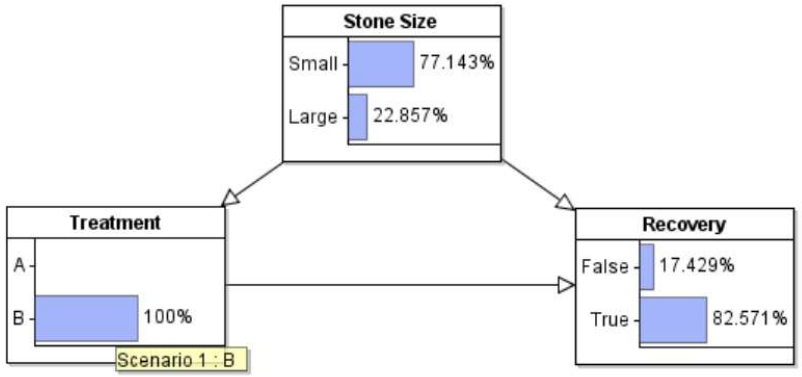
We simply simulate the effect of an intervention in the model by cutting the link from stone size to treatment



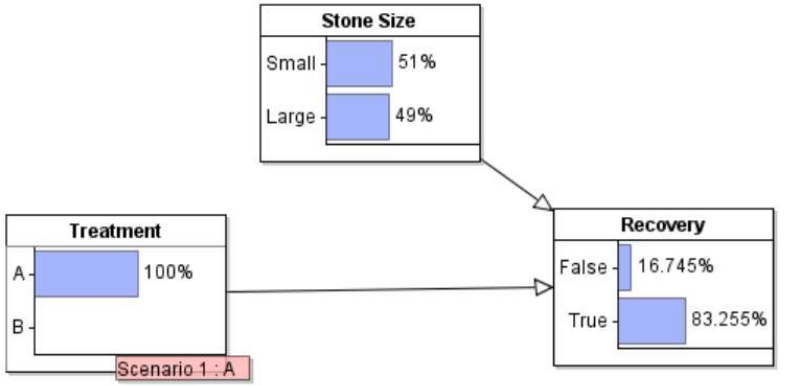
Interventions (rung 2)

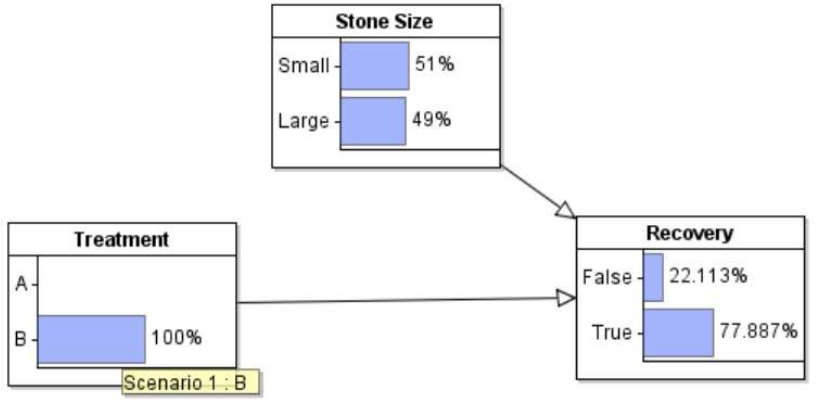
From observational model (does not measure the real effect of an intervention)





Correct effect of an intervention (simulates RCT)

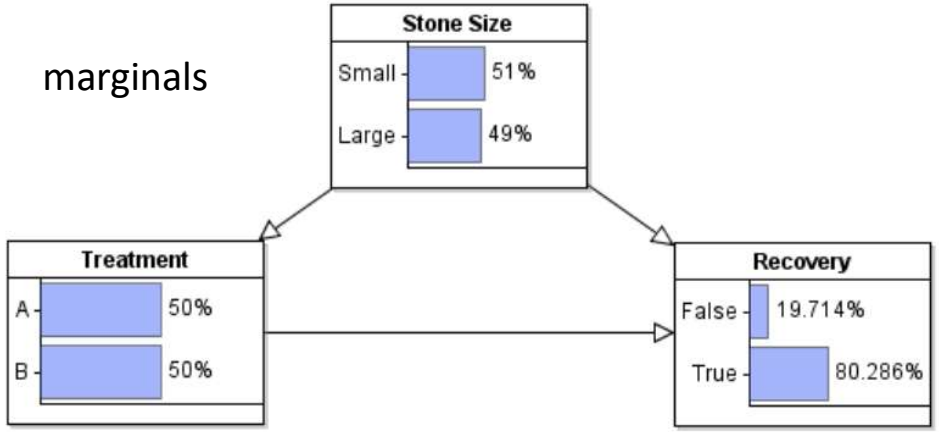




…and if we do not know the stone size

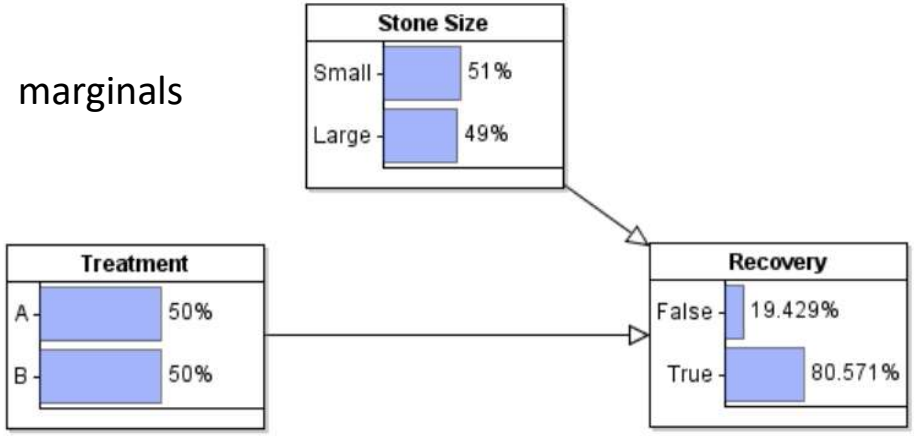
Observational model

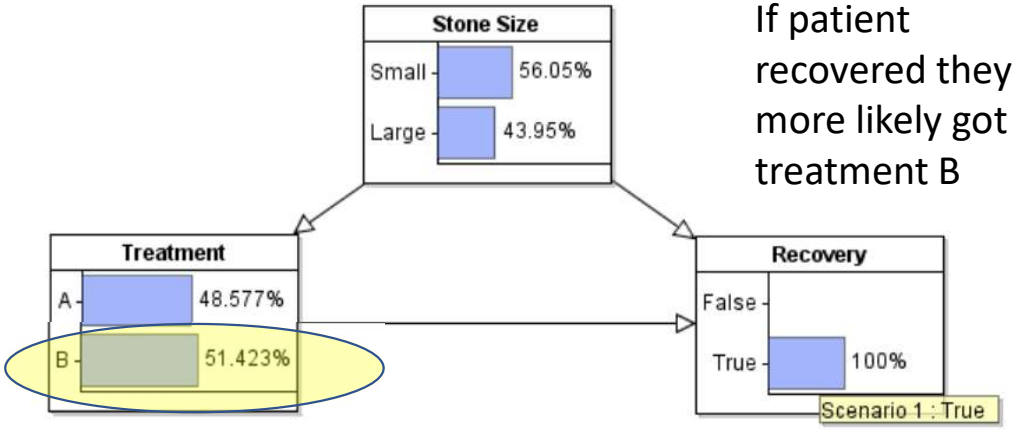
marginals



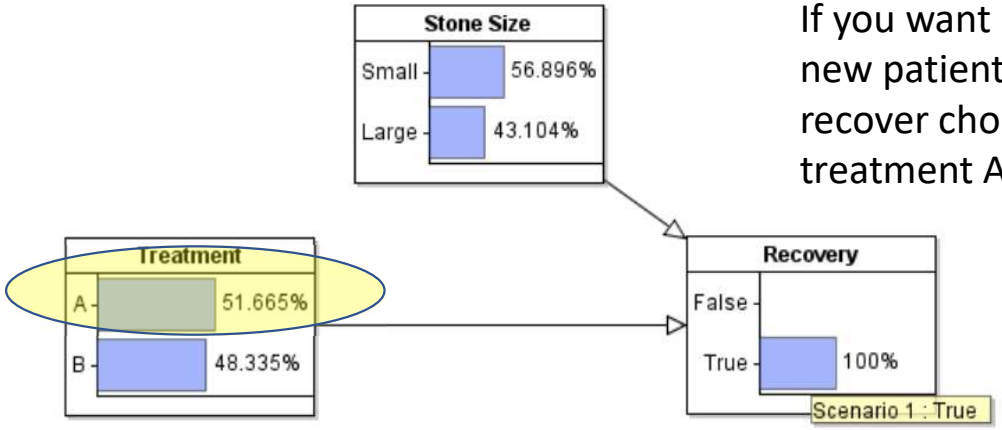
Intervention Model

marginals





If patient recovered they more likely got treatment B



If you want a new patient to recover choose treatment A

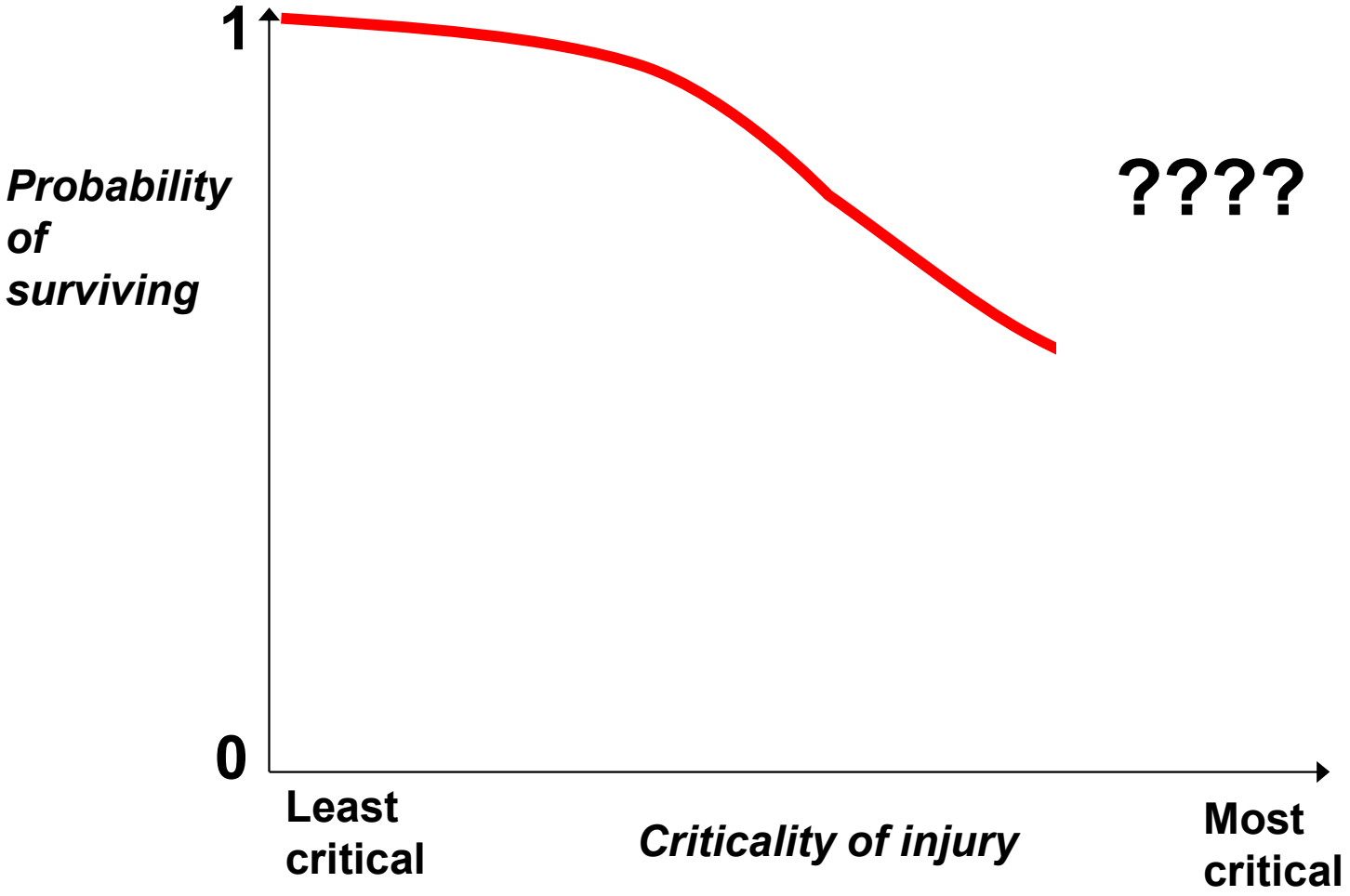
If I do not know patient stone size which treatment should we give?

Trauma injury Case Study

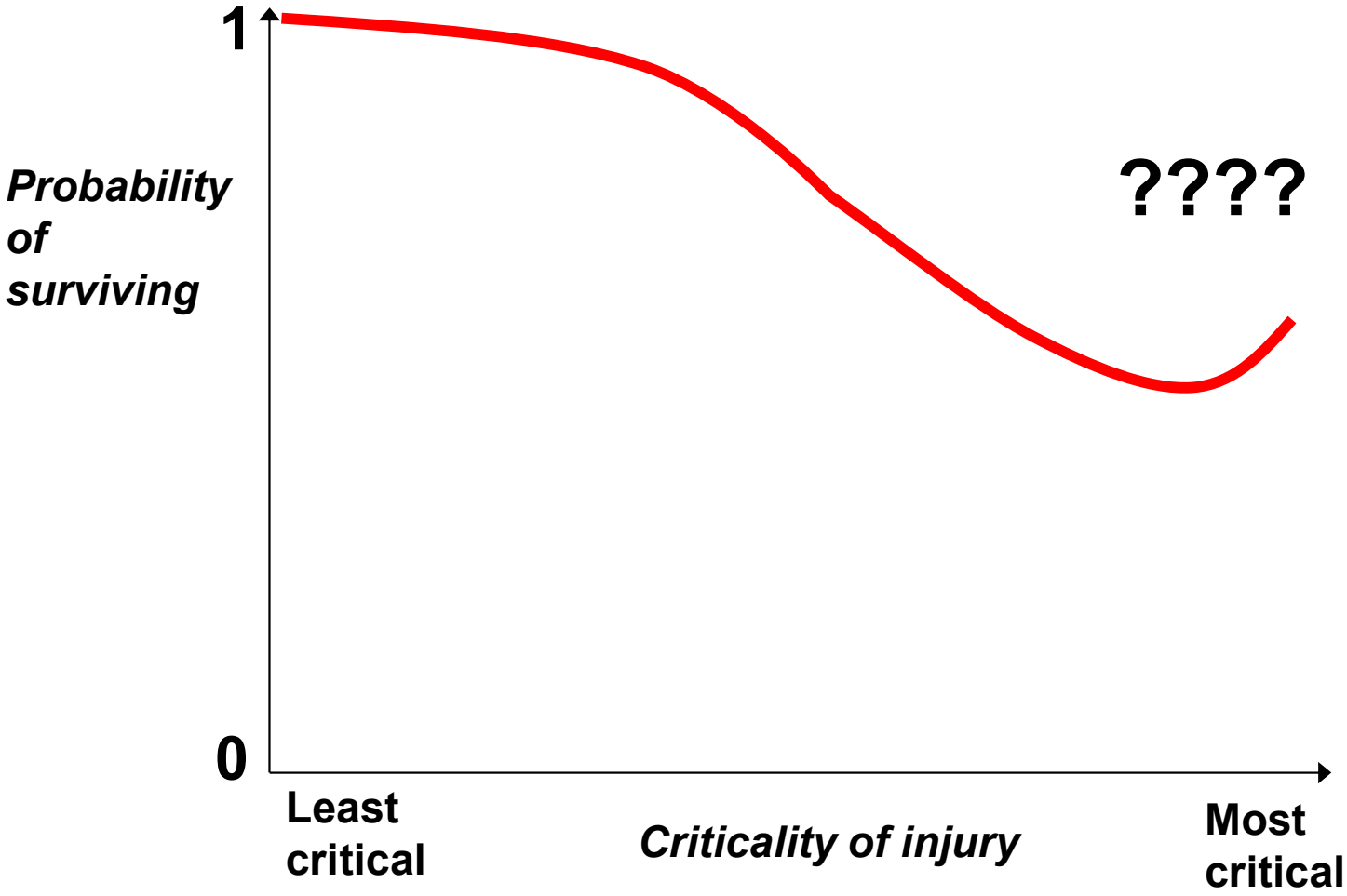
A typical data-driven study

Age	Delay in arrival	Injury type	Brain scan result	Arterial pressure	Pupil dilation	Outcome (death y/n)
17	25	A	N	L	Y	N
39	20	B	N	M	Y	N
23	65	A	N	L	N	Y
21	80	C	Y	H	Y	N
68	20	B	Y	M	Y	N
22	30	A	N	M	N	Y
…	…	…		..	…	…

