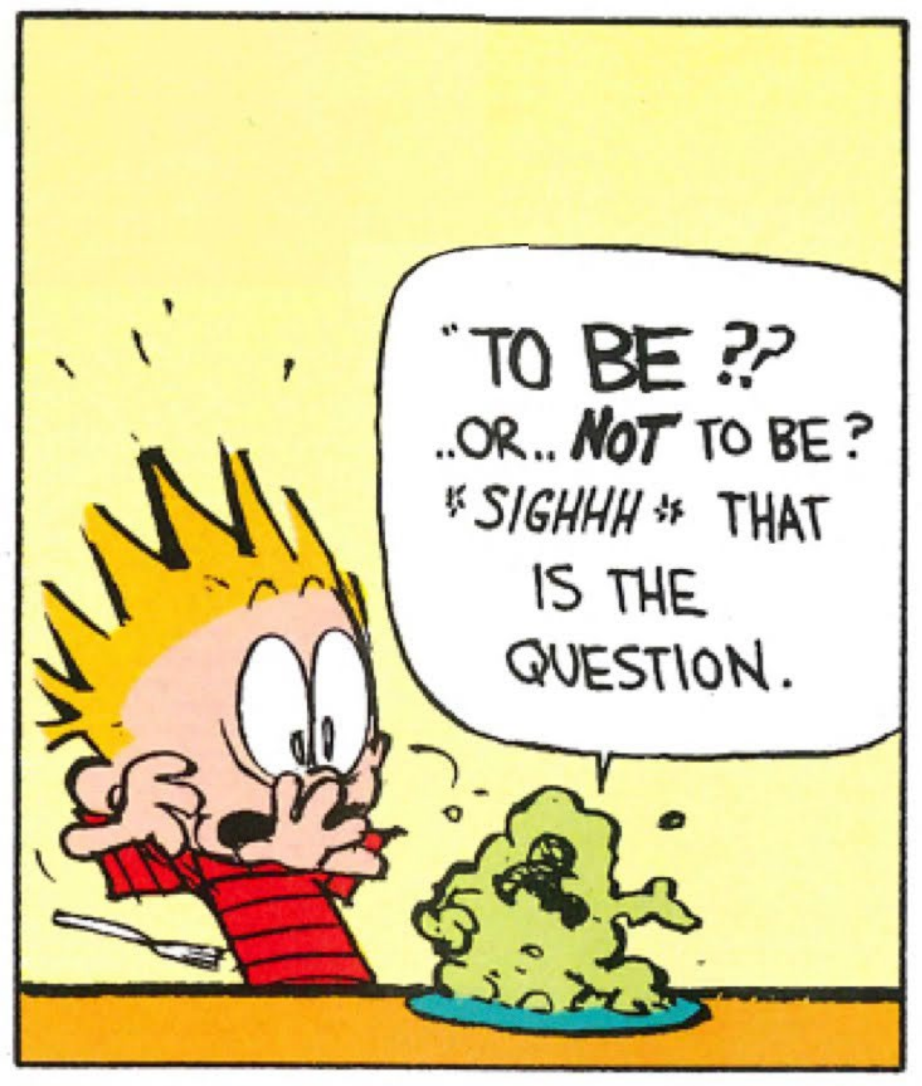
Structural equation modeling in food science

Quentin Read

USDA ARS SEA Statistician

August 1, 2023



Fill in the blank:

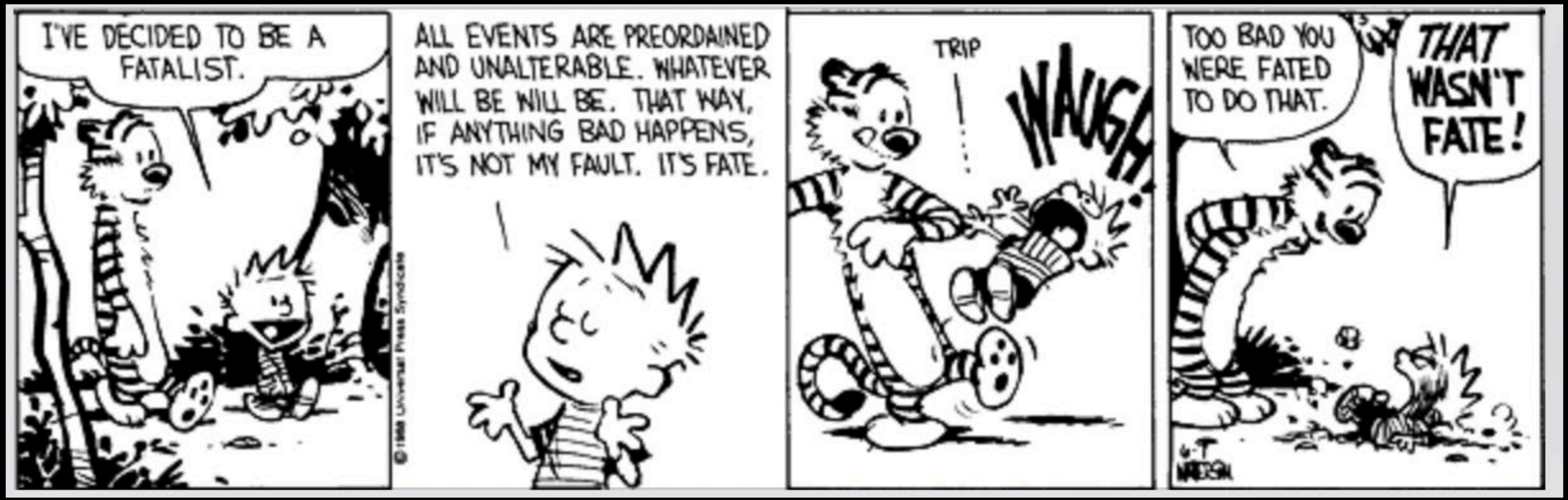
Correlation is not _______.

It is true that measuring correlation cannot prove causation

But today we will explore a method to infer causation in complex systems using data

We want to answer the big question:

WHY?

