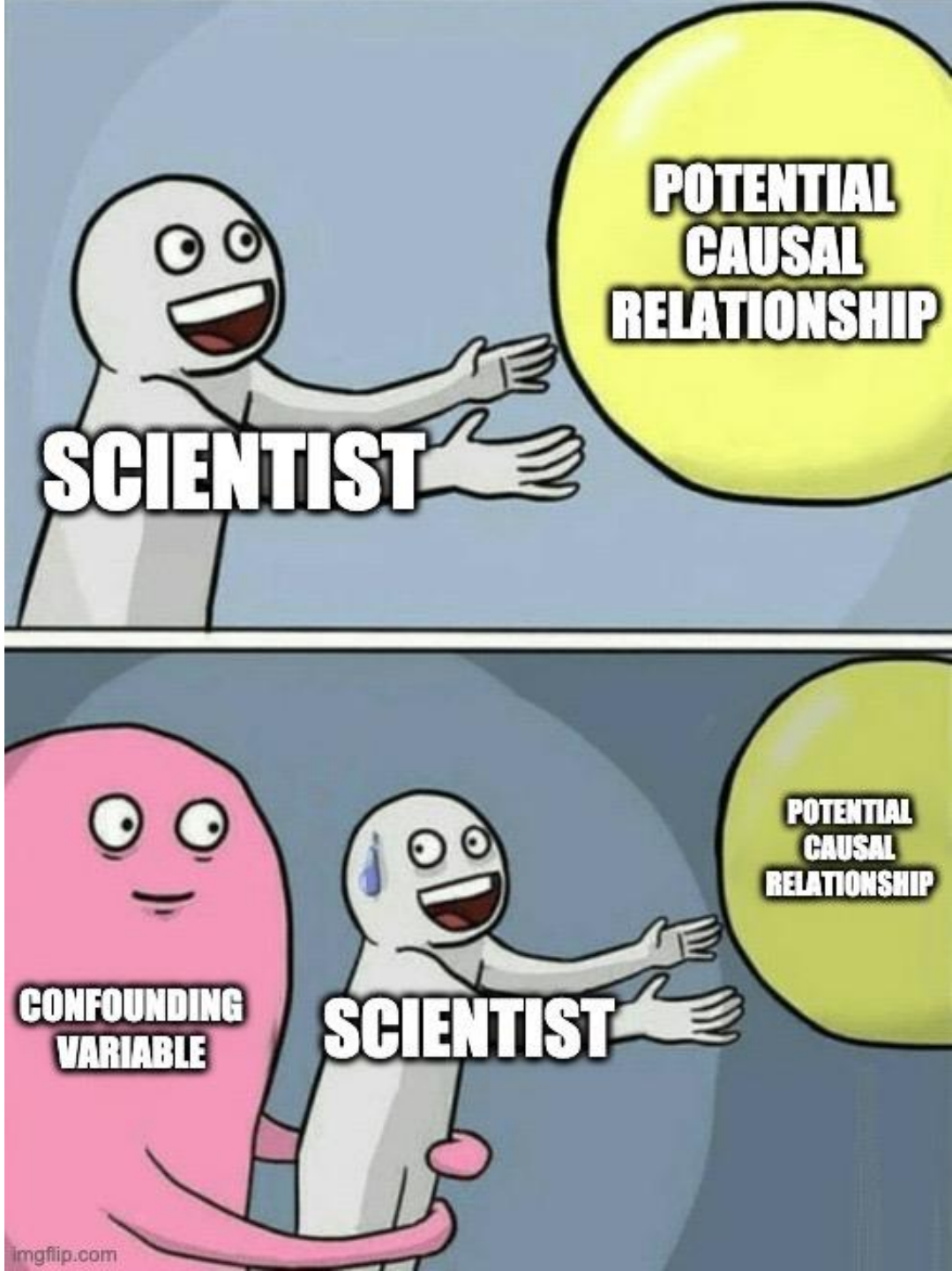
Correlated data, nuisance variables

ENTMLGY 6707
Entomological Techniques and Data Analysis

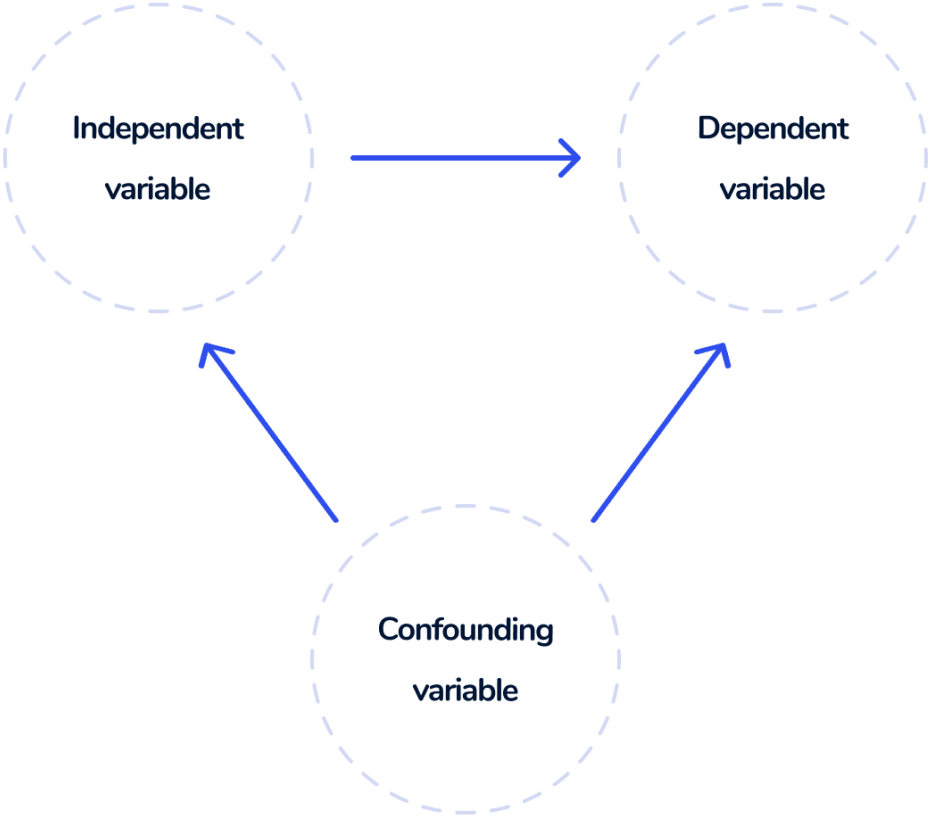


Learning objectives

1. Define, identify, and account for nuisance variables
2. Recognize potential correlations in (the residuals of) response variables
3. Compare and contrast fixed and random effects

