Three Stats Pitfalls Facing the New Data Scientist

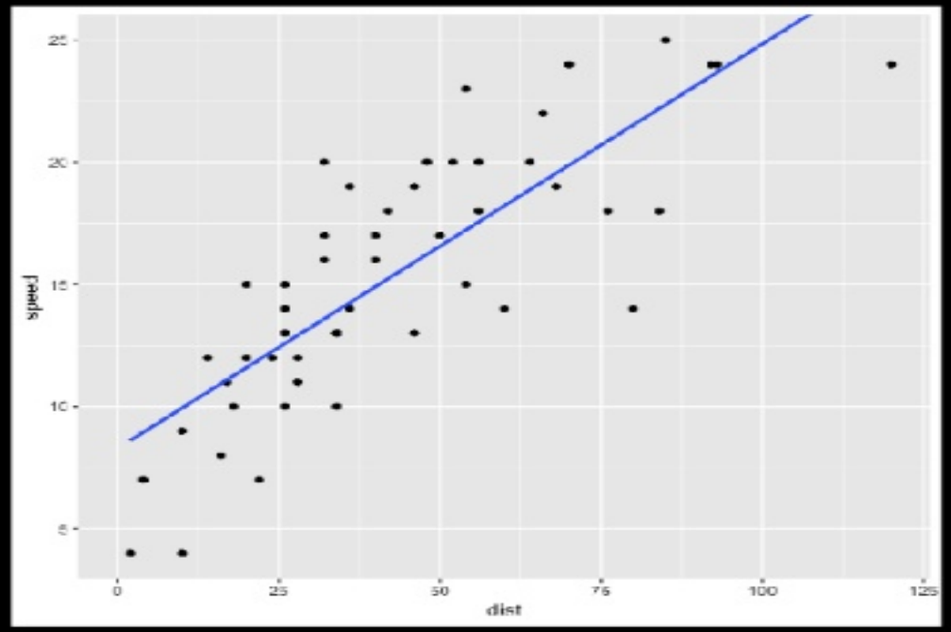
Sean Owen

Do I Know You?

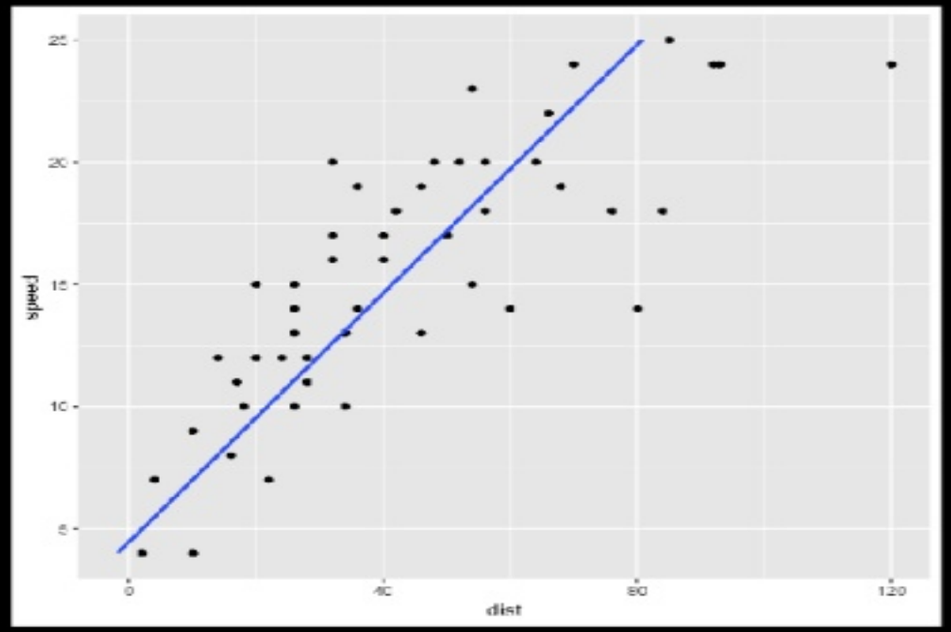
- Apache Spark committer, PMC
- “Advanced Analytics with Spark”
- Recently: Director, Data Science @ Cloudera

Correlation is not causation.
But then, what is causation?

