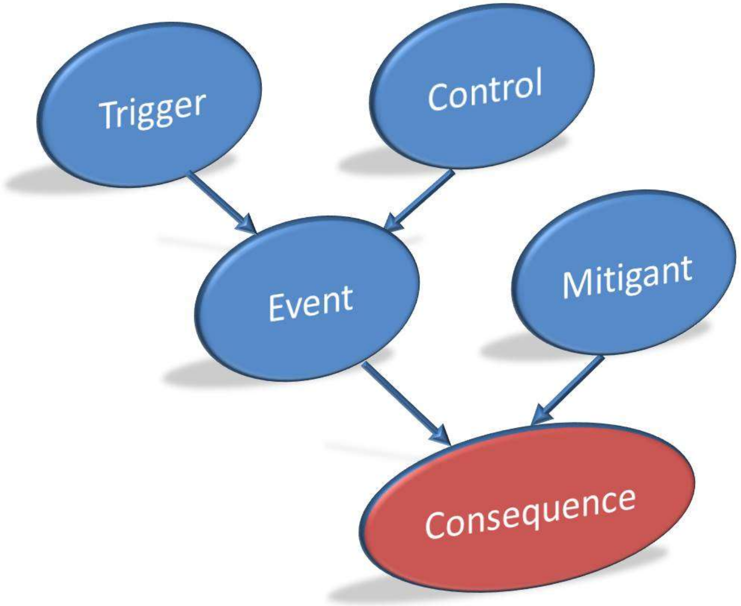
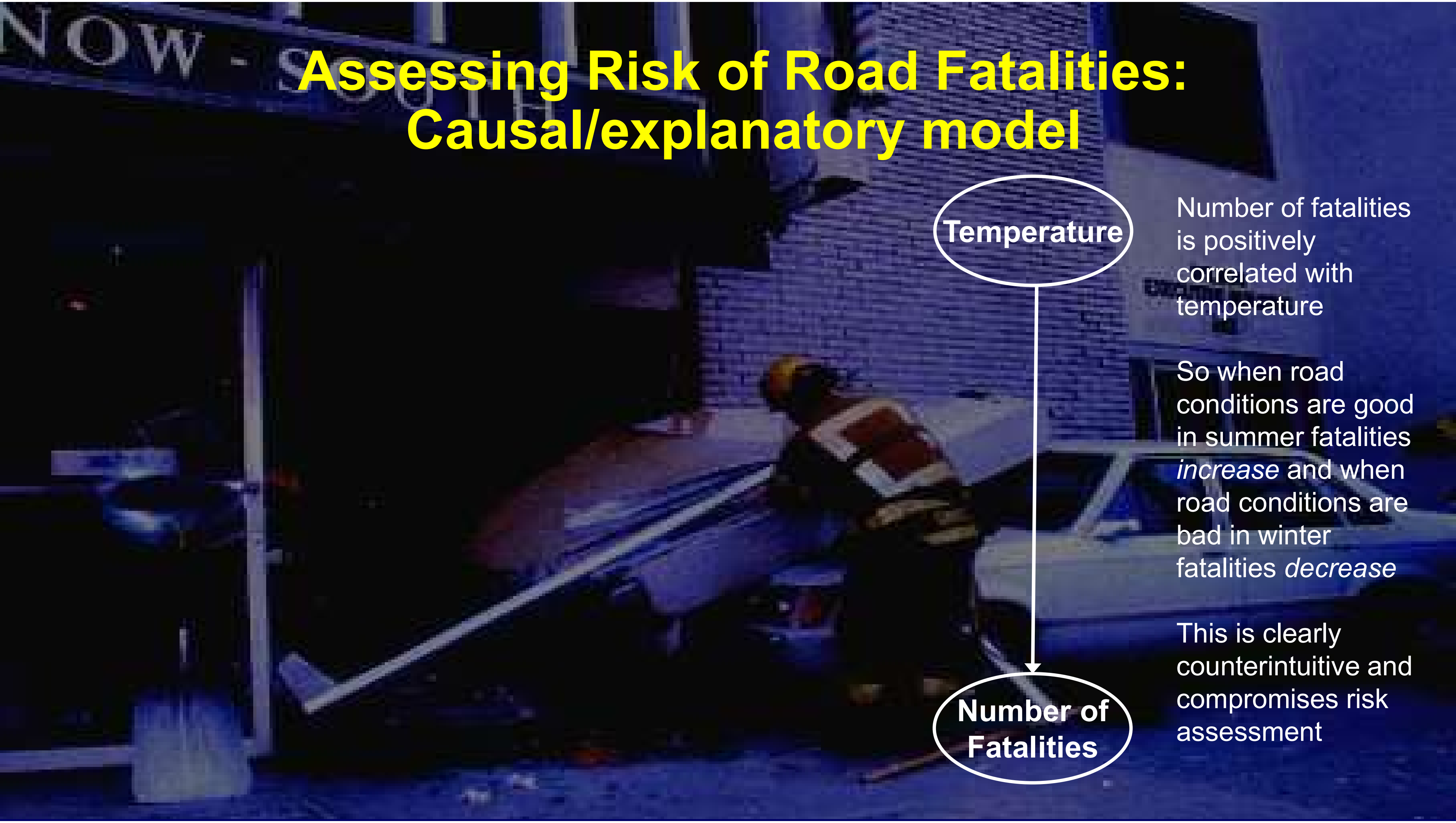


Risk and Decision Making for Data Science and AI

Lesson 6 Understanding risk data through causal paradoxes, explanations and models

Norman Fenton
@ProfNFenton

Assessing Risk of Road Fatalities: Causal/explanatory model



Temperature

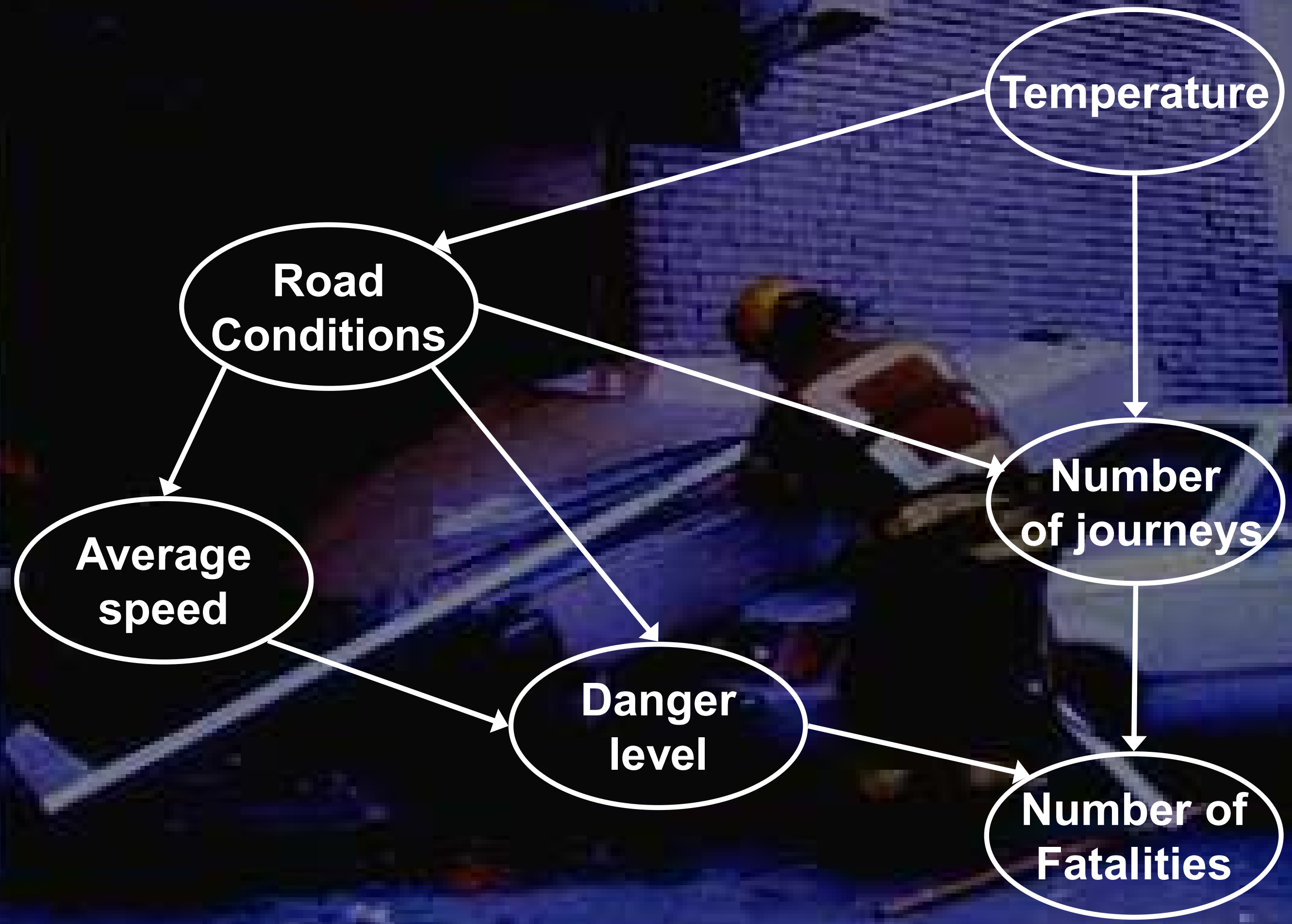
Number of Fatalities

Number of fatalities is positively correlated with temperature

So when road conditions are good in summer fatalities *increase* and when road conditions are bad in winter fatalities *decrease*

This is clearly counterintuitive and compromises risk assessment

Assessing Risk of Road Fatalities: Causal/explanatory model



True or False?

Babies born underweight are less likely to survive if the mother smoked during pregnancy

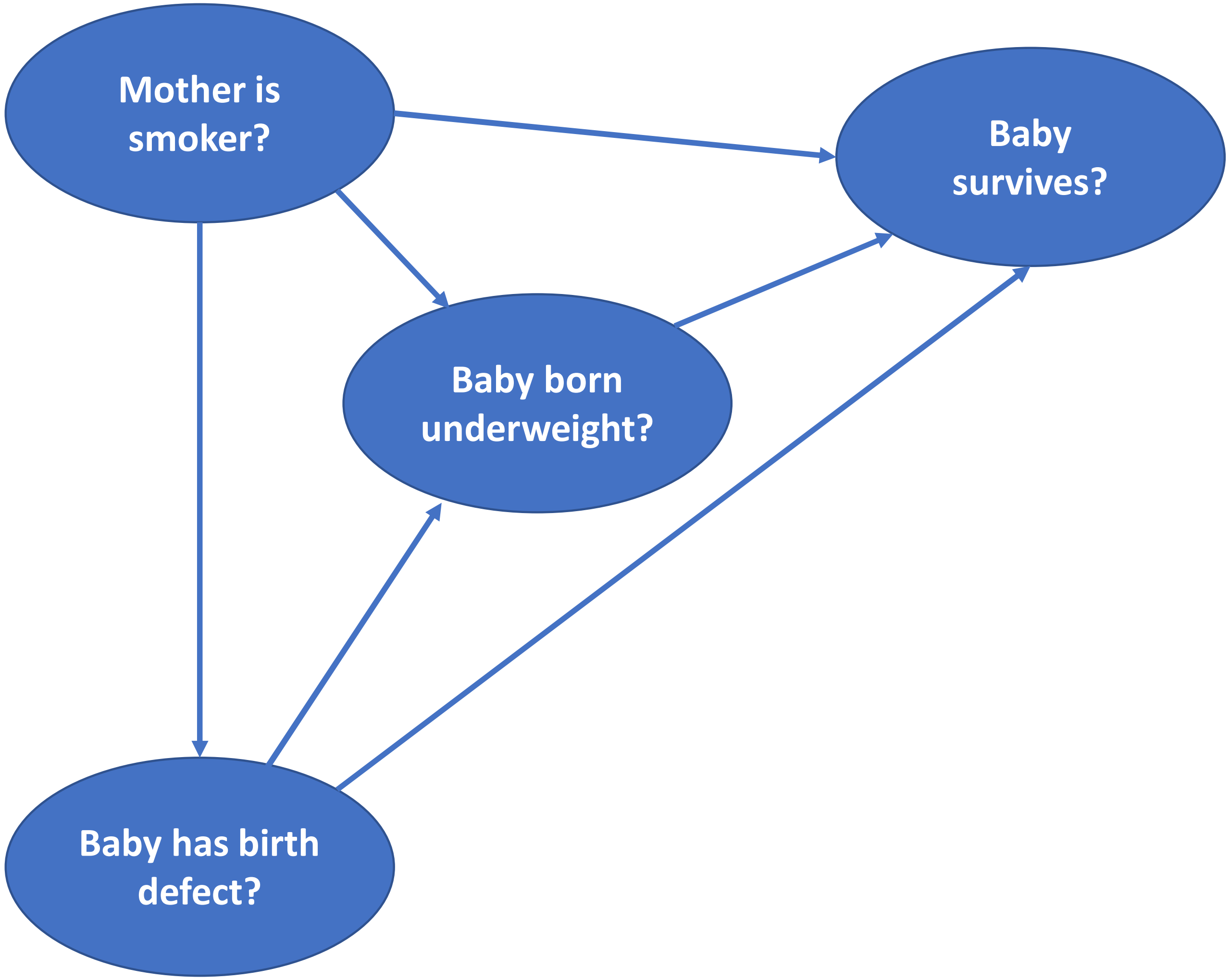
The smoking ‘birth-weight’ paradox

Babies born underweight are ***more*** likely to survive if the mother smoked during pregnancy

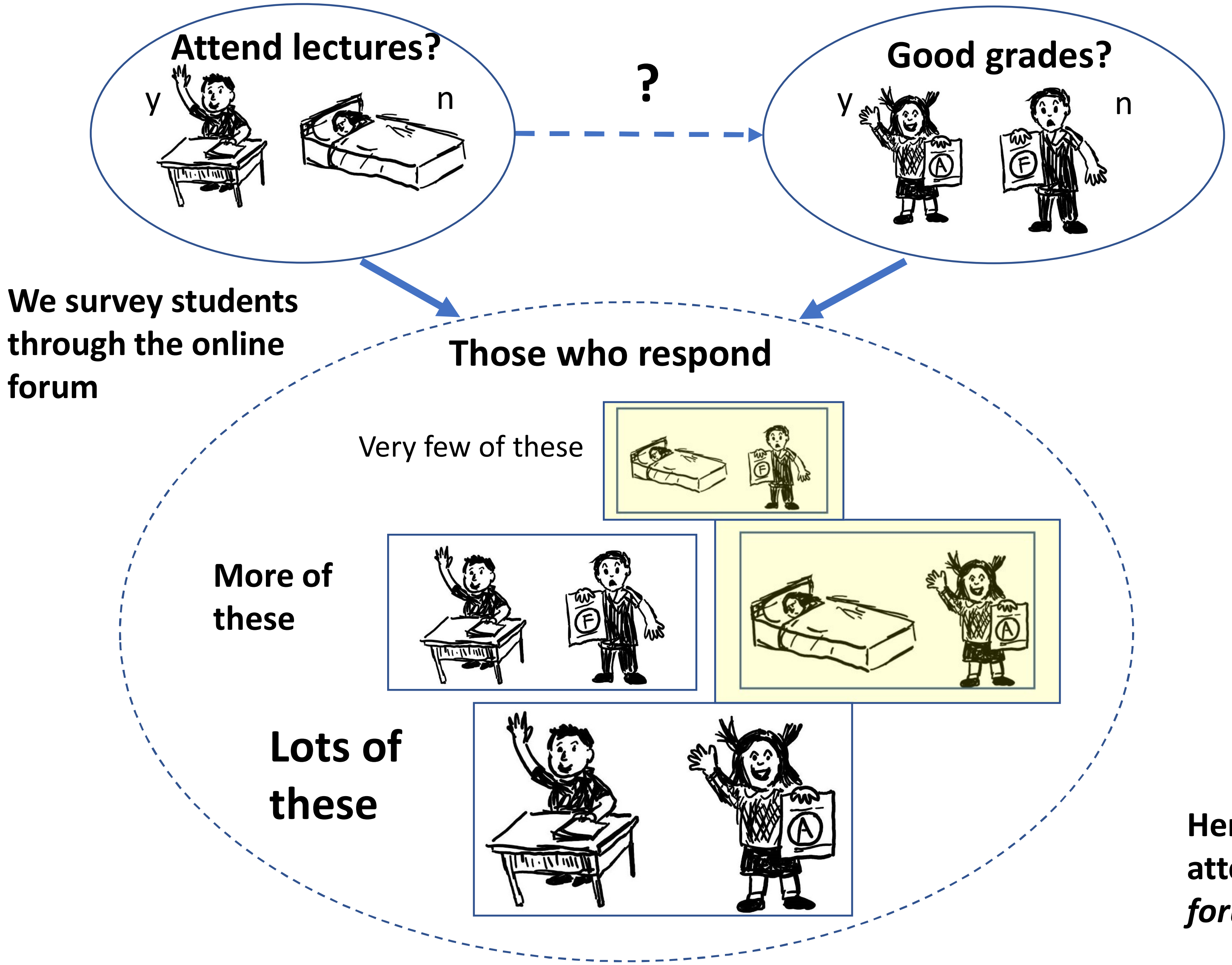
1959 public health study of pre- and post-natal data on over 15,000 births in San Francisco

Does this mean that mothers who are likely to give birth to underweight babies should take up smoking?

