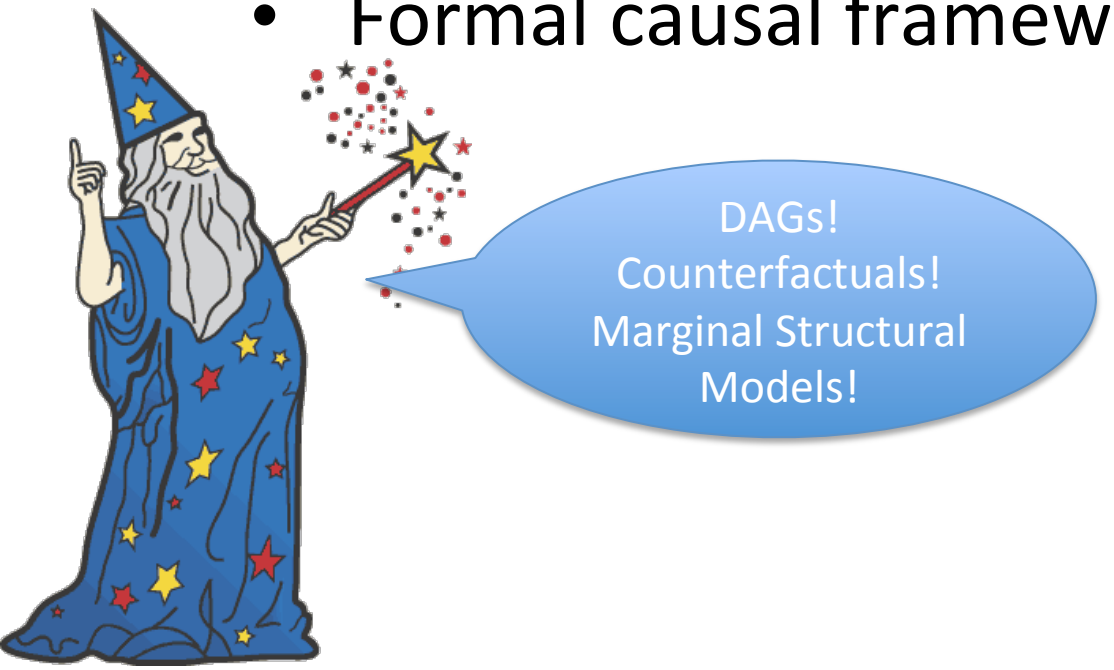
# Lecture 1:
# Causal versus Statistical Inference
# "The Roadmap"

# We must ask causal questions

- Not all questions are causal, but many are…
  - How things work
  - How best to intervene to improve health
- Answering causal questions is HARD
  - Formal causal frameworks to the rescue…




# Here Be Dragons….

- Overconfidence in assumptions
  - I can draw a causal graph

  -> My graph is useful! My graph represents reality!

- Misinterpretation of results
  - My observational analysis now replicating a trial…

  -> Why does my trial not match my analysis results?

- Complexity obscuring common sense

$$E(Y(j+m)_{\overline{A}(j-1)\underline{a}(j)}|\overline{A}(j-1)=\overline{a}(j-1),\overline{S}(j))=E(Y(j+m)_{\overline{a}(j-1)\underline{a}(j)}|\overline{S}(j)_{\overline{a}(j-1)})$$

  -> ???

# Formal Causal Frameworks are a Tool

- What are they useful for?

1. Make uncertainty explicit
2. Frame better questions
3. Understand assumptions
4. Improve study design
5. Design statistical analyses that come closer to answering our questions
6. Interpret results appropriately

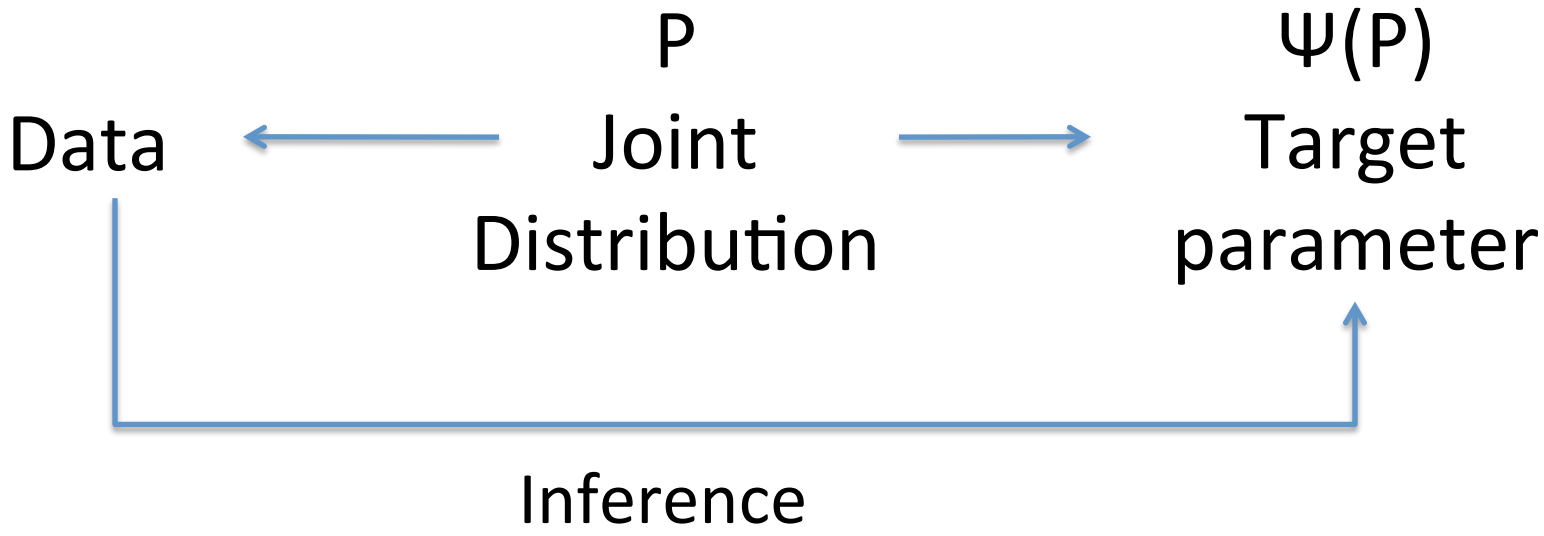
# Outline and References

- Causal versus statistical inference
  - Judea Pearl on the Basic Distinction
    - J. Pearl Tutorial: cs.ucla.edu/~judea/jsm12
    - J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys, 2009* 3: 96—146.
- A general Roadmap for causal questions
    - M Petersen, M van der Laan, Causal Models and Learning from Data: Integrating Causal Modeling and Statistical Estimation. *Epidemiology 2014* 25(3): 418-426