Which Best-Fit Line Is Best?




```
lm(speed ~ dist, cars)
```




```
lm(dist ~ speed, cars)
```

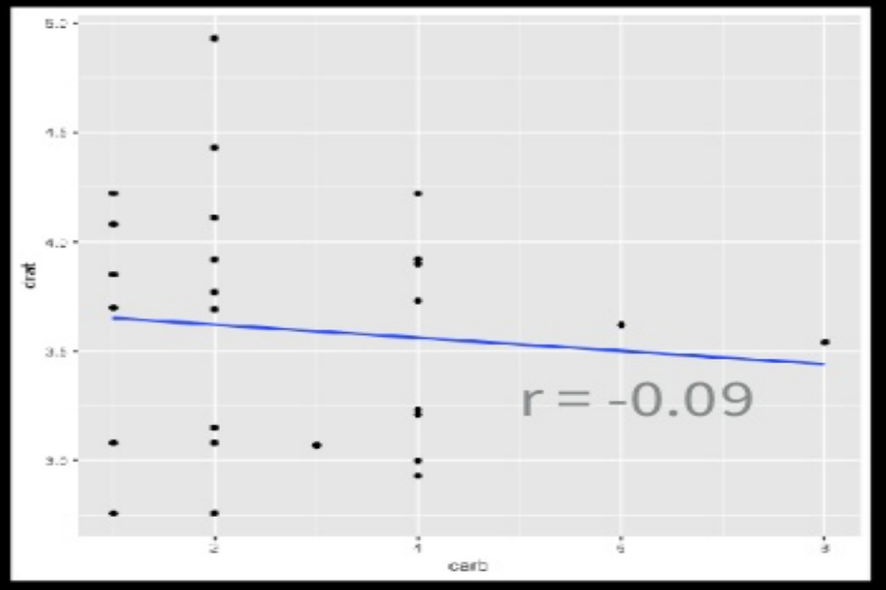
Which Treatment is Better?

	Treatment A	Treatment B
Small Stones	**93% (81/87)**	87% (234/270)
Large Stones	**73% (192/263)**	69% (55/80)
	78% (273/350)	**83% (289/350)**

Now, Which Treatment is Better?

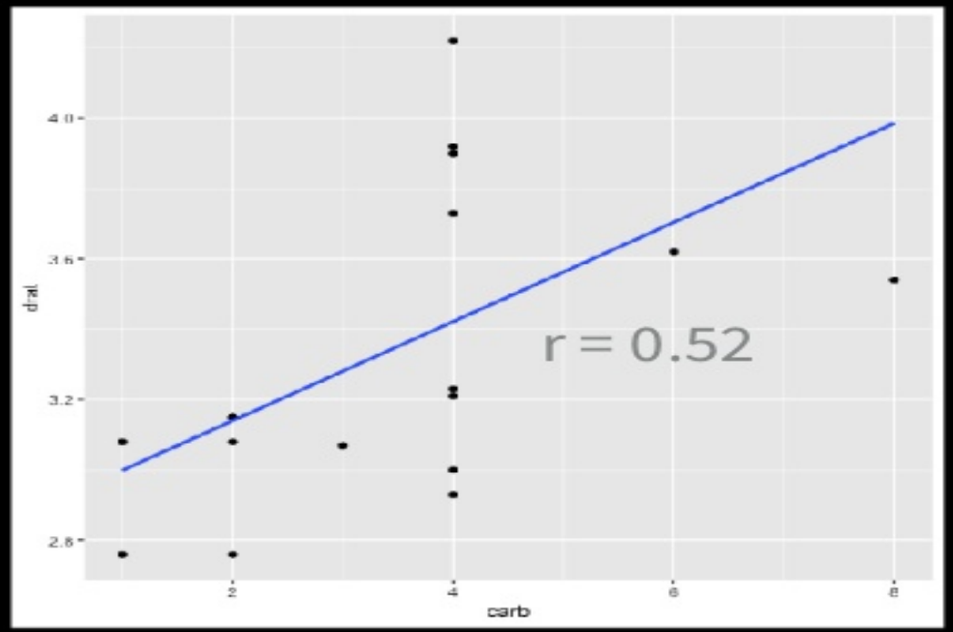
	Treatment A	Treatment B
Low Blood pH	**93% (81/87)**	87% (234/270)
High Blood pH	**73% (192/263)**	69% (55/80)
	78% (273/350)	**83% (289/350)**