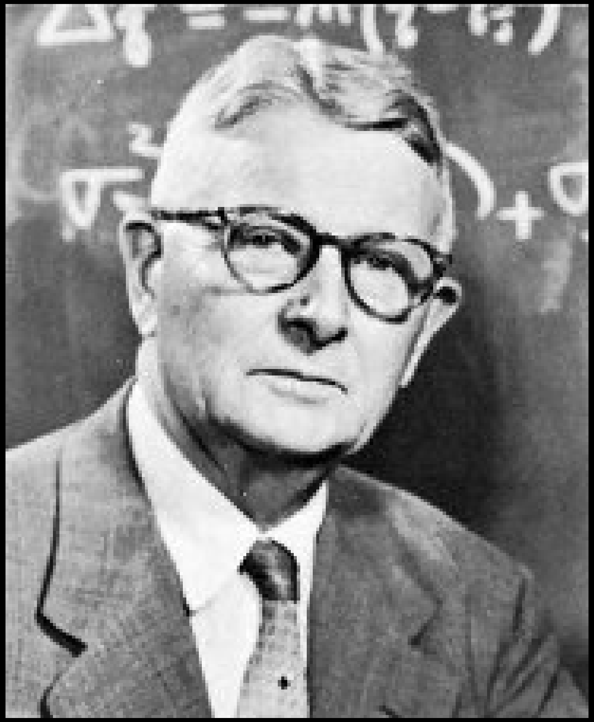


Everything causes everything else

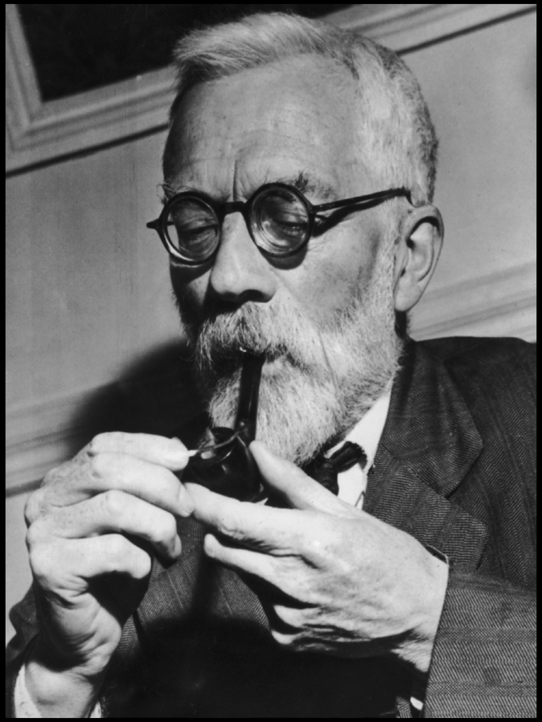
According to physics, causal influences can travel as fast as the speed of light

Instead of pretending we cannot make inferences about causes, we need to carefully describe how we think causes in our system are operating

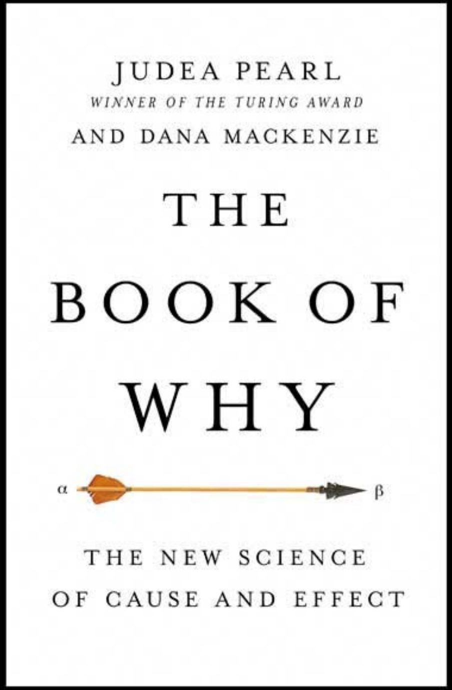
Very brief history of causal modeling



Sewall Wright
path analysis (1918)