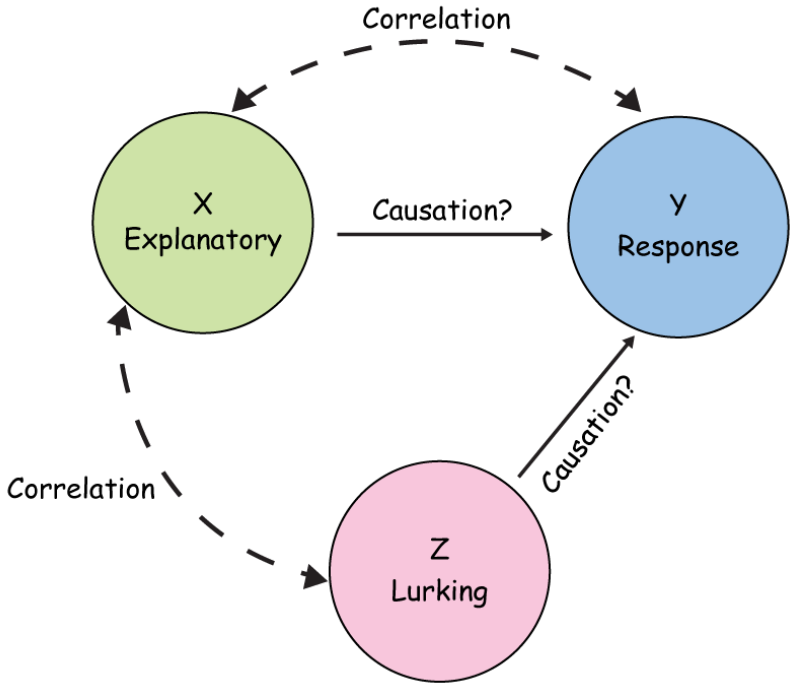
Definition

Confounding variable: influences response and predictor, potentially leading to spurious associations. They obscure…or confound…the relationship between response and predictors.



Definition

Nuisance variable: influences (= explains variation in) response variable but not directly relevant to the hypothesis/research question; can be difficult to remove from the experiment. If the NV is unknown (or difficult to measure), we hope randomization will account for its effects. Otherwise, we use blocking and/or fit the NV as a predictor.



Calcworkshop.com

Time can be a confounding variable…

Beetle survival on diet A and diet B.

Rep A	Rep B
Obs 1-4	Obs 9-12
Obs 5-8	Obs 13-16

Time

…so, change the design…

Beetle survival on diet A and diet B.

Rep A	Rep B
Obs 1-4	Obs 9-12
Obs 5-8	Obs 13-16

Time

BUT…now time is a nuisance variable

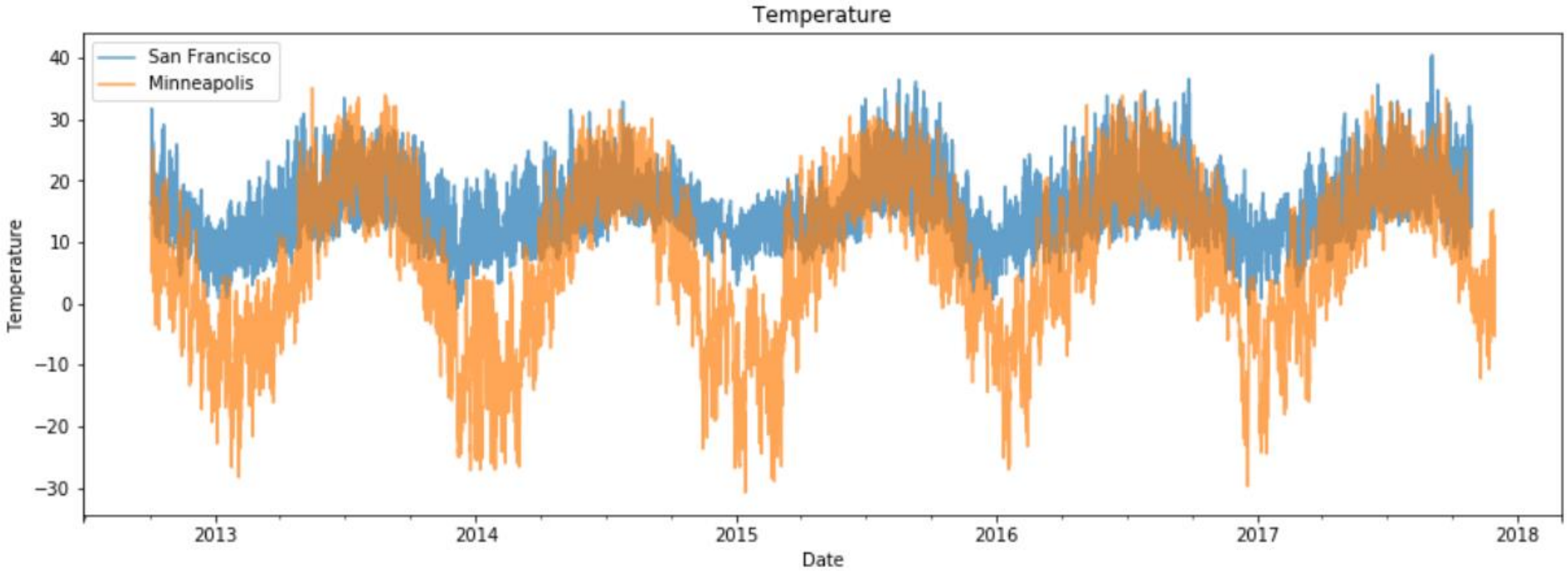
Beetle survival on diet A and diet B.