This clearly sounds wrong – but the ‘Paradox’ was not resolved until 2006


Mother is smoker?
Baby survives?
Baby born underweight?
Baby has birth defect?

Students who skip lectures get better grades?



In reality students who skip lectures are LESS likely to get good grades…but we want to show that by an empirical study

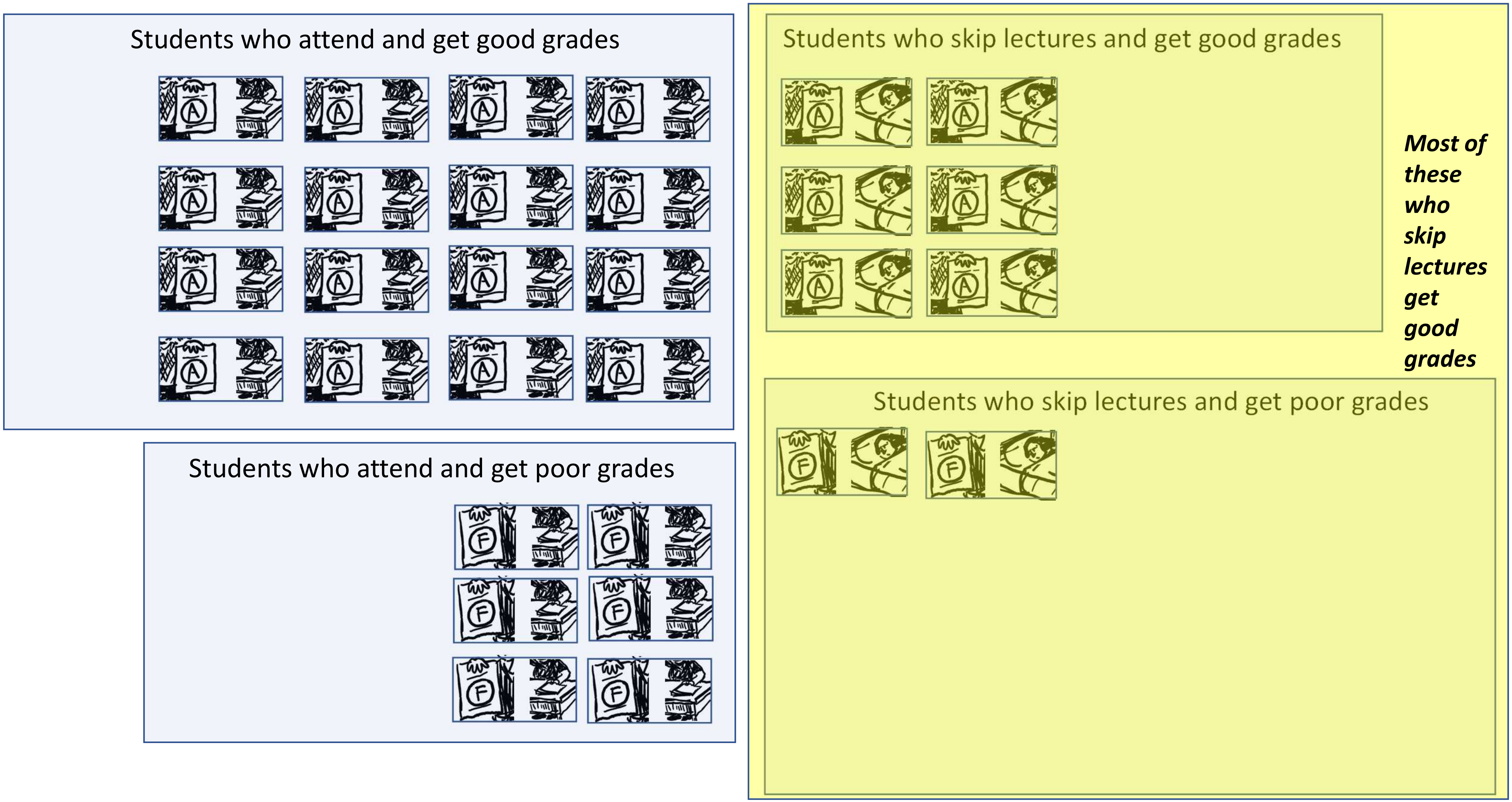
Those who are unengaged (those who do not attend lectures *and* get poor grades) are unlikely to respond to the forum

So unengaged students are mainly 'censored out' of the data

Hence, most of the students who do not attend lectures – *but respond to the forum* - get good grades

The FULL set of students:

Students who complete the forum survey



Assuming we can estimate the probability of a student responding we can use the (biased) sample data to determine the 'true' population

Let *R* = 'responds', *A* = 'attends lectures' and *G* = good grade

We estimate:
Prob (*R* | *A, G*) = 0.9
Prob (*R* | not *A*, *G*) = 0.5
Prob (*R* | *A*, not *G*) = 0.5
Prob (*R* | not *A*, not *G*) = 0.1

Suppose the responses we get are:
Number of (*A*, *G*) = 270
Number of (not *A*, *G*) = 114
Number of (*A*, not *G*) = 101
Number of (not *A*, not *G*) = 27

Then the 'true' totals are
Number of (*A*, *G*) = 300
Number of (not *A*, *G*) = 228
Number of (*A*, not *G*) = 202
Number of (not *A*, not *G*) = 270

As the 'true' total sums to 1000, we conclude
Prob(*A* and *G*) = 0.3
Prob(not *A* and *G*) = 0.228
Prob(*A* and not *G*) = 0.202
Prob (not *A* and not *G*) = 0.27

And from these we also get
Prob(*A*) = Prob(*A* and *G*) + Prob(*A* and not *G*) = 0.502
Prob(not *A*) = 0.498

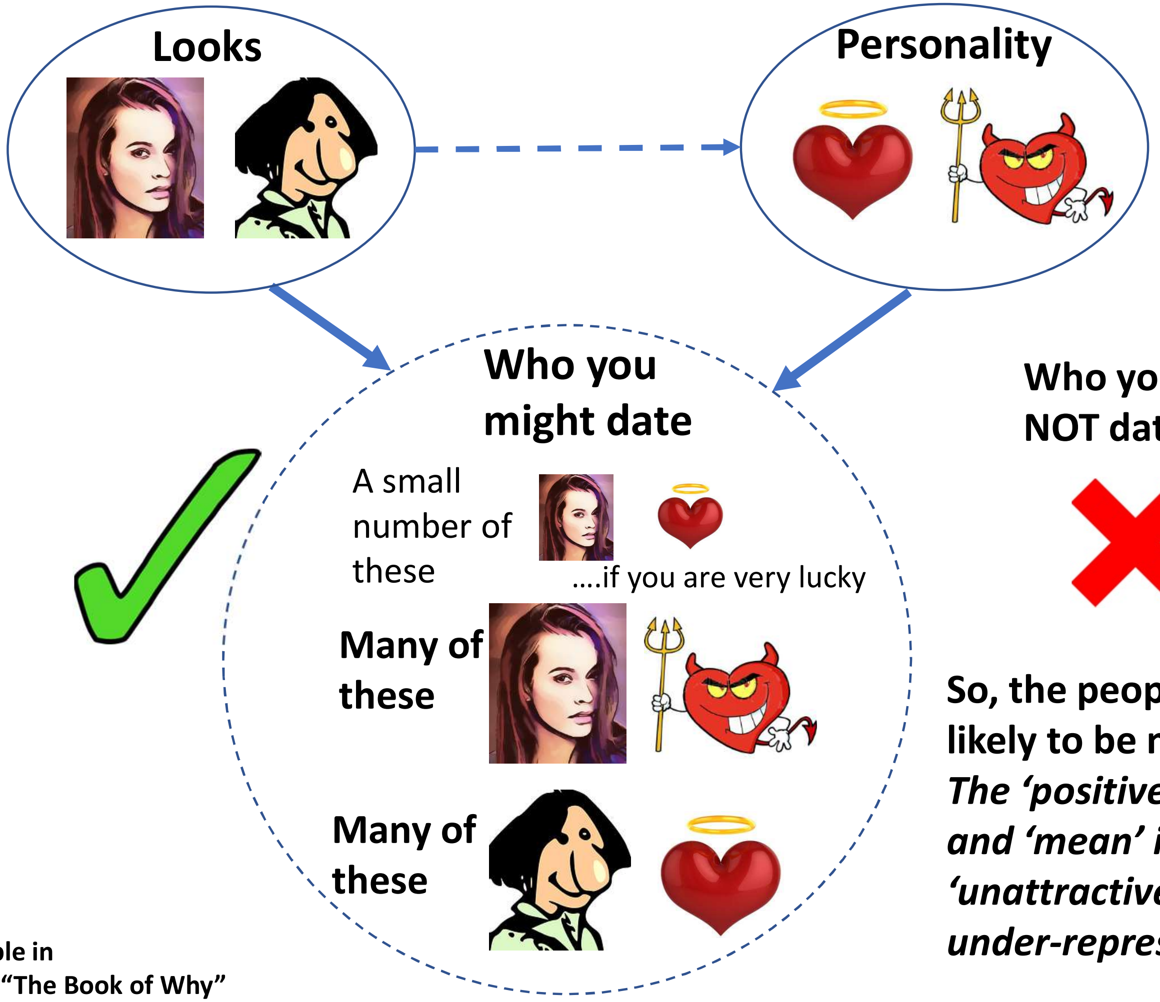
From this we can get the 'true' conditional probabilities we want

$$\text{Prob}(G\ |A) = \frac{P(A \text{ and } G)}{P(A)} = \frac{0.3}{0.502} = 0.598$$

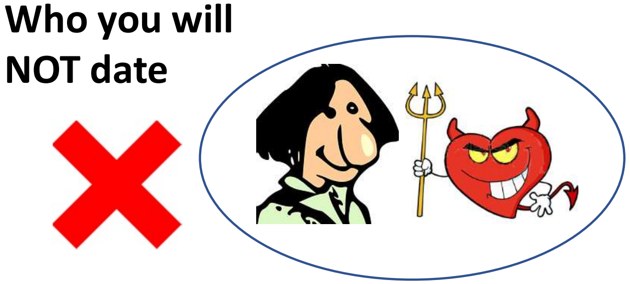
$$\text{Prob}(G\ |\text{not } A) = \frac{P(\text{not } A \text{ and } G)}{P(\text{not } A)} = \frac{0.228}{0.498} = 0.458$$

Are ‘attractive’ people more likely to be ‘mean’ than ‘unattractive’ people?

Why do we believe ‘attractive’ people are more likely to be ‘mean’ than ‘nice’ (Berkson’s paradox)?



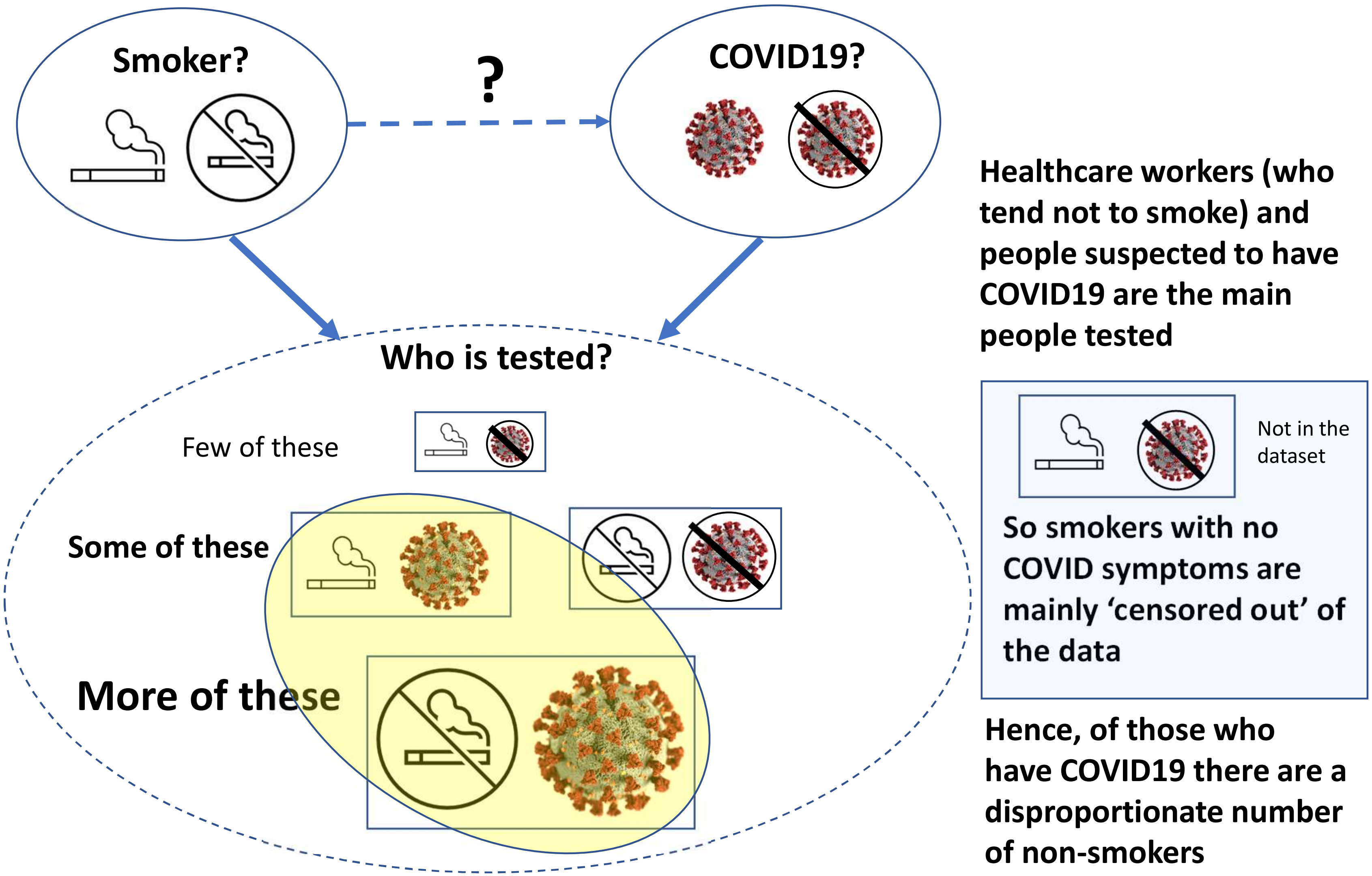
You will generally date somebody for their ‘looks’ or their ‘personality’



So, the people you date who are attractive are more likely to be mean than nice.
The ‘positive association’ between being ‘attractive’ and ‘mean’ is an illusion created by censoring out the ‘unattractive, mean people’, i.e. mean people are under-represented.