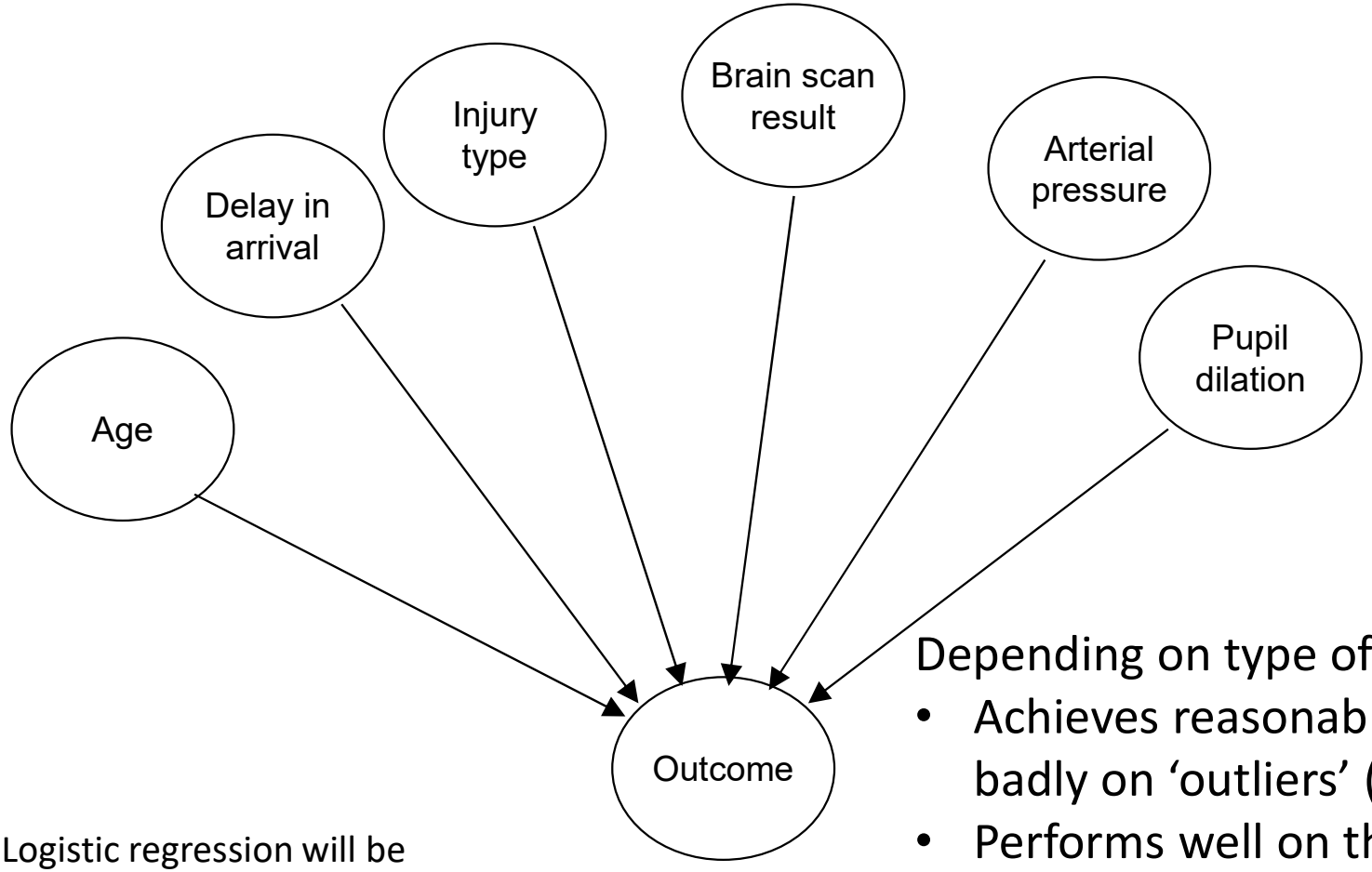
Traumatic head injury: Some observational data



Traumatic head injury: Some observational data



Regression model* learnt purely from data ('supervised learning')

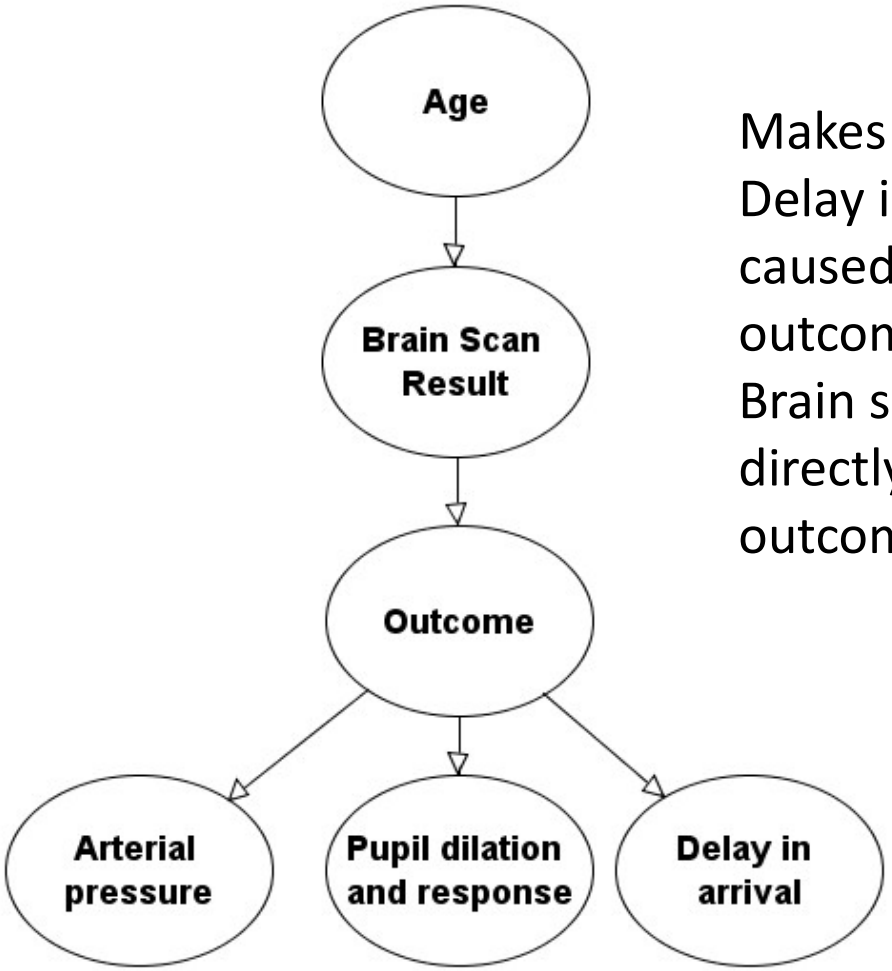


Depending on type of regression:

- Achieves reasonable accuracy but performs badly on 'outliers' (e.g fails to model 'uptick')
- Performs well on this data but is overfitted (will be poor for new data)

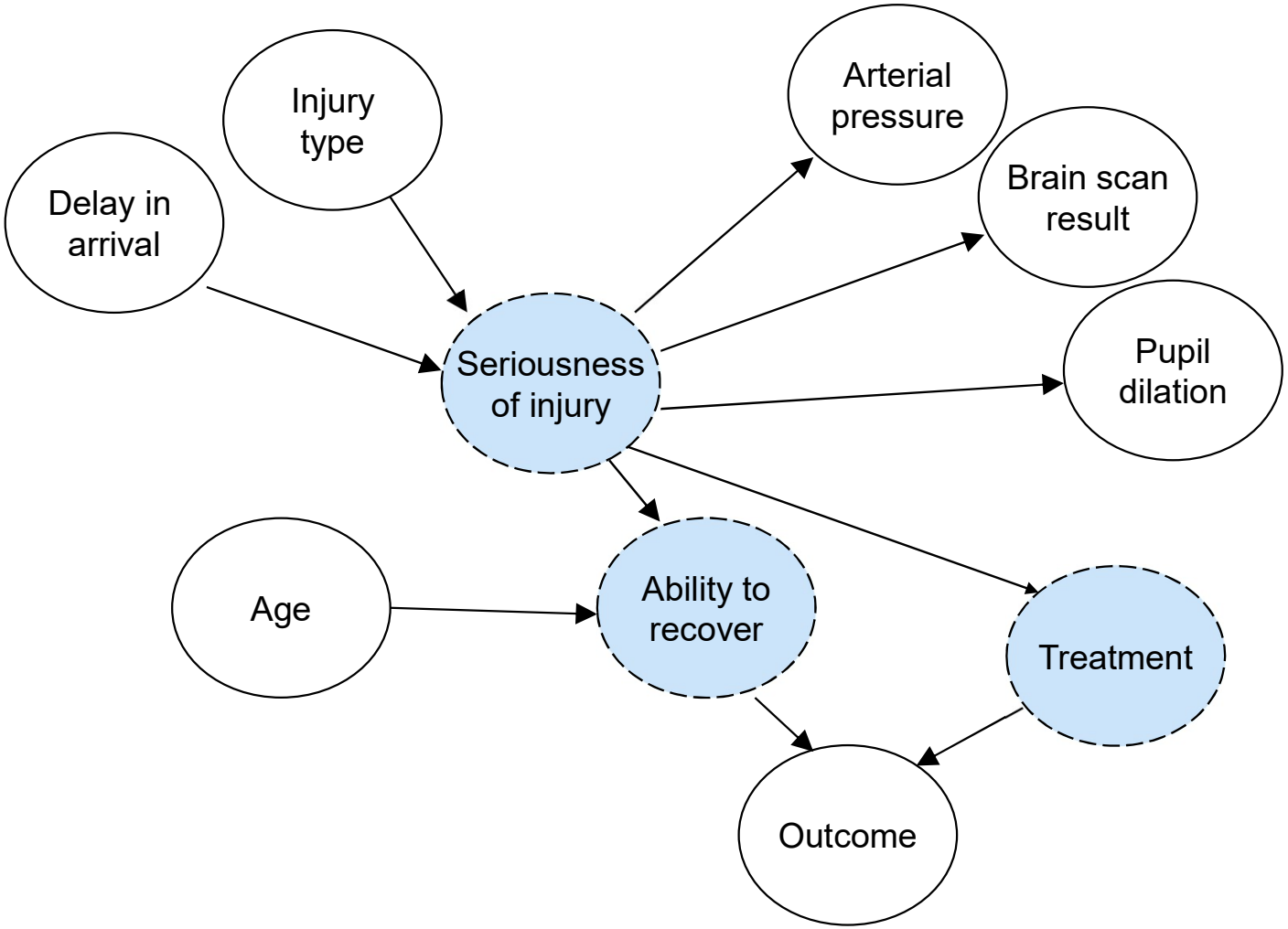
*Logistic regression will be covered in Lesson 8

Structural BN learnt purely from data



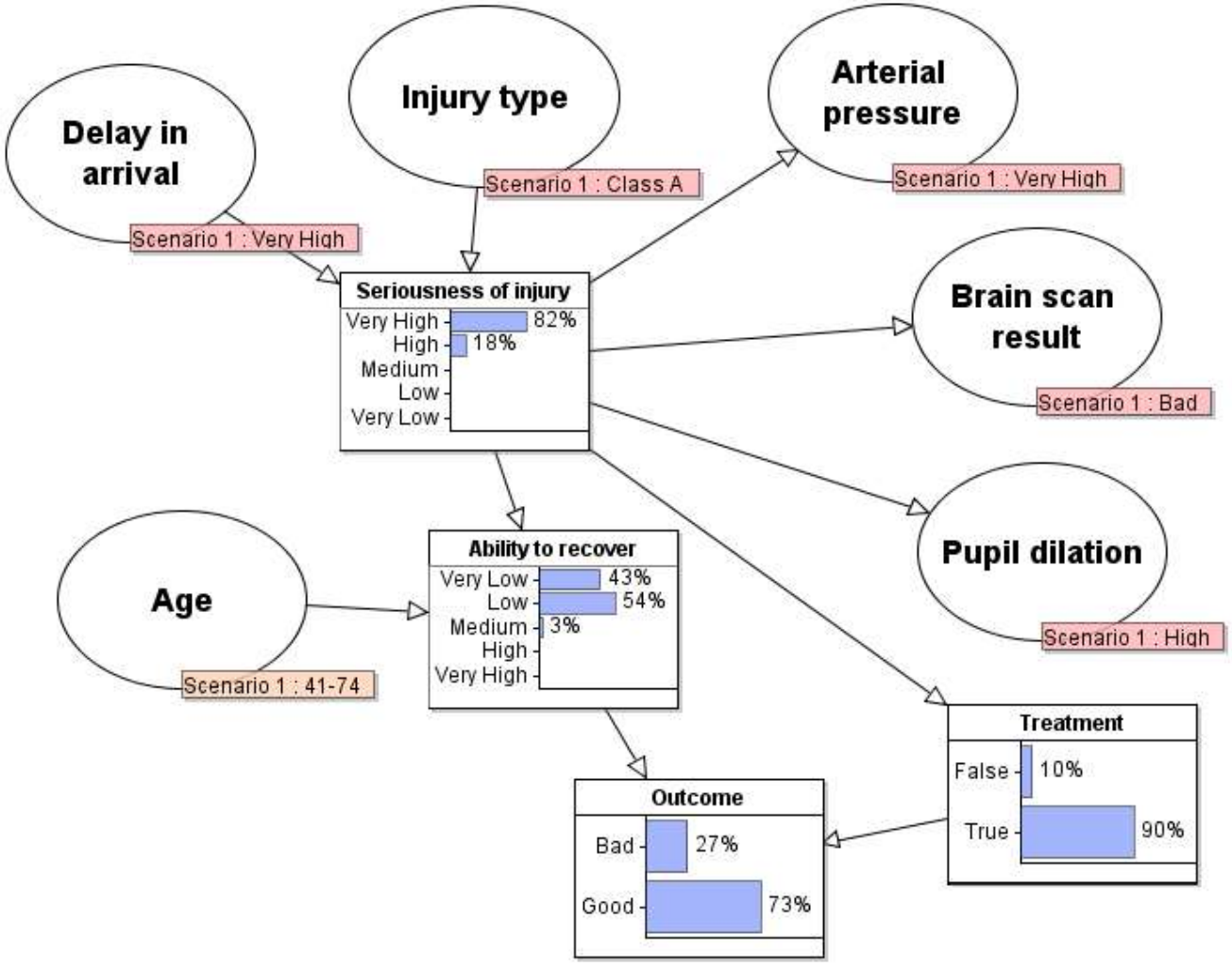
Makes no causal sense
Delay in arrival cannot be caused or influenced by the outcome
Brain scan results cannot directly influence or cause the outcome etc

Expert causal BN with hidden explanatory and intervention variables



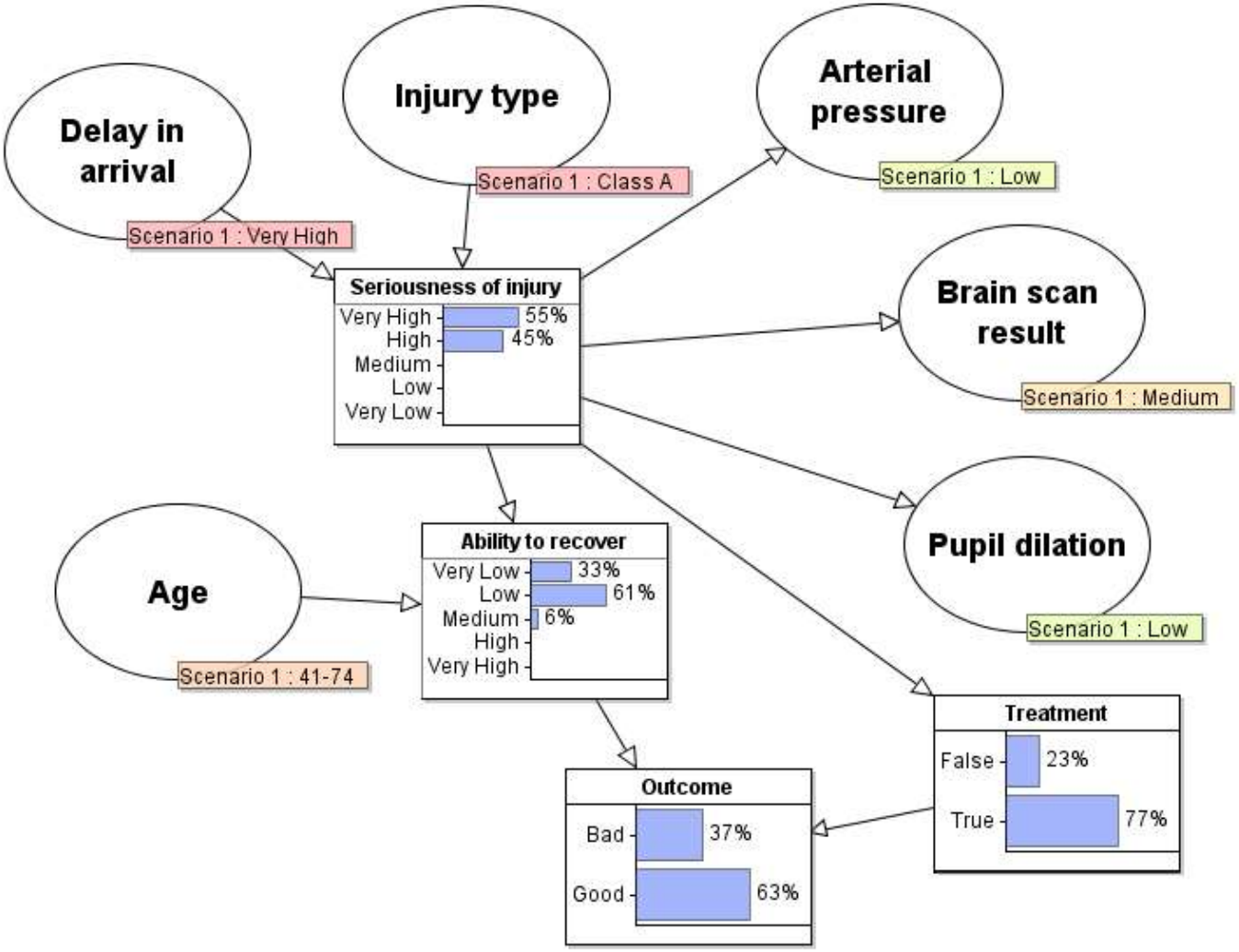
More accurate (properly handles outliers), more useful for decision support

Results in worst case scenario



Better than expected outcome predicted because of likely intervention

A less bad scenario….

