

Bridging statistical and mathematical modelling

or

The most important things I could think to teach
you about data, maths/stats, and infections in
90 minutes given your other lectures

Chris Wymant

github.com/ChrisHIV/teaching

These materials + more:
maths refresher, statistical modelling
refresher, Stan, how to write a paper



Terminology apology

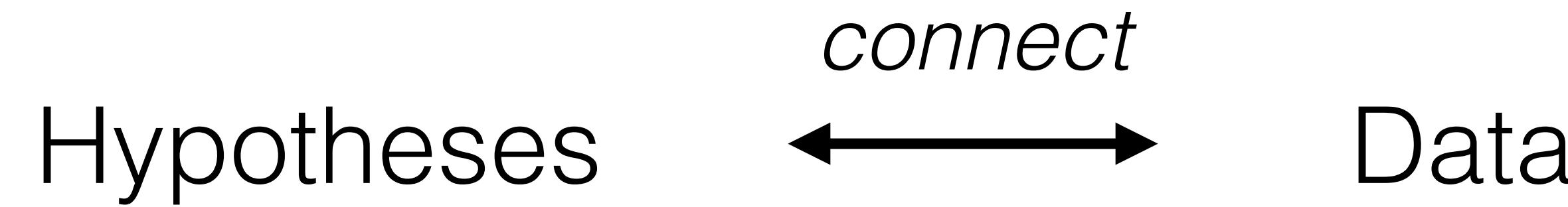
By statistical modelling I actually only mean likelihood-based statistical modelling, using my own experience. Comments may generalise to other types of statistics, machine learning etc. YMMV

Lecture structure

1. Revisiting probability
2. Statistical and/or mathematical modelling
3. Counterfactuals
4. Random effects case study: HIV immune system decline

Part 1: Revisiting probability: Feeling at home with the basic laws

Your job (a subjective simplification)



Probability is *the* natural language connecting hypotheses with (quantitative) data.

Learn to speak that language well.

Probability fundamentals: your lectures so far

Probability refresher - Some terminology

- Sample point:
 - A possible outcome of a random experiment
- Sample space:
 - The collection of all sample points
 - i.e. all possible outcomes of a random experiment
 - often called Ω

Probability refresher - Some terminology

- Event:
 - Collection of possible outcomes
- Distribution:
 - Assignment of values to events satisfying the probability axioms

Probability refresher – The three axioms of probability

1. The probability of an event A is a value $P(A)$ between 0 and 1
2. $P(\Omega)=1$
3. If A_1, A_2, \dots are mutually exclusive events, then

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$

Probability refresher – Random variables

- Variables whose possible values are numerical outcomes of a random phenomenon
- Subject to variation due to chance
- A probability is associated to each of the possible values the variable can take
- There are two types of random variables:
 - Discrete
 - Continuous

Probability refresher – Discrete random variables

- Can take only a number of distinct values
- In general (but not always), they are counts
 - e.g. number of defective bulb lights in a box
- Have a probability mass function
 - Assigns probabilities to the possible values of the random variable $P(X=x)$, where X is the random variable and x the possible value

! The probabilities must follow some requirements:

$$p_i \geq 0 \quad \forall i \quad \text{Probabilities must be positive numbers}$$

$$\sum_{i=1}^n p_i = 1 \quad \text{The sum of the probabilities must be 1}$$

Probability refresher – Discrete random variables

! The probabilities must follow some requirements:

$$p_i \geq 0 \quad \forall i \quad \text{Probabilities must be positive numbers}$$

$$\sum_{i=1}^n p_i = 1 \quad \text{The sum of the probabilities must be 1}$$

Values	$P(X=x)$
1	0.20
3	0.50
7	0.30

The probability of the random variable being exactly 1 is 0.2

Probability refresher – Continuous random variables

- Can take an infinite number of possible values
- In general, they are measurements
 - e.g. the time required to run a mile
- Have a probability density function
 - specifies the probability that the value of the random variable falls within a specific range
 - it is represented by the area under the density function (integral)

! The PDF needs to satisfy the following requirements:

$$\text{Probabilities must be positive} \quad f(x) \geq 0 \quad \forall x \quad \int_{-\infty}^{+\infty} f(x) dx = 1 \quad \text{The area under the entire density curve must be 1}$$

Probability refresher – Continuous random variables

! The PDF needs to satisfy the following requirements:
 Probabilities must be positive $f(x) \geq 0 \quad \forall x \quad \int_{-\infty}^{+\infty} f(x) dx = 1 \quad$ The area under the entire density curve must be 1

- For example, considering a random variable X measuring the depth of a lake in various spots $f(x)$
- The probability that X takes on a value in the interval $P(a \leq X \leq b)$ is the area under the density function curve
- The probability that X takes an exact value of x is 0



Definition of probability, $0 \leq P(A) \leq 1$, sum to 1



Probability definitions

$P(A)$ = probability of A.

$$0 \leq P(A) \leq 1$$

Frequentist probability is defined only for random variables: outcomes of repeatable stochastic experiments. e.g. data, but not parameters.

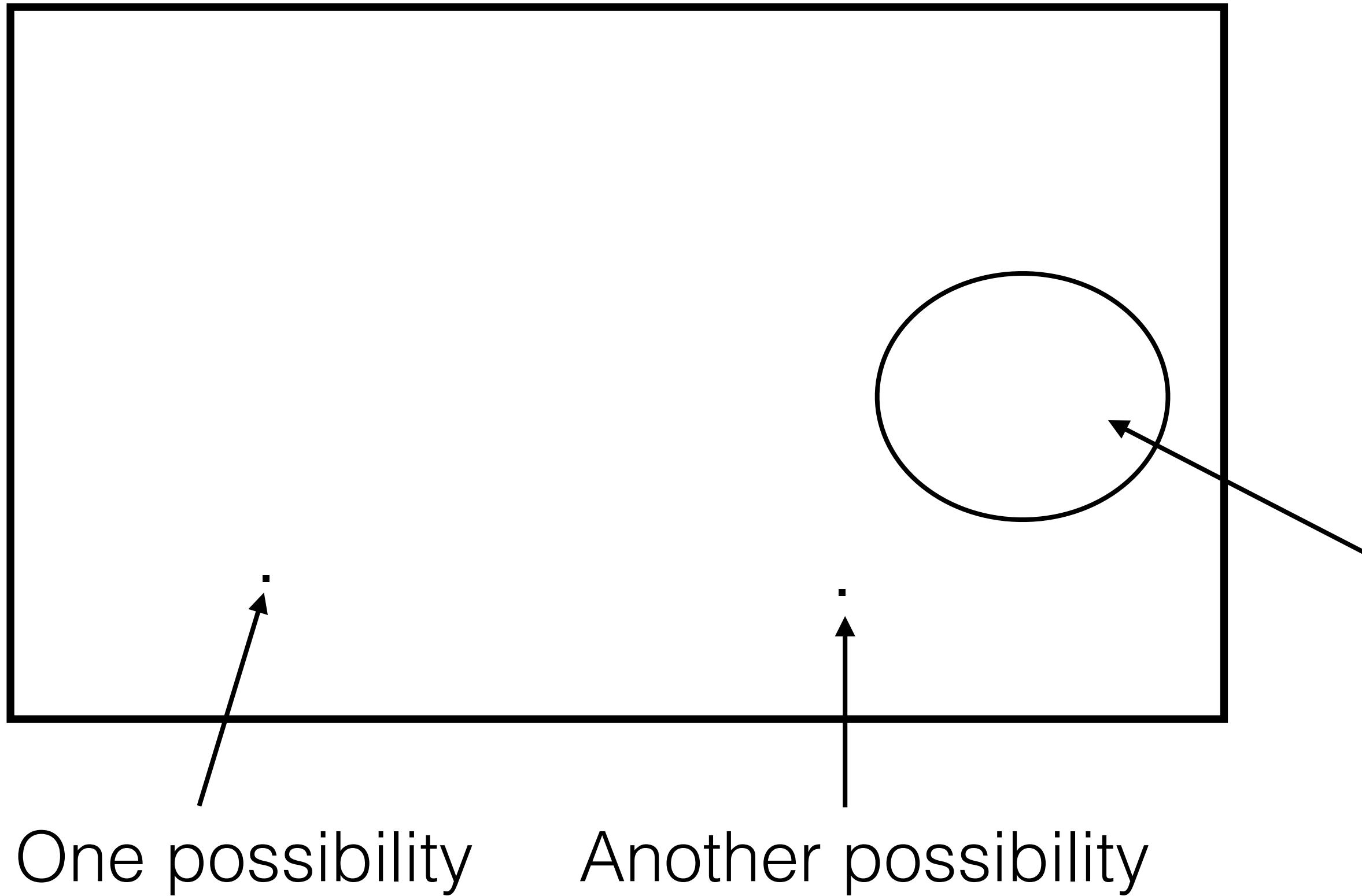
$$P(A) = \lim_{N \rightarrow \infty} \left(\frac{\text{number of } A \text{ outcomes}}{N} \right)$$

Bayesian probability: extend the definition above to *also* include degree of belief, allowing quantification of our uncertainty. Can now talk about $P(\text{parameter})$.

(Alternatively, some say Bayesians extend the concept of ‘random variables’ to include parameters, as well as data. Less clear.)

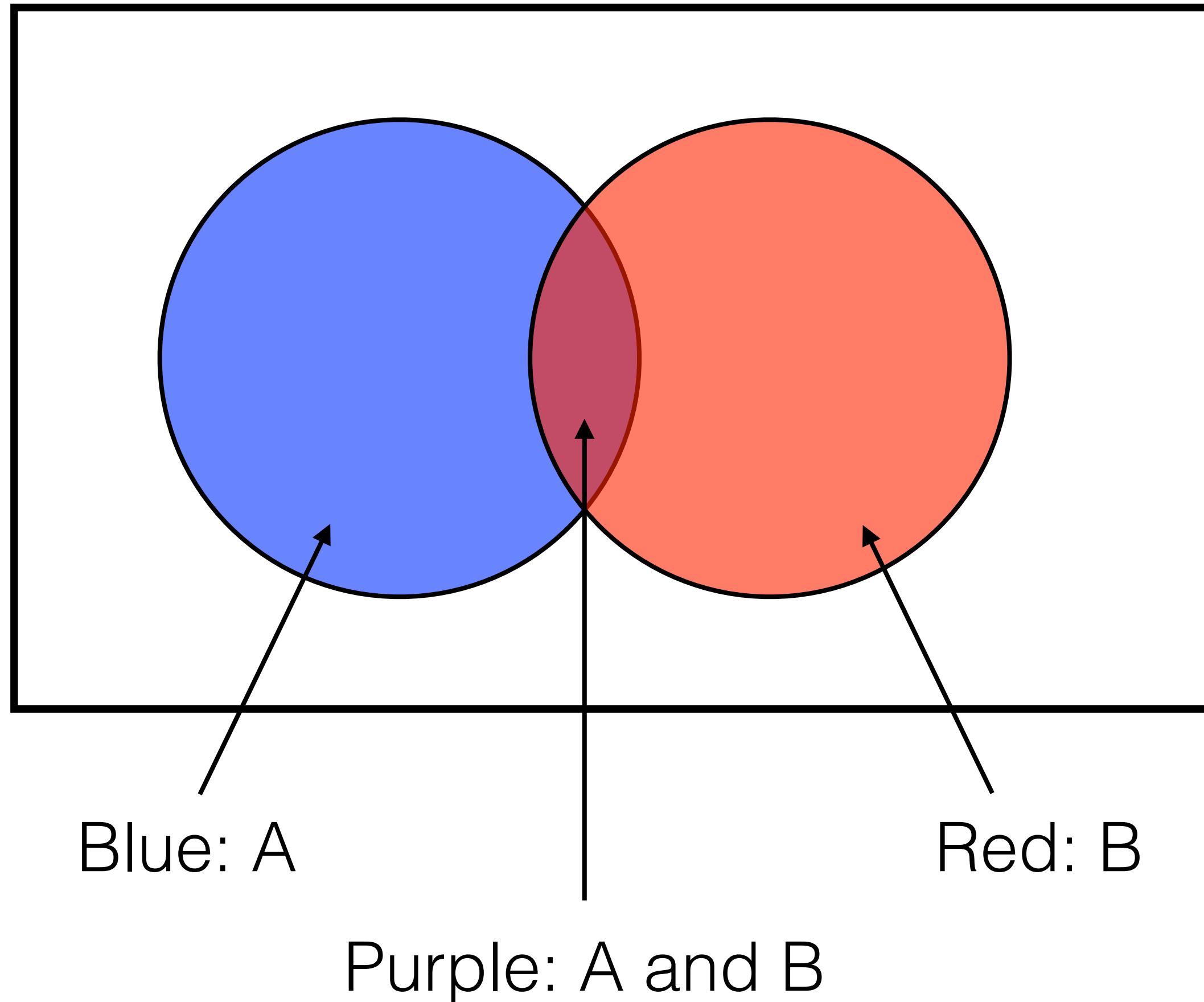
Venn diagrams

Space of everything that's possible



A collection of possibilities that we group together and consider as a single composite possibility. Area represents its probability.

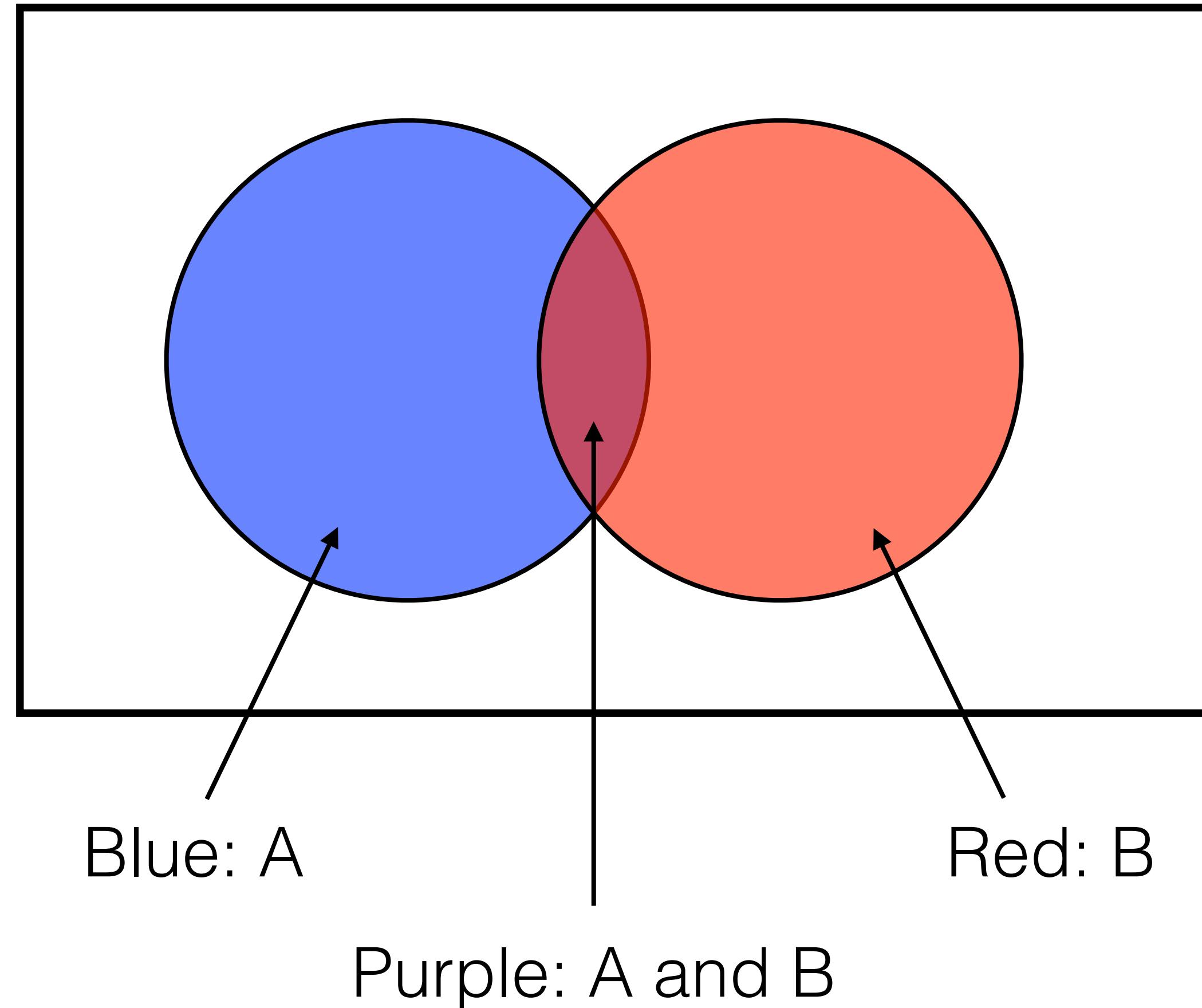
“and”, “or”



$$P(A) = P(A \text{ and } B) + P(A \text{ and not-}B)$$
$$P(B) = P(B \text{ and } A) + P(B \text{ and not-}A)$$

$$\therefore P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

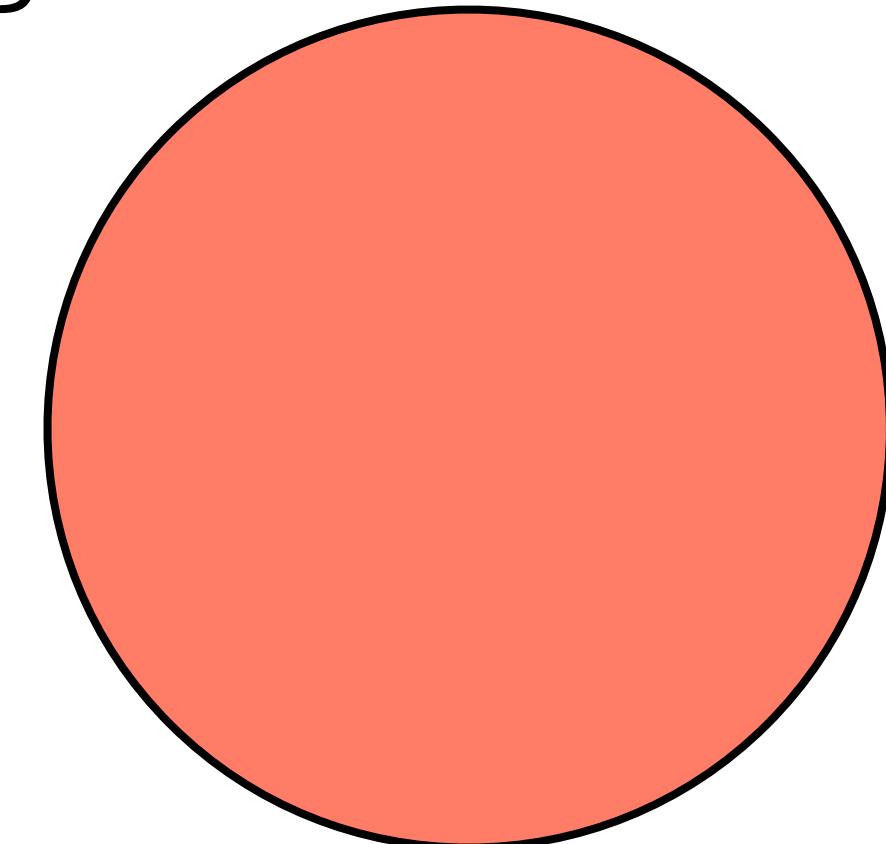
Conditional probability



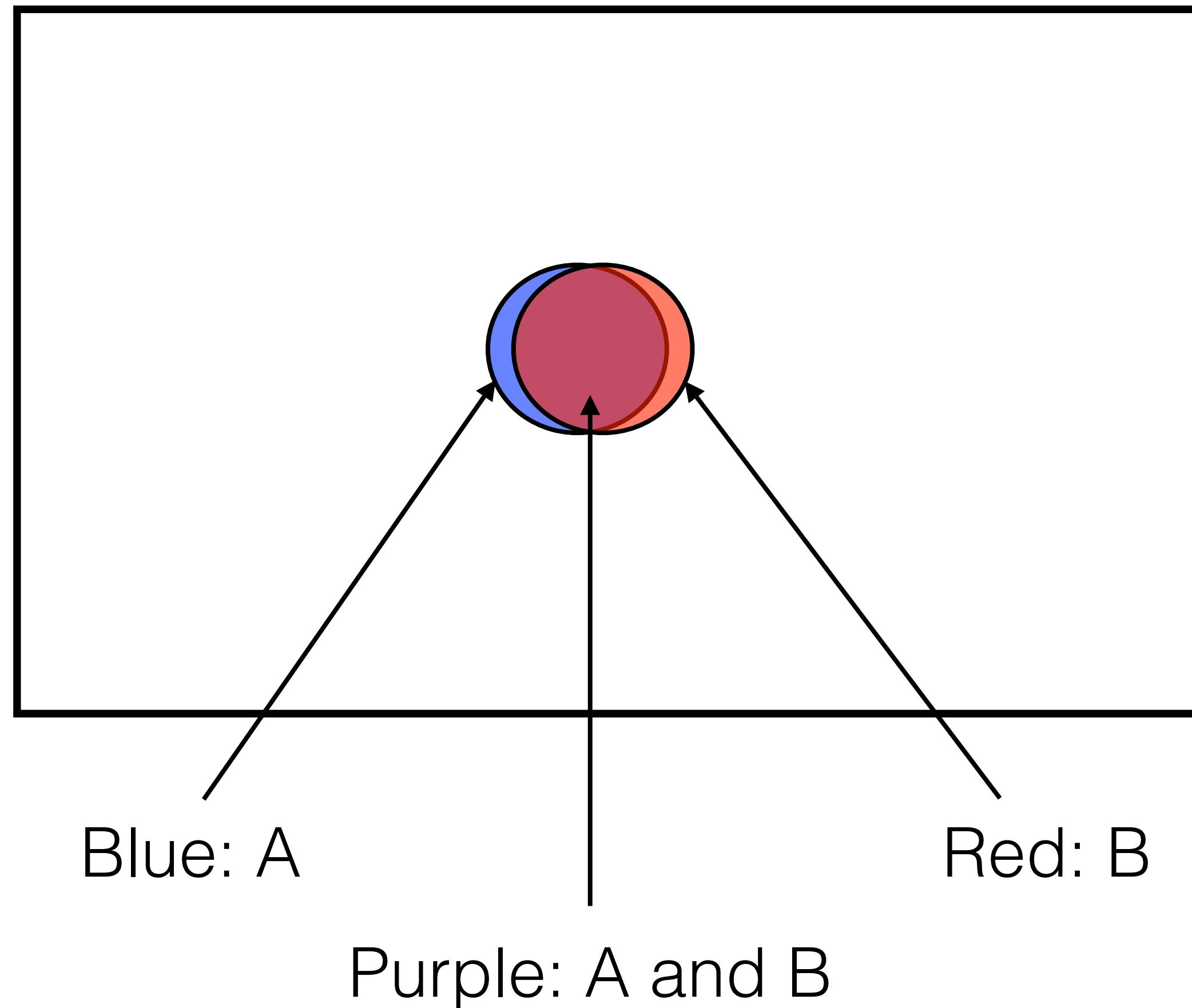
| means “given (that)” or
“conditional upon”

Definition: $P(A|B) = P(A \text{ and } B) / P(B)$
= fraction of outcome space that's A
within the space that's B

$$= \text{area of } \text{Purple} / \text{area of } \text{Red}$$



Conditional probability: given what?



“Lies, damn lies, and statistics”
or, “the $|$ operator changes
everything”

$P(A)$ and $P(A|B)$ can be very
different!

Ask (them/yourself) *given what*
every time you encounter a
probability.