Based on an example in
Pearl & Mackenzie “The Book of Why”

Why non-smokers may seem to be more at risk of COVID19



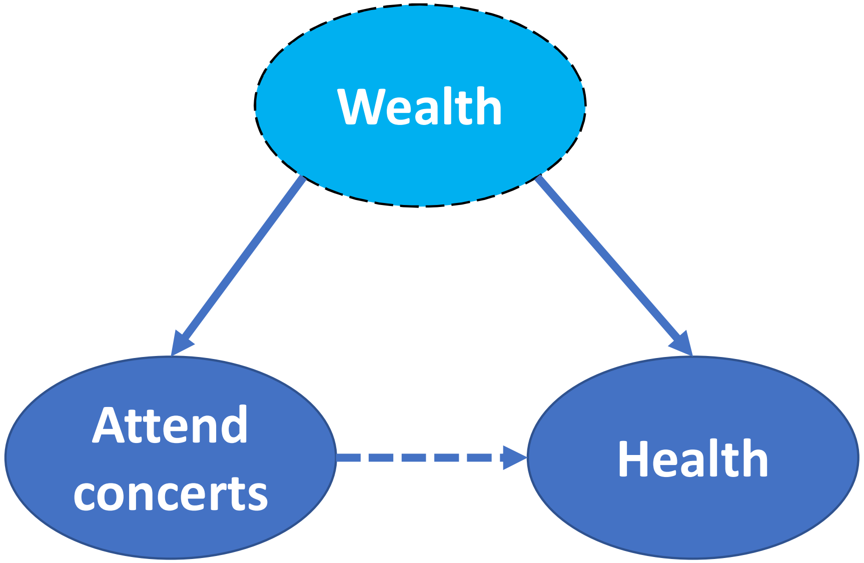
Berkson’s paradox (Collider bias)

Because we are relying on a dataset which over-represents some subjects and under-represents others, we conclude that the true causal relationship between two factors is the opposite of what it is in reality (in this case ‘skipping lectures results in better grades’)

The ‘paradox’ can also lead us to conclude that, for two completely unrelated factors (such as ‘personality’ and ‘looks’) one has a causal influence on the other (‘good looking people are more likely to have a bad personality’)

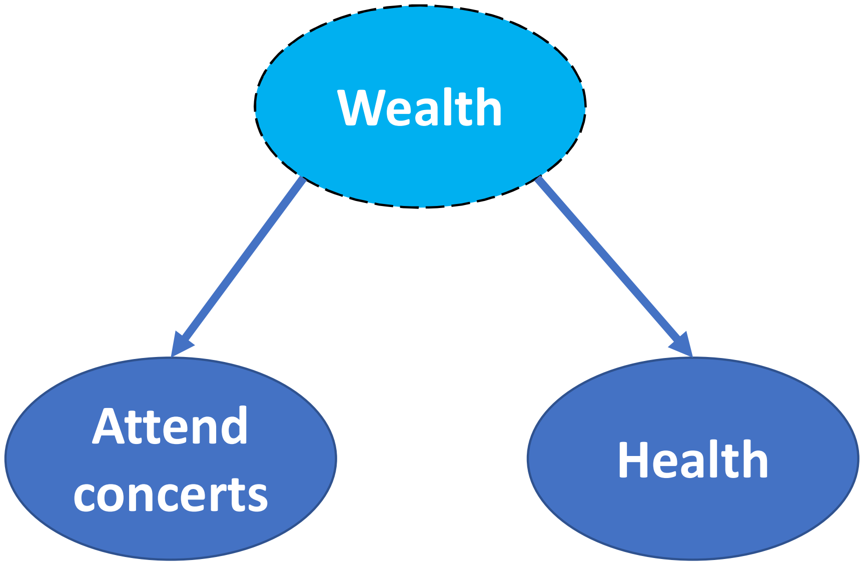
By using a causal model that explicitly identifies the collider ‘variable’ we can overcome the paradox without having to get any new data

Very different hidden causal explanations for relationships that seem causal but are not

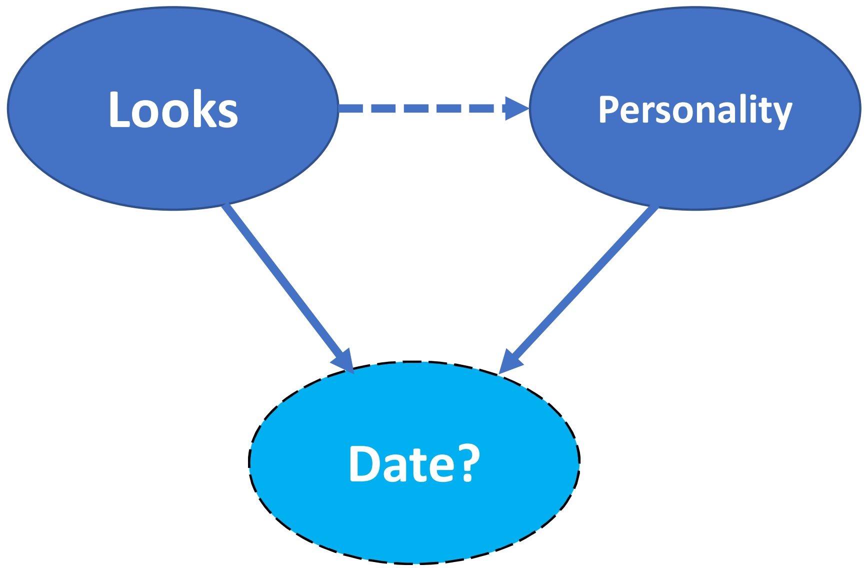


Vacuous 'causal' relationship is explained by common causal factor (wealth)

Very different hidden causal explanations for relationships that seem causal but are not

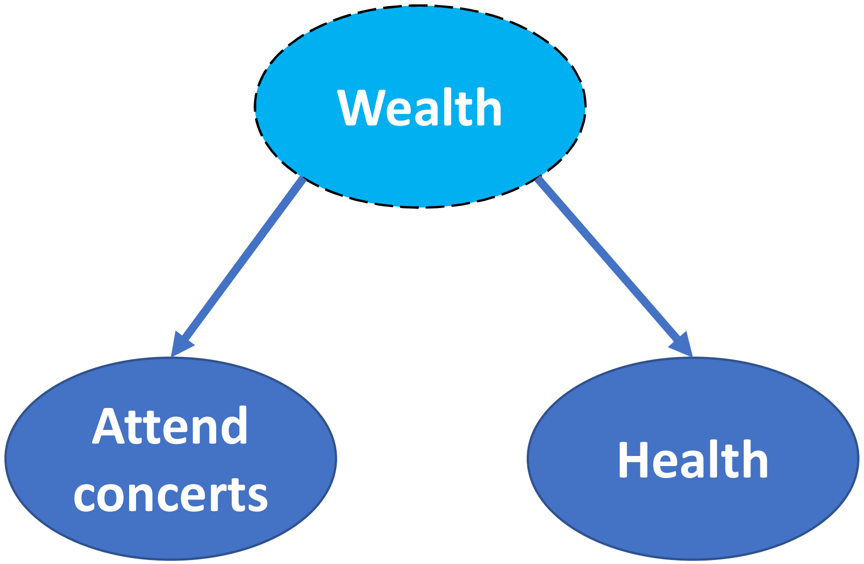


Vacuous 'causal' relationship is explained by common causal factor (wealth)

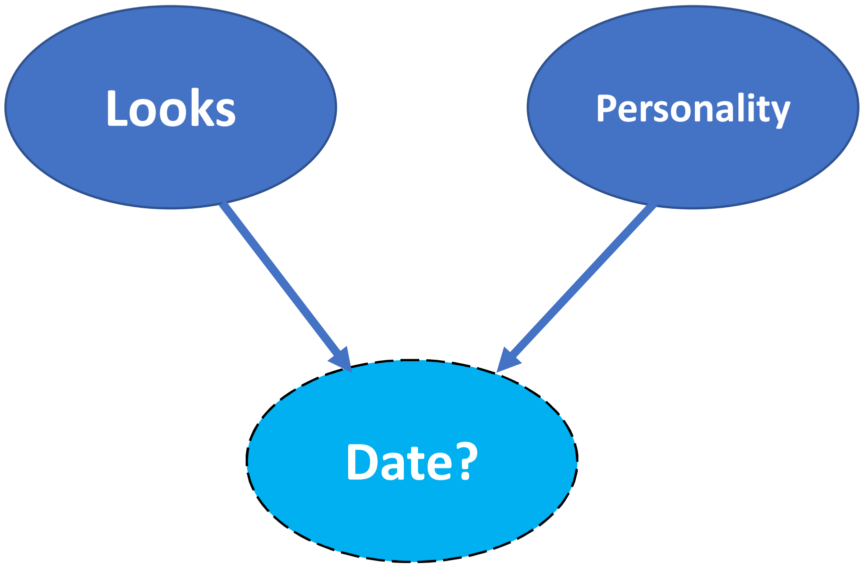


Vacuous 'causal' relationship is explained by a shared constraint (Date?) that censors the observed data

Very different hidden causal explanations for relationships that seem causal but are not



Vacuous 'causal' relationship is explained by common causal factor (wealth)

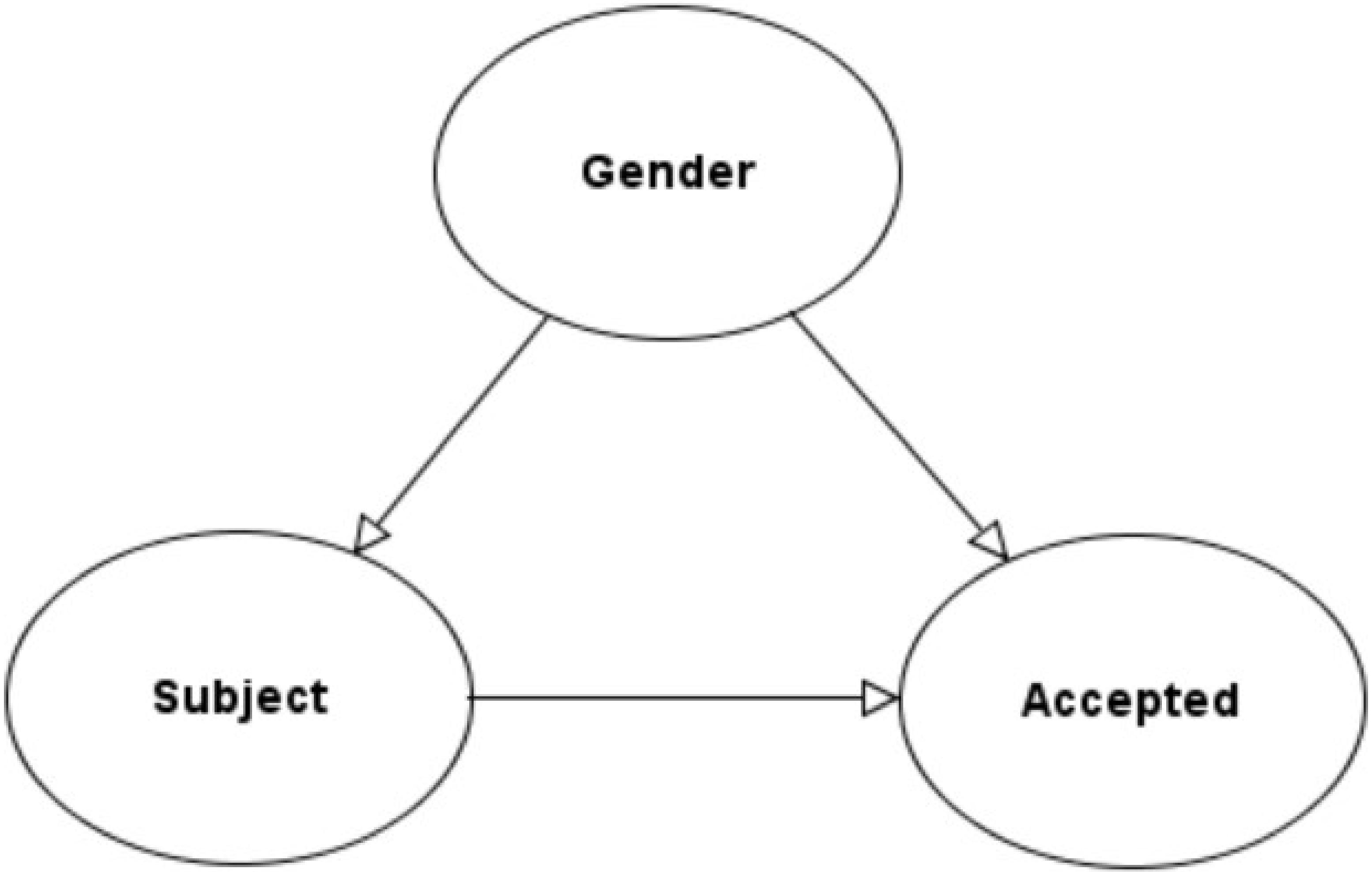


Vacuous 'causal' relationship is explained by a shared constraint (Date?) that censors the observed data

Simpson’s Paradox revisited

Cambridge Admissions Data

	WOMEN				MEN		
	Applied	Accepted	%		Applied	Accepted	%
Computer Science	26	7	27%		228	58	25%
Economics	240	63	26%		512	112	22%
Engineering	164	52	32%		972	252	26%
Medicine	416	99	24%		578	140	24%
Veterinary medicine	338	53	16%		180	22	12%
TOTAL	**1,184**	**274**	**23%**		**2,470**	**584**	**24%**



Gender
Subject
Accepted

Simpson's paradox

New drug is being tested on a group of 800 people (400 men and 400 women) with a particular disease. Half of the people (randomly selected) are given the drug and the other half are given a placebo (a *randomized controlled trial*)

Overall drug trial results

Drug taken	**No**	**Yes**
Recovered		
No	240	200
Yes	160	200
Recovery rate	***40%***	***50%***

So the drug is clearly effective

…isn't it?

