> Is there any additional value in knowing a variable, once I already know all of the other predictor variables?

This might help us identify when we are collecting useless extra data.

If we look at the model in 5.1.3, we see how we build a model to check the ways in which two variables, or factors, influence the model.

Causal Workflow

The point of causal modeling is to create a model *outside of statistics* to describe what we think is happening in our research world.

It is about causation, i.e., we specifically want to see if some intervention is bringing about a change in the predicted variable. If we don't care about causation then this model wouldn't apply. But note in my experience that is usually the type of question data science is curious about.

title: Causal Models
date: May 2023
author: Neil Ernst
marp: true

Causal Workflow: paths

1. Identify the data we have collected or will collect (lines of code, complexity, hours worked, etc). Since the universe is itself causal, we clearly cannot collect data on everything.
2. Consider unobserved (latent) variables - we won't or cannot measure it (e.g. for privacy reasons), but we need to account for it.
3. Decompose the DAG into three elemental paths: **forks** ($X \leftarrow Z \rightarrow Y$) **pipes** ($X \rightarrow Z \rightarrow Y$) and **colliders** ($X \rightarrow Z \leftarrow Y$)

Pipes

- If ignore Z, then X and Y are associated. Stratify by Z, X and Y not associated. All of the information about X comes to Y through Z. No additional information about X if we know Z.

Forks

- Same structure as pipes. X and Y are associated because they both have information only from Z. Once you learn Z, learning X doesn't give more data on Y.
- Need more data to distinguish forks from pipes.

Collider

- Ignore Z, X and Y NOT associated. Stratify by Z, X and Y are associated.
- X and Y are independent causes of Z.
- X and Y have mutual information: learning X tells me something about Y