Rep A	Rep B
Obs 1-4	Obs 9-12
Obs 5-8	Obs 13-16

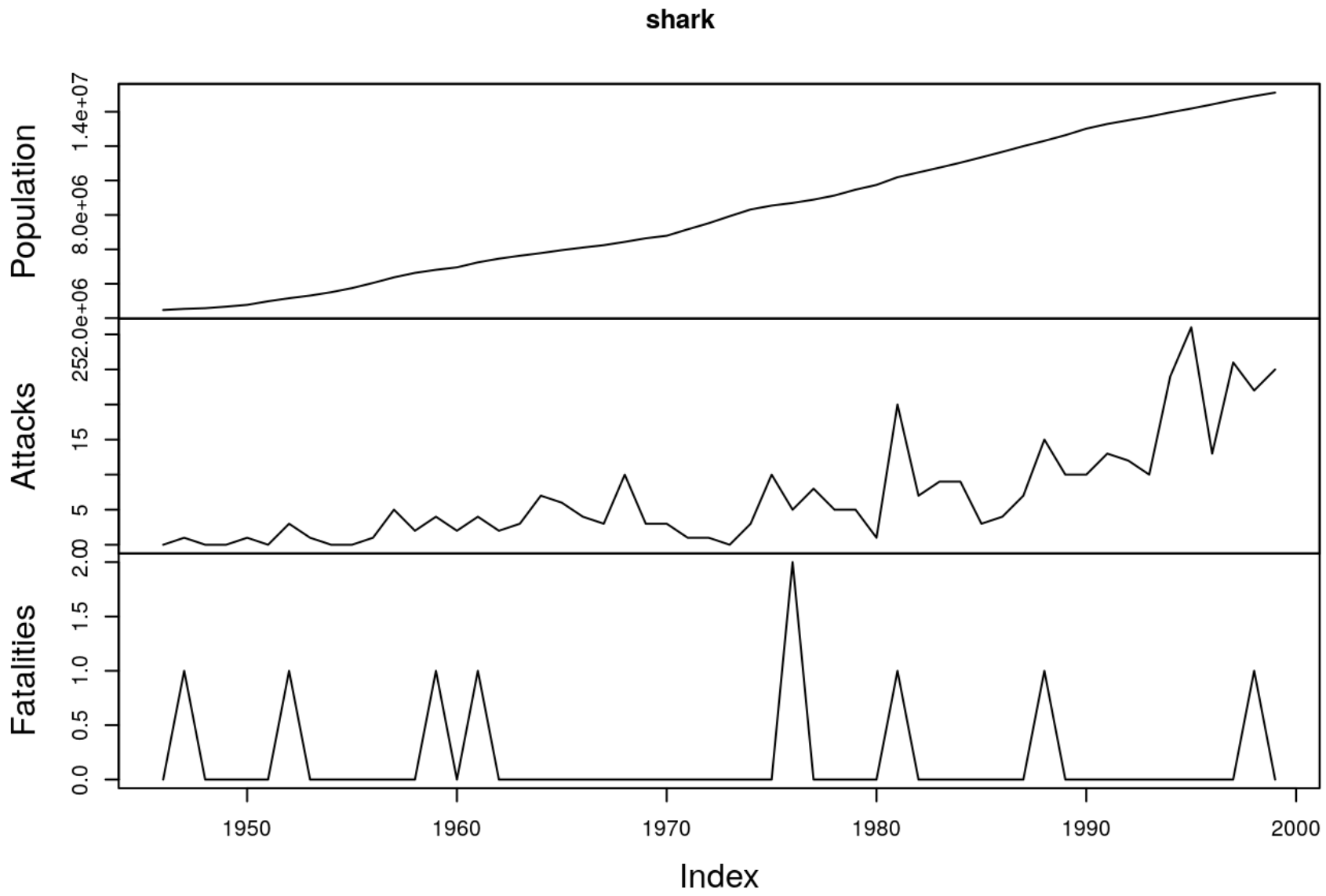
Time

Activity (don't submit):
Identify one confounding variable and/or one nuisance variable that could *potentially* drive variation in a response variable in your study system.