Drug trial results with sex of patient included

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%***

In every subcategory the drug is worse than placebo…..HOW?

Unluckily, more men than women were given the drug

So can we fix the problem.....?

To ‘fix’ the previous problem in new study we make sure equal number of men and women chosen to receive drug:

- **200 men take drug; 200 men take placebo;**
- **200 women take drug; 200 women take placebo**

Sex	**Female**		**Male**	
Drug taken	No	Yes	No	Yes
Recovered				
No	108	92	84	56
Yes	92	108	116	144
Recovery rate	***46%***	***54%***	***58%***	***72%***

Drug taken	**No**	**Yes**
Recovered		
No	192	148
Yes	208	252
Recovery rate	***52%***	***63%***

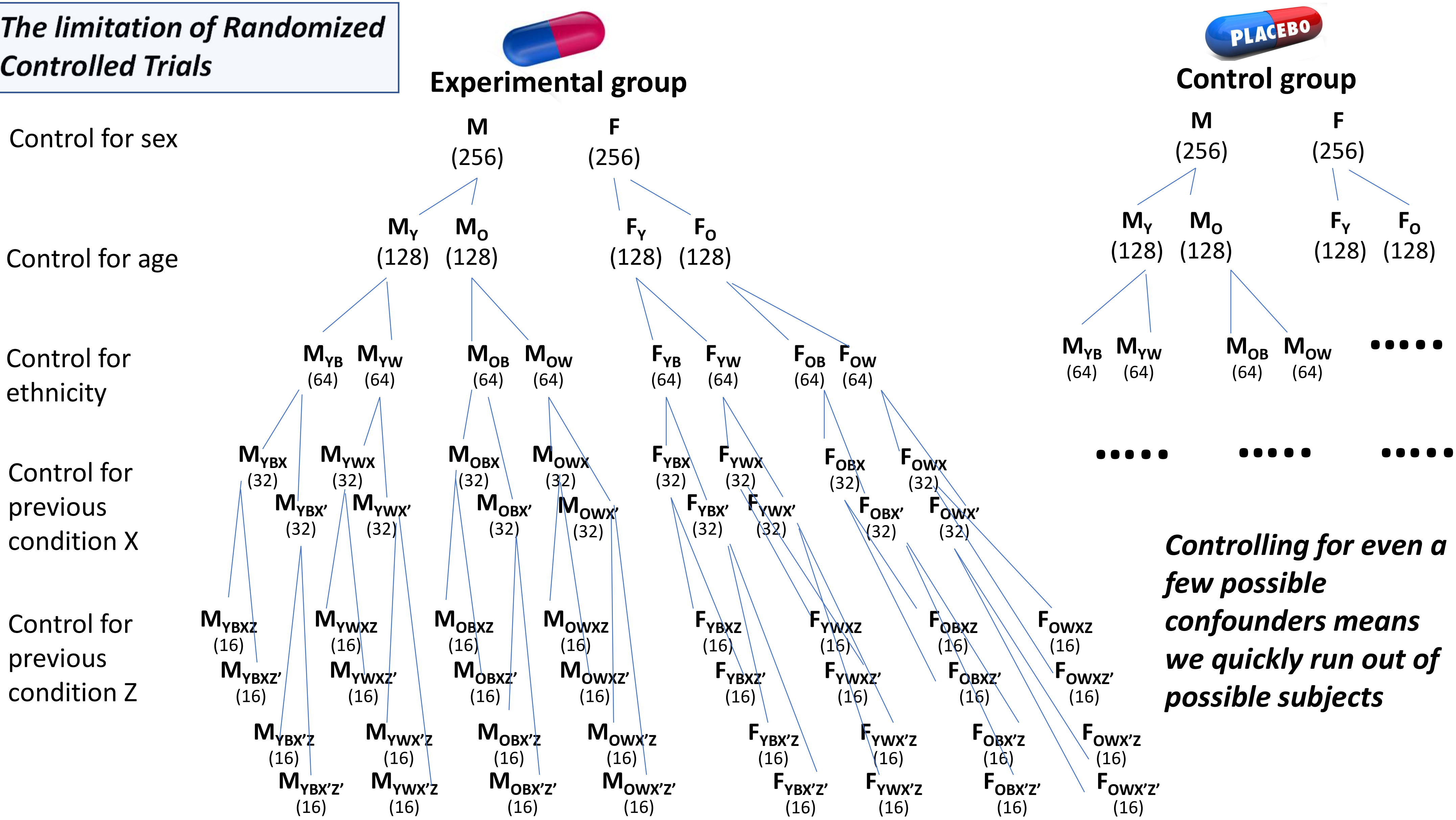
But when we ‘drill down’ further considering another factor *age* ....

Age	**40+**				**< 40**			
Sex	Female		Male		Female		Male	
Drug taken	No	Yes	No	Yes	No	Yes	No	Yes
Recovered								
No	96	28	80	24	12	64	4	32
Yes	64	12	80	16	28	96	36	128
Recovery rate	***40%***	***30%***	***50%***	***40%***	***70%***	***60%***	***90%***	***80%***

In every subcategory the drug is worse than placebo.....HOW?

Unluckily, more people aged <40 were given the drug than people aged 40+

The limitation of Randomized Controlled Trials



Controlling for even a few possible confounders means we quickly run out of possible subjects

Observational Studies

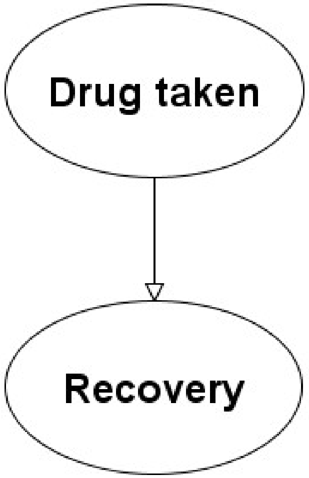
Collect whatever data are available on the topic of interests.

Without ‘controls’ any conclusions are considered dubious. They are even more susceptible (than RCTs) to:

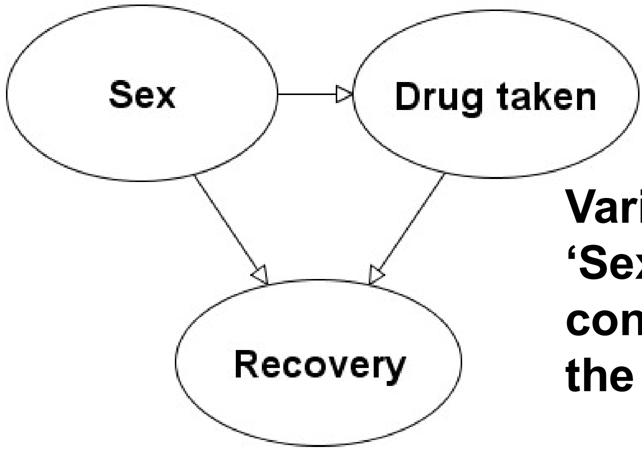
- **Simpson’s paradox**: e.g men and young people may be *systematically* more likely to take a particular drug/treatment than women and old people (so effect of confounder is systematic in the data)
- **Berkson’s paradox (collider bias /censored data)**: e.g. clinicians specialising in a particular type cancer collect extensive data on patients diagnosed with that type of cancer …. but have little or no similar data on patients not diagnosed with it

But by considering a causal model for the data, observational studies can be used to simulate RCTs and resolve the problems

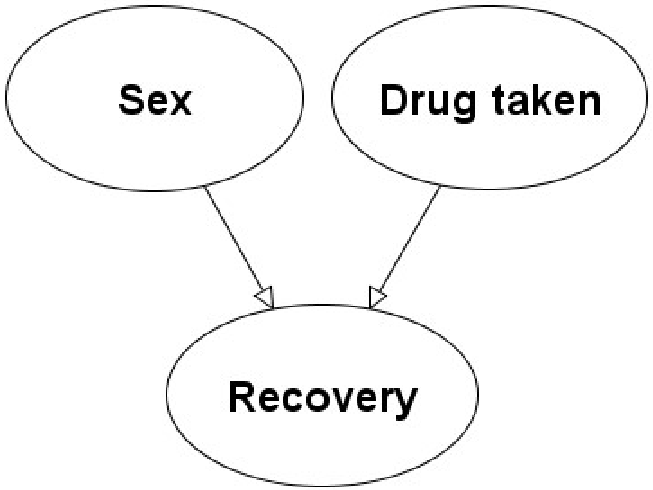
Causal explanation of Simpson



What we really want to know overall



Variable ‘Sex’ confounds the results



We need to break the dependency between ‘Sex’ and ‘Treatment’

A real medical study: treatment of kidney stones

Compare outcome (success rate) of two treatments A and B for patients with kidney stones
700 patients, half of whom had treatment A and half of whom had treatment B

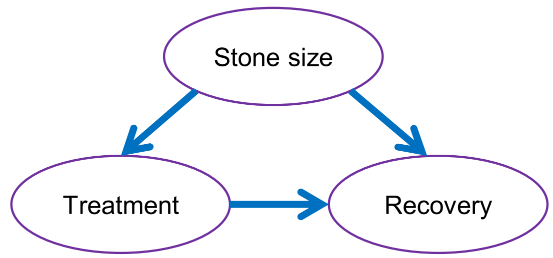
	Overall	Patients with small stones	Patients with large stones
Treatment **A**	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment **B**	**83%** (289/350)	87% (234/270)	69% (55/80)

Treatment B is **better overall** (recovery rate: 83%)

But **worse in every subcategory** when we split data by the size of kidney stones

For patients with small stones   A is better (93% compared to 87%)

For patients with large stones   A is better (73% compared to 69%)

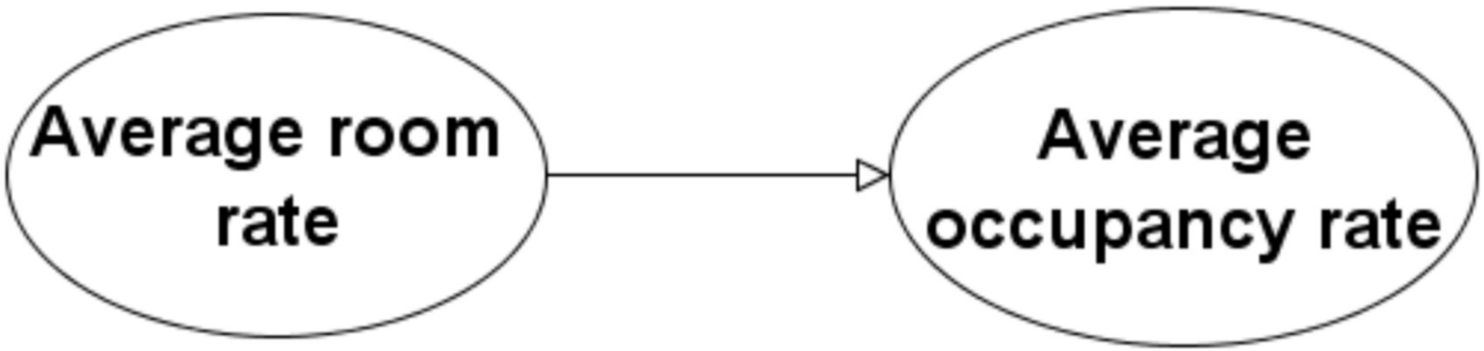


Patients with large stones were more likely to be given treatment A

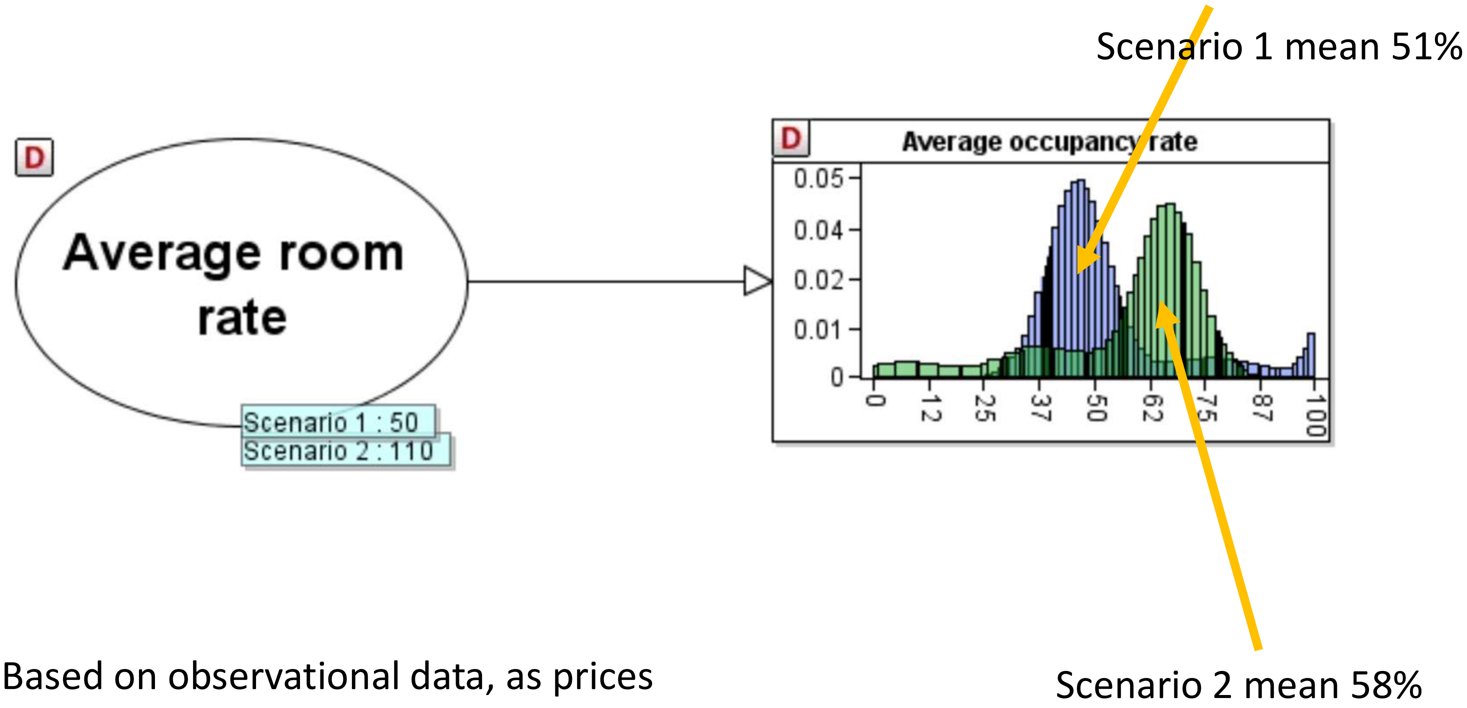
....inevitable kind of limitation of 'observational studies'

Hotel pricing

How does pricing impact on occupancy?