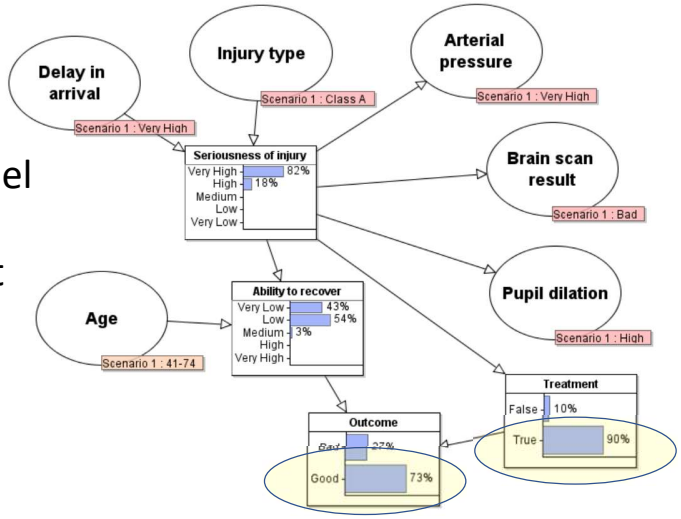


Note: outcome predicted is worse because less likely intervention

But we can use the BN to simulate the intervention

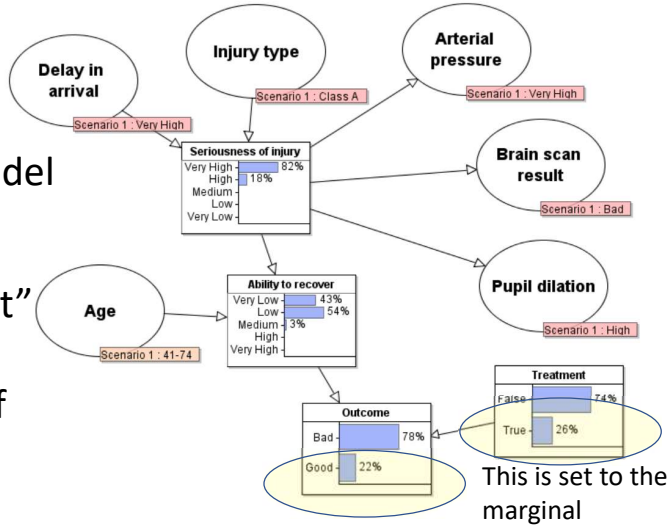
Observational model

shows Treatment very likely to be given in this scenario



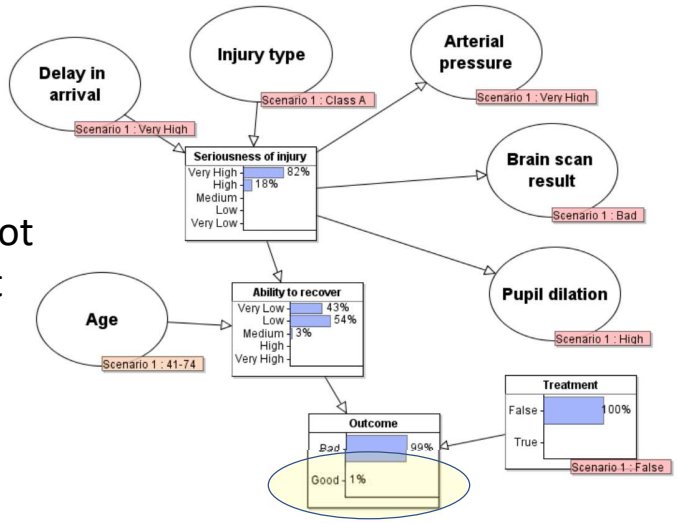
Interventional model

By breaking the link to "Treatment" we can simulate the 'real' effect of the intervention

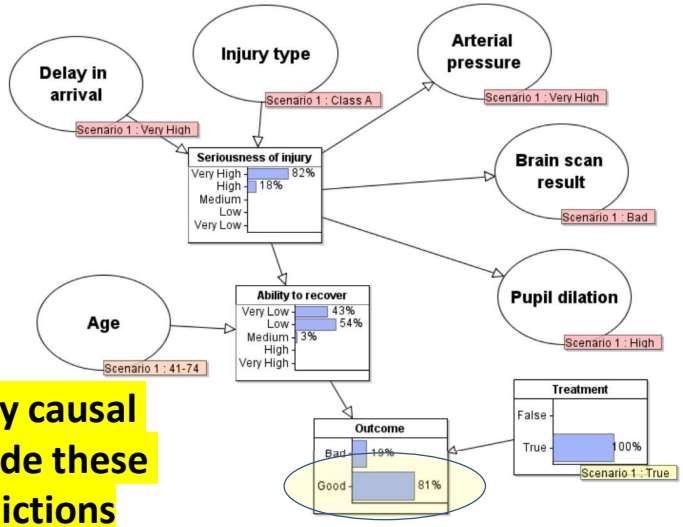


This is set to the prior marginal

…if we decide not to do treatment



…if we decide to do treatment



As Pearl said, only causal models can provide these insights and predictions

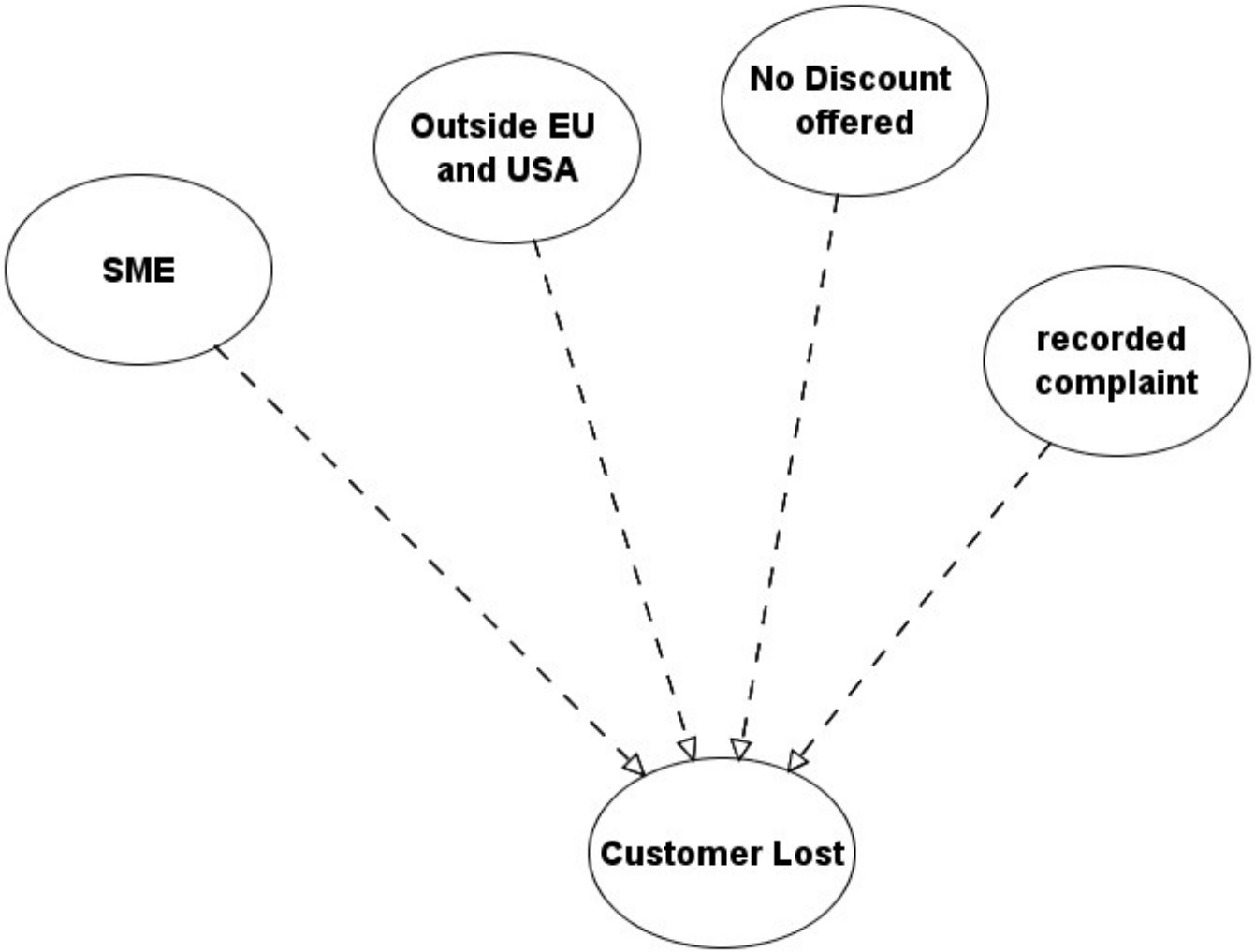
Customer retention

	"RISK FACTORS"					
Customer number	SME	Outside EU/USA	No Discount offered	Recorded complaint		Customer Lost
1	N	N	N	N		N
2	N	N	Y	N		Y
3	N	N	N	N		N
4	N	N	N	N		N
5	Y	N	N	N		N
6	N	N	Y	N		N
7	N	N	N	Y		Y
8	N	N	N	N		Y
9	Y	N	N	N		N
10	N	N	N	N		N
11	N	N	N	N		N
12	N	Y	N	N		N
....						
9999	Y	N	N	Y		Y
10000	N	Y	N	N		N

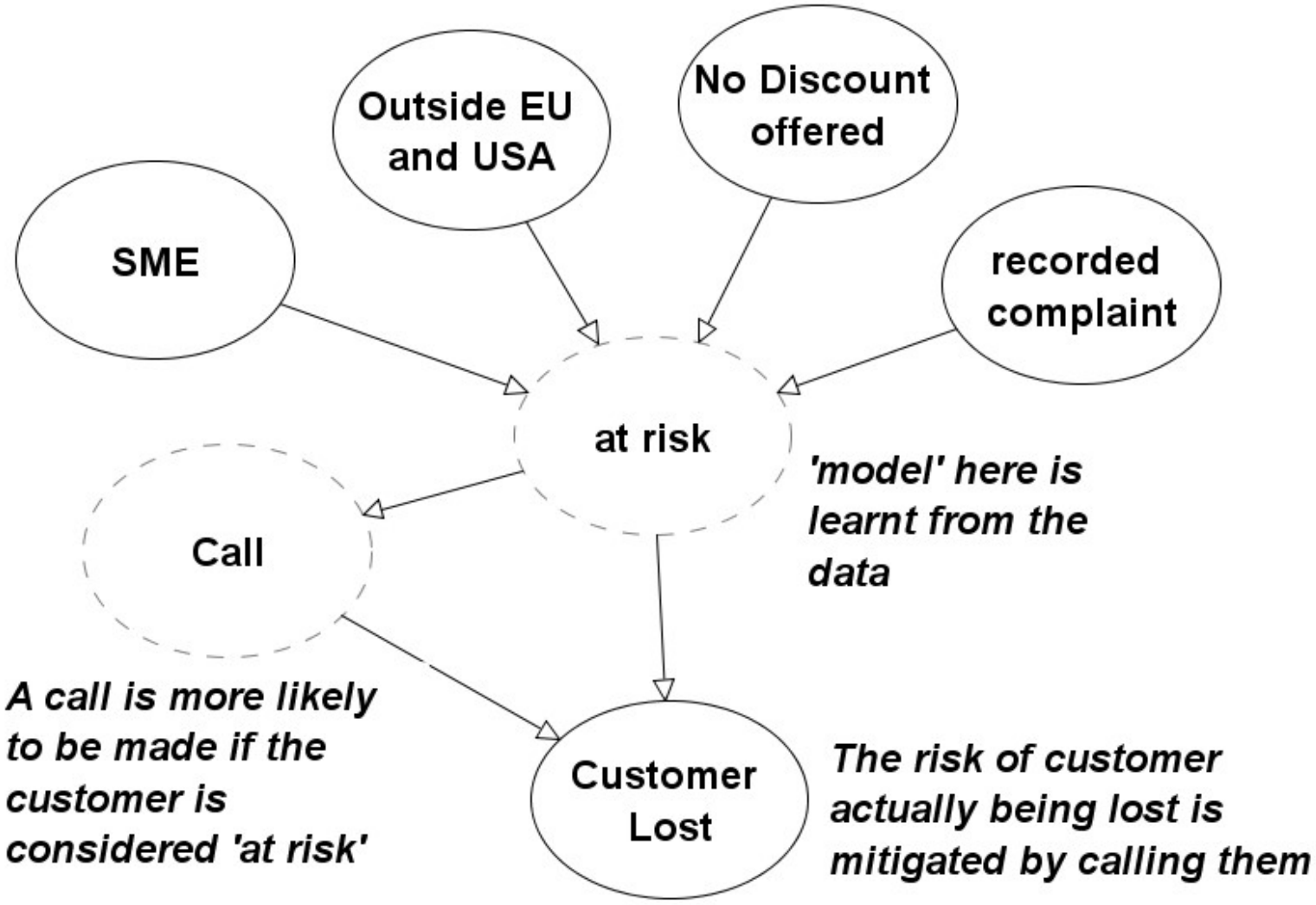
Risk Factor ‘profiles’

"RISK FACTOR PROFILE"						
SME	Outside EU/USA	No Discount offered	Recorded complaint	Number with this profile	Total with this profile lost	% of those with this profile lost
N	N	N	N	3213	270	8.4
N	N	N	Y	357	96	27.0
N	N	Y	N	2142	523	24.4
N	Y	N	N	567	108	19.1
Y	N	N	N	1377	226	16.4
N	N	Y	Y	238	76	31.8
N	Y	N	Y	63	19	30.2
N	Y	Y	N	378	108	28.6
Y	Y	N	N	243	58	23.8
Y	N	Y	N	918	253	27.6
Y	N	N	Y	153	45	29.4
N	Y	Y	Y	42	14	33.1
Y	N	Y	Y	102	33	32.8
Y	Y	N	Y	27	9	31.7
Y	Y	Y	N	162	49	30.5
Y	Y	Y	Y	18	6	33.7

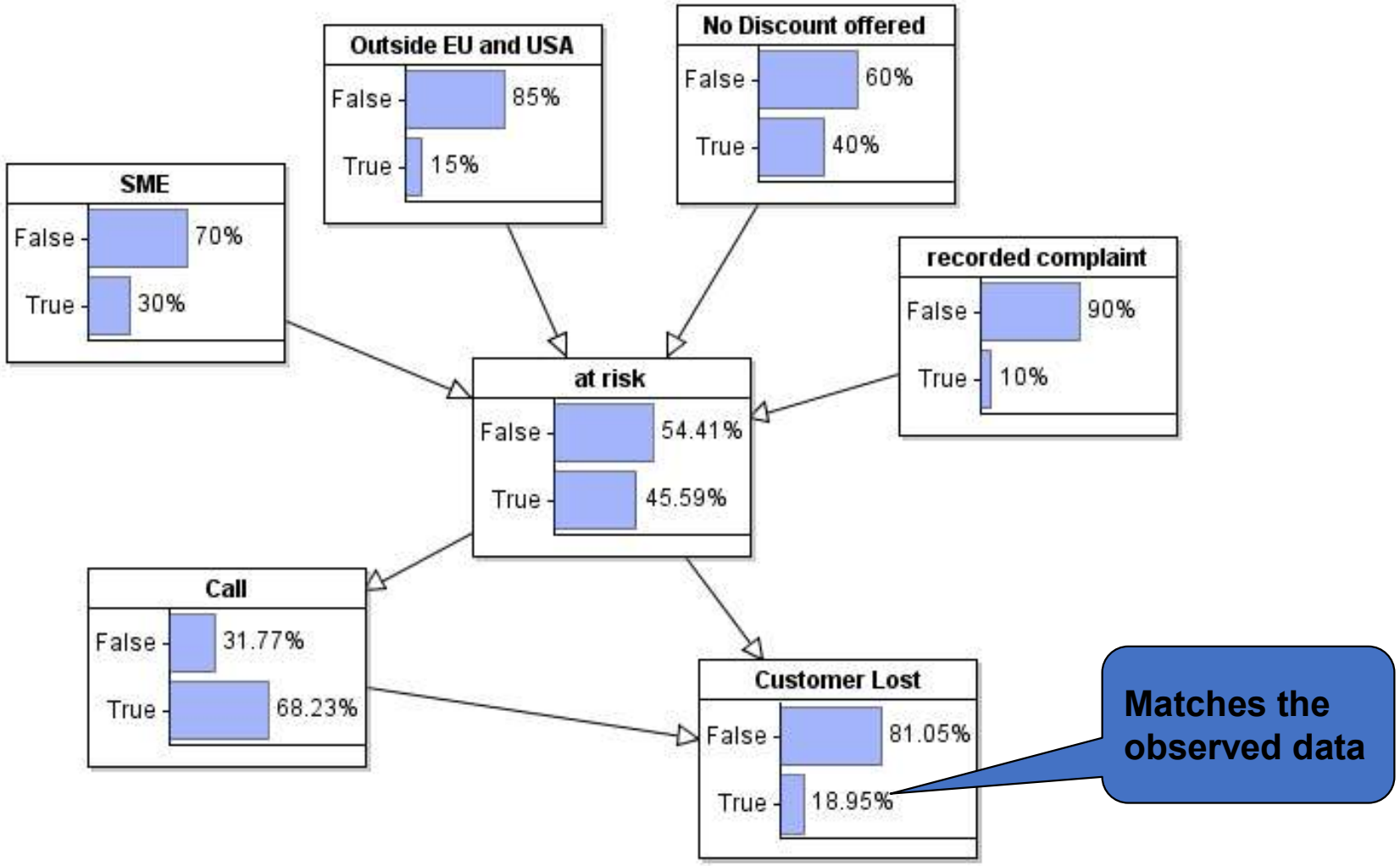
Temptation to build this model (‘supervised learning’)



