The need for hierarchical* models to infer things from naturally grouped data

Chris Wymant

Disclaimer: these slides contain my work-in-progress understanding. I’m far from an expert.

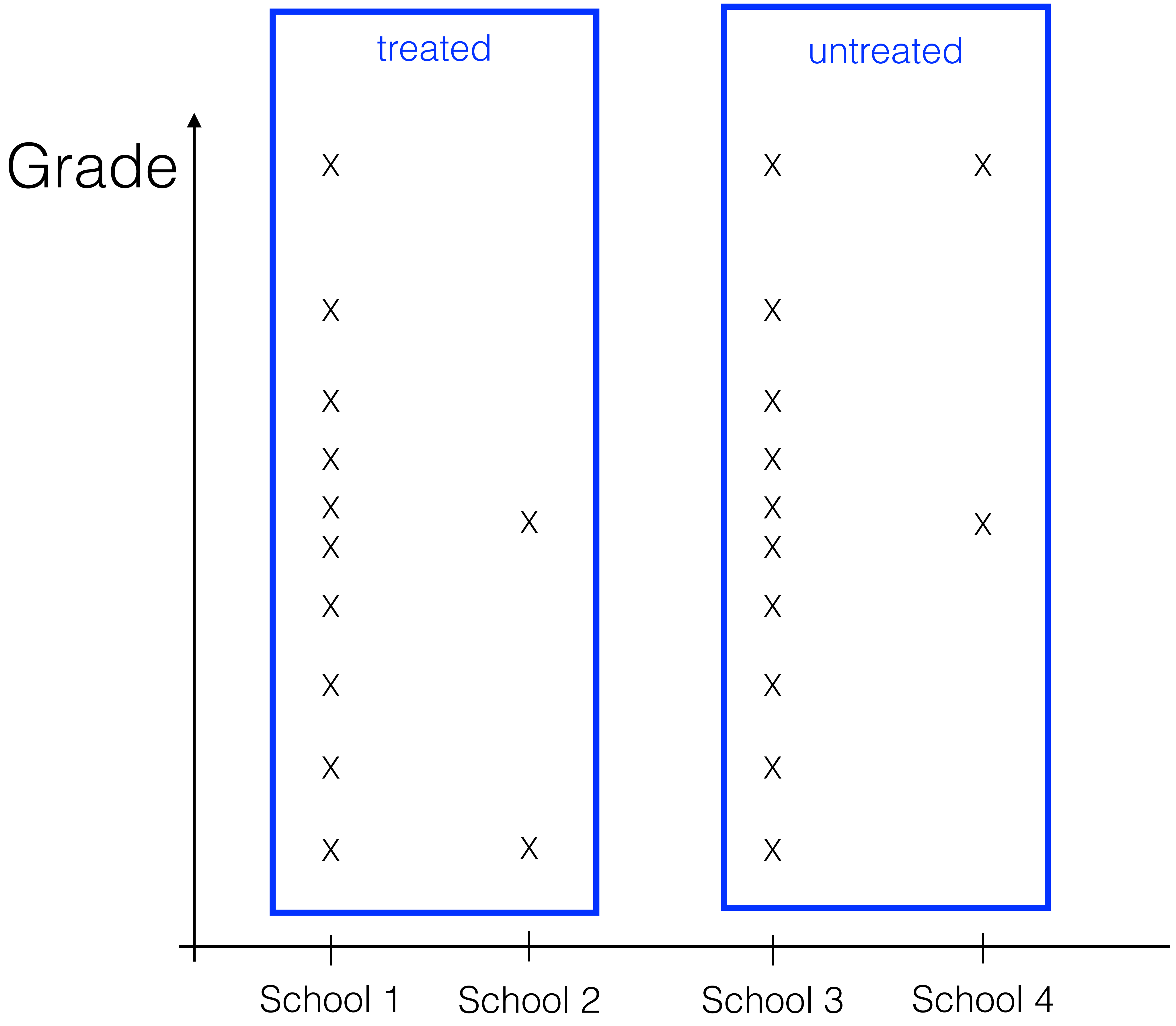
* a.k.a. multi-level, mixed effects or random effects

See also “Inferring things from (quantitative) data” and lectures on probability at github.com/ChrisHIV/teaching

We’re interested in the compatibility between data and hypotheses. This is typically not binary - compatible or incompatible - but is a question of degree. Probability is the natural language for quantifying it.

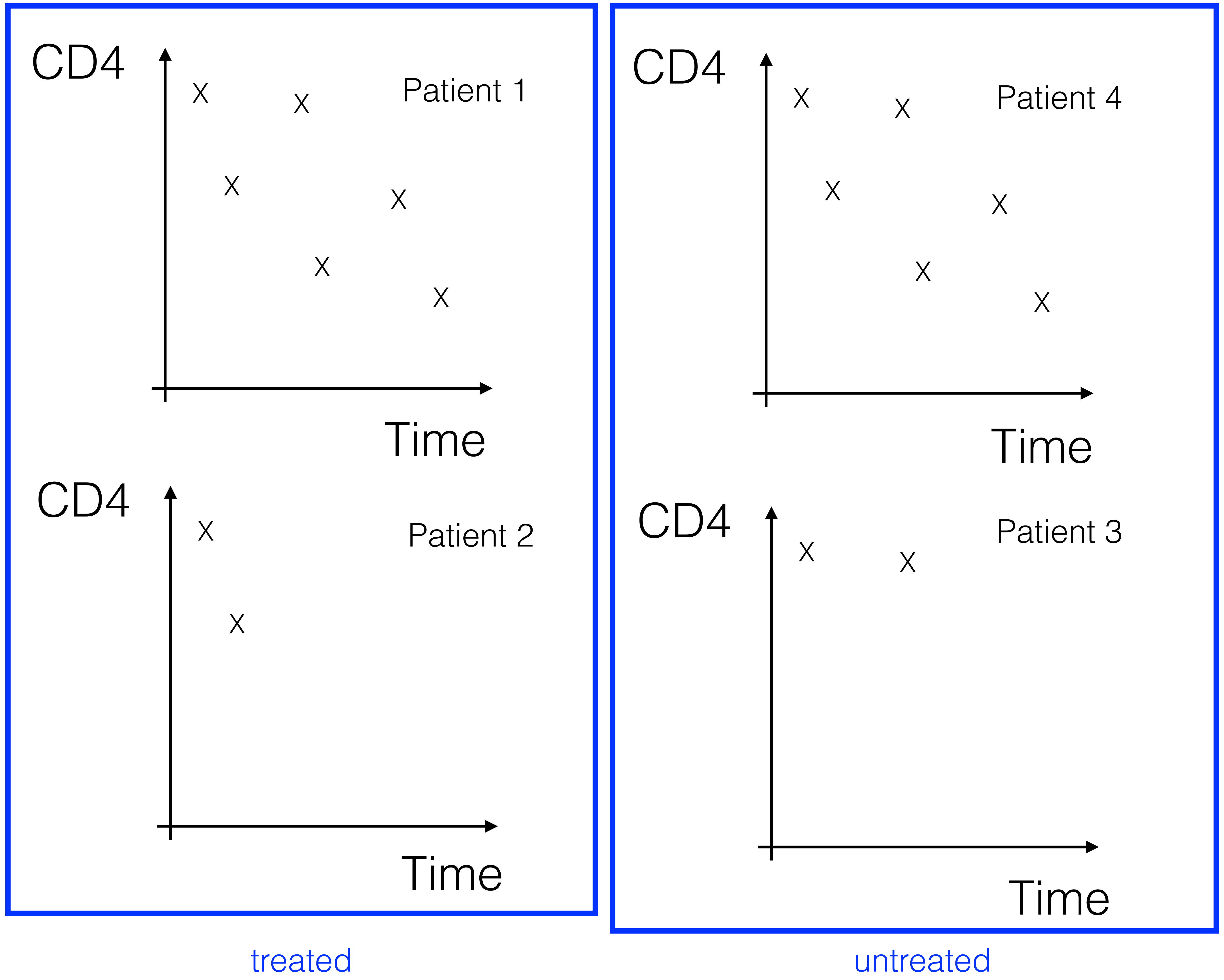
Motivating example one

Grades in treated vs untreated schools



Motivating example two

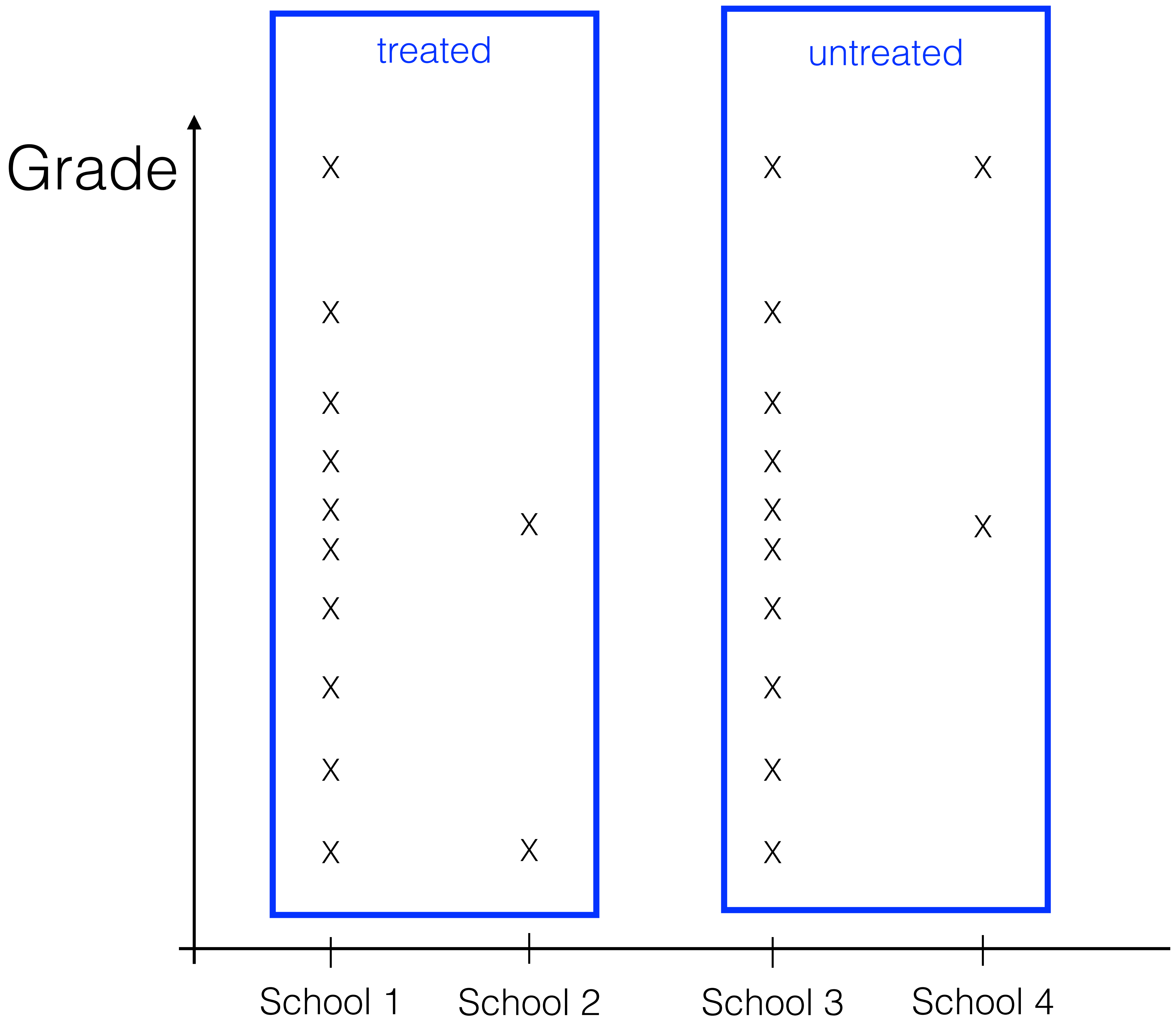
Rate of CD4 cell decline in “treated” vs untreated HIV patients (imagine a treatment not as dramatic as usual ART)