Carburetors and Axle Ratio Uncorrelated

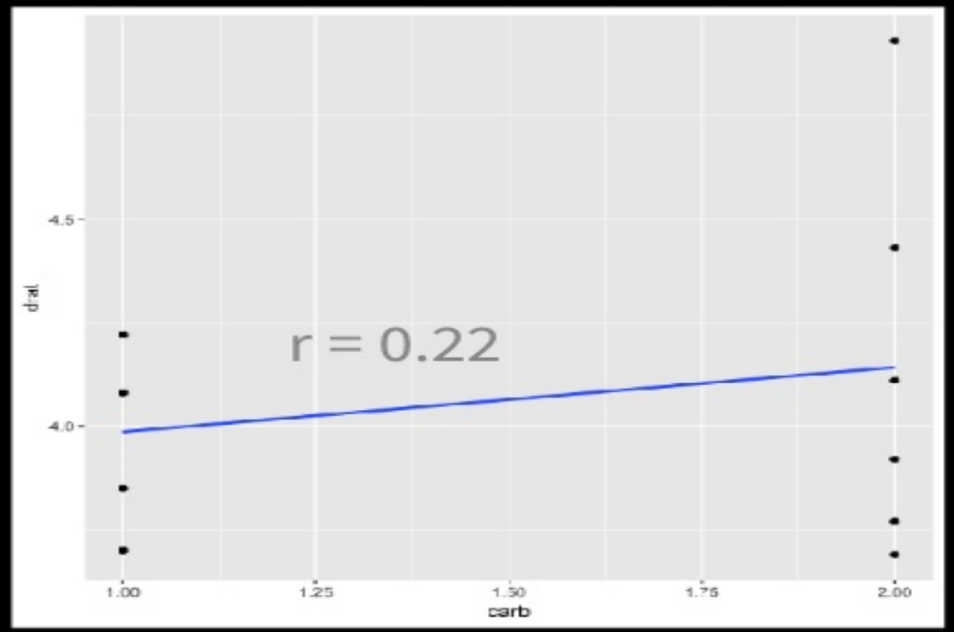



```
lm(drat ~ carb, mtcars)
```

. . . Except in Both Halves of the Data?




```
lm(drat ~ carb, mtcars[
  which(mtcars$cyl >= 6),])
```



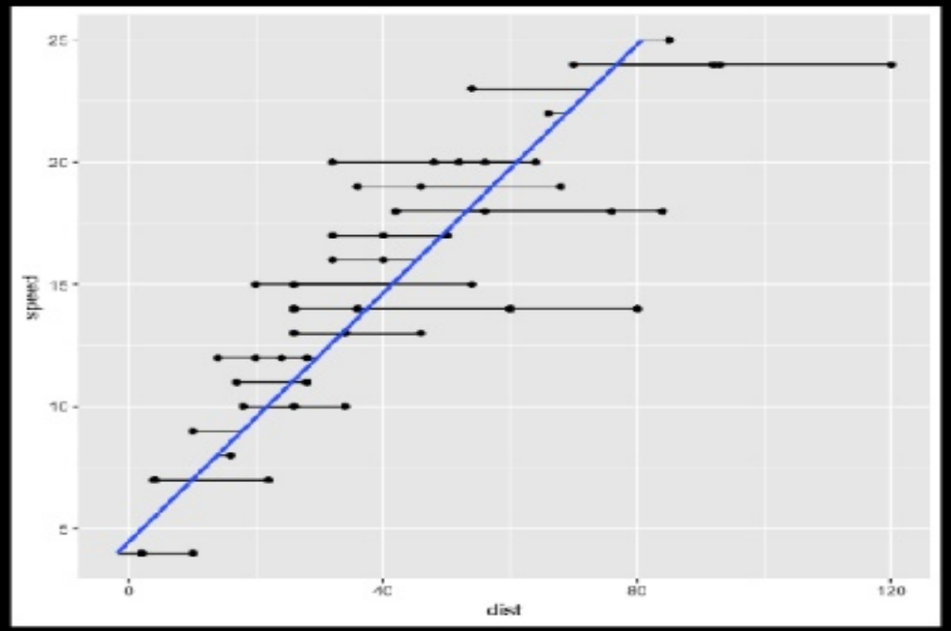

```
lm(drat ~ carb, mtcars[
  which(mtcars$cyl < 6),])
```