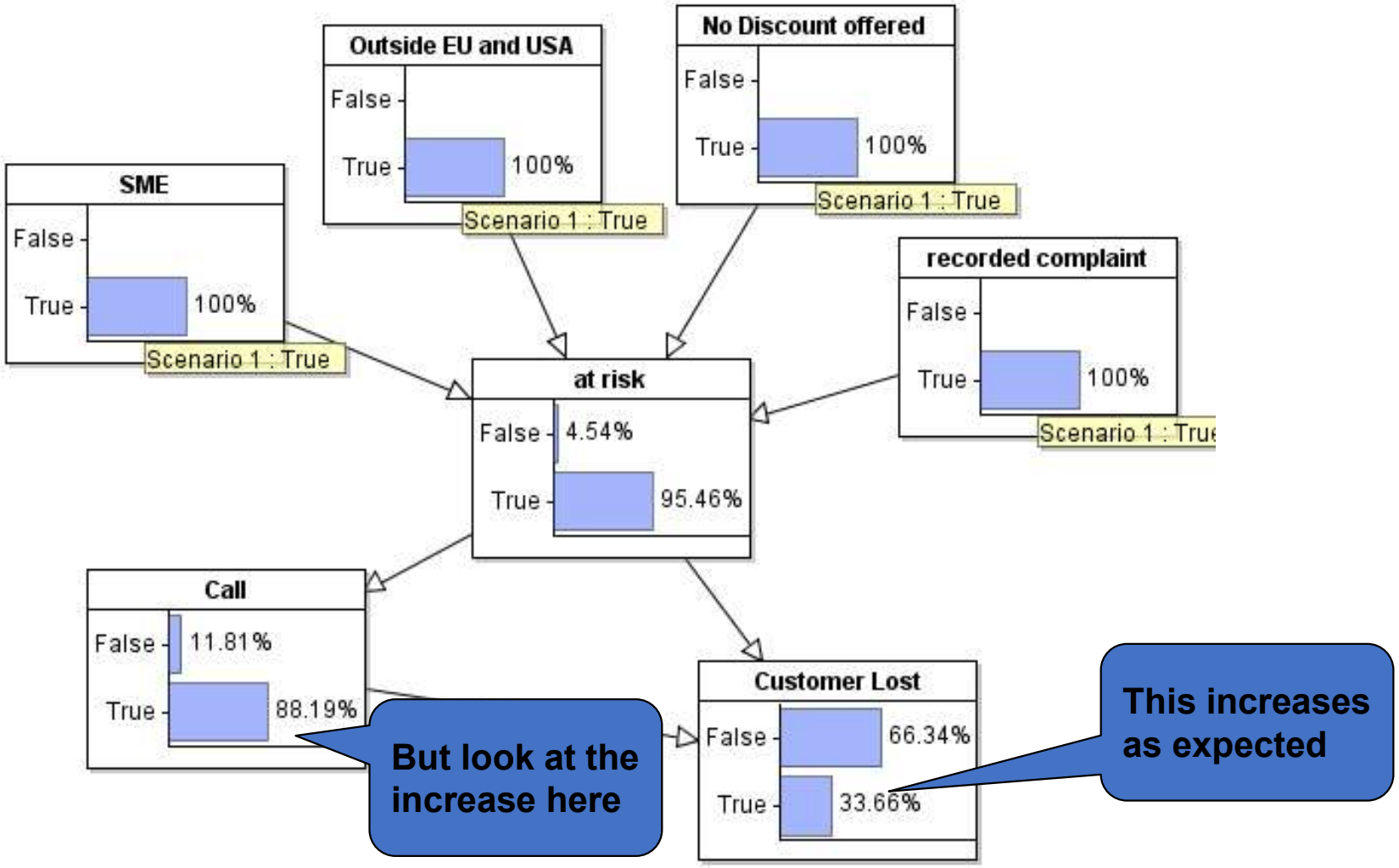
Causal BN built with data and knowledge



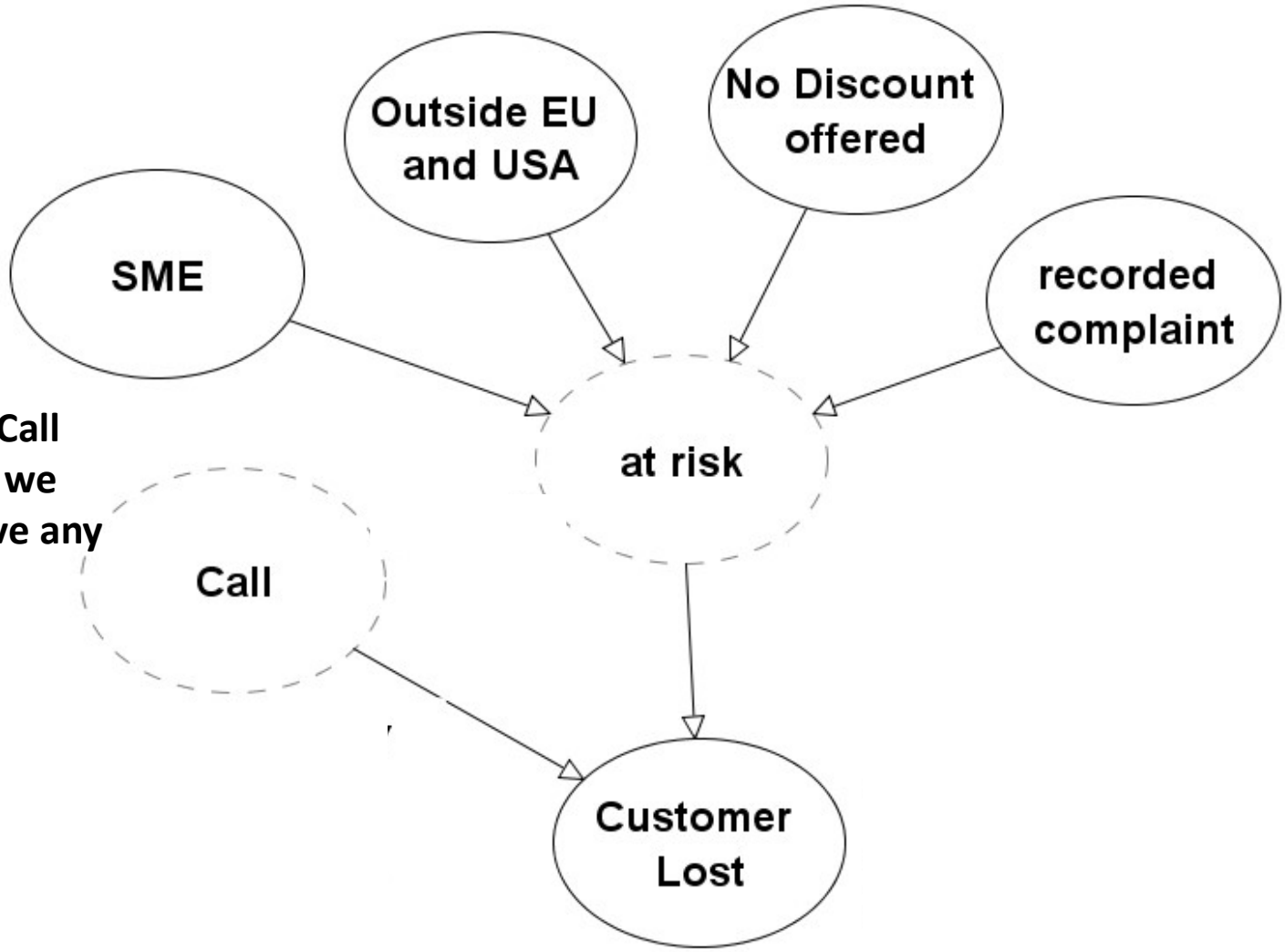
Model with marginal probabilities



When risk factors are all true

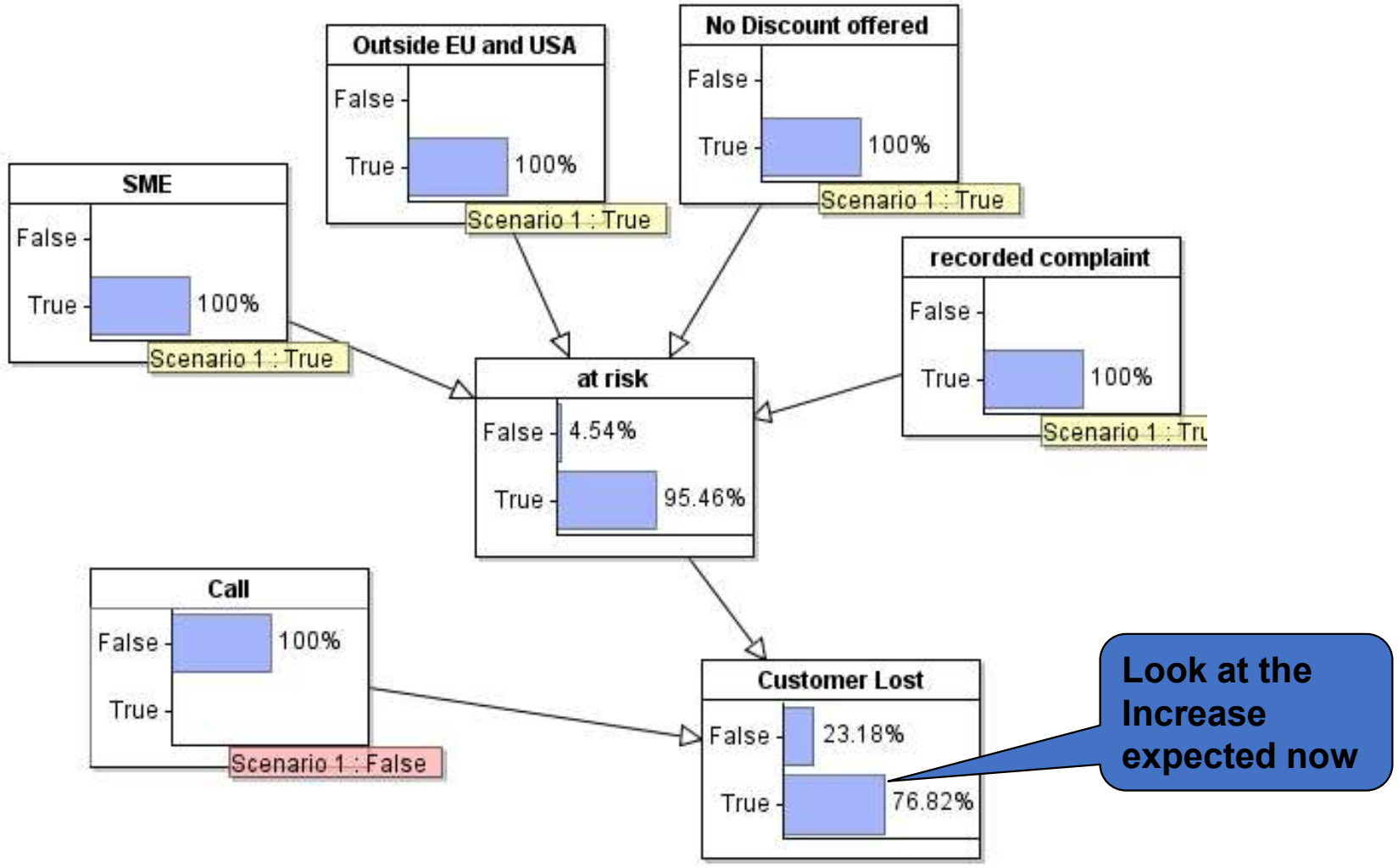


The intervention model

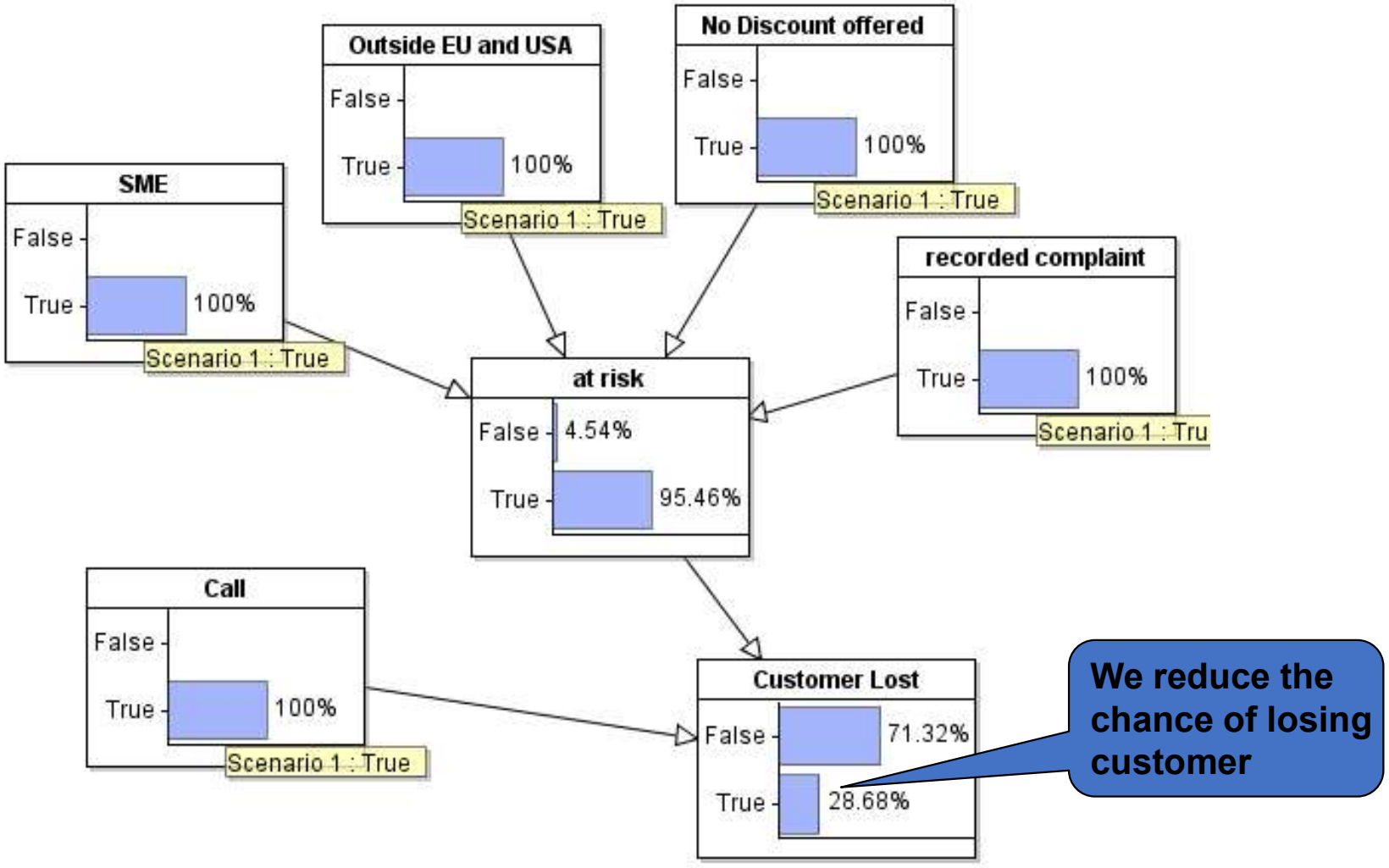


To simulate a Call 'intervention' we have to remove any link to Call

Using the intervention model: If we *do not* make the call …



Using the intervention model: If we *do* make the call …



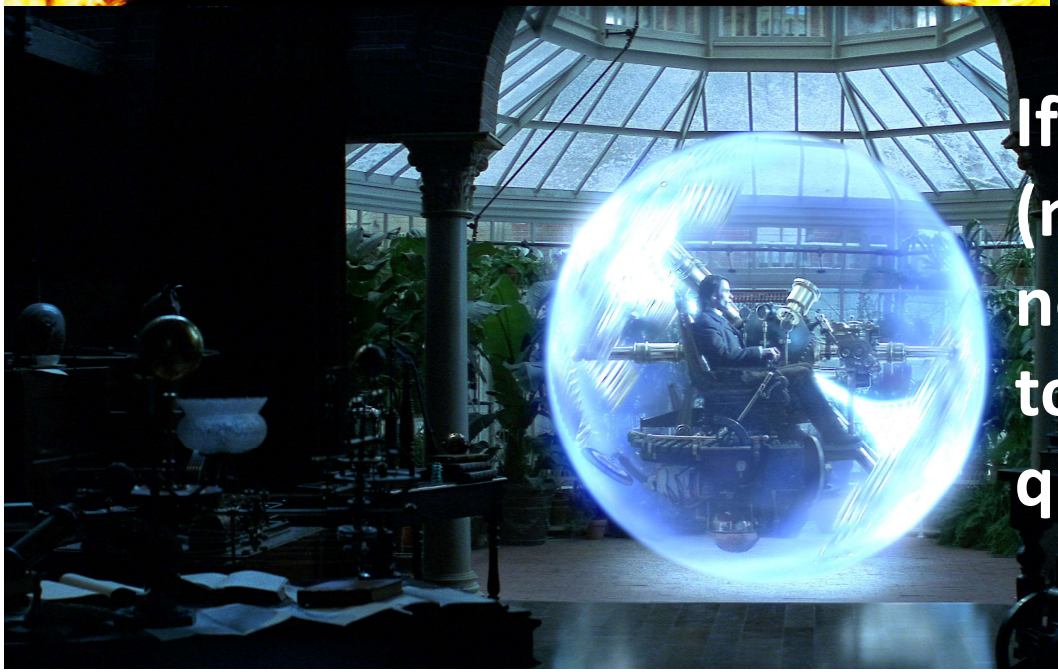
Counterfactual reasoning (rung 3)

We know Hitler did rise to power and one of the outcomes of this was that Churchill became PM. Would Churchill still have become PM if Hitler had never risen to power?

We know this patient survived after being given treatment T. Would this patient still have survived if he had not been given treatment T?

I got home late after taking the M25. Would I have got home earlier if I had taken the A406?