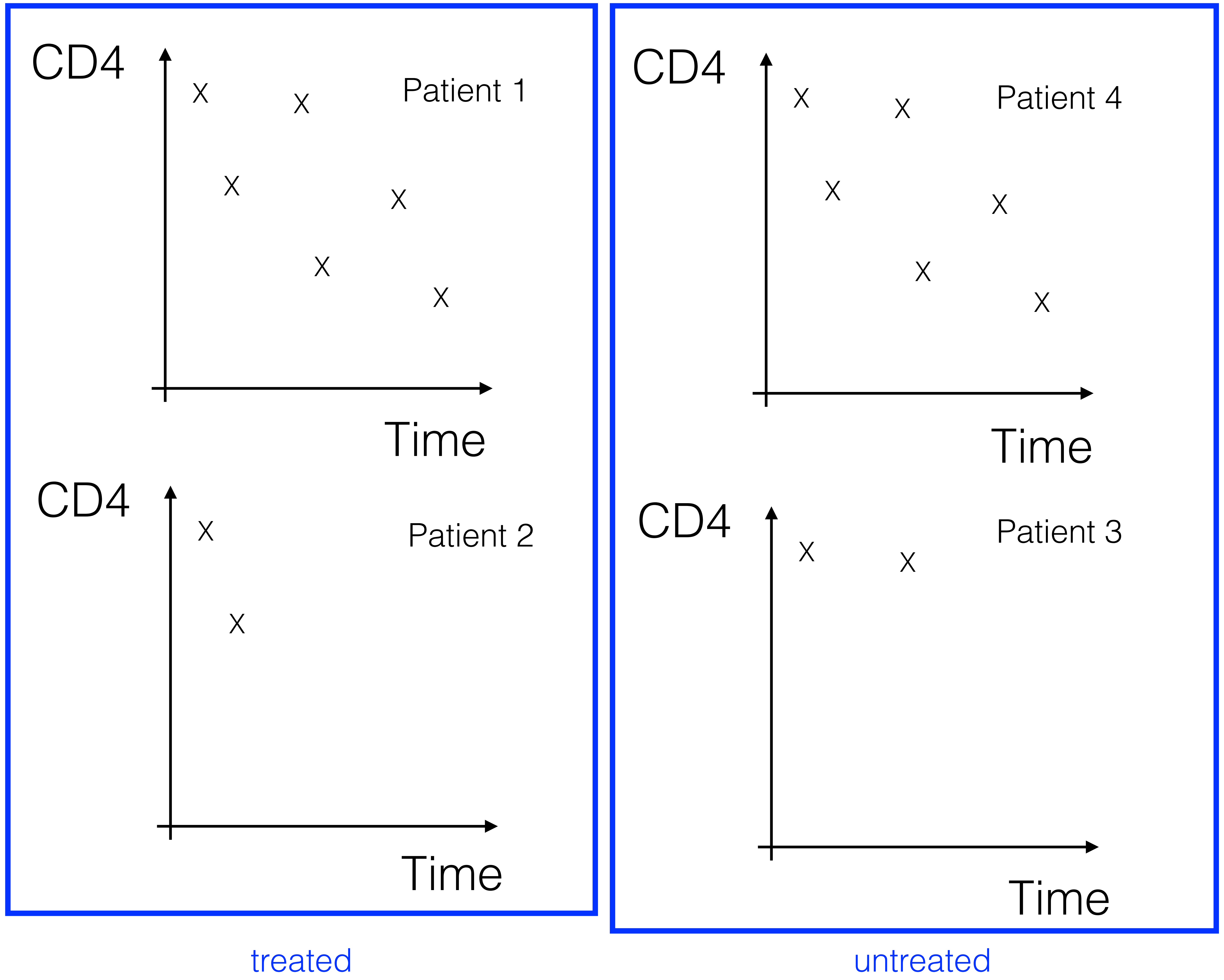


Full pooling of data?

Treat all data from within the same group as coming from the same distribution? No. There’s reason to believe that there are systematic differences between schools and between patients’ immune responses to HIV, a priori and reflected in the data.

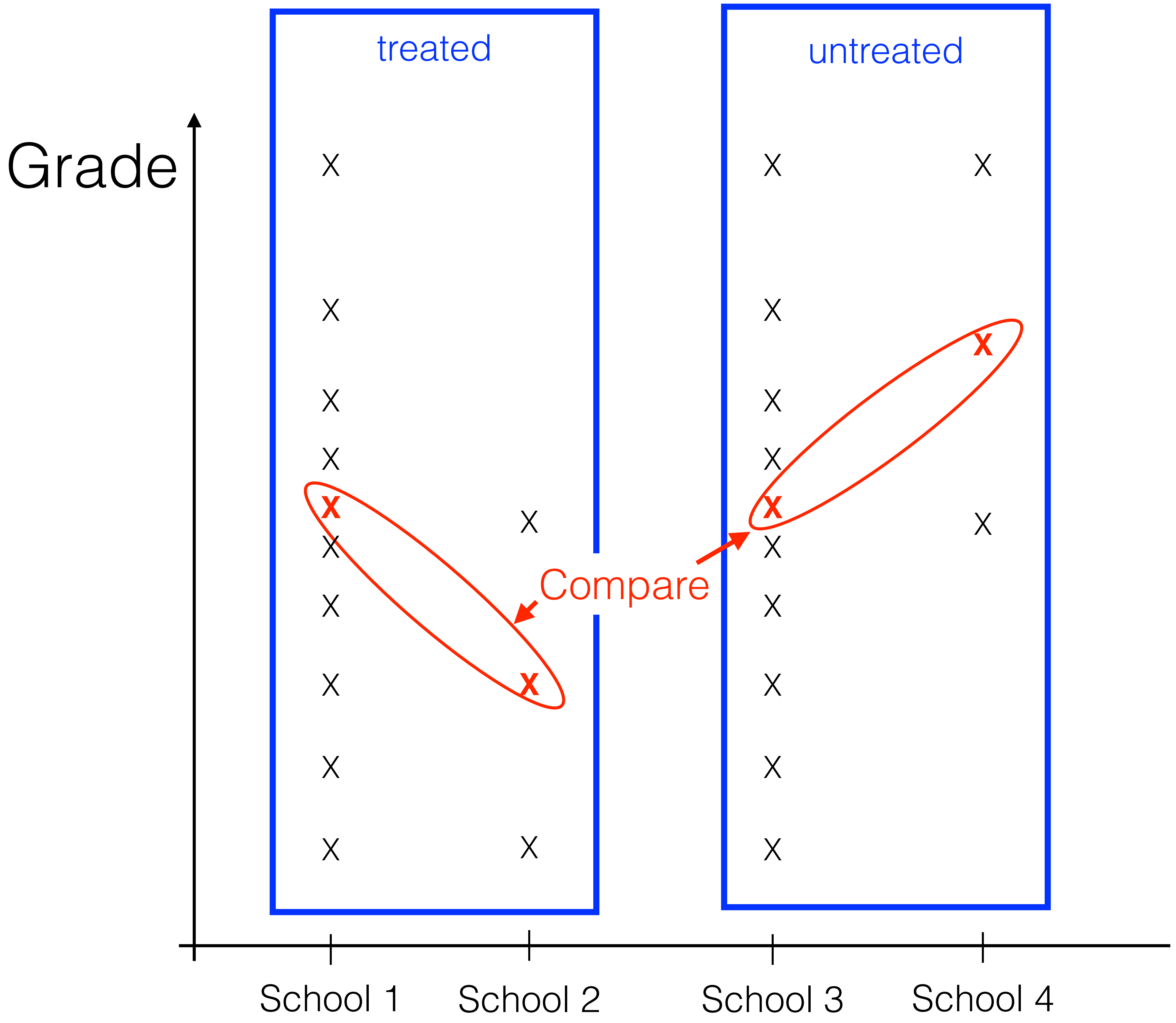
A fully pooled model is underfitting the data.
One systematic difference between groups should not be considered as getting an unlikely observation within a group again and again and again.



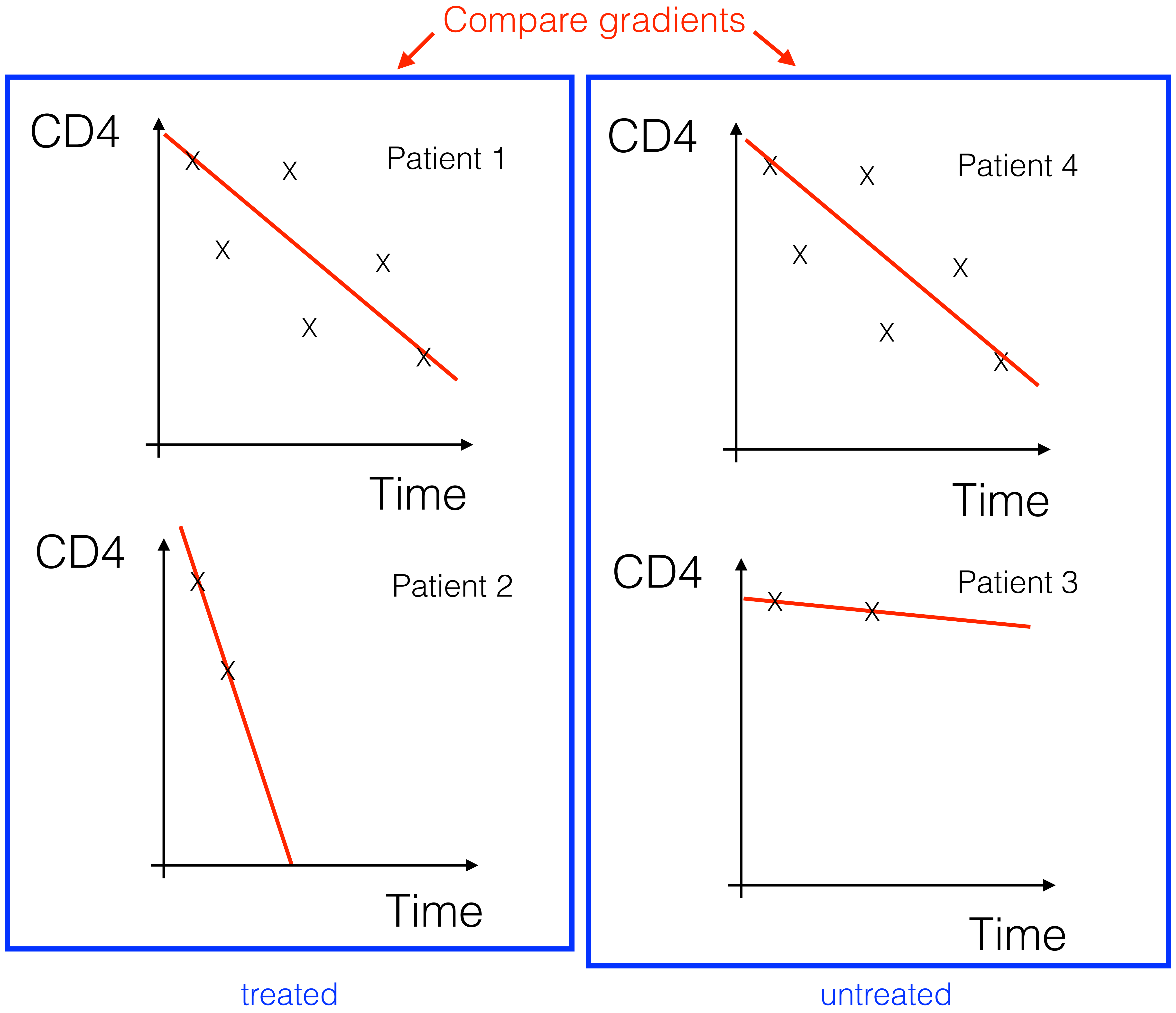


No pooling of data?

