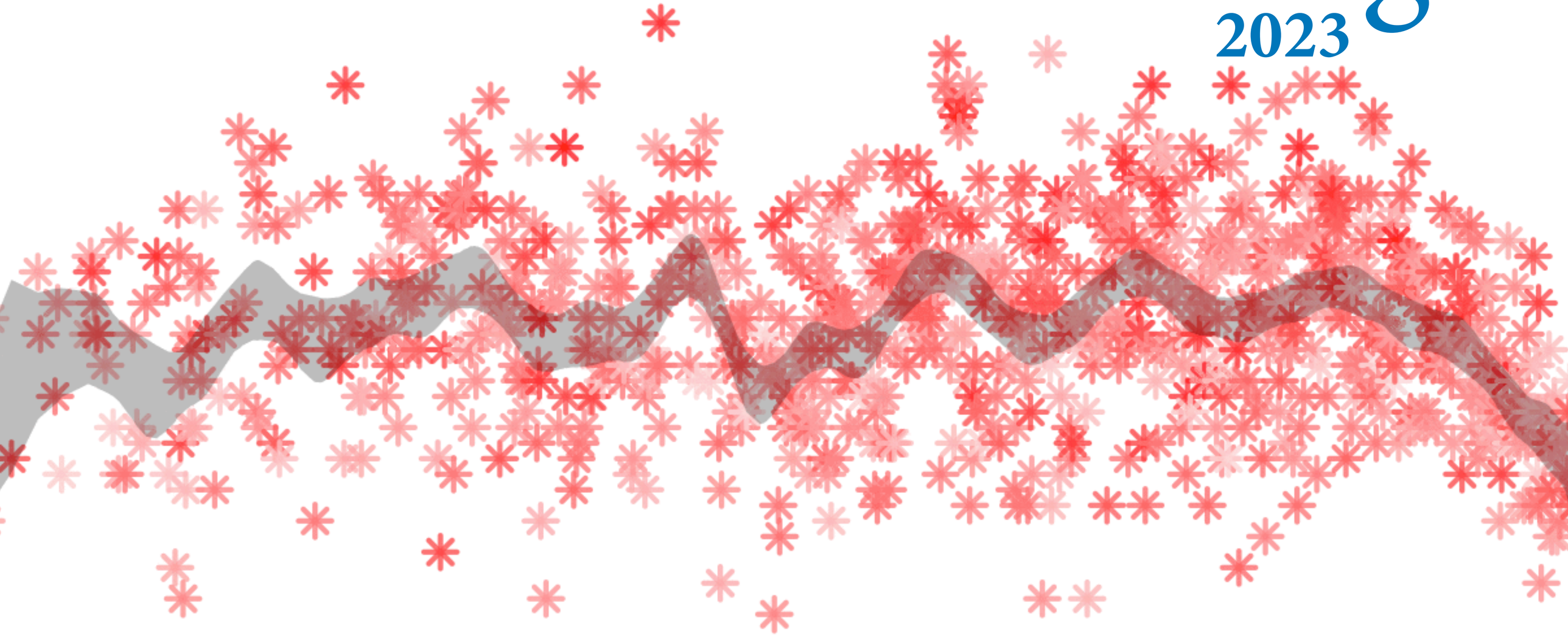
Statistical Rethinking

2023



1. The Golem of Prague

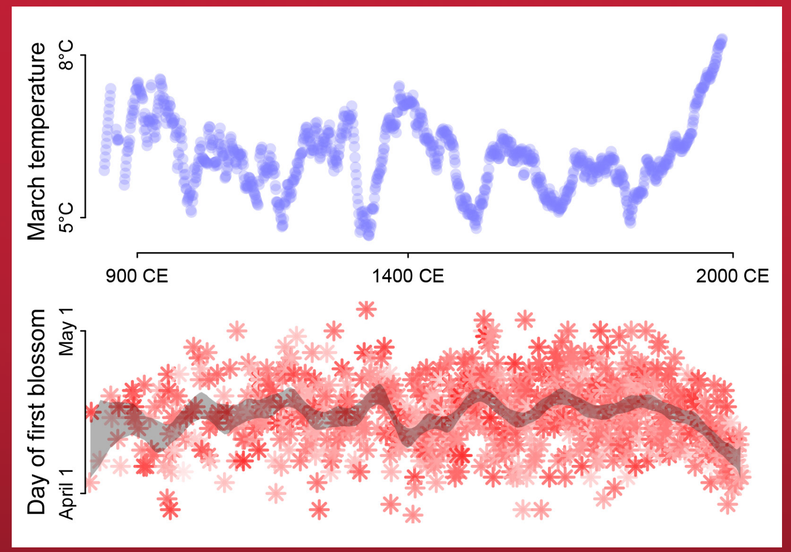
Second Edition

Texts in Statistical Science

Statistical Rethinking

A Bayesian Course with Examples in R and Stan

THIRD EDITION



Richard McElreath

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Statistical Rethinking

McElreath

CRC

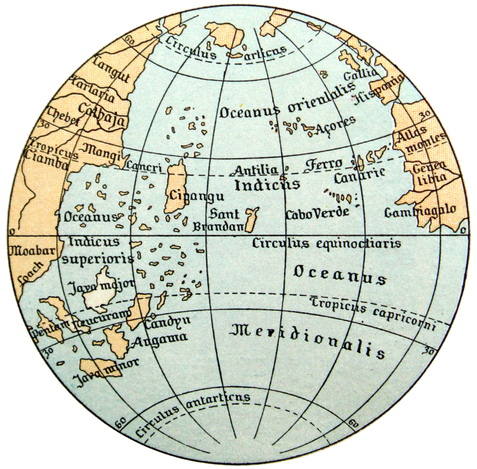


FIGURE 2.1. Illustration of Martin Behaim's 1492 globe, showing the small world that Colombo anticipated. Europe lies on the right-hand side. Asia lies on the left. The big island labeled "Cipangu" is Japan.

worlds, using the strengths of qualitative insight to criticize small world quantitative insight. The precision and transparency of the small world is powerful and essential. But it is very difficult to incorporate all of our scientific knowledge and expertise into the small world. So often we realize how bad a model is only when we see how badly it behaves in the broader scientific context.

In this chapter, you will begin to build Bayesian models. Almost all of the work in this chapter takes place in the small world. We'll start with a simple goal (estimand): *What proportion of the Earth's surface is covered by water?* This is a descriptive estimand, but it still requires causal assumptions about how the sample arises. Once those assumptions are in place, we'll use Bayesian inference to produce an estimate.

Rethinking: Fast and frugal in the large world. The natural world is complex, as trying to do science serves to remind us. Yet everything from the humble tick to the industrious squirrel to the idle sloth manages to frequently make adaptive decisions. But it's a good bet that most animals are not Bayesian, if only because being Bayesian is expensive and depends upon having a good model. Instead, animals use various heuristics that are fit to their environments, past or present. These heuristics take adaptive shortcuts and so may outperform a rigorous Bayesian analysis, once costs of information gathering and processing (and overfitting, Chapter 6) are taken into account.[35] Once you already know which information to ignore or attend to, being fully Bayesian is a waste. It's neither necessary nor sufficient for making good decisions, as real animals demonstrate. But for human animals, Bayesian analysis provides a general way to discover relevant information and process it logically.

2.1. The garden of forking data

Our goal in this section is to build Bayesian inference up from humble beginnings, so there is no superstition about it. Bayesian inference is really just counting and comparing of possibilities. We make causal assumptions, we enumerate the implications of those assump-

Changes and updates

Fewer examples, more depth

Detailed workflow, testing

Interventions, post-stratification

Foreground measurement, missing

Sensitivity analysis

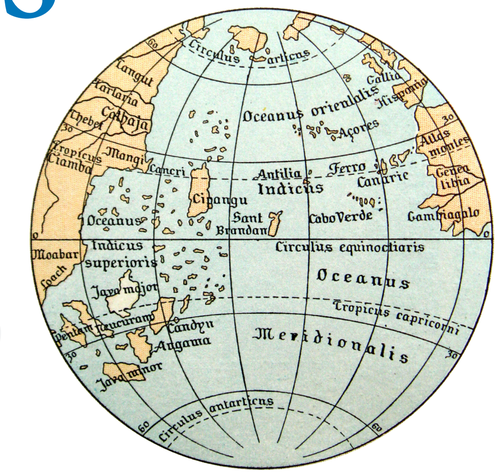


FIGURE 2.1. Illustration of Martin Behaim's 1492 globe, showing the small world that Colombo anticipated. Europe lies on the right-hand side. Asia lies on the left. The big island labeled "Cipangu" is Japan.

worlds, using the strengths of qualitative insight to criticize small world quantitative insight. The precision and transparency of the small world is powerful and essential. But it is very difficult to incorporate all of our scientific knowledge and expertise into the small world. So often we realize how bad a model is only when we see how badly it behaves in the broader scientific context.

In this chapter, you will begin to build Bayesian models. Almost all of the work in this chapter takes place in the small world. We'll start with a simple goal (estimand): *What proportion of the Earth's surface is covered by water?* This is a descriptive estimand, but it still requires causal assumptions about how the sample arises. Once those assumptions are in place, we'll use Bayesian inference to produce an estimate.

Rethinking: Fast and frugal in the large world. The natural world is complex, as trying to do science serves to remind us. Yet everything from the humble tick to the industrious squirrel to the idle sloth manages to frequently make adaptive decisions. But it's a good bet that most animals are not Bayesian, if only because being Bayesian is expensive and depends upon having a good model. Instead, animals use various heuristics that are fit to their environments, past or present. These heuristics take adaptive shortcuts and so may outperform a rigorous Bayesian analysis, once costs of information gathering and processing (and overfitting, Chapter 6) are taken into account.[35] Once you already know which information to ignore or attend to, being fully Bayesian is a waste. It's neither necessary nor sufficient for making good decisions, as real animals demonstrate. But for human animals, Bayesian analysis provides a general way to discover relevant information and process it logically.

DAGS

DAGS
GOLEMS

BAGS
GOLEMS
OWLS

DAGS
GOLEMS
OWLS


BAYES
FREQUENTISM


CAUSAL
INFERENCE
imgflip.com

Science Before Statistics

For **statistical models** to produce scientific insight, they require additional **scientific (causal) models**

The **reasons** for a statistical analysis are not found in the data themselves, but rather in the **causes** of the data

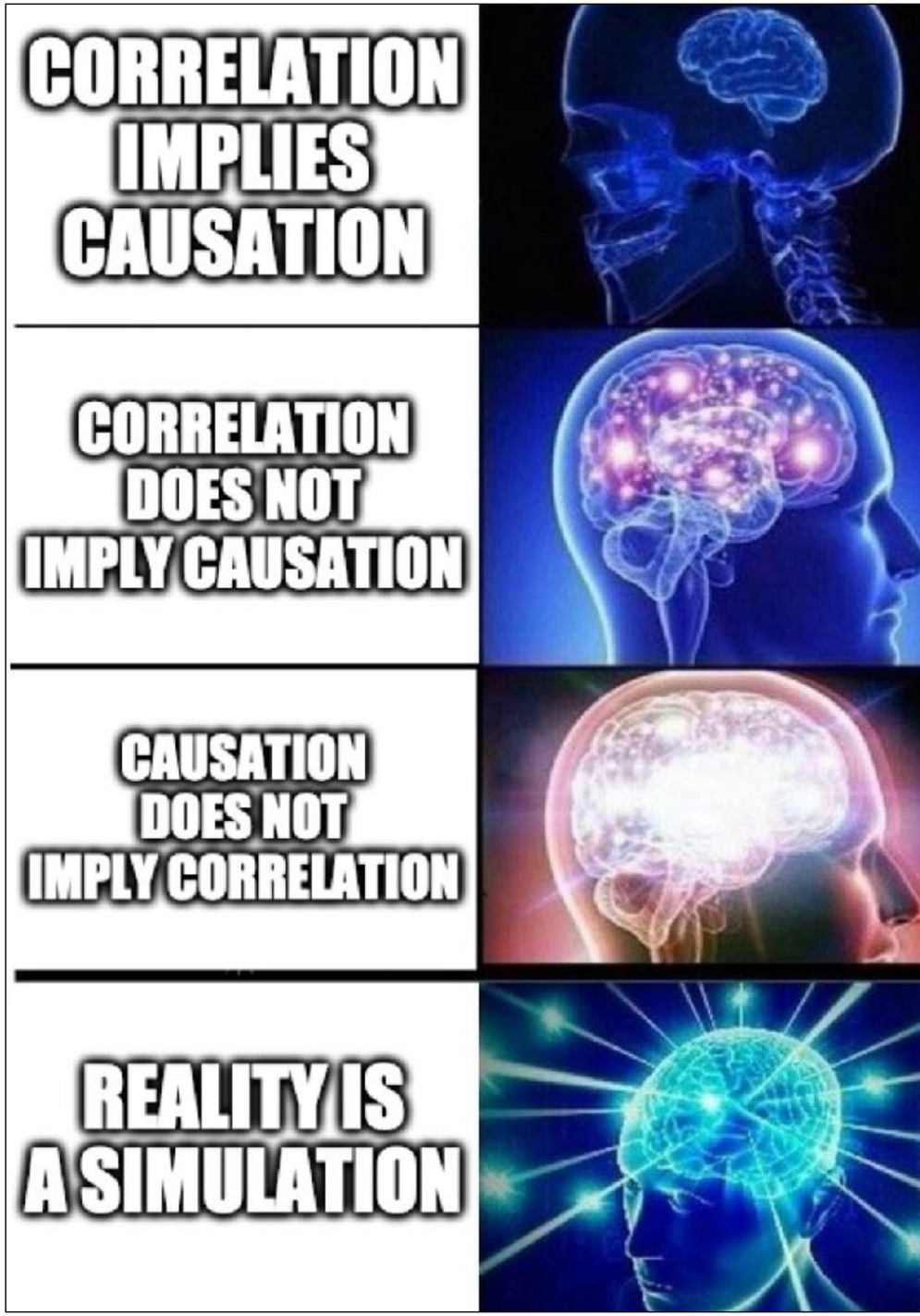
The **causes** of the data cannot be extracted from the data alone. No causes in; no causes out.

What is Causal Inference?

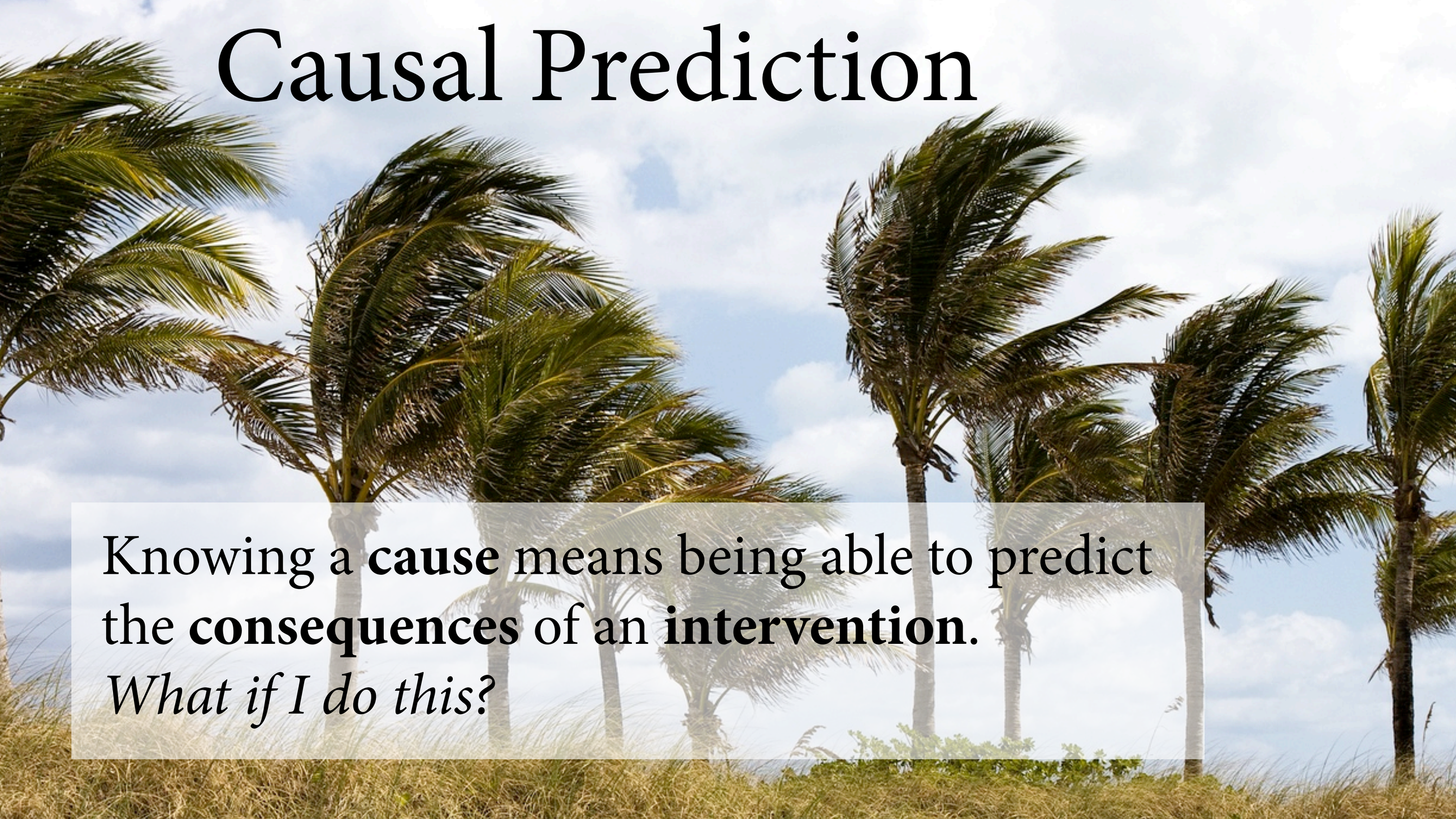
More than **association** between variables