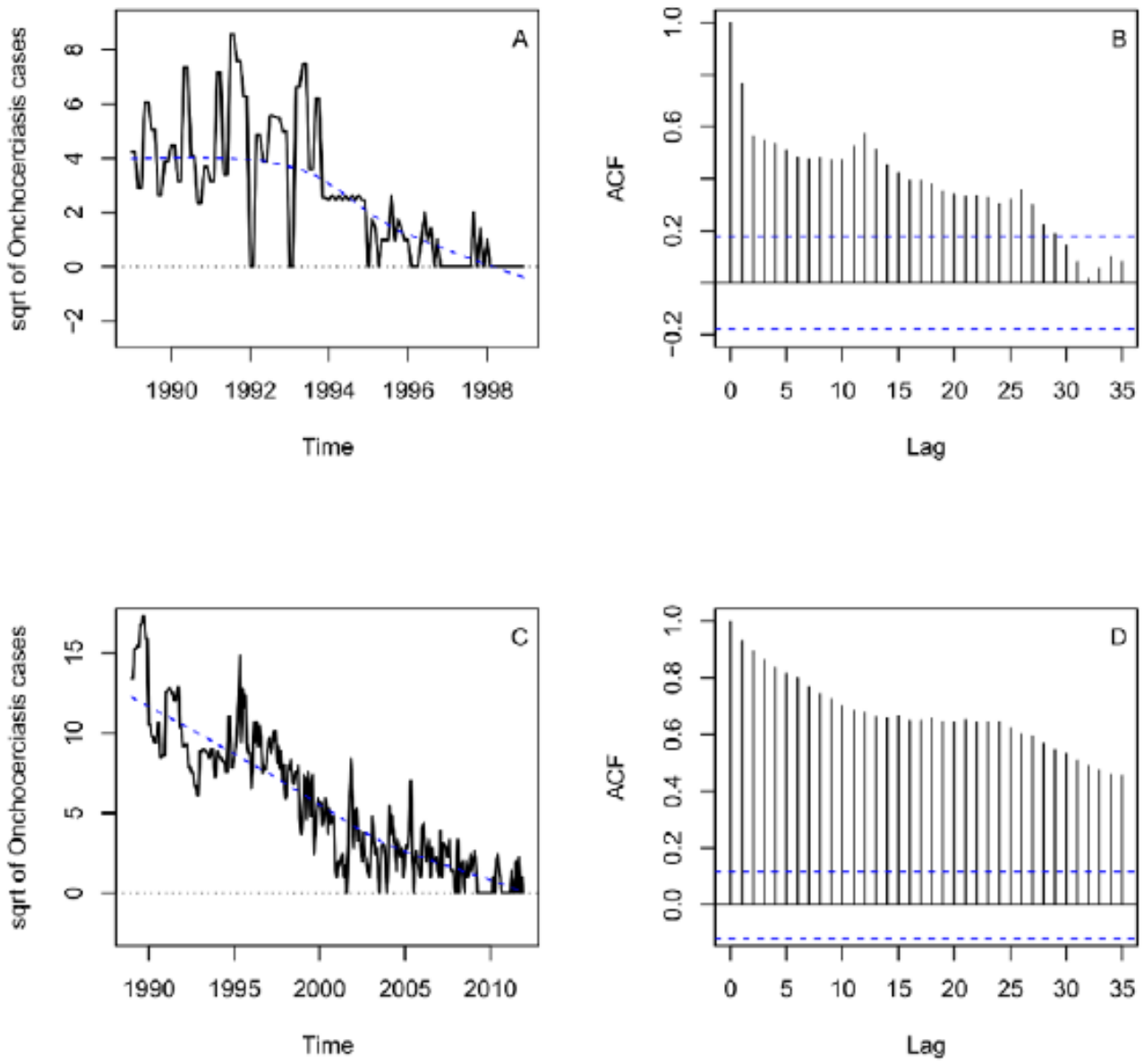
A few slides/comments on spatial and temporal analyses…



María García Gumbao
Product Data Analyst at Glovo



Shark Attacks in Florida
http://www.seec.uct.ac.za/time-series-analysis



OPEN ACCESS Freely available online

PLOS | NEGLECTED TROPICAL DISEASES

Time Series Analysis of Onchocerciasis Data from Mexico: A Trend towards Elimination

Edgar E. Lara-Ramírez[1], Mario A. Rodríguez-Pérez[1], Miguel A. Pérez-Rodríguez[1], Monsuru A. Adeleke[2], María E. Orozco-Algarra[3], Juan I. Arrendondo-Jiménez[3], Xianwu Guo[1]*

1 Centro de Biotecnología Genómica, Instituto Politécnico Nacional, Reynosa, Tamaulipas, México, **2** Public Health Entomology and Parasitology Unit, Department of Biological Sciences, Osun State University, Osogbo, Osun, Nigeria, **3** Centro Nacional de Vigilancia Epidemiológica y Control de Enfermedades, Secretaria de Salud, México Distrito Federal, México

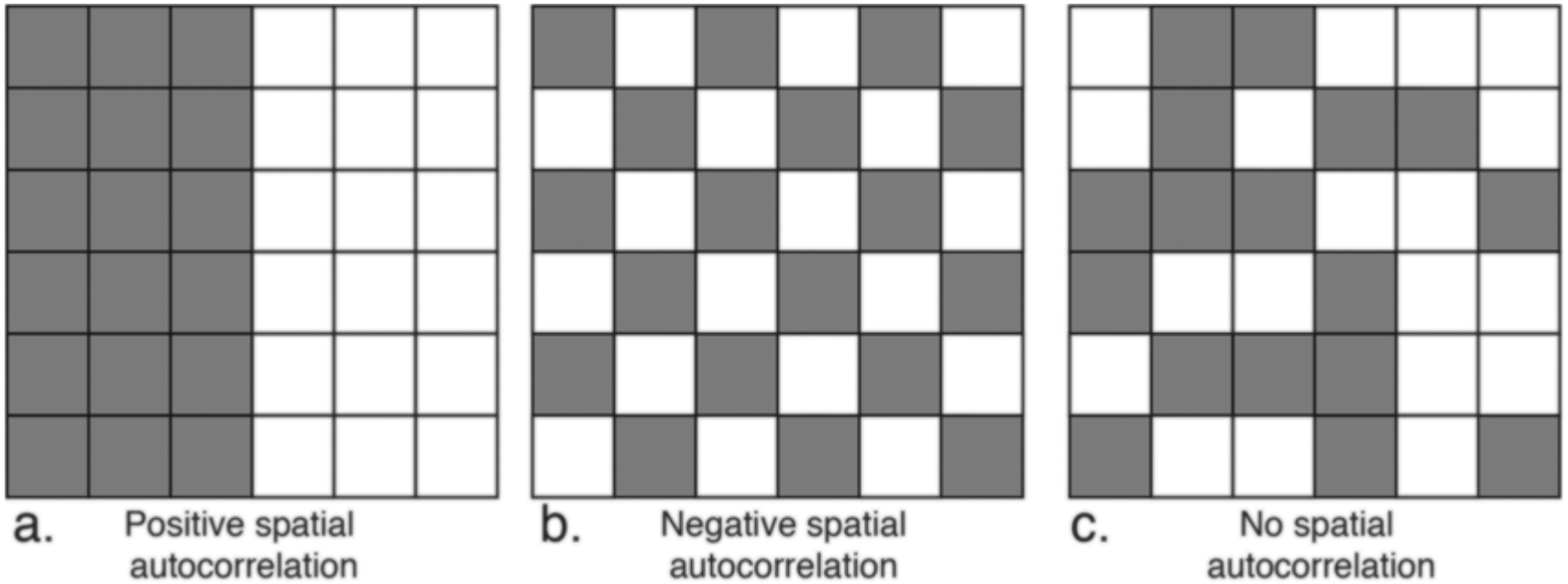
Year	Cases	Cases_nolag	Cases_lag1	Cases_lag2	Cases_lag3
1991	15	15	NA	NA	NA
1992	18	18	15	NA	NA
1993	10	10	18	15	NA
1994	8	8	10	18	15
1995	6	6	8	10	18
1996	5	5	6	8	10
1997	4	4	5	6	8

Space as a nuisance variable

Beetle survival on cultivar A and cultivar B.