3 Answers

Resolution: Causation

- Humans reason causally
- Data doesn’t contain causal information
- Data correlations consistent with multiple causal models
- Correct inference requires adding causal model

One Consistent with Causal Knowledge

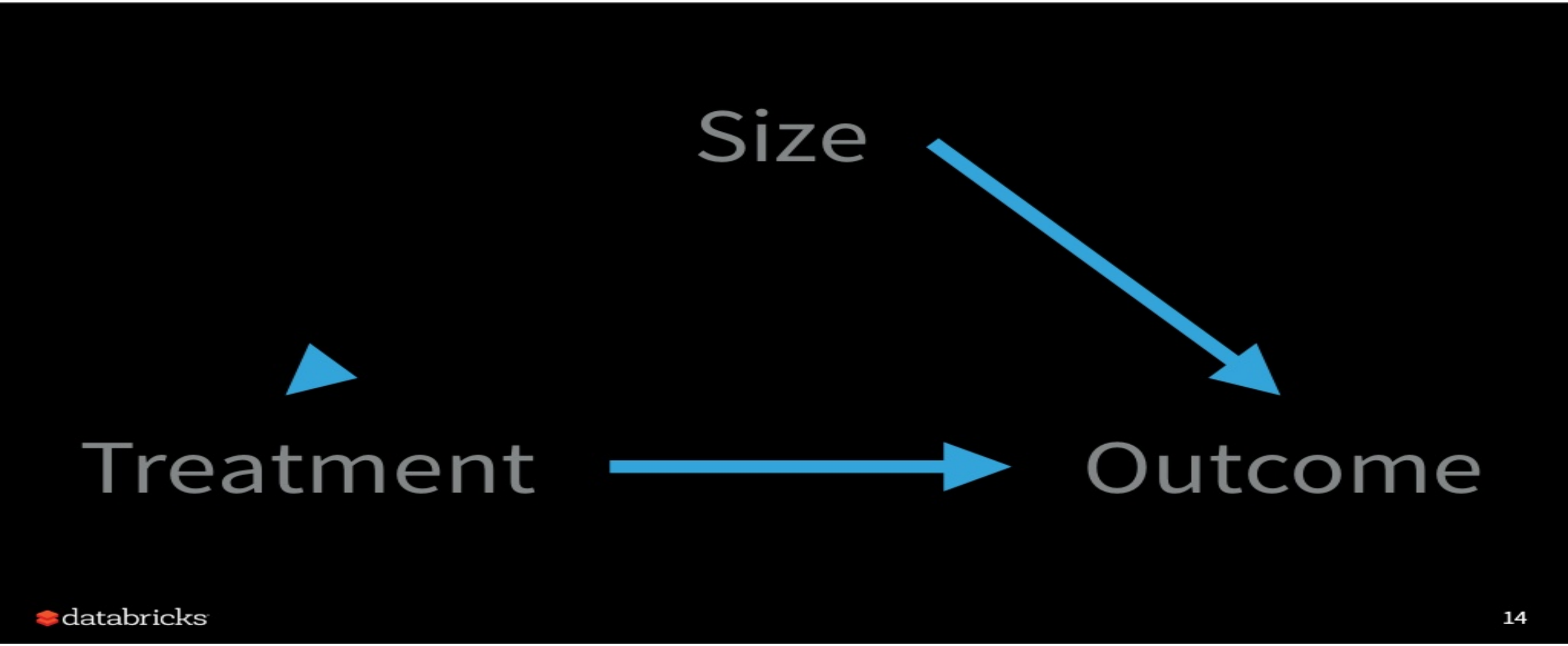