Causal inference is **prediction** of intervention

Causal inference is **imputation** of missing observations

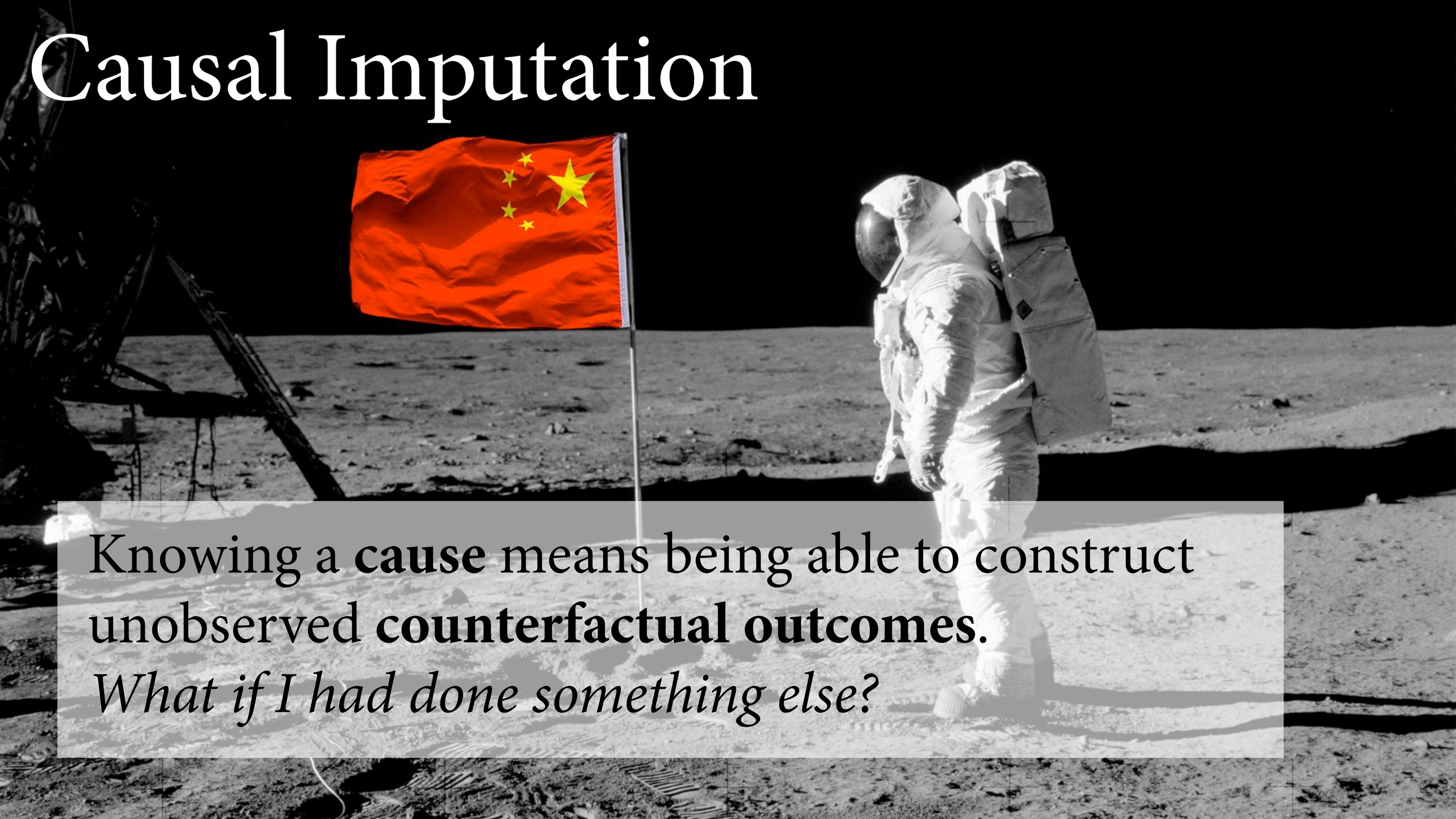


Causal Prediction

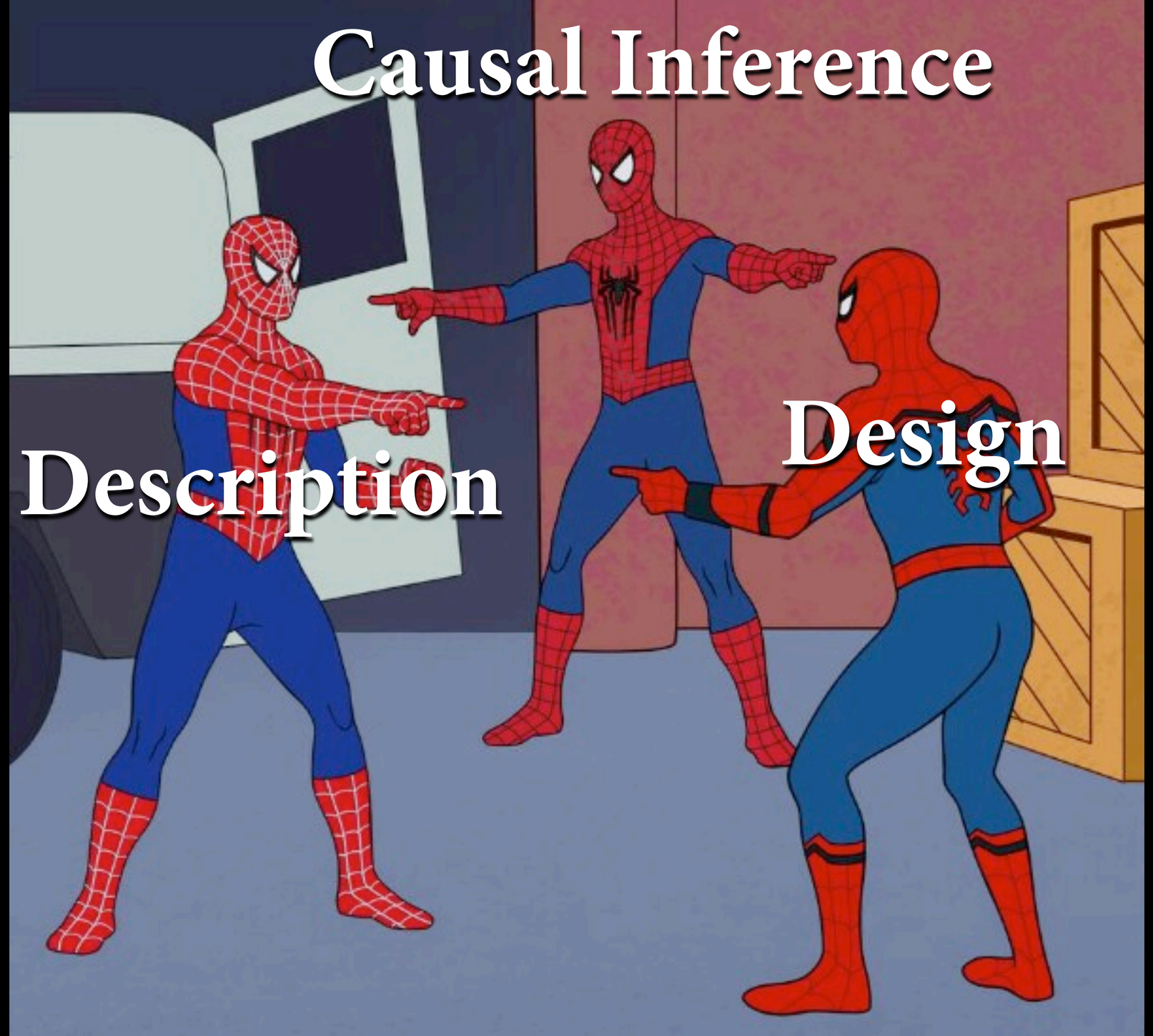


Knowing a **cause** means being able to predict the **consequences** of an **intervention**.
What if I do this?

Causal Imputation



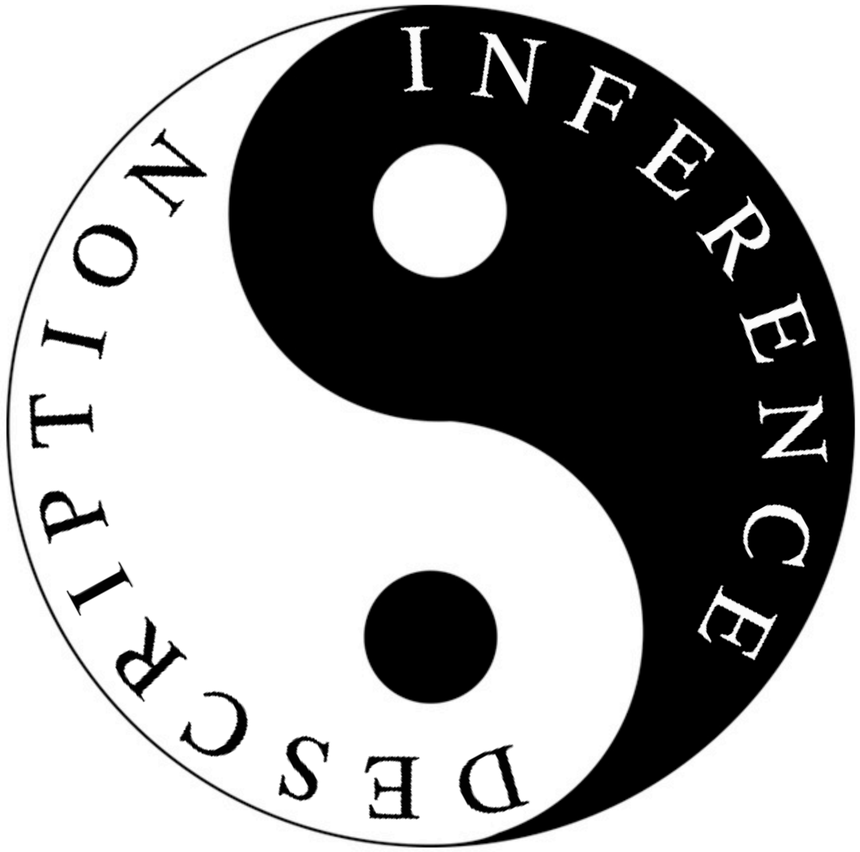
Knowing a **cause** means being able to construct unobserved **counterfactual outcomes**.
What if I had done something else?


Causal Inference
Description
Design

Causes Are Not Optional

Even when goal is **descriptive**, need causal model

The **sample** differs from the **population**; describing the population requires causal thinking about why