We could estimate the **per-school mean** for schools 1-4 separately. However, (a) these aren't directly of interest: need a second step comparing them for treated and untreated schools; (b) perhaps not optimal to model each school as completely independent of all the others?



Can estimate the **per-patient decline** mean for patients 1-4 separately. Same issues as the school example.



Interlude: the law of total probability

Possibilties being “mutually exclusive” means *at most one* of them can be true. (To get the probability that any of them is true, you can simply add their individual probabilities, because there’s no overlap and so no double-counting of probability.)

Possibilities being “collectively exhaustive” means *at least one* of them must be true.

Possibilties being “mutually exclusive and collectively exhaustive” (ME&CE) implies *exactly one* of them must be true. If some possibilies are ME&CE, the sum of their probabilities is 1.

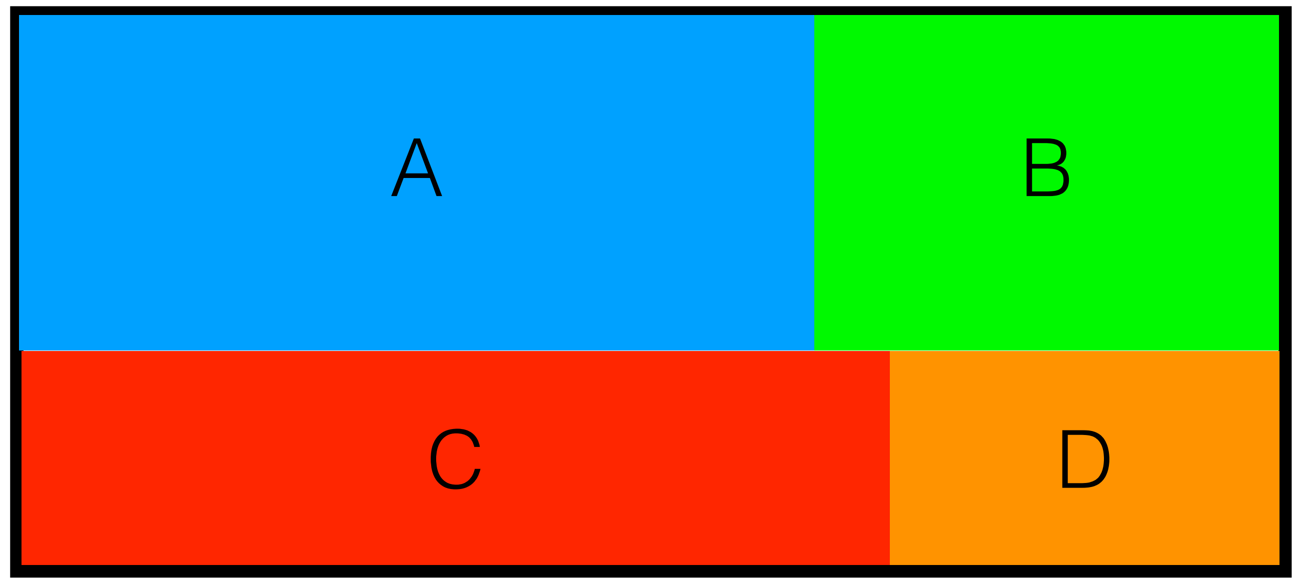
Example sets of possibilities for the result of rolling a die once:

	Mutually exclusive	Not mutually exclusive
Collectively exhaustive	●Result is 1-3 ●Result is 4-6	●Result is 1-4 ●Result is 4-6
Not collectively exhaustive	●Result is 1 ●Result is 2	●Result is 1-2 ●Result is 2-3

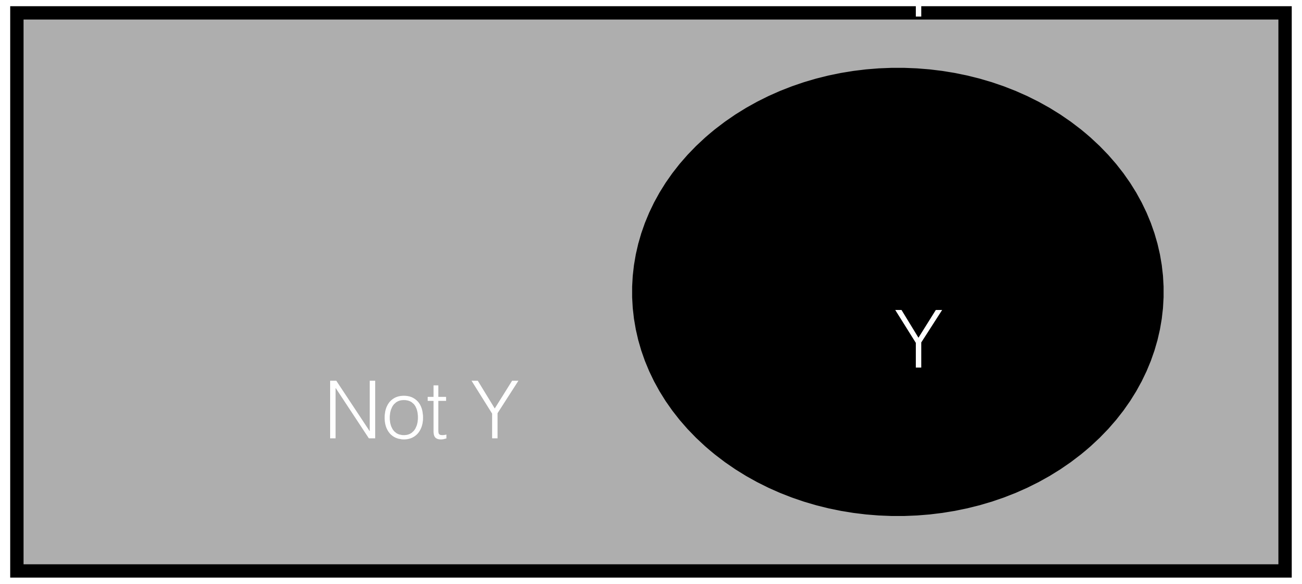
You can take the collection of all things that are possible and split it into ME&CE groups in different ways.
e.g. the day of the week today = Mon, Tue, … Sun, and
The mean temperature here today is <10°C, or ≥10°C