Site 1	Site 2
Obs 1-4	Obs 9-12
Obs 5-8	Obs 13-16

Spatial gradient

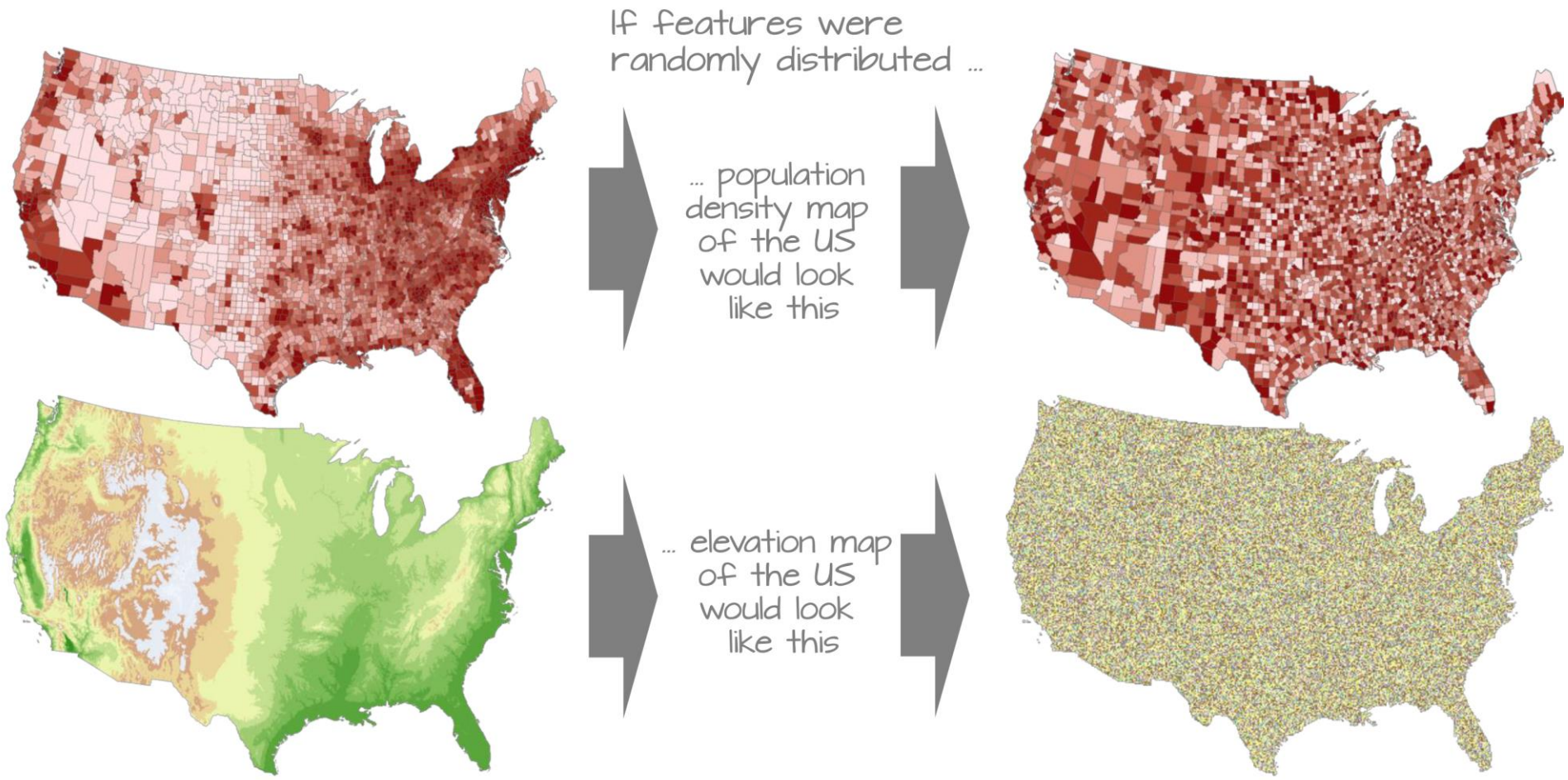


Microbial Ecology (2021) 81:874–883
https://doi.org/10.1007/s00248-020-01608-4

ENVIRONMENTAL MICROBIOLOGY

Scale-Dependent Influences of Distance and Vegetation on the Composition of Aboveground and Belowground Tropical Fungal Communities

André Boraks[1] · Gregory M. Plunkett[2] · Thomas Morris Doro[3] · Frazer Alo[3] · Chanel Sam[3] · Marika Tuiwawa[4] · Tamara Ticktin[1] · Anthony S. Amend[1]



https://mgimond.github.io/Spatial/spatial-autocorrelation.html

Accounting for nuisance variables

Site 1	Site 2
Obs 1-4	Obs 9-12
Obs 5-8	Obs 13-16

Option 1 (fixed-effects only model):
fit1 <- lm(y ~ site + trt, data= df)

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i$$

Site effect

Treatment effect

Option 2 (mixed-effects models):
library(lme4)
fit1 <- lmer(y ~ trt + (1|site), data=df)

$$Y_i = \beta_0 + b_i + \beta_1 X_i + \varepsilon_i$$

Site "effect"

Treatment effect

The modeling frameworks differ conceptually

Option 1 (fixed-effects only model):

```
fit1 <- lm(y ~ site + trt, data= df)
```

"Fitting 'group' as a fixed effect in model M1 assumes the 'group' means are all independent of one another and share a common residual variance."