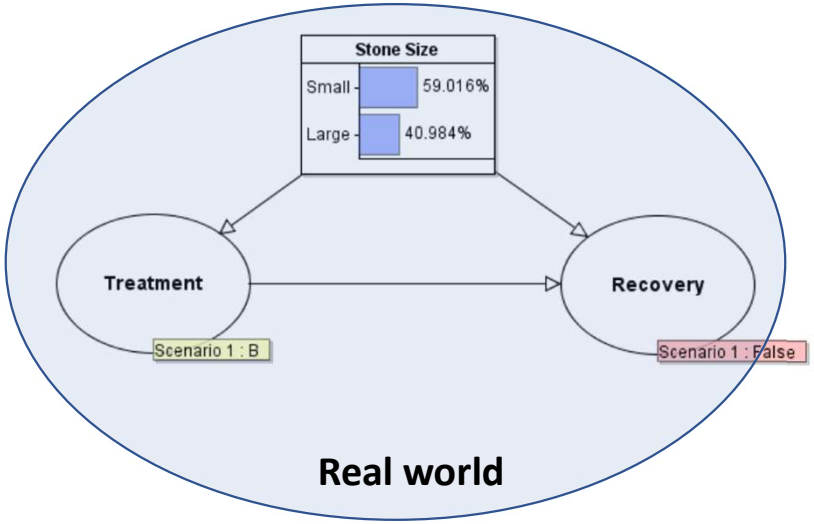
The Good News



If we have some data (not even a lot) we do not need a time machine to answer counterfactual questions

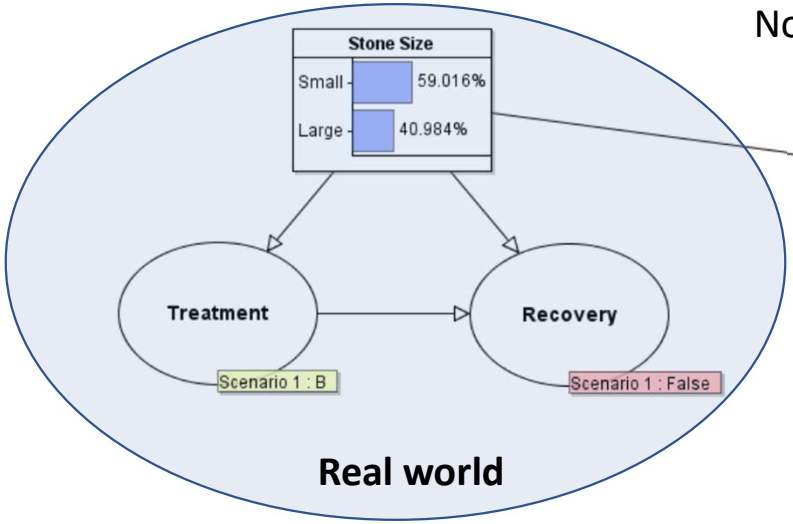
Counterfactual reasoning (rung 3)

In ‘real world’ patient had treatment B but did not recover
From this we learn the stone size was most likely small (but because the patient did not recover this is less likely that for most patients taking treatment B)

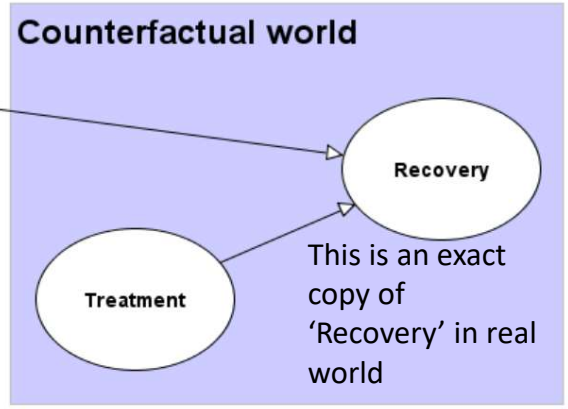


Counterfactual reasoning (rung 3)

In ‘real world’ patient had treatment B but did not recover
From this we learn the stone size was most likely small (but because the patient did not recover this is less likely that for most patients taking treatment B)

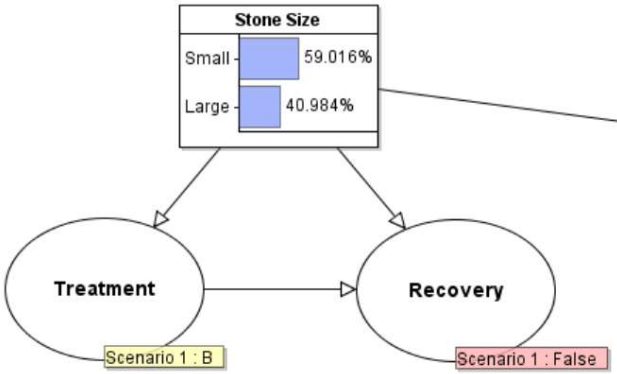


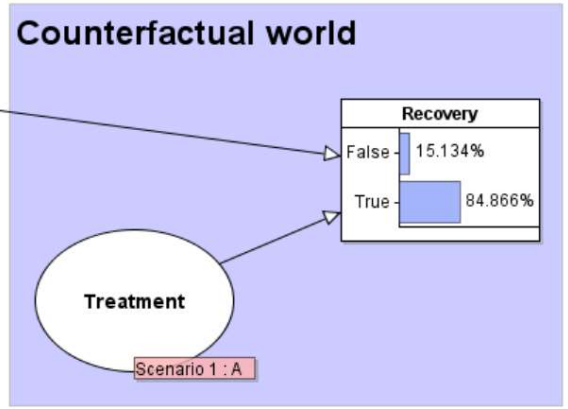
Now we create a counterfactual world



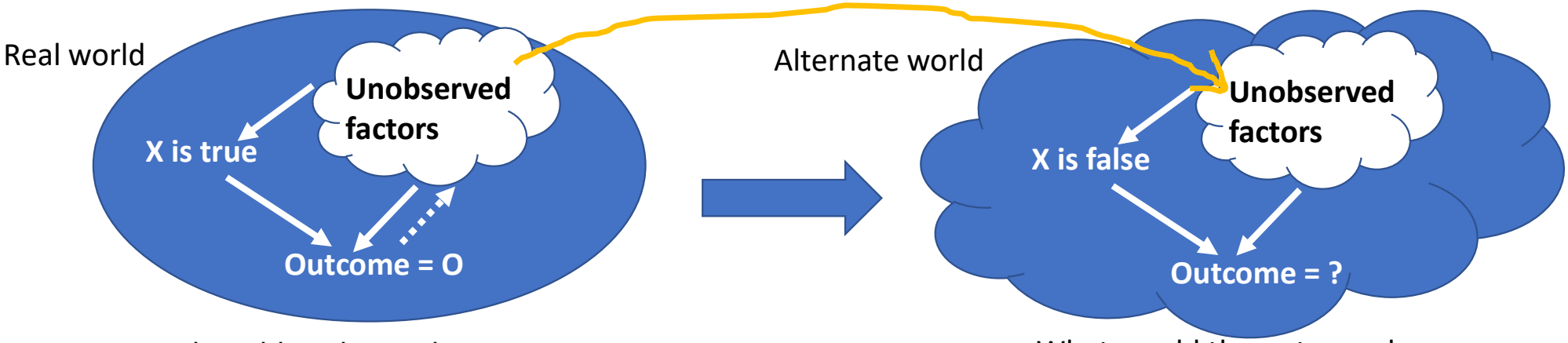
In counterfactual world we can simulate effect of treatment A on this patient (with updated information about stone size).

The patient would very likely have recovered





Counterfactual reasoning (rung 3)



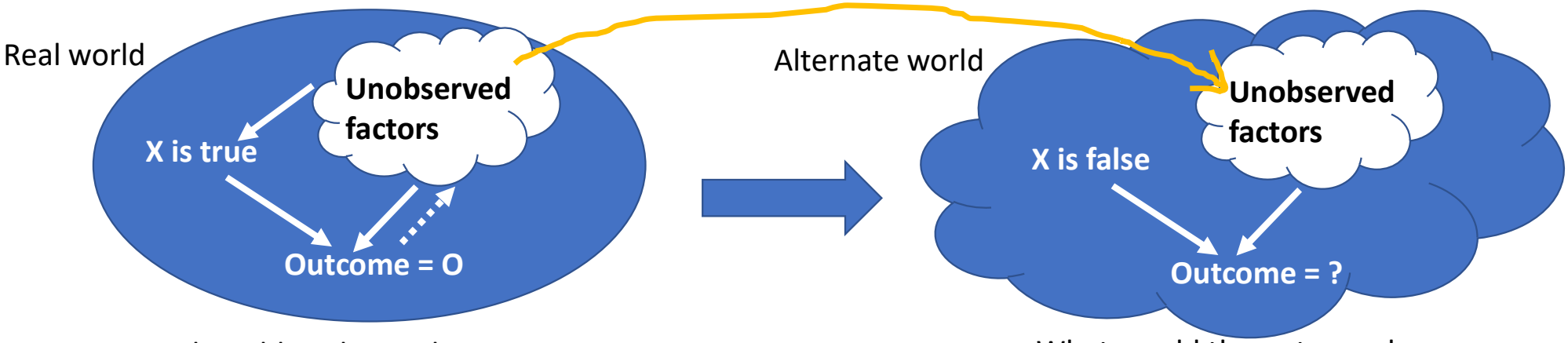
In real world we know the outcome was O when X was true

What would the outcome have been if X had been false?

STEP 1 Learn from what happened in the 'real world': Knowing outcome was O when X was true updates our knowledge about all the unobserved factors in the real world; e.g. because the patient did not recover with treatment B, we learnt that the stone size was likely larger than originally believed.

STEP 2 Create 'alternate' world: In alternate world everything is the same except the changed prior beliefs about the unobserved factors.

Counterfactual reasoning (rung 3)



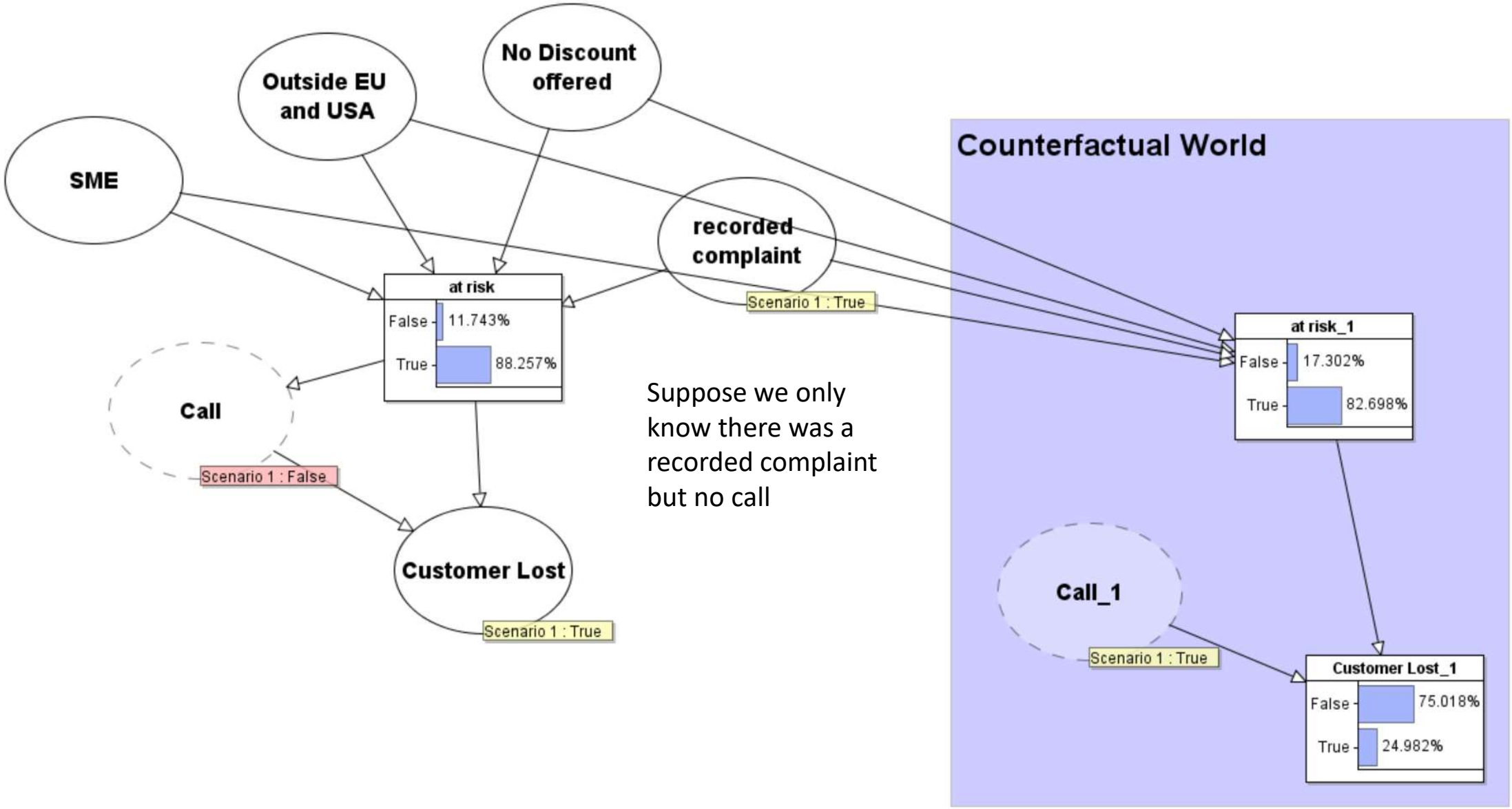
In real world we know the outcome was O when A was true

What would the outcome have been if A had been false?

STEP 1 Learn from what happened in the 'real world': Knowing outcome was O when X was true updates our knowledge about all the unobserved factors in the real world; e.g. because the patient did not recover with treatment B, we learnt that the stone size was likely larger than originally believed.

STEP 2 Create 'alternate' world: In alternate world everything is the same except the changed prior beliefs about the unobserved factors.

STEP 3: Simulate different intervention in the alternate world: simulate the effect of making X false, i.e. simulate an intervention on X by breaking all links to X in the alternate world.

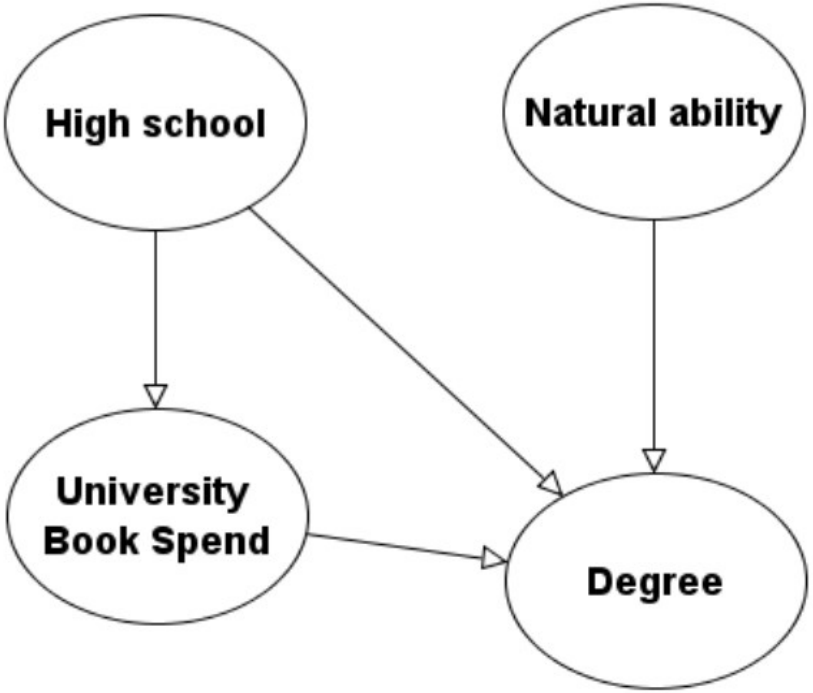


Suppose we only know there was a recorded complaint but no call

MSc Student degrees

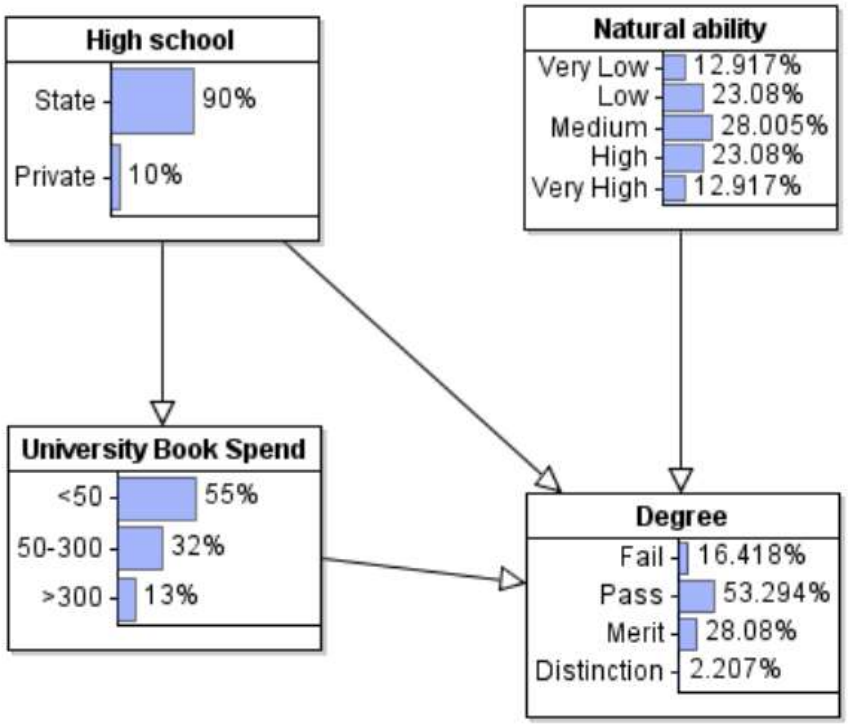
Unobservable variable
('exogenous')