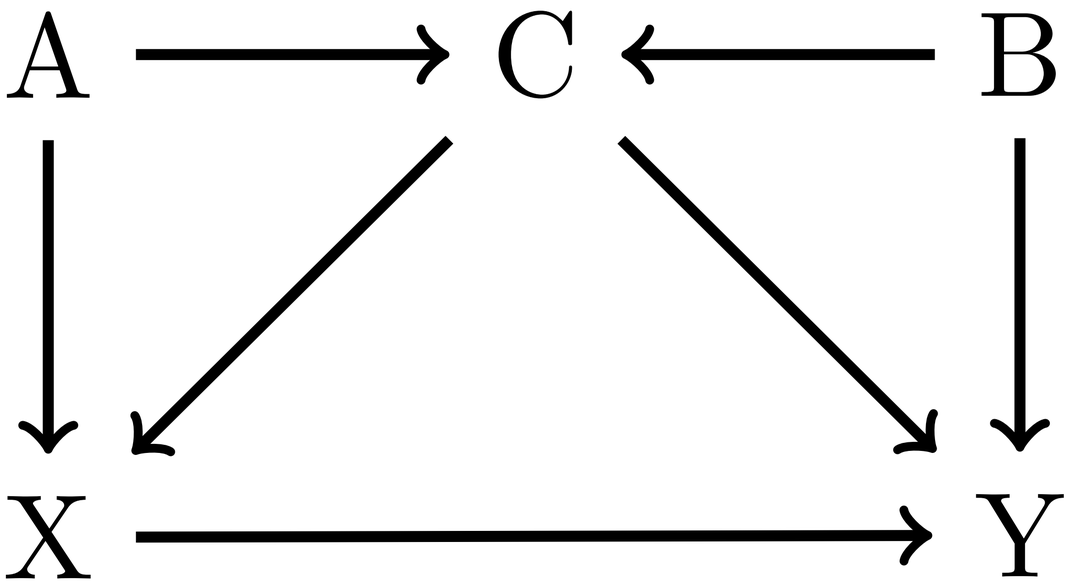


DAGs

Directed Acyclic Graphs

Heuristic causal models

Clarify scientific thinking

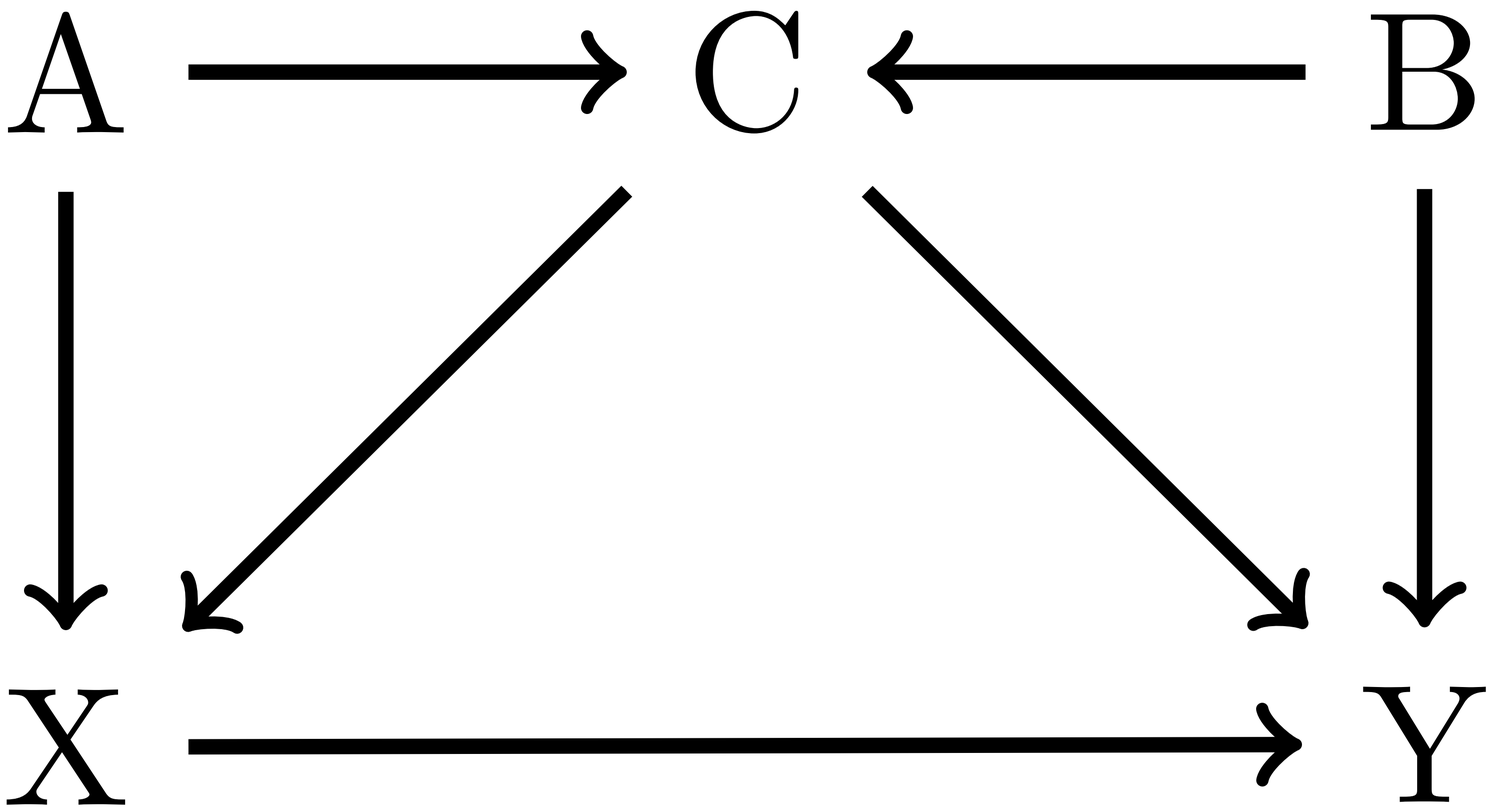


Analyze to deduce appropriate statistical models

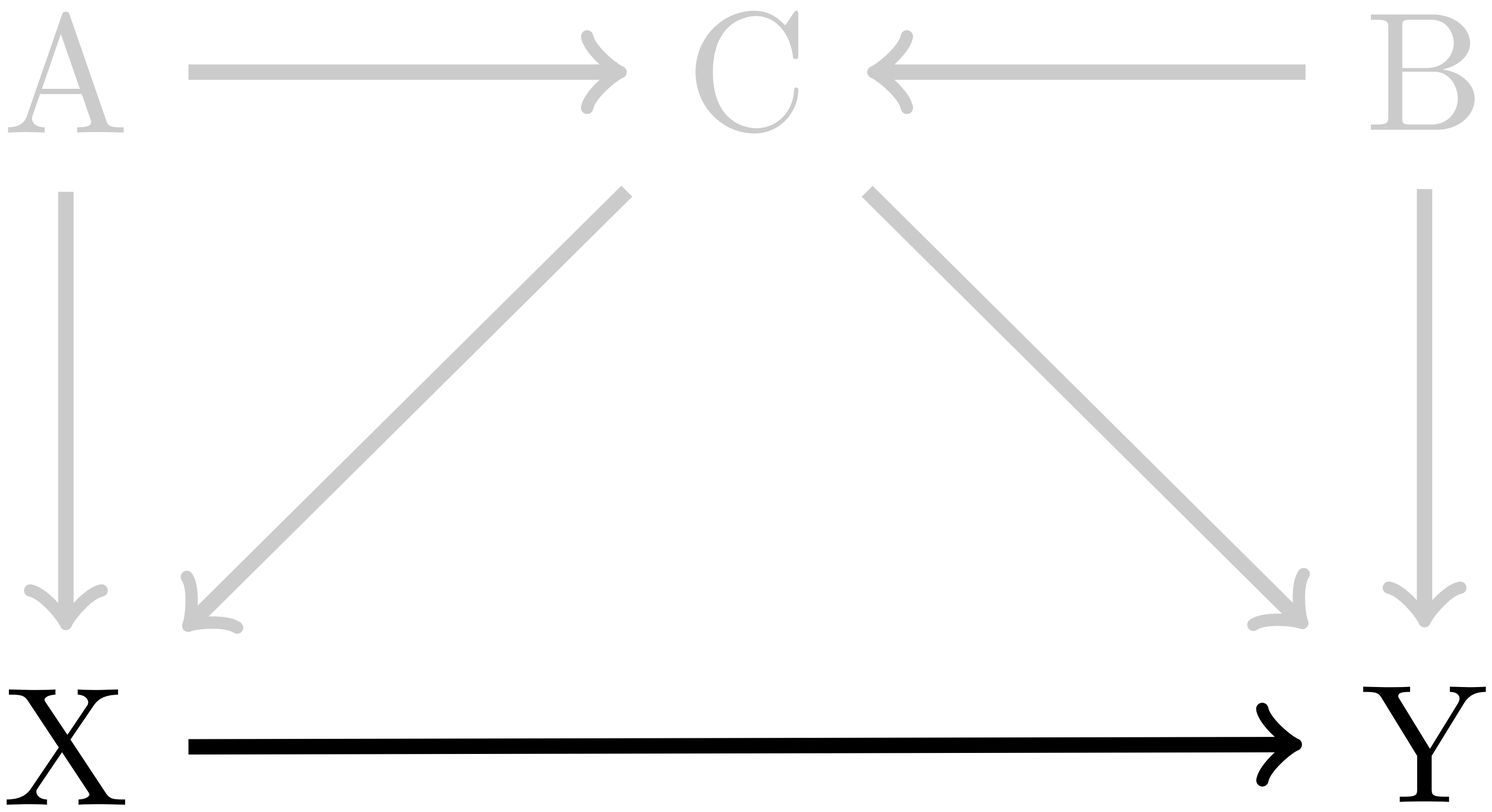
"What can we decide, without additional assumptions?"