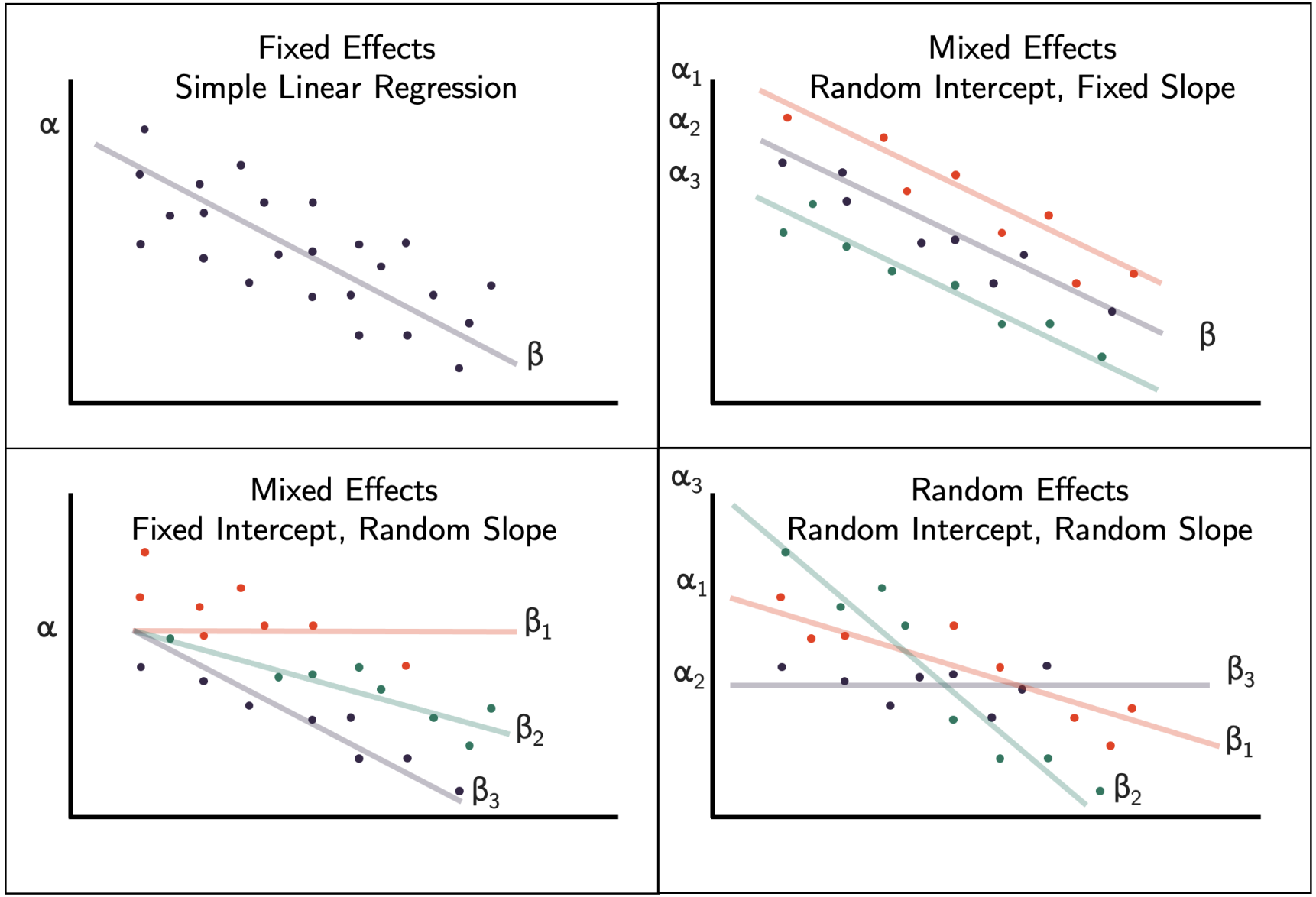
Option 2 (mixed-effects models):

```
library(lme4)
fit1 <- lmer(y ~ trt + (1|site), data=df)
```

"Conversely, fitting group as a random intercept model assumes that the measured group means are only a subset of the realised possibilities drawn from a 'global' set of population means that follow a Normal distribution with its own mean and variance."

Harrison et al. (2018), A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ6:e4794; DOI 10.7717/peerj.4794

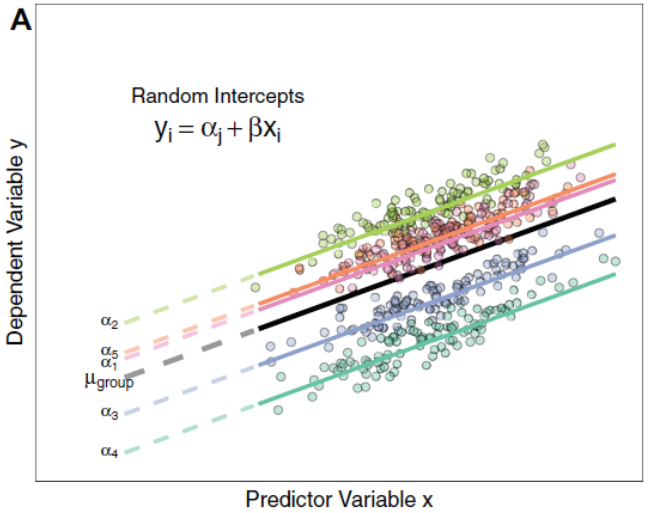
Random effects



Random effects

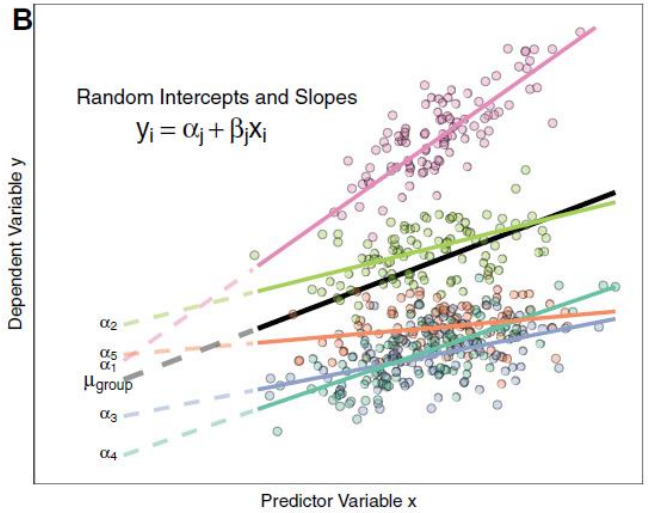
Random intercept

lmer(y ~ trt+ (1|site), data=df)