Independence

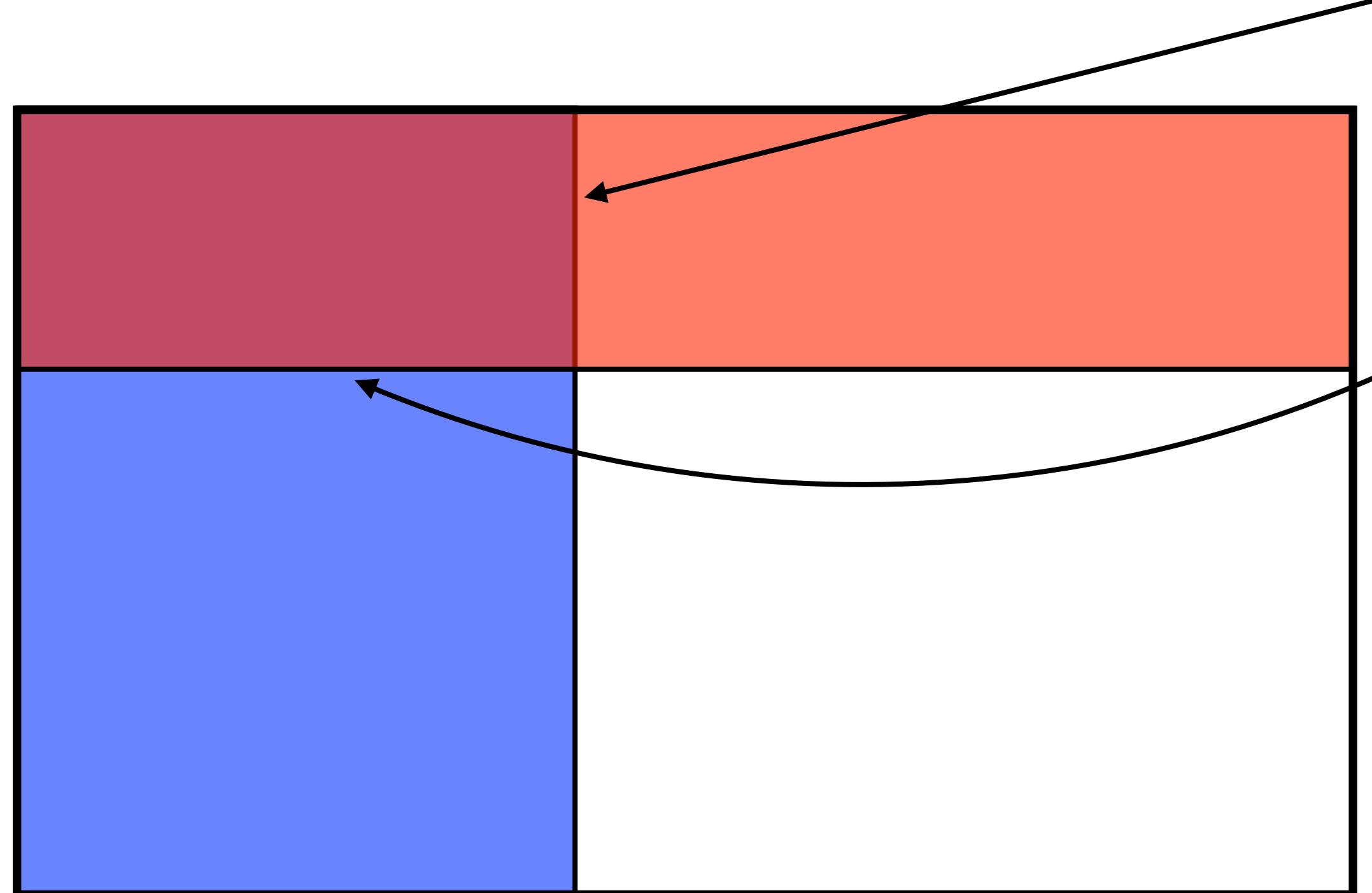
A and B are outcomes of independent events

\Leftrightarrow knowing B is wholly uninformative regarding A and vice versa

$\Leftrightarrow P(A | B) = P(A)$ and $P(B | A) = P(B)$

$\Leftrightarrow P(A \text{ and } B) = P(A)P(B)$

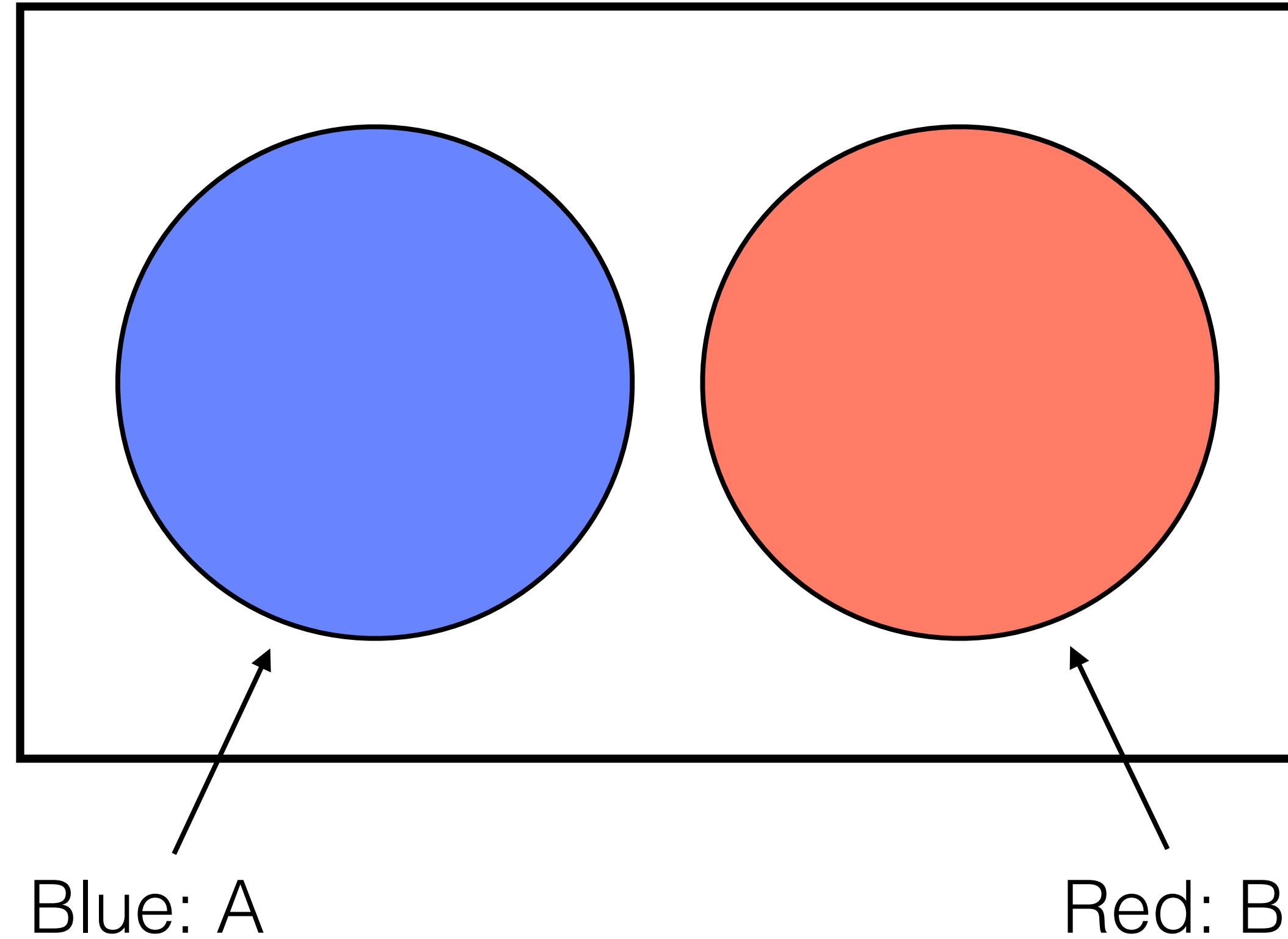
Fraction of *red* that's (also) *blue*
= fraction of *everything* that's *blue*



Fraction of *blue* that's (also) *red*
= fraction of *everything* that's *red*

Very common for likelihoods to be
built using an assumption of
conditional independence:
 $P(\text{all data} | \text{params}) =$
 $P(\text{datum 1} | \text{params}) \times$
 $P(\text{datum 2} | \text{params}) \times \dots$

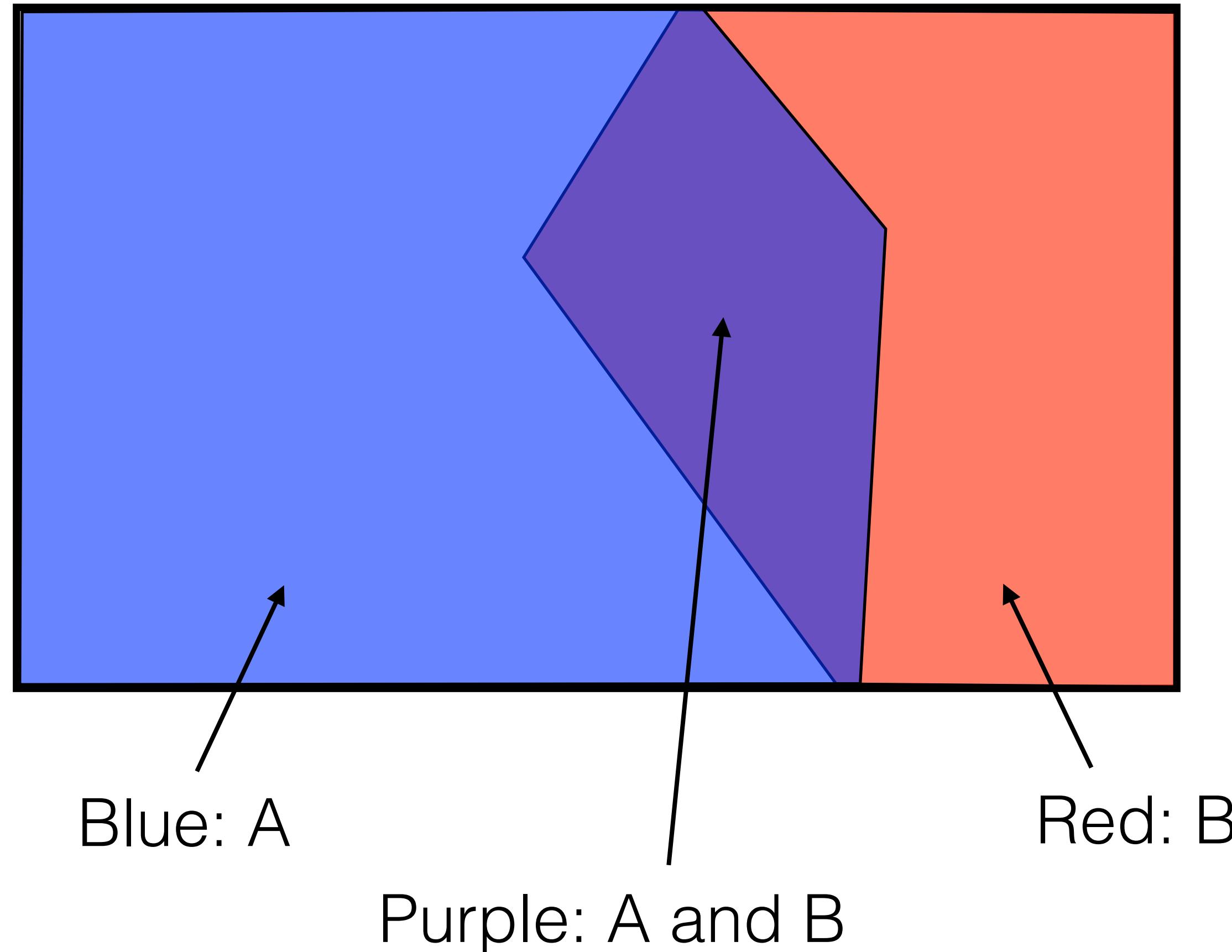
Mutual exclusivity



A and B are “mutually exclusive”
↔ at most one of them is true
↔ $P(A \text{ and } B) = 0$
↔ $P(A \text{ or } B) = P(A) + P(B)$

Because, recall, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

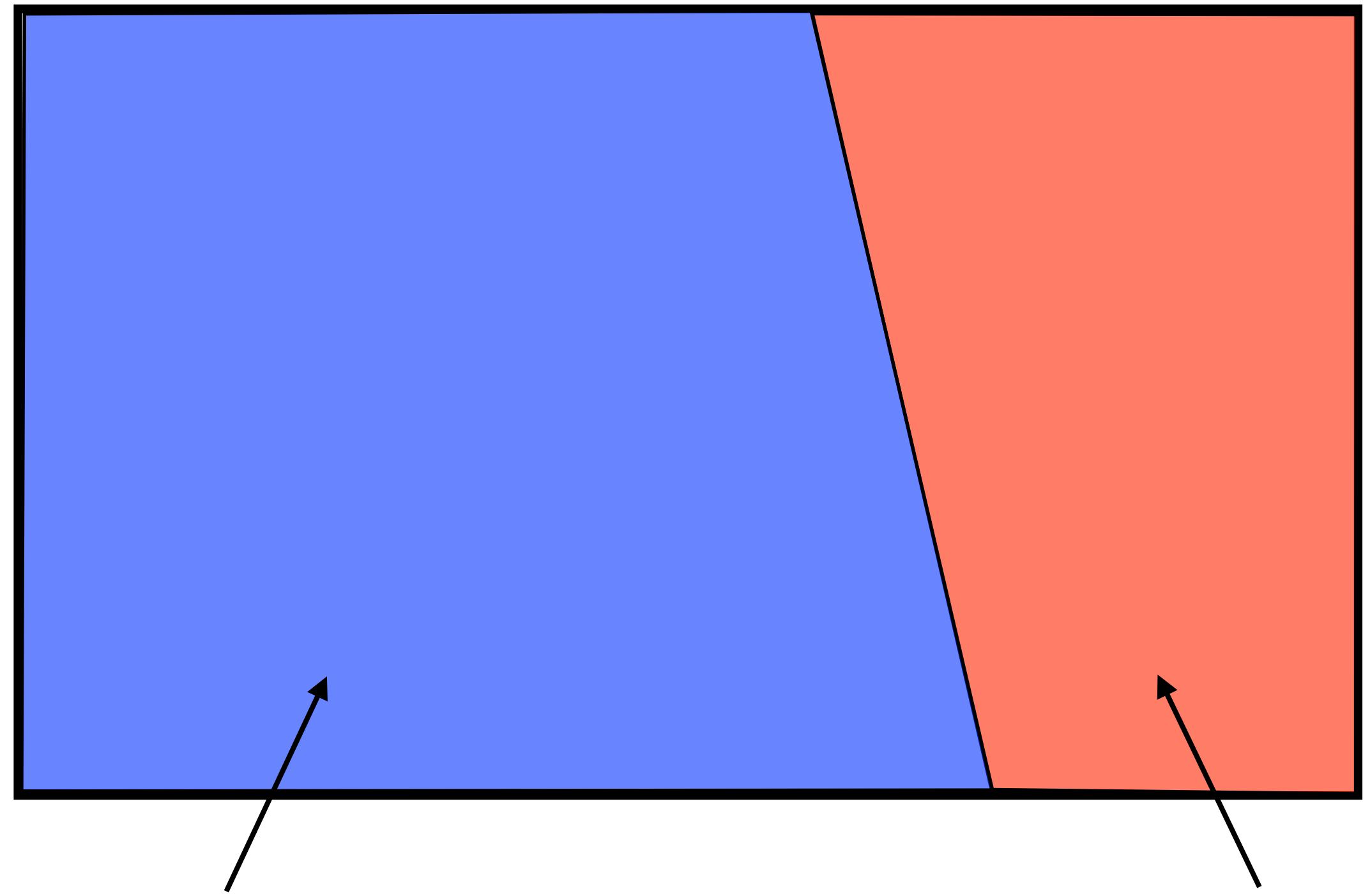
Collective exhaustion



A and B are “collectively exhaustive”
 \Leftrightarrow at least one of them is true
 $\Leftrightarrow P(A \text{ or } B) = 1$
 $\Leftrightarrow P(\text{neither } A \text{ nor } B) = 0$

Because, recall, the sum of probabilities is 1

Mutually exclusive and collectively exhaustive



A and B are mutually exclusive and collectively exhaustive (ME&CE)

\Leftrightarrow at most one of them is true AND at least one of them true

\Leftrightarrow exactly one of them is true

$\Rightarrow P(A) + P(B) = 1$

Example sets of possibilities for the result of rolling a die once:

Because, recall, ME $\Leftrightarrow P(A \text{ or } B) = P(A) + P(B)$,
and CE $\Leftrightarrow P(A \text{ or } B) = 1$

	Mutually exclusive	Not mutually exclusive
Collectively exhaustive	<ul style="list-style-type: none">Result is 1-3Result is 4-6	<ul style="list-style-type: none">Result is 1-4Result is 4-6
Not collectively exhaustive	<ul style="list-style-type: none">Result is 1Result is 2	<ul style="list-style-type: none">Result is 1-2Result is 2-3

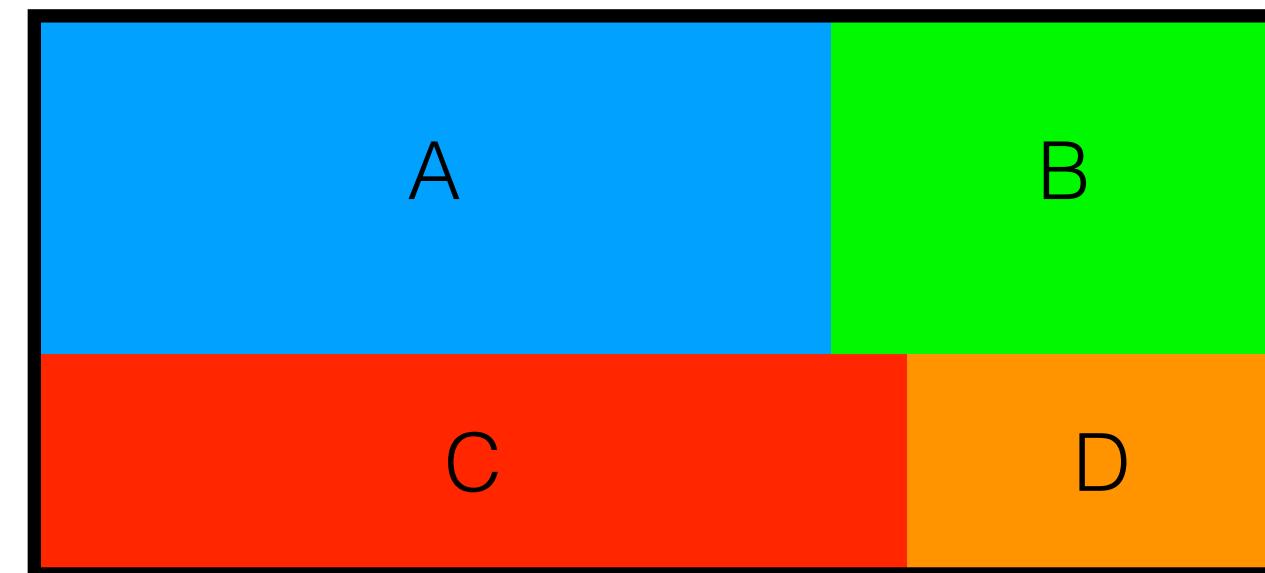
You can take the collection of all things that are possible and split it into ME&CE groups in different ways.

e.g. the day of the week today = Mon, Tue, ... Sun, and
The mean temperature here today is $<10^{\circ}\text{C}$, or $\geq10^{\circ}\text{C}$

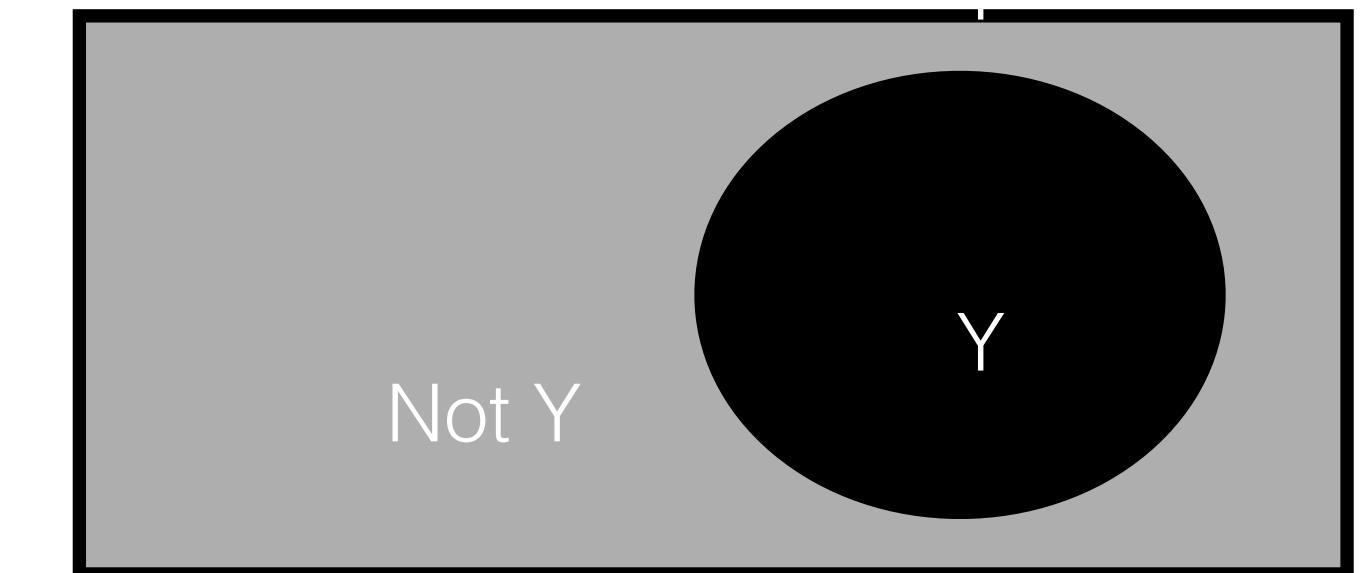
Space of everything that's possible



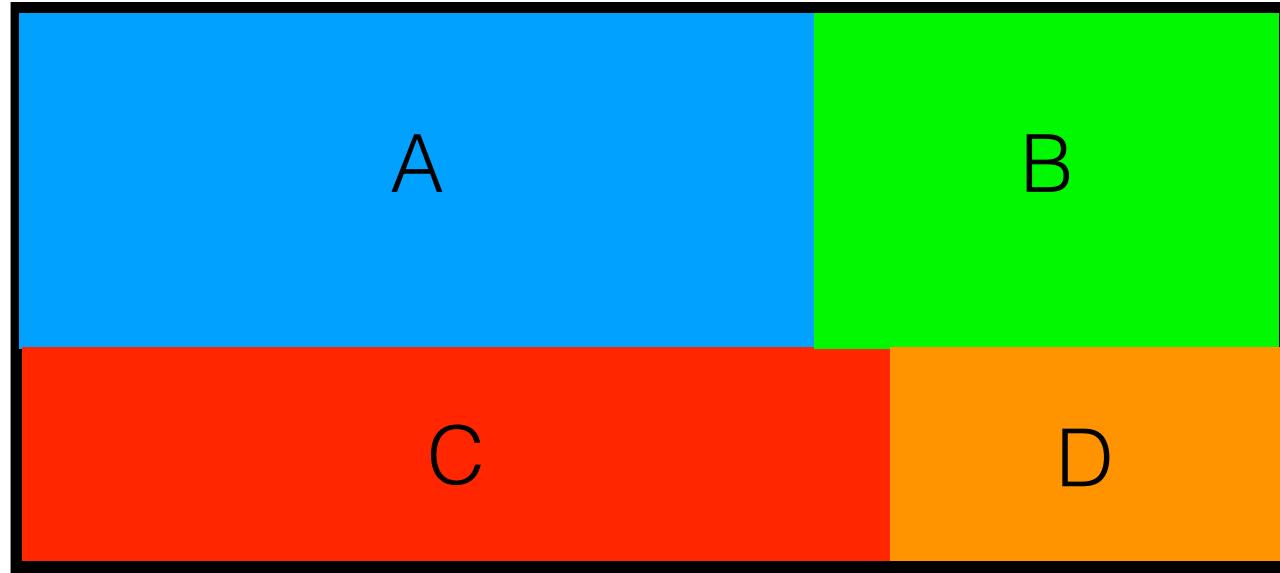
One way of splitting into ME&CE groups: A, B, C or D