# Statistical parameters

- Provide a summary of a data distribution based on a sample drawn from that distribution
  - Ex: We sample individuals from some underlying population and on each subject observe: A=vitamin use, Y=breast cancer
  - Can estimate association between A and Y in underlying population
    - Ex: P(Y=1|A=1) and P(Y=1|A=0)

# Traditional Statistical Inference Paradigm




Example: How likely were women who took vitamins to have breast cancer?
P(Y=1|A=1)

Adapted from: cs.ucla.edu/~judea/jsm12

# Statistical parameters

- Provide a summary of a data distribution based on a sample drawn from that distribution
  - How did the risk of breast cancer differ between women who took vitamins and those who did not?
    - P(Y=1|A=1)-P(Y=1|A=0)
- Tells us about probabilities of past events
  - Extend this to future events only if we assume the experimental conditions didn't change

# However…

- Many, many central questions in epidemiology and biostatistics (and political science, and economics, and education, and sociology, and medicine, and…) ask what would happen if conditions did change…
  - The natural "experiment" (or process) that generated our data rarely corresponds to the ideal experiment we are interested in…

# Example

- In the “experiment” that generated our data, only a self-selected group (women motivated and with the means to do so) took vitamins
- How would breast cancer rates change if we instituted a policy in which all women took vitamins?
  - How would $P(Y=1|A=1)$ change?
- Or we randomly assigned women to take vitamins?
  - How would $P(Y=1|A=1)-P(Y=1|A=0)$ change?

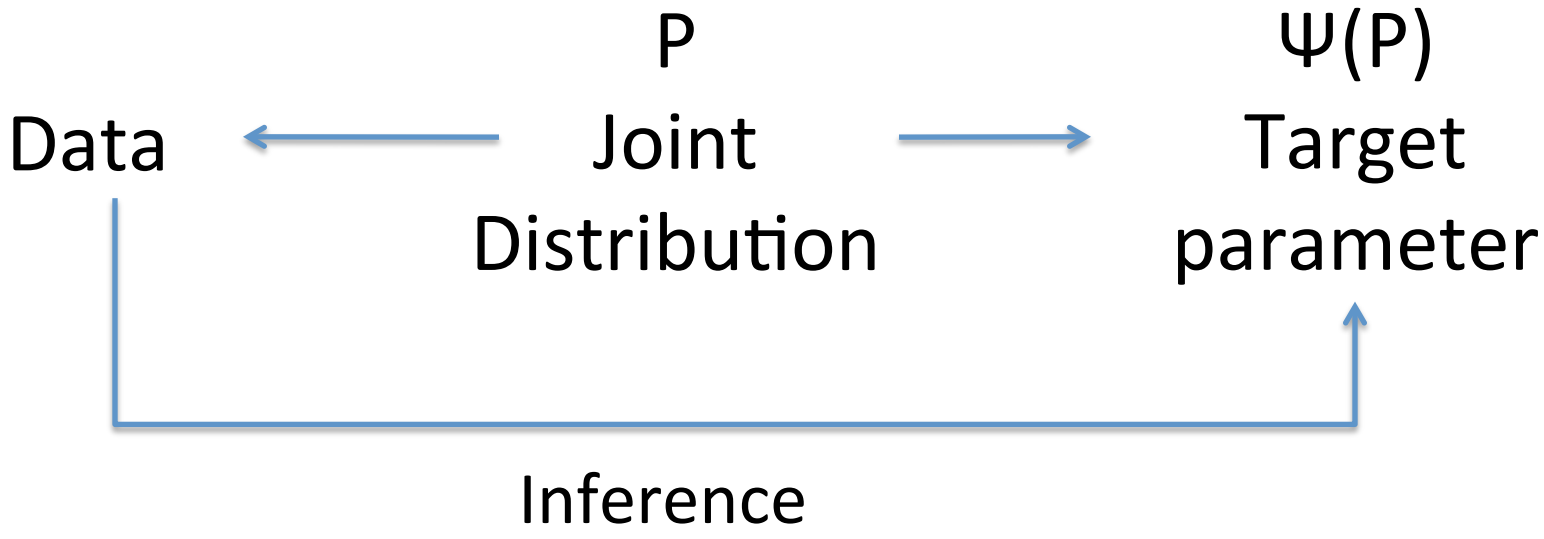
# Randomized controlled trials are not excepted…

- In the experiment that generated our data, we randomly assigned patients to take a drug versus a placebo
  - A lot of people dropped out of the study
  - Many people didn't comply with their assigned treatment
- We observed a minimal reduction in mortality in the treated versus control arm
- What would the mortality difference between the arms have been if we had prevented loss to follow up and enforced compliance?

# Causal questions are questions about what happens when we change the way data are generated

- Causal Parameters: Provide a summary of how a data distribution would change if the experimental conditions changed
- This is a very general way to think about the goal of causal inference – inference about parameters of a distribution we do not (fully) observe!

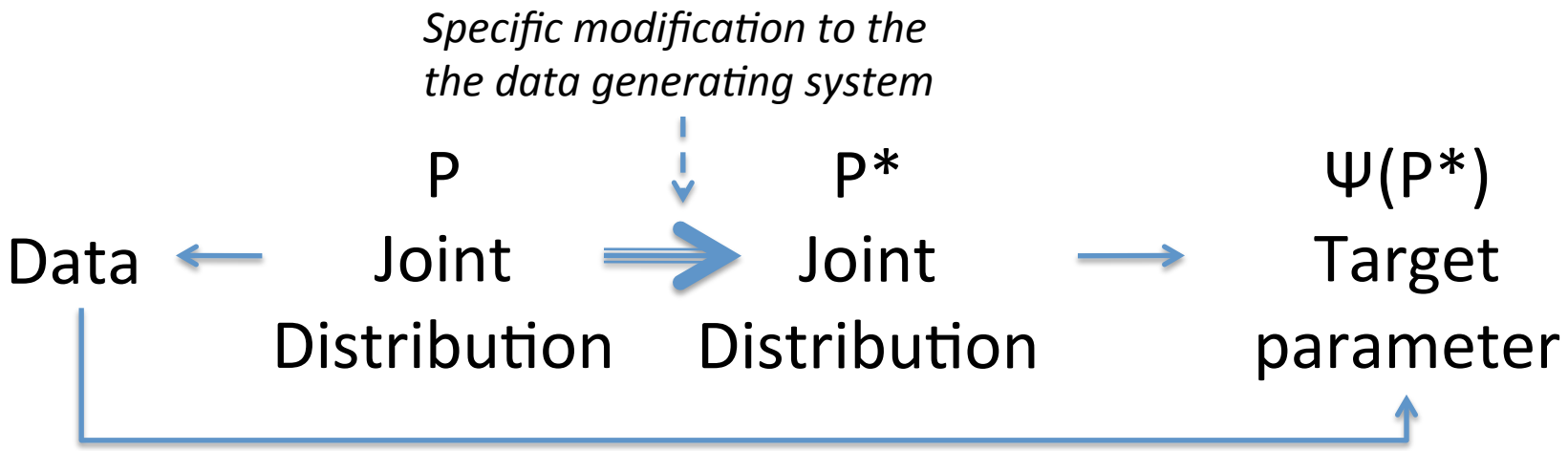
# Traditional Statistical Inference Paradigm




Example: How likely were women who took vitamins to have breast cancer?
P(Y=1|A=1)

Adapted from: cs.ucla.edu/~judea/jsm12

# From a Statistical to Causal Analysis




Inference

Example: If all women had taken vitamins, how likely would they have been to have breast cancer?