Space of everything that’s possible

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

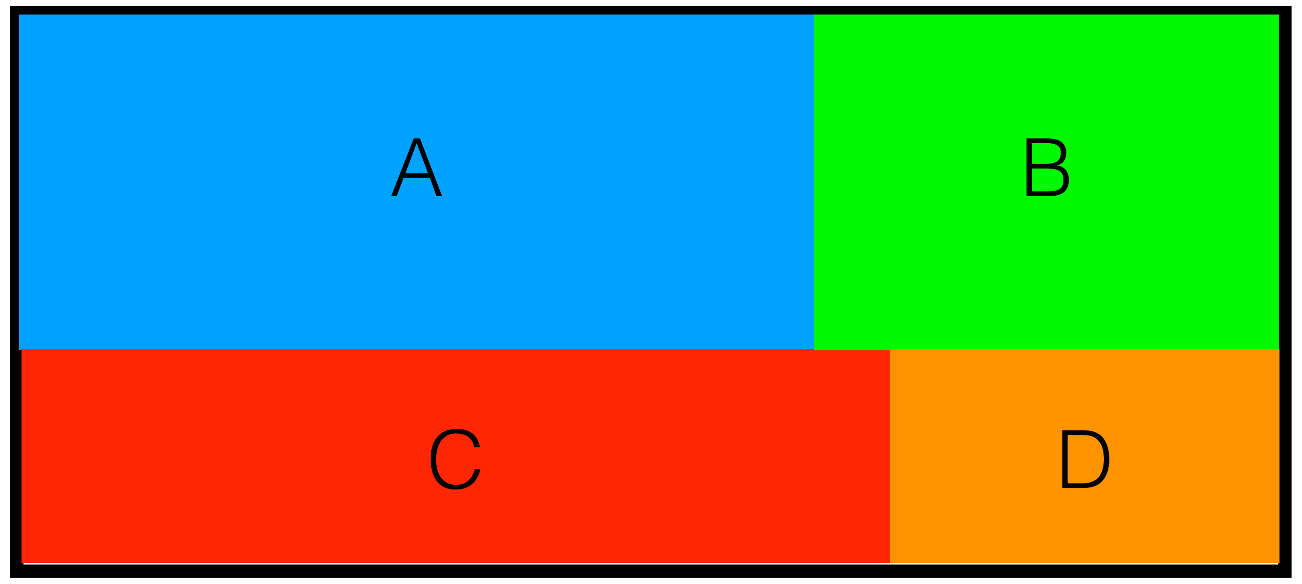


Another way of splitting into ME&CE groups: Y or not Y

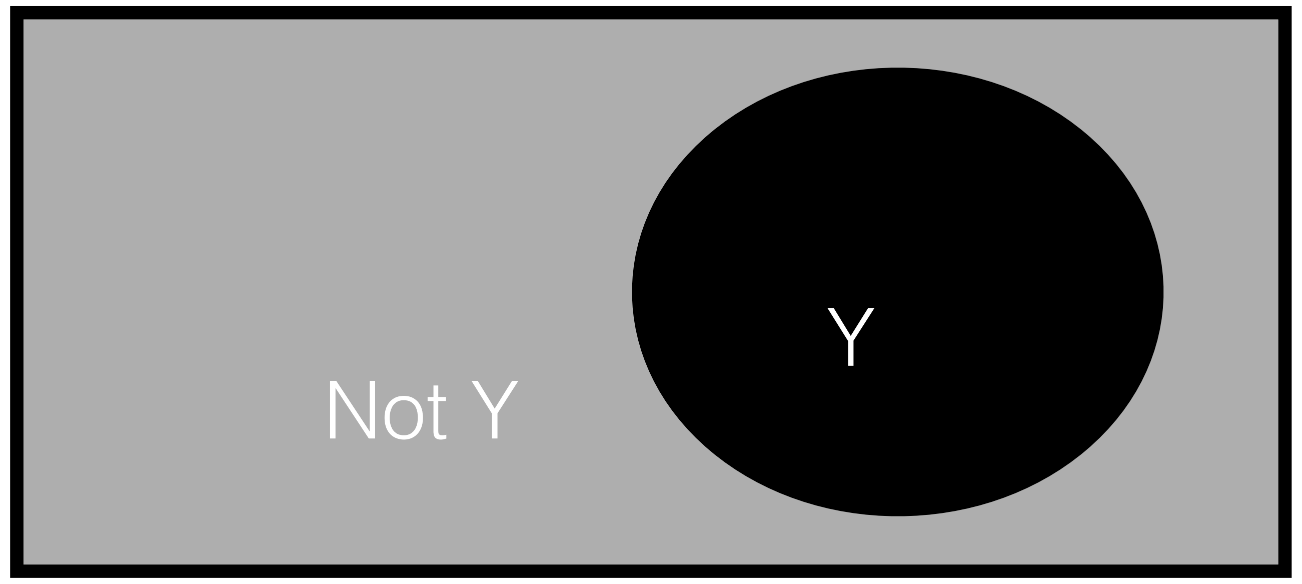


Where area = probability

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

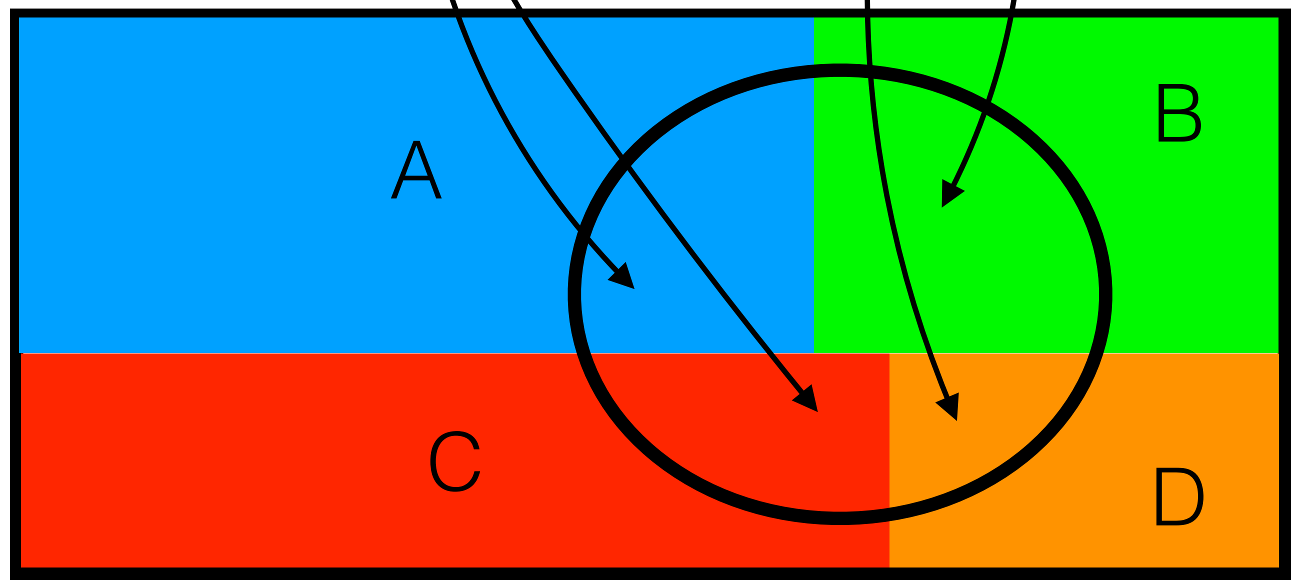


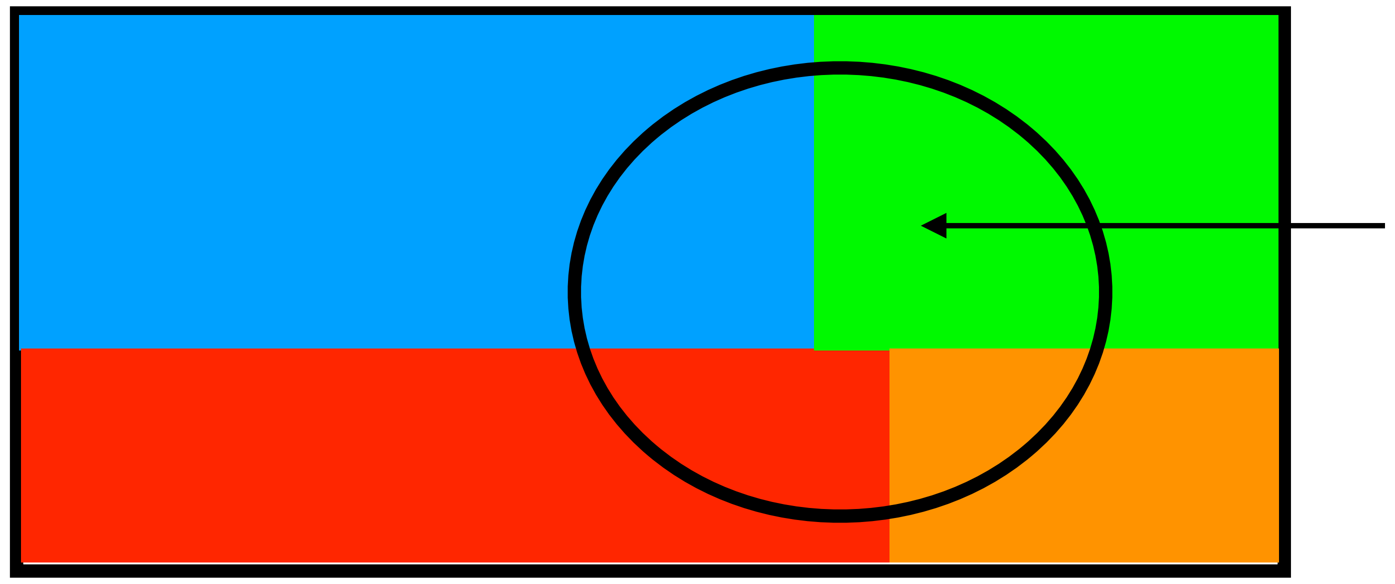
Another way of splitting into ME&CE groups: Y or not Y



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

P(Y) = P(Y and A)
+P(Y and B)
+P(Y and C)
+P(Y and D)





P(Y and B)
= fraction of whole space that's Y × fraction of Y that's B
= P(Y) P(B | Y)
= P(B | Y) P(Y)

So in the end,

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

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, … A_N (any way you like except "Y or not Y" which would result in something true but unhelpful)
- then you have

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

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

Example: P(I want a beer) = P(I want a beer | it's a weekday) P(it's a weekday) +
P(I want a beer | it's the weekend) P(it's the weekend*)

*defined to exclude Friday night

Do say:

- "We should consider all possibilities that are compatible with our outcome of interest, weighting by how likely they are."
- "Integrate over the nuisance parameters" (when your answer depends on A, and a range of values for A are possible, but you want to provide one overall value that's not conditional upon A)

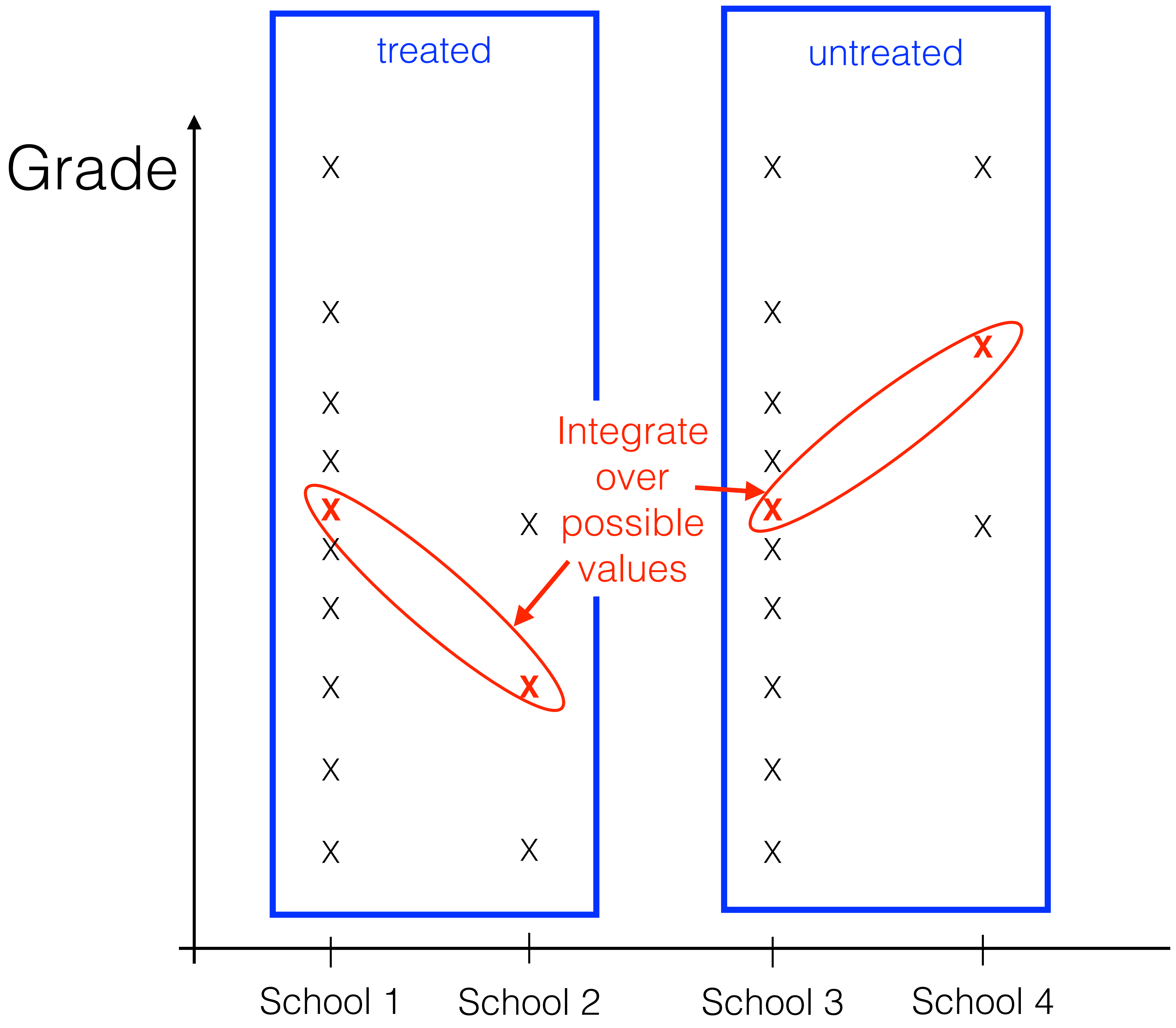
And very similarly, when Y depends on X only through A_i:

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

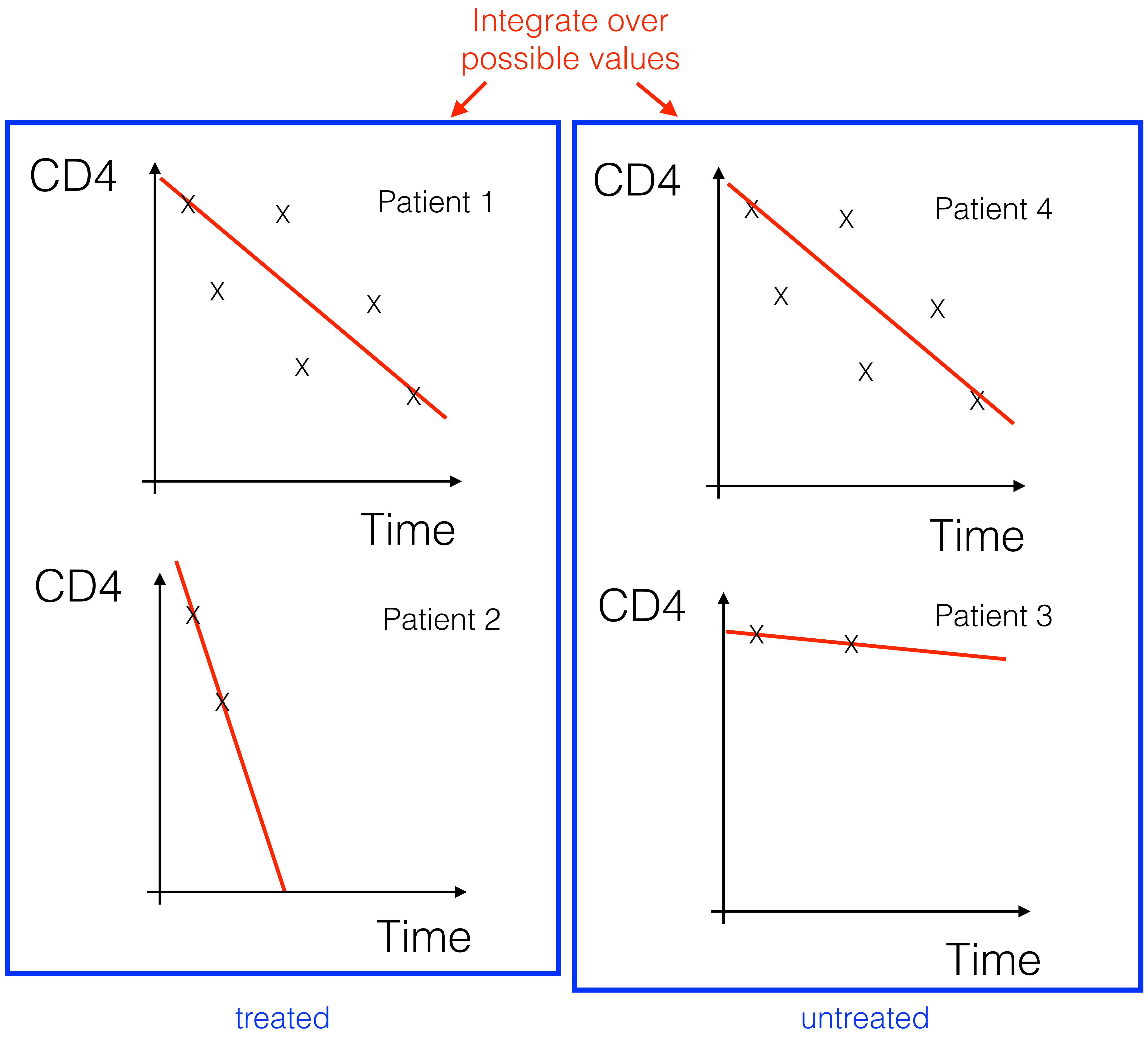
Do say:

- “The probability that a particle is at x_2 at t_2 given that it was at x_1 at t_1 is given by the path integral: the sum of probabilities for all possible paths from x_1 at t_1 to x_2 at t_2.” (An integral over the infinite-dimensional space of functions, rather than over a finite-dimensional space of nuisance parameter values.)

Group-level parameters as *nuisance parameters*



Integrate over possible values



Integrate over possible values

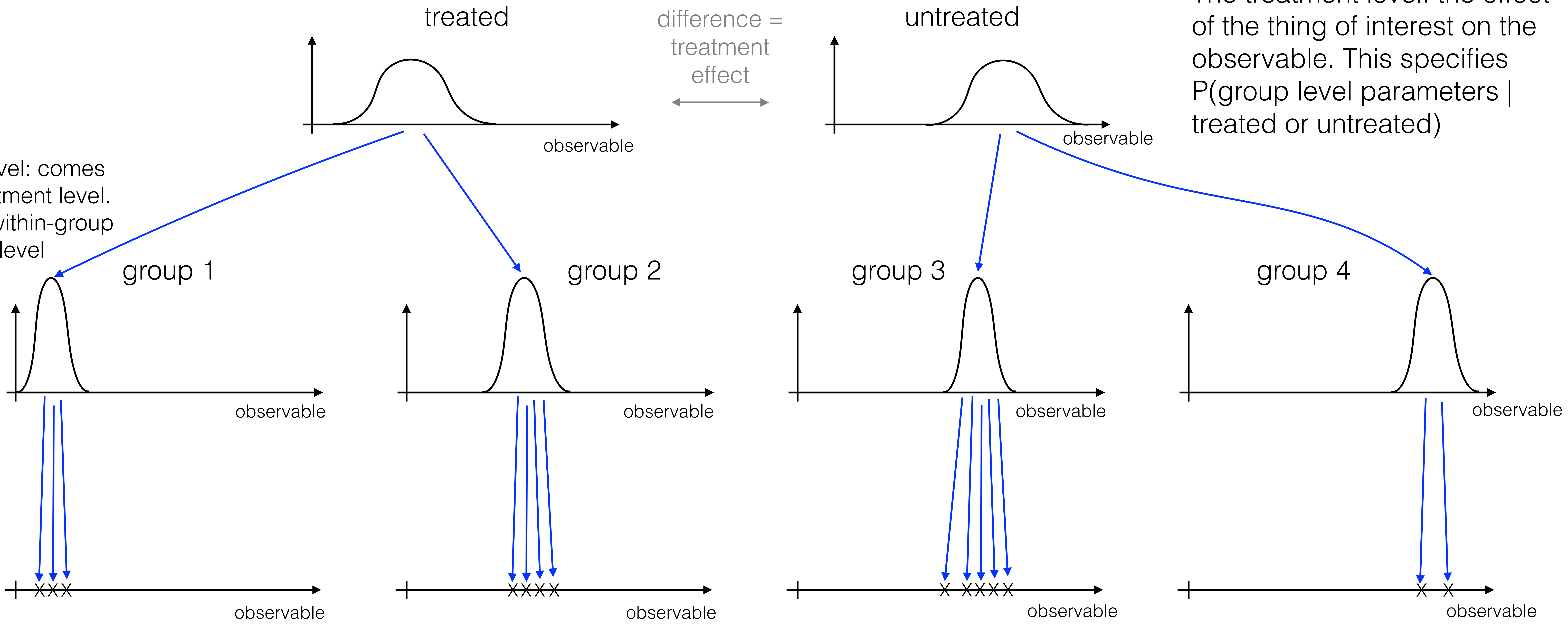
Group-level parameter values are *nuisance parameters*. As per the law of total probability, they should be integrated over weighted by how likely they are. But how likely are they? We need to specify a model.

Generally, assume the value for each group has the same distribution of others, based on ignorance/exchangeability (c.f. individuals in a common population). This shares information between groups - “partial pooling” - instead of examining each one independently of the others. We then estimate the parameters of this distribution. Commonly, assume a normal distribution.

The hierarchy

The treatment level: the effect of the thing of interest on the observable. This specifies P(group level parameters | treated or untreated)

The group level: comes from the treatment level. Specifies P(within-group data | group-level parameter)



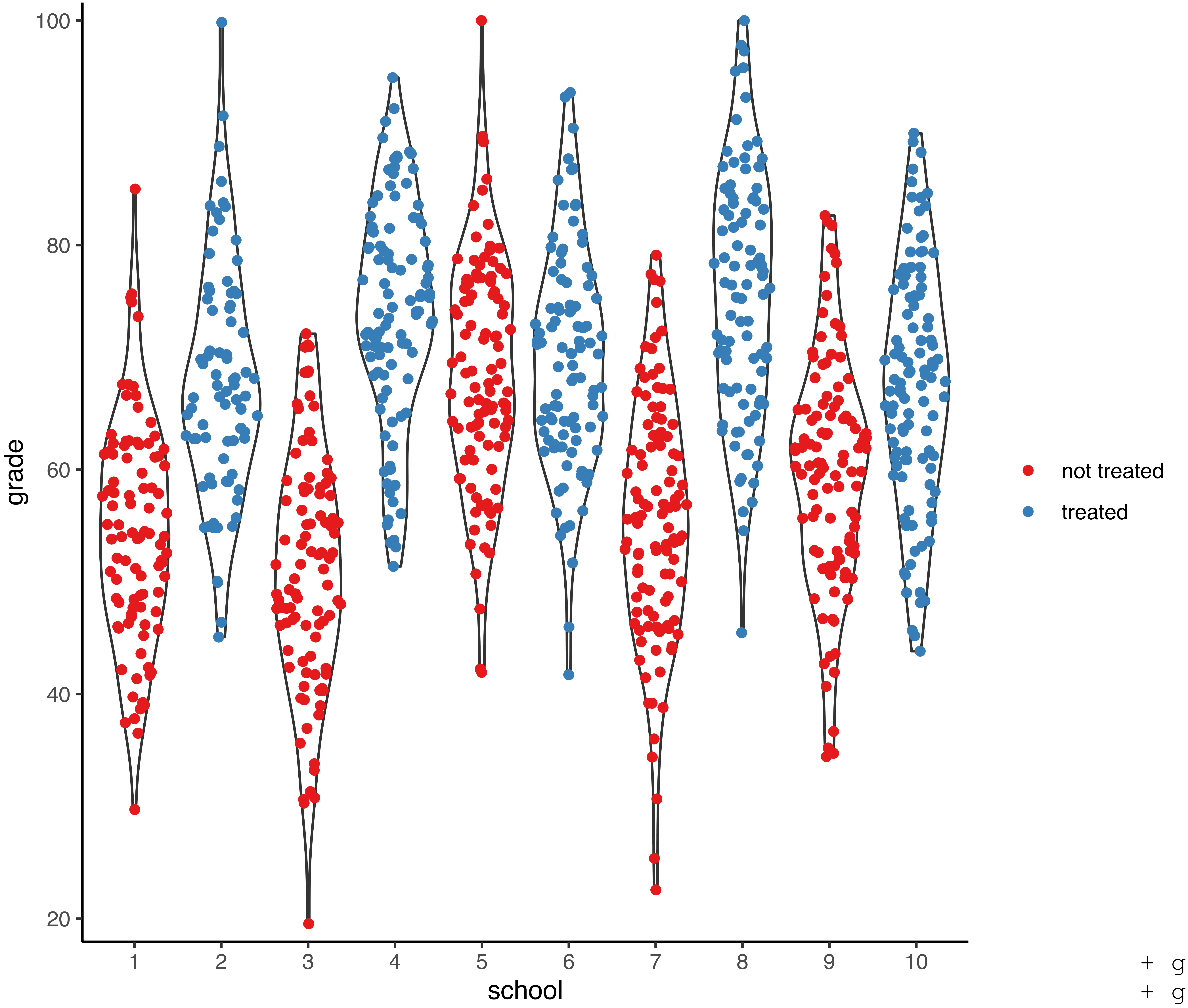
Data within each group is drawn from its own distribution

The group-level parameters (GLPs, defining a distribution for each group) are an intermediate between the data and the object of interest: the distribution for treated vs untreated.
P(data | treatment effect) = P(data | GLPs) P(GLPs | treatment effect) *integrated over GLPs*