Gateway to scientific modeling

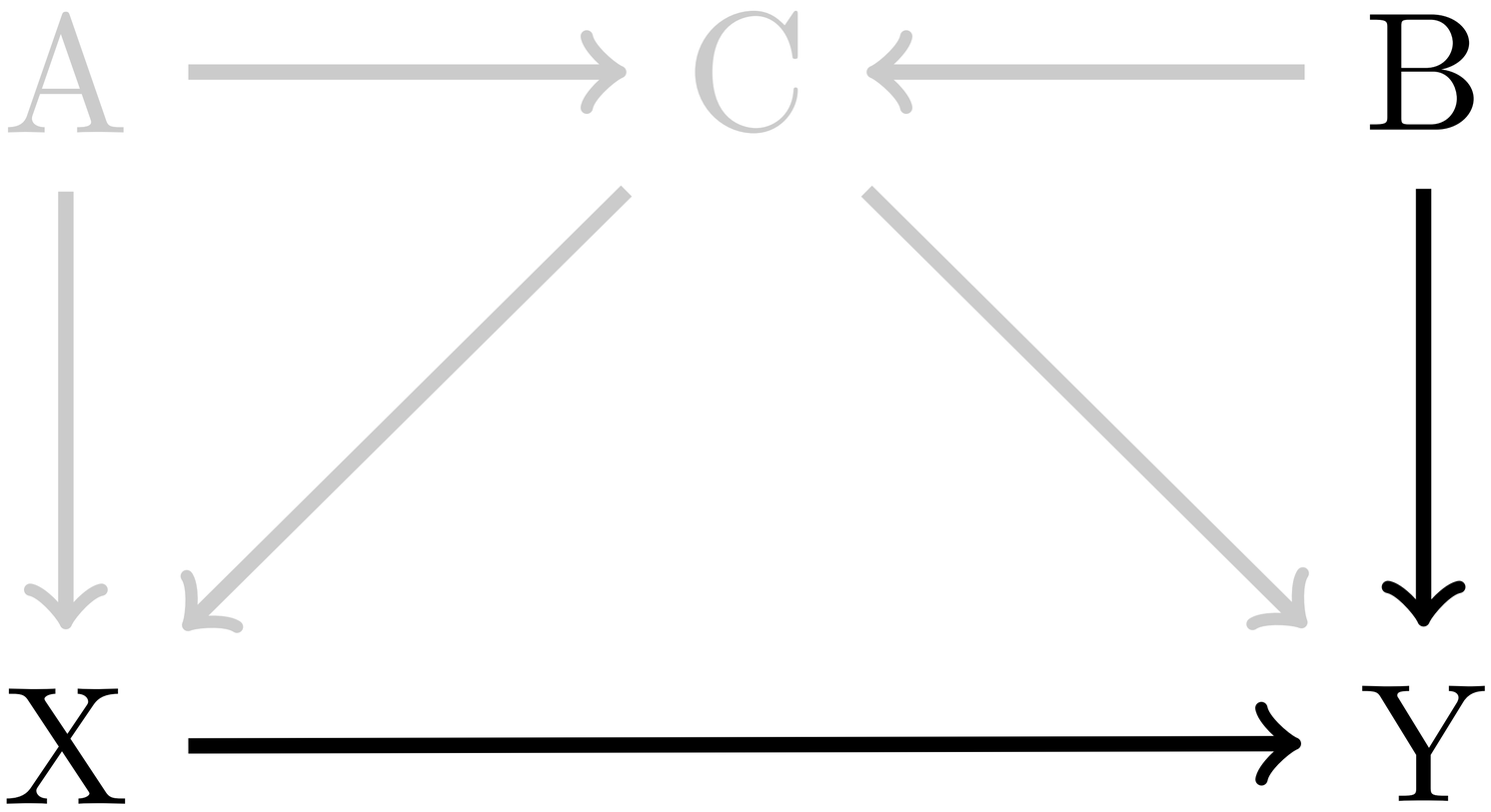
DAGs



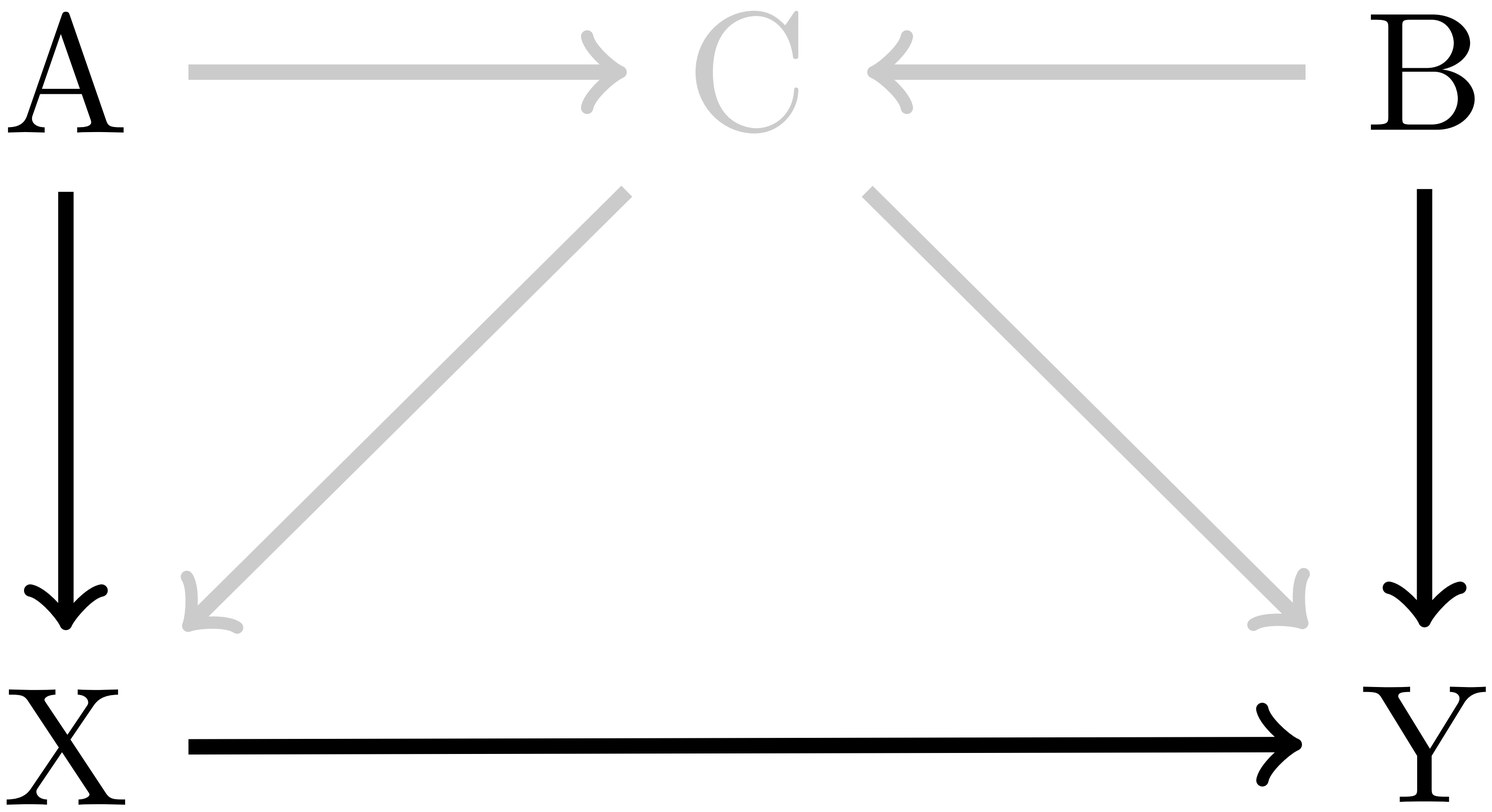
DAGs



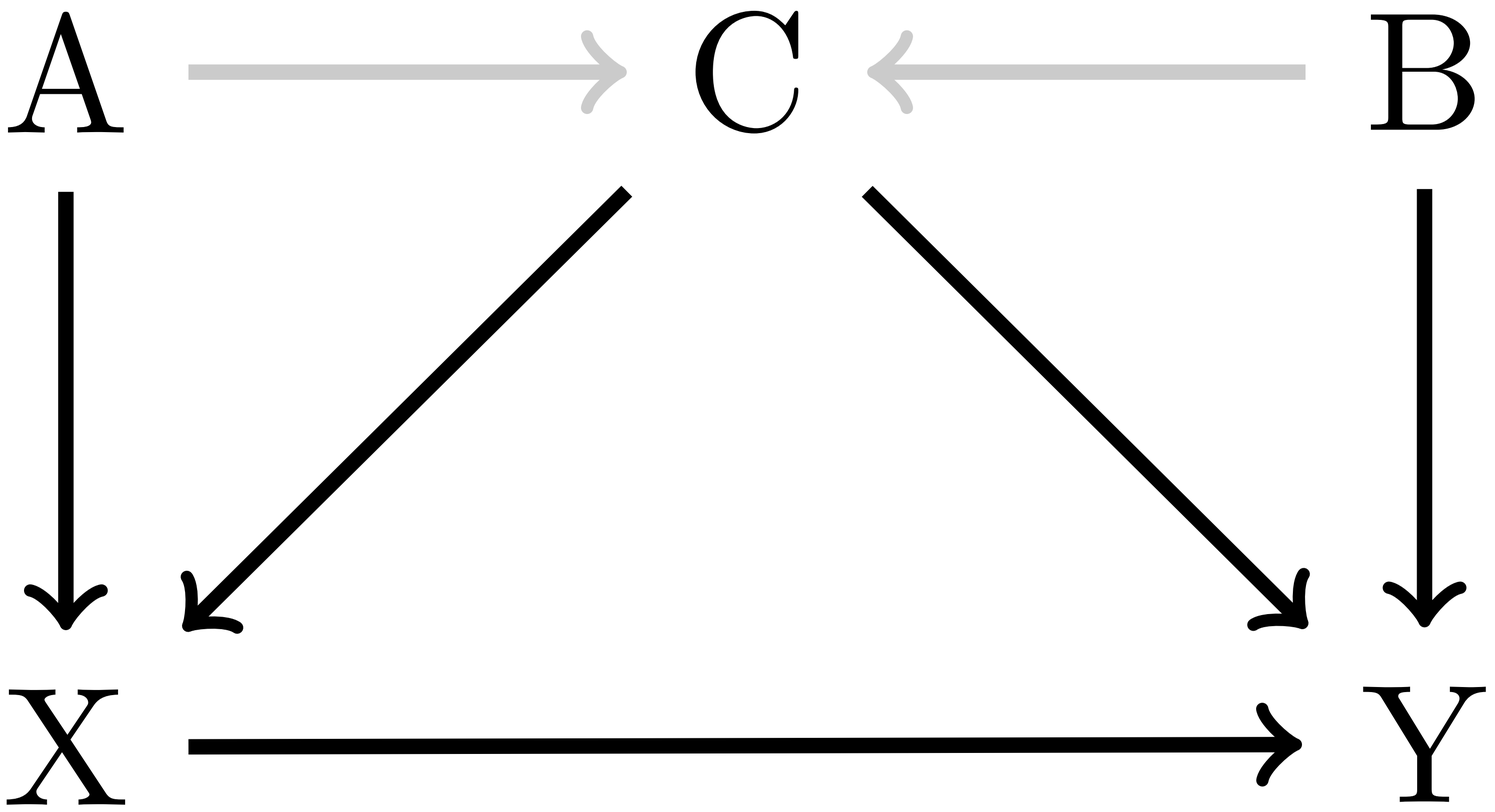
DAGs



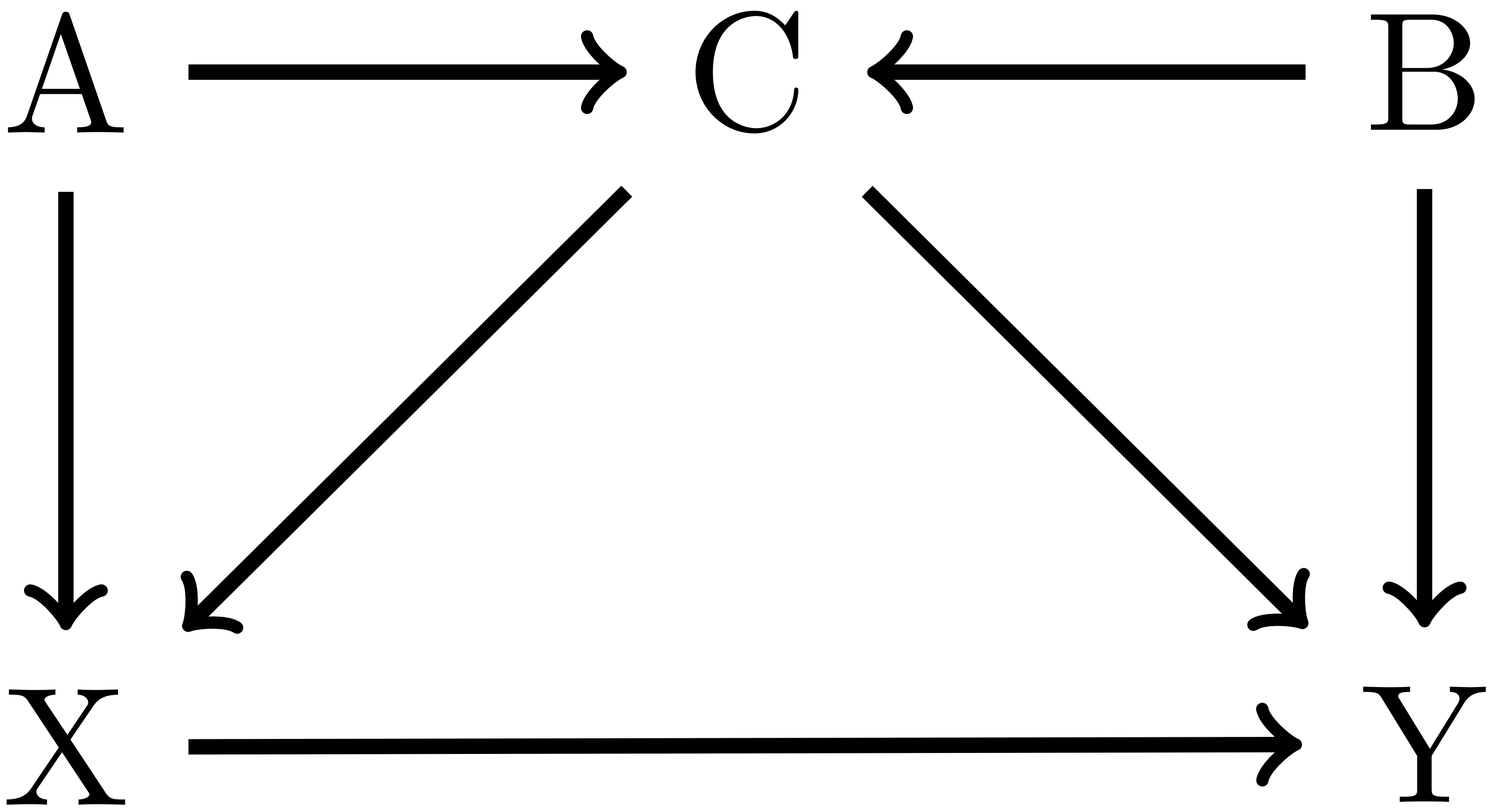
DAGs



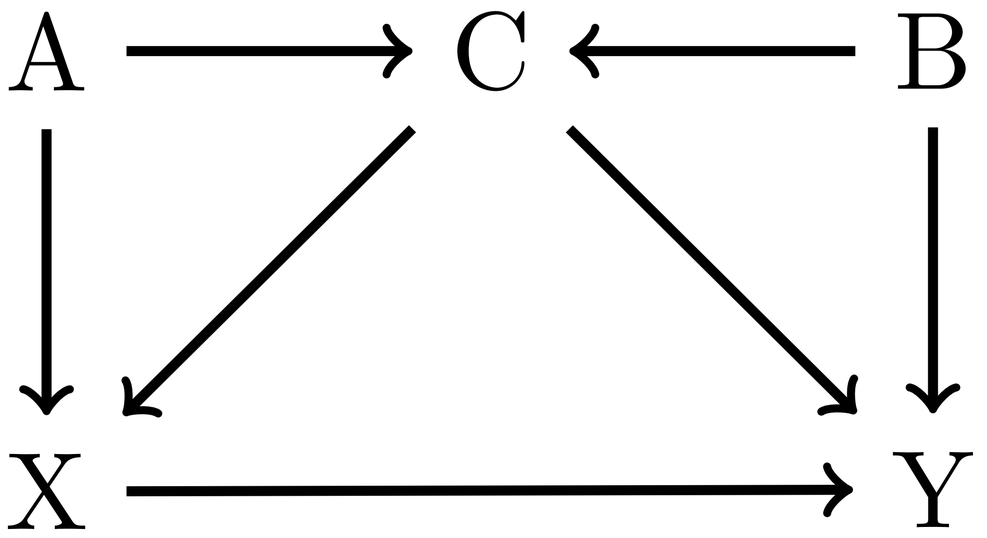
DAGs



DAGs



DAGs



Different queries, different models

Which control variables?

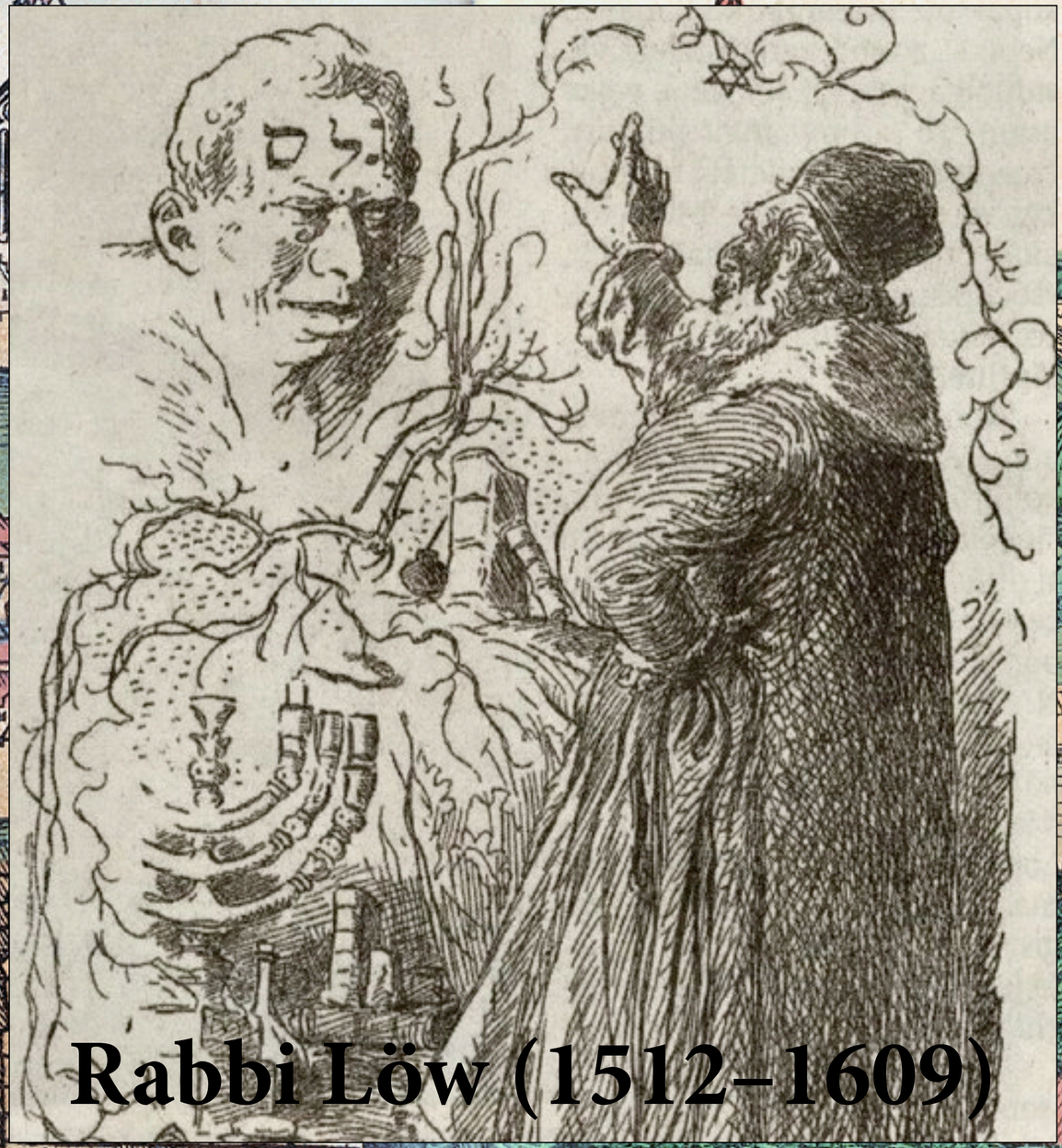
Absolute not safe to add everything — bad controls

How to test/refine the causal model?

DAGs are intuition pumps: get head out of data, into science

DAGS
GOLEMS

Prague 16th century



Rabbi Löw (1512–1609)

Golems

Clay robots

Powerful

No wisdom or foresight

Dangerous

“Breath of Bones: A Tale of the Golem” (2014)

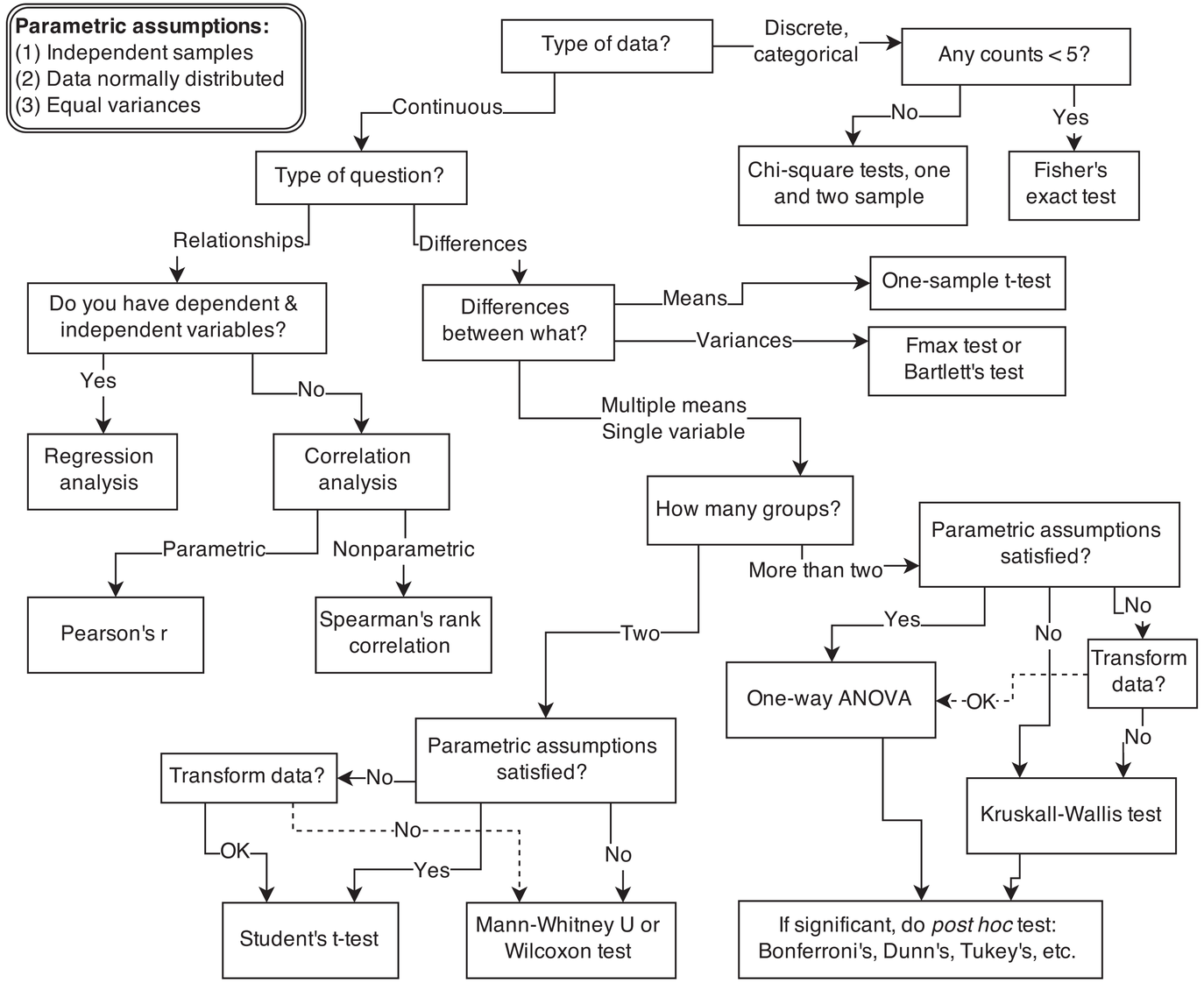
Statistical Models

Clay robots

Powerful

No wisdom or foresight

Dangerous



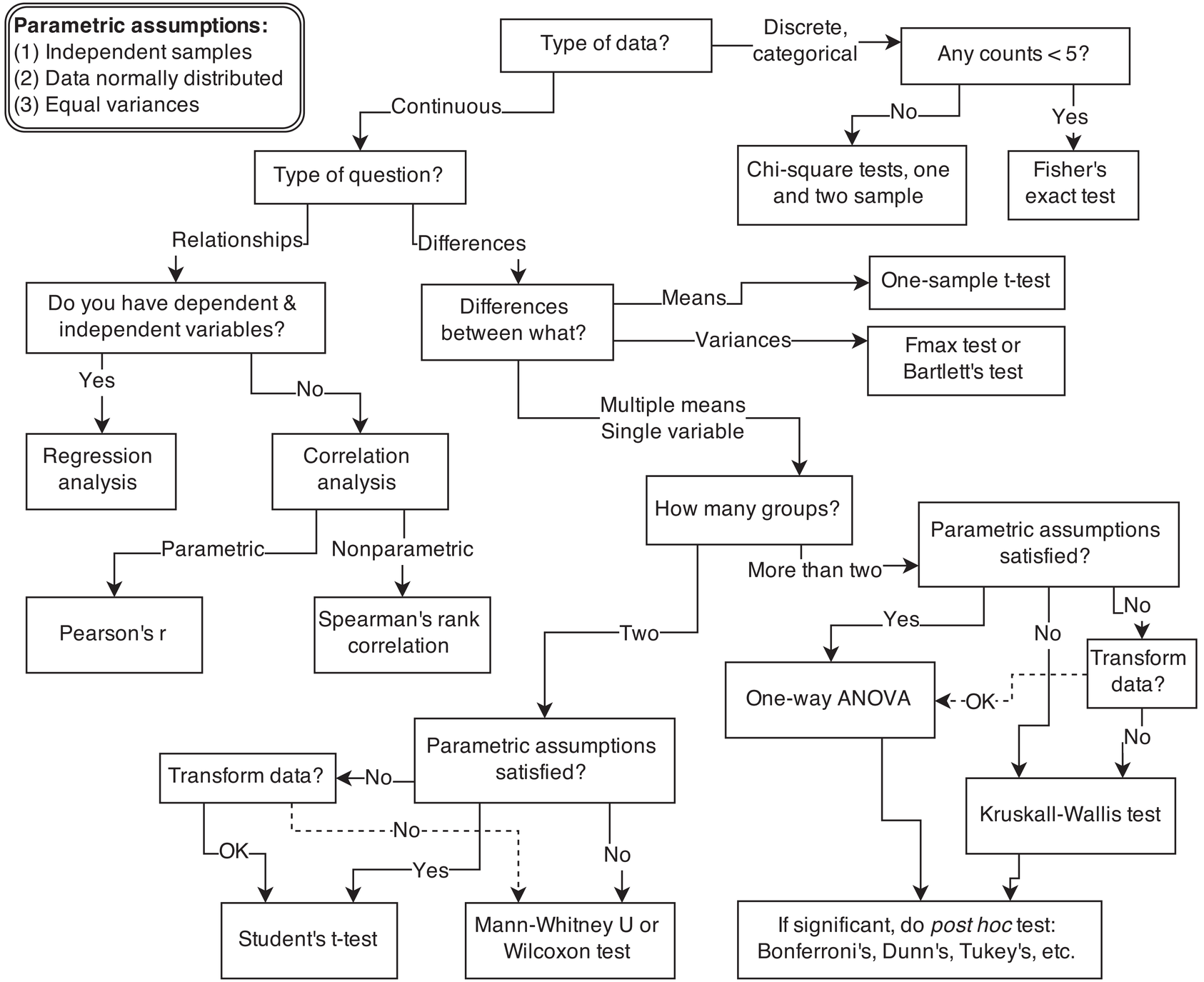

Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances
Type of data?
Discrete, categorical
Any counts < 5?
No
Yes
Chi-square tests, one and two sample
Fisher's exact test
Continuous
Type of question?
Relationships
Differences
Do you have dependent & independent variables?
Yes
No
Regression analysis
Correlation analysis
Parametric
Nonparametric
Pearson's r
Spearman's rank correlation
Differences between what?
Means
One-sample t-test
Variances
Fmax test or Bartlett's test
Multiple means Single variable
How many groups?
Two
More than two
Parametric assumptions satisfied?
Yes
No
One-way ANOVA
OK
Transform data?
No
Kruskall-Wallis test
Parametric assumptions satisfied?
Transform data?
No
OK
No
Yes
No
Student's t-test
Mann-Whitney U or Wilcoxon test
If significant, do post hoc test: Bonferroni's, Dunn's, Tukey's, etc.