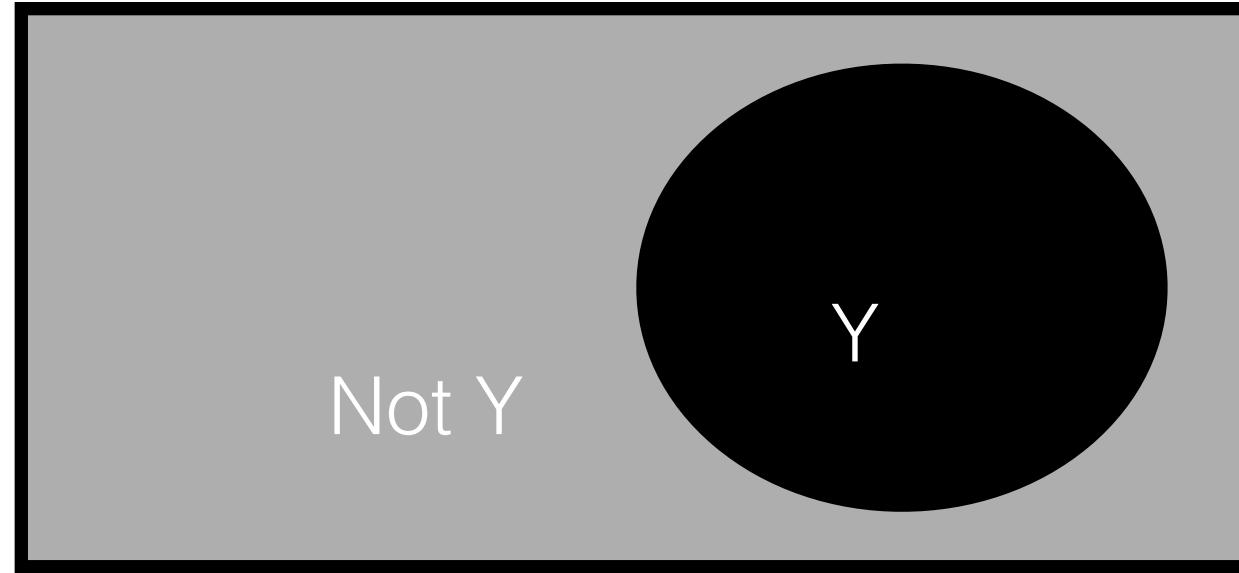
Another way of splitting into ME&CE groups: Y or not Y



One way of splitting into ME&CE groups: A, B, C or D



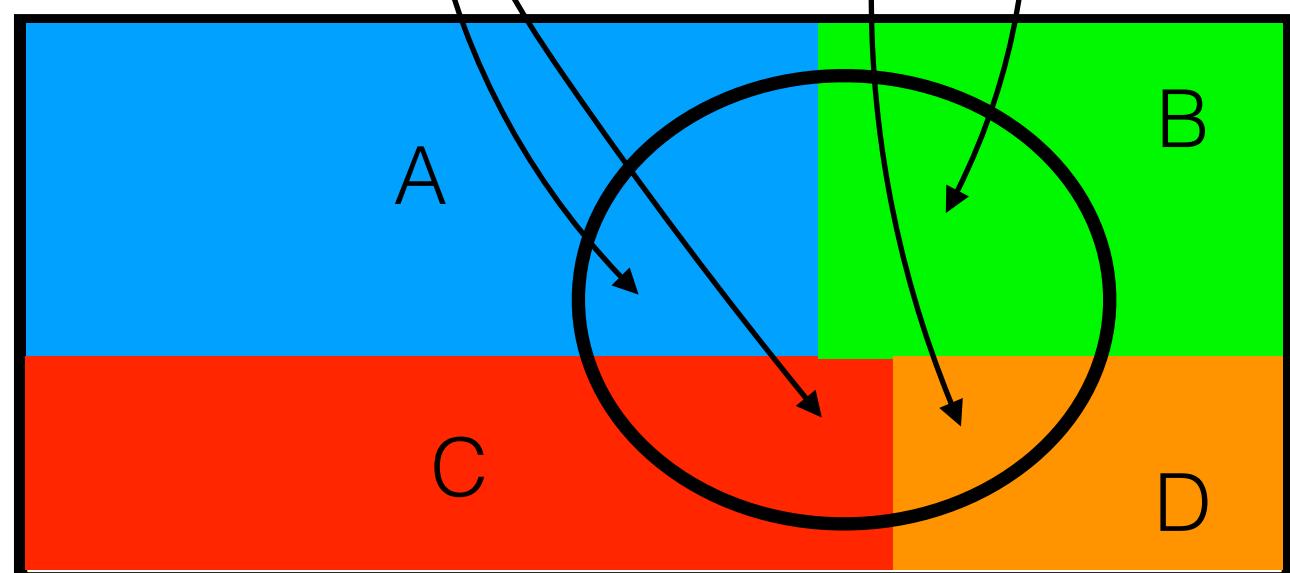
Another way of splitting into ME&CE groups: Y or not Y



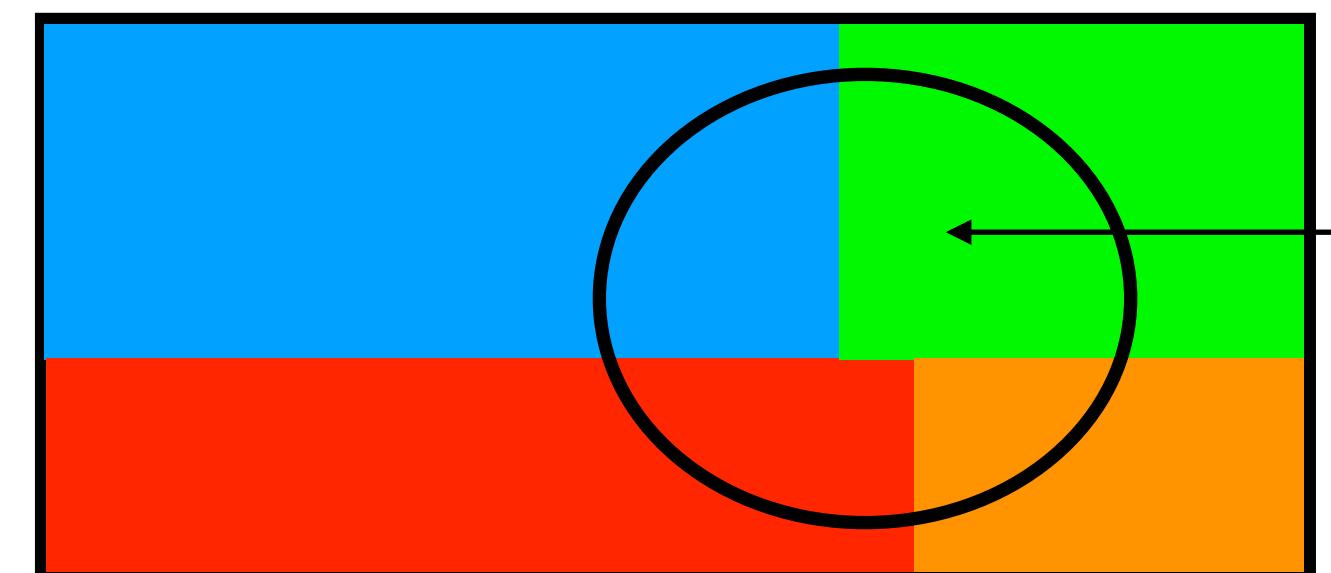
Because exactly one of A, B, C or D is true, we can say

$$P(Y) = P(Y \text{ and } A)$$

$$\begin{aligned} &+ P(Y \text{ and } B) \\ &+ P(Y \text{ and } C) \\ &+ P(Y \text{ and } D) \end{aligned}$$



Because ABCD are ME&CE for the whole space, they are ME&CE for any subset of the space



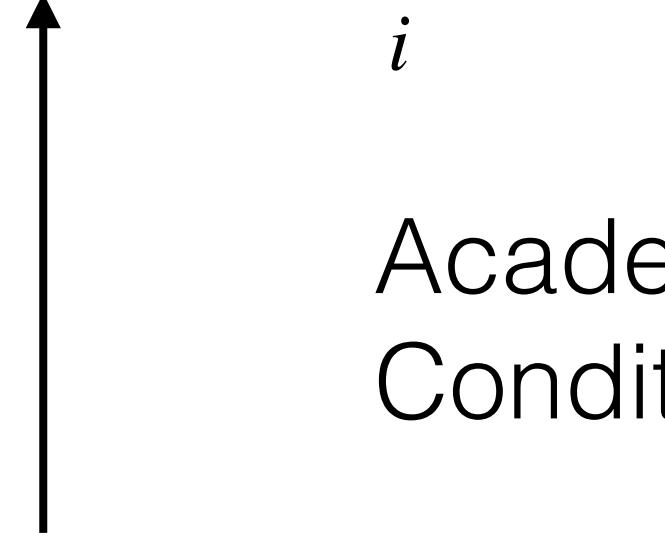
$$\begin{aligned} &P(Y \text{ and } B) \\ &= \text{fraction of whole space that's B} \times \\ &\quad \text{fraction of B that's Y} \\ &= P(B) P(Y | B) \end{aligned}$$

So in the end,

$$\begin{aligned} P(Y) &= P(Y | A) P(A) \\ &+ P(Y | B) P(B) \\ &+ P(Y | C) P(C) \\ &+ P(Y | D) P(D) \end{aligned}$$

Using the *law of total probability*:

- Decide on your possibility of interest, Y ,
- Decide on some way of splitting up the space of everything that's possible into a set of ME&CE possibilities A_1, A_2, \dots, A_N (any way you like except "Y or not Y" which would result in something true but unhelpful)
- then you have

$$P(Y) = \sum_i P(Y|A_i)P(A_i)$$


Academics' favourite answer to anything: "it depends".
Conditional probabilities are easier to calculate.

Decision makers need to make a decision, so "it depends on something we don't know for sure" is not helpful. Remove the dependence on A_i (sometimes called a *nuisance parameter*.)

and its integral equivalent for continuous A rather than discrete A_i .

Summing/integrating over A_i in proportion to $P(A_i)$ is called *marginalising* over A_i .

Sensitivity to unknowns

$$P(Y) = \sum_i P(Y|A_i)P(A_i)$$

Let $P(Y)$ be the degree of belief that the best course of action is Y .
Imagine $P(Y | A_i)$ is narrow for any particular i . How narrow is $P(Y)$?

	$P(Y A_i)$ is <u>insensitive</u> to A_i	$P(Y A_i)$ is <u>sensitive</u> to A_i
$P(A_i)$ is narrow	$P(Y)$ is narrow	$P(Y)$ is moderately broad
$P(A_i)$ is broad	$P(Y)$ is moderately broad	$P(Y)$ is very broad

The Rules of Probability

sum rule $p(X) = \sum_Y p(X, Y)$ (1.10)

product rule $p(X, Y) = p(Y|X)p(X).$ (1.11)

Here $p(X, Y)$ is a joint probability and is verbalized as “the probability of X and Y ”. Similarly, the quantity $p(Y|X)$ is a conditional probability and is verbalized as “the probability of Y given X ”, whereas the quantity $p(X)$ is a marginal probability and is simply “the probability of X ”. These two simple rules form the basis for all of the probabilistic machinery that we use throughout this book.

From the product rule, together with the symmetry property $p(X, Y) = p(Y, X)$, we immediately obtain the following relationship between conditional probabilities

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$
 (1.12)

which is called *Bayes’ theorem* and which plays a central role in pattern recognition and machine learning. Using the sum rule, the denominator in Bayes’ theorem can be expressed in terms of the quantities appearing in the numerator

$$p(X) = \sum_Y p(X|Y)p(Y).$$
 (1.13)

Frequentist and/or Bayesian

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

↑ Bayes' Theorem or Bayes' Rule. *Always true.*

Frequentists can and do use this, though only when both H and E are random variables.

Doing Bayesian inference means H can be a hypothesis or parameter (and E evidence). Useful.

Frequentist and/or Bayesian

Let D denote the hypothesis of having a disease,
let \bar{D} denote “not D” i.e. not having the disease,
and let “+” denote the observation of a positive
test result.

$P(+ | D)$ and $P(+ | \bar{D})$ are just the test sensitivity and 1-specificity.

Bayesian doctor:

- $P(D)$ is the prior probability of their current patient having the disease. Use the fraction of people in the general population.
- $P(D | +)$ is the posterior probability of this patient having the disease. This object *is* the inference drawn.

Frequentist doctor:

- $P(D)$ is the fraction of people in the general population having the disease.
- $P(D | +)$ is the fraction of positive patients that have the disease. Use this object to draw an inference.

$$\begin{aligned} P(D | +) &= \frac{P(+) | D) P(D)}{P(+) } \\ &= \frac{P(+) | D) P(D)}{P(+) | D) P(D) + P(+) | \bar{D}) P(\bar{D})} \end{aligned}$$

(Having marginalised $P(+)$ over
the ME&CE possibilities D and \bar{D})

Chris: skip?

Part 1 summary:

- understand laws of (conditional) probability
- ask *given what* for each probability statement you encounter
- Use statements of conditional probability to build connections between data and hypotheses, and to communicate your method & results

Part 2: Statistical and/or
mathematical modelling

Models

A model: a simplified view of complex reality.

Creating and using models is essentially the only way we understand the world: we must choose which factors are relevant to our question and which are not.

e.g. “Where there’s smoke there’s fire”

e.g. “For disease X, treatment Y is best”

All models are wrong but some are useful.



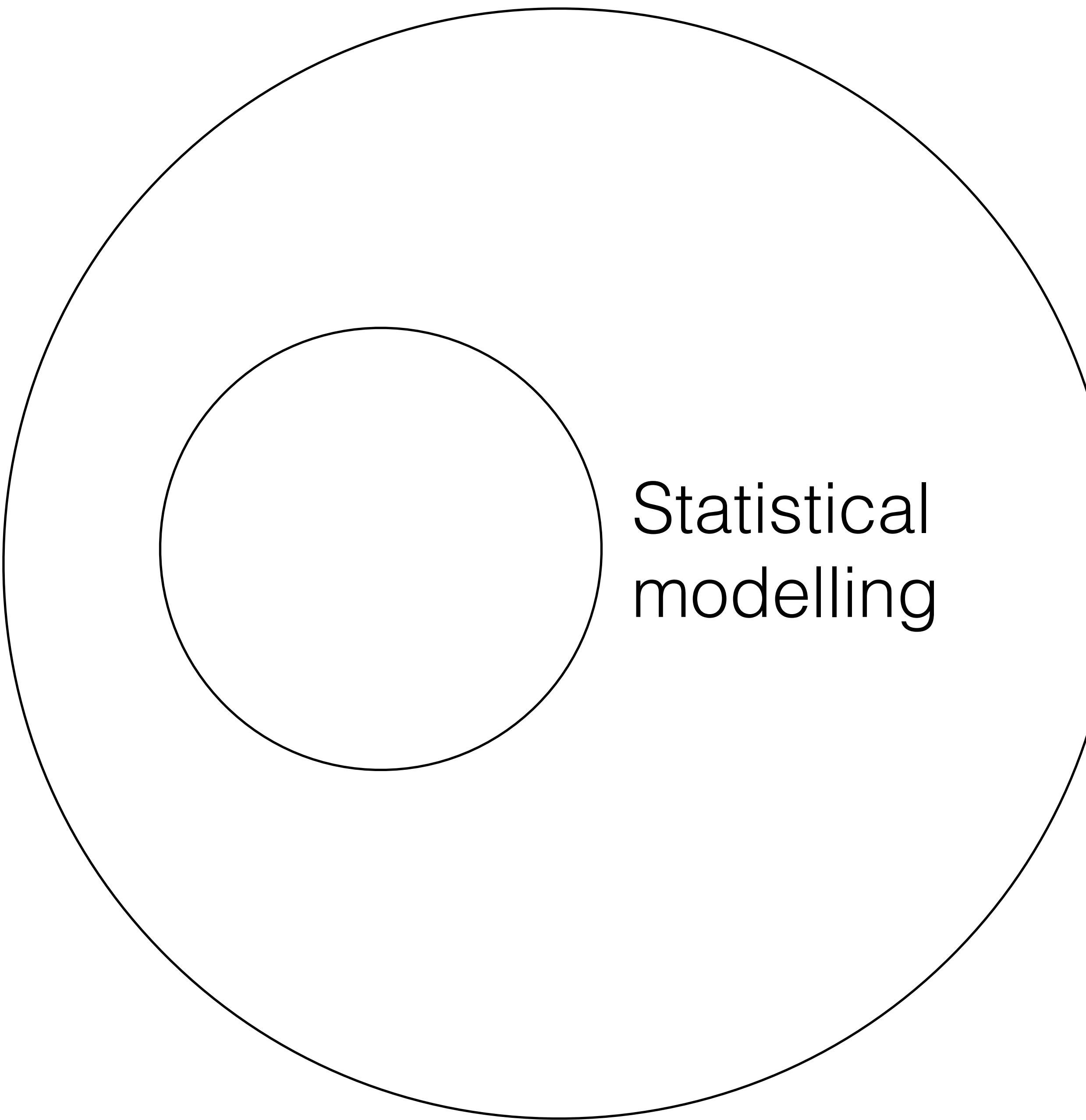
A model of what's in this photo



model
city

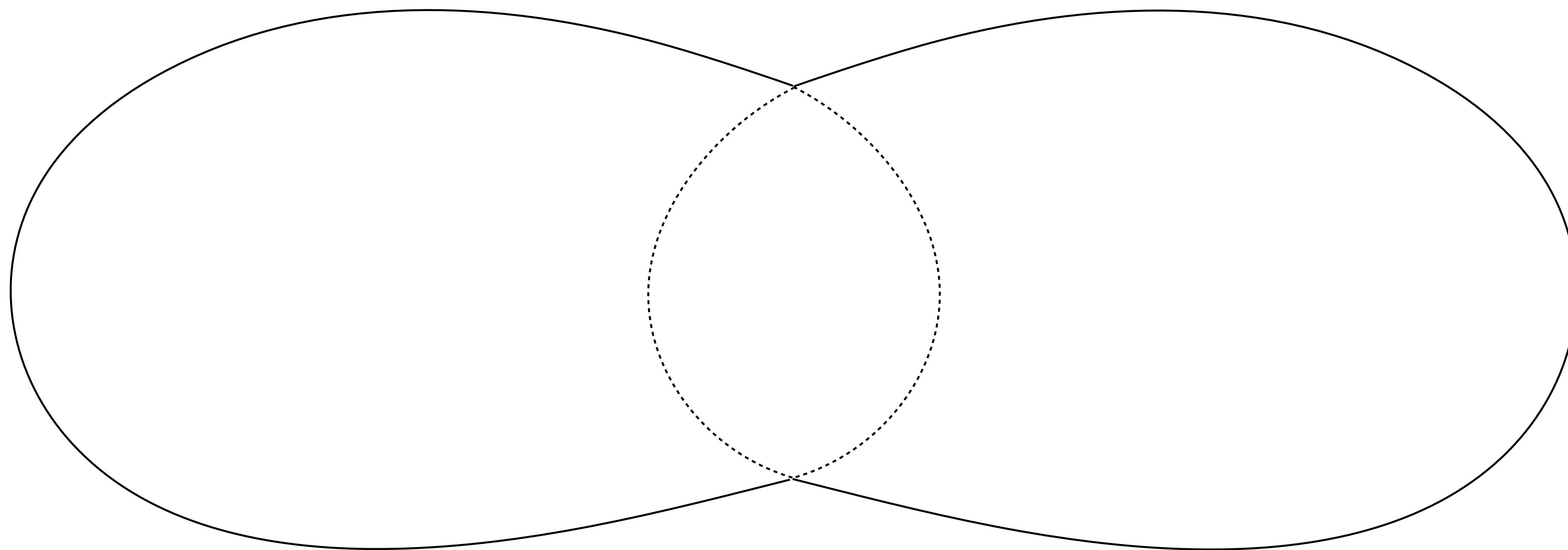
Image: Ittyblox

In theory,



because statistics
uses maths.

In practice,



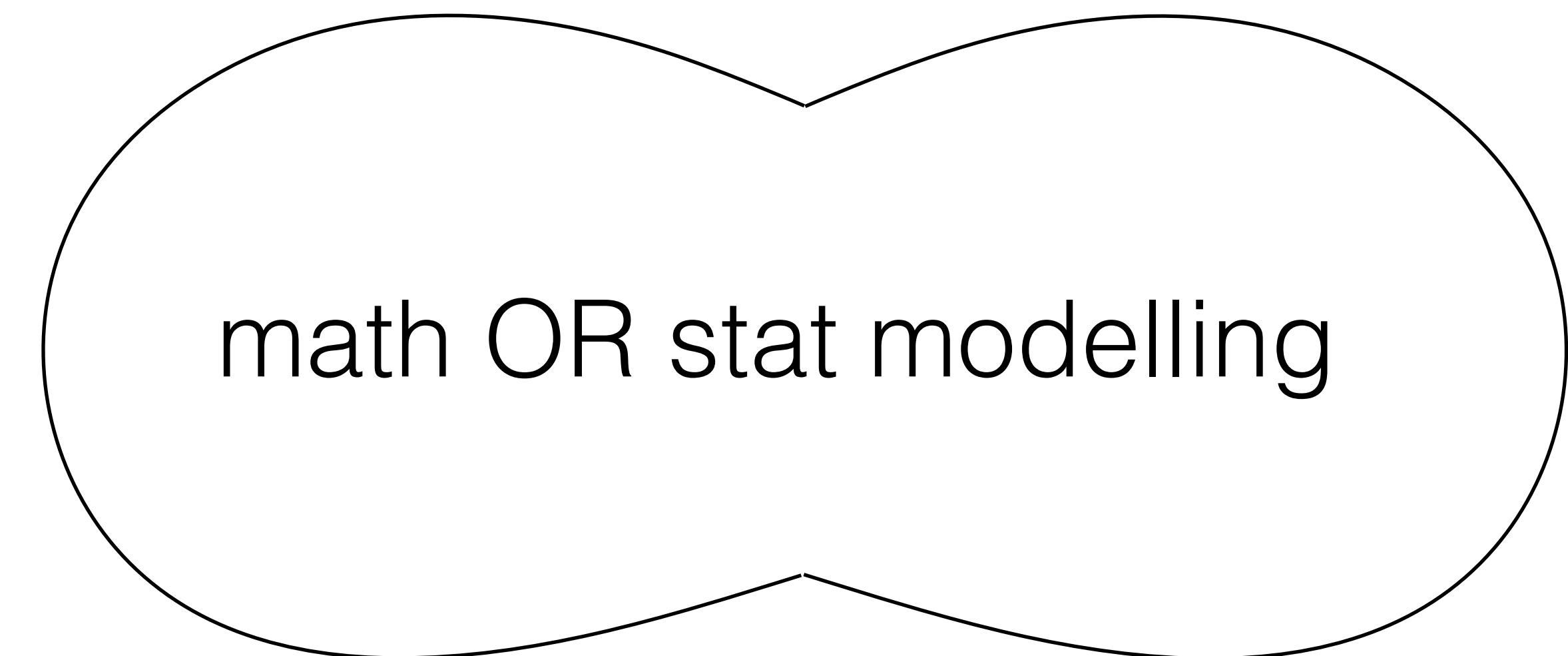
“Mathematical modelling”

“Statistical modelling”

Math and stat modelling: similarities

both involve

- making a model that involves maths,
- *parameterising* the model: coming up with a possible way of expressing the parameters that control the model
- often thinking about what the parameters should be in light of current knowledge
- learning from what the model tells you.