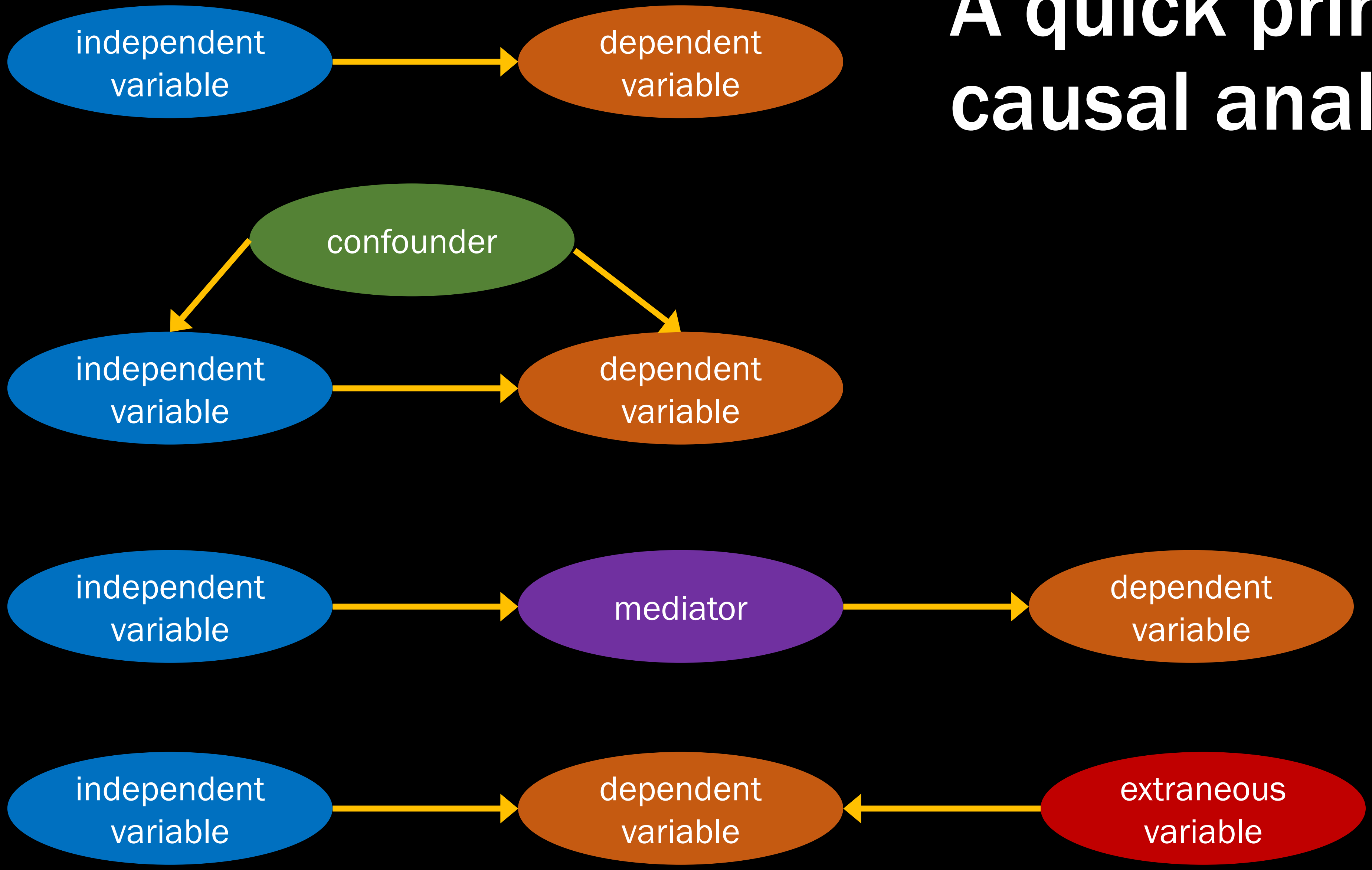


R. A. Fisher
stalled the development of causal models



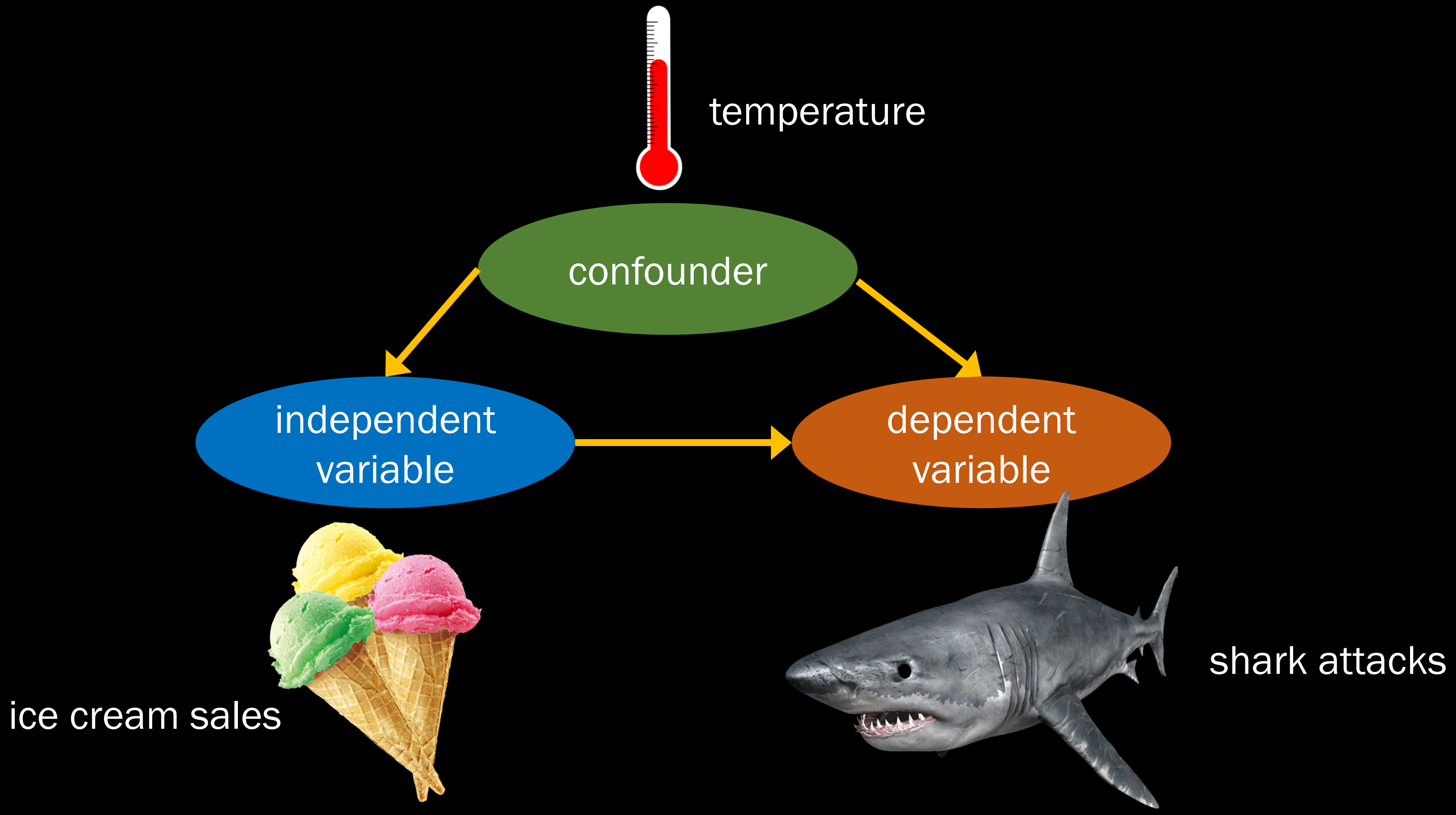
Judea Pearl
Use of DAGs for causal inference (2000-present)

A quick primer on causal analysis



Confounder

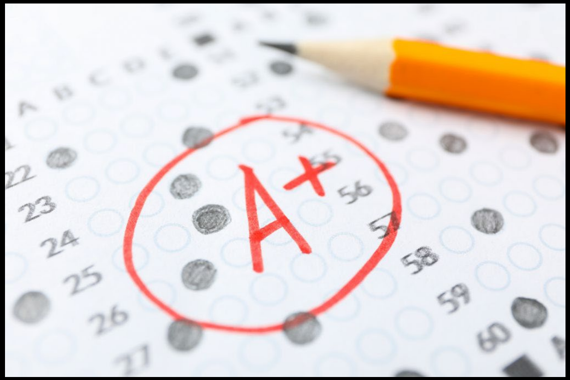
Influences both independent and dependent variable



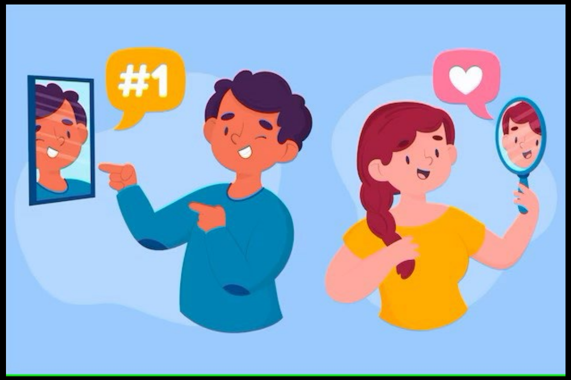
Mediator

Independent variable influences the mediator, which influences the dependent variable

independent variable → mediator → dependent variable



grades



self-esteem

happiness

Extraneous variable

Influences the dependent variable, but not through a path involving the independent variable

Will be part of the "noise" or residuals of the model

blood alcohol level

braking distance

weather conditions

independent variable → dependent variable ← extraneous variable

What the DAG?