Simulate hierarchical data in R…

Full code + comments + plotting etc.
github.com/ChrisHIV/teaching

```r
num_students <- 1000
num_schools <- 10
stddev_students <- 10
stddev_schools <- 5
grade_untreated_mean <- 60
treatment_effect <- 10

df <- tibble(student = 1:num_students,
             school = sample(1:num_schools,
                             size = num_students,
                             replace = TRUE),
             treated = school %% 2 == 0)

school_effects <- rnorm(num_schools,
                        mean = 0,
                        sd = stddev_schools)

df$grade_expected <-
  grade_untreated_mean +
  map_dbl(df$school, function(school_) {school_effects[[school_]]}) +
  if_else(df$treated, treatment_effect, 0)

df$grade <- rnorm(n = num_students,
                  mean = df$grade_expected,
                  sd = stddev_students)
df$grade <- pmin(df$grade, 100)
df$grade <- pmax(df$grade, 0)
```



```
+ ggplot::geom_violin()
+ ggforce::geom_sina()
```

…do frequentist estimation with lme3…

```
lmm <- lmer(data = df,
            grade ~ treated + (1 | school))
summary(lmm)
confint(lmm)
```

lme3: Bates et al. 2015

…or Bayesian inference with Stan

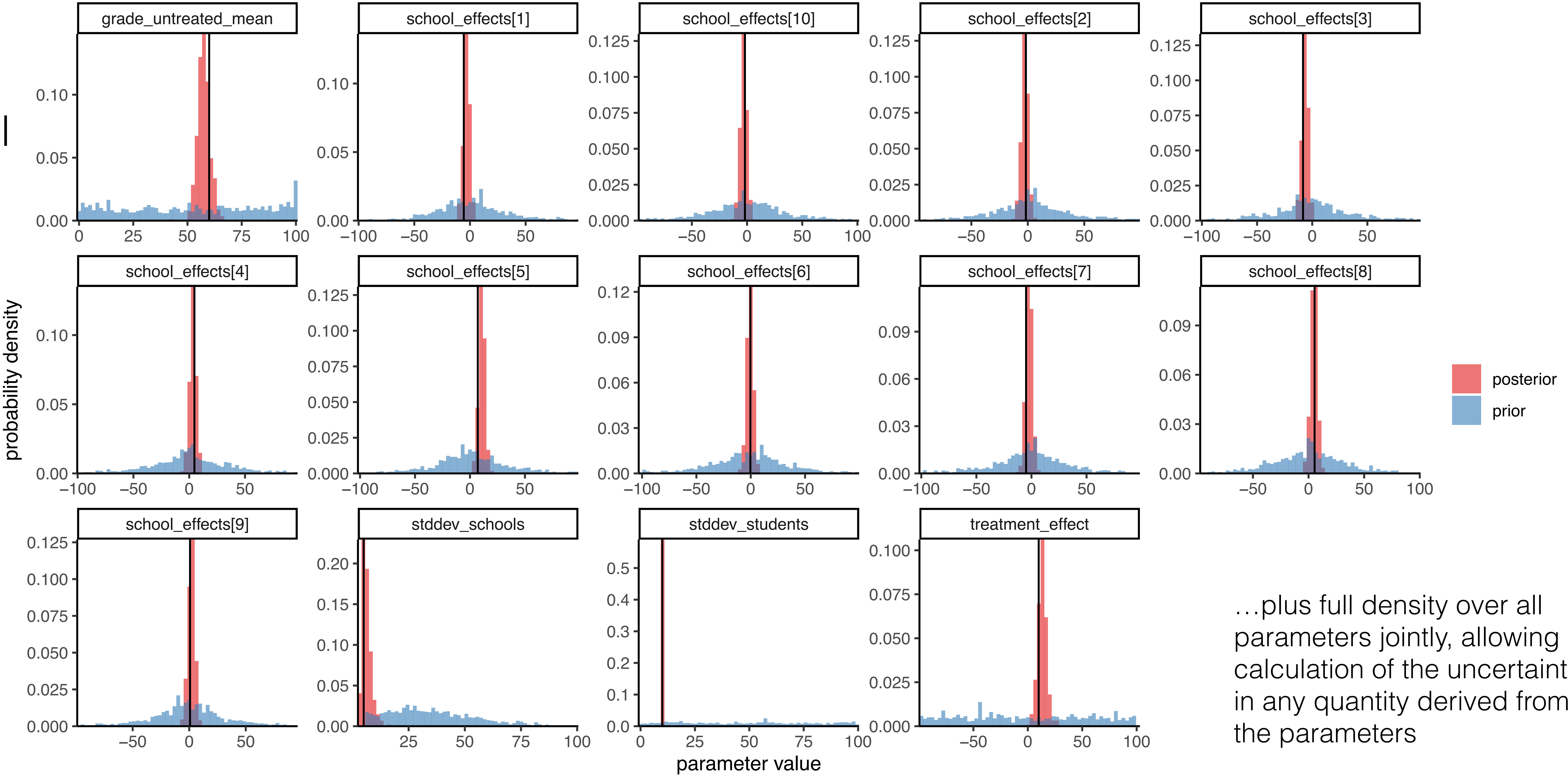
```
data {
  int<lower = 1> num_schools;
  int<lower = num_schools> num_students;
  int<lower = 1,  upper = num_schools> school[num_students];
  int<lower = 0,  upper = 1> treated[num_students];
  real<lower = 0, upper = 100> grade[num_students];
}

parameters {
  real<lower = 0,    upper = 100> stddev_students;
  real<lower = 0,    upper = 100> stddev_schools;
  real<lower = 0,    upper = 100> grade_untreated_mean;
  real<lower = -100, upper = 100> treatment_effect;
  real<lower = -100, upper = 100> school_effects[num_schools];
}

model {
  real grade_expected[num_students];
  for (student in 1:num_students) {
    grade_expected[student] =
      grade_untreated_mean +
      school_effects[school[student]] +
      treated[student] * treatment_effect;
  }

  school_effects ~ normal(0, stddev_schools);

  grade ~ normal(grade_expected, stddev_students);
}
```

← Inefficient parameterisation.
See code online & later pro tip.

lme3:

	2.5th % CI	REML estimate	97.5th % CI	Truth
stddev_students	9.9	10.4	10.8	10
stddev_schools	3.2	5.4	8.2	5
grade_untreated_mean	53	58	63	60
treatment_effect	6	13	20	10

Stan marginals (black vertical line is truth):



…plus full density over all parameters jointly, allowing calculation of the uncertainty in any quantity derived from the parameters

Frequentist: the group-level values are ‘random effects’ that are integrated over without being estimated, unlike fixed effects which are estimated. ‘Mixed effects’ models include both.

Bayesian: no conceptual difference. They’re all just parameters.

Lme3 code: very quick to write

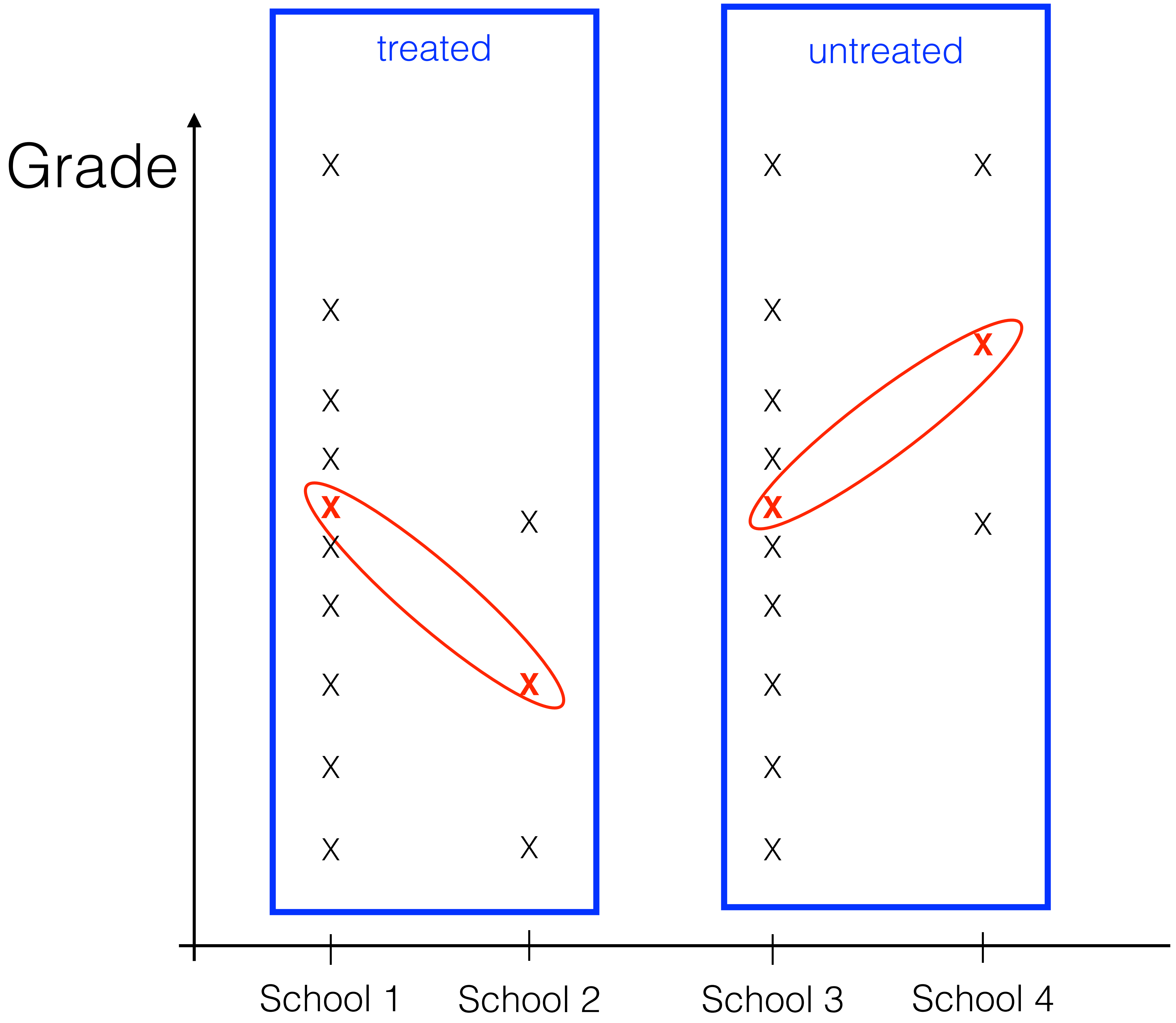
Stan code: constructing the model explicitly piece by piece

- forces you, and allows others, to understand your model
- allows much flexibility