Adapted from: cs.ucla.edu/~judea/jsm12

# Causal inference requires something extra…

- There is nothing in the distribution of the data alone that tells us how it should change when conditions change
  - Which parts of the distribution will stay the same, which will change…?
- To make causal inferences we have to make assumptions about the processes that generated the data
  - These are not statistical assumptions!
    - They may have implications that are testable statistically, but that is not the same thing…
    - We will return to this in detail

# What's special about causal inference?

- Causal questions require some knowledge of the data generating process
  - They cannot be computed from the data alone or the distributions that govern it
  - Example: We see that headache (A) and brain tumor (B) are associated
  - No information in data alone (A,B) about the direction of underlying data-generating process
    - Based on distribution of the data alone, no way to know if curing the brain tumor will improve the headache, if curing the headache will improve the brain tumor, or neither

# Judea Pearl on defining causation

- Associational concept=any relationship that can be defined in terms of a joint distribution of observed variables
  - Correlation, conditional independence, dependence, likelihood, propensity score…
  - Testable in principle
- Causal concept: any relationship that cannot be defined in terms of this distribution alone
  - Randomization, confounding, instrumental variable, attribution, effect…
  - Not testable in principle (without experimental control)
    - Ie we can only test them if we can intervene on the system and see what happens
    - Even then there are complications…

# Outline

- Causal versus Statistical Inference
  - Judea Pearl on the Basic Distinction
- A General Roadmap for tackling causal questions rigorously
  - The framework for the class
  - Today, an introduction

# A roadmap for causal inference

1. Specify **Causal Model** representing real background knowledge
2. Specify **Causal Question**
3. Specify **Observed Data** and link to causal model
4. **Identify** : Knowledge + data sufficient?
5. Commit to an **estimand** as close to question as possible, and a **statistical model** representing real knowledge.
6. **Estimate**
7. **Interpret** Results

# Before we get started….

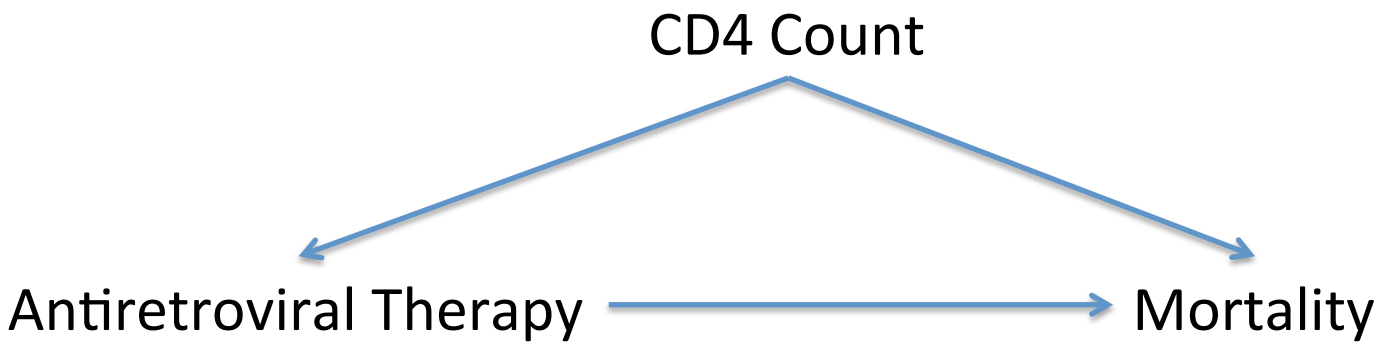
- Need to specify the question
  - What system (including target population) that you want to study, and what you want to learn about it?
- Example: Does antiretroviral therapy (ART) reduce mortality (and if so, by how much)?
  - For which patients? Initiated when?
  - ART defined how? Specific drugs? Duration?
  - Which outcomes? All-cause mortality?
  - Outcomes assessed over what time frame?

# 1. Specify **Causal Model**

- Good Practice: Use background knowledge
  - Confounding, selection bias, measurement error….
- Causal Model: A formal language for expressing it
  - Makes knowledge more powerful
- Moving beyond intuition…
  - We take for granted that intuition is not enough to help us make sense of complex data- that is why we have statistics
  - Same for causal reasoning. Intuition breaks down…

# Causal Models Express Knowledge

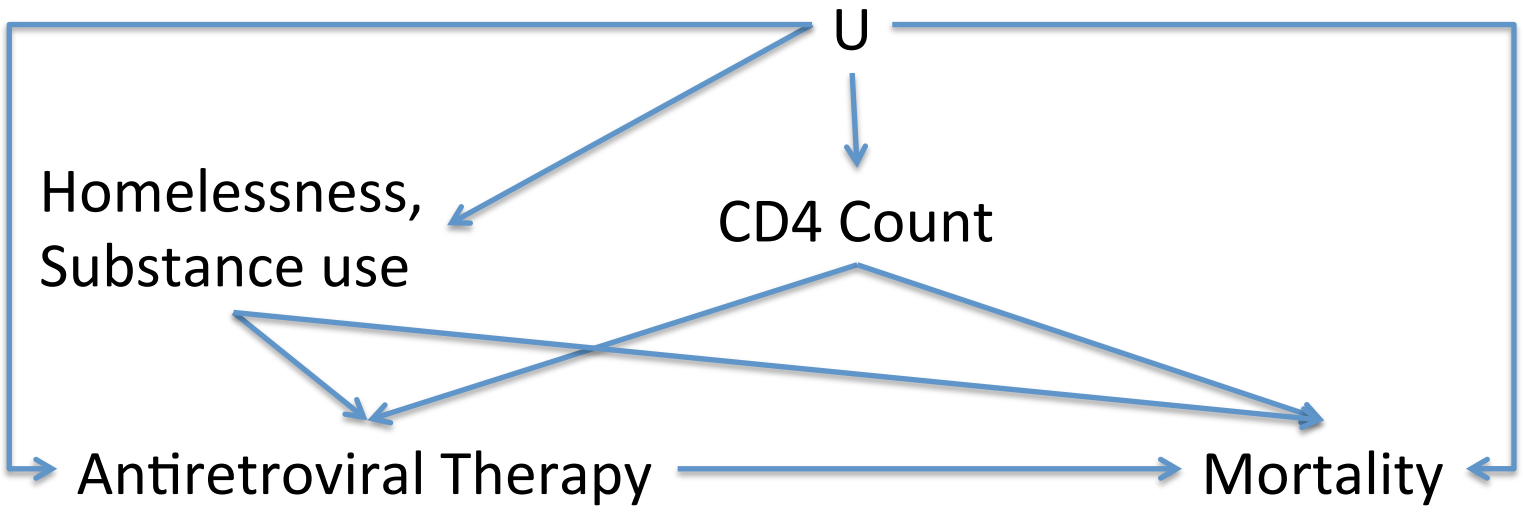
- This class: Structural Causal Models
  - Unification of causal graphs, (non-parametric) structural equation models, counterfactuals