The diagrams we have seen so far are called DAGs

Directed Acyclic Graph

Directed: arrows have directions on them: the direction of causal influence

Acyclic: No loops (if you start at A, there is no path that gets back to A)

Graph: A mathematical structure consisting of pairwise relationships between a set of objects

Complex DAGs can be built out of the main “building blocks” we just saw

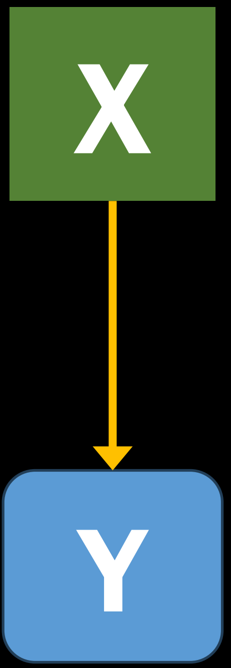
Structural equation models (SEM)

System of equations representing the relationships between variables (DAG in equation form)

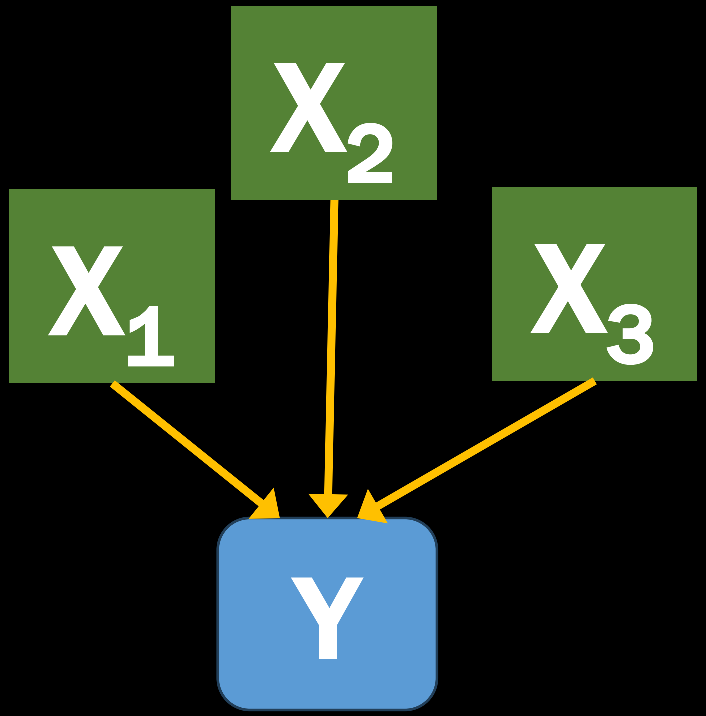
Variables in the model can be “x” variables in one of the equations and “y” variables in another

They can be caused/explained by one variable and cause/explain other variables in turn