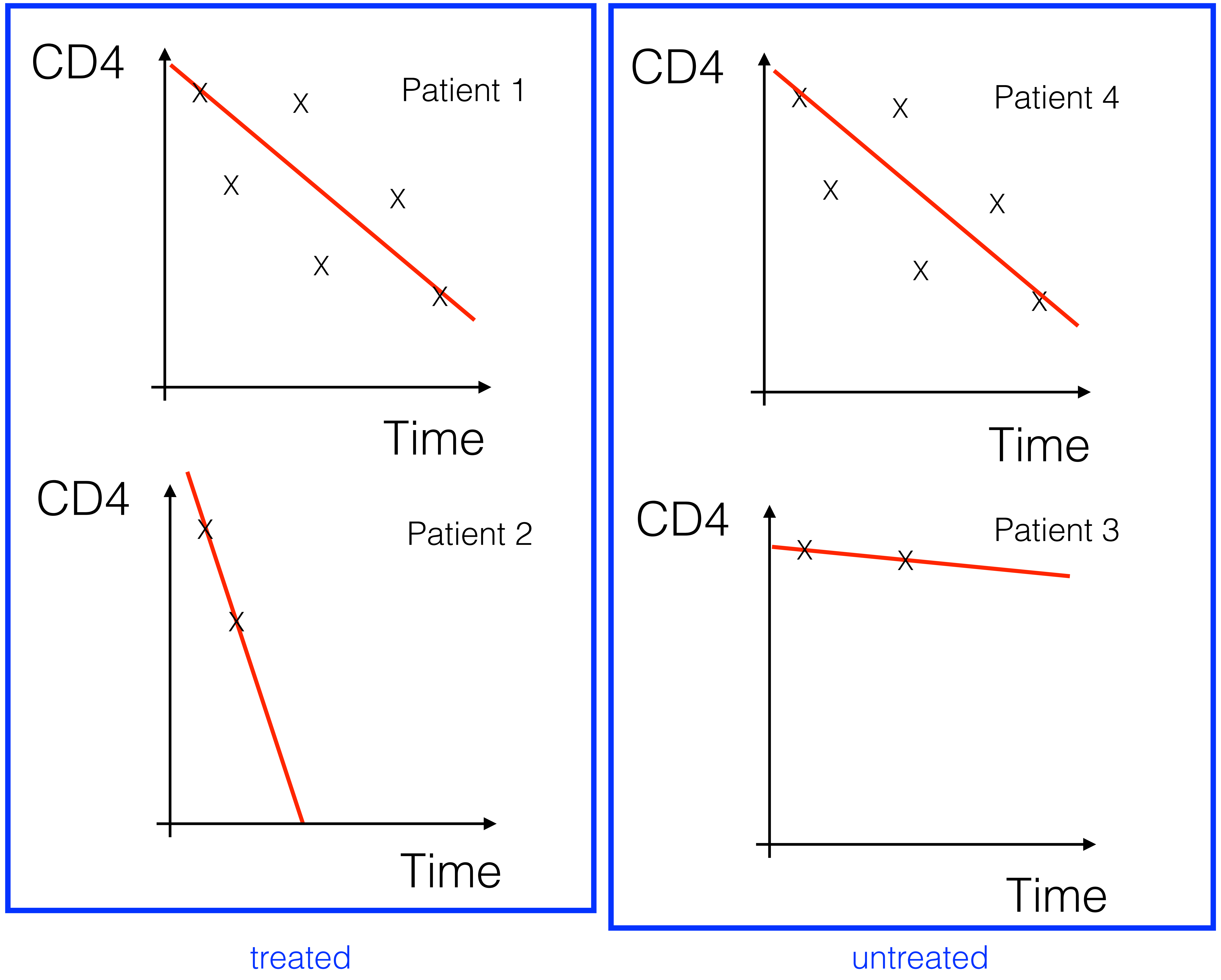
Partial pooling

Partial pooling means each group contributes to the likelihood in proportion to how informative it is, and information is shared between groups.

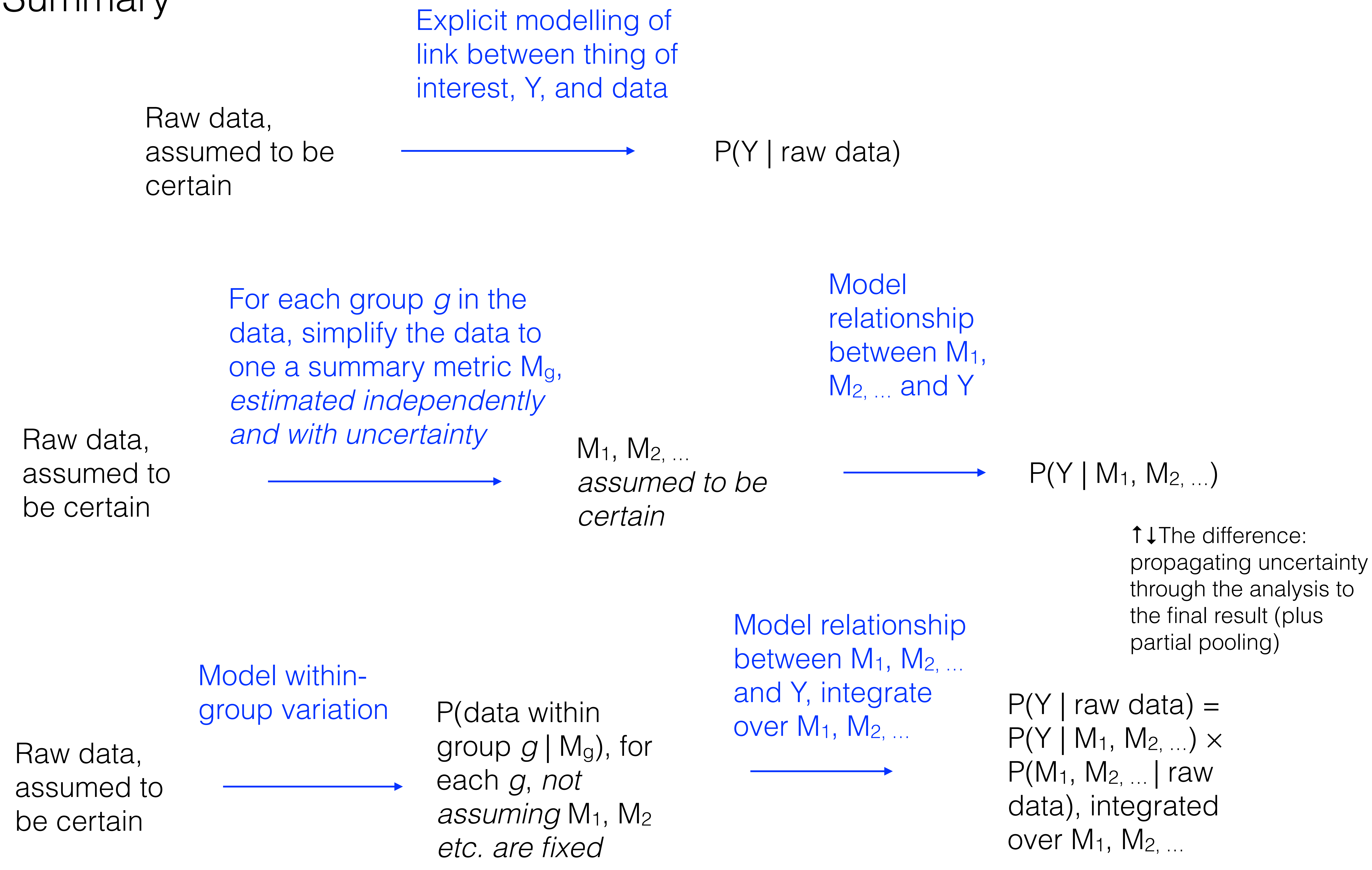
Schools 2 and 4 have outlier means, based on limited data. The variability we see within schools 1 and 3 means this could be observational noise.



Patients 2 and 4 have outlier gradients, based on limited data. The variability we see within schools 1 and 3 means this could be observational noise.



Summary



Raw data, assumed to be certain

Explicit modelling of link between thing of interest, Y, and data

P(Y | raw data)

Raw data, assumed to be certain

For each group *g* in the data, simplify the data to one a summary metric M_g, *estimated independently and with uncertainty*

M_1, M_2, … *assumed to be certain*

Model relationship between M_1, M_2, … and Y

P(Y | M_1, M_2, …)

↑↓The difference: propagating uncertainty through the analysis to the final result (plus partial pooling)

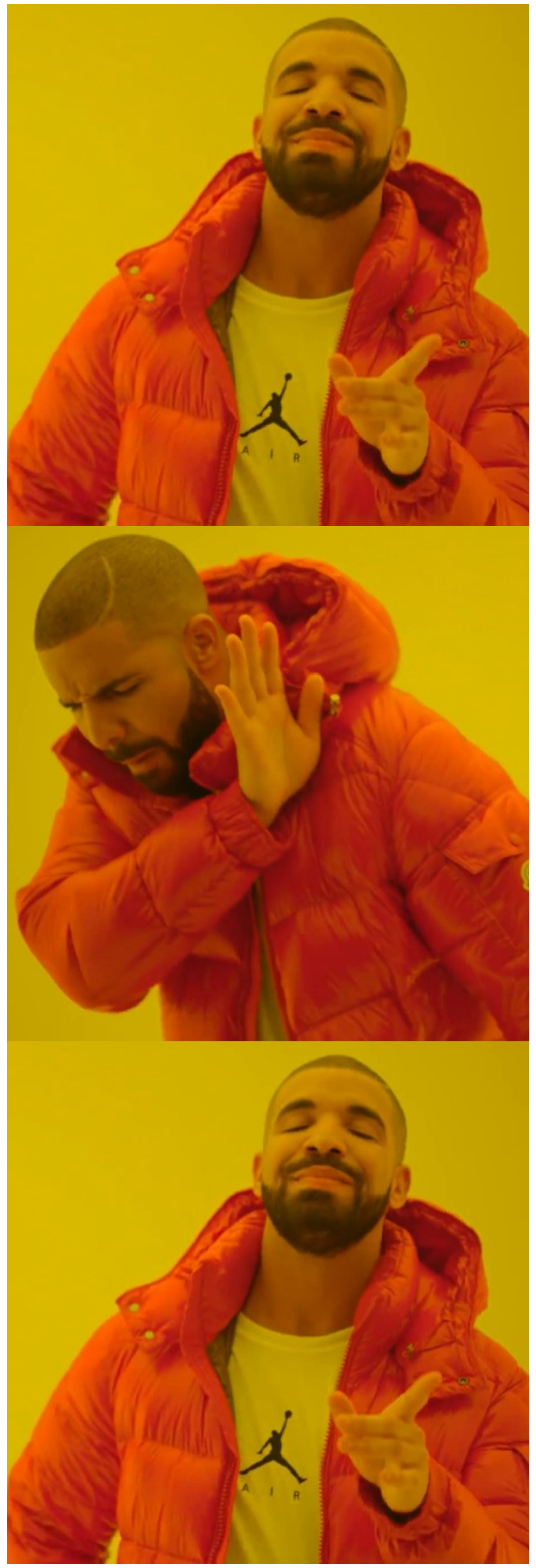
Raw data, assumed to be certain

Model within-group variation

P(data within group *g* | M_g), for each *g*, *not assuming* M_1, M_2 *etc. are fixed*