- Other formal causal frameworks:
- Counterfactual/Potential Outcome, Single World Intervention Graphs, FFRCISTG, Decision Theoretic….

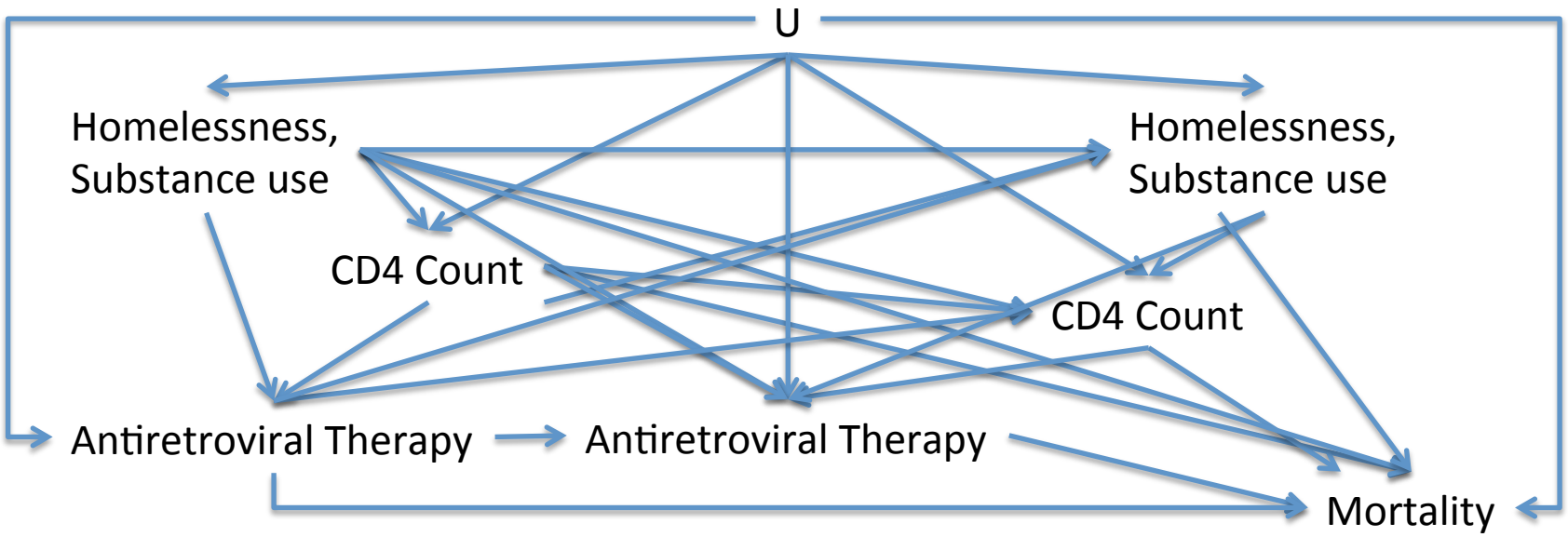
# Causal Models Express Uncertainty

- We have real knowledge, **but it is limited**
  - Which variables affect each other?
  - Unmeasured factors?
  - Functional form of those relationships?



# All causal models are not wrong…

- Causal models should represent real knowledge
  - We often need to make additional assumptions in order to make progress
  - Keep this process separate
- Beware the temptation to oversimplify



# All causal models are not wrong…

- Causal models should represent real knowledge
  - We often need to make additional assumptions in order to make progress
  - Keep this process separate
- Beware the temptation to oversimplify