Math modelling: *if this then what*

- *if this*: pick the structure of the model, pick values for the parameters controlling the model.
- *then what*: what does the model imply given those assumptions? Often: what dynamic behaviour results.
- Model may involve probability ('stochastic') or not ('deterministic')

Stat modelling: which *if?*

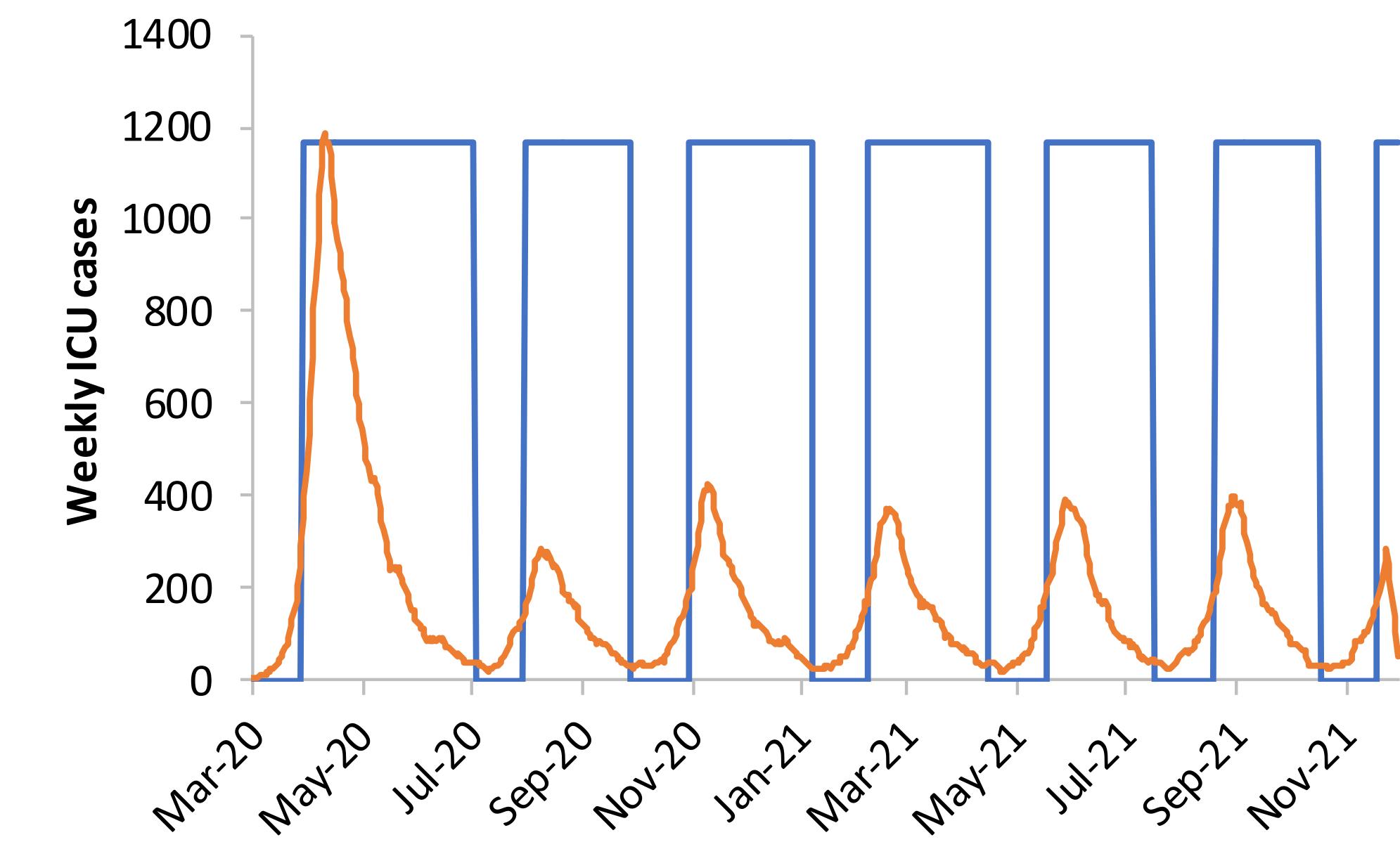
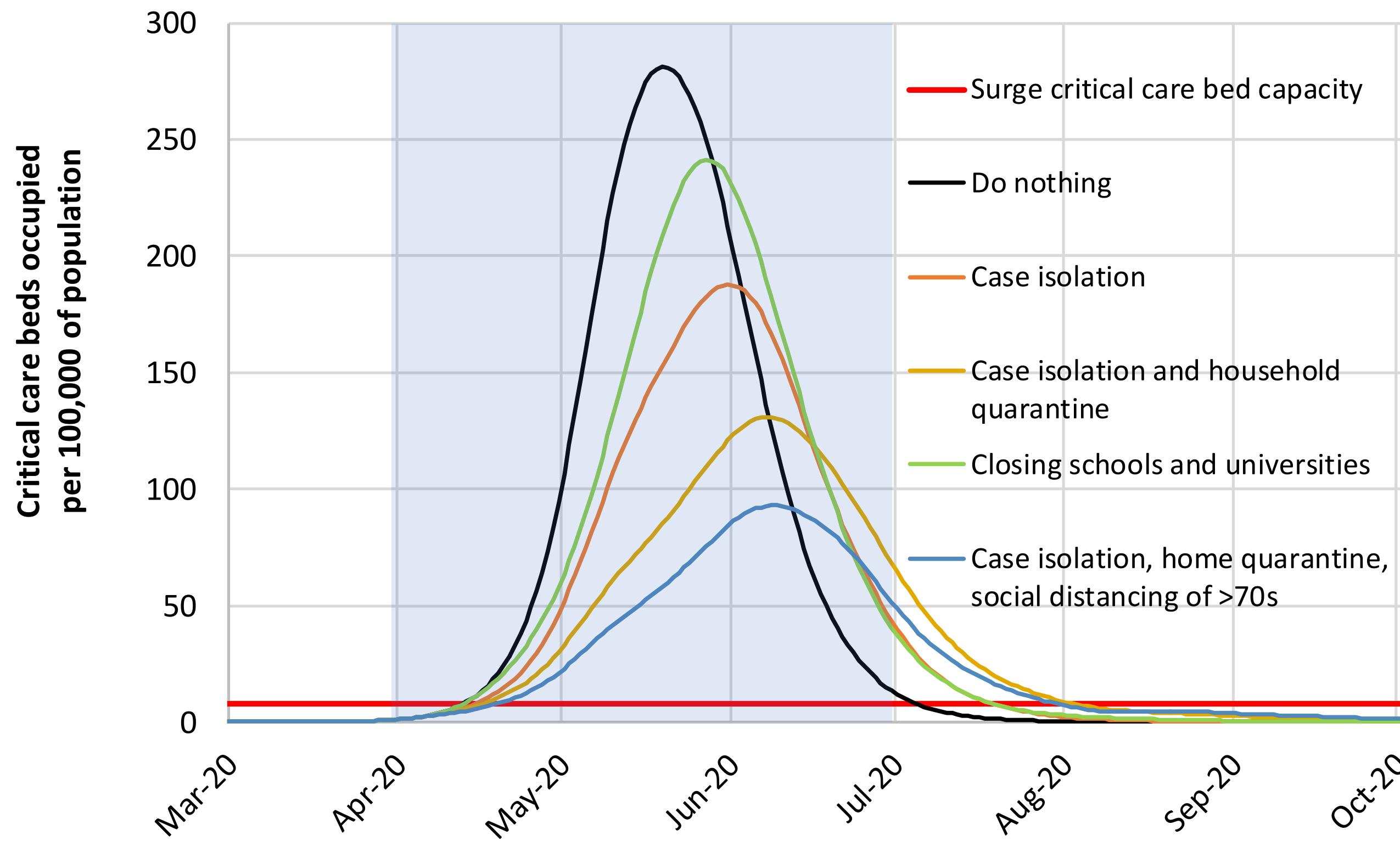
- Model the data-generating process
- Learn about the parameters of that process, using the data it has generated
- Model requires probability

Explore what these equations imply for the relationship between the variables

$$P(y_i | \hat{y}_i, \sigma^2) = N(y_i | \hat{y}_i, \sigma^2)$$
$$\hat{y}_i = mx_i + c$$

Fit to observations of x and y ; learn values of m , c and σ

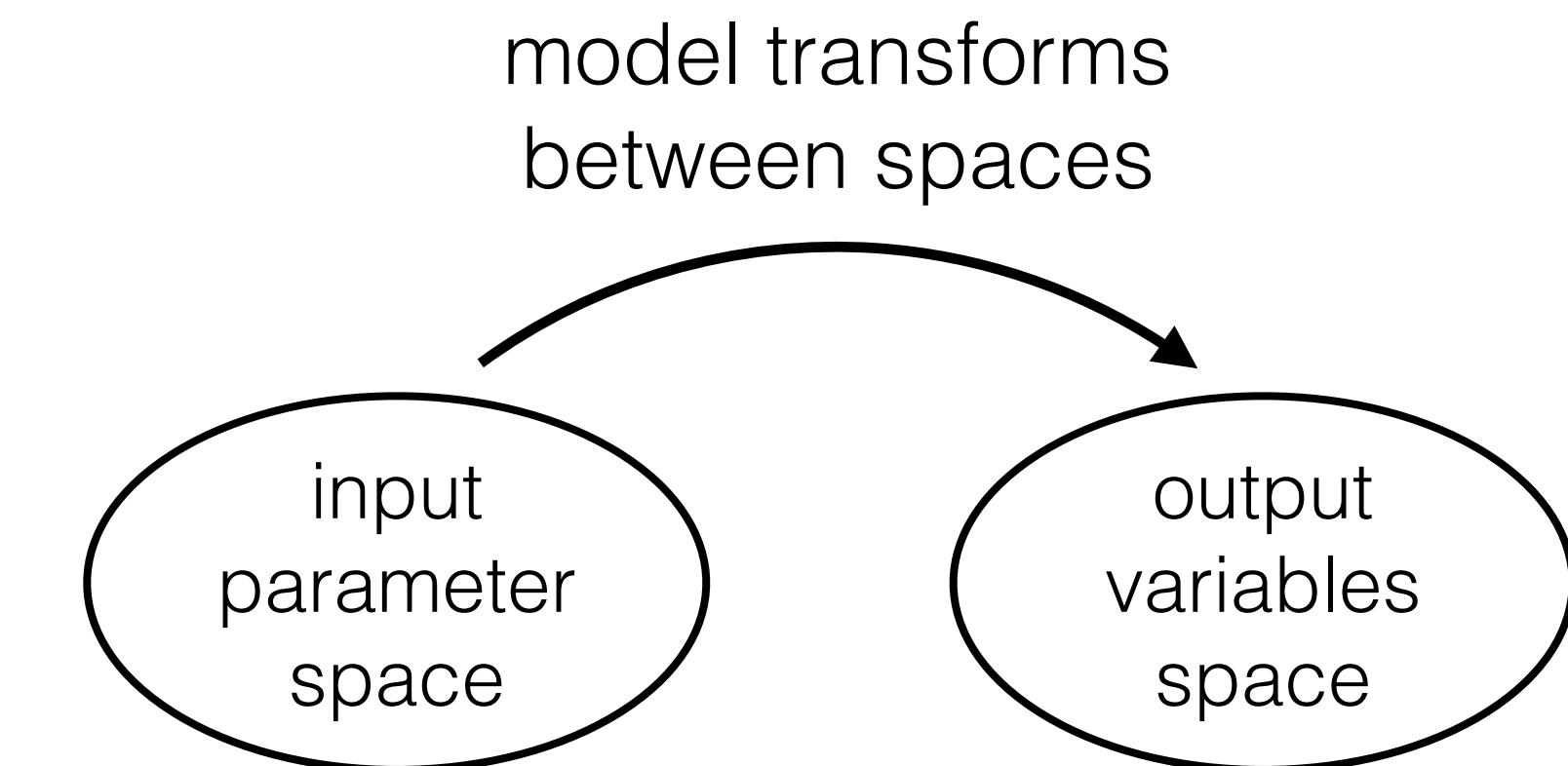
Exploring $y = mx + c\dots$ sounds a bit easy and useless no?



Ferguson et al March 2020, "Imperial Report 9"

Math vs stat modelling: transforming spaces

A mathematical model is a transformation from the input parameter space to some other space of output variables, by modelling the relation between spaces.

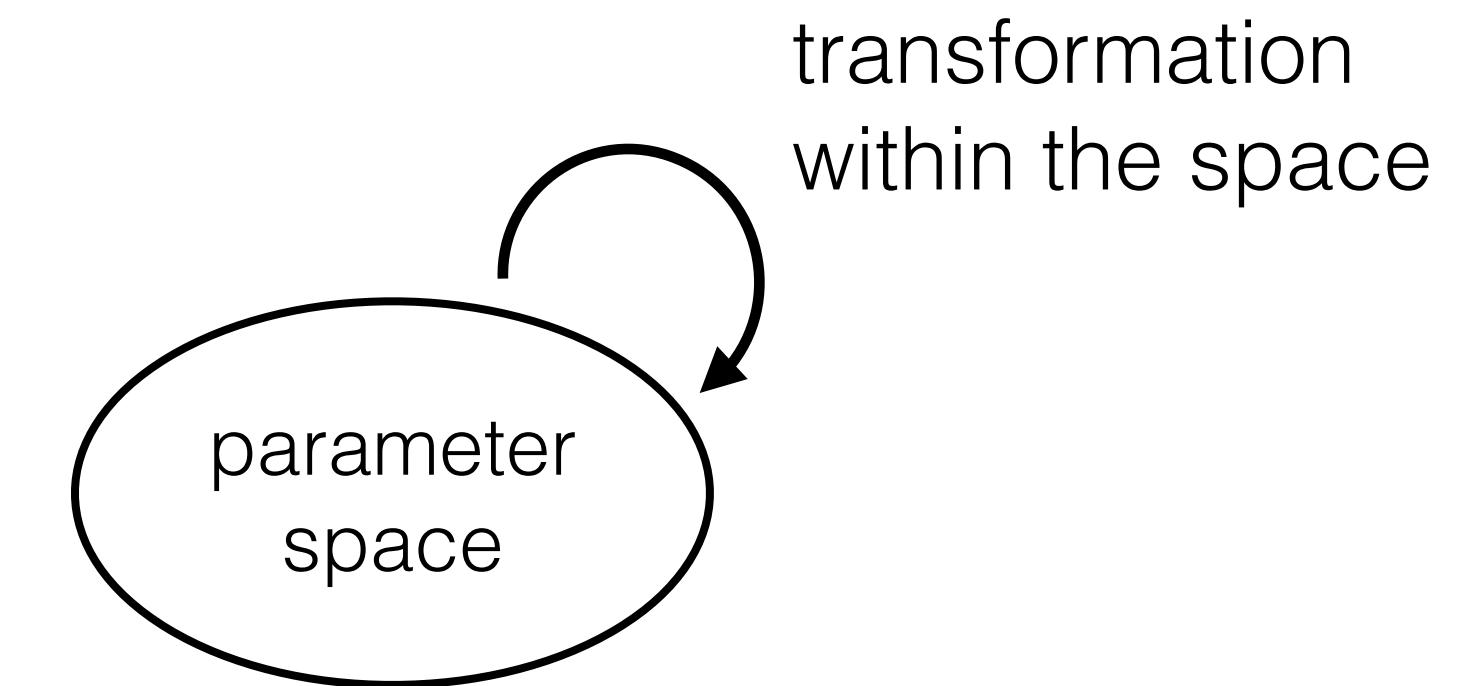


"If this... then... what"

A statistical model is a transformation within the same parameter space, reflecting our learning from the data.

Bayesians: transform the prior distribution to the posterior distribution.

Frequentists: transform the full space to the subset not rejected.



Math vs stat modelling: Bayesian view examples

SIR math model:

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta IS}{N} \\ \frac{dI}{dt} &= \frac{\beta IS}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

transformation
between spaces

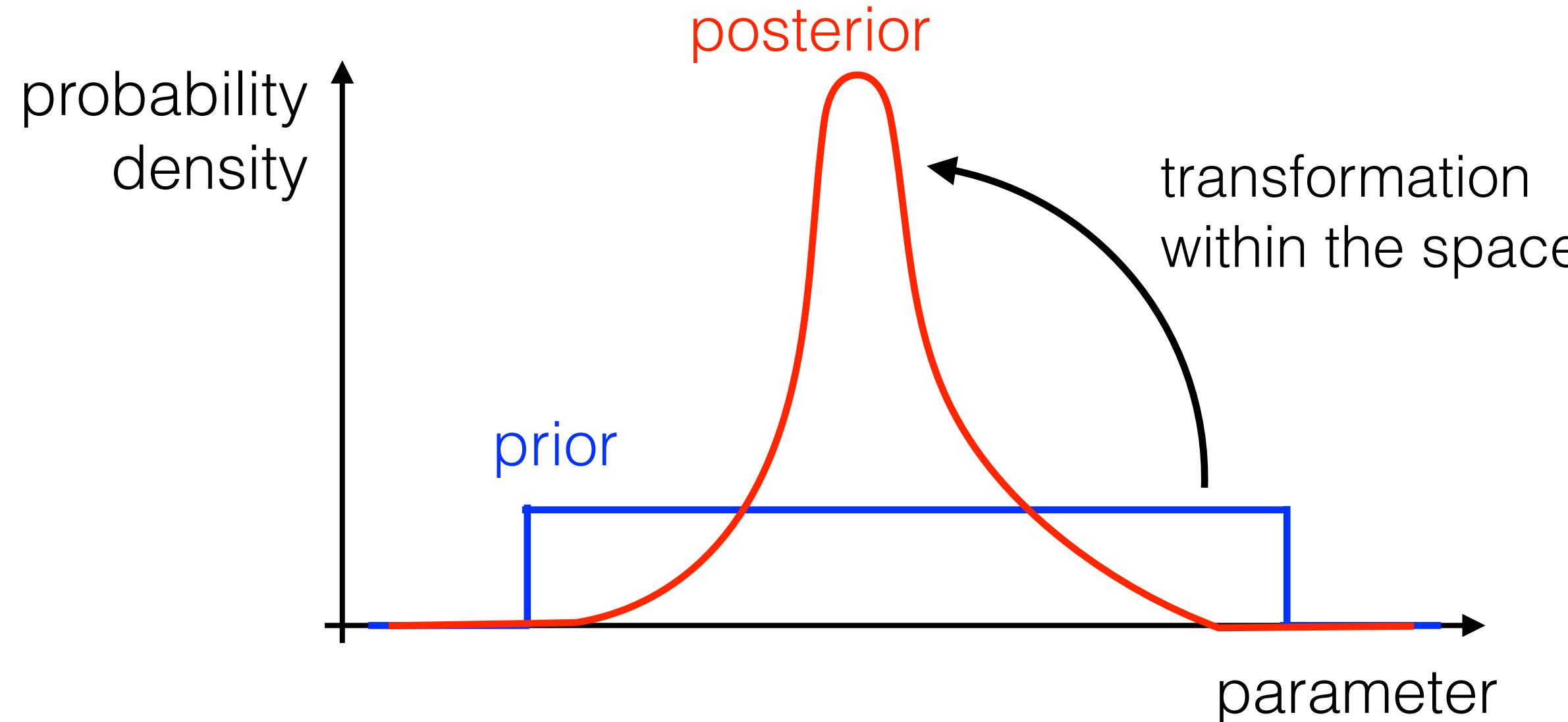
$\beta, \gamma,$
 $S_{t=0}, I_{t=0}, R_{t=0}$

5D space

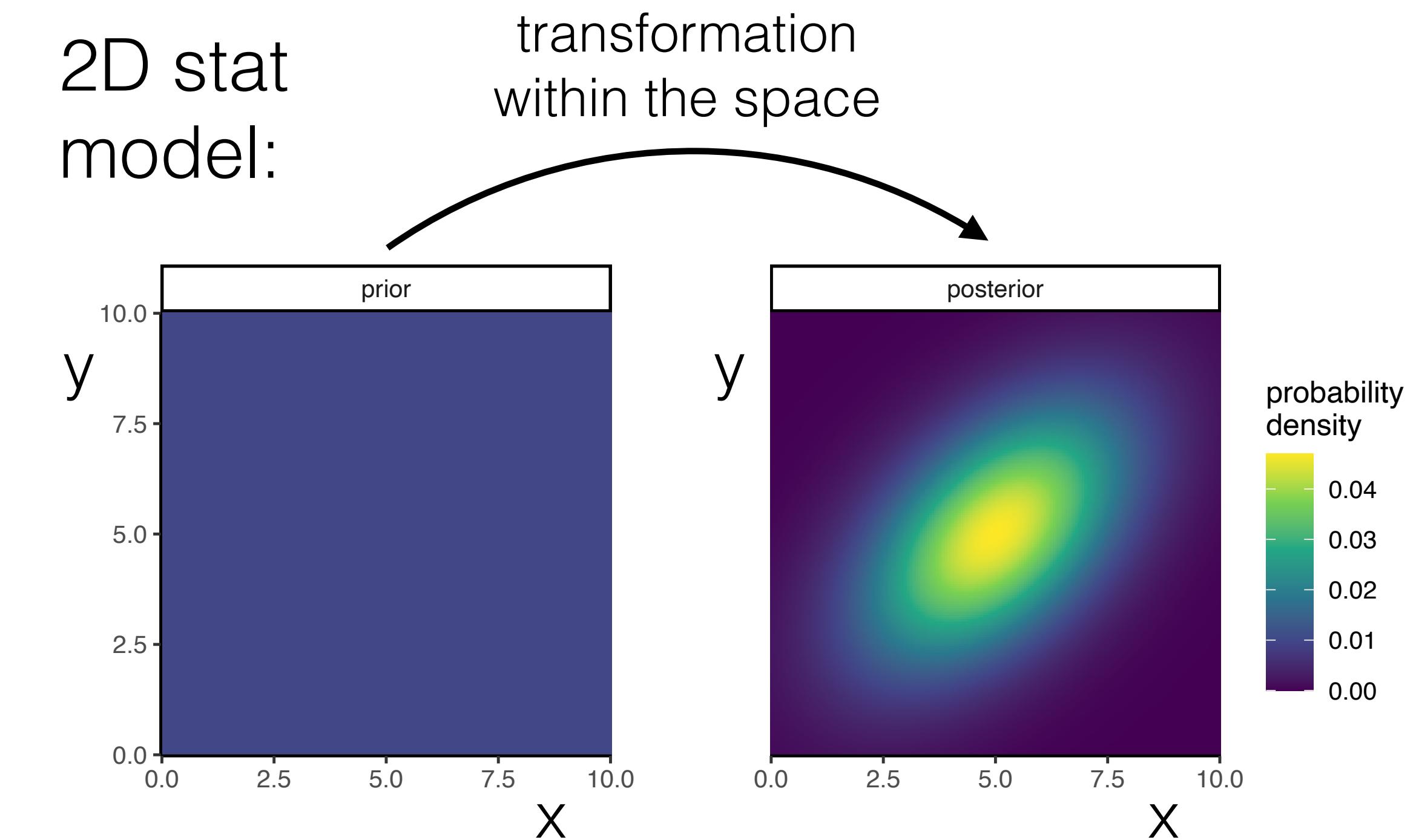
$S(t), I(t), R(t)$

[space of possible functions]³

1D stat model:



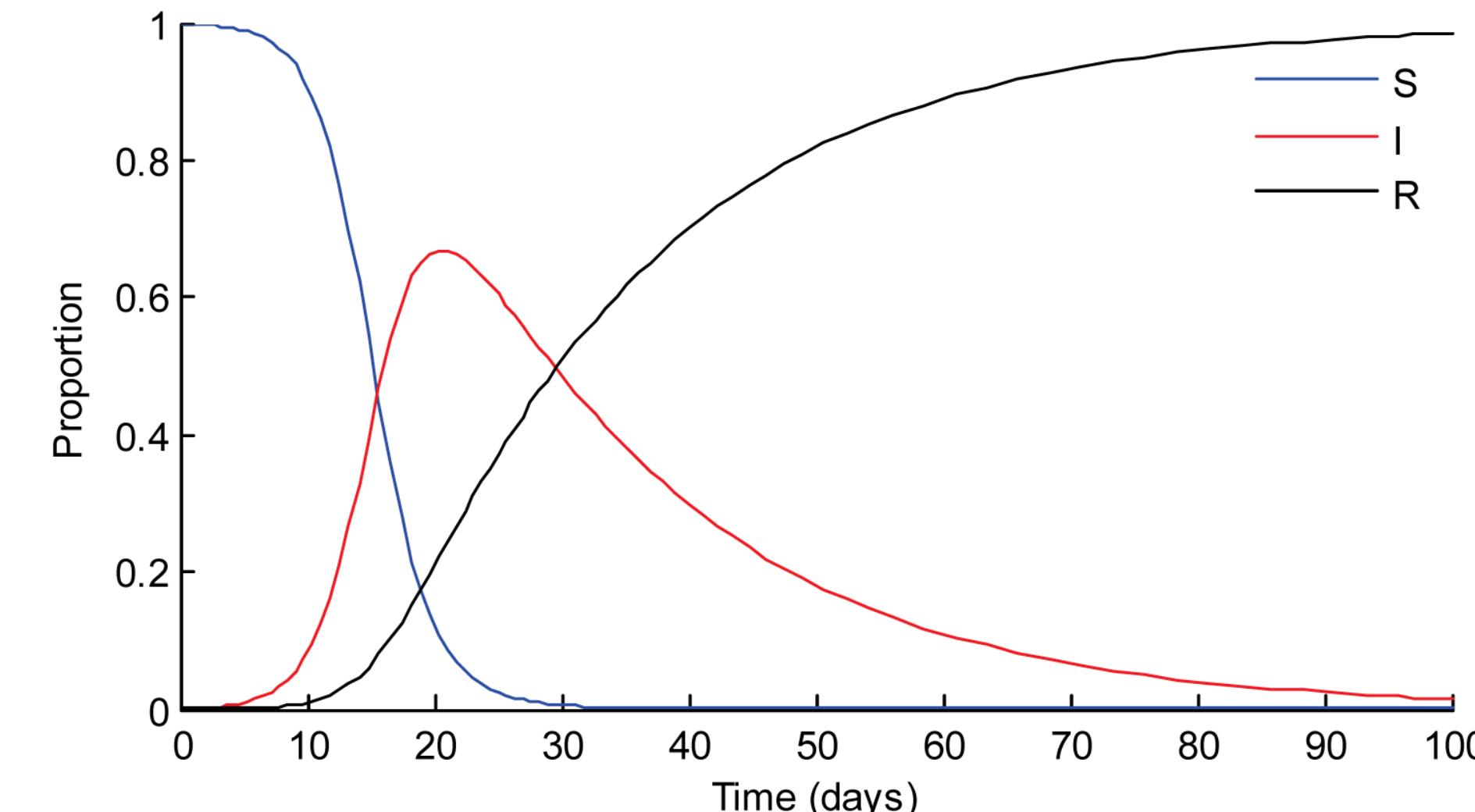
2D stat model:



Deterministic vs stochastic math models

Deterministic SIR:

$$\frac{dS}{dt} = -\frac{\beta IS}{N}$$
$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

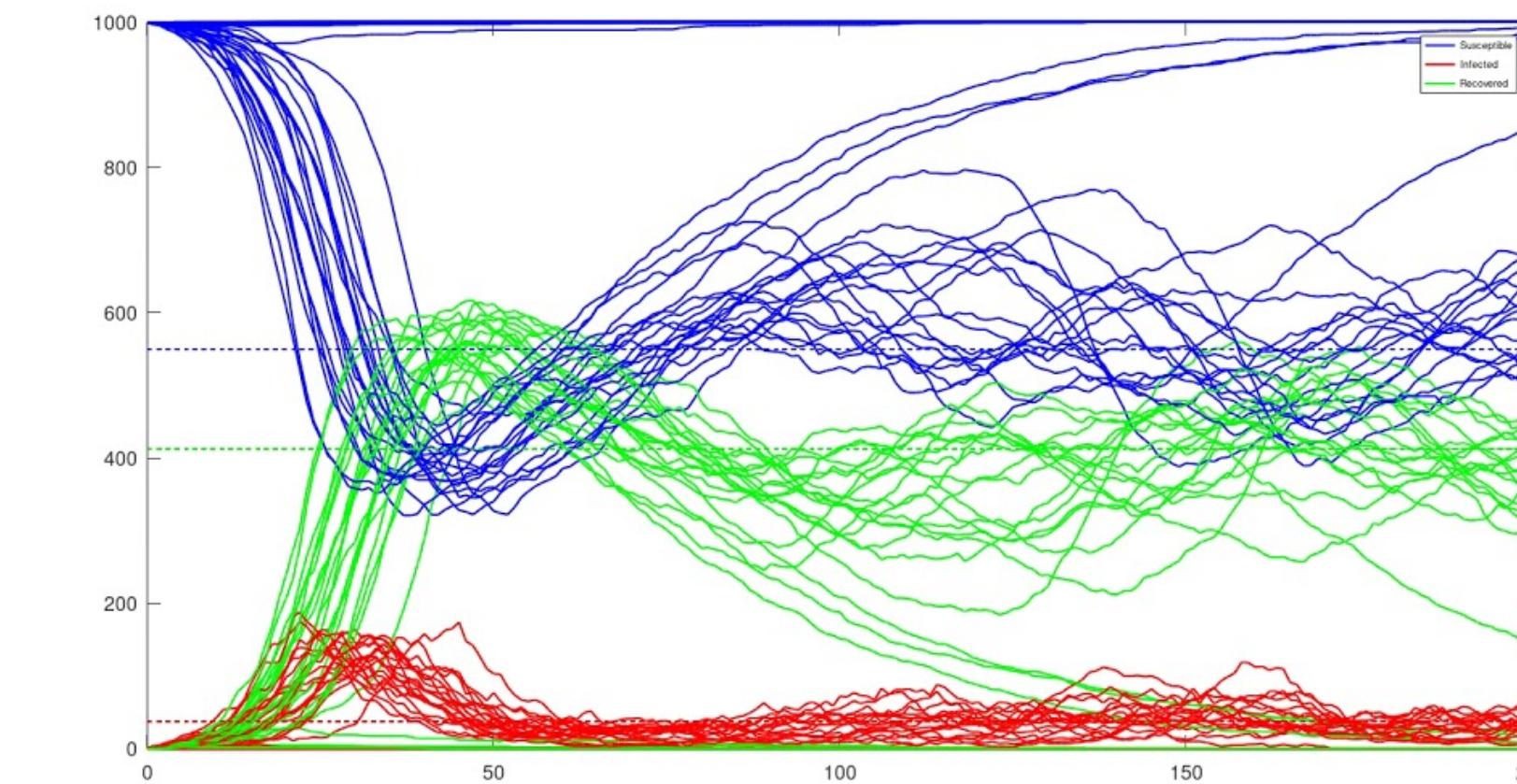


Luz et al, PLoS NTD 2010

Stochastic models:

Instead of specifying derivatives, specify a probability distribution for the finite change in each variable at each finite time step.

$$P(\delta S) = \dots$$
$$P(\delta I) = \dots$$
$$P(\delta R) = \dots$$



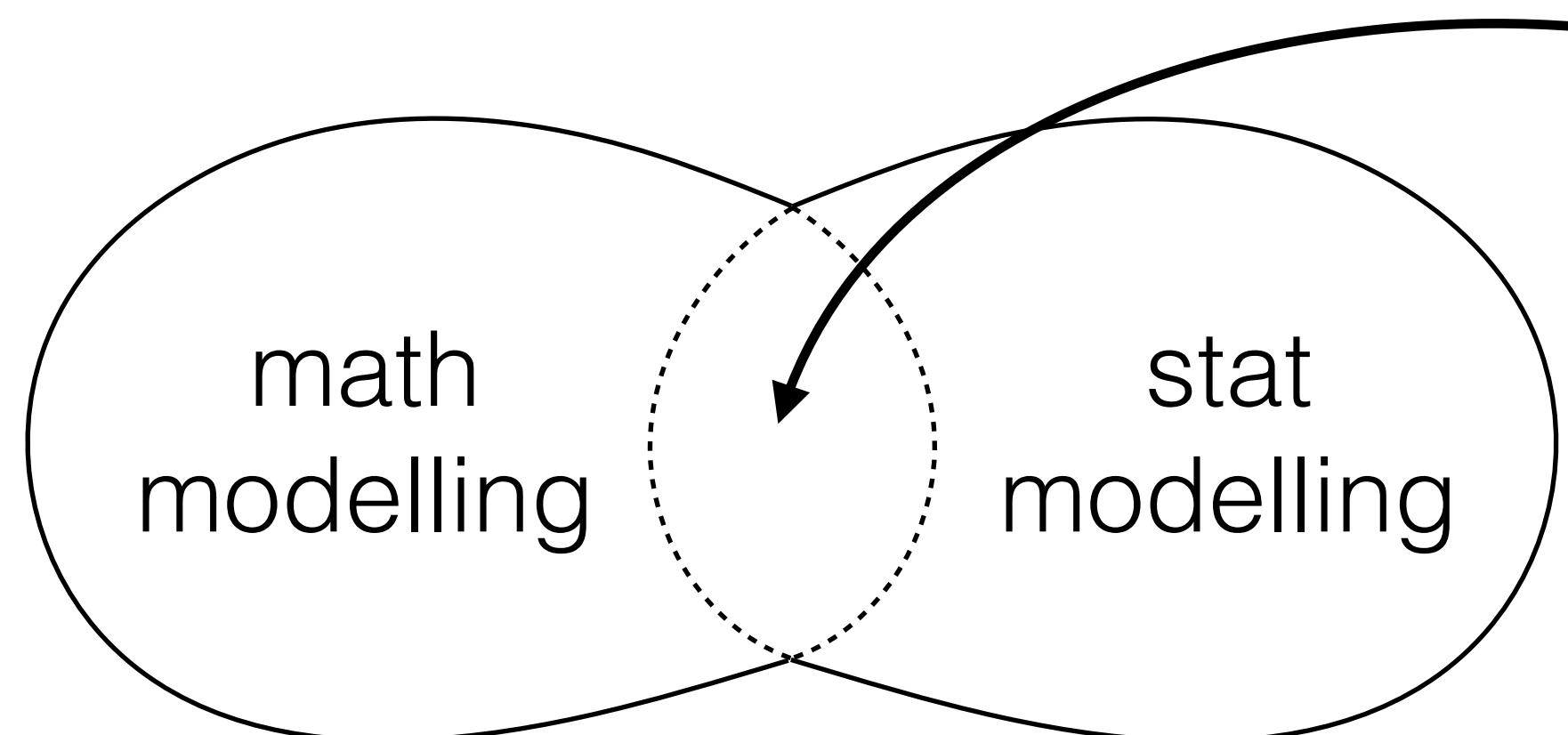
Mark Panaggio [video](#)

Math model mappings & uncertainty

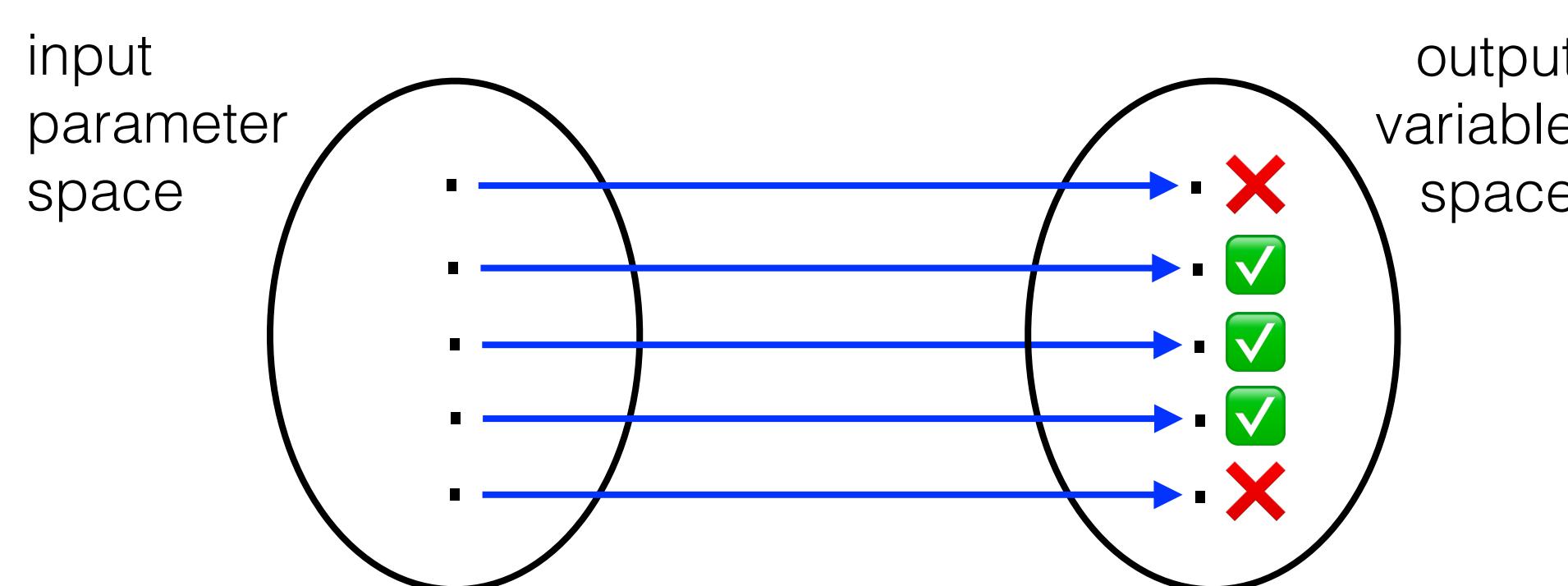
	Deterministic model	Stochastic model
Input = 1 parameter point	<p>input parameter space</p> <p>A diagram showing two circles representing sets. A horizontal blue arrow points from the left circle to the right circle. There is one point in each circle.</p> <p>Output = 1 point, capturing no uncertainty</p>	<p>output variable space</p> <p>A diagram showing two circles representing sets. A horizontal blue arrow points from the left circle to the right circle. From the right side of the arrow, three blue arrows branch out to different points within the right circle.</p> <p>Output = distribution, capturing inherent/stochastic/ontological uncertainty in the process</p>
Input = distribution over parameters	<p>A diagram showing two circles representing sets. Three horizontal blue arrows originate from different points in the left circle and point to different points in the right circle.</p> <p>Output = distribution, capturing epistemological uncertainty in the parameters</p>	<p>A diagram showing two circles representing sets. Three horizontal blue arrows originate from different points in the left circle and point to different points in the right circle. From each of these three points in the right circle, three blue arrows branch out to different points within the right circle.</p> <p>Output = distribution, capturing both inherent/stochastic/ontological uncertainty in the process and epistemological uncertainty in the parameters</p>

Best not to say ‘confidence intervals’ for this uncertainty. Instead e.g. “the central 95% of uncertainty capturing...”

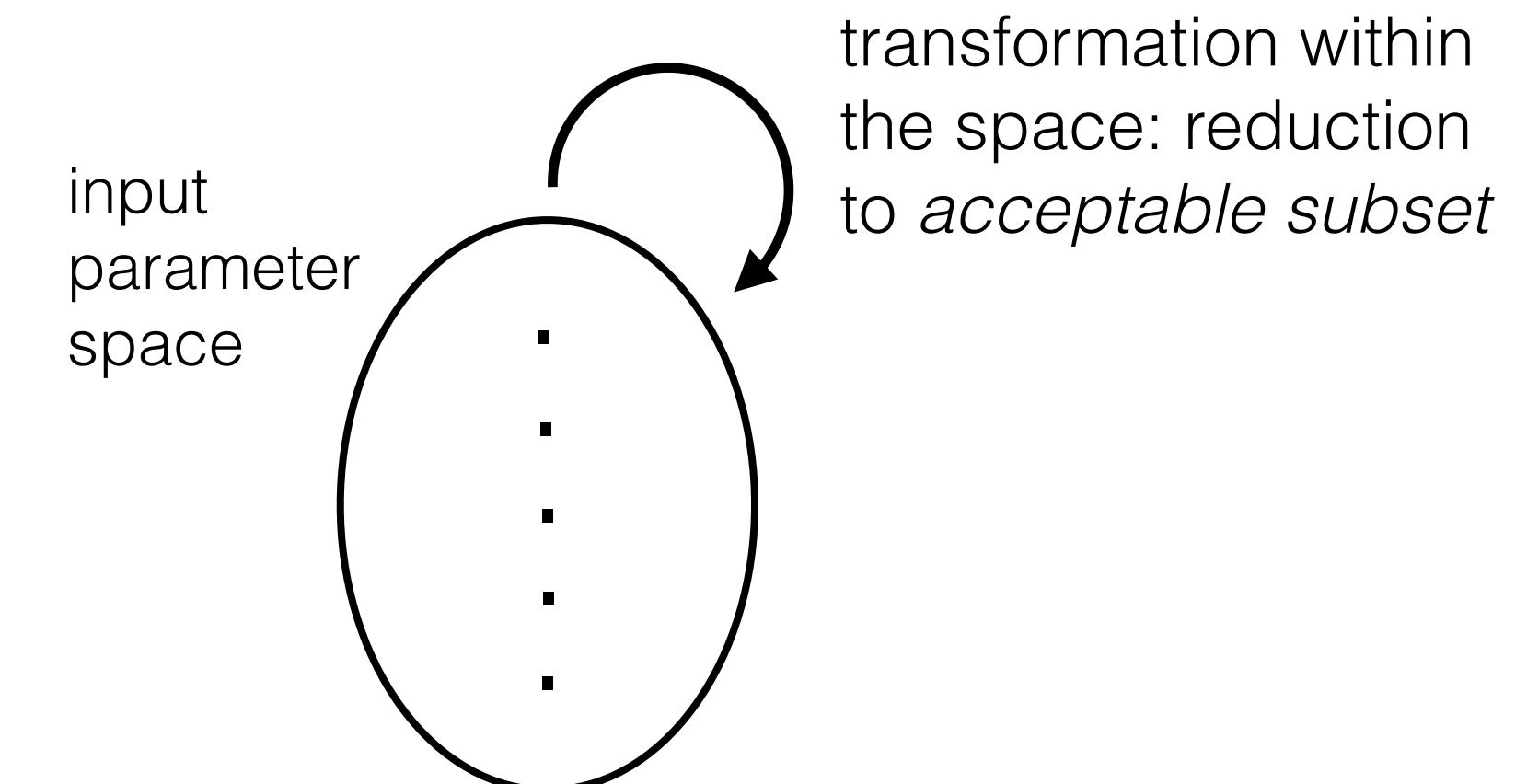
The grey area



Calibration of a math model is the process of choosing one or more points in the input parameter space, based on whether their transformation to the output variable space matches data.



is equivalent to



Calibration is usually done as a preliminary step to counterfactual modelling: first asking “which if” for some parameters, then asking “if this then what” for other parameters. e.g. first “what parameters describe pathogen transmission so far”, and then “if we intervene, how much will we reduce transmission”.

Part 3: Revisiting counterfactuals: a cross-cutting concept in saying anything

Three tasks in data science:

1. Description (e.g. summaries, clustering)
2. Prediction (inputs → outputs)
3. Counterfactual prediction / causal inference
(what if)

Hernán, Hsu & Healy 2019

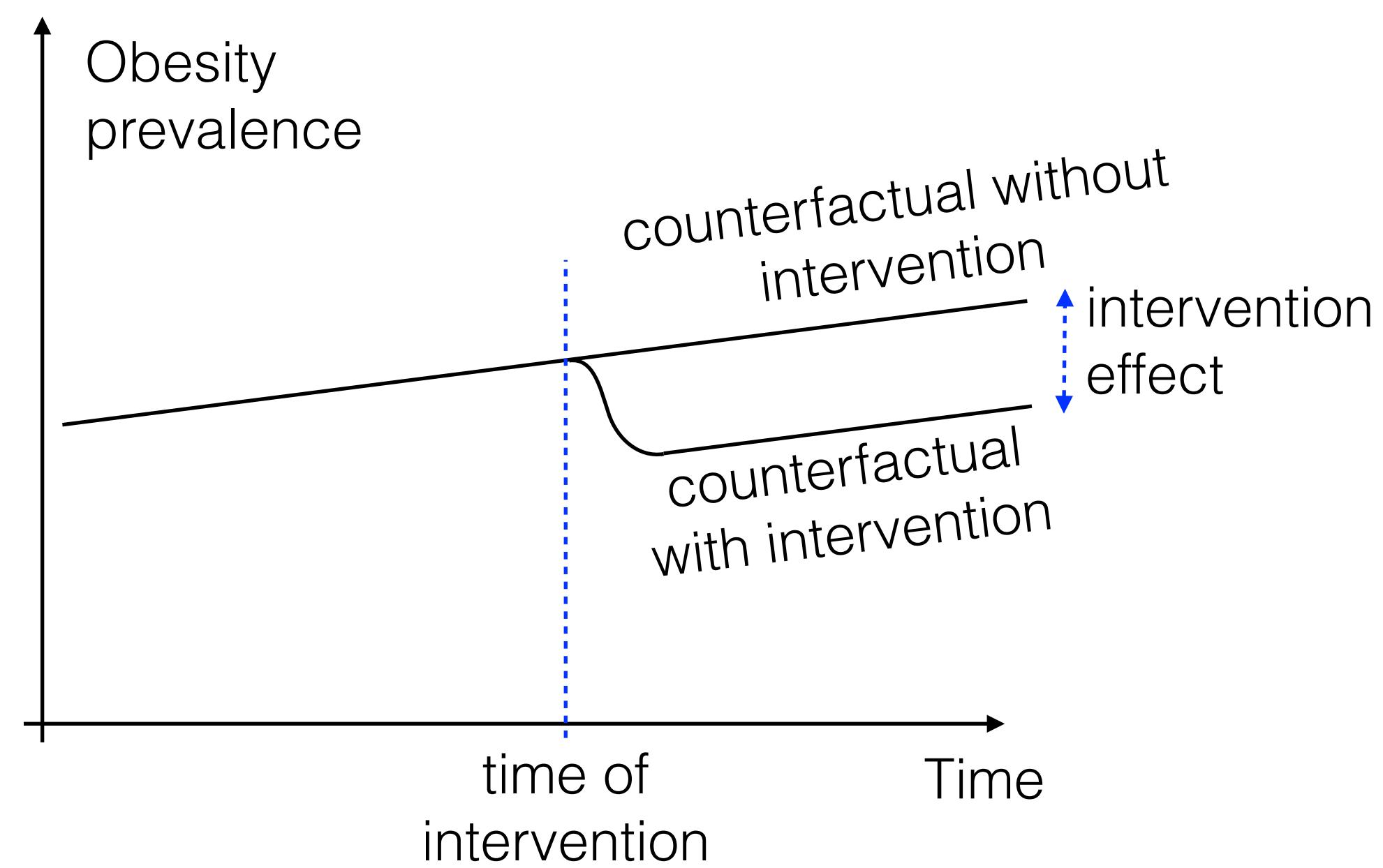
Think in terms of counterfactuals

Counterfactuals / scenarios / potential outcomes: things that could have happened (past) or could happen (future).