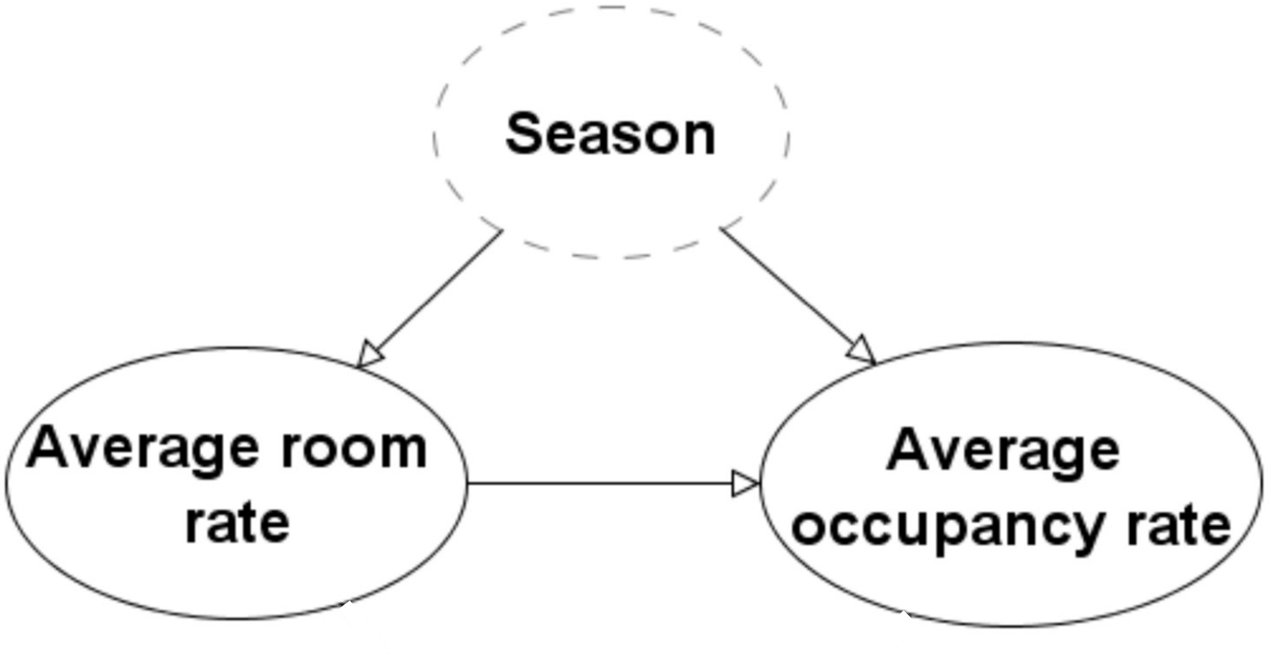


Hotel room rates

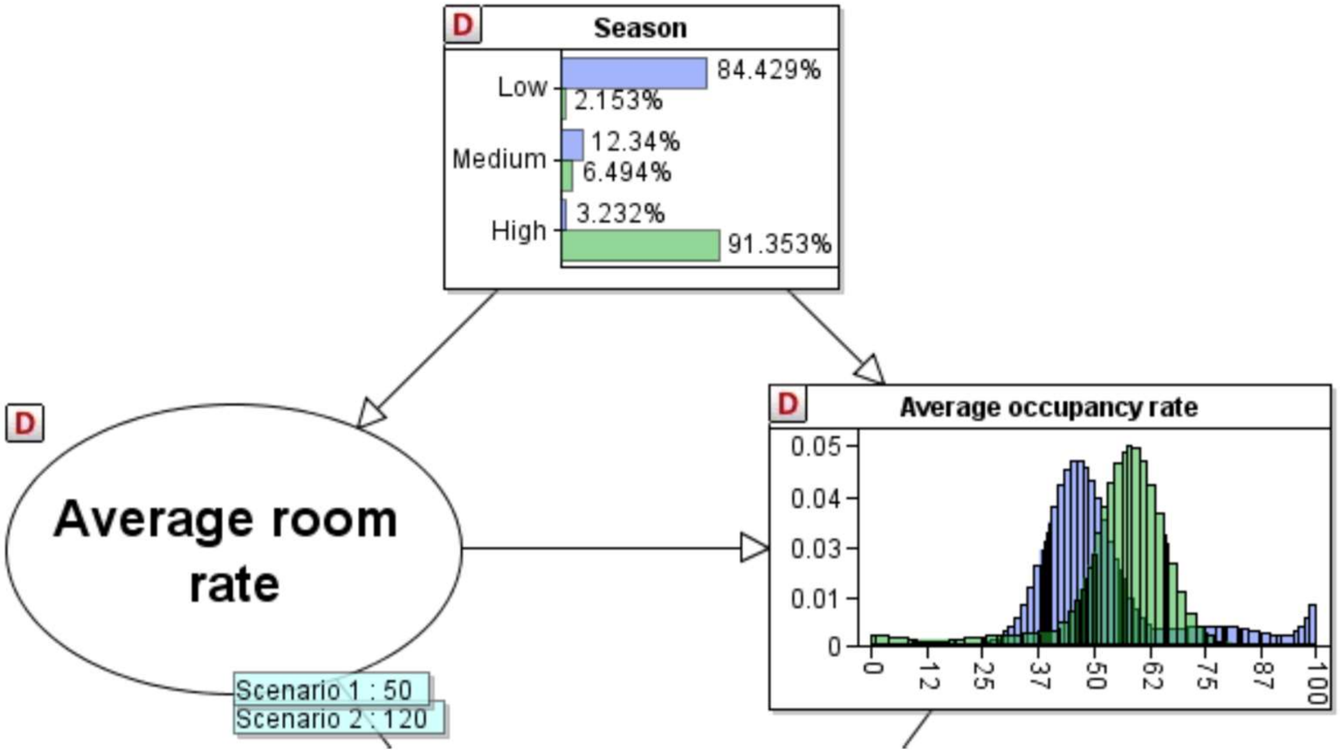


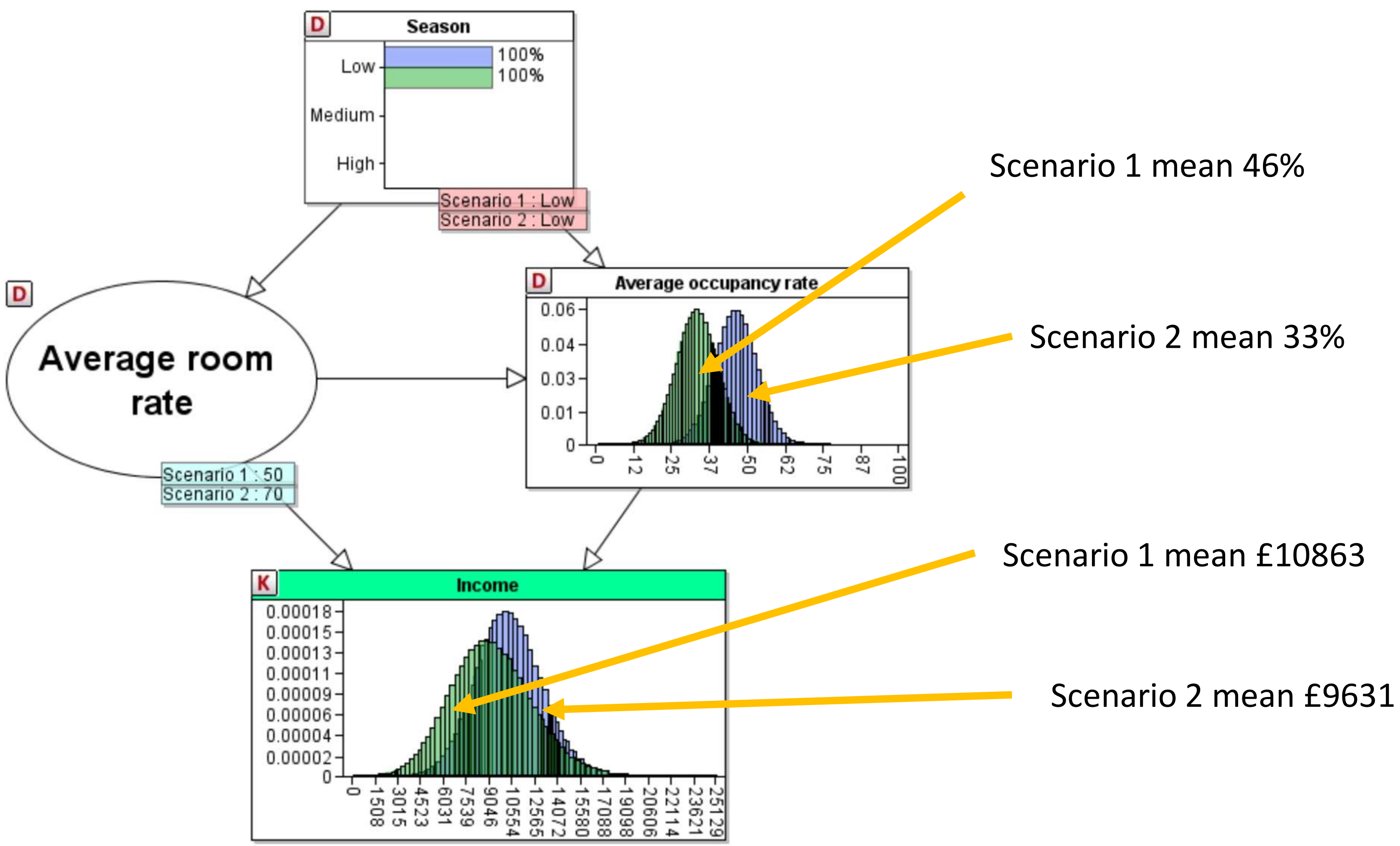
Based on observational data, as prices increase the occupancy rate increases. So we should increase prices to maximize occupancy?

Missing causal factor



Higher rates in high season




Season
Low
Medium
High
100%
100%
Scenario 1 : Low
Scenario 2 : Low
Average room rate
Scenario 1 : 50
Scenario 2 : 70
Average occupancy rate
0.06
0.04
0.03
0.01
0
0
12
25
37
50
62
75
87
100
Income
0.00018
0.00015
0.00013
0.00011
0.00009
0.00006
0.00004
0.00002
0
0
1508
3015
4523
6031
7539
9046
10554
12565
14072
15580
17088
19098
20606
22114
23621
25129
Scenario 1 mean 46%
Scenario 2 mean 33%
Scenario 1 mean £10863
Scenario 2 mean £9631

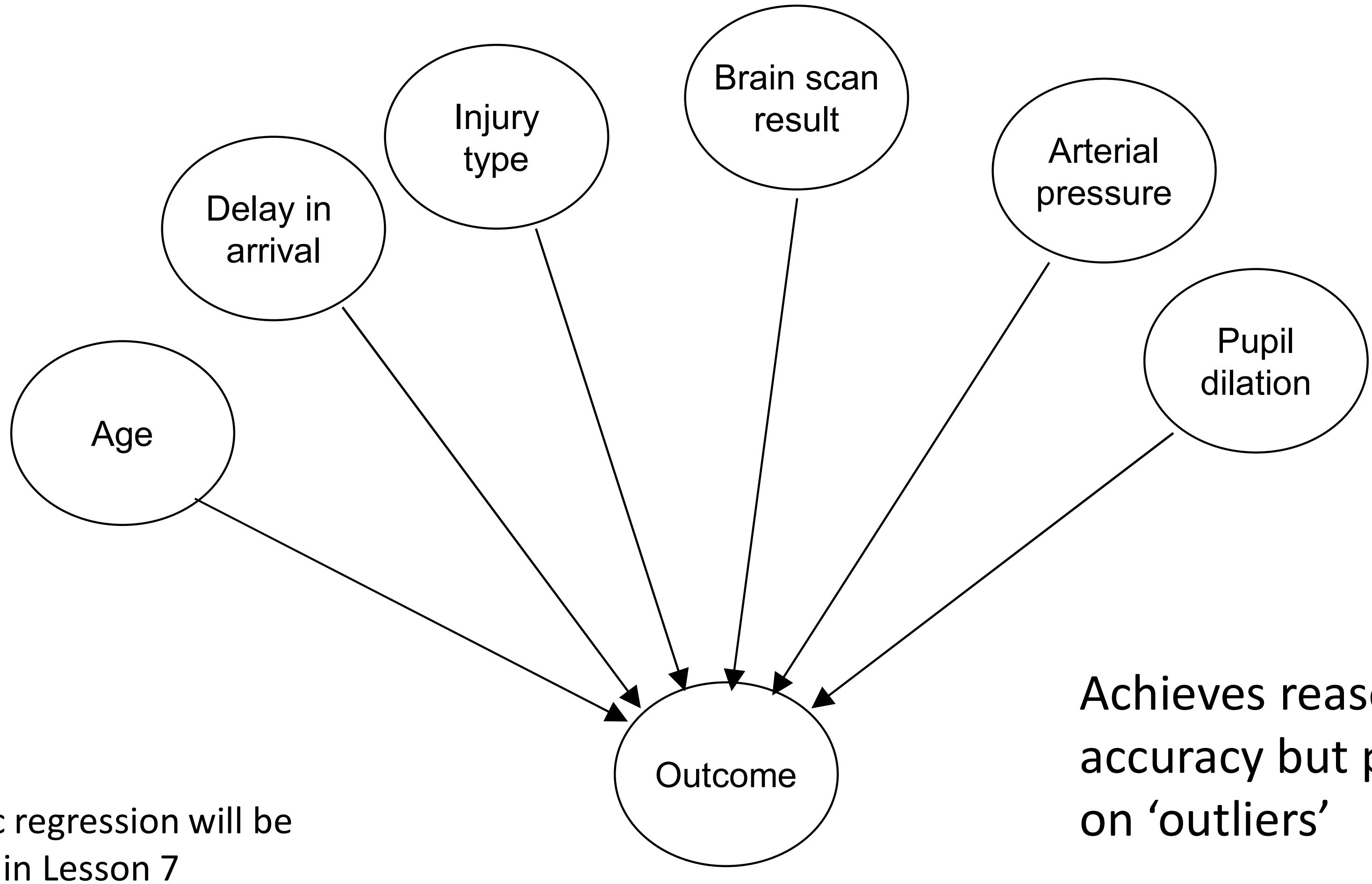
Summary

- The conclusions made in many empirical studies are often compromised because of causal paradoxes
- Where there are either ‘hidden’ confounders or confounders that are not properly ‘controlled’, then the overall conclusions may be completely the opposite of reality (e.g. based on the data, Drug A may be deemed more effective than Drug B even though the opposite is true).
- This ‘reversal’ of the correct results when we drill down to take proper effect of confounders is called Simpson’s paradox
- Unobserved confounders can also lead us to wrongly conclude that relationships exist when they do not
- Where the data under-represents some factors and/or over-represents others (‘censoring’), the overall conclusions may again be completely the opposite of reality.
- This ‘reversal’ of the correct results when we take proper effect of ‘censoring’ in the data is called Berkson’s paradox (or collider paradox)
- Censored data can also lead us to wrongly conclude that relationships exist when they do not
- Many empirical studies are compromised by both confounders and censored data
- In theory randomized control trials (RCTs) can avoid all of these problems, but in practice even if we knew all the possible confounding factors we cannot get enough data/subjects to control for them
- By casting a study (based only on observational data rather than a RCT) as a causal model that incorporates potential confounders (observable or not) and colliders, we can avoid the paradoxes and make reliable conclusions