**Yuck!!!**

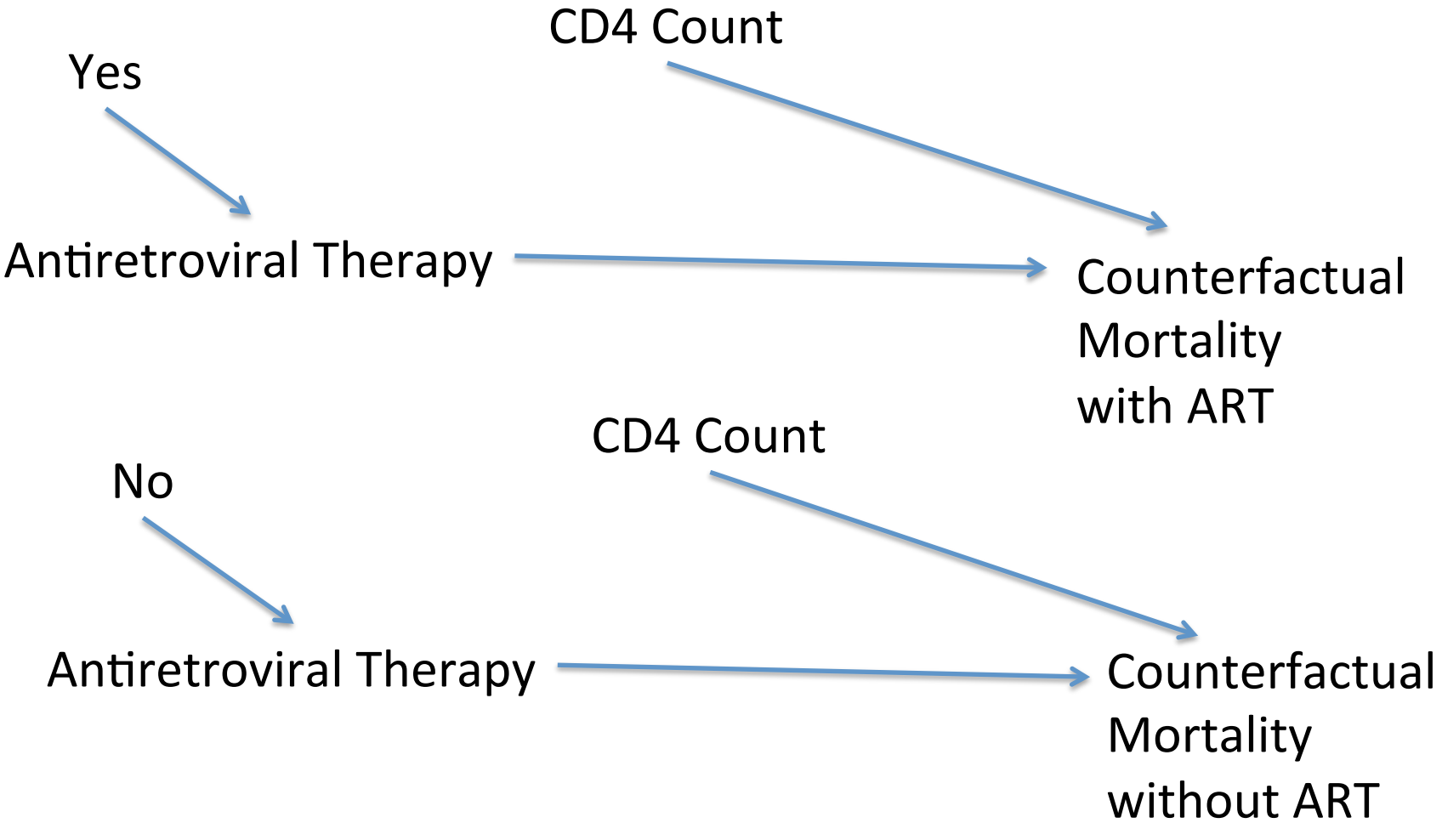
- Use a simpler model? Not unless your knowledge justifies it
  - Can use a different language to express your knowledge and lack thereof…

# **2.** Specify **Causal Question**

- Good Practice: State your scientific question clearly
- Causal Framework: A language to translate your question into a formal query
  - Ex: Using counterfactuals
  - Forces you to be explicit about exactly which "experiment" (change(s) to the system of interest) would let you answer your question

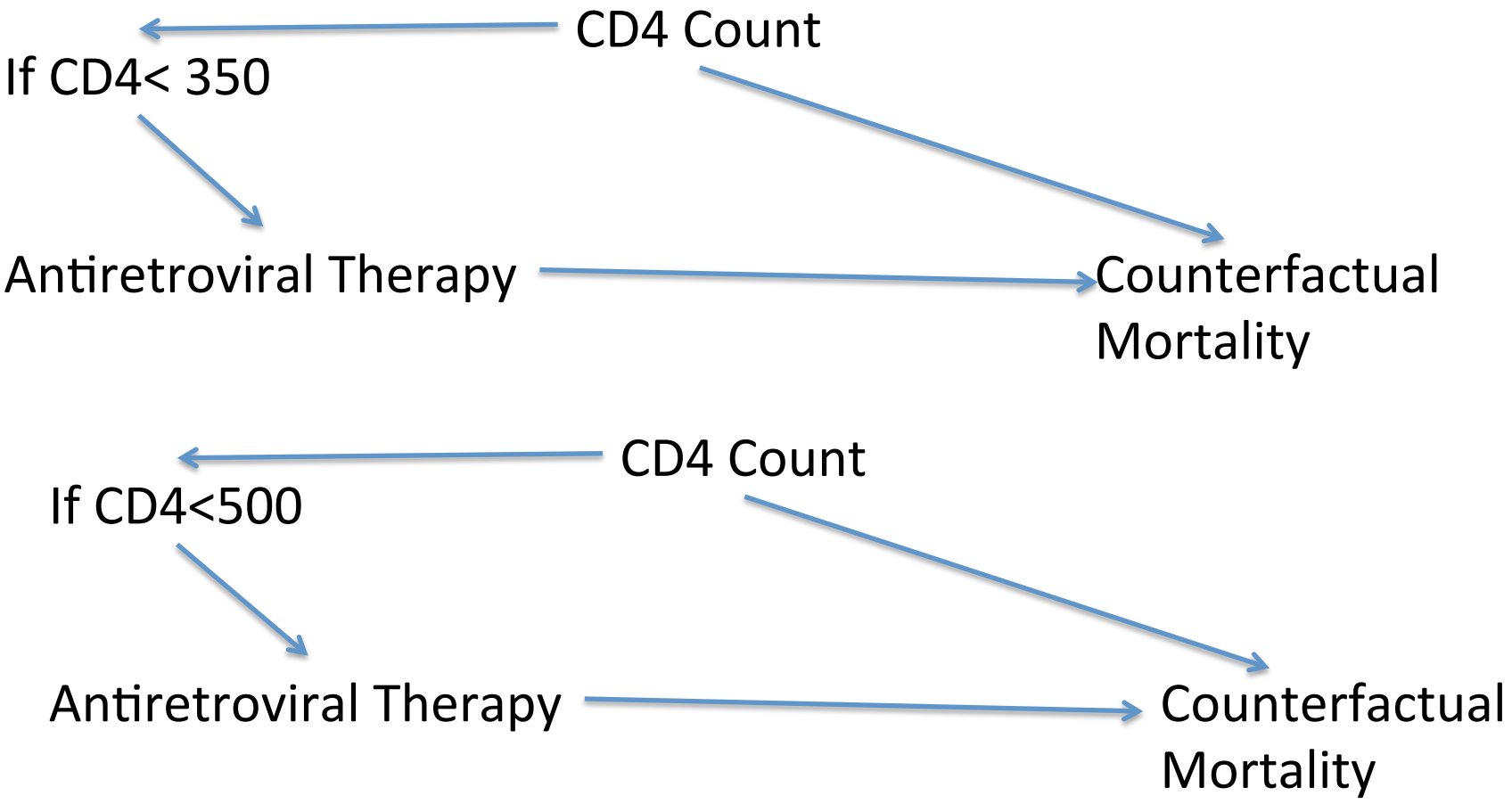
# Counterfactual Interventions

- Example: Simple static interventions



# Very flexible- Make the target quantity match the question

- Example: Dynamic Regimes



# Many more options!

- Longitudinal Effects
  - Effect of Antiretroviral Therapy (ART) over time
- Direct Effects
  - Effect of ART not mediated by CD4 changes
- Missing Data and Censoring
  - Interventions to prevent losses to follow up
- Stochastic Interventions
  - Shift the delay from diagnosis to ART start
- Can also summarize the distribution of the counterfactual outcomes in lots of ways…