An intervention: an action that causes two counterfactuals to begin diverging from that point in time.

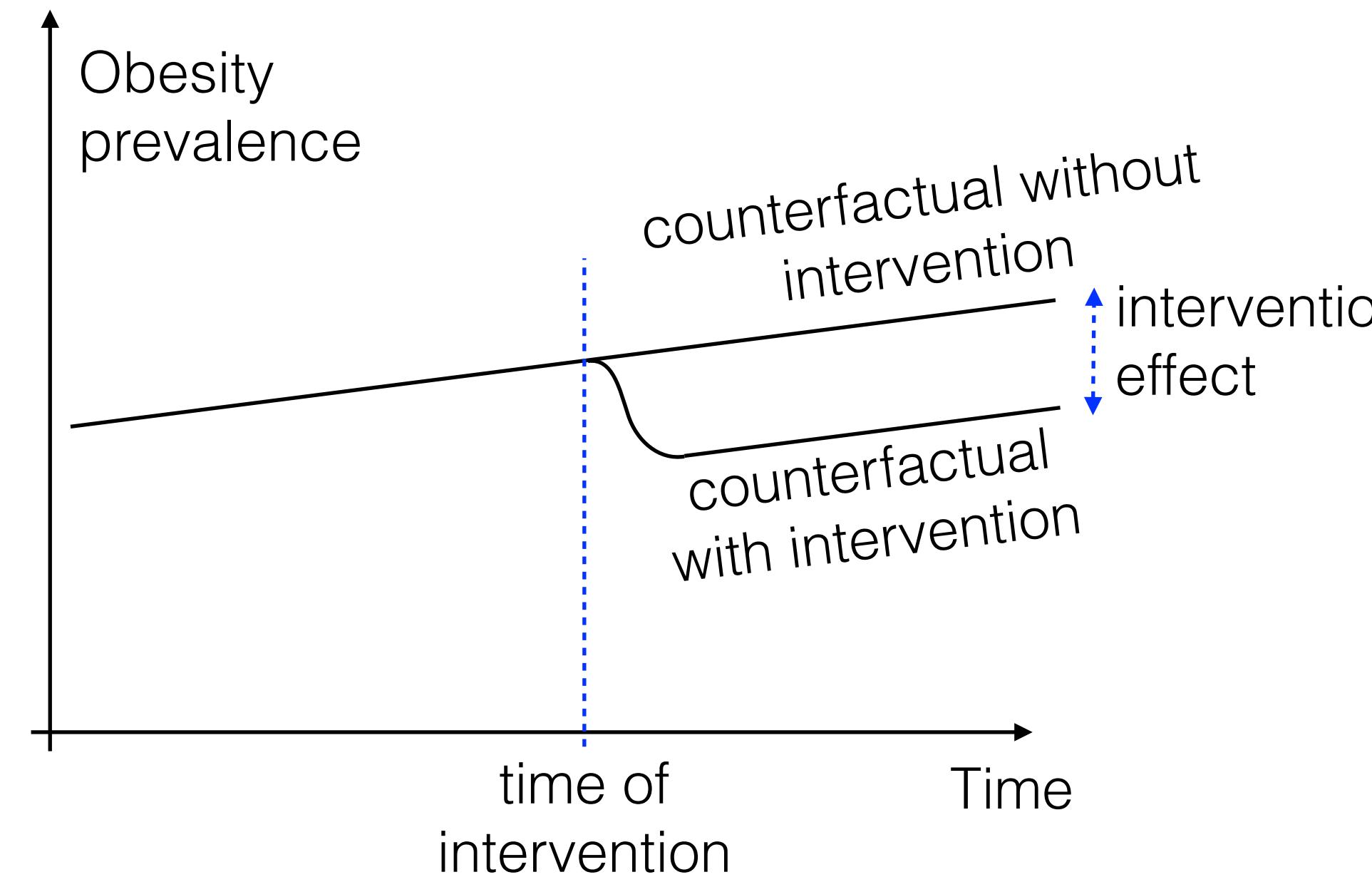
A comparison of *precisely* defined counterfactuals is necessary to:

- define causality and related words like “affects”, “because of” (i.e. attribution). e.g. “obesity causes poor health” is an ill-defined statement because the intervention is not specified. Reducing obesity by chopping off arms would not improve health.
- define the meaning of some common but loaded words like *should*: “we should do X” rests upon defining some measure of value and finding this is higher when doing X than when doing some precisely defined alternative to X.

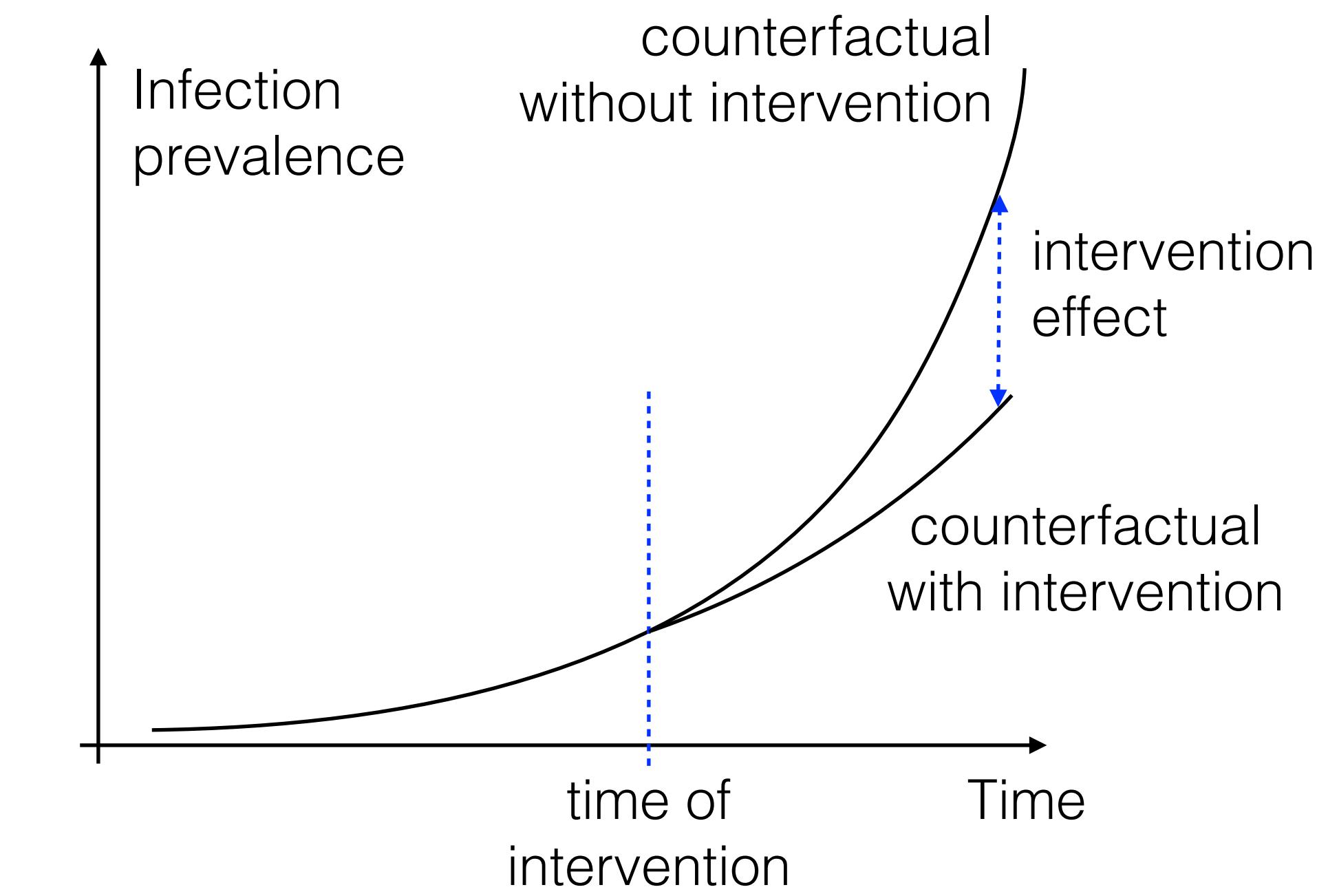


Interacting systems

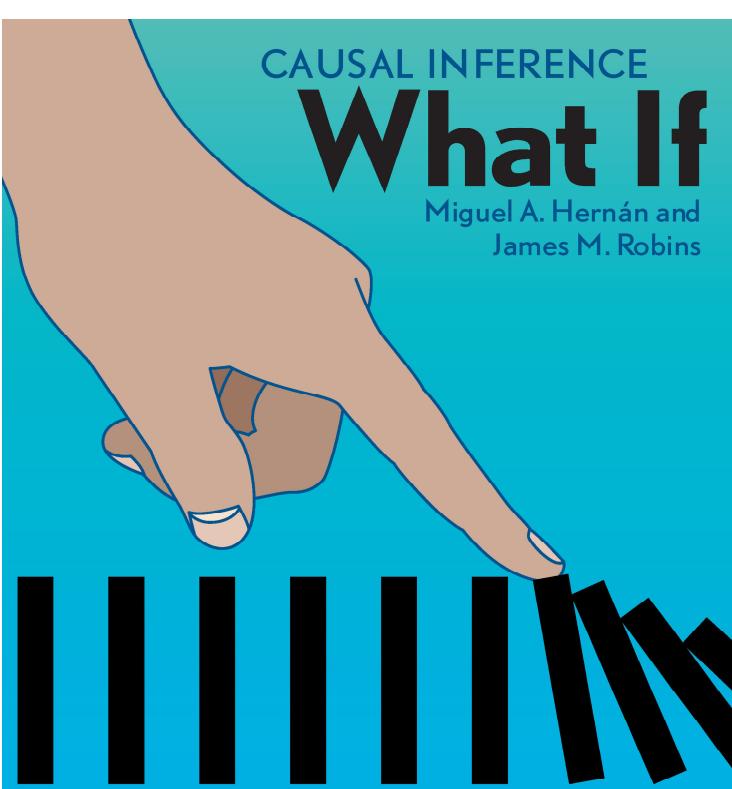
Non-communicable diseases:



Communicable (infectious) diseases:



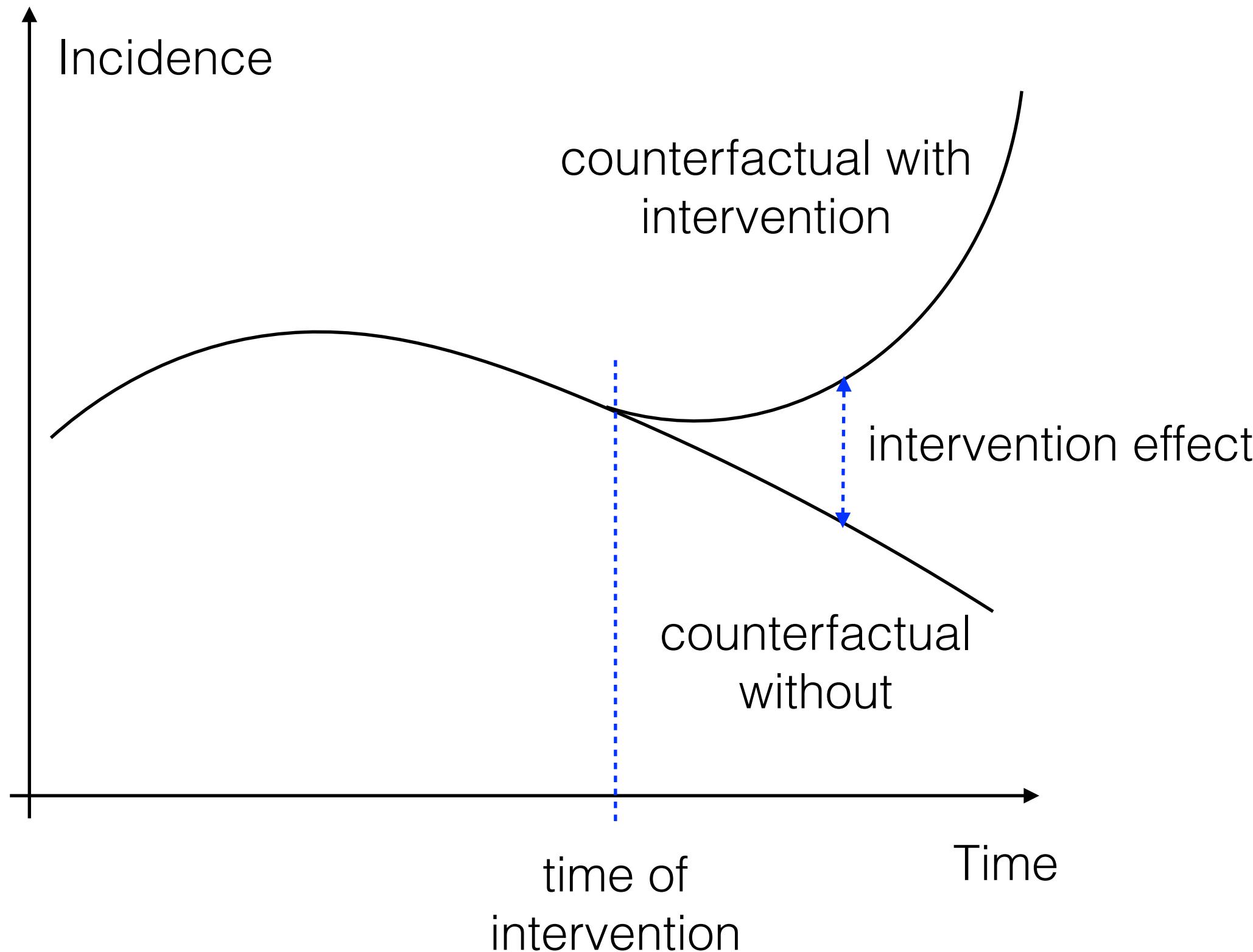
Interventions in interacting systems
can put you on a diverging path,
taking you far away from the
← unseen counterfactual



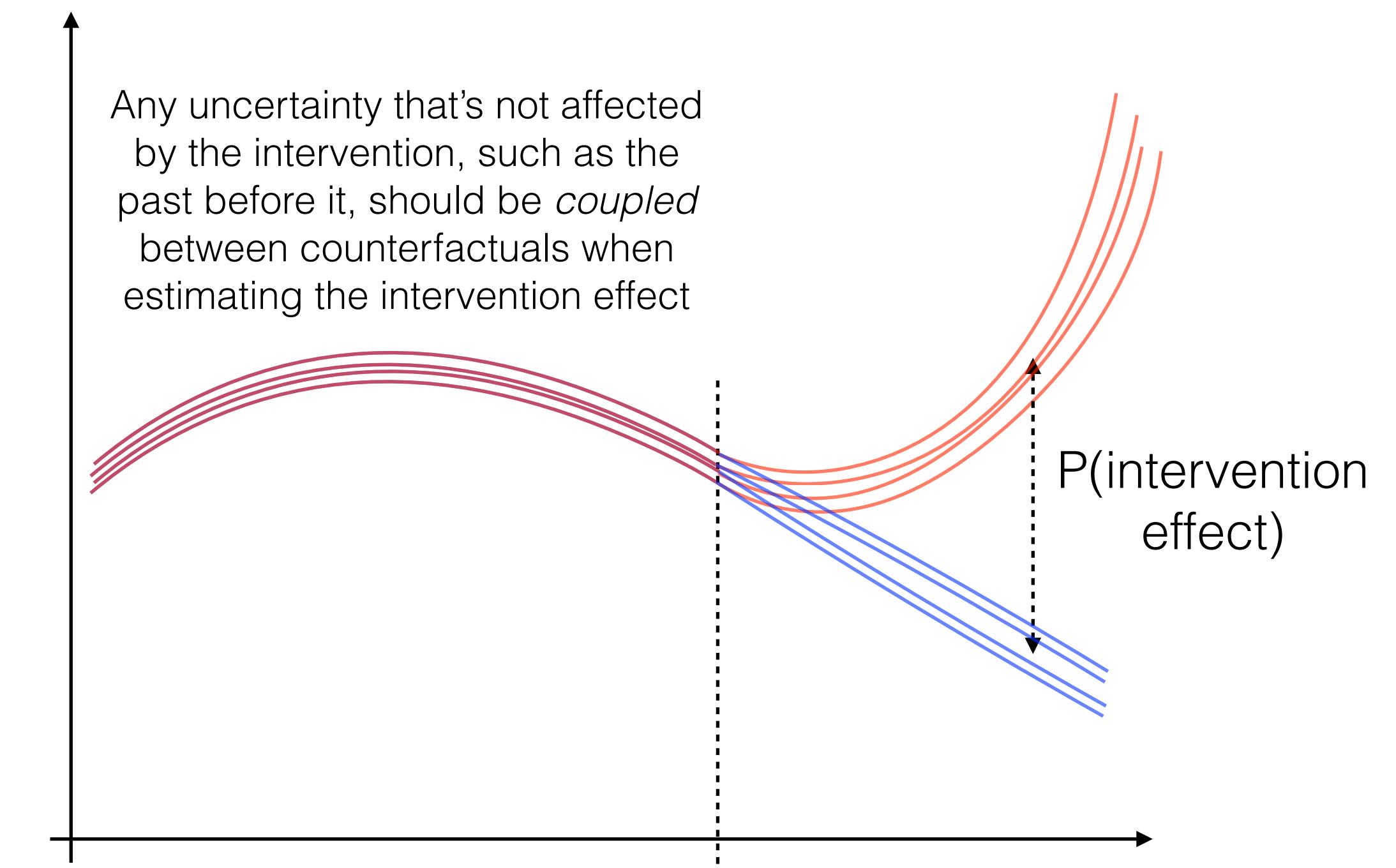
- “The intervention failed to control spread” ✓ but 😢
- “Things got worse after the intervention” ✓ but 😢
- “Things got worse because of the intervention” XXX

Rule of life: aim to
say things that are
both true and helpful

Certain dynamics (precisely known and deterministic)



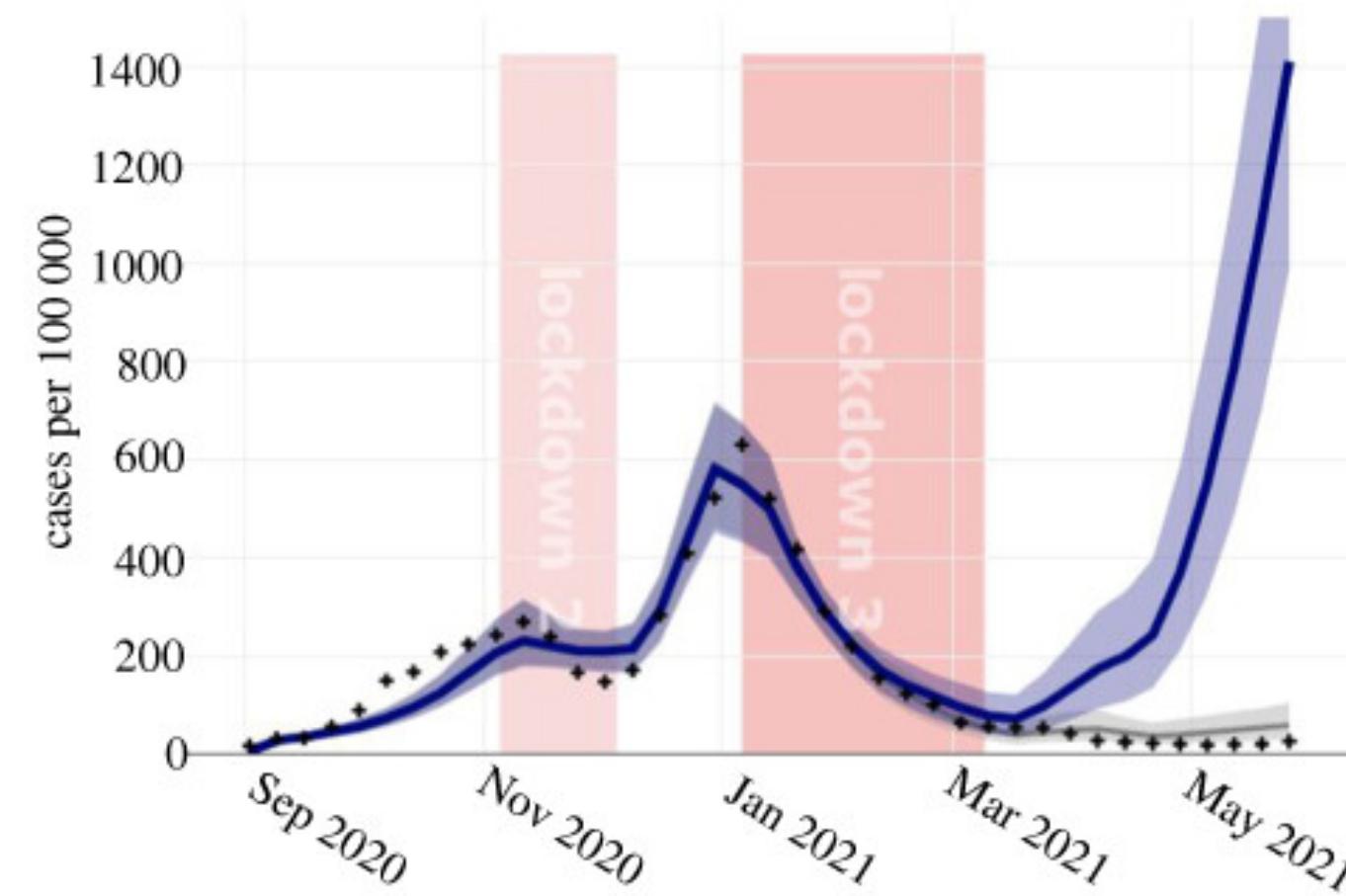
Uncertain dynamics (imprecisely known and/or stochastic)