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
…	…	…		..	…	…

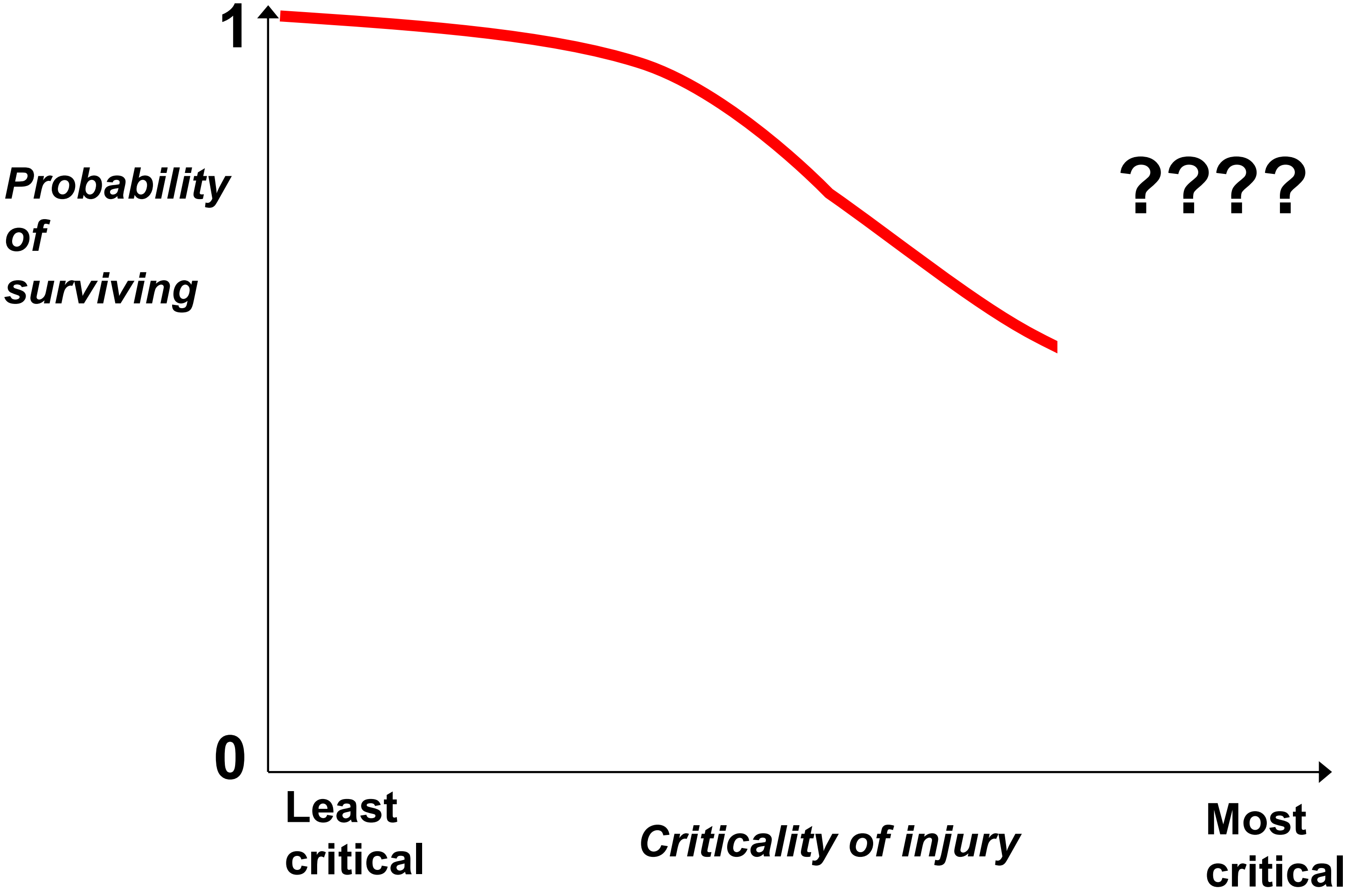
Regression model* learnt purely from data ('supervised learning')



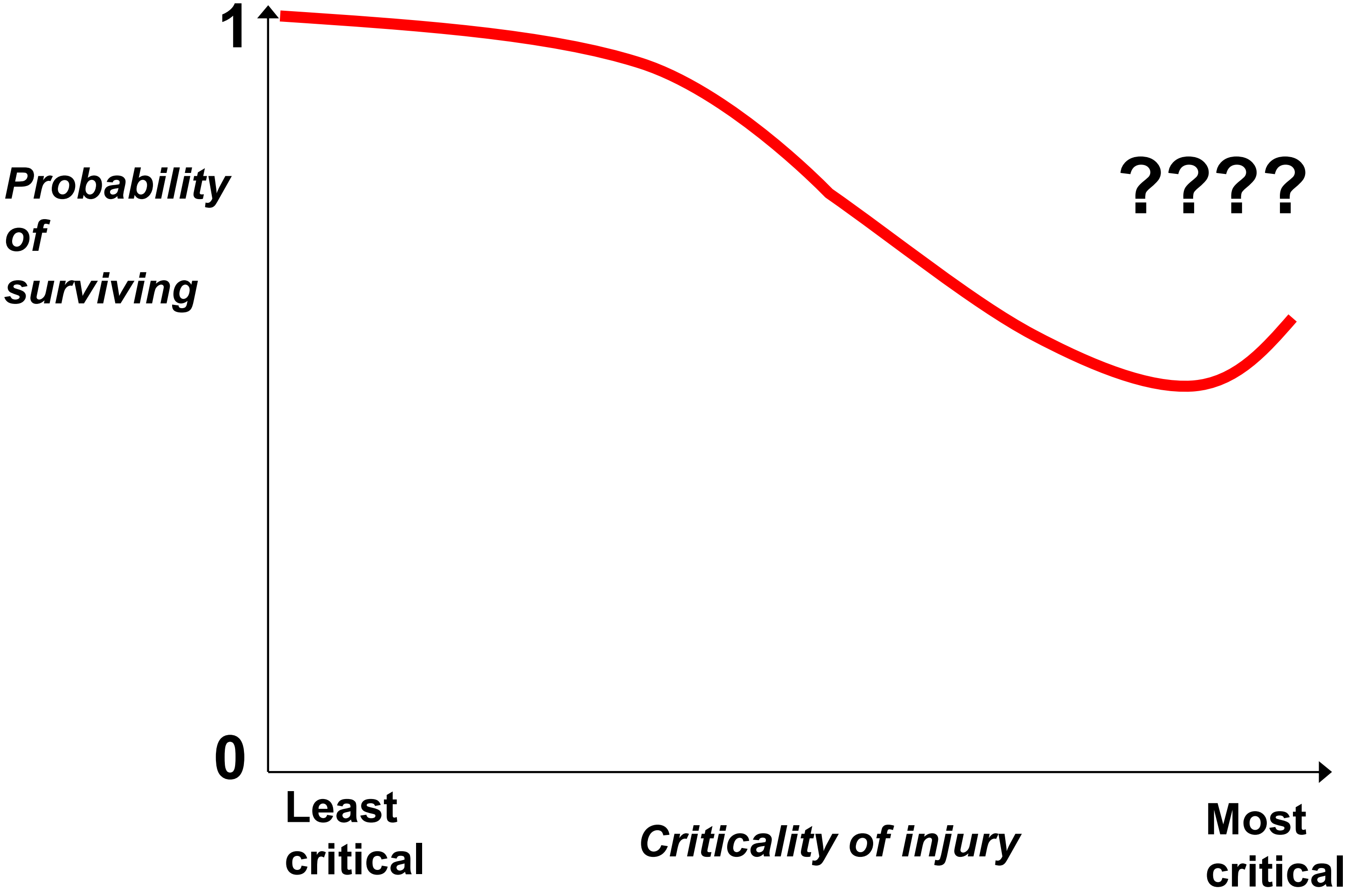
Achieves reasonable accuracy but performs badly on 'outliers'

*Logistic regression will be covered in Lesson 7

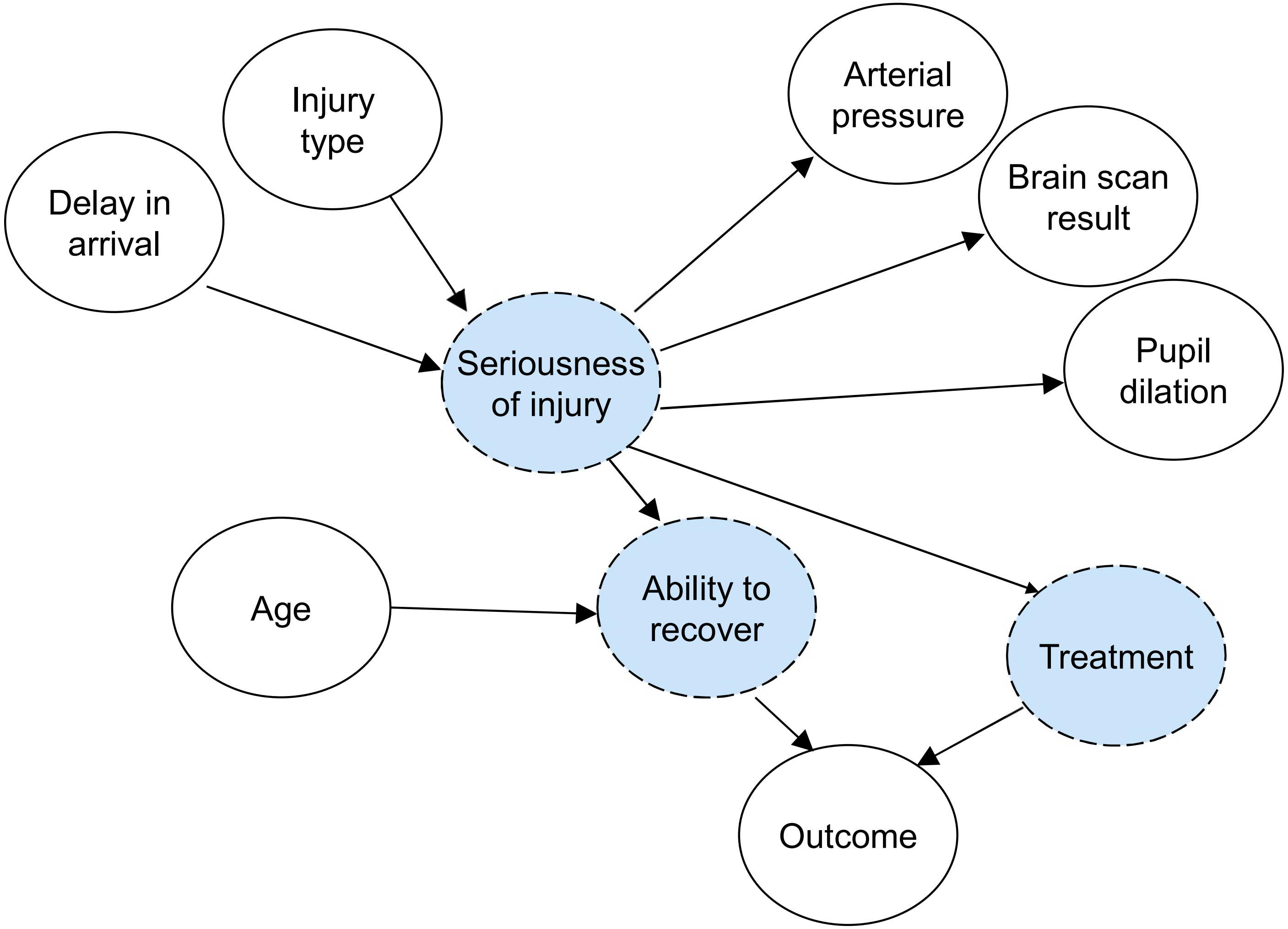
Traumatic head injury: Some observational data