Figure 1.1

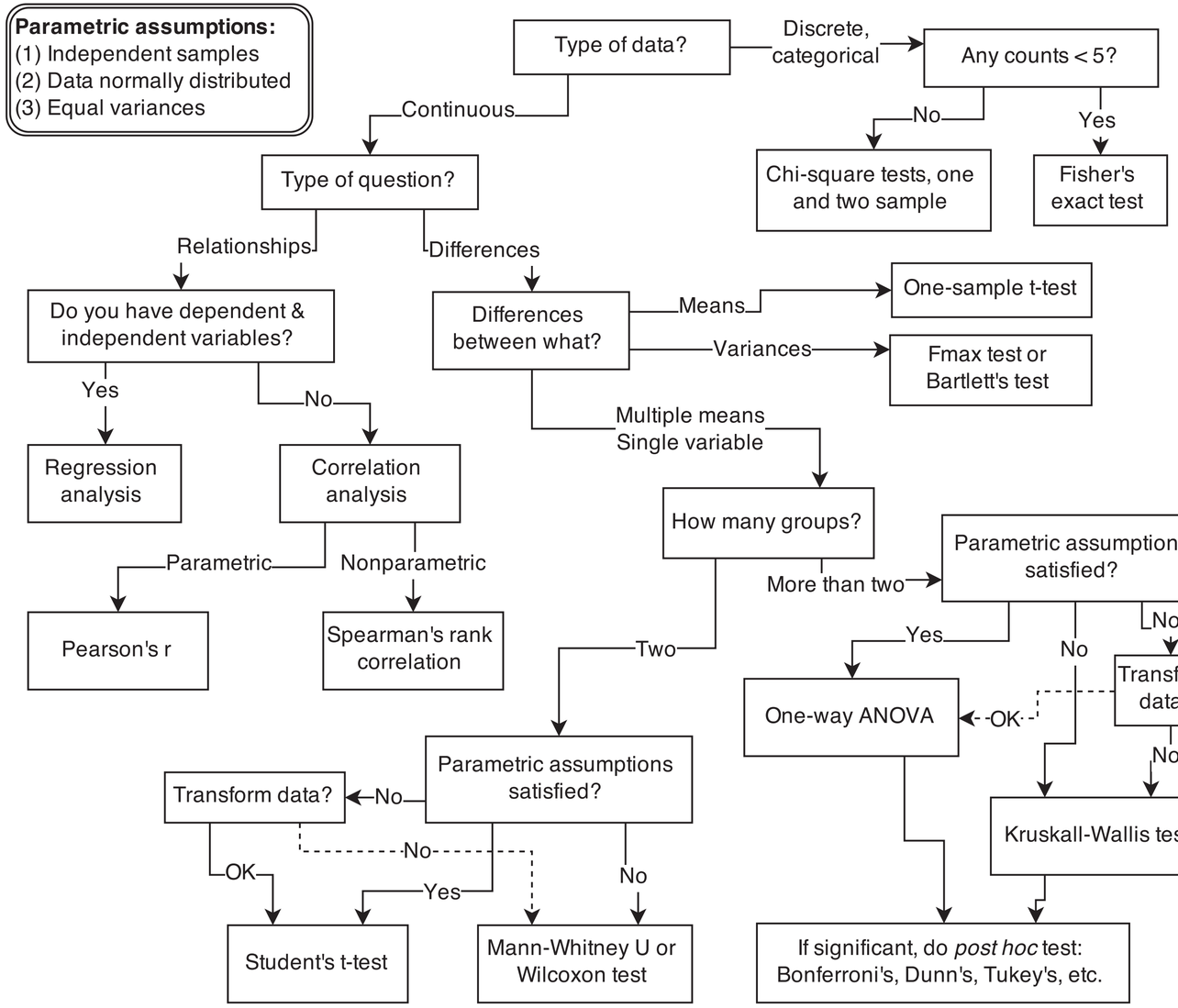
Statistical Models

Incredibly limiting

Focus on rejecting null hypotheses

Relationship between research and test not clear

Industrial framework



Null Models Rarely Unique

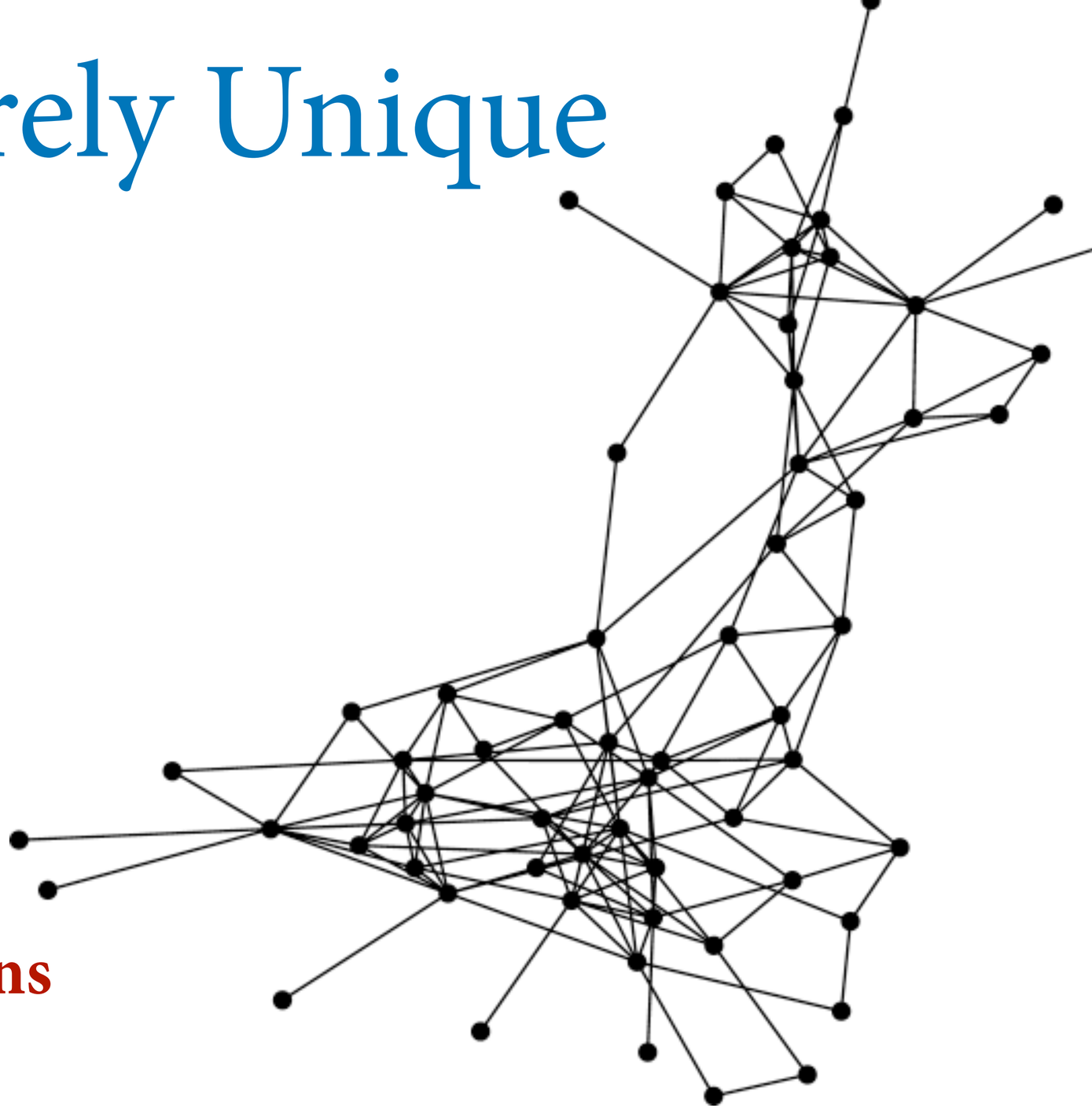
Null population dynamics?

Null phylogeny?

Null ecological community?

Null social network?

Problem: Many processes produce similar distributions



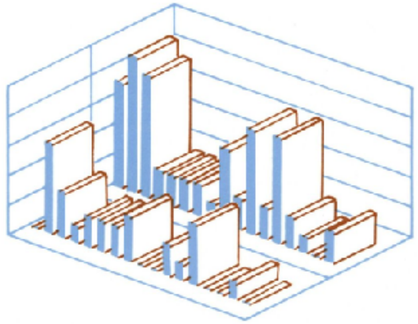
The neutral theory of molecular evolution

Motoo Kimura

Oxford Series in Ecology and Evolution

The Causes of Molecular Evolution

John H. Gillespie



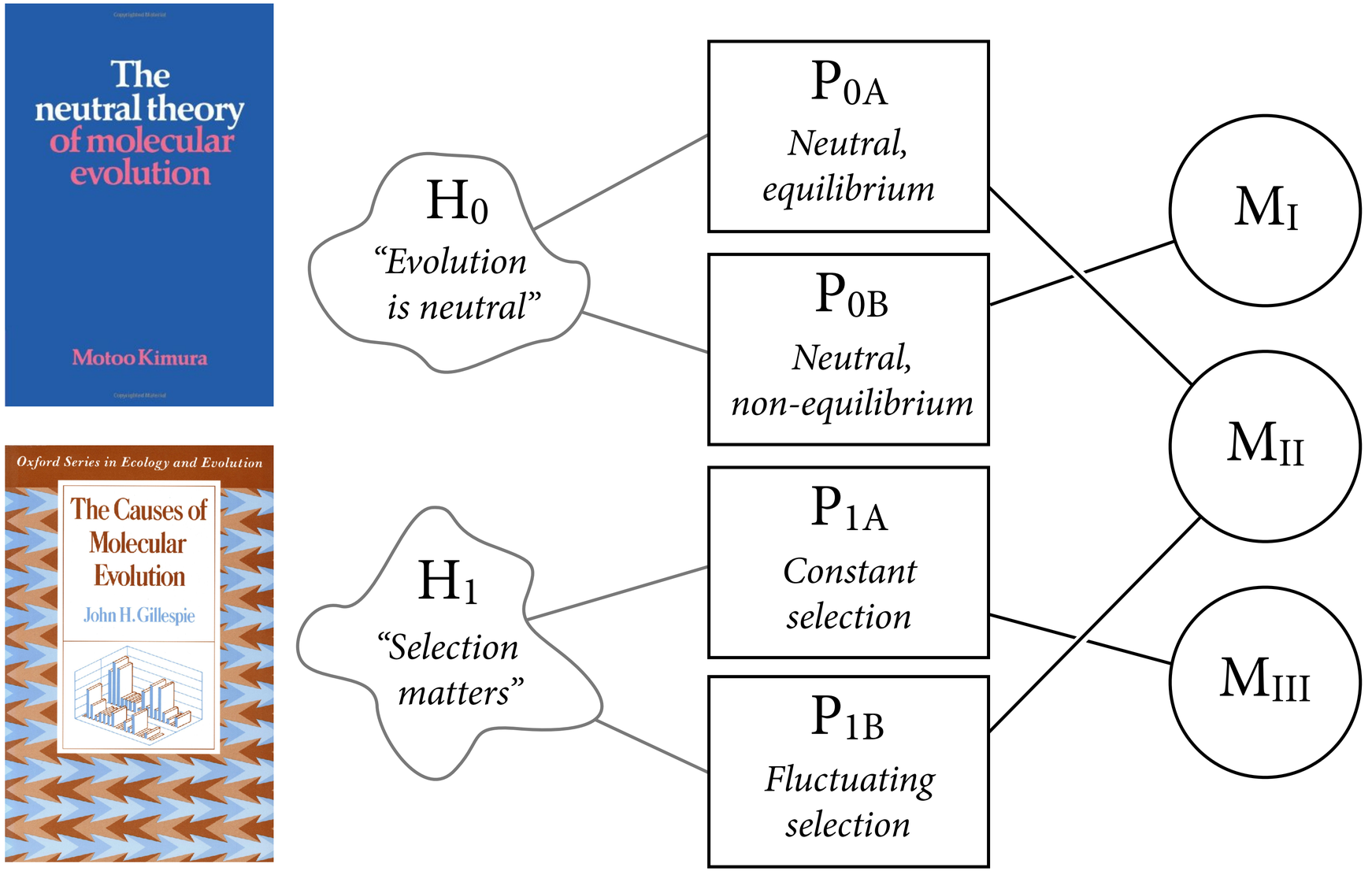

Hypotheses
Process models
Statistical models
The neutral theory of molecular evolution
Motoo Kimura
Oxford Series in Ecology and Evolution
The Causes of Molecular Evolution
John H. Gillespie
H0
"Evolution is neutral"
H1
"Selection matters"
P0A
Neutral, equilibrium
P0B
Neutral, non-equilibrium
P1A
Constant selection
P1B
Fluctuating selection
MI
MII
MIII

Figure 1.2

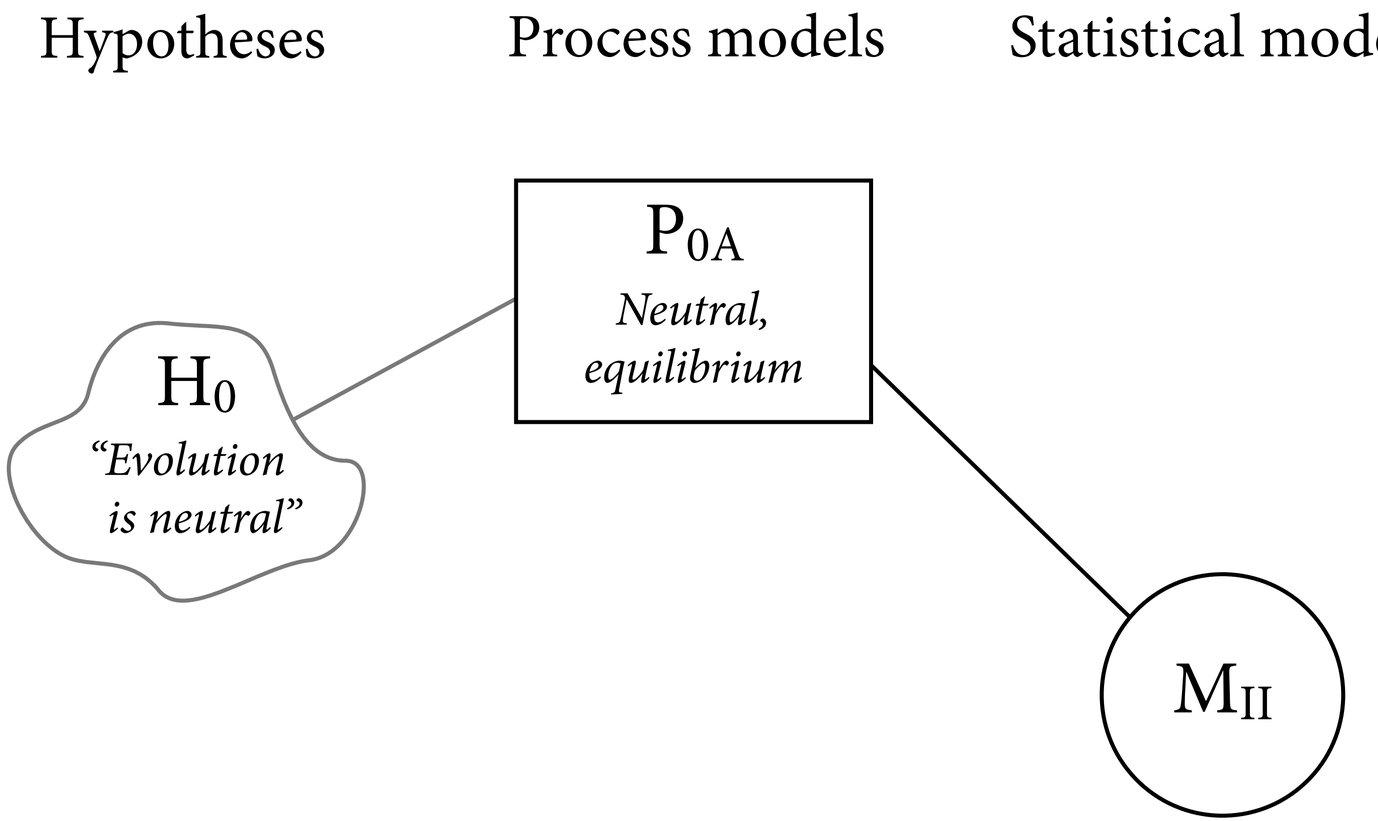

Hypotheses
Process models
Statistical models
P_{0A}
Neutral, equilibrium
H_0
"Evolution is neutral"
M_{II}