Random intercept and slope

lmer(y ~ trt + (trt|site), data=df)

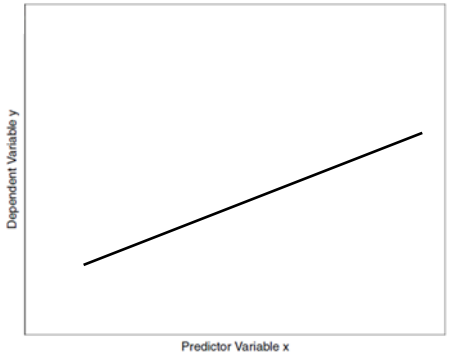


You will sometimes see "random factors" mentioned in ANOVA, but these take on a different meaning/interpretation (http://www.stat.columbia.edu/~gelman/research/published/banova7.pdf)

Note that "trt" above could be continuous or categorical – mixed-effects models are extremely flexible.

To reduce confusion, some folks suggest referring to random intercepts and slopes as variable intercepts and variable slopes, respectively.

Fixed effects only



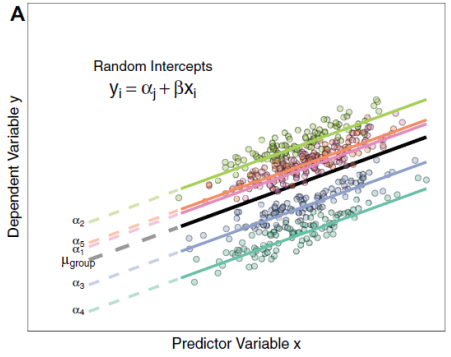

```
> fit0 <- lm(circumference~age, data=Orange)
> summary(fit0)

Call:
lm(formula = circumference ~ age, data = Orange)

Residuals:
    Min      1Q  Median      3Q     Max 
-46.310 -14.946  -0.076  19.697  45.111 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 17.399650   8.622660   2.018   0.0518 .  
age          0.106770   0.008277  12.900 1.93e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.74 on 33 degrees of freedom
Multiple R-squared:  0.8345,	Adjusted R-squared:  0.8295 
F-statistic: 166.4 on 1 and 33 DF,  p-value: 1.931e-14
```

library(lmerTest)

The lme4 package will NOT return *p*-values by default, apparently because there is disagreement on the best way to calculate degrees of freedom when fitting random effects.

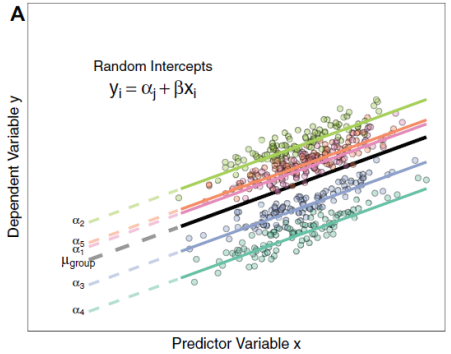
The lmerTest package enables use of Satterthwaite's approximation (default) and other methods, and thus will result in *p*-values being displayed for lme4 models.

Random intercept



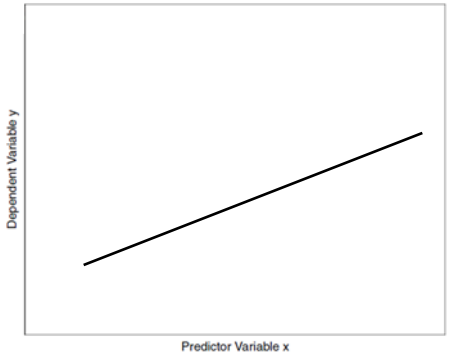

```
> library(lme4)
> library(lmerTest)
> fit1 <- lmer(circumference~age+(1|Tree), data=Orange)
> summary(fit1)
Linear mixed model fit by REML. t-tests use
  Satterthwaite's method [lmerModLmerTest]
Formula: circumference ~ age + (1 | Tree)
   Data: Orange

REML criterion at convergence: 303.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.8781 -0.6743  0.2320  0.5053  1.5416

Random effects:
 Groups   Name        Variance Std.Dev.
 Tree     (Intercept) 389.6    19.74
 Residual             232.9    15.26
Number of obs: 35, groups:  Tree, 5
```

Random intercept



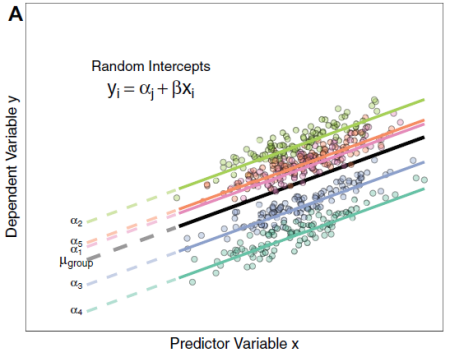

```
Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept) 17.399650  10.423696  6.528443   1.669    0.142
age          0.106770   0.005321 29.000000  20.066   <2e-16

(Intercept)
age         ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
    (Intr)
age -0.471
```