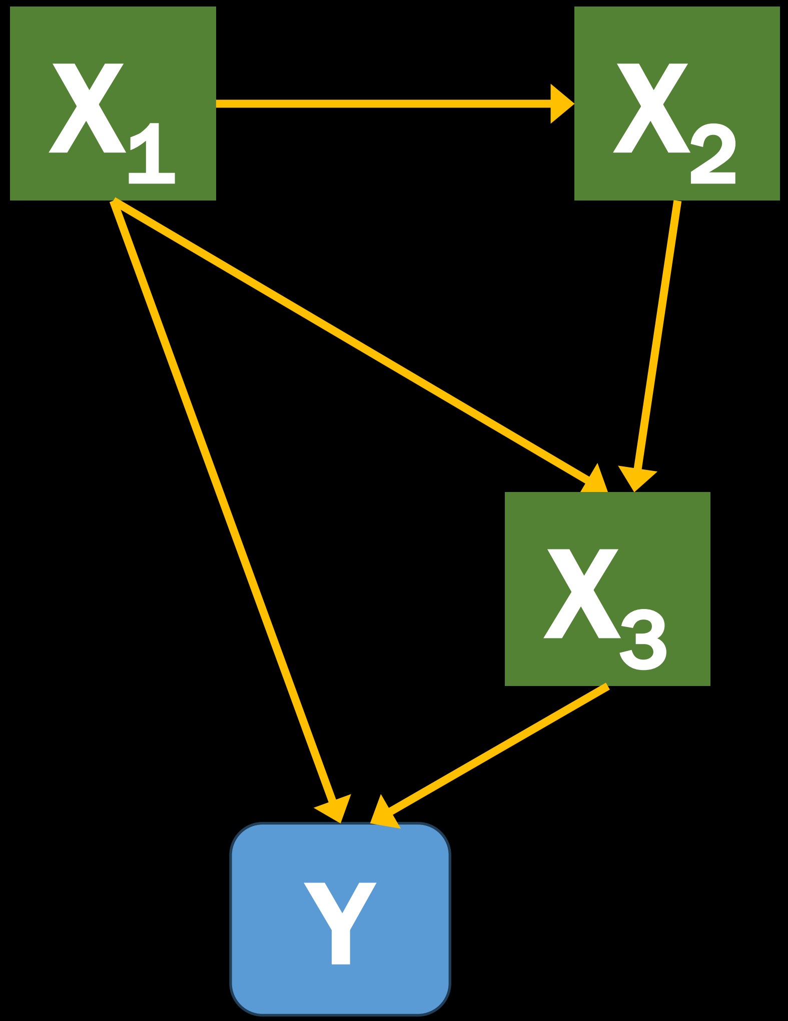
SEM versus simpler regression models



Simple linear regression



Multiple linear regression

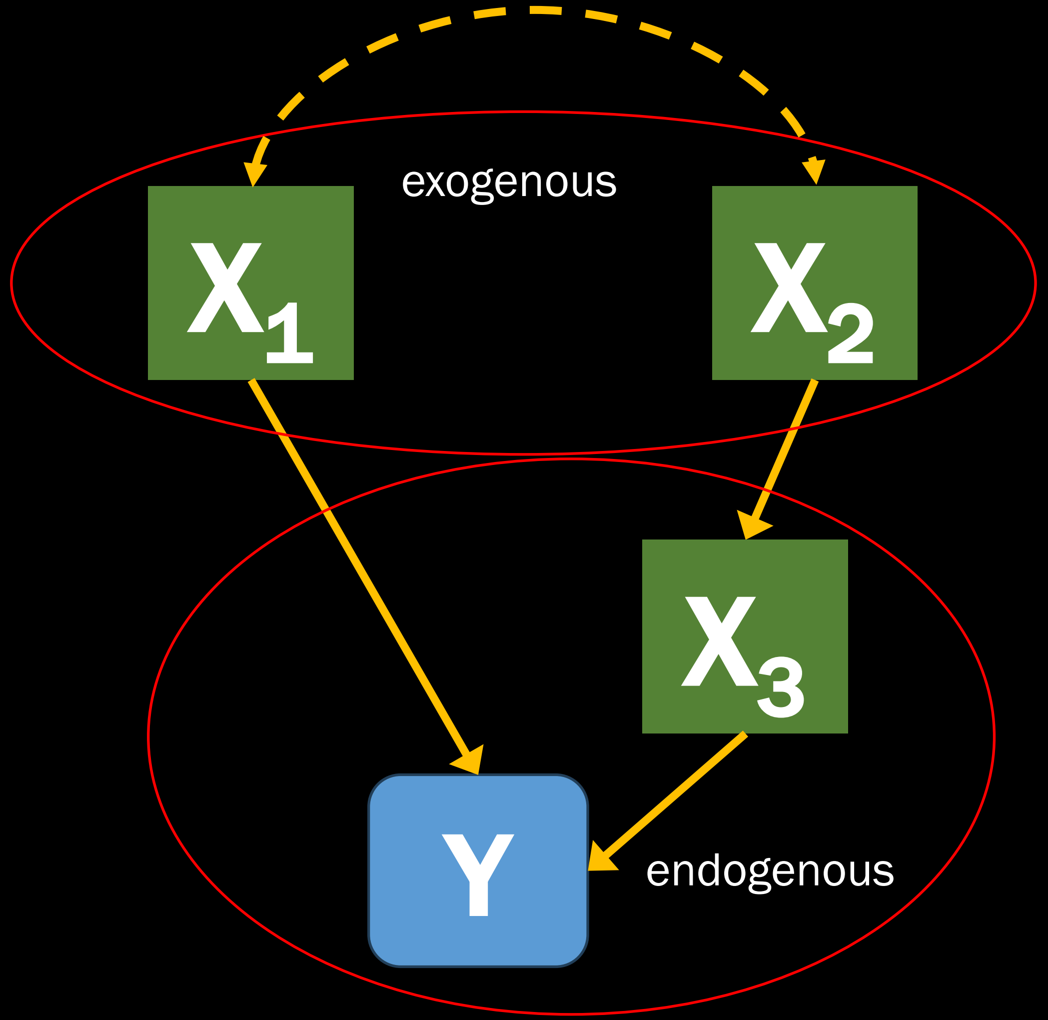


SEM

Endogenous and exogenous variables

Endogenous variables are at least partially explained by other variables in the model

Exogenous variables are not … but they may have residual covariance not otherwise explained by the model



Path analysis

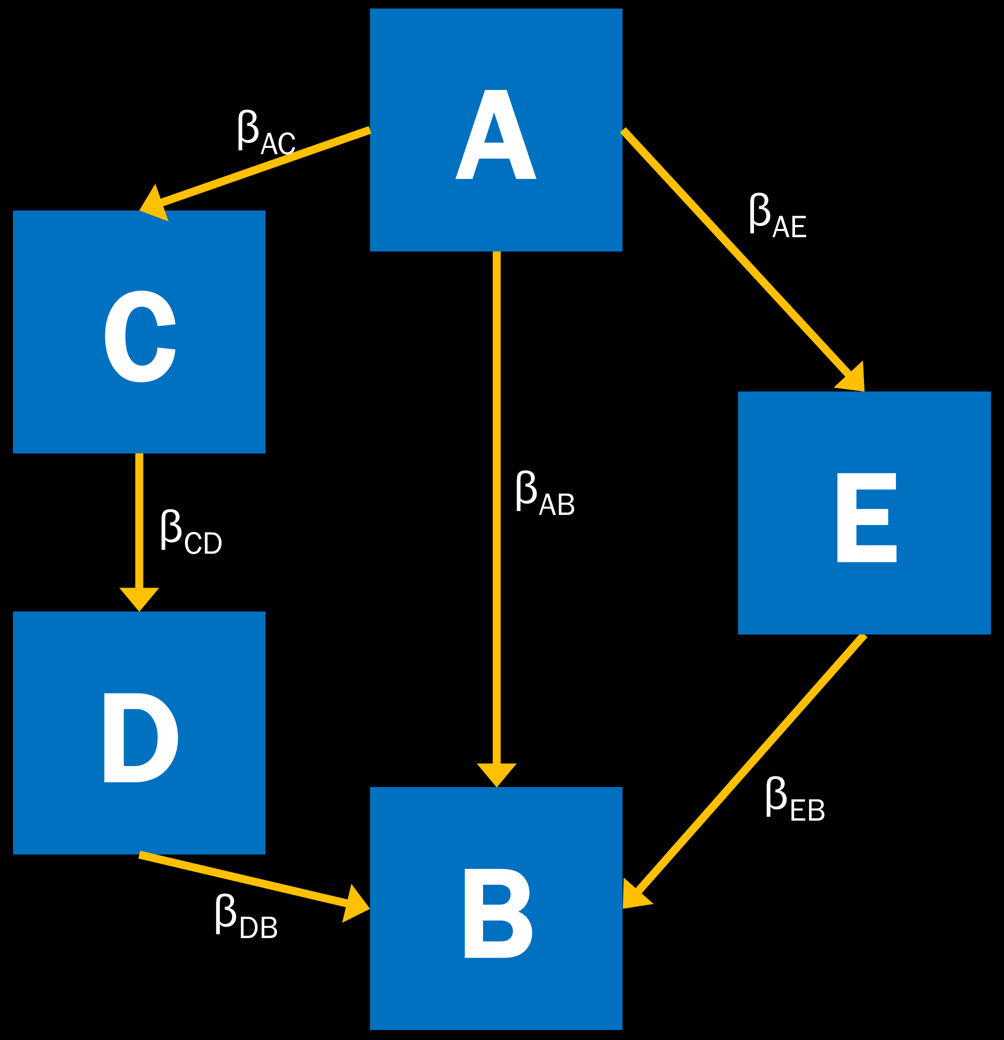
The net effect of a path from A to B is the product of all the effects on that path