Traumatic head injury: Some observational data

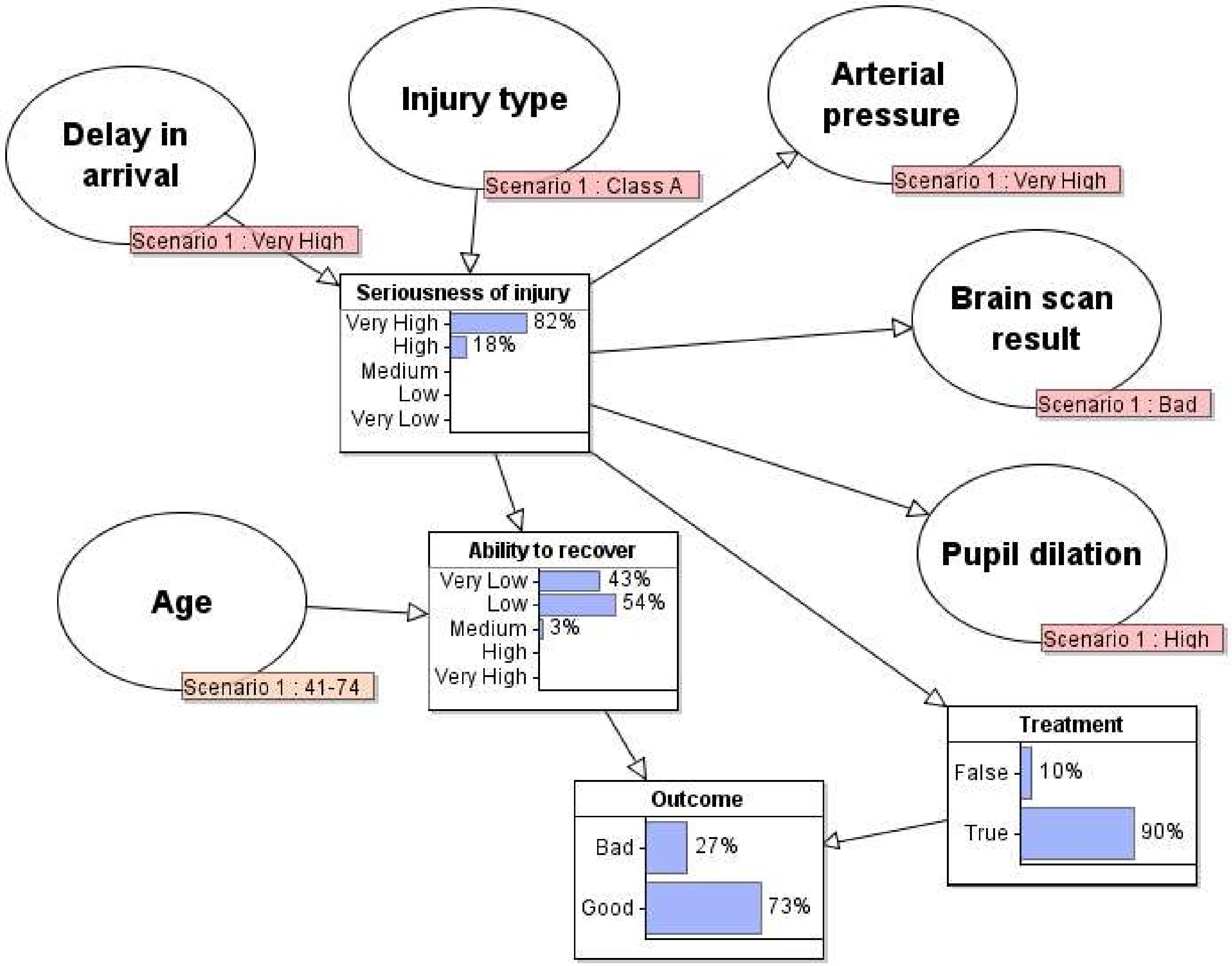


Expert causal BN with hidden explanatory and intervention variables



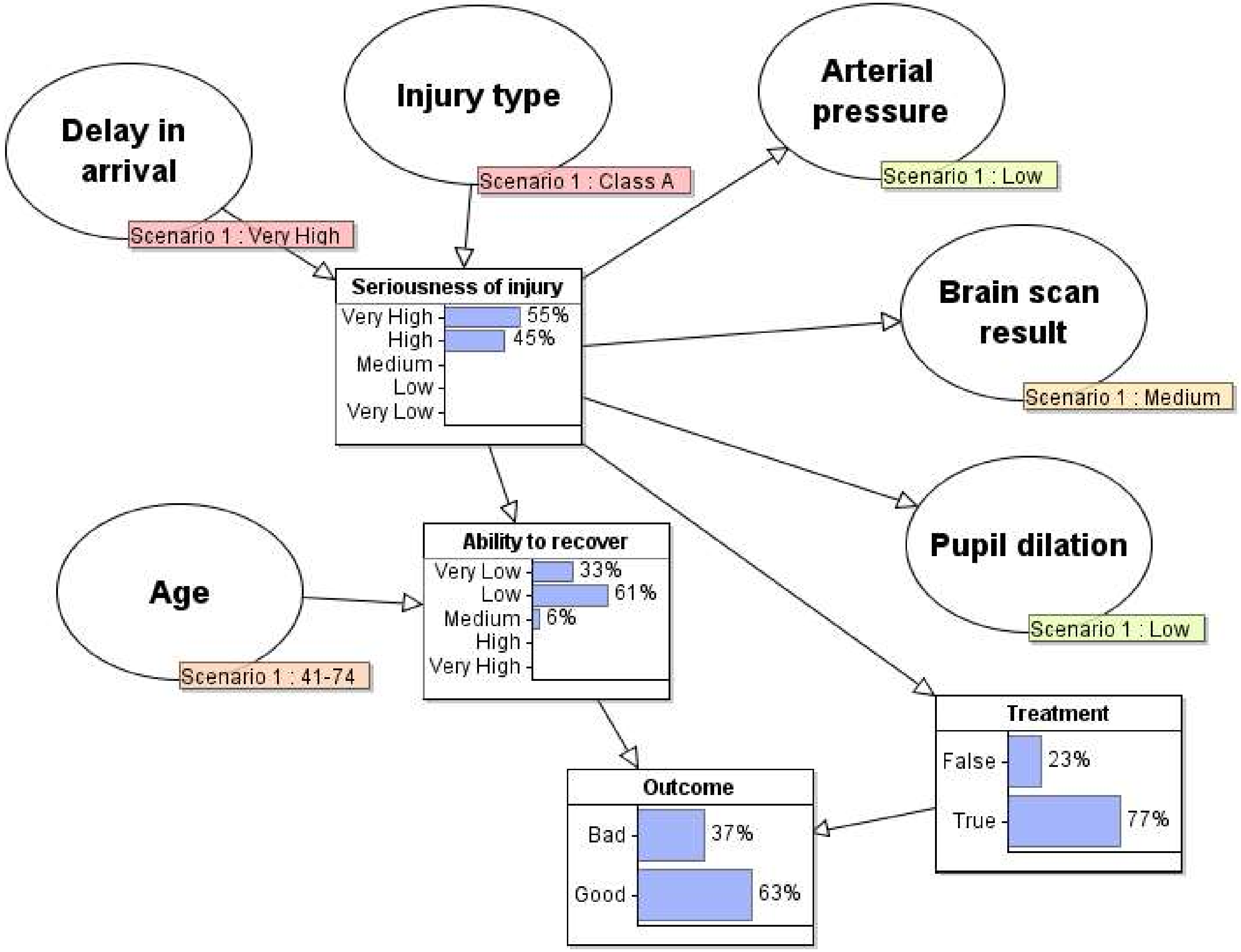
More accurate (properly handles outliers), more useful for decision support

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

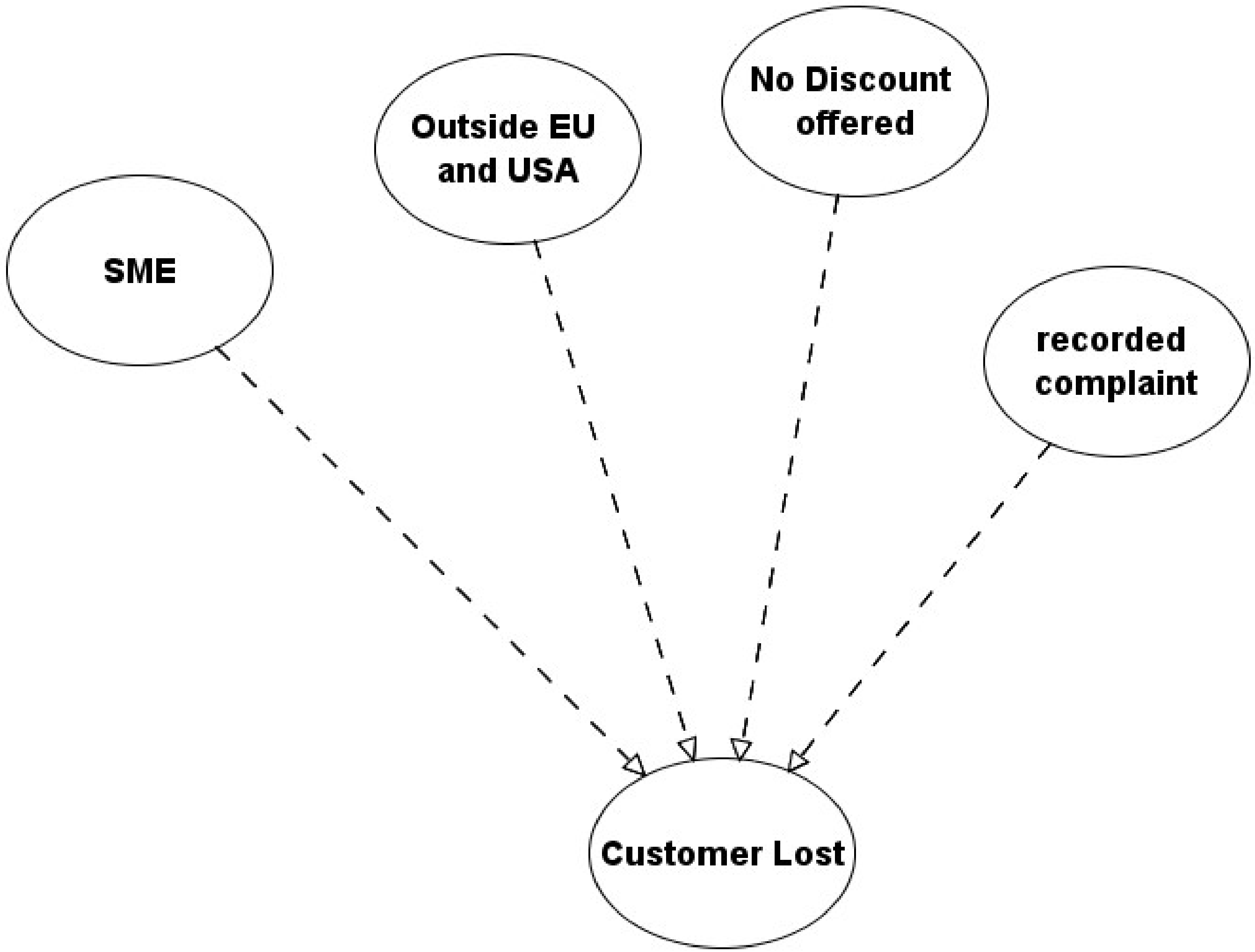
Customer retention

	"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

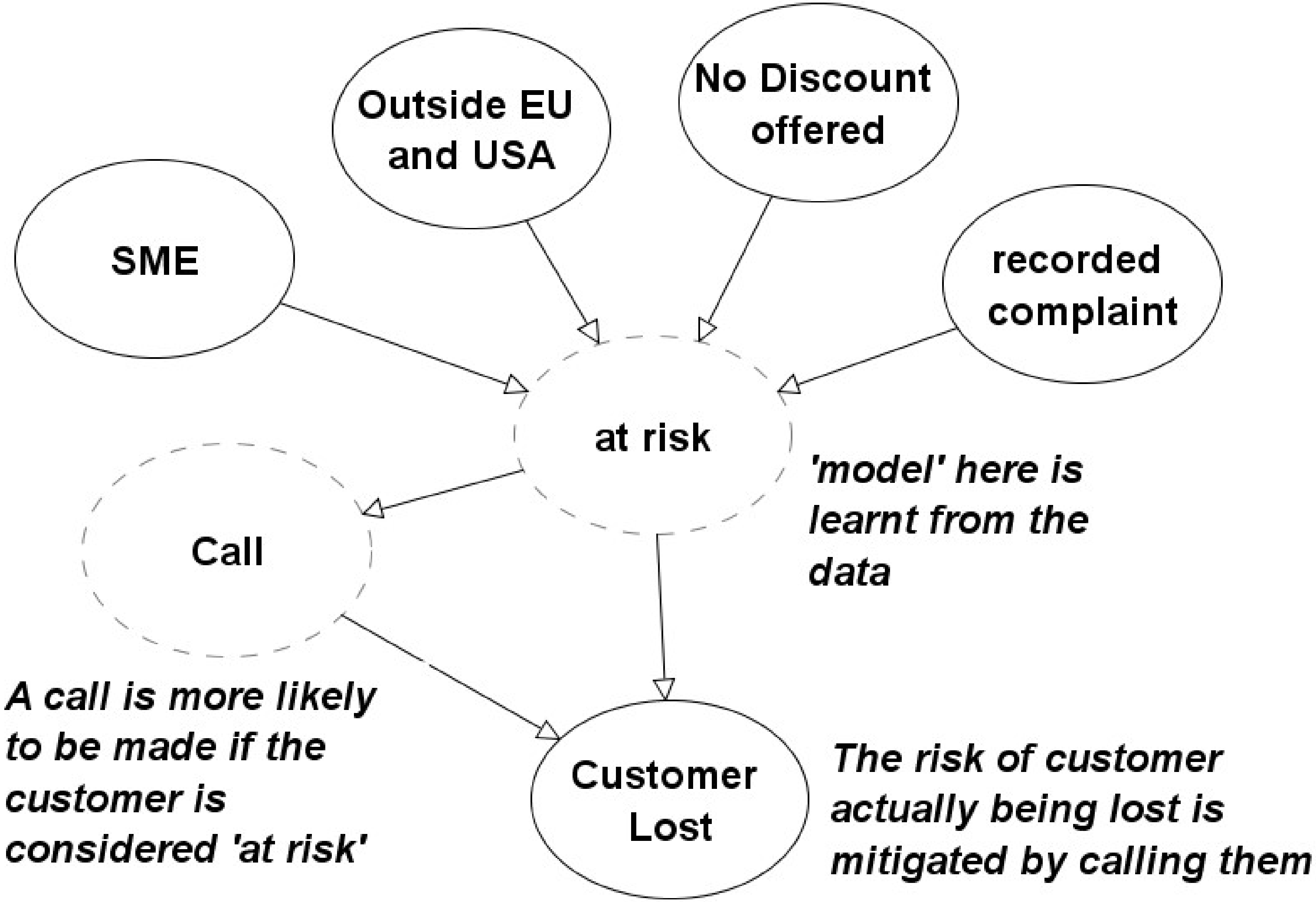
Risk Factor ‘profiles’

"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

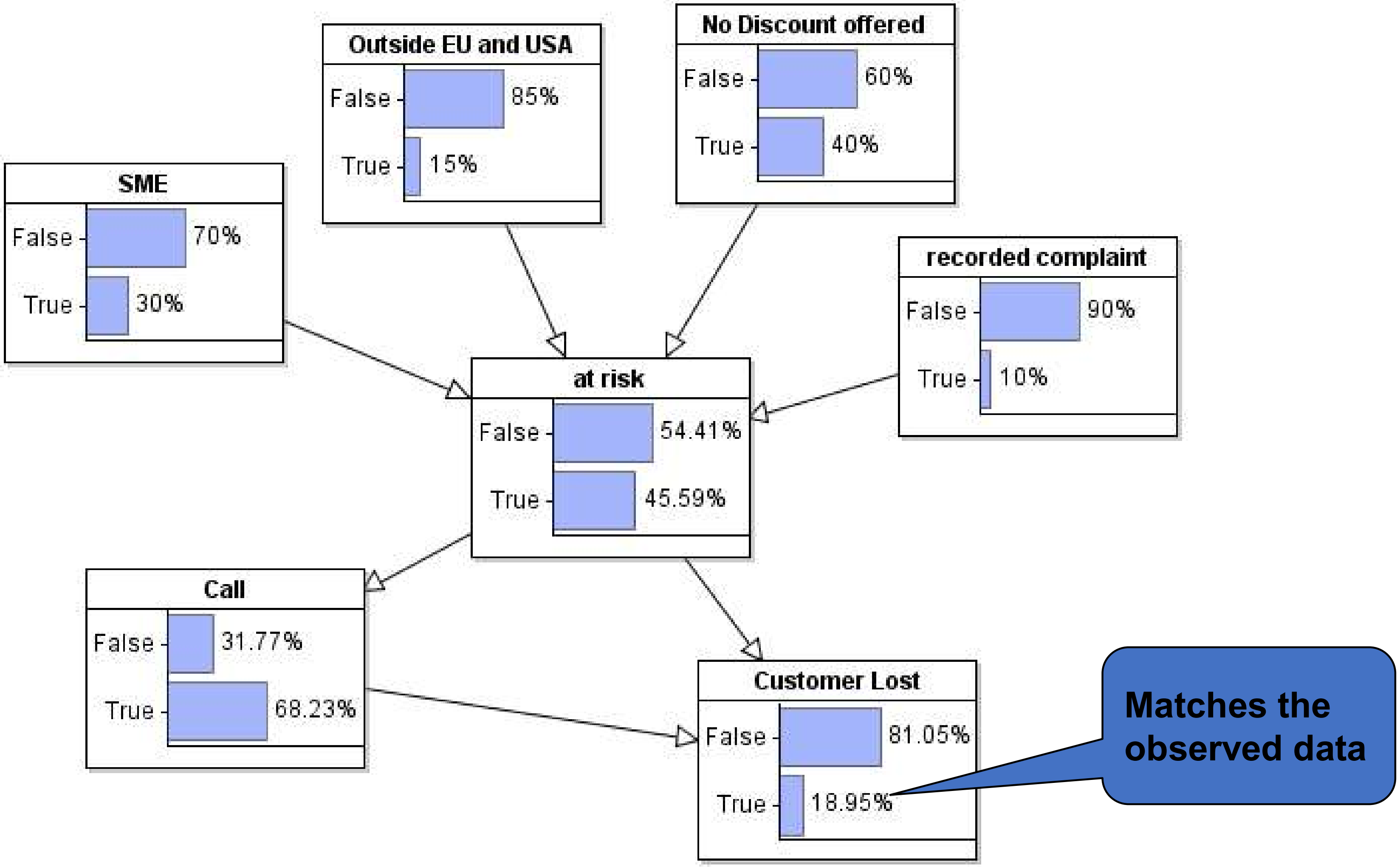
Temptation to build this model ('supervised learning')