Variables which may be observable and which we may wish to 'intervene' on



Outcome variable

Basic causal model

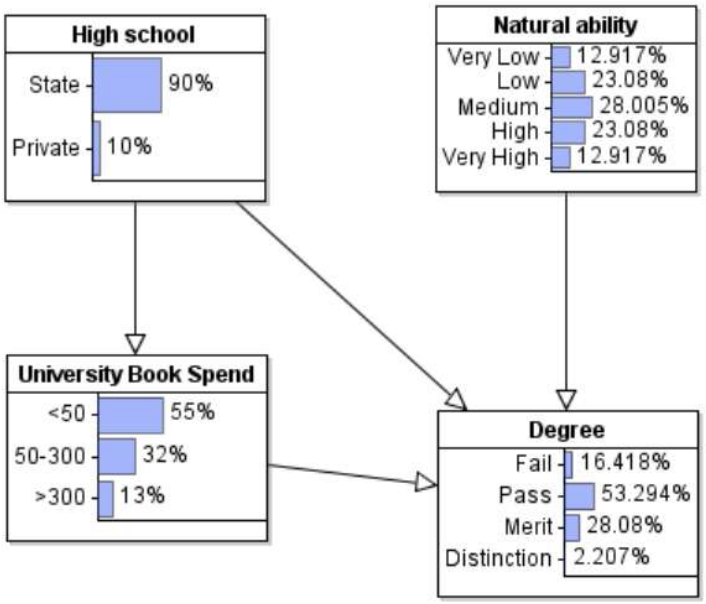


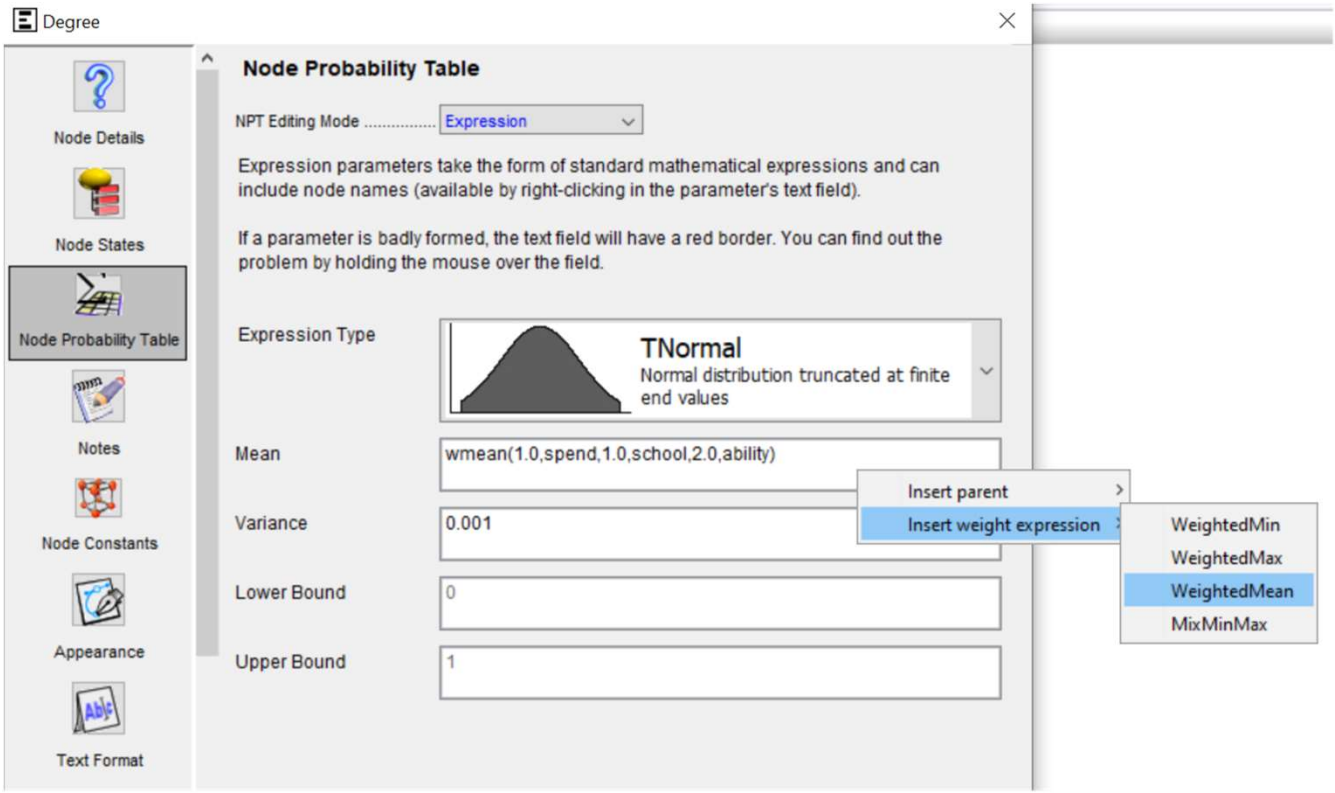
Marginal probabilities

MSc Student degrees

An aside: note the use of RANKED NODES in this example

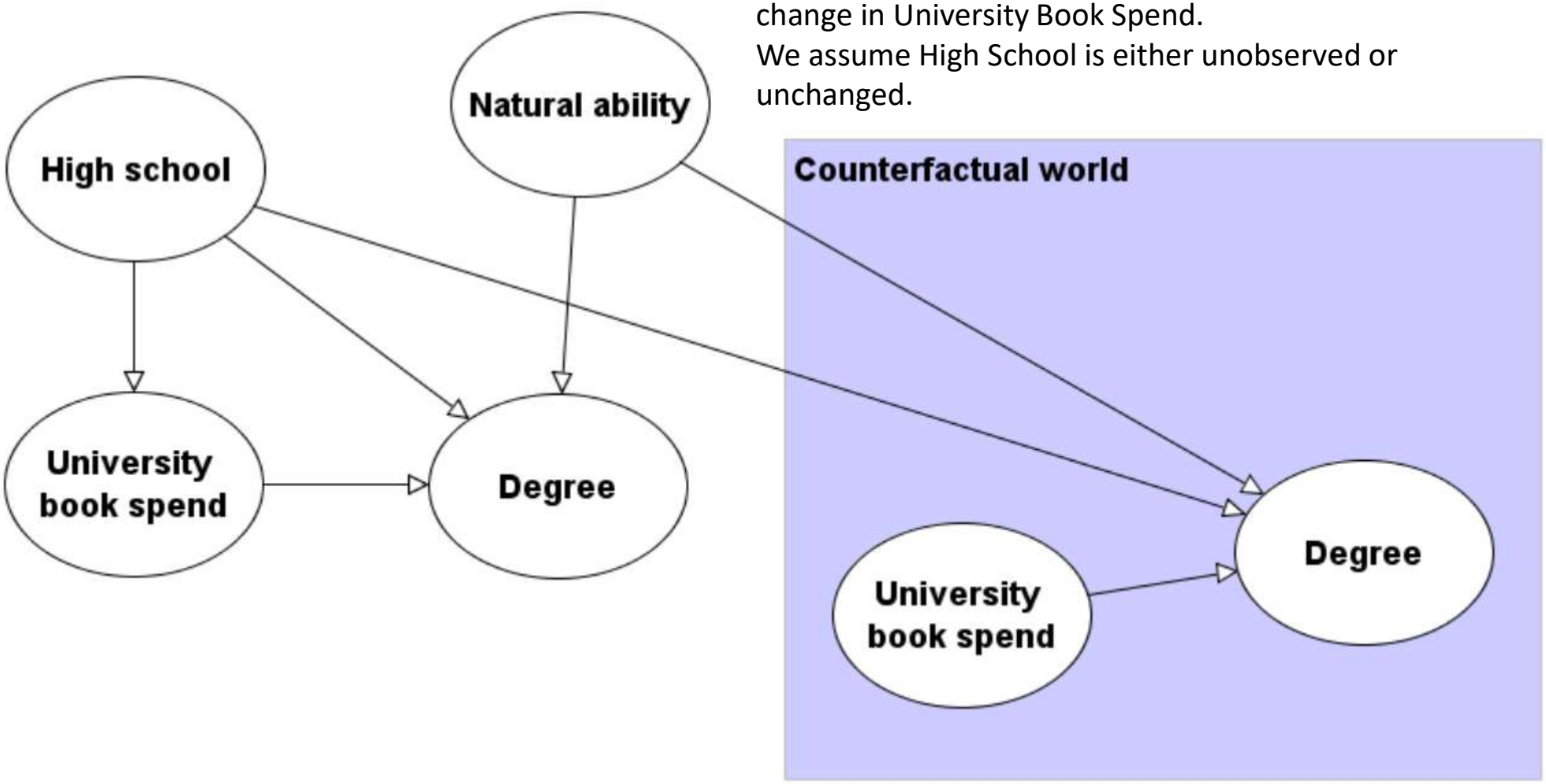
This is an extremely efficient way to get good approximations of otherwise intractable probability tables (NPTs).
By defining all the nodes here as ranked nodes the NPT for Degree is defined by a simple weighted mean function



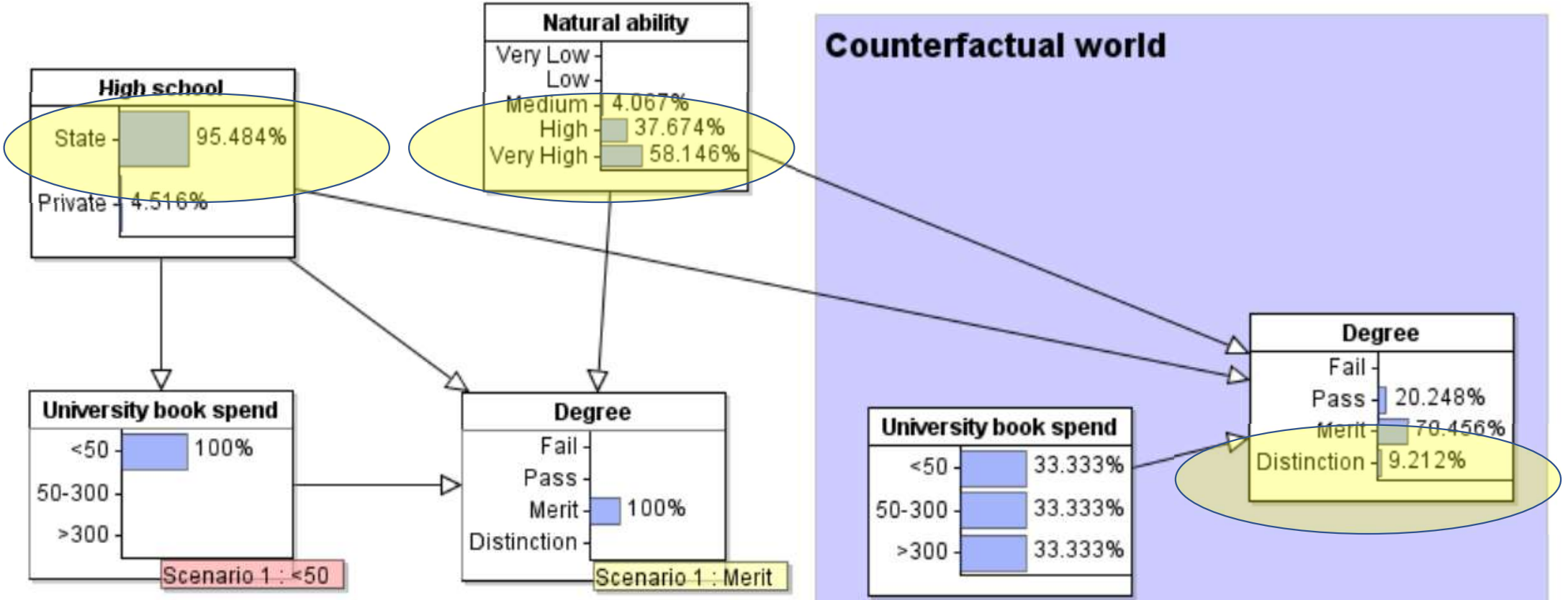


Degrees: counterfactual (1 intervention)

In this case we seek to answer counterfactuals involving change in University Book Spend.
We assume High School is either unobserved or unchanged.



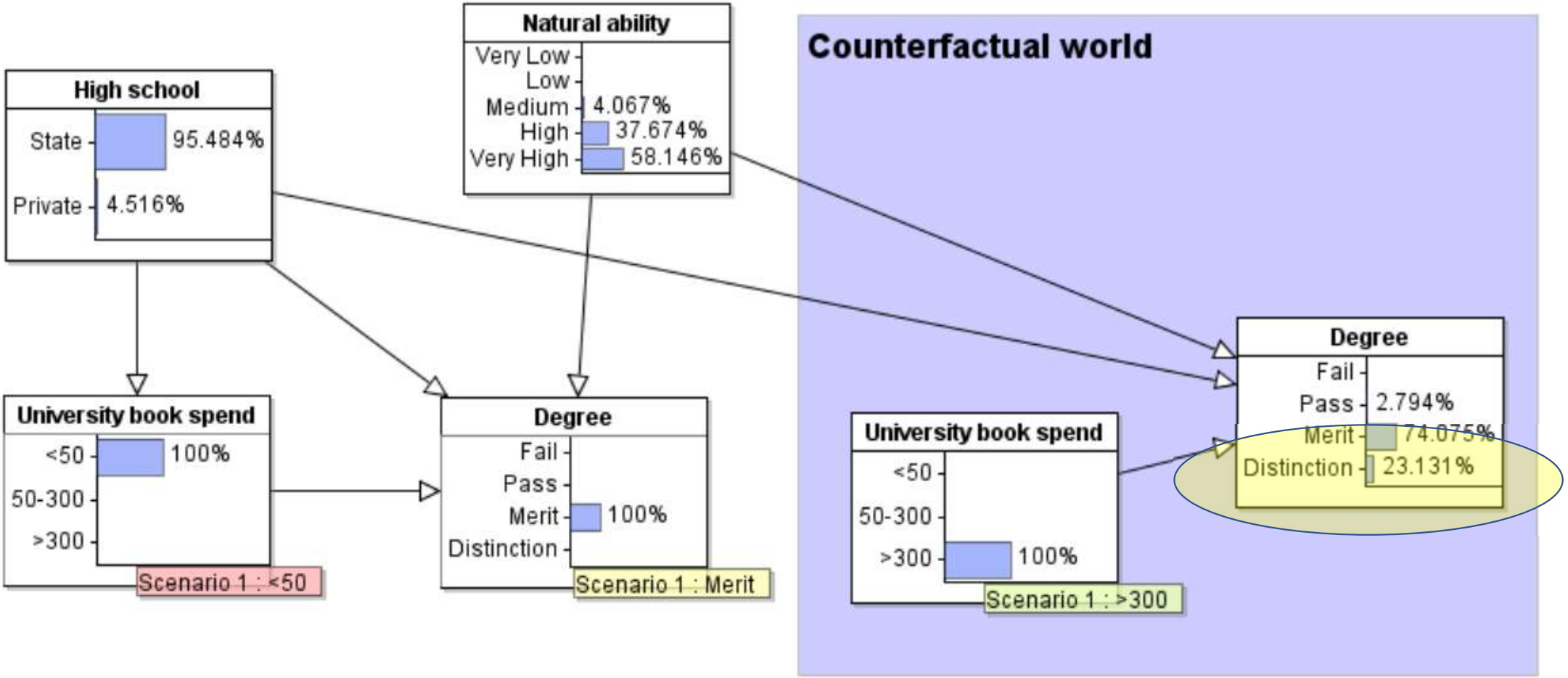
Degrees: counterfactual (1 intervention)



Student achieves a Merit despite spending less than £50
Leads to increase in probability that the student came from State school and had high natural ability

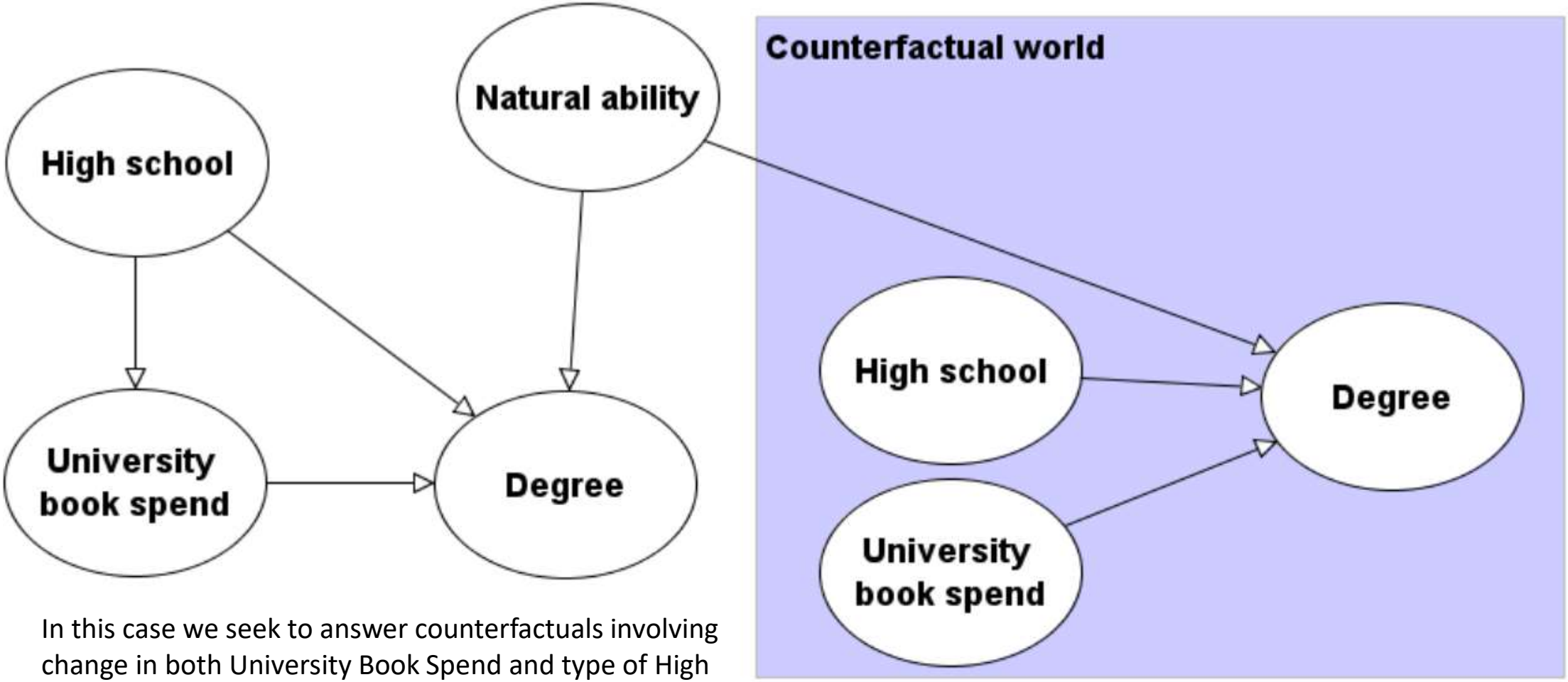
Without knowing book spend in counterfactual world this student is still unlikely to get Distinction