Fixed effects only



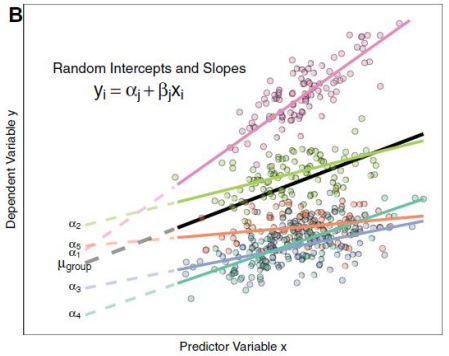

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.399650   8.622660   2.018   0.0518 .
age          0.106770   0.008277  12.900 1.93e-14 ***
```

Random intercept



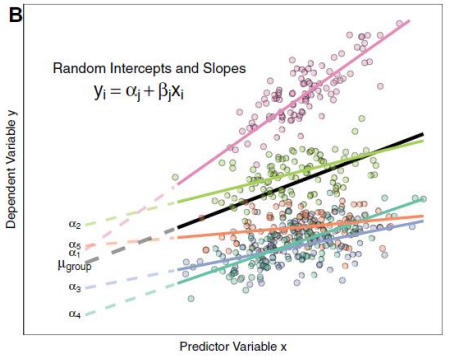

```
Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept) 17.399650  10.423696  6.528443   1.669    0.142
age          0.106770   0.005321 29.000000  20.066   <2e-16
```

Random intercept and slope




```
> fit2 <- lmer(circumference~age+(age|Tree), data=Orange)
Warning messages:
1: In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  unable to evaluate scaled gradient
2: In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
3: Model failed to converge with 1 negative eigenvalue: -1.5e+05
> summary(fit2)
Linear mixed model fit by REML. t-tests use
  Satterthwaite's method [lmerModLmerTest]
Formula: circumference ~ age + (age | Tree)
   Data: Orange

REML criterion at convergence: 281.1

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.09099 -0.50176 -0.07625  0.71181  1.63662
```

Random intercept and slope




```
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 Tree     (Intercept) 8.312e+00  2.88310
          age         5.083e-04  0.02255 0.99
 Residual             1.016e+02 10.07726
Number of obs: 35, groups:  Tree, 5

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept) 17.39965    3.88098  7.11618   4.483  0.00274 **
age          0.10677    0.01068  3.38479   9.999  0.00125 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
    (Intr)
age 0.037
optimizer (nloptwrap) convergence code: 0 (OK)
unable to evaluate scaled gradient
Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

Simpson's Paradox

When a trend appears in a dataset comprised of several groups, but then disappears or reverses when the grouping structure is accounted for (slightly paraphrased from Wikipedia).

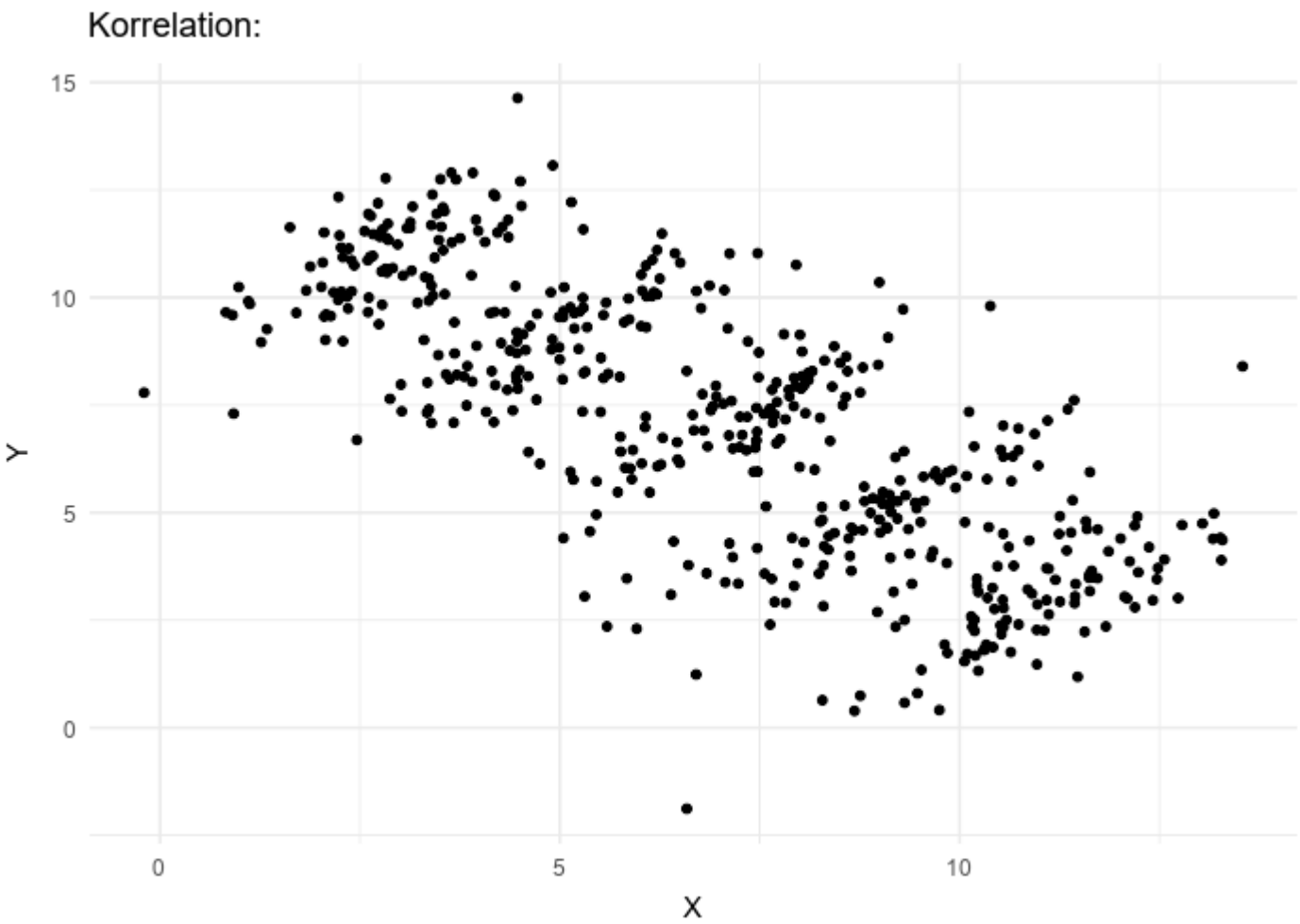


Figure by Pace~svwiki - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=62007681