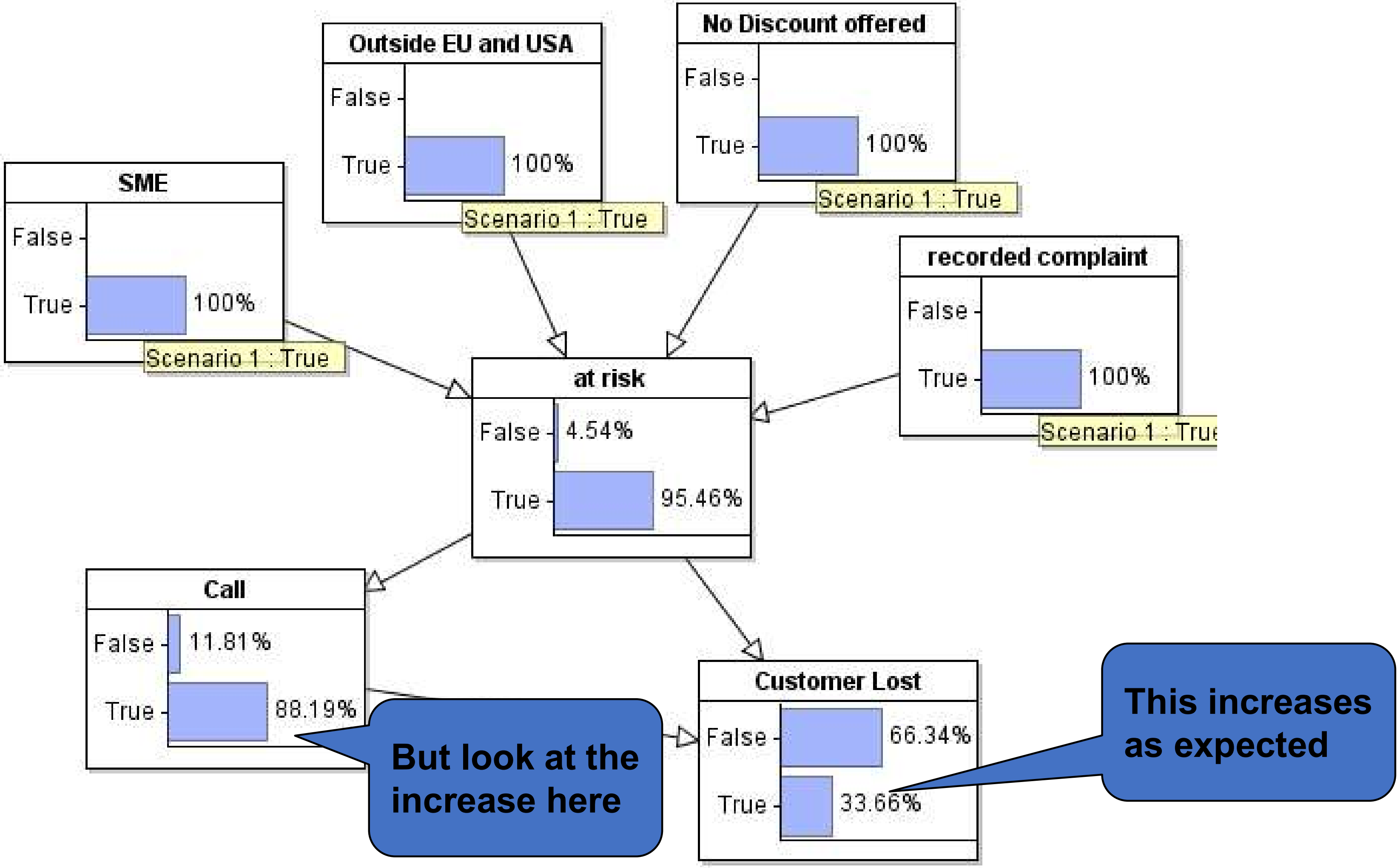
Causal BN built with data and knowledge



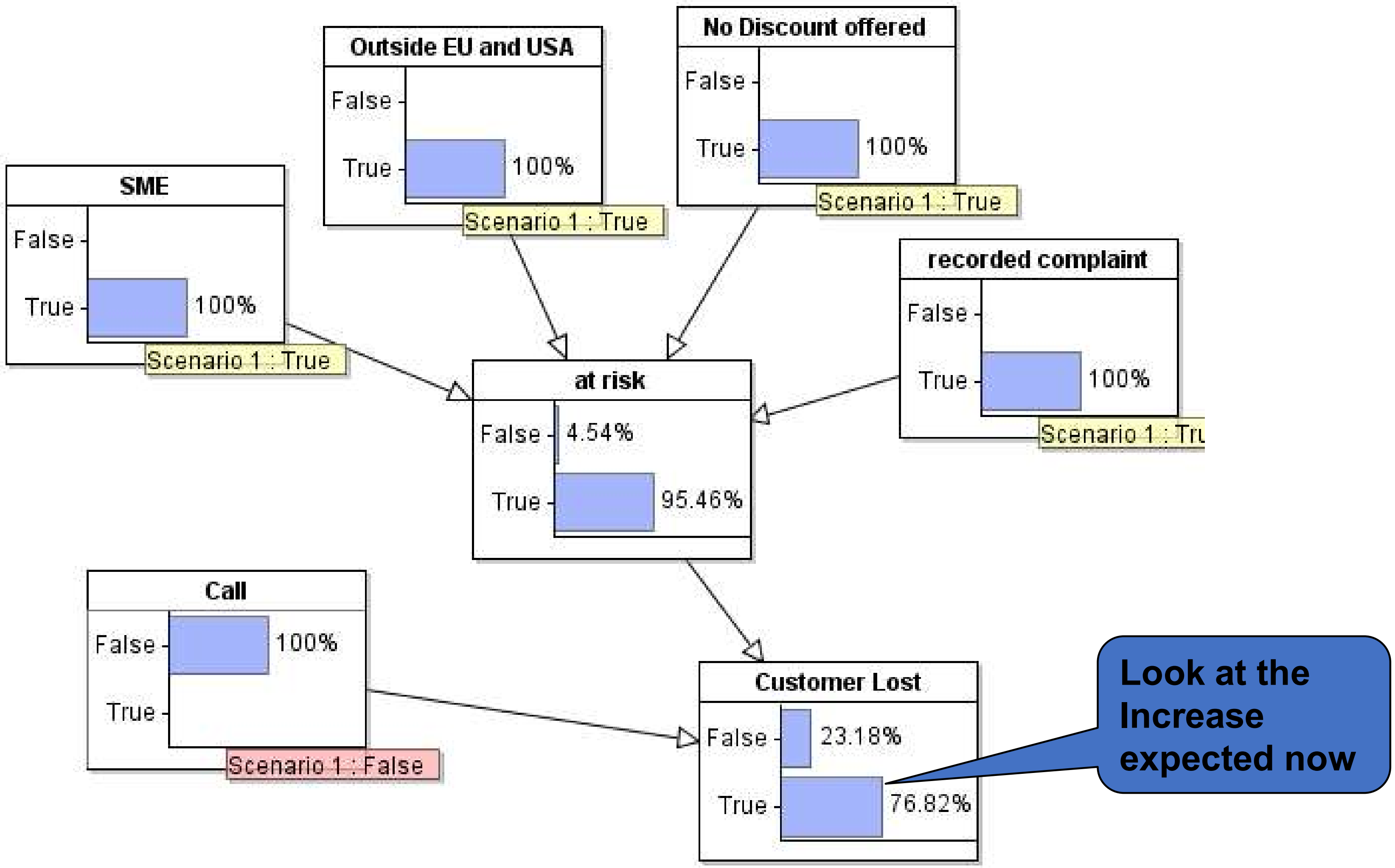
Model with marginal probabilities



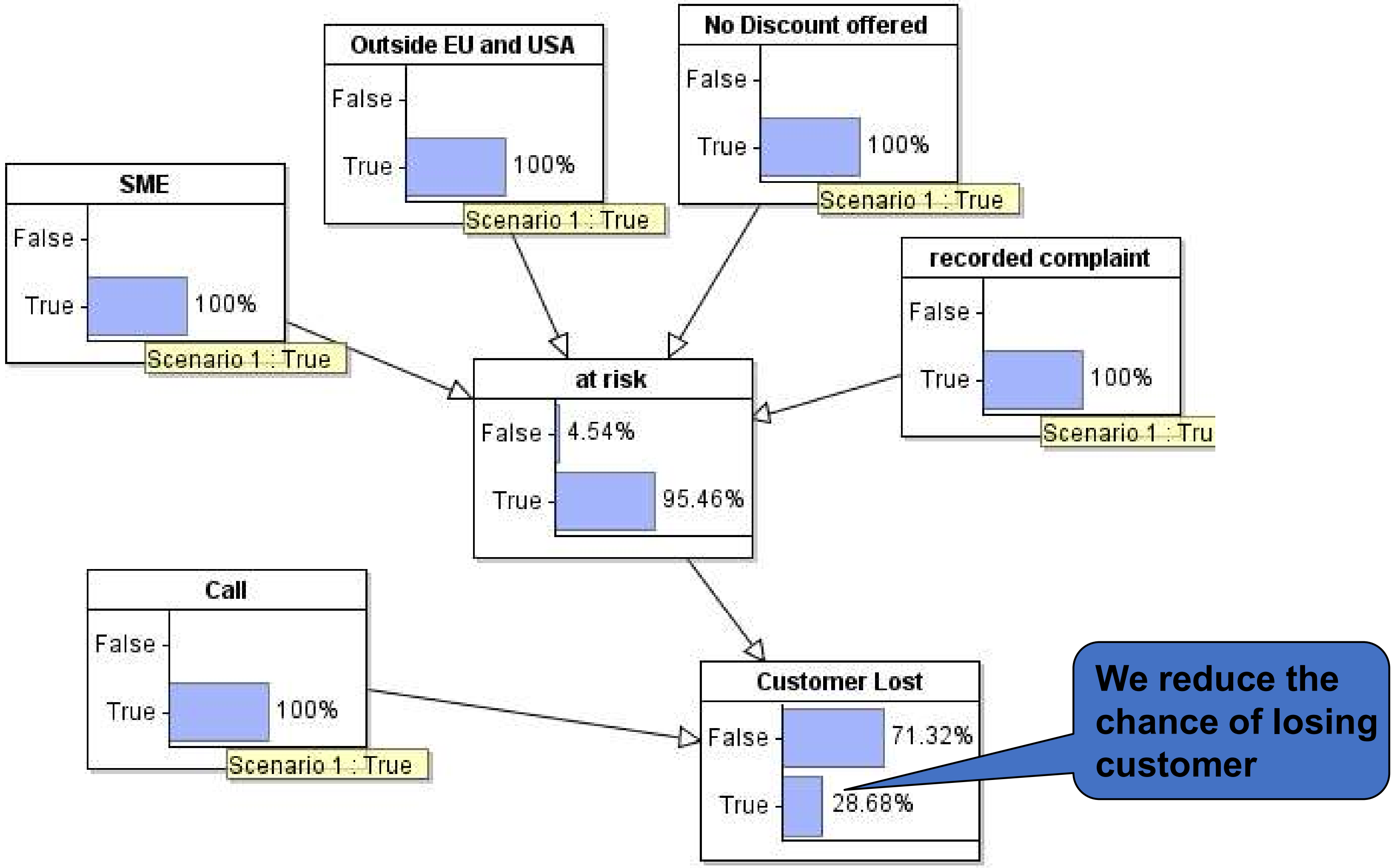
When risk factors are all true



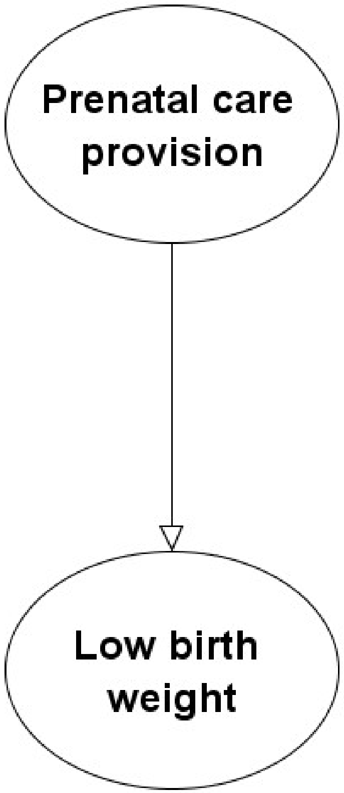
If we do not make the call …



But if we make the call …

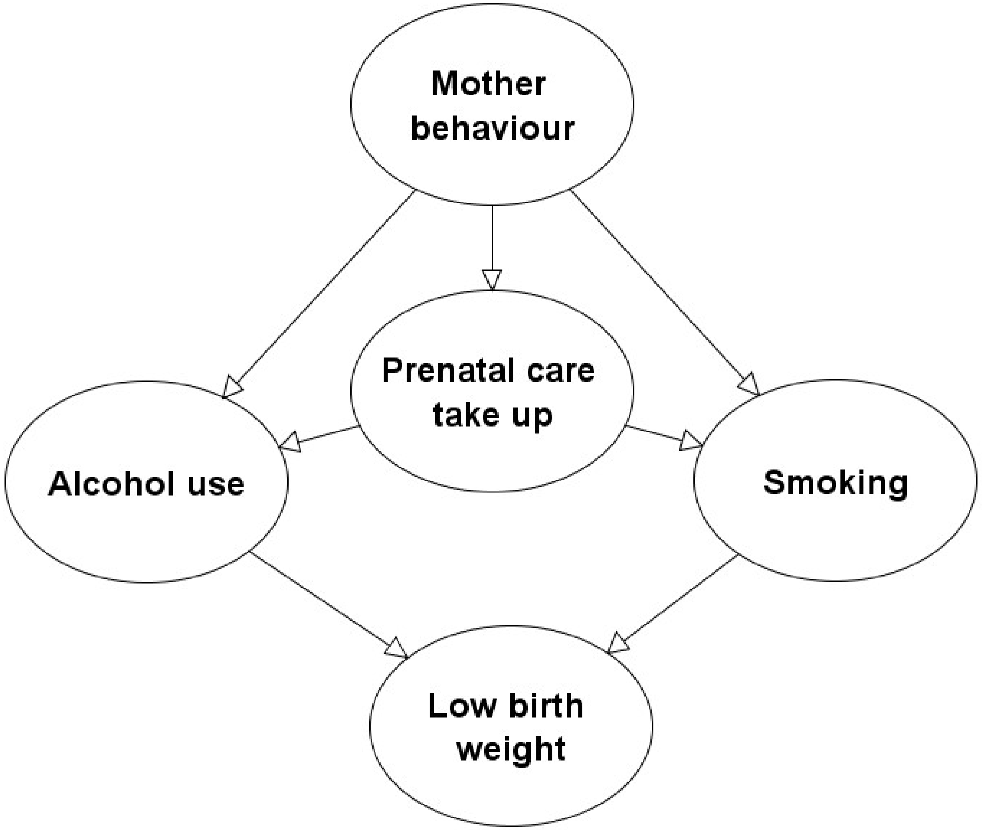


When ideology and causation collide



- Causal explanations can be proposed based on ideological position
- Correlation between prenatal care provision and weight of new born babies
- Mothers receiving low levels of prenatal care were disproportionately poor
- Media blamed social failure to provide enough prenatal care to poor women and called for increased resources

When ideology and causation collide



- Independent of income, smoking and alcohol abuse more prevalent amongst mothers who did not get prenatal support
- Additional factor not included – personal responsibility.
- A deficit in personal responsibility may explain failure to seek prenatal care.