Degrees: counterfactual (1 intervention)



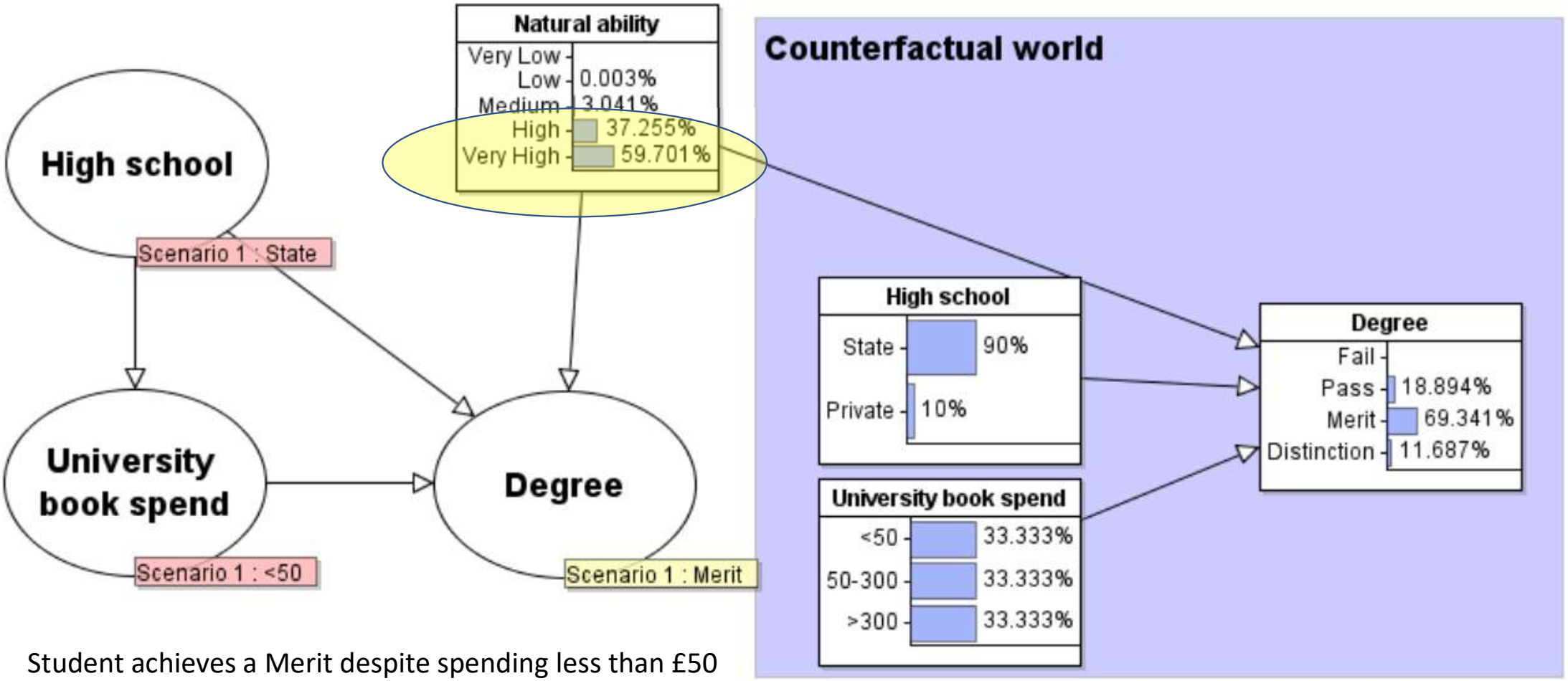
In the counterfactual world we assume the student spends >£300 on books.
We conclude there is a 23% probability this student would have got a Distinction if she had spent extra on books (a significant increase)

Degrees: counterfactual (2 interventions)



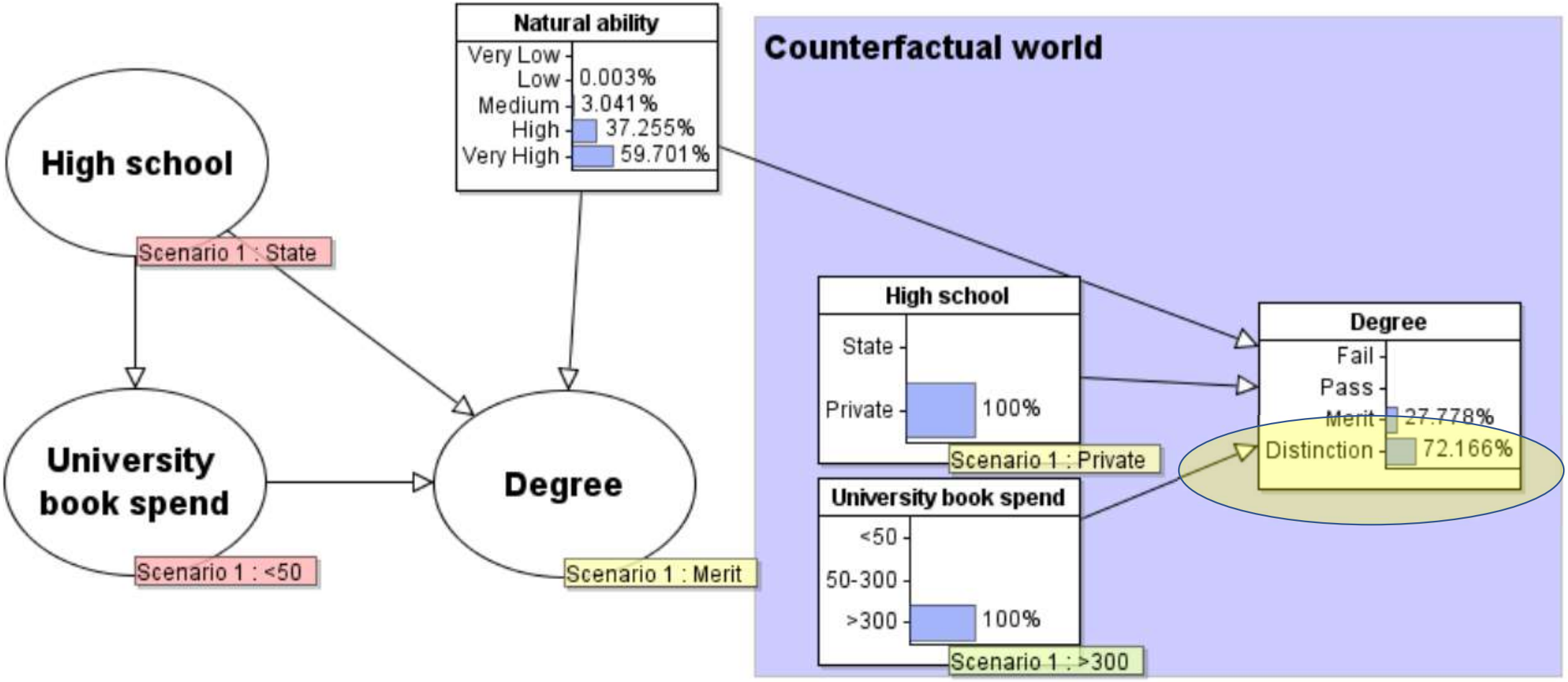
In this case we seek to answer counterfactuals involving change in both University Book Spend and type of High School.

Degrees: counterfactual (2 interventions)



Student achieves a Merit despite spending less than £50 and known to come from State school.
Leads to strong probability student had high natural ability

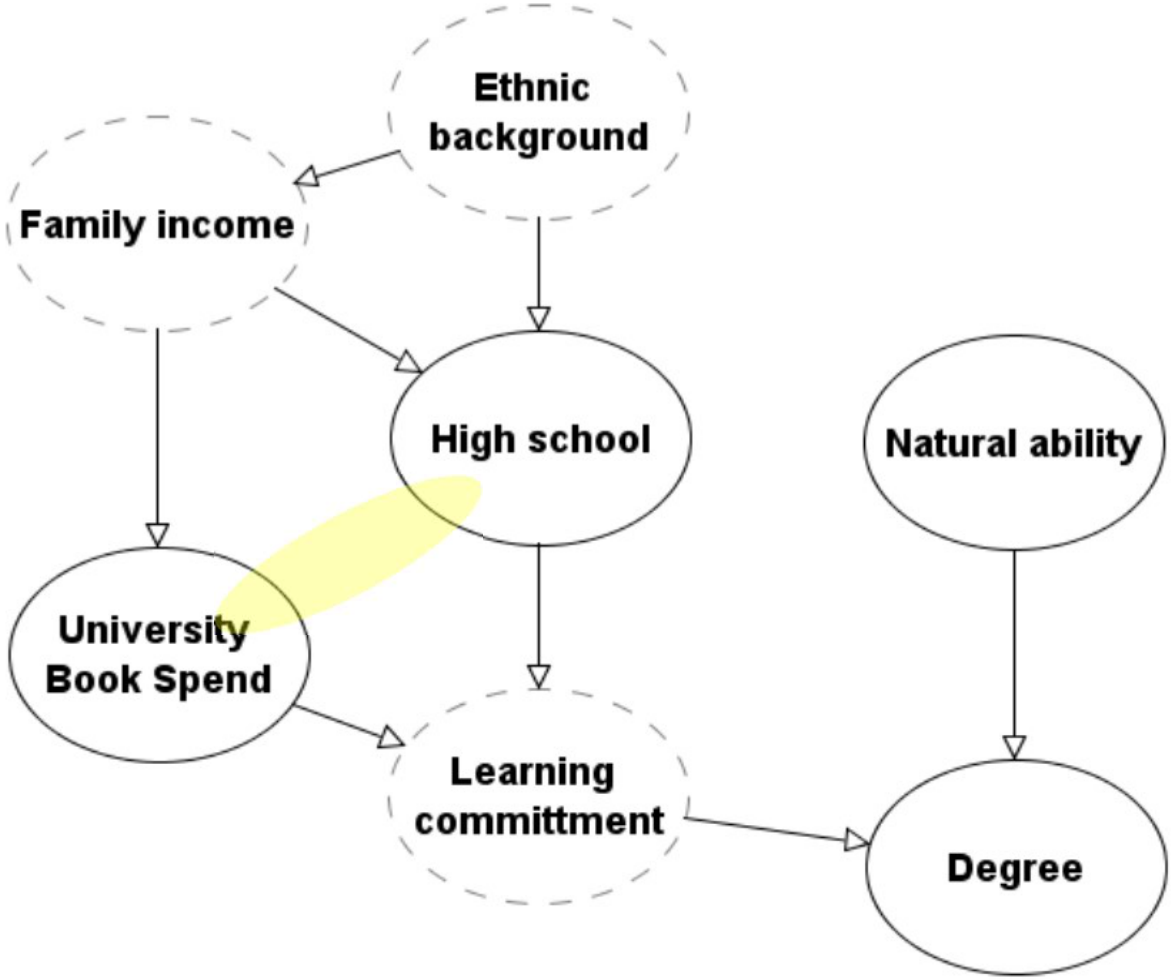
Degrees: counterfactual (2 interventions)



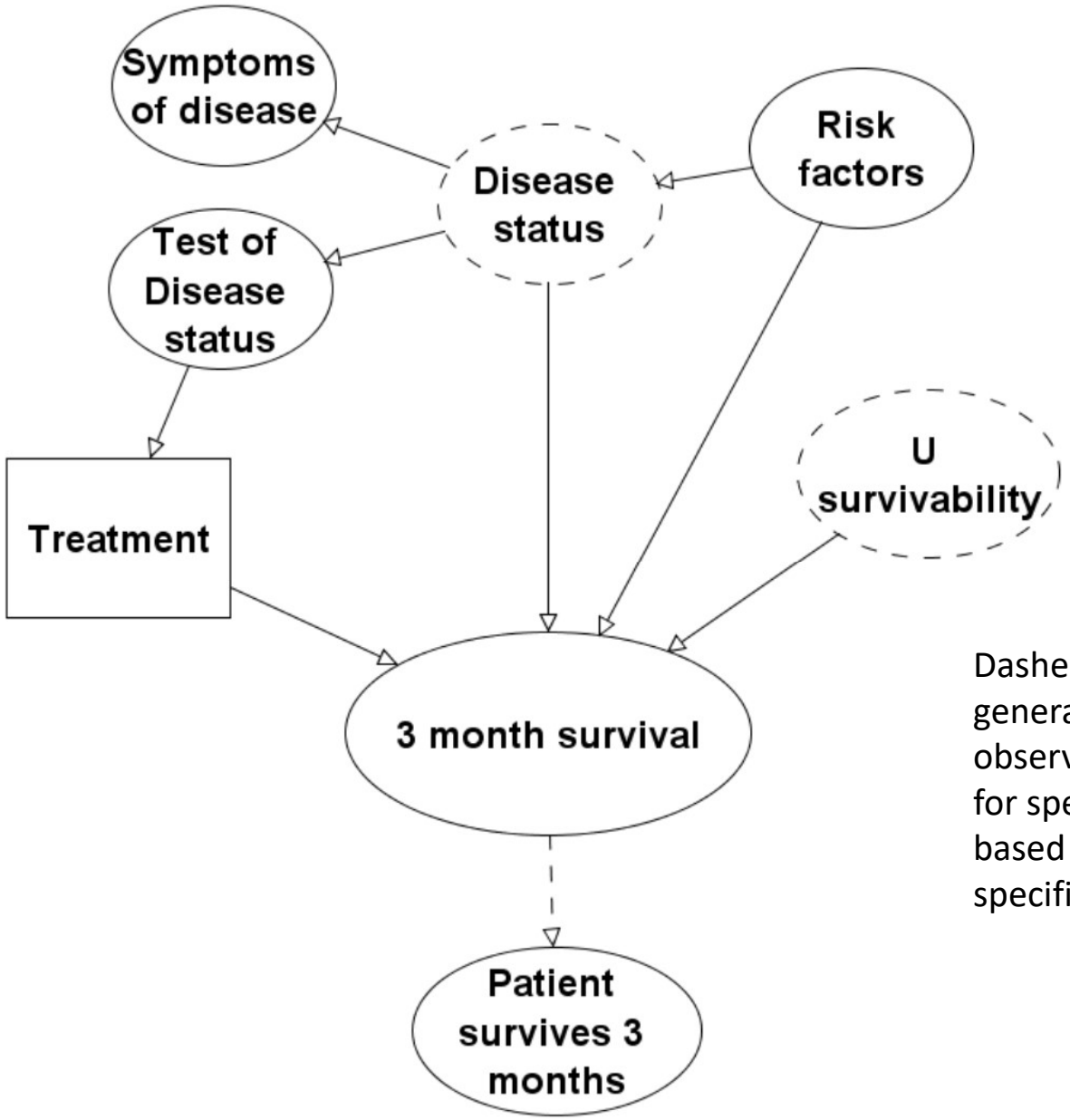
In the counterfactual world we assume the student spends >£300 on books and went to Private School. There is a 72% probability this student would have got a Distinction

More complete causal model of Degree outcome

Note the original link from High School to Book spend is due to the common causal factor ‘family income’.