$+ \times + = +$

$- \times - = +$

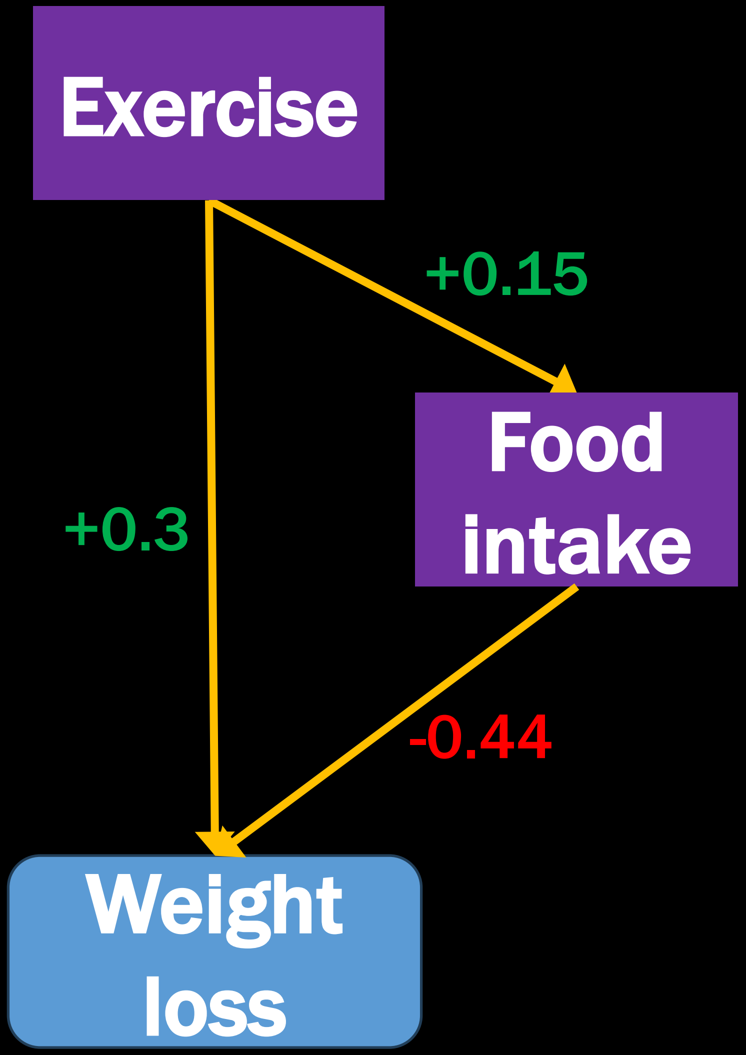
$+ \times - = -$

The total effect of A on B is the sum of all the paths that go from A to B, direct and indirect



$\beta_{AB} + \beta_{AC}\,\beta_{CD}\,\beta_{DB} + \beta_{AE}\,\beta_{EB}$

Path analysis example



What is the effect of exercise on weight loss?