Figure 1.2

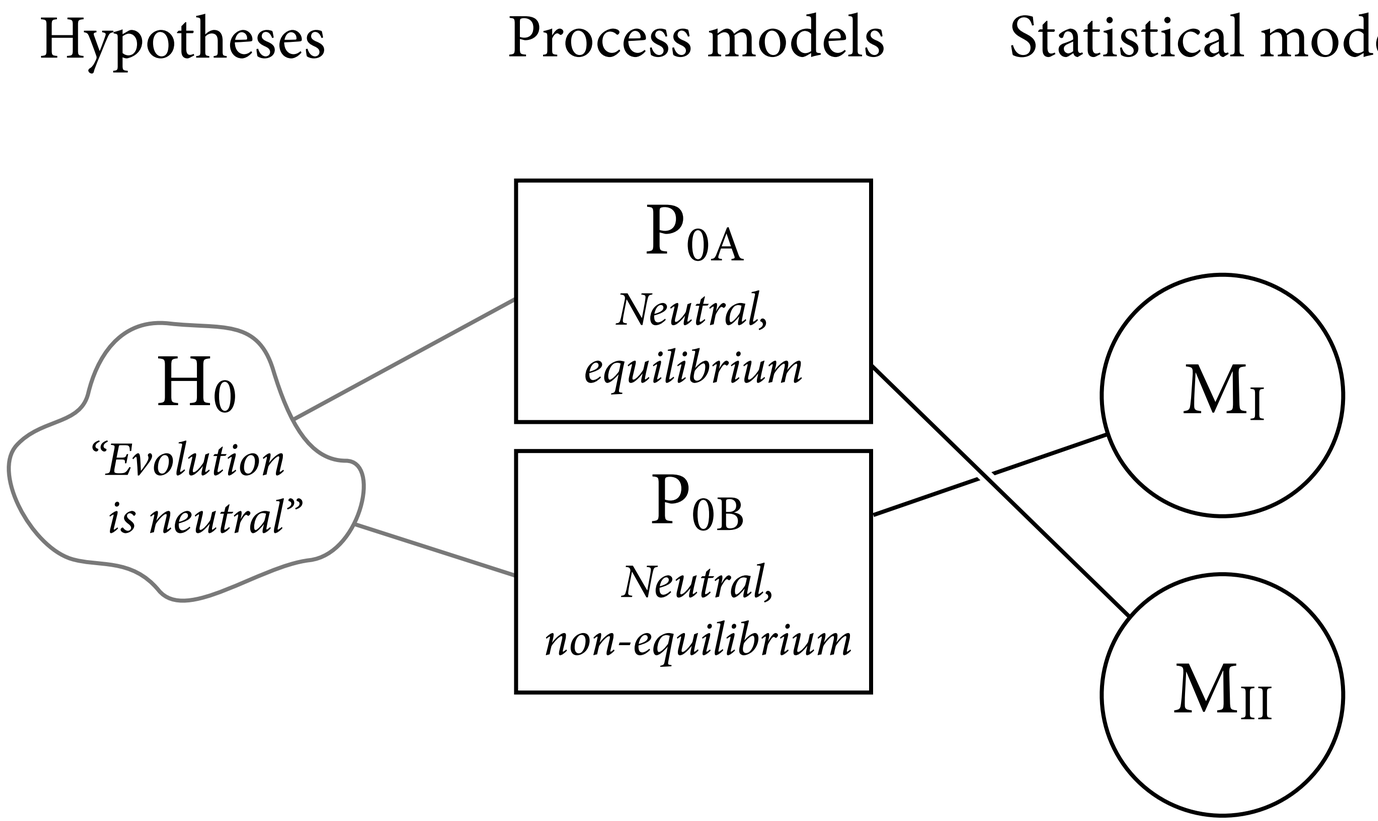

Hypotheses
Process models
Statistical models
H0
"Evolution is neutral"
P0A
Neutral, equilibrium
P0B
Neutral, non-equilibrium
MI
MII

Figure 1.2

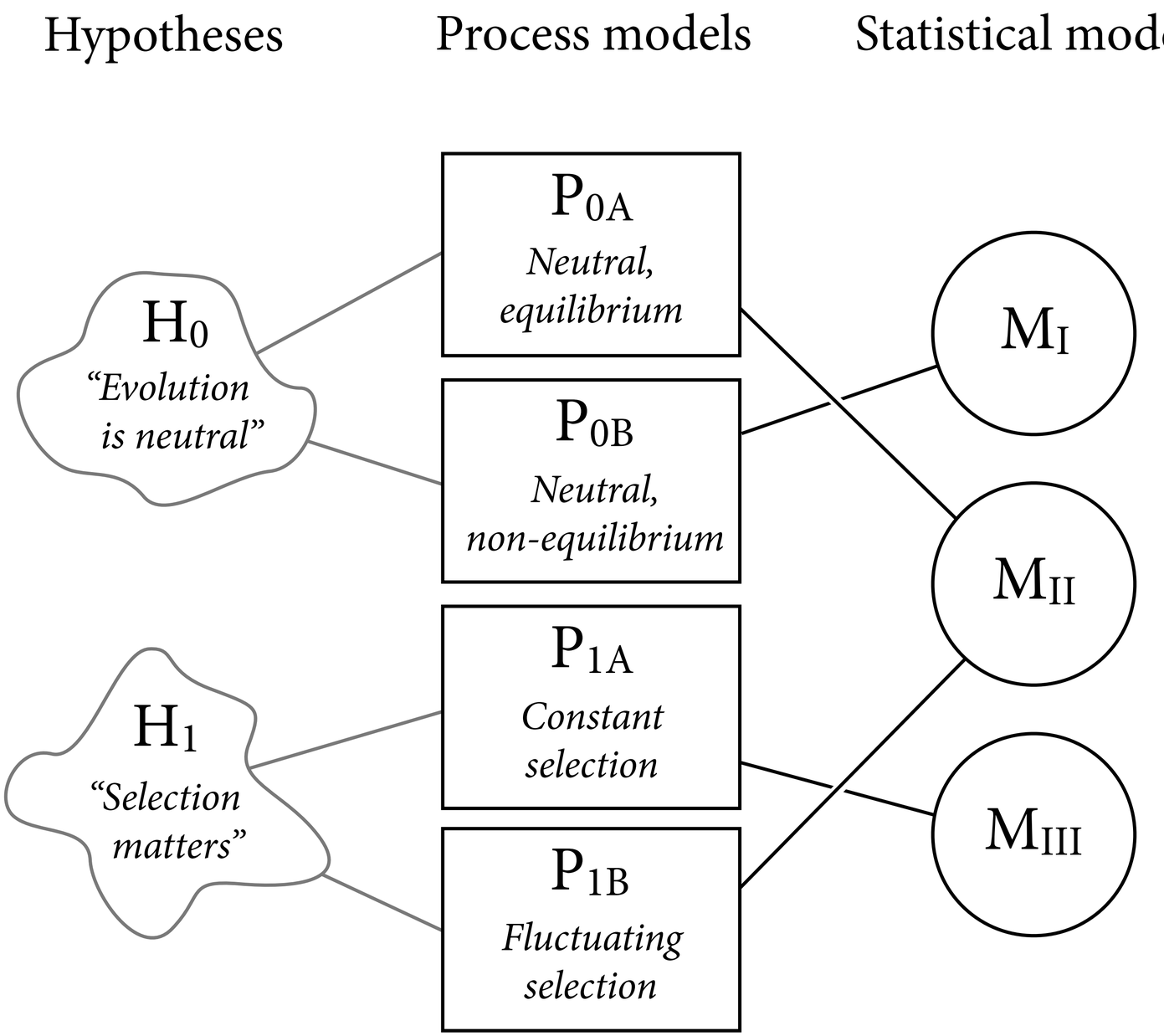

Hypotheses
Process models
Statistical models
H0
"Evolution is neutral"
H1
"Selection matters"
P0A
Neutral, equilibrium
P0B
Neutral, non-equilibrium
P1A
Constant selection
P1B
Fluctuating selection
MI
MII
MIII

Figure 1.2

The Unified Neutral Theory of
BIODIVERSITY AND BIOGEOGRAPHY

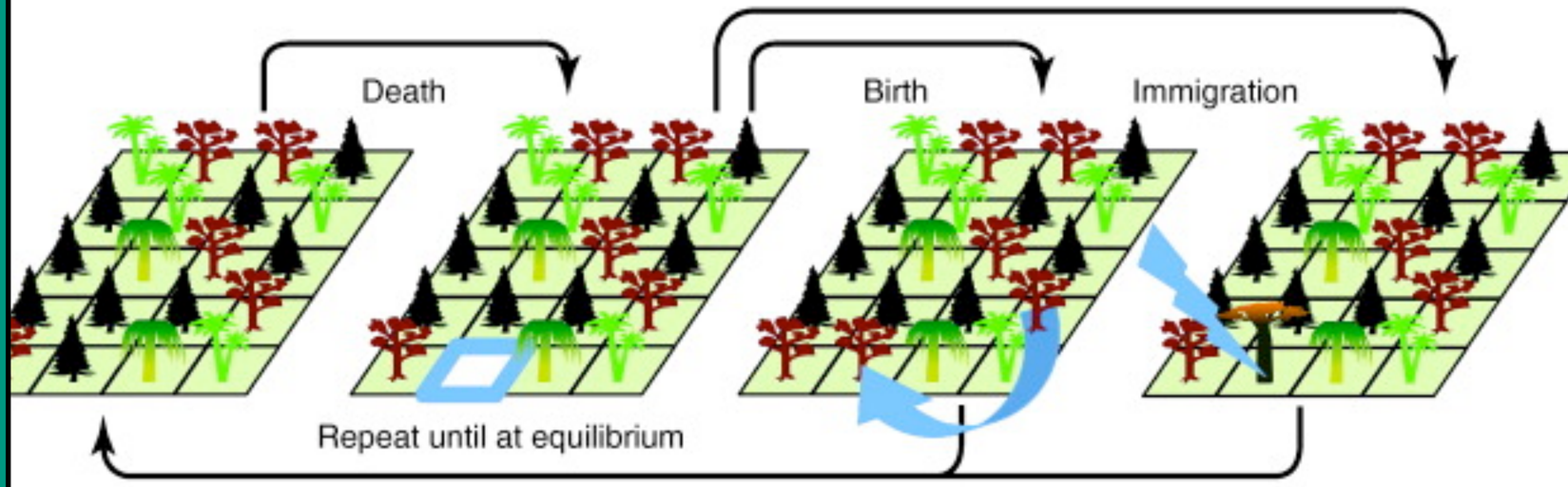
STEPHEN P. HUBBELL

MONOGRAPHS IN POPULATION BIOLOGY • 32

The coherence problem with the Unified Neutral Theory of Biodiversity

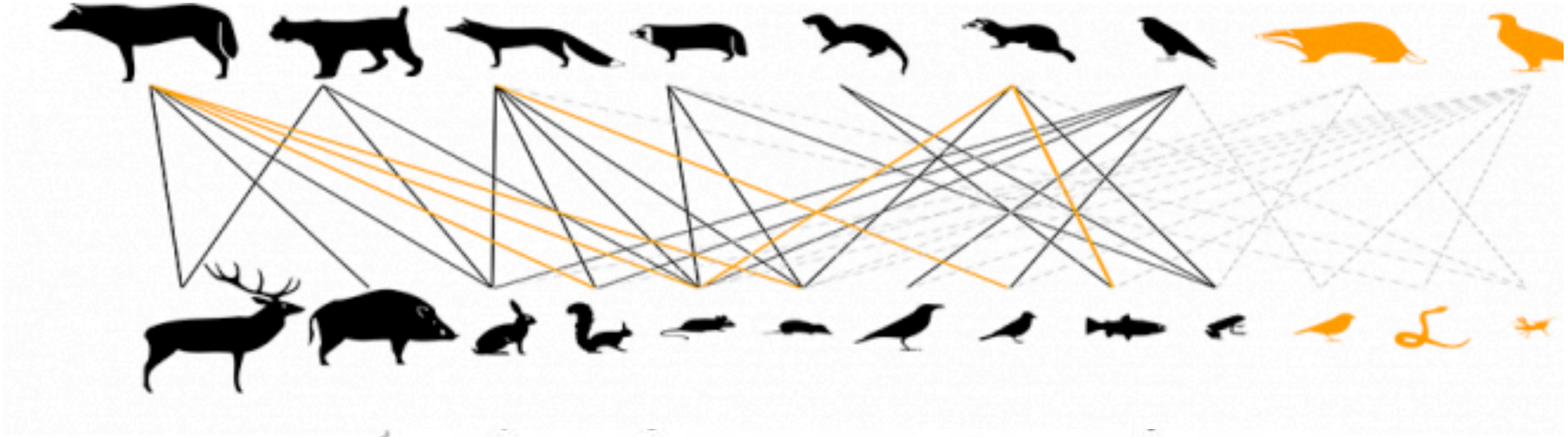
James S. Clark

Equal probability is not a theory, but lack of one; it does not include or exclude any process relevant to coexistence of competitors. Models lacking explicit species can make useful predictions, but this does not support neutral theory.



TRENDS in Ecology & Evolu

No null ecology



"The 'null hypothesis' analysis by Connor and Simberloff is characterized by hidden structure, inefficiency, lack of common sense, imprudence, and statistical weakness, and ultimately by a scandalous disregard for their own procedure," write Diamond and Gilpin. "We

Locations

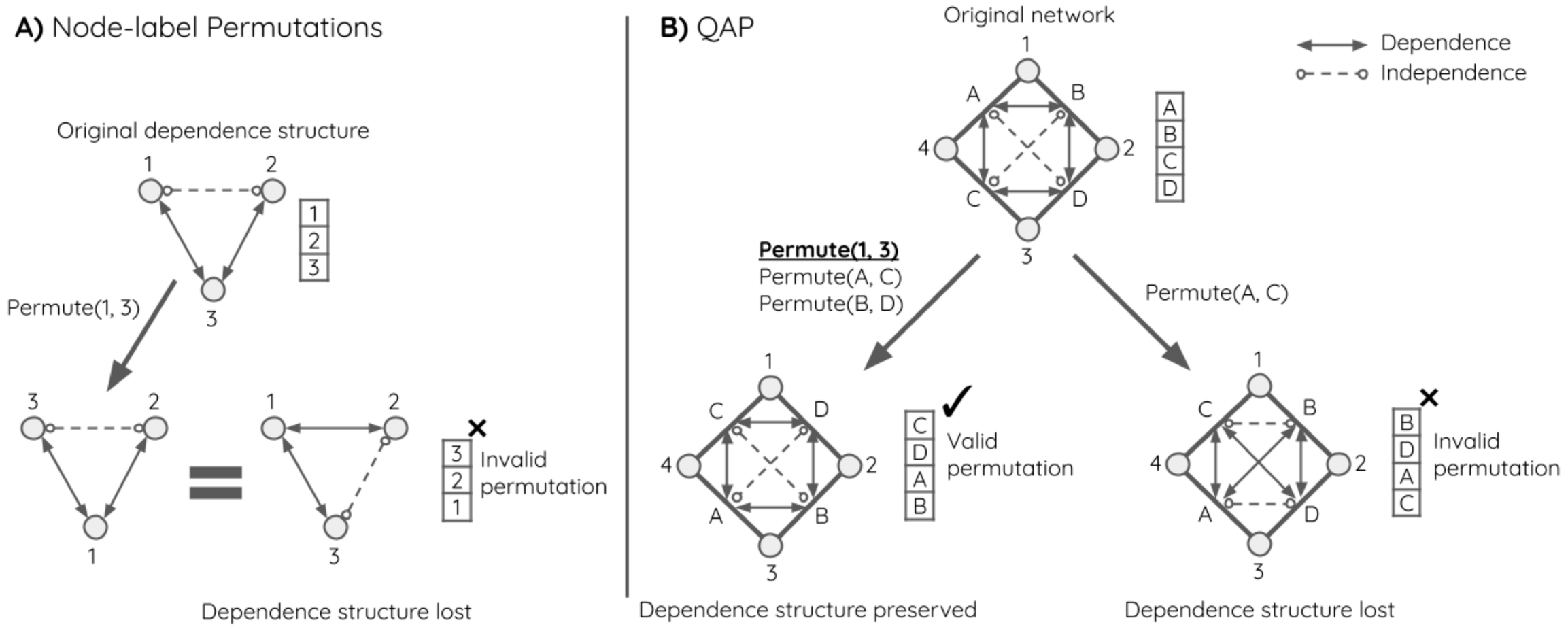
Species

0	0	1	0	1	0	1	1	1	1	0	1	0	0	0	0	1	1	1	0
1	1	0	1	0	1	0	0	0	0	1	0	1	1	1	1	0	0	0	1
0	0	1	0	1	1	1	1	0	0	1	0	0	1	0	0	1	1	1	0
1	1	0	1	0	0	0	0	1	1	0	1	1	0	1	1	0	0	0	1
1	0	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	0	1	1
0	1	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	1	0	0
1	0	0	1	1	1	0	1	0	1	1	0	0	1	0	0	1	1	0	0
0	1	1	0	0	0	1	0	1	0	0	1	1	0	1	1	0	0	1	1
0	0	1	0	1	0	1	1	1	0	0	1	0	0	1	1	0	1	1	0
1	1	0	1	0	1	0	0	0	1	1	0	1	1	0	0	1	0	0	1
1	1	0	1	0	1	1	1	1	0	0	0	0	1	0	1	1	0	0	0
0	0	1	0	1	0	0	0	0	1	1	1	1	0	1	0	0	1	1	1
1	0	0	0	0	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0
0	1	1	1	1	0	0	0	0	1	1	1	0	0	1	1	0	0	0	1
1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	0	0	1	1	0
0	0	1	1	0	0	1	1	0	1	1	0	0	0	1	1	1	0	0	1
0	0	0	1	0	1	0	1	1	0	1	0	1	1	1	0	1	0	1	0
1	1	1	0	1	0	1	0	0	1	0	1	0	0	0	1	0	1	0	1
0	1	1	0	1	1	0	1	0	1	0	0	0	1	0	1	0	1	0	1
1	0	0	1	0	0	1	0	1	0	1	1	1	0	1	0	1	0	1	0