# 3. Specify **Observed Data** and its link to the Causal Model

Causal Model:
What we know

Observed Data:
What we Measure

- Causal model describes the set of processes that may have given rise to the observed data
- Implies the set of distributions possible for the observed data: The Statistical Model

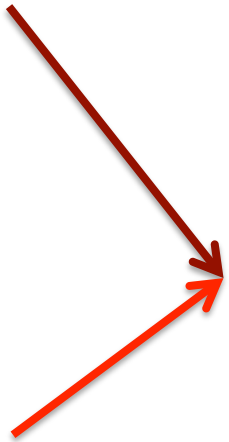
# All statistical models are not wrong…

- Statistical Models should represent real knowledge
- Good Practice: Choose an "appropriate" statistical model…
- Causal Framework: Can help us choose statistical models that reflect our uncertainty
  - Often put no restrictions on the joint distribution of our observed variables: Non-parametric
  - Sometimes we do have real knowledge… By all means use it!

# A Roadmap….

1. **Causal Model**
Representing background knowledge and uncertainty

**3. Observed Data**
Process that generated the data described by the causal model



**Statistical Model**
Possible distributions for the Observed data

**2. Question**
Translate the scientific question into a formal causal quantity (using counterfactuals)

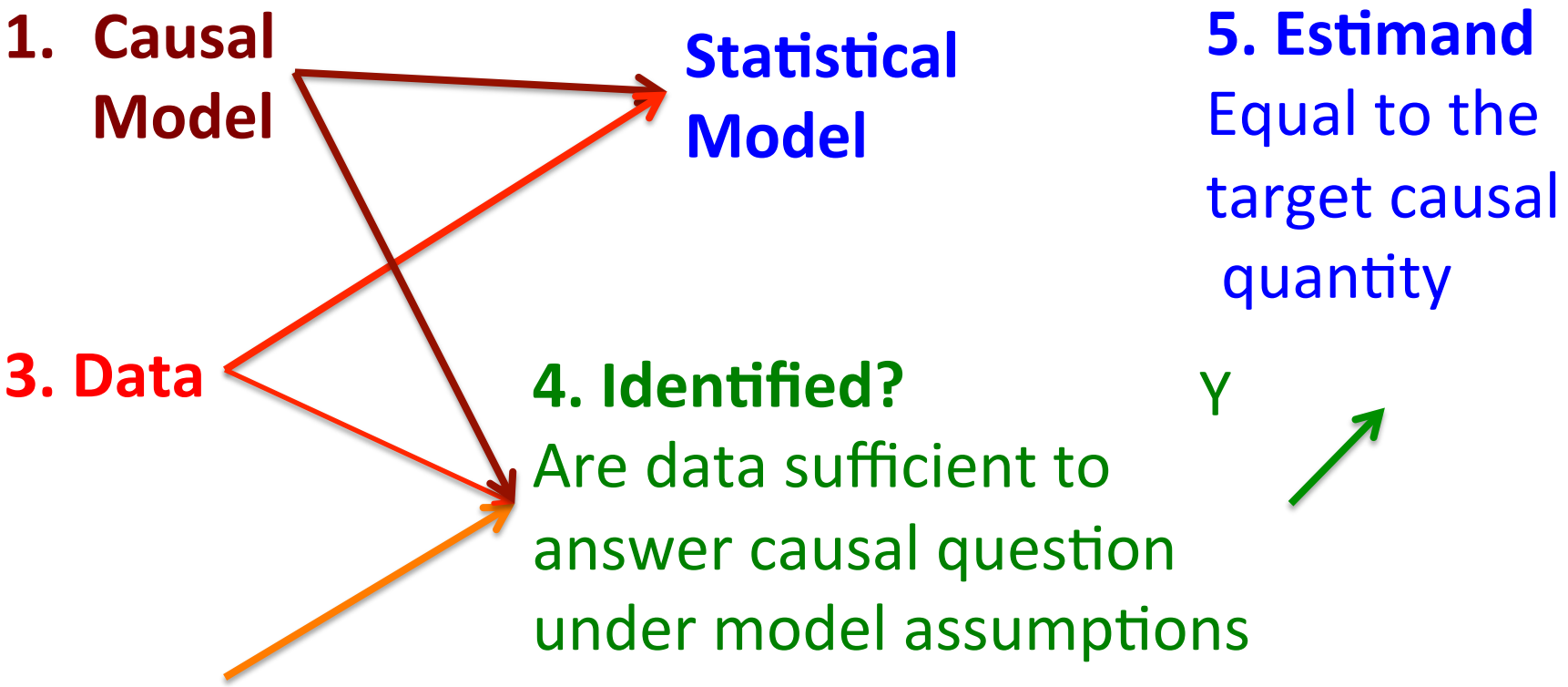
# **4. Identify**.
# Knowledge + Data Sufficient?

- Our question is stated of in terms of counterfactual quantities
  - What would things have looked like under different conditions
- Can we translate our target causal quantity into a statistical parameter?
  - A parameter of the observed data distribution, or "**estimand**"
- If so, which estimand?

# Intuition only gets you so far….

- Good Practice: "Control for confounding"
- Causal Framework
  - Which variables to "control for"
    - Adjusting for some pre-treatment variables is harmful!
  - Even when no single adjustment set is sufficient, your target parameter may still be identified
    - Longitudinal effects and time dependent confounding
    - Effect mediation and direct effects
    - Instrumental variables…
    - Many more…..
- **Many of these estimands are not intuitive**

# A Roadmap….


1. Causal Model
Statistical Model
5. Estimand
Equal to the target causal quantity
3. Data
4. Identified?
Are data sufficient to answer causal question under model assumptions
Y
2. Question
Translate the scientific question into a formal causal quantity (using counterfactuals)