$$dist_i = \beta_0 + \beta_1 speed_i + \epsilon_i$$

Speed  Distance

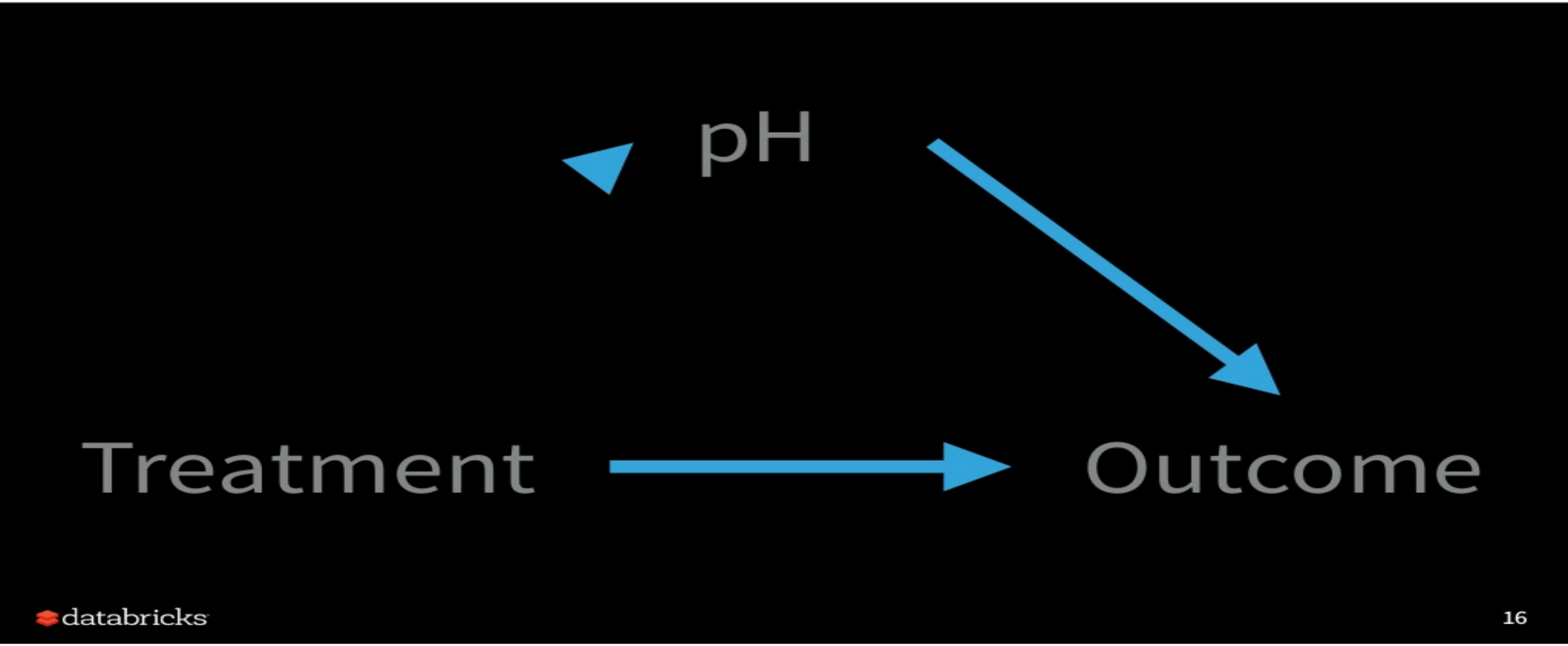
Controlling Confounders is Right

	Treatment A	Treatment B
Small Stones	**93% (81/87)**	87% (234/270)
Large Stones	**73% (192/263)**	69% (55/80)
	78% (273/350)	**83% (289/350)**