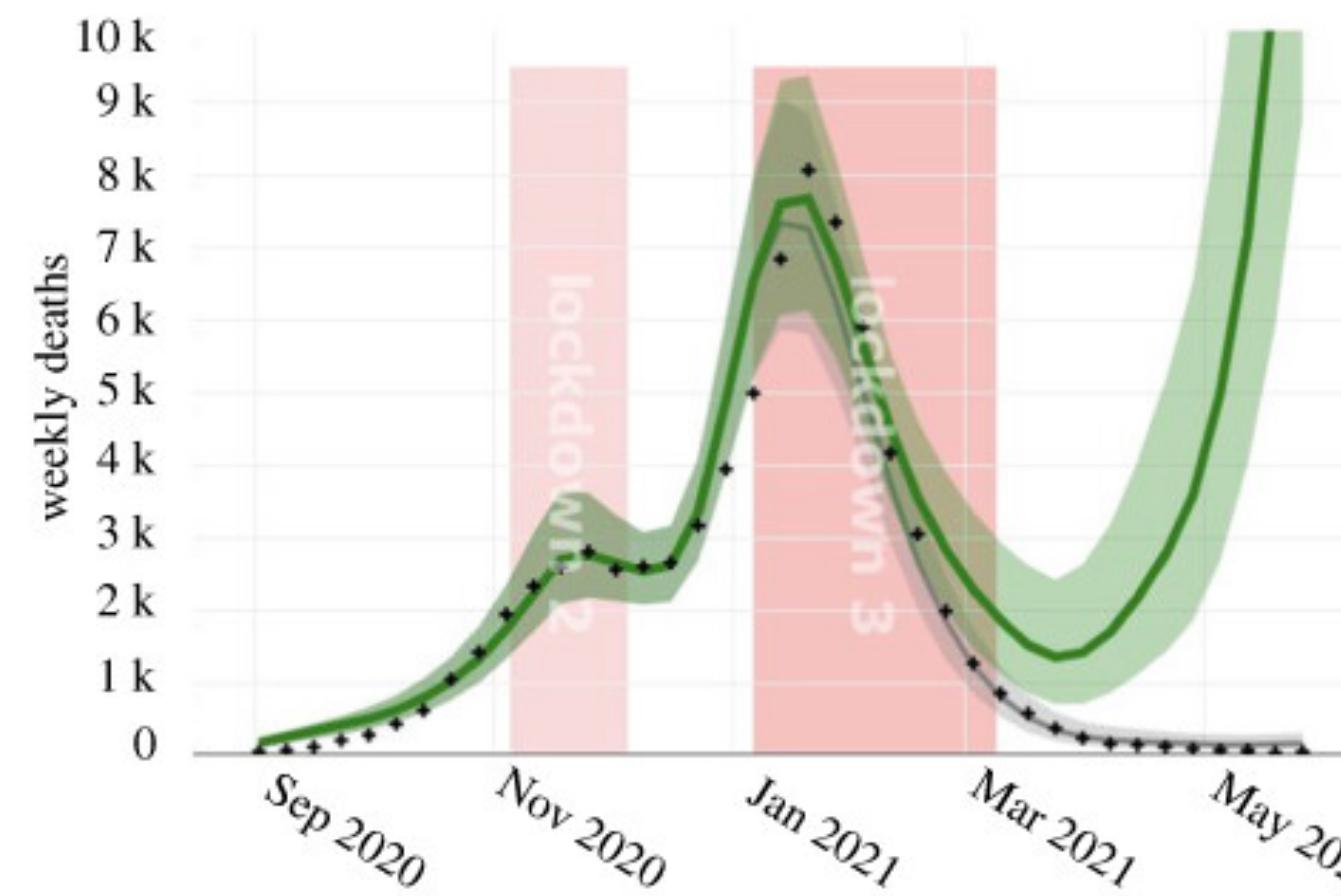
$$\text{Cumulative intervention effect on } X = \int (\text{effect of intervention on } dX/dt) dt$$

$$\text{e.g. cumulative cases averted} = \int (\text{incidence in factual} - \text{incidence in counterfactual}) dt$$

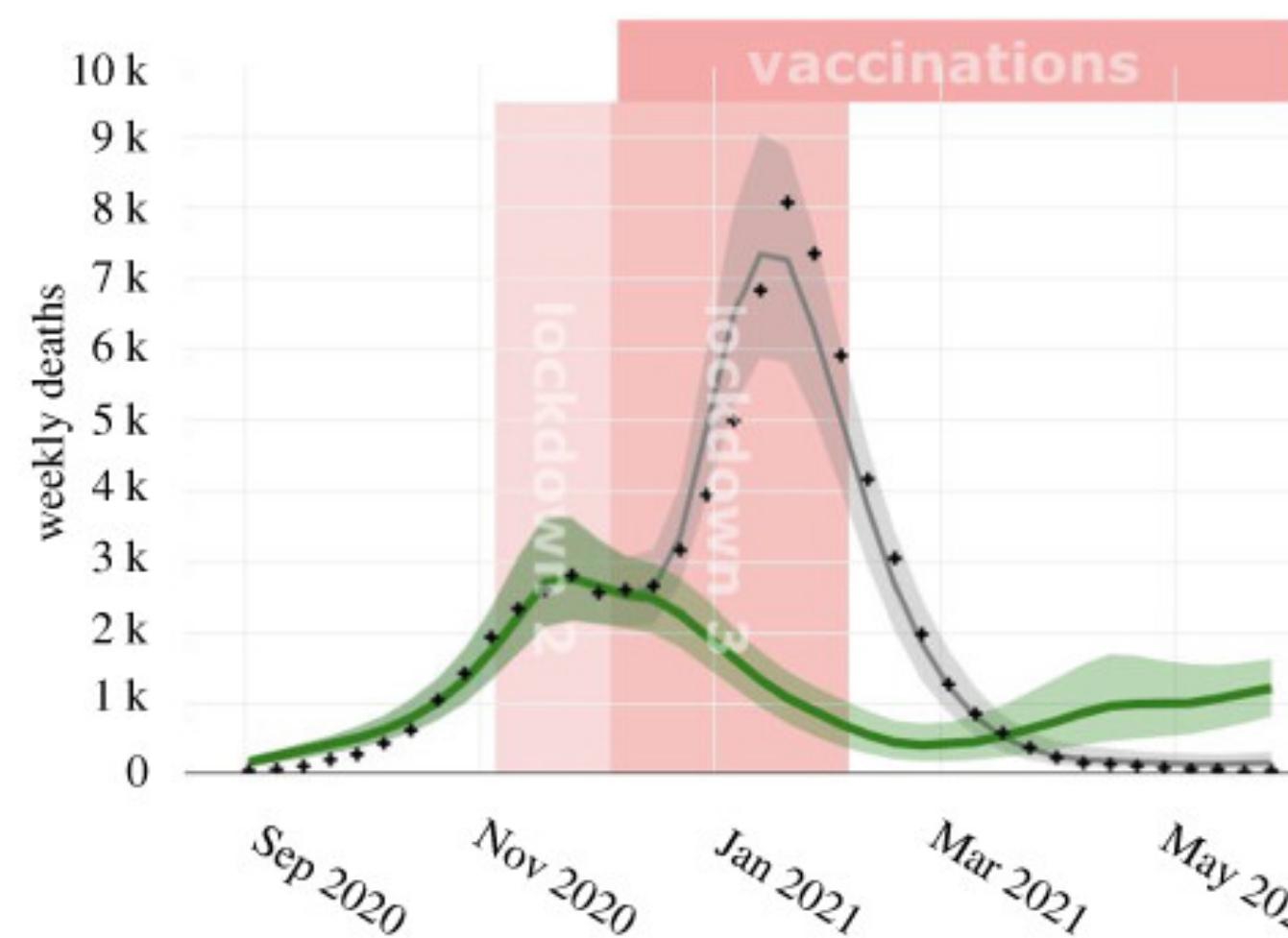
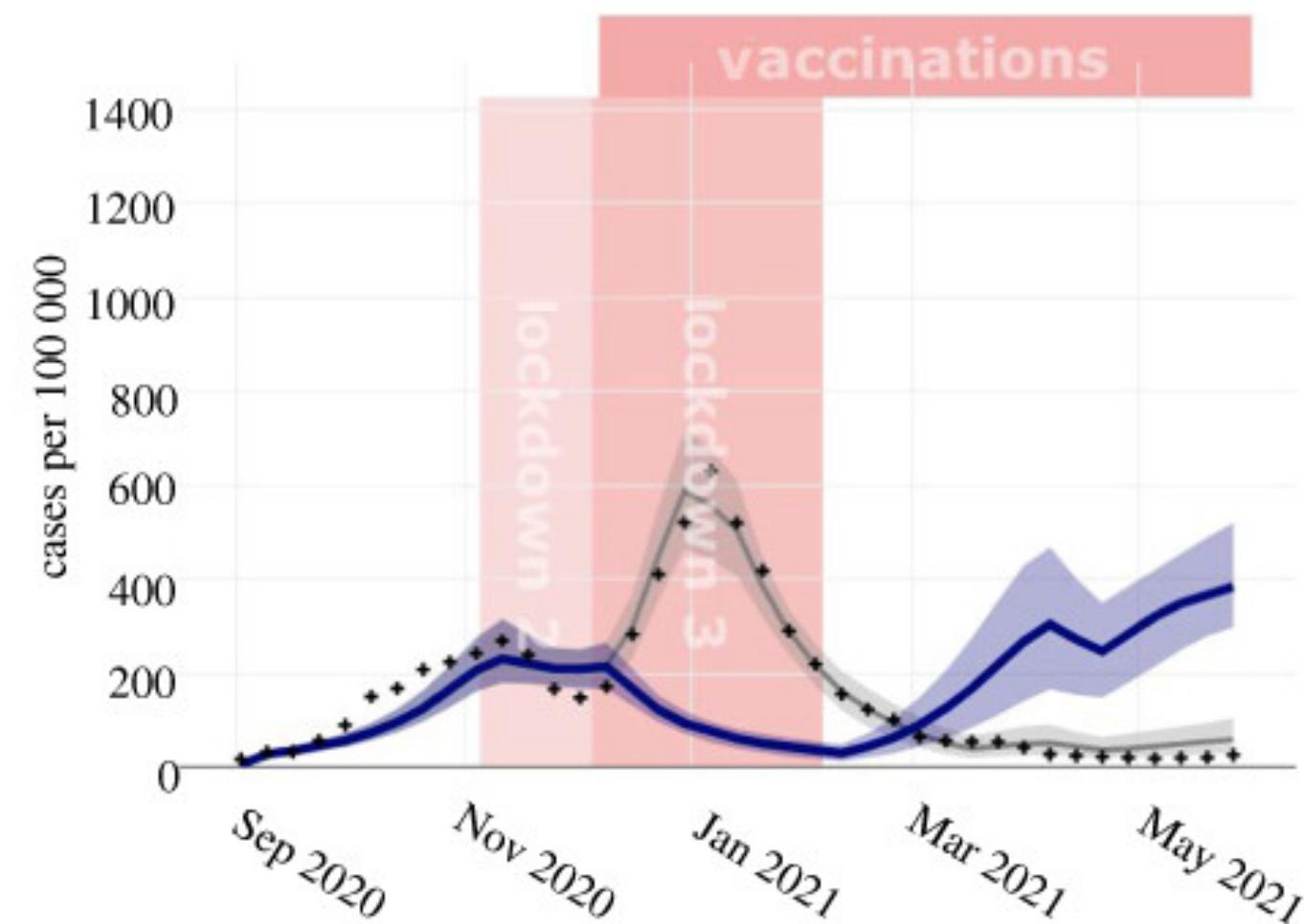
cases



deaths



Factual vs counterfactual comparison 1:
what if we'd had no vaccines?



Factual vs counterfactual comparison 2:
what if we'd done lockdown 3 earlier?

“total deaths would have been
approximately 30 k (24 k–38 k) lower
between January and May 2021”

Points = data

Grey = model fit to data (the factual)

Blue = counterfactual model, cases

Green = counterfactual model, deaths

Part 4: random effects case study: HIV immune system decline

New HIV variant discovered in the Netherlands



Discovery of HIV variant shows virus can evolve to be more severe — and contagious



A 'highly virulent' HIV strain is 'no cause for alarm,' scientists say

The newly identified, more infectious strain of HIV likely began circulating in the Netherlands in the 1990s and responds well to treatment, according to researchers.



ALJAZEERA

News ▾ Ukraine war Features Economy Opinion Video

Is there a 'new' HIV variant?

The Human Immunodeficiency Virus (HIV) is one of the fastest mutating viruses ever studied. Now a team of scientists, led by Oxford University, with key contributions from the Dutch HIV Monitoring Foundation, have identified a strain of HIV, being called the "VB" variant, which has been found to be highly [virulent](#).

Forbes

Newly Discovered HIV Variant Can Cause Patients To Develop AIDS Twice As Fast, Researchers Say



PRESS STATEMENT

Identification of fast-spreading HIV variant provides evidence of urgency to halt the pandemic and reach all with testing and treatment

UNAIDS



Wymant et al, Science 2022
beehive.ox.ac.uk/hiv-lineage

EL PAÍS

LA PANDEMIA DE VIH/SIDA >

Descubierta una nueva variante del VIH más contagiosa y virulenta

El hallazgo, en un centenar de personas en Países Bajos, es una constatación de que los virus pueden evolucionar hacia formas más agresivas

BBC Menu

Science In Action

Science in Action Home Episodes Podcast Join us on Facebook Follow us on Twitter

Listen now

Identifying a more infectious HIV variant

A new but treatable variant of HIV has been uncovered in the Netherlands. A reminder that as viruses evolve, even after 40 years they can become more virulent and infectious.

Available now 30 minutes

SPIEGEL Wissenschaft

Zufallsfund bei Studie

Aggressivere Variante von HIV in den Niederlanden entdeckt

Bei einer Langzeitstudie sind Wissenschaftler einer bisher unbekannten Variante des HIV-Virus auf die Spur gekommen. Sie ist wohl leichter übertragbar. Dennoch geben Experten Entwarnung.



NIEDERLANDE

Aggressivere HIV-Variante entdeckt

Im Rahmen einer Langzeitstudie ist ein Forschungsteam in den Niederlanden auf eine bisher unbekannte Variante des HI-Virus gestoßen, die unter anderem leichter übertragbar ist. Ein Grund zur Sorge soll diese aber nicht sein.



Un nouveau variant du VIH plus virulent identifié aux Pays-Bas, mais "pas de raison de s'alarmer"



Hiv-variant biedt les voor corona-aanpak

Random effects recap I

	Frequentist	Bayesian
Fixed effect parameter p_f:	A parameter that controls a stochastic process. We don't talk about the probabilities of such parameters (no no): we can say $P(\dots p_f)$ but never $P(p_f \dots)$. $p_f \xrightarrow{\text{stochastic process}} \text{data}$	A parameters whose prior is independent of all other parameters in our model $\xrightarrow{\text{prior}} p_f \xrightarrow{\text{stochastic process}} \text{data}$
Random effect parameter p_r:	Acts as a parameter for a stochastic process, but is itself the outcome of a stochastic process, and so has an associated probability. $p_f \xrightarrow{\text{stochastic process}} p_r \xrightarrow{\text{stochastic process}} \text{data}$	A parameters whose prior depends on other parameters in our model $\xrightarrow{\text{prior}} p_f \xrightarrow{\text{prior}} p_r \xrightarrow{\text{stochastic process}} \text{data}$

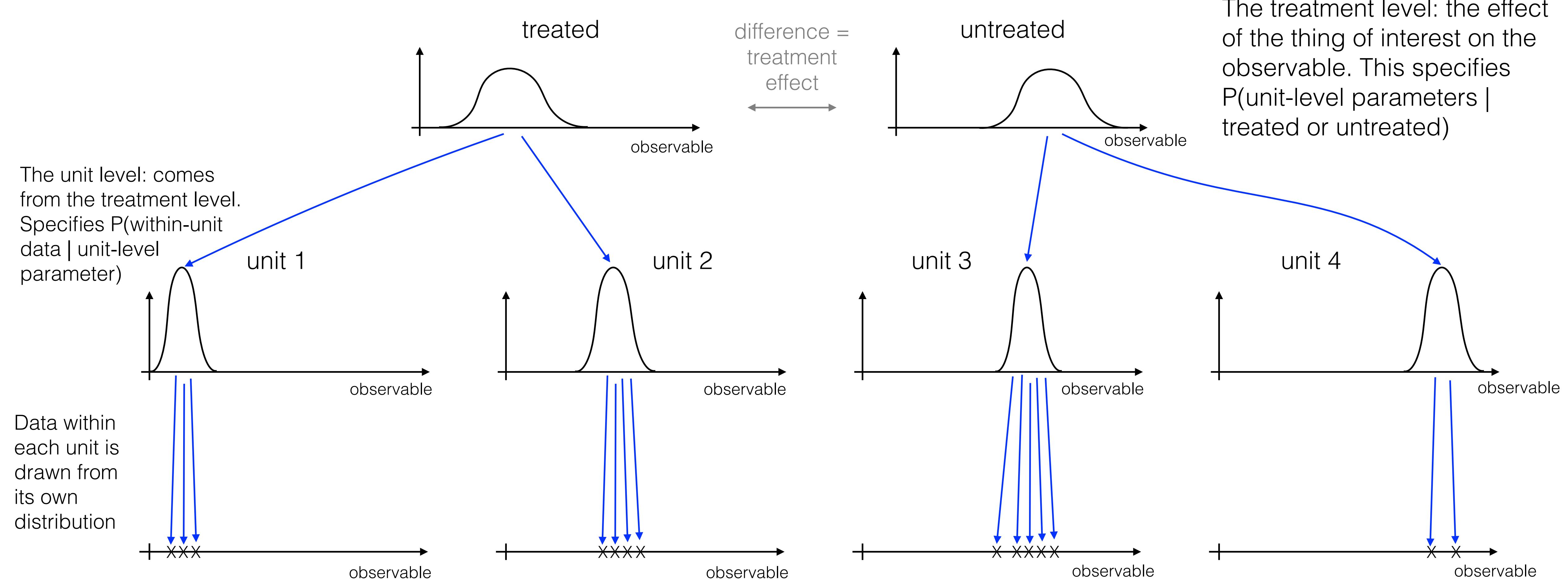
Random effects recap II

	Frequentist	Bayesian
When might you use random effects	<p>When data is generated for each unit, and there is an obvious sampling process for obtaining units from a larger population. e.g. sample some cafés from a population of all cafés, then sample some coffee waiting times from each café.</p>	<p>As with Frequentists, but also whenever it is useful to have a prior with correlated rather than independent parameters, ‘partially pooling’ what they learn from data. e.g. using regression coefficients for each age group, typically we expect these to have roughly similar values. We can encode that assumption and let the data tell us how similar.</p>
How do you estimate the values of random effects	<p>Point estimate: e.g. Best Linear Unbiased Predictors, BLUPs. Uncertainty: I don’t know.</p>	<p>Exactly the same way you estimate fixed effects: with their posteriors.</p>

Random effects recap III

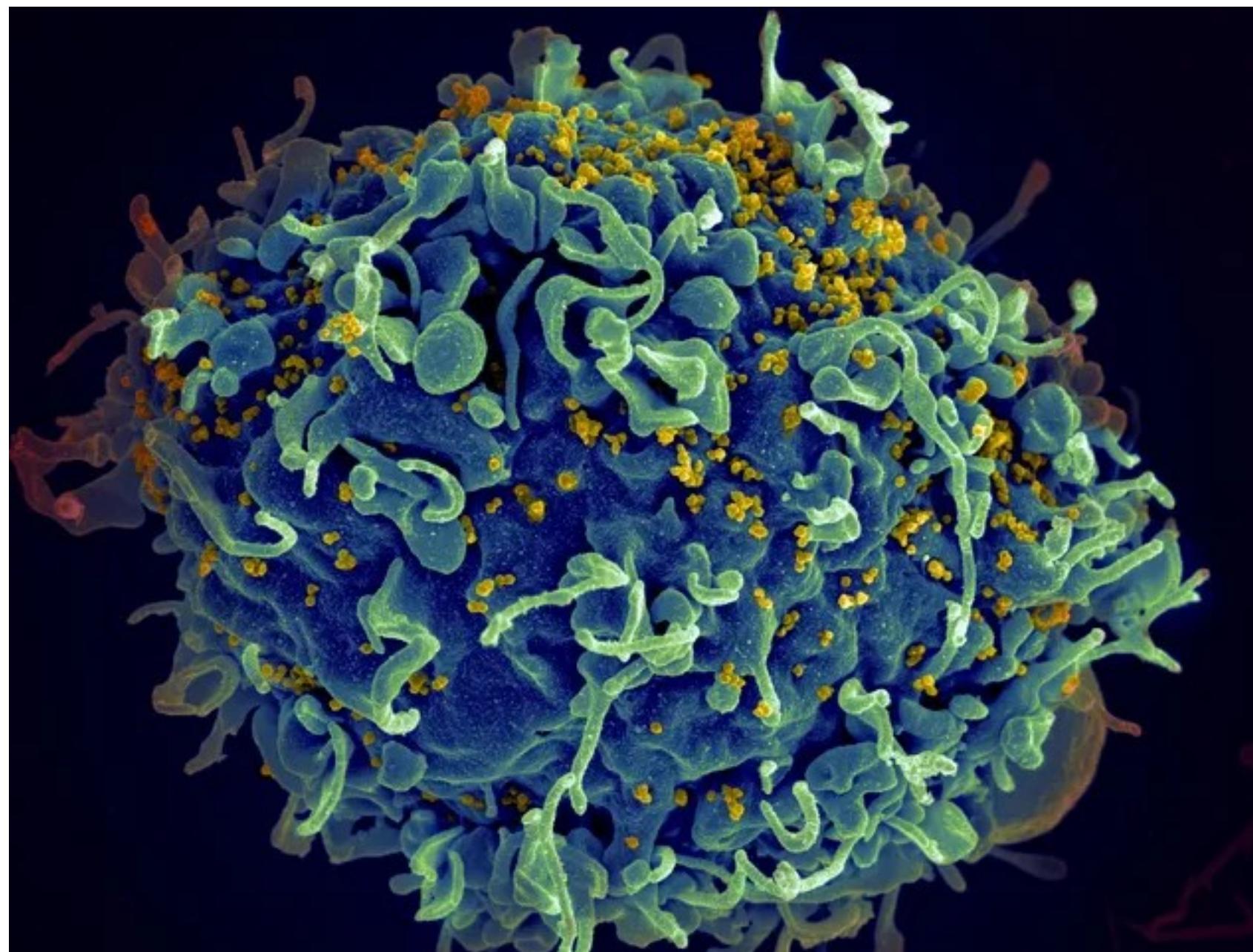
	Frequentist	Bayesian
Working with fixed-effect parameters only: p_f	Likelihood = $P(\text{data} \mid p_f)$	Likelihood = $P(\text{data} \mid p_f)$ Prior = $P(p_f)$ NB $P(p_f)$ is a distribution that may be described with parameters, but these are kept fixed throughout the analysis.
Working with fixed- and random-effect parameters: p_f and p_r	Marginal likelihood = $\int_{p_r} P(\text{data} \mid p_f, p_r) P(p_r \mid p_f) dp_r$ <p style="text-align: center;">↑ Full likelihood ↑ Sampling distribution for p_r</p>	Likelihood = $P(\text{data} \mid p_f, p_r)$ Prior = $P(p_f, p_r)$ = $P(p_r \mid p_f) P(p_f)$

Models with random effects: ‘hierarchical’ or ‘multilevel’



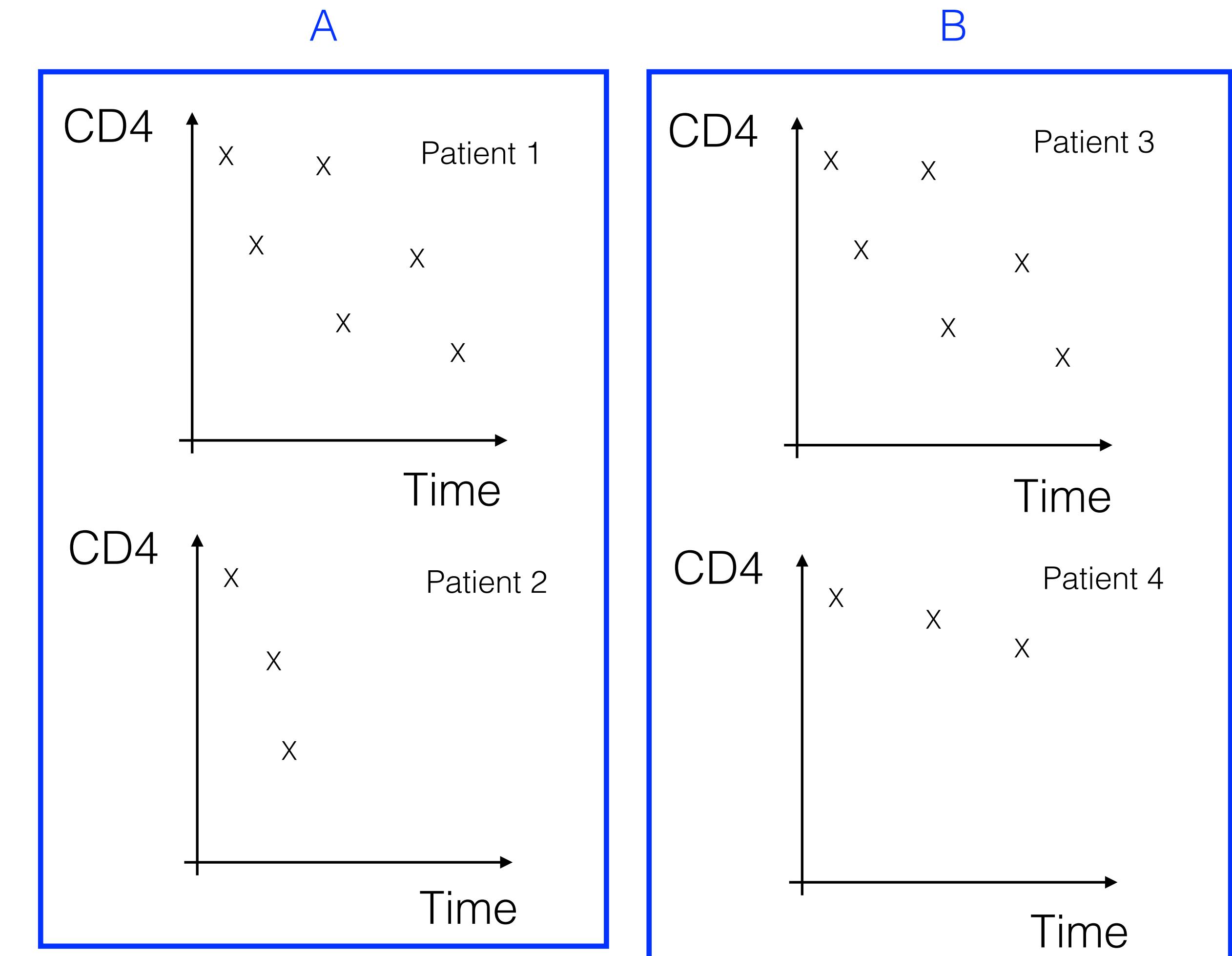
CD4+ T cell decline

HIV virions kill CD4+ T cells, causing their number to decline (until antiretroviral treatment is started).