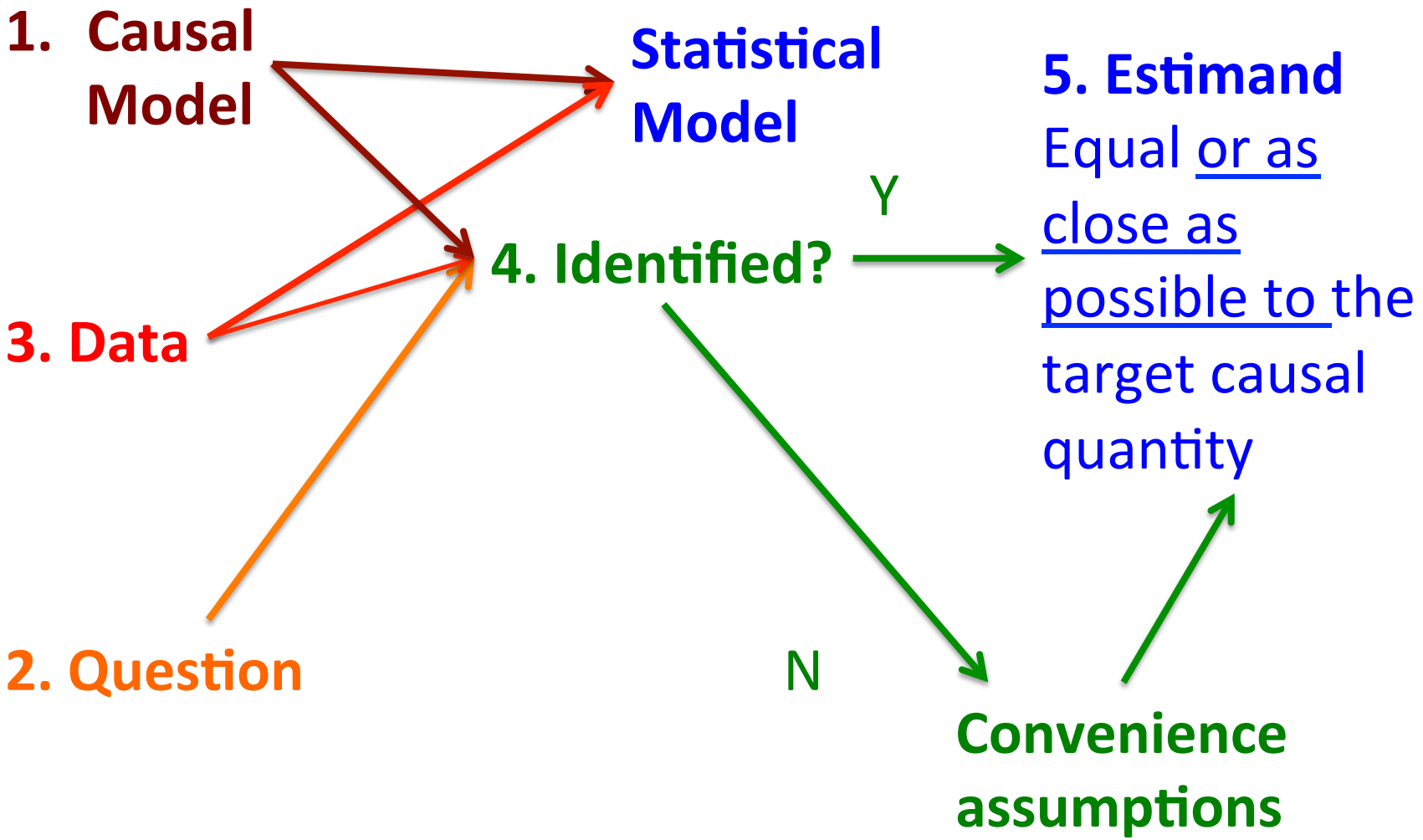
# In many (most?) cases, data + model are NOT sufficient

- Good Practice
  - Get more data
  - Do the best job you can with data you have, and understand limitations
- Formal Causal Framework:
  - Which data/how to change design
  - What additional assumptions are needed: "**convenience assumptions**"?
  - Estimand that comes closest to answering our question with the data we have

# Example: "Convenience assumptions"

- Physicians decided whether to prescribe ART based in part on some variables we measured
  - Ex. CD4 count, Homelessness
- We suspect that they also prescribed ART based on variables we did not measure
  - Ex. Perceived ability to take the drugs on time
- In order to proceed with estimation, we make a convenience assumption
  - Formal version of "no- unmeasured confounding"

# A Roadmap….


1. Causal Model
Statistical Model
5. Estimand
Equal or as close as possible to the target causal quantity
Y
4. Identified?
3. Data
2. Question
N
Convenience assumptions

# 6. Estimate

- Causal framework got us here…but this step is purely statistical
- Many choices of estimator
- This class
  - Regression-based approaches
  - Inverse probability weighting
  - Targeted maximum likelihood estimation