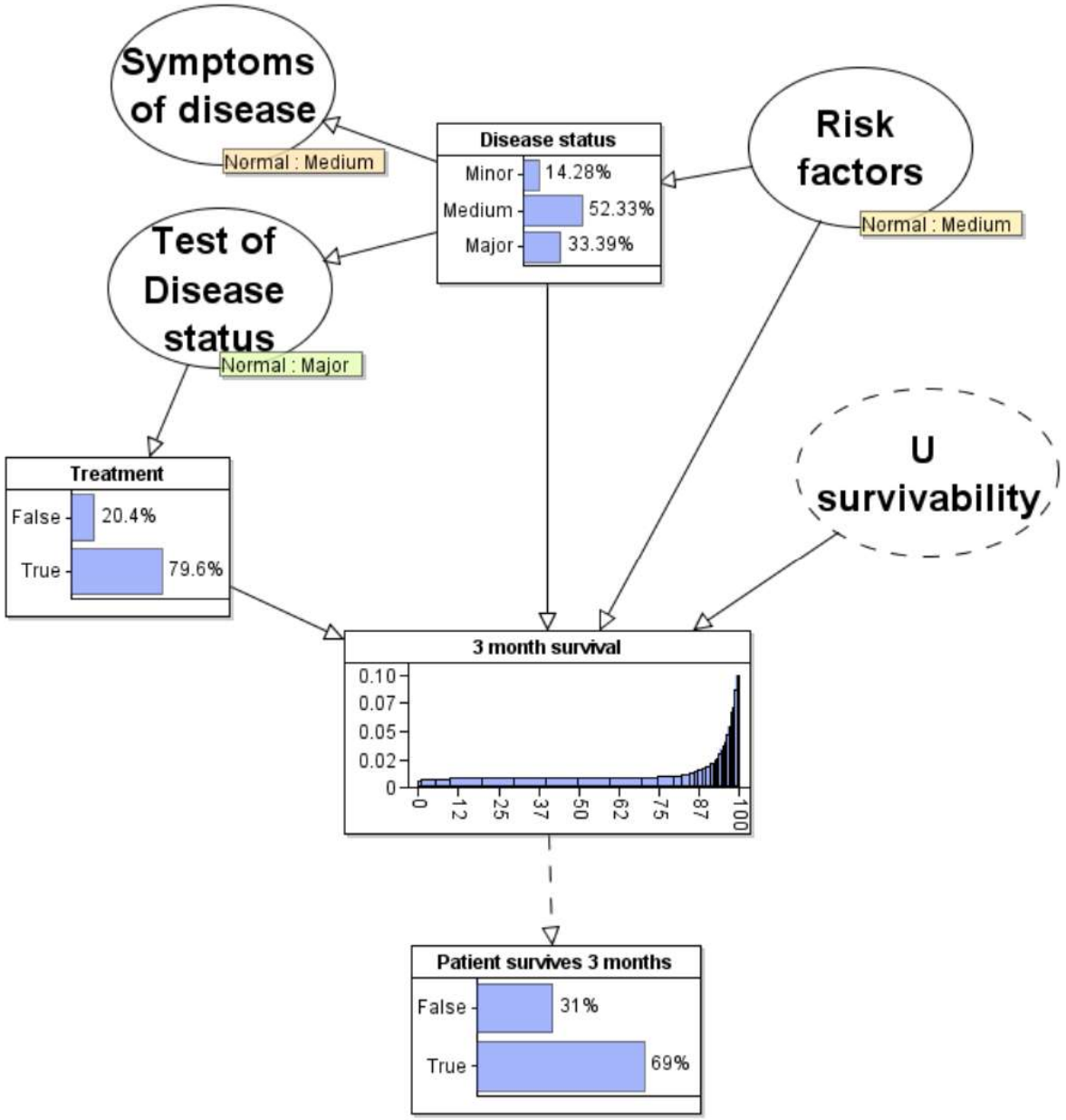


General structure of more advanced model

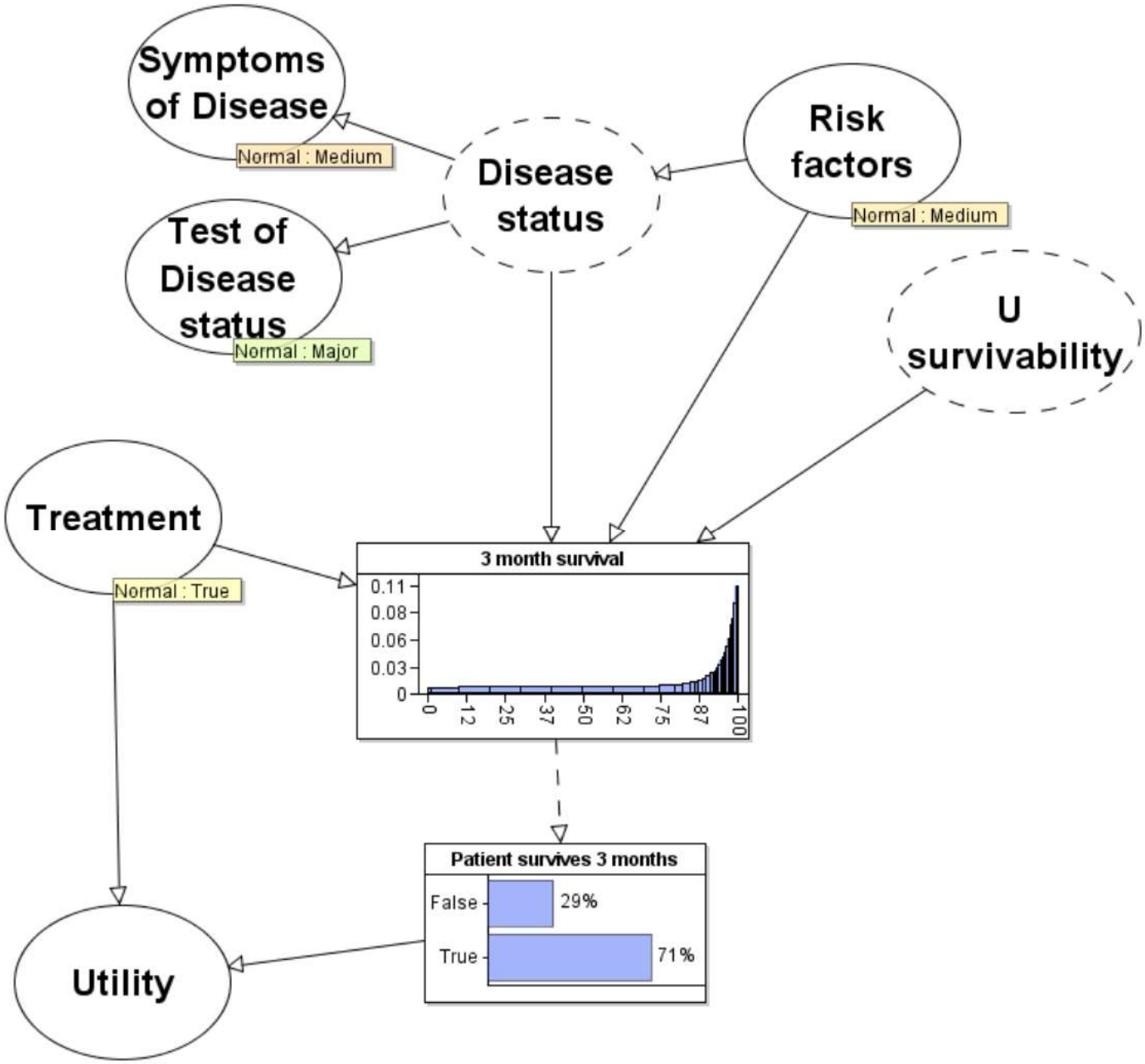


Dashed nodes are generally not directly observable but 'learnt' for specific patients based on patients specific observations

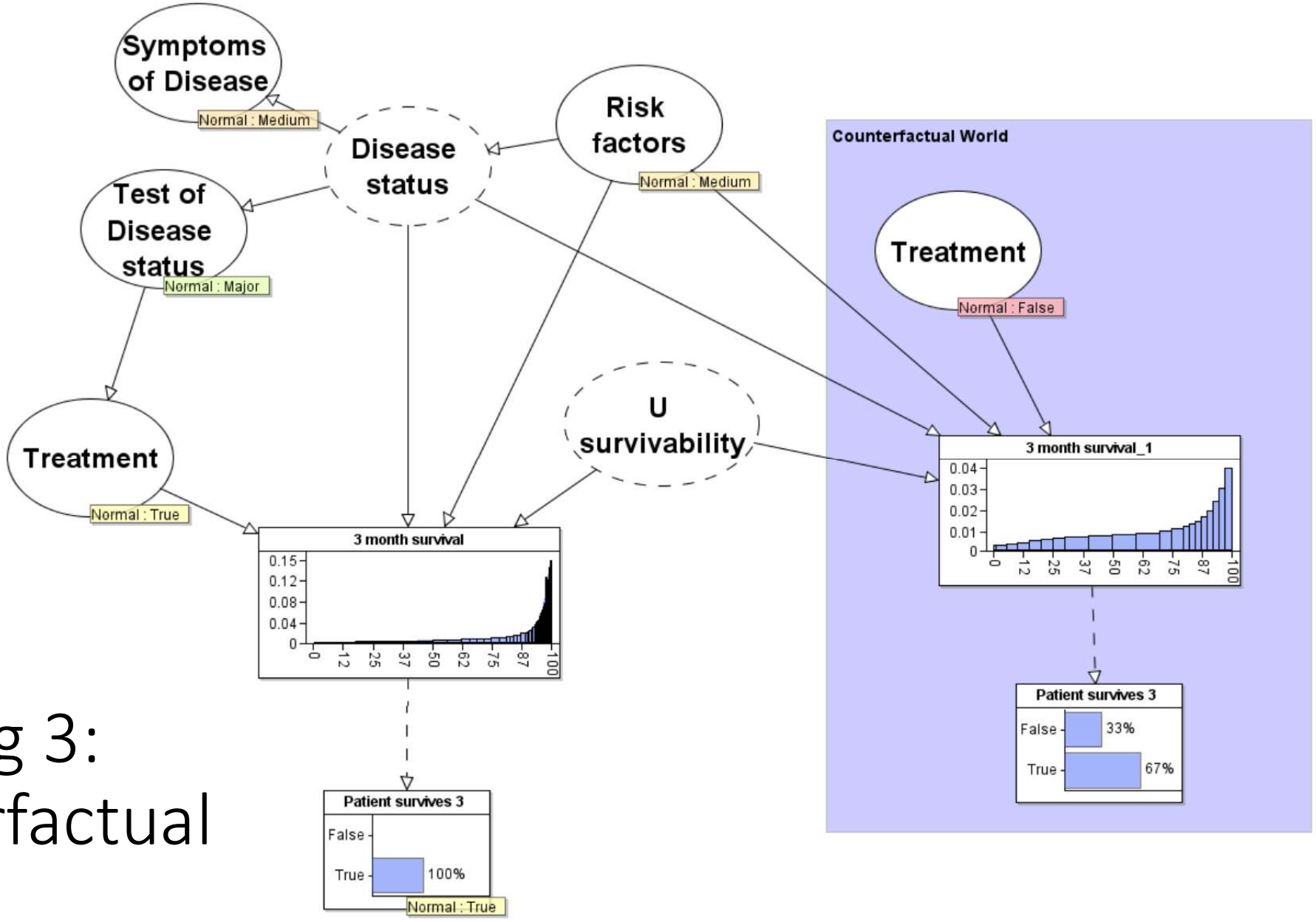
Rung 1: Observational



Rung 2: Intervention



Rung 3: Counterfactual



Counterfactuals and ‘fairness’/‘bias’

We are interested in the processes for making decisions about people. For example:

- Should a prisoner be released on bail
- Should an applicant be accepted for a particular job
- Should an applicant be accepted on an MSc programme
- What grade to award a student on an MSc

If the process is based purely on data, requiring no human intervention, we say the process is a decision-making *algorithm*

How to define when a decision-making process or algorithm is fair?

Counterfactuals and ‘fairness’/‘bias’

Let X be the set of attributes associated with people that can impact the outcome of the decision-making process.

Let A be any member of X that is agreed to be a ‘protected attribute’.

Suppose that, for a particular person p who has non-protected attribute values $X1_p$, $X2_p$, .., Xn_p and value A_p for protected attribute, the outcome of the process is O_p.

If, in the counterfactual world where we change only the value of A_p, the outcome *is different to* O_p then the process is defined to be **biased with respect to the protected attribute A.**

Informally (but less accurately) the process must produce the exact same outcome for any two people if they have the same values for all non-protected attributes but differ for some protected attribute.

*An algorithm that is not biased is defined as ‘**fair**’.*

Example: Job application

X1= Degree class, X2=Years of relevant experience, X3=University attended

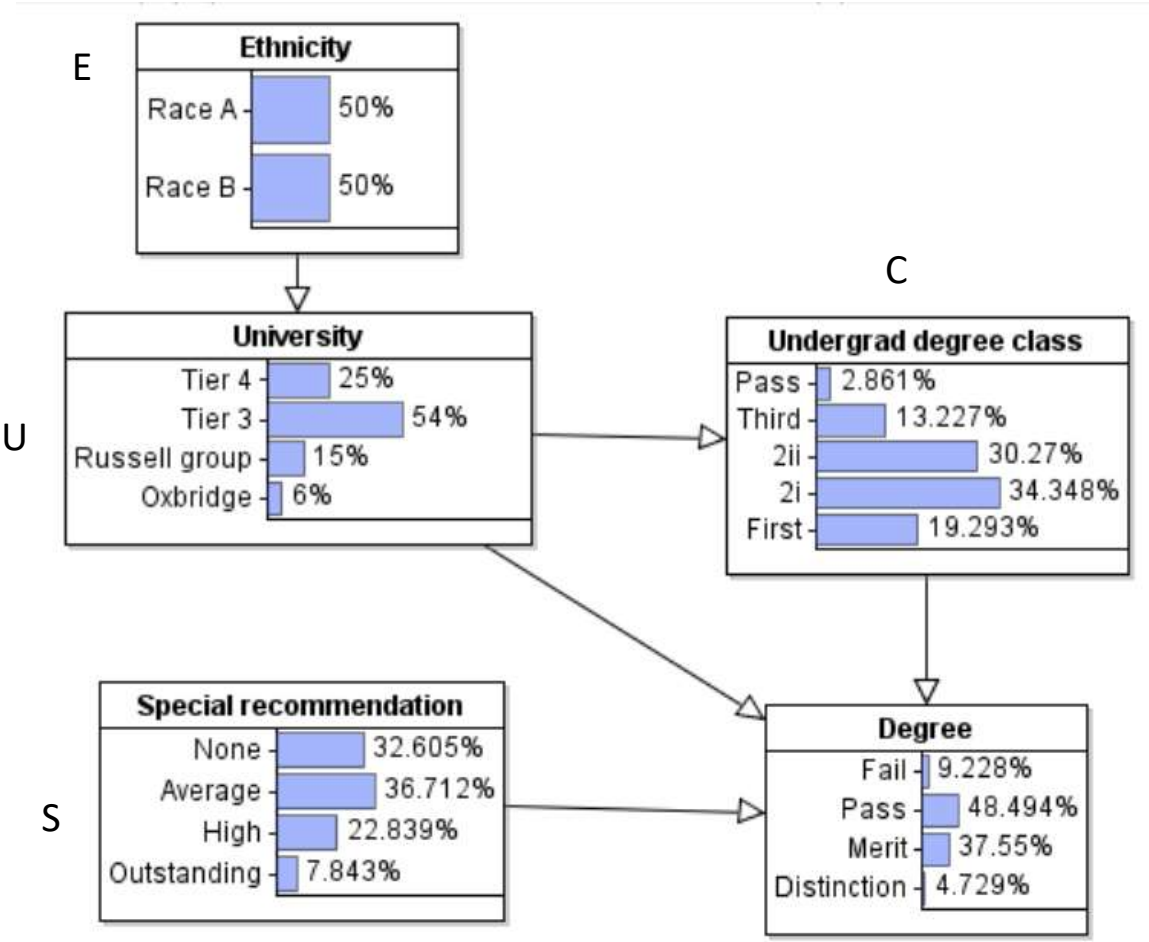
A = Sex (M/F) – other typical examples ‘ethnicity’, ‘disability’

For Fred $X1_{fred}$=“2i”; $X2_{fred}$=“8”; $X3_{fred}$=“Oxford”; A_{fred}=“M”.
Outcome O_{fred}=“Accept”

In counterfactual world
$X1_{fred}$=“2i”; $X2_{fred}$=“8”; $X3_{fred}$=“Oxford”; A_{fred}=“F”.
Outcome O_{fred}=“Reject”

For Jane: $X1_{jane}$=“2i”; $X2_{jane}$=“8”;
$X3_{jane}$=“Oxford”; A_{jane}=“F”.
Outcome O_{jane}=“Reject”

Example: Model for predicting MSc degree outcome



Suppose a university uses this model from past observational data of MSc students to decide whether to accept students on to its MSc.
e.g. based on knowing one or more of the attributes E, U, C, S they run the model and if Degree failure probability is less than 10% they accept student

Note: There is no direct link from E into degree otherwise, since E is a protected attribute. it would certainly result in bias

Case 1: As long as the answer to U is **required** for the decision, the algorithm is **NOT biased** with respect to E because once U is known it blocks any information from E influencing any of the other variables including Degree

Case 2: Suppose only C and S are required for the decision. Then the algorithm **IS biased** with respect to E

But beware other definitions of algorithm bias'

The most problematic (and common) definitions based on the different 'error rates' for protected attributes

e.g in our MSc admissions algorithm even if we **require** attributes U, C, S to be provided (so the algorithm **cannot be biased with respect to ethnicity**) the following are both true:

The probability the algorithm wrongly rejects a Race B student is greater than the probability the algorithm wrongly rejects a Race A student

and

The probability the algorithm wrongly accepts a Race A student is greater than the probability the algorithm wrongly rejects a Race B student

So the algorithm 'errs in favour' of Race A both with respect to false negatives and false positives.

Sounds bad. But it is a statistical inevitability for any (protected) attribute that correlates with any non-protected attribute used in the algorithm.

E.g. any algorithm that includes an attribute like 'income', or 'profession' will inevitably be 'biased' in the above sense against people of a particular ethnicity and sex simply because those correlate with 'income', or 'profession'.

See: "Bias in AI Algorithms" https://probabilityandlaw.blogspot.com/2018/06/bias-in-ai-algorithms.html
and the brief article: "Criminally Incompetent Academic Misinterpretation of Criminal Data" http://doi.org/10.13140/RG.2.2.32052.55680

Summary

- Causal models enable us to model observational data in such a way that we can both make accurate ‘observational’ predictions from them and explain counterintuitive observations.
- Causal models enable us to model interventions from observational data alone and so reach rung 2 of Pearl’s ladder. In a BN model this involves ‘breaking all links’ into the intervention variables.
- Causal models enable us to model counterfactuals from observational data alone and so reach rung 3 of Pearl’s ladder. In a BN model this involves creating a ‘twin counterfactual model’
- Models learnt purely from data cannot achieve any of the above
- To properly define algorithm bias and fairness we have to use the notion of counterfactuals
- Beware, however, that there are very commonly used definitions of algorithm bias that provably ensure every algorithm is classified as biased