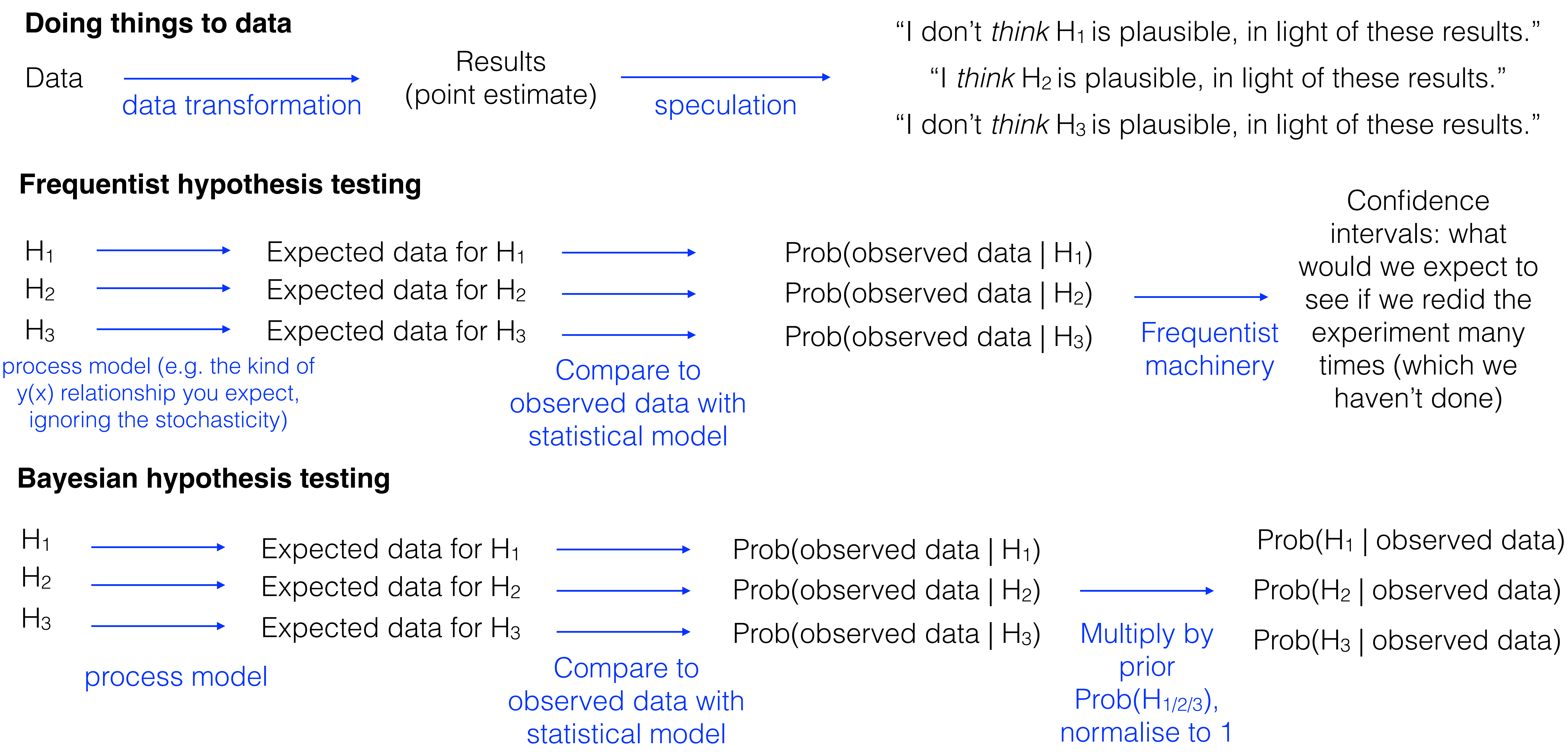
Model relationship between M_1, M_2, … and Y, integrate over M_1, M_2, …

P(Y | raw data) = P(Y | M_1, M_2, …) × P(M_1, M_2, … | raw data), integrated over M_1, M_2, …



P(Copyright by Drake) > 0

Throwback to my ill-informed, opinionated taxonomy of inference



Doing things to data

Data $\xrightarrow{\text{data transformation}}$ Results (point estimate) $\xrightarrow{\text{speculation}}$

"I don't *think* H_1 is plausible, in light of these results."

"I *think* H_2 is plausible, in light of these results."

"I don't *think* H_3 is plausible, in light of these results."

Frequentist hypothesis testing

H_1 → Expected data for H_1 → Prob(observed data | H_1)

H_2 → Expected data for H_2 → Prob(observed data | H_2)

H_3 → Expected data for H_3 → Prob(observed data | H_3)

process model (e.g. the kind of y(x) relationship you expect, ignoring the stochasticity)

Compare to observed data with statistical model

Frequentist machinery → Confidence intervals: what would we expect to see if we redid the experiment many times (which we haven't done)

Bayesian hypothesis testing

H_1 → Expected data for H_1 → Prob(observed data | H_1)

H_2 → Expected data for H_2 → Prob(observed data | H_2)

H_3 → Expected data for H_3 → Prob(observed data | H_3)

process model

Compare to observed data with statistical model

Multiply by prior Prob($H_{1/2/3}$), normalise to 1 →

Prob(H_1 | observed data)

Prob(H_2 | observed data)

Prob(H_3 | observed data)

Ask not what you can do to your data, but how you can turn hypotheses into data i.e. what is the data-generating process.

Calculating a single average to summarise each group is *doing things to data.*

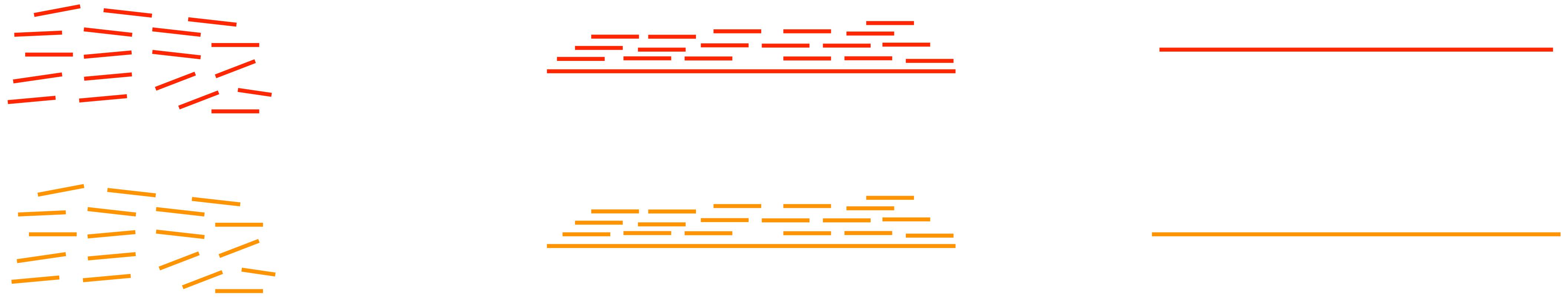
JFKlibrary.org

The exception to that: pragmatism

Online reader: we're talking about bioinformatic and phylodynamic analysis of fragments of genetic sequence data here.

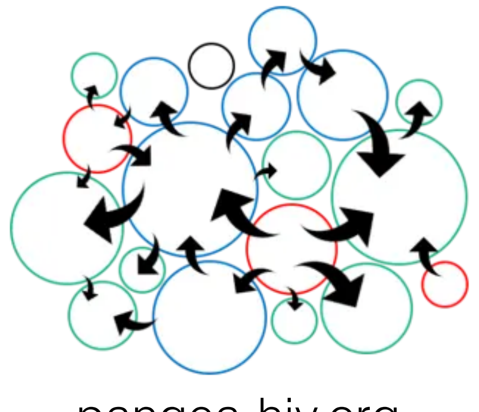
Doing things to data, sigh

many fastq files → mapping etc. → many bam files → call most common base → many consensus sequences



Phylodynamic modelling, integrating over evolutionary histories compatible with a given hypothesis and the consensuses *assumed to be certain*

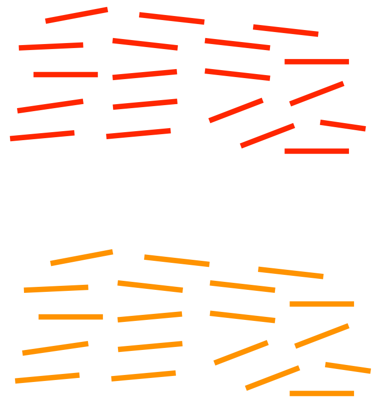
hypotheses of interest



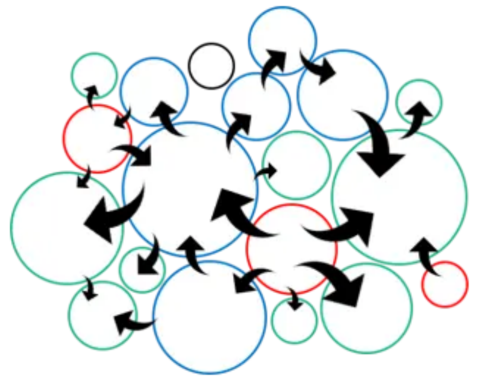
pangea-hiv.org

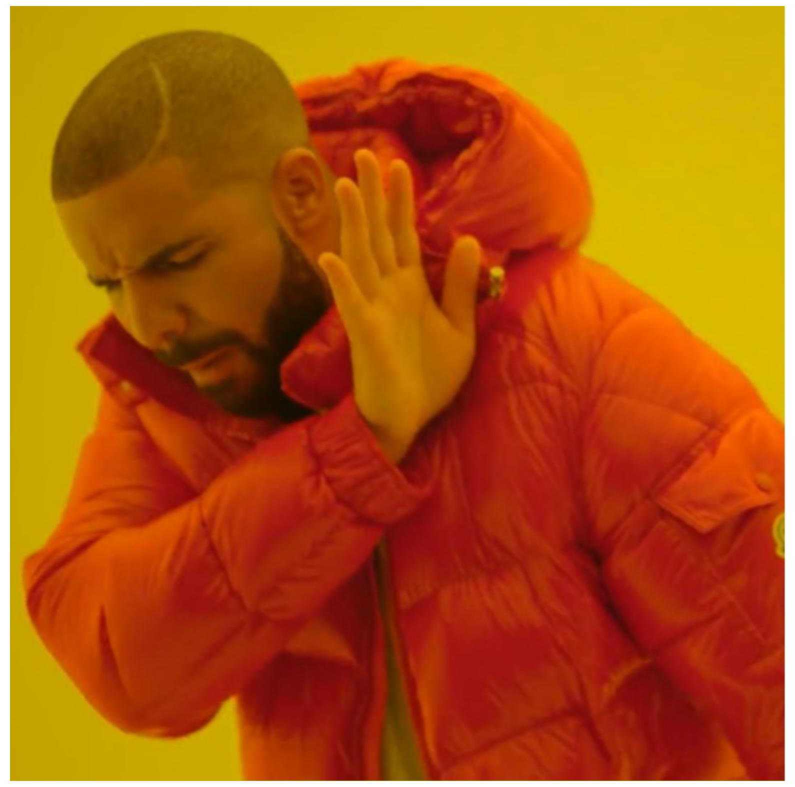
Phylodynamic modelling as before, also integrating over all possible within-host evolution processes (i.e. quasispecies space) and sequence amplificiation processes

many fastq files



hypotheses of interest





Some reading on hierarchical models

Lecture 6 of https://ben-lambert.com/bayesian-lecture-slides/

Chapter 13 of *Statistical Rethinking* textbook, on my desk

Chapter 5 of *Bayesian Data Analysis* textbook, on my desk
and free online http://www.stat.columbia.edu/~gelman/book/

Pro tip: non-centred parameterisations might increase the efficiency of the MCMC

```
real school_effects[num_schools];
school_effects ~ normal(0, stddev_schools);
```

→

```
real school_effects_standardised[num_schools];
school_effects_standardised ~ normal(0, 1);
school_effects = school_effects_standardised *
                 stddev_schools;
```

Different parameterisations of the same model, i.e. identical mathematically.
But the right-hand version makes the two parameters that the MCMC explores more independent of each other, making the posterior geometry easier to explore.