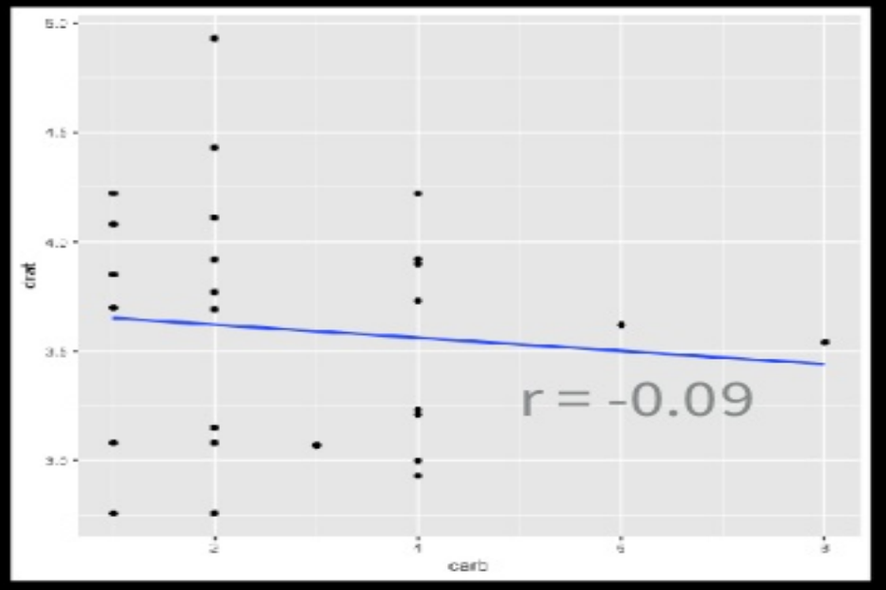

Size
Treatment
Outcome

Controlling Mediators is Wrong

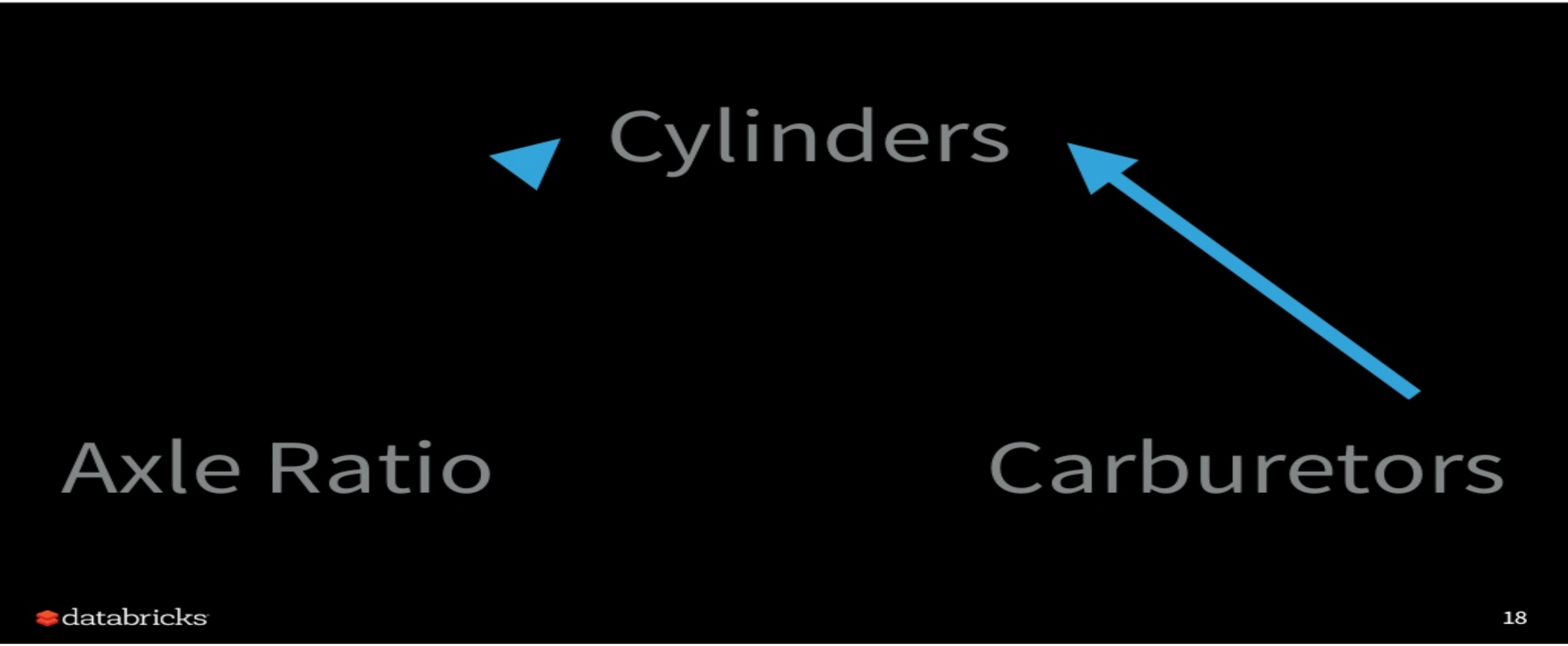
	Treatment A	Treatment B
Low Blood pH	**93% (81/87)**	87% (234/270)
High Blood pH	**73% (192/263)**	69% (55/80)
	78% (273/350)	**83% (289/350)**