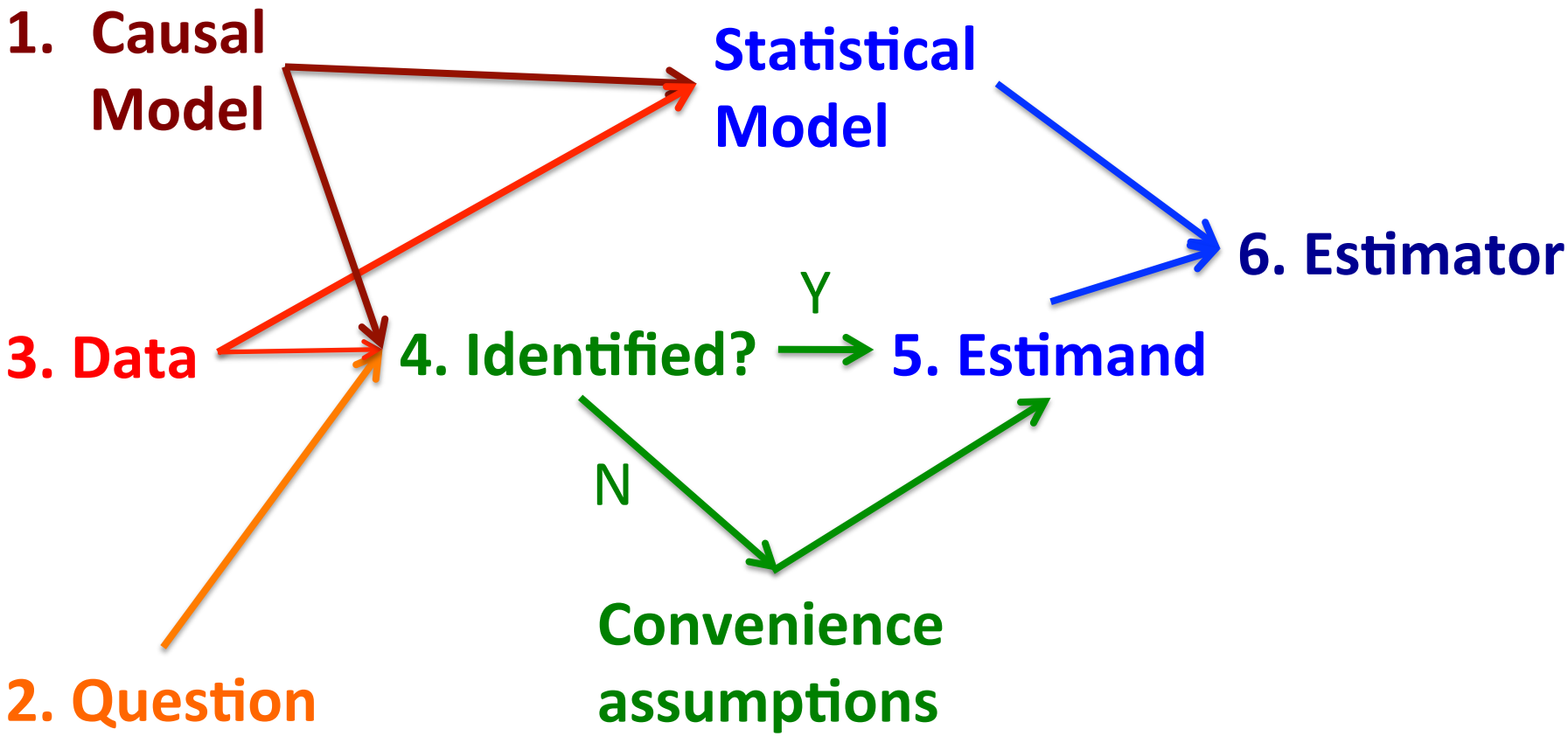
# Many Choices of Estimator

- The **estimator should respect knowledge**, as expressed in the statistical model
  - This is not the point to introduce new assumptions
  - Ex: Assume linear relationship between CD4 and mortality
- Estimators have different statistical properties
  - Bias, variance, robustness, consistency….
  - Given a statistical model and estimand, we can study these properties and pick the best for our problem

# Causal inference can be hard, but so can statistics

- Good Epidemiological Practice: The question should drive the statistical analysis
- Causal Framework : Knowledge + data + question often = hard statistical problems
- Many estimands do not correspond to a coefficient in a single regression
  - Complexity is not an end in itself
  - Complex statistical methods are sometimes necessary to get the best possible answer to real world questions

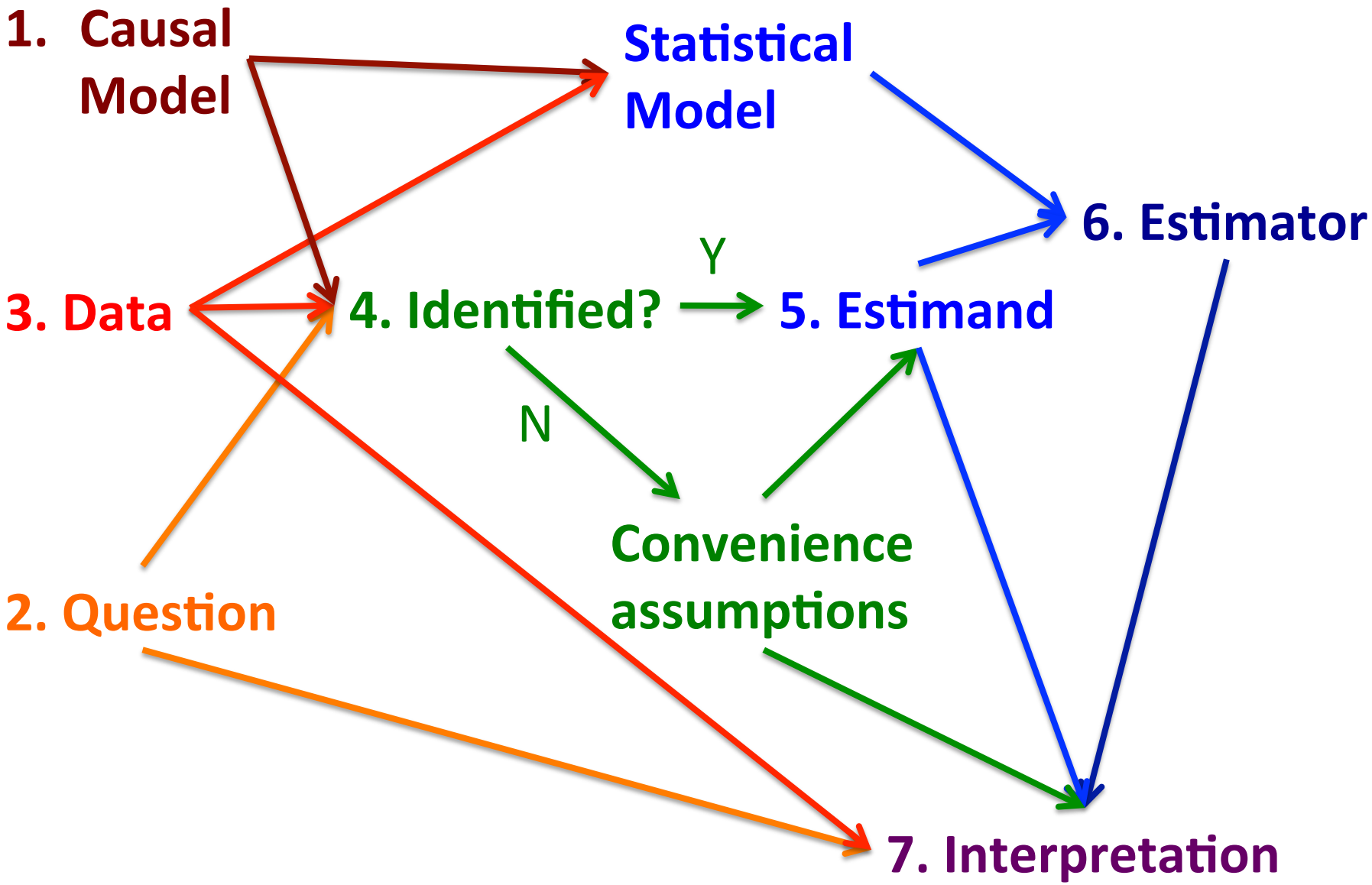
# A Roadmap….


1. Causal Model
Statistical Model
6. Estimator
Y
3. Data
4. Identified?
5. Estimand
N
Convenience assumptions
2. Question

# 7. **Interpret** Results

- Statistical Interpretation?
  - With correct statistical model and good estimator
- Causal Interpretation?
  - If causal model + convenience assumptions are true
  - Makes explicit what these are
- Impact of a real world intervention?
  - More assumptions…
    - Stability of conditions and population
    - Interference
    - Correspondence between hypothetical and real world intervention…

# A Roadmap….


1. Causal Model
Statistical Model
6. Estimator
Y
3. Data
4. Identified?
5. Estimand
N
Convenience assumptions
2. Question
7. Interpretation


*"As we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns -- the ones we don't know we don't know…."*

- Causal frameworks used well
  - Keep us uncomfortably aware of how little we know
  - While not freezing us into panicked inaction
- A systematic approach to using them can help statistics improve public health