Yellow = HIV virions
Image: US NIH

Imagine we want to compare CD4 decline between two groups of patients, A and B.

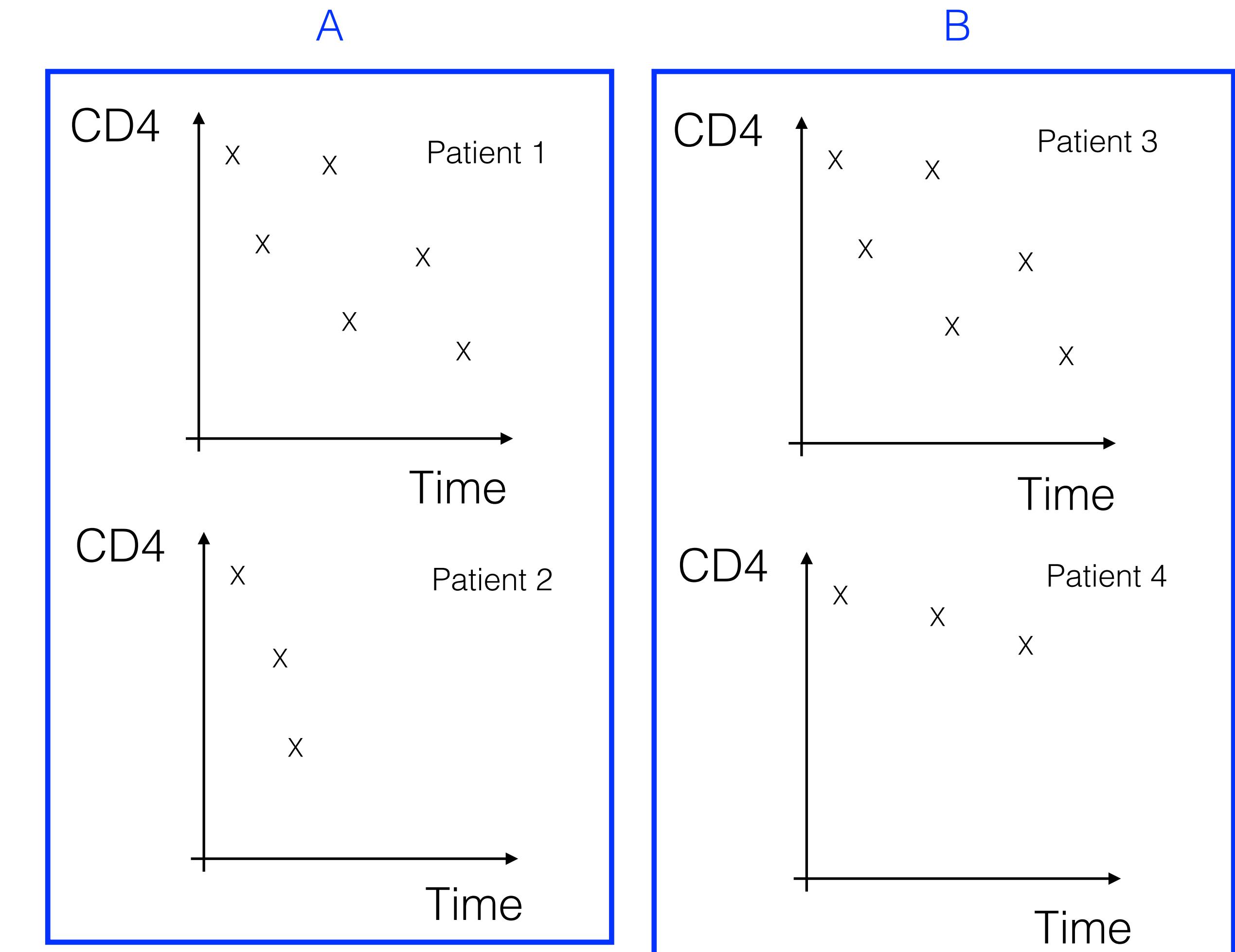


Full pooling of data?

Full pooling = treating all data from within the same group (at the same time) as coming from the same distribution.

Bad idea: there are systematic differences between patients in the same group.
Ignoring these is under-fitting the data.

If we want to predict a new observation at the same point for two different individuals in the same group, we should not make the same prediction for both.

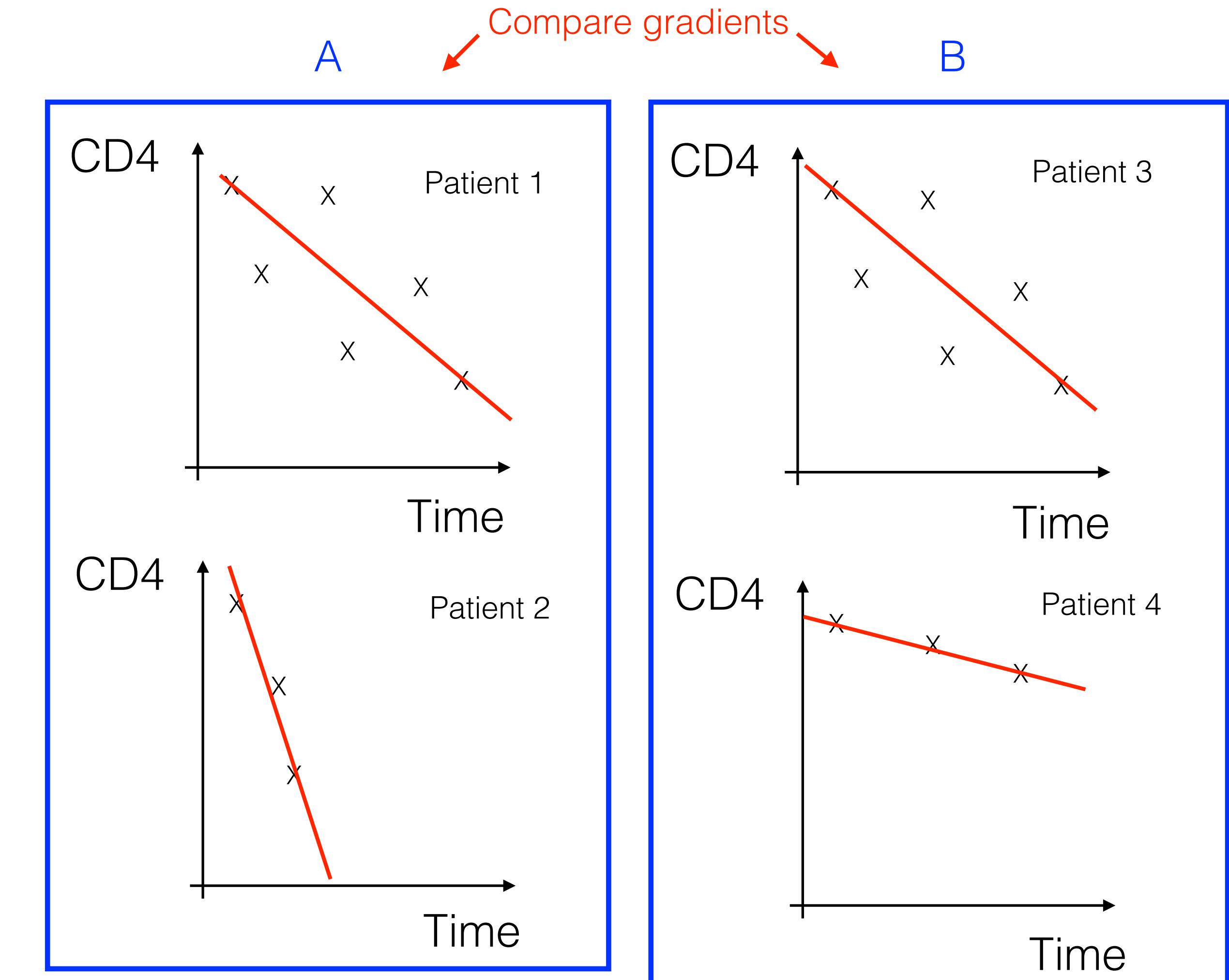


No pooling of data?

We could estimate the **per-patient decline**

for patients 1-4 separately. However,

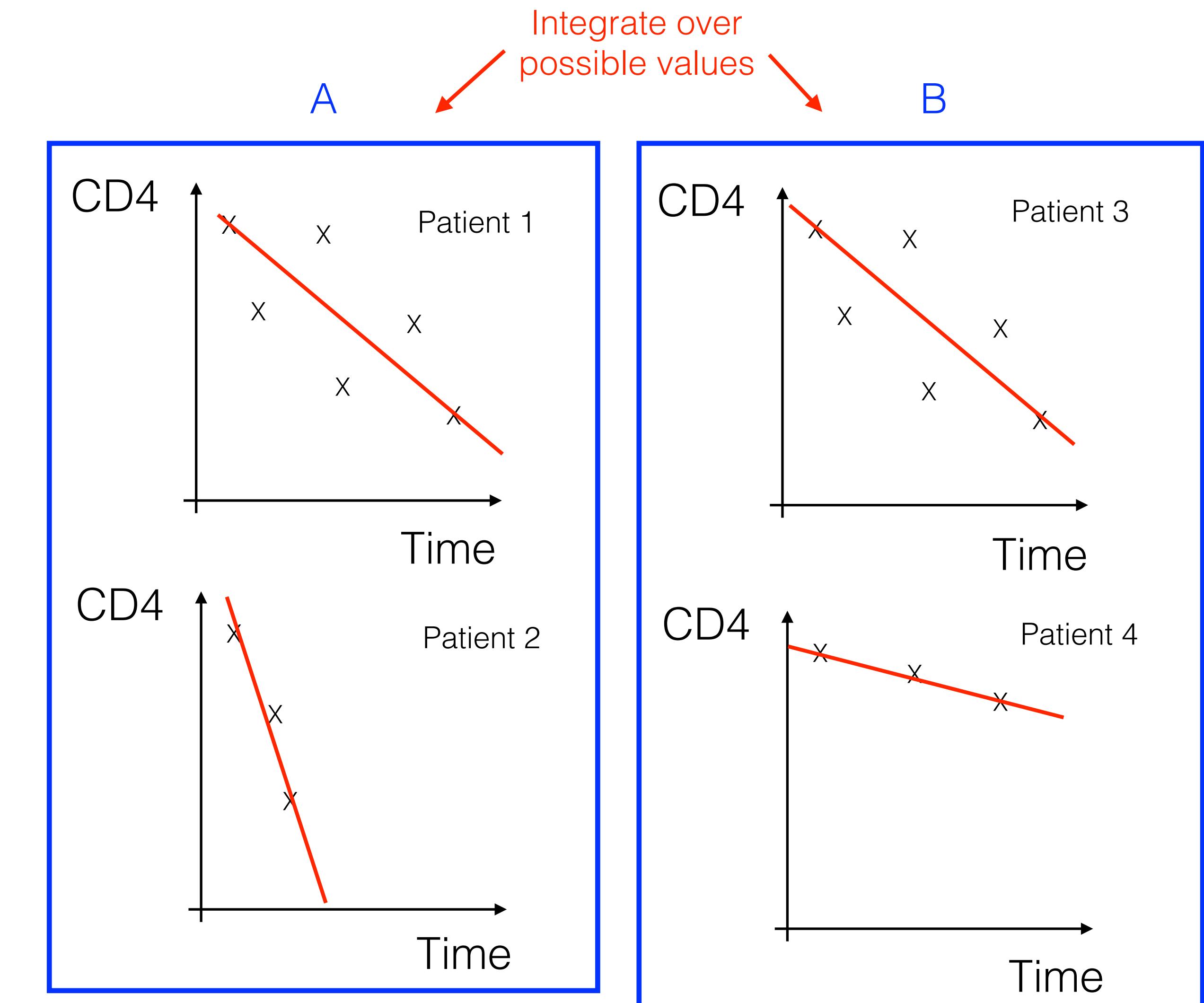
- i. these aren't directly of interest: we'd need a second step comparing these estimates for A vs B;
- ii. silly to do estimation for different patients completely independently, i.e. saying that what we learn from all *other* patients is wholly uninformative for *this* patient.



Group-level parameters as *nuisance parameters*

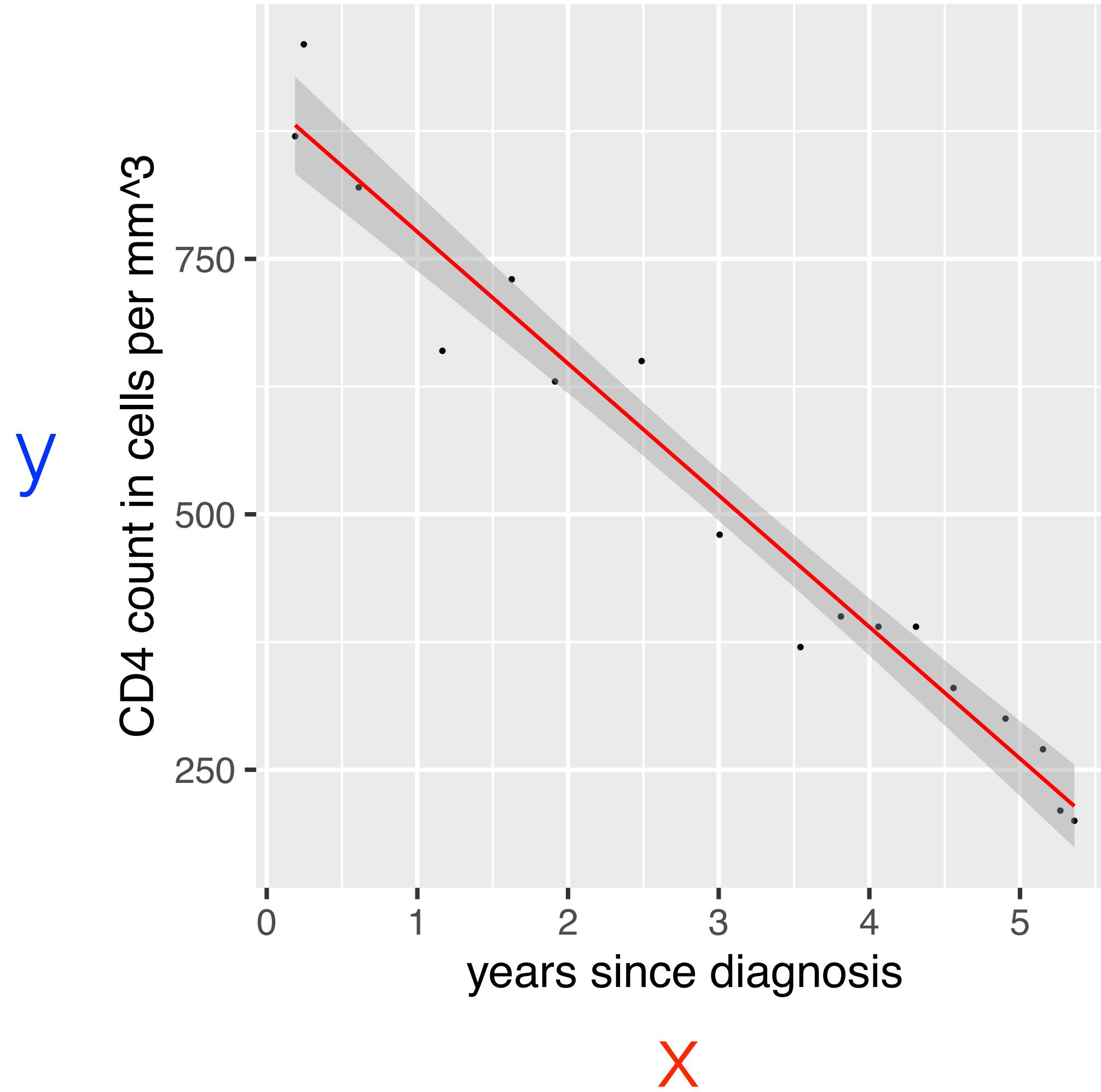
Individual-level parameter values are effectively *nuisance parameters* here. Per the law of total probability, they should be integrated over weighted by how likely they are. But how likely are they? We need to specify a model.

Generally, assume the value for each group has the same distribution of others: assume exchangeability of patients. This shares information between patients - “partial pooling” - instead of examining each one independently of the others. We then estimate the parameters of this distribution.



Effect of VB variant on CD4 before ART: model

$$y = mx + c$$



Let m and c be different for:

- VB and not-VB (fixed effect)
- males and females (fixed effect)
- each age category (fixed effect, c only)
- every individual (random effect: variation constrained to be normally distributed)

Terminology apology: “the lineage” = the VB variant

```
lmm <- lmer(data = df_cd4_decline,
  # Model CD4 counts as...
  cd4_count ~
    # a linear function of time,
    years_since_diagnosis +
    # with a fixed effect of age on the intercept,
    age_diagnosed +
    # a fixed effect of sex on both intercept and slope,
    years_since_diagnosis * sex +
    # a fixed effect of the lineage on both intercept and slope,
    years_since_diagnosis * in_lineage +
    # and a random effect of the individual on both intercept and slope
    (years_since_diagnosis | id_paper))
```

```
lmer(cd4_count ~ years_since_diagnosis +
    age_diagnosed +
    years_since_diagnosis * sex +
    years_since_diagnosis * in_lineage +
    (years_since_diagnosis | id))
```

Revisit this slide for practical exercise 3b

For individual i , the expected CD4 count changes linearly with time ($x_{i,n}$ = time of i 's n th CD4 count) 

$$P(y_{i,n} | \bar{y}_{i,n}, \epsilon^2) = N(y_{i,n} | \bar{y}_{i,n}, \epsilon^2)$$

$$\bar{y}_{i,n} = m_i x_{i,n} + c_i$$

$$m_i = m + \gamma G_i + \lambda L_i + s_i$$

n th CD4 count for individual i is normally distributed around some expected value with a variance parameter common to all observations

The intercept for i

Γ = effect of sex on intercept

Λ = effect of lineage on intercept

$A_{i,age}$ = 0-or-1 coding of i 's age group

a_{age} = effect of age group on intercept

r_i = random effect of "being i " on intercept

$$c_i = c + \Gamma G_i + \Lambda L_i + \sum_{\text{ages}} a_{age} A_{i,age} + r_i$$

$$P\left(\begin{pmatrix} r_i \\ s_i \end{pmatrix} | \sigma_r, \sigma_s, \rho\right) = N\left(\begin{pmatrix} r_i \\ s_i \end{pmatrix} | \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_r^2 & \rho\sigma_r\sigma_s \\ \rho\sigma_r\sigma_s & \sigma_s^2 \end{pmatrix}\right)$$

The slope for i

G_i = sex of i (0 or 1), γ = effect of sex on slope

L_i = lineage of i (0 or 1), λ = effect of lineage on slope

s_i = random effect of "being i " on slope

The random effects are normally distributed around zero. Correlation $\rho \in [-1, 1]$ between the effects on slope and intercept for same individual i ; no correlation between random effects for different individuals i and j .

Bayesian implementation of the same model with Stan

```

data {
    ...
}

parameters {
    ...
    real slope_pat_scale;
    real inter_pat_scale;
    real rho;
    vector[2] inter_and_slope_per_pat_unscaled[num_pats];
}

transformed parameters {
    ...
    vector[num_data] cd4_expected;
    cd4_expected = inter_ref + slope_ref * time +
        design_matrix_for_inter * beta_inter +
        design_matrix_for_slope * beta_slope .* time;
    for (i in 1:num_data) {
        cd4_expected[i] = cd4_expected[i] +
            inter_and_slope_per_pat_unscaled[pat[i], 1] * inter_pat_scale +
            time[i] * inter_and_slope_per_pat_unscaled[pat[i], 2] * slope_pat_scale;
    }
}

model {
    inter_and_slope_per_pat_unscaled ~ multi_normal(zeros, Rho);
    cd4 ~ normal(cd4_expected, sd_error);
}

```

$$\left(\begin{pmatrix} r_1/\sigma_r \\ s_1/\sigma_s \end{pmatrix}, \begin{pmatrix} r_2/\sigma_r \\ s_2/\sigma_s \end{pmatrix}, \dots \right)$$

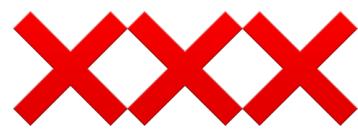
$\bar{y}_{i,n} = m_i x_n + c_i$
 $m_i = m + \gamma G_i + \lambda L_i + s_i$
 $c_i = c + \Gamma G_i + \Lambda L_i + \sum_{\text{ages}} \alpha_{\text{age}} A_{i,\text{age}} + r_i$

$\frac{s_i}{\sigma_s} \times \sigma_s$
 $x \times \frac{r_i}{\sigma_r} \times \sigma_r$

$\left(\begin{pmatrix} r_i/\sigma_r \\ s_i/\sigma_s \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \right)$

$y_{i,n} \sim N(\bar{y}_{i,n}, \epsilon^2)$

Closing thought: your question



Data

- do things to it
- get result
- speculate what it means



What questions *could* you ask

- which question *should* you ask
- what are the possible answers
- to what extent do the data discriminate between them



↑ Reposted by Peter Tennant, PhD
Sean Mackinnon @seanpmackinnon.bsky.social · 17h
Stats consulting is constantly like:

Them: I want you to run (complex stats)

Me: OK, what's your research question though?

Them: ...A (complex stat)?

Me: Research question?

Them: You know, like the stats in this journal article. Something reviewers will like.