Direct effect

+0.3

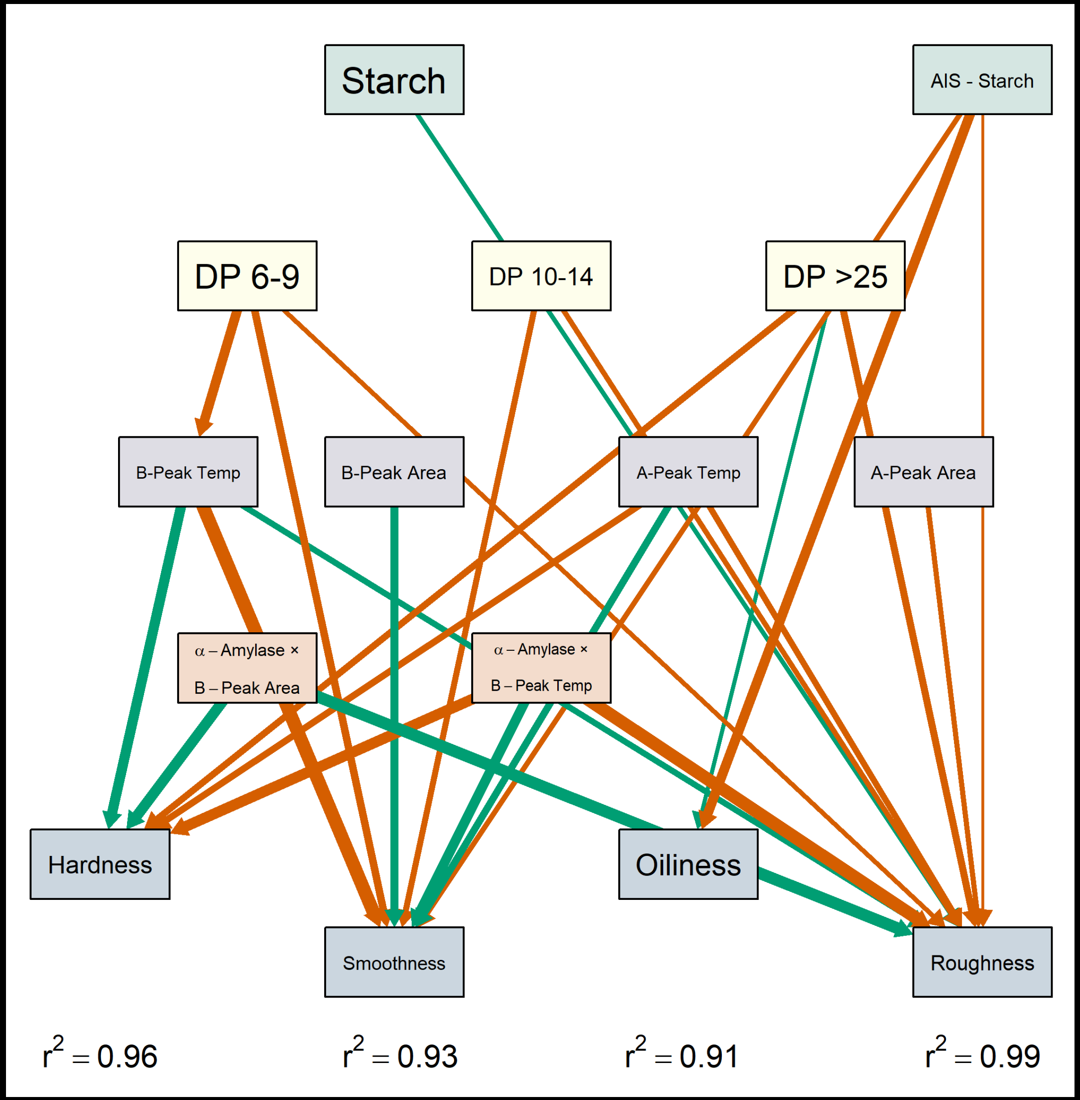
Indirect effect via food intake

+0.15 × -0.44 = -0.07

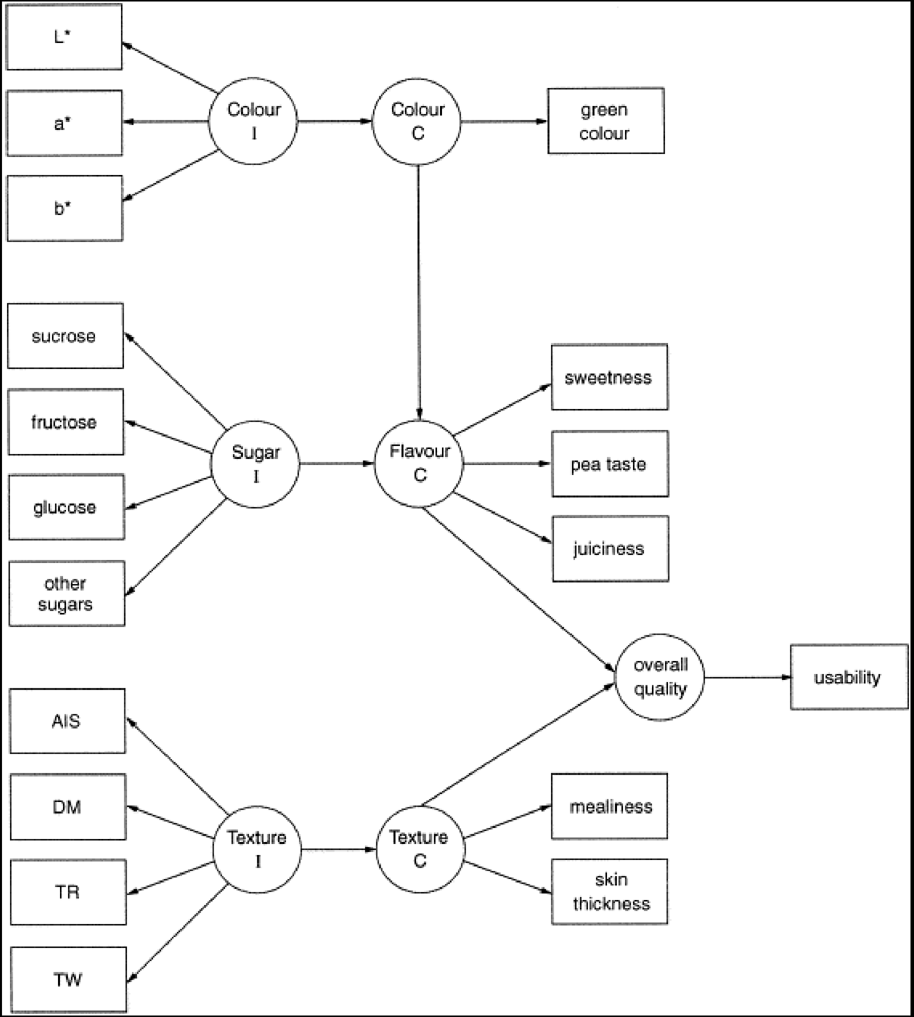
Total effect (direct effect + indirect effect)

+0.3 + (+0.15 × -0.44) = +0.23

SEM is ideal for complex systems like food chemistry!

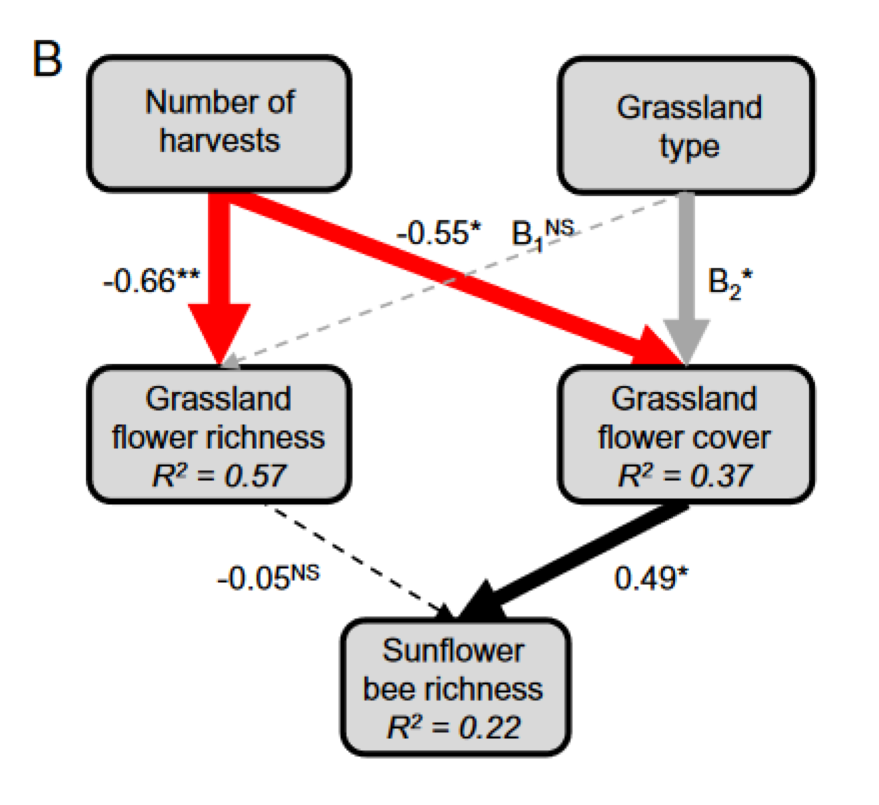


SEM can be used to estimate “latent variables”



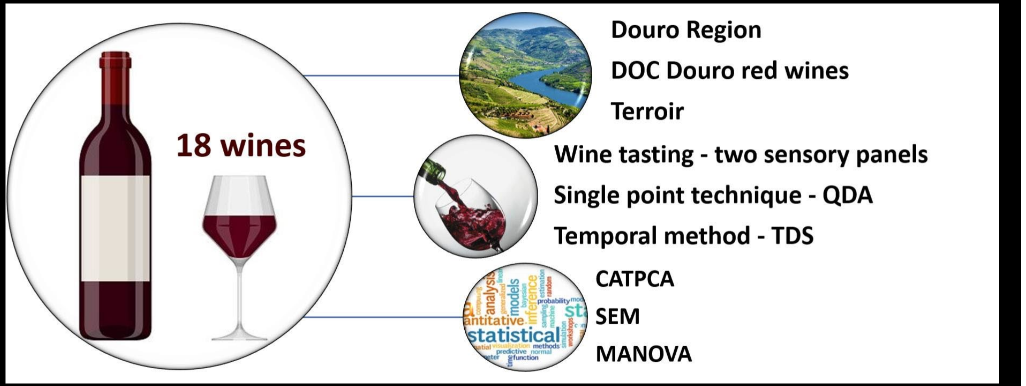
Bech et al. 2000, *Food Quality & Preference*