pH
Treatment
Outcome

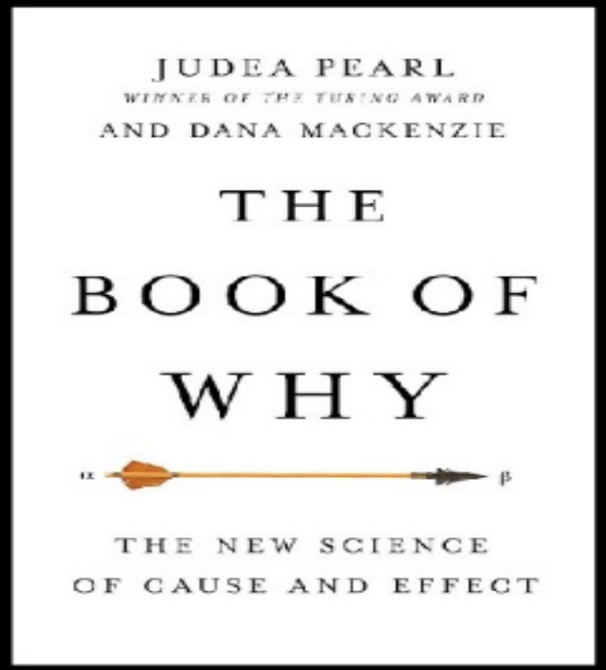
Colliders Create Correlation

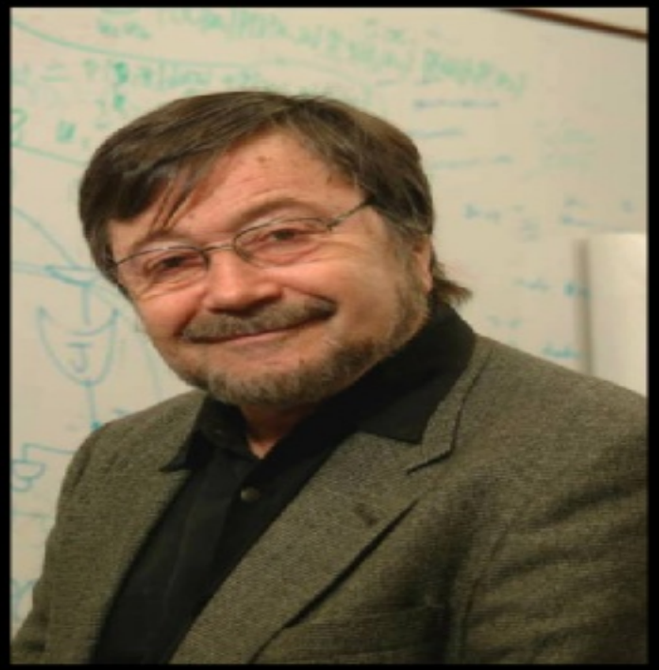



```
lm(drat ~ carb, mtcars)
```


Cylinders
Axle Ratio
Carburetors

Causation and do-Calculus



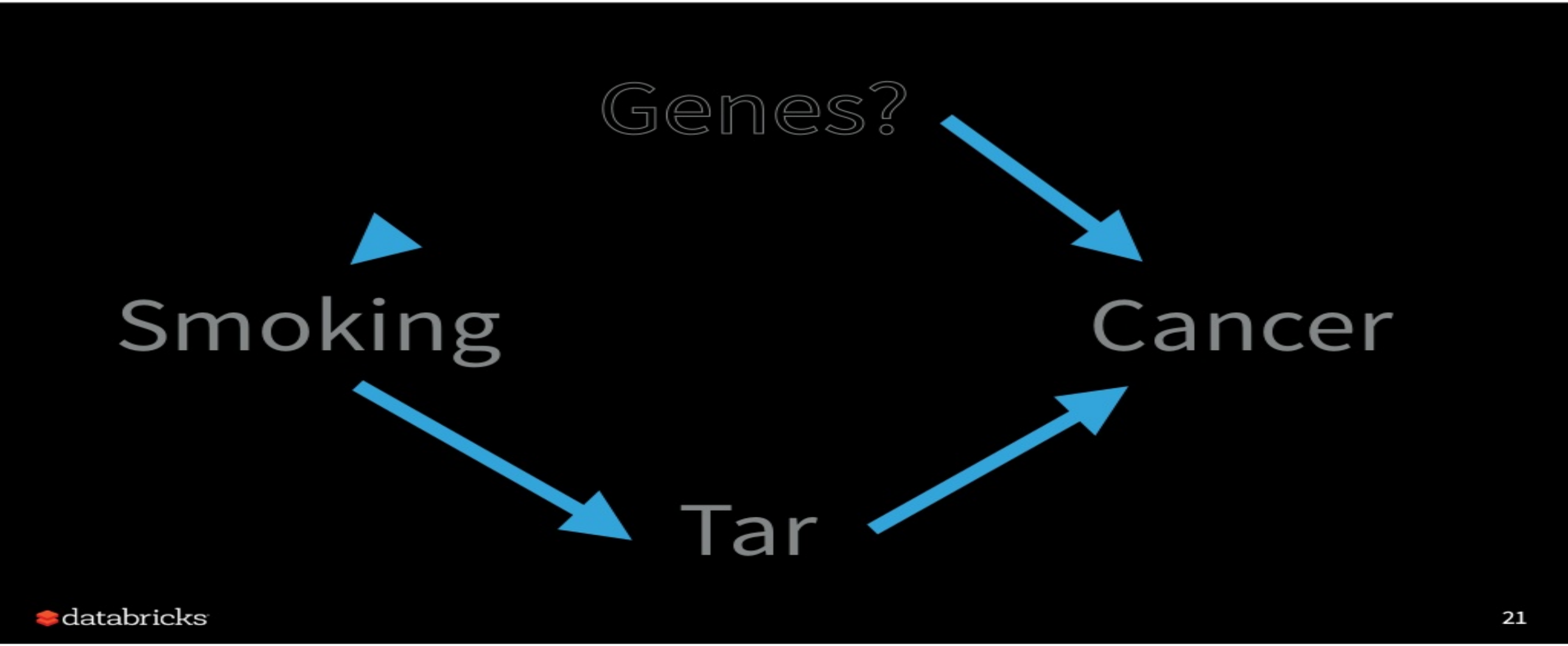


do-Calculus

$$P(Y|X) \neq P(Y|do(X))$$

Just because it's more often raining when you walk outside with an umbrella ...

... doesn't mean that you carrying an umbrella makes it more likely to be raining.


Genes?
Smoking
Cancer
Tar

$$
\begin{aligned}
P(C \mid do(S)) &= \sum_t P(C \mid do(S), t) P(t \mid do(S)) \\
&= \sum_t P(C \mid do(S), do(t)) P(t \mid do(S)) \\
&= \sum_t P(C \mid do(S), do(t)) P(t \mid S) \\
&= \sum_t P(C \mid do(t)) P(t \mid S) \\
&= \sum_{s'} \sum_t P(C \mid do(t), s') P(s' \mid do(t)) P(t \mid S) \\
&= \sum_{s'} \sum_t P(C \mid t, s') P(s' \mid do(t)) P(t \mid S) \\
&= \sum_{s'} \sum_t P(C \mid t, s') P(s') P(t \mid S)
\end{aligned}
$$

Conclusion

- Must bring causal info to data for proper interpretation
- Know common causal pitfalls!
- PGMs help reason about causal effects
- Do-calculus can clarify reasoning about intervention


Thank You
@sean_r_owen
databricks