Conor & Simberloff 1979, Diamond & Gilpin 1982

Null networks & fantastical beasts

Network permutation methods: low power, high false positives

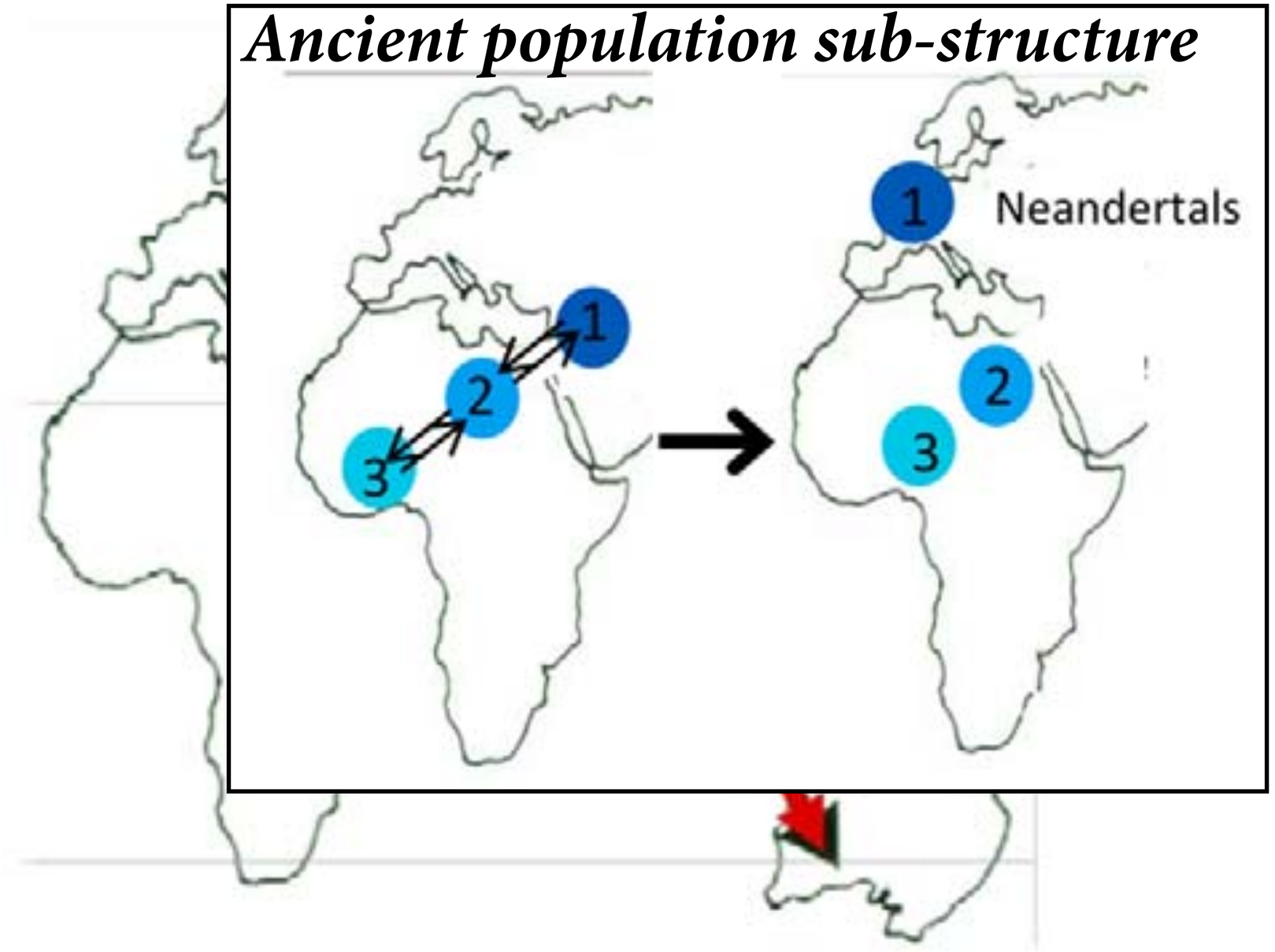


Hart et al 2021

Neandertal-Human interbreeding

Sankararaman et al 2012

Neandertal-Human interbreeding



Sankararaman et al 2015

Hypotheses and Models

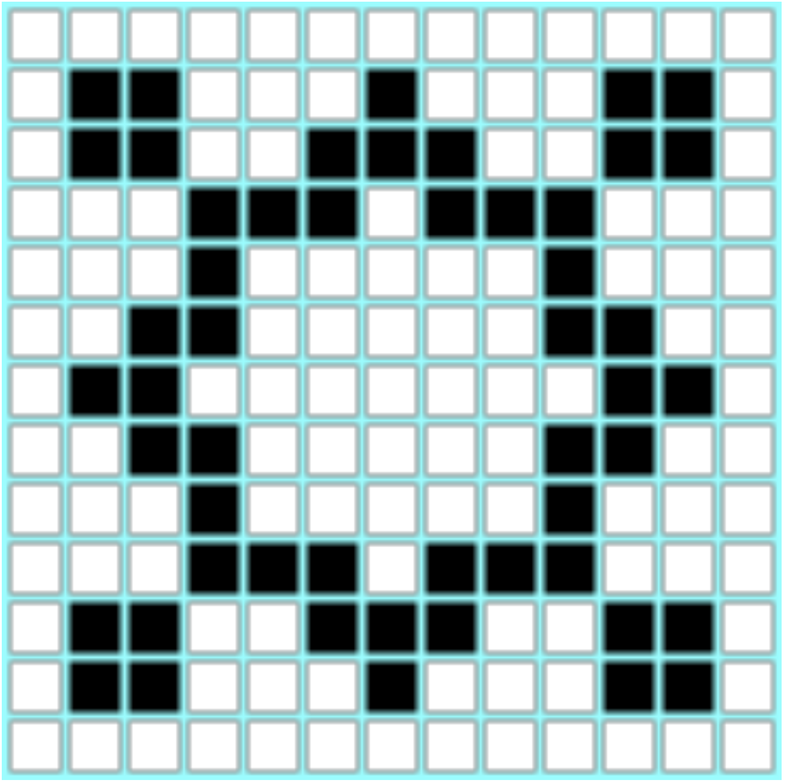
Research requires more than null robots

Also requires:

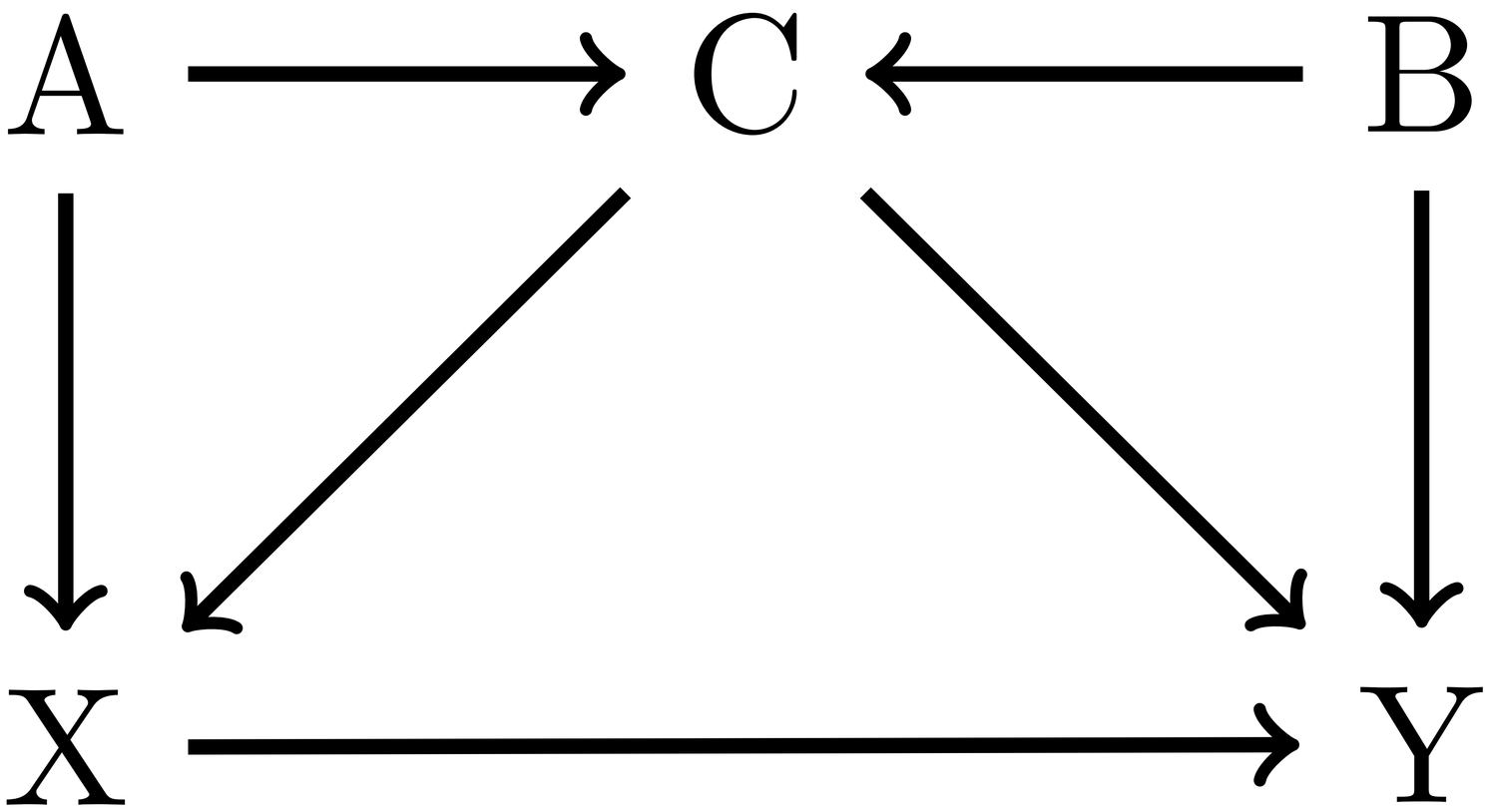
Generative causal models

Statistical models justified by generative models & questions (estimands)

An effective way to produce estimates



Justifying “controls”



Justifying “controls”

$Y \sim X$

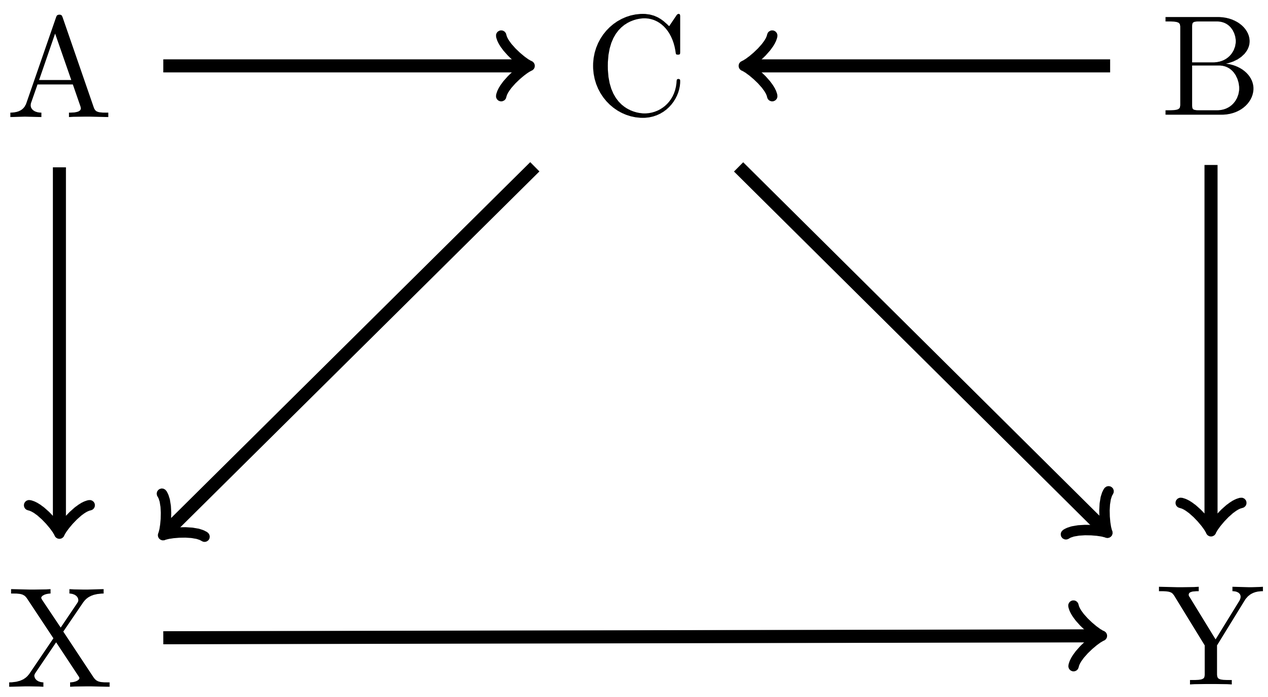
$Y \sim X + A$

$Y \sim X + A + B$

$Y \sim X + C$

$Y \sim X + A + C$

$Y \sim X + B + C$



Justifying “controls”

Y ~ X

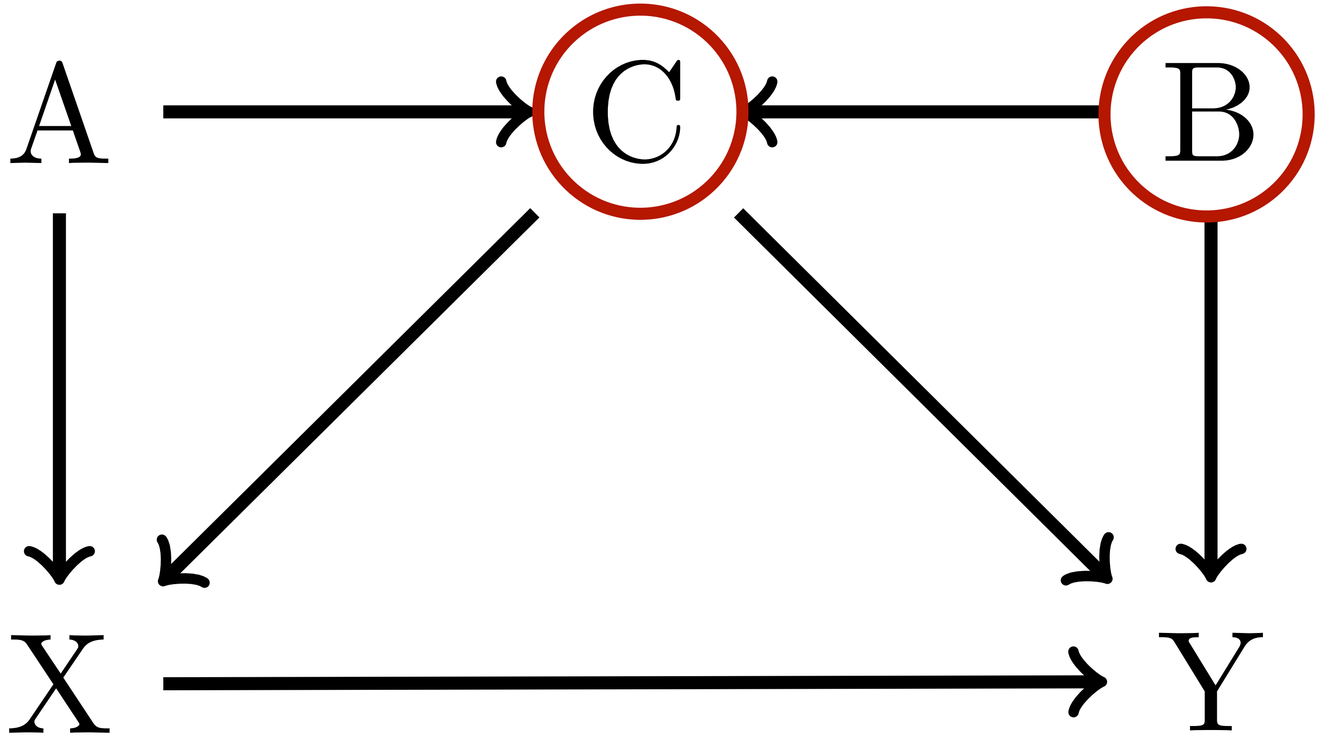
Y ~ X + A

Y ~ X + A + B

Y ~ X + C

Y ~ X + A + C

Y ~ X + B + C



“Adjustment set”