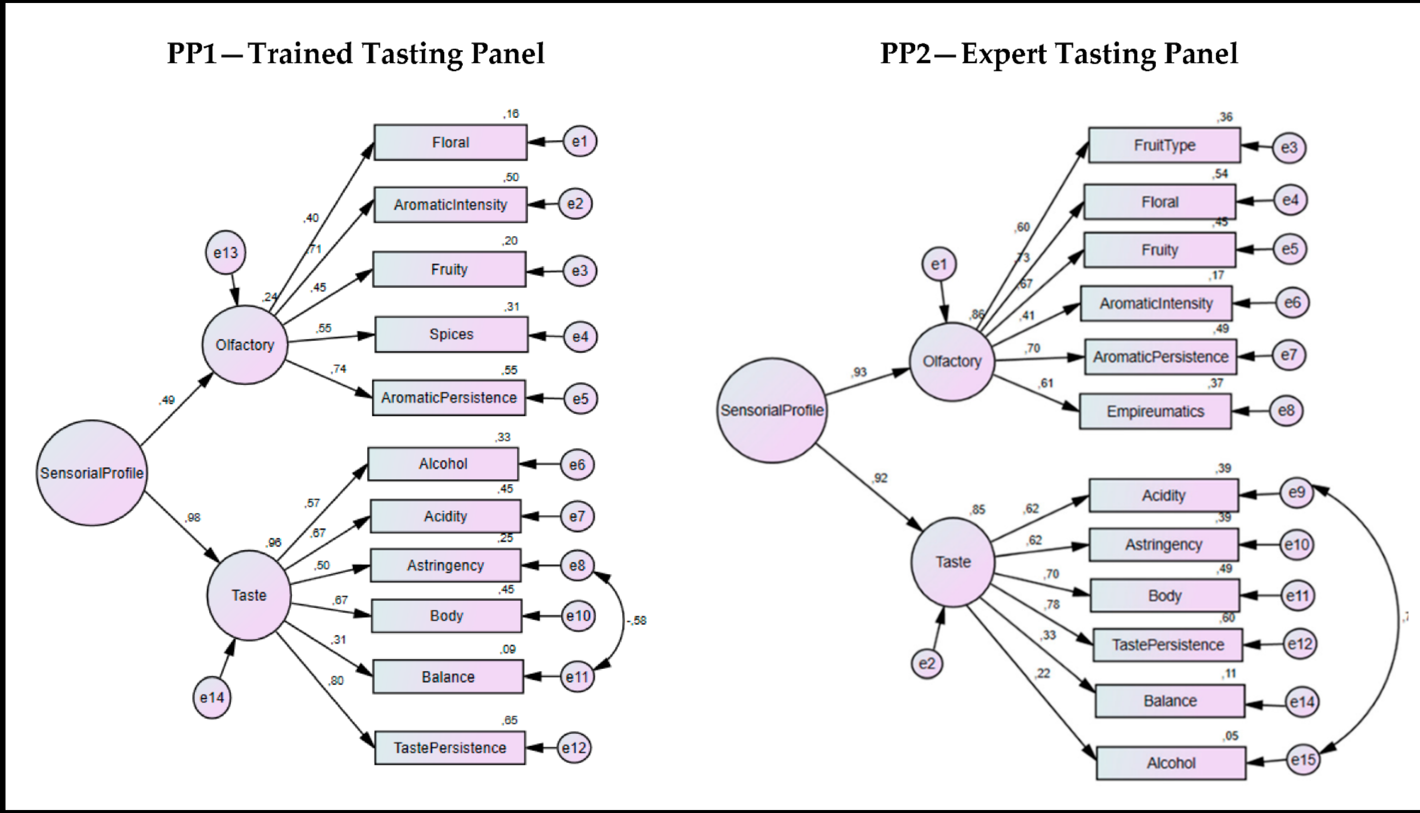
SEM is common in disciplines like agroecology



Scheper et al. 2023, *PNAS*

Is SEM underutilized in food science?





Correia, Amorim, & Vilela 2022, *Foods*

Warnings about SEM

SEM is just another linear model

- Relationships between variables are linear
- Assumes normally distributed errors

You must standardize variables to be able to compare effects among different units

SEM is not magic

We can use SEM to test between different causal hypotheses, but it cannot magically get causation from correlation

It can say which of several causal hypotheses is most consistent with the data

The trick is asking the right question (informed by theory and expert knowledge of how you think the system works)

SEM is not for fishing

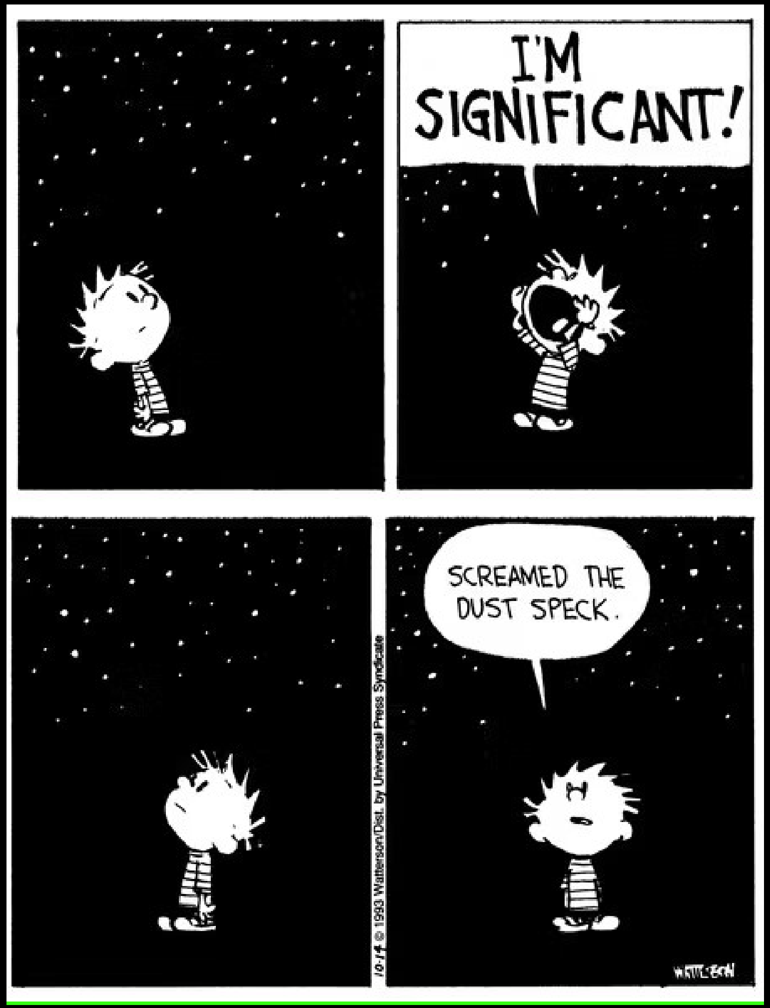
Not for blind model selection

> "throwing variables at a wall and seeing what's significant"

Use your expert knowledge and/or theory to construct some well-thought-out and defensible causal hypotheses

Build them manually by drawing the boxes and arrows of a DAG to help you visualize

Then translate them into a statistical model and see which ones have the most support from the data



Advanced SEM

Nonlinear relationships
- Quadratic, asymptotic, …

Non-normally distributed error
- Binary, categorical, …

Random effects
- Multilevel SEM

All of these are areas of active development!

SEM is not just for observational data

Some say if you have an experiment you do not need SEM

But many experiments have a lot of covariates, complex interactions, confounding influences … SEM can help with that!

Software

R packages: `lavaan`, `blavaan` (Bayesian), `piecewiseSEM` (random effects, nonlinear relationships), `ggdag` (drawing DAGs)

SAS: `PROC CALIS`

DAGitty.net (drawing and analyzing DAGs) usda-ree-ars.github.io/SEAStats

quentinread.com

quentin.read@usda.gov

Scan this QR code to visit the SEAStats page with stats lessons and FAQs!