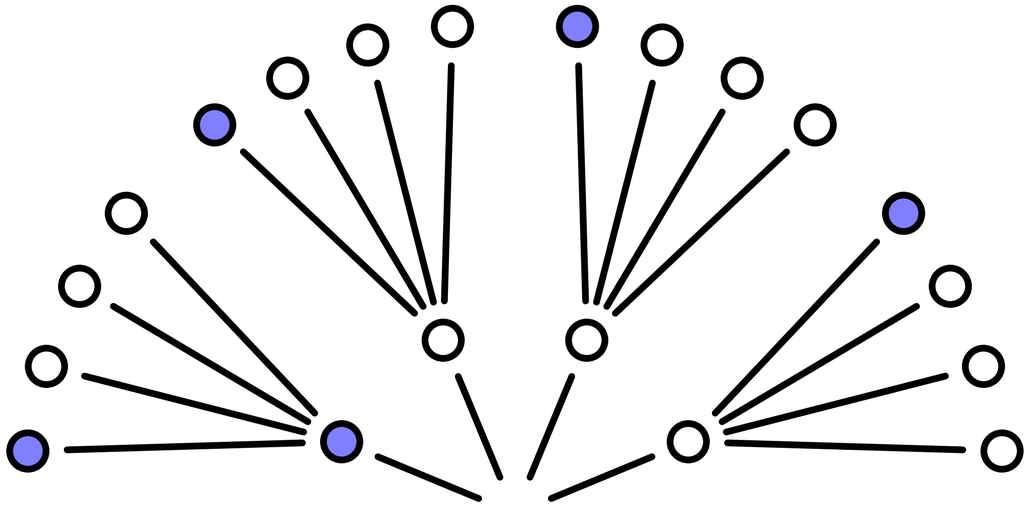
Finite data, infinite problems

DAG is not enough

Need generative model to design/ debug inference

Need a strategy to derive estimate and uncertainty

Easiest approach: Bayesian data analysis



Bayes is practical, not philosophical

Simple analyses: little difference, adds mess

Realistic analyses: huge difference

Measurement error, missing data, latent variables, regularization

Bayesian models are *generative*

Statistics wars are over

Bayes no longer controversial or marginalized

Bayesian methods routine

Waiting for teaching to catch up

The action is in machine learning, which has different battles

BAGS
GOLEMS
OWLS

HOW TO DRAW AN OWL

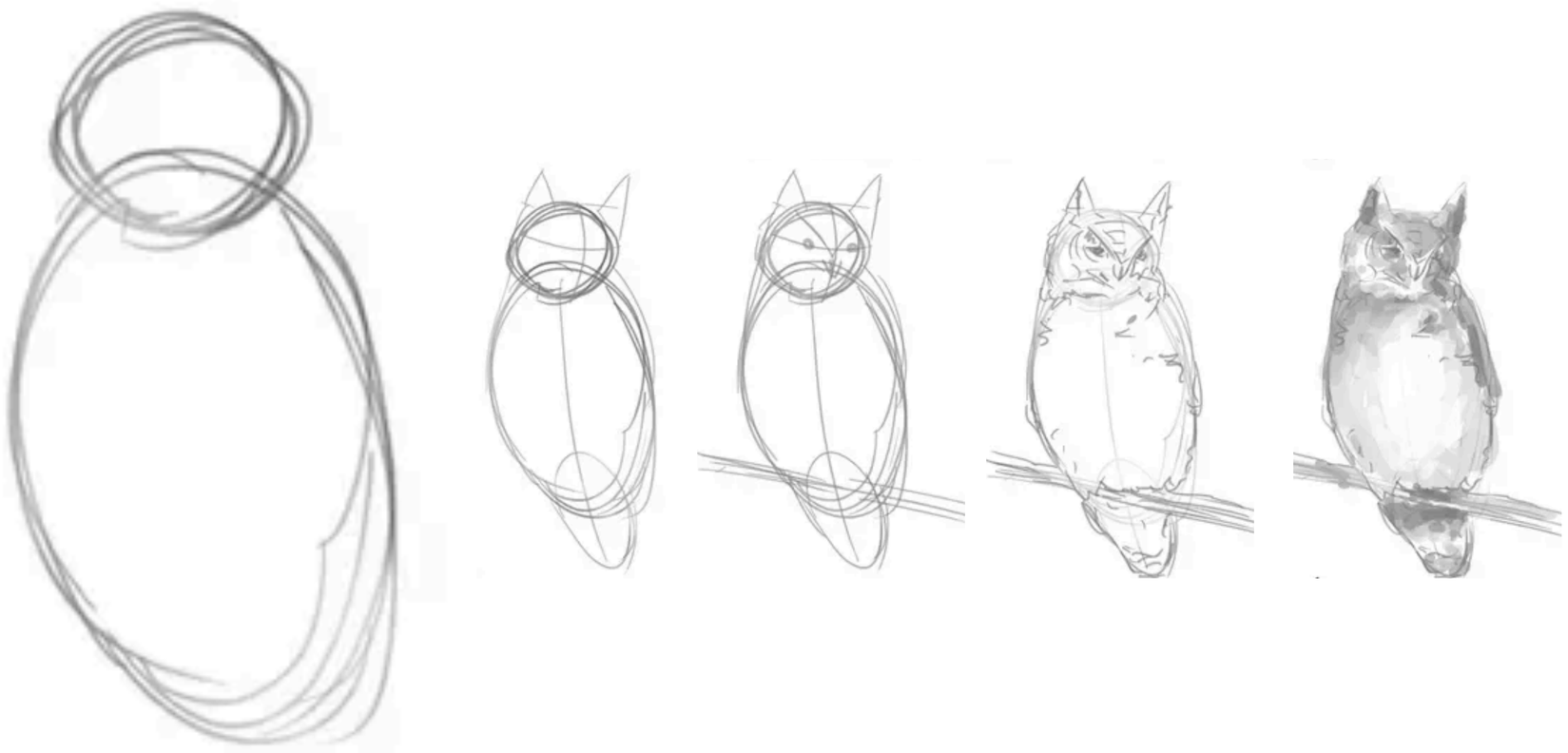
1. Draw some circles

HOW TO DRAW AN OWL

1. Draw some circles

2. Draw the rest of the owl

HOW TO DRAW AN OWL



1. Draw some circles

2. Draw the rest of the owl

1

```
# function to toss a globe covered p by water N times
sim_globe <- function( p=0.7 , N=9 ) {
    sample(c("W","L"),size=N,prob=c(p,1-p),replace=TRUE)
}
```

2

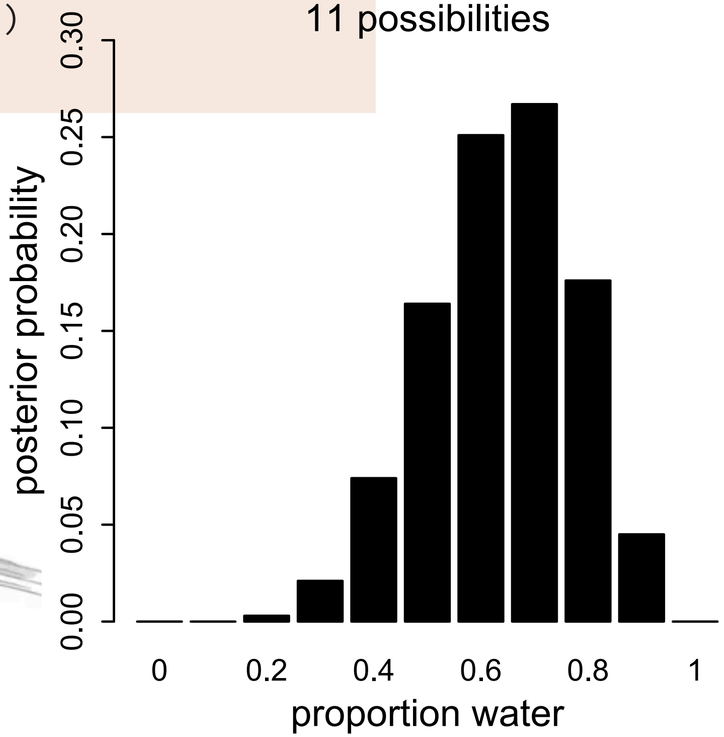
```
# function to compute posterior distribution
compute_posterior <- function( the_sample , poss=c(0,0.25,0.5,0.75,1) ) {
    W <- sum(the_sample=="W") # number of W observed
    L <- sum(the_sample=="L") # number of L observed
    ways <- sapply( poss , function(q) (q*4)^W * ((1-q)*4)^L )
    post <- ways/sum(ways)
    bars <- sapply( post, function(q) make_bar(q) )
    data.frame( poss , ways , post=round(post,3) , bars )
}
```

3

```
compute_posterior( sim_globe() )
```

```
  poss ways  post
1 0.00    0 0.000
2 0.25  243 0.291 ######
3 0.50  512 0.612 ############
4 0.75   81 0.097 ##
5 1.00    0 0.000
```

4



Drawing the Bayesian Owl

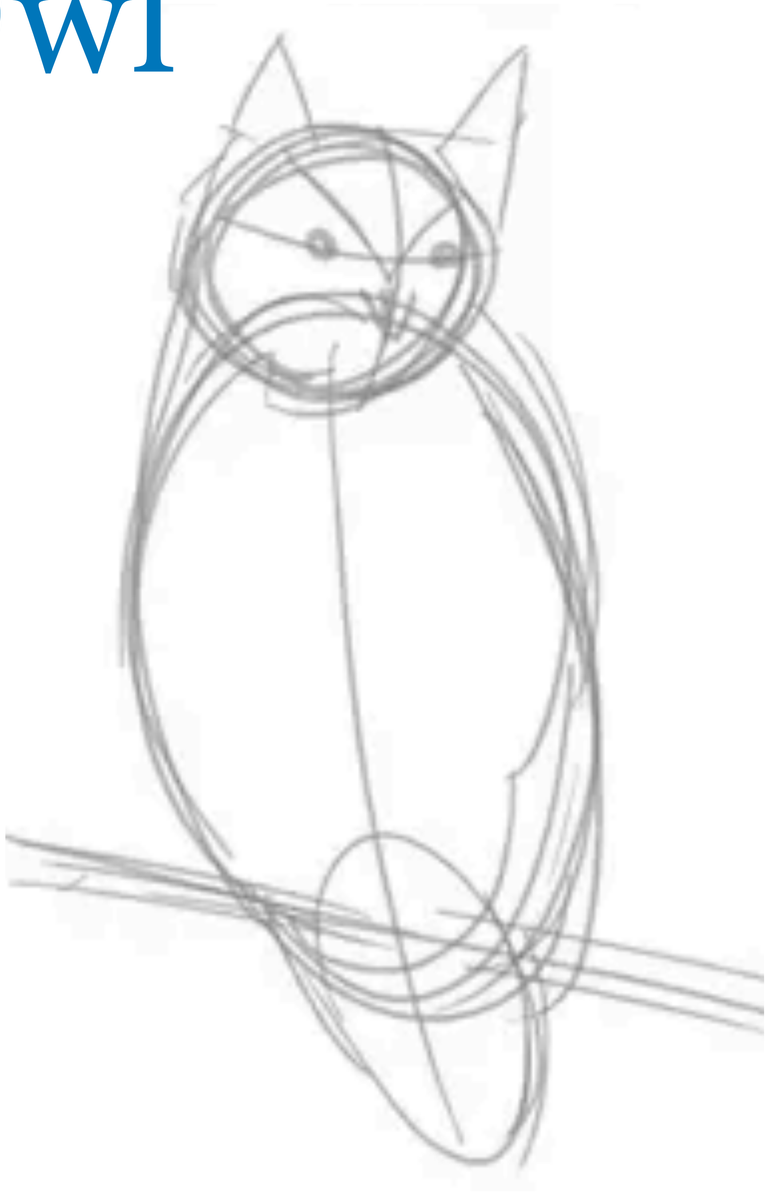
Scientific data analyses:
Amateur software engineering

Three modes:

Understand what you are doing

Document your work, reduce error

Respectable scientific workflow



Drawing the Bayesian Owl

1. Theoretical estimand

Drawing the Bayesian Owl

1. Theoretical estimand

2. Scientific (causal) model(s)

Drawing the Bayesian Owl

1. Theoretical estimand

2. Scientific (causal) model(s)

3. Use 1 & 2 to build statistical model(s)

Drawing the Bayesian Owl

1. Theoretical estimand

2. Scientific (causal) model(s)

3. Use 1 & 2 to build statistical model(s)

4. Simulate from 2 to validate 3 yields 1

Drawing the Bayesian Owl

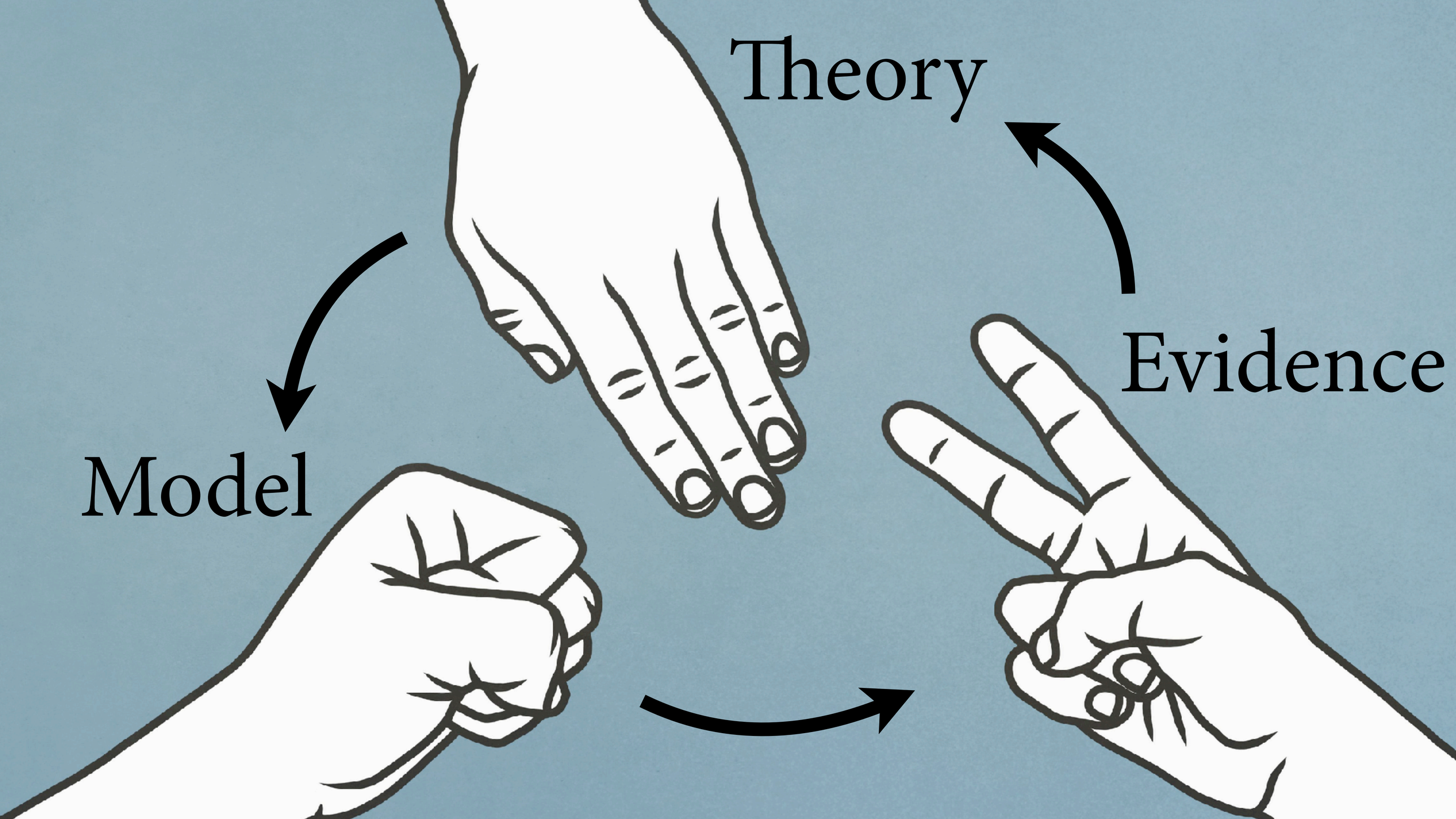
1. Theoretical estimand
2. Scientific (causal) model(s)
3. Use 1 & 2 to build statistical model(s)
4. Simulate from 2 to validate 3 yields 1
5. Analyze real data

DAGs, Golems & Owls

DAGs: Transparent scientific assumptions to
justify scientific effort
expose it to useful critique
connect theories to golems

Golems: Brainless, powerful statistical models

Owls: Documented procedures, quality assurance


Theory
Model
Evidence

Course Schedule

Week 1	Bayesian inference	Chapters 1, 2, 3
Week 2	Linear models & Causal Inference	Chapter 4
Week 3	Causes, Confounds & Colliders	Chapters 5 & 6
Week 4	Overfitting / Interactions	Chapters 7 & 8
Week 5	MCMC & Generalized Linear Models	Chapters 9, 10, 11
Week 6	Integers & Other Monsters	Chapters 11 & 12
Week 7	Multilevel models I	Chapter 13
Week 8	Multilevel models II	Chapter 14
Week 9	Measurement & Missingness	Chapter 15
Week 10	Generalized Linear Madness	Chapter 16

https://github.com/rmcelreath/stat_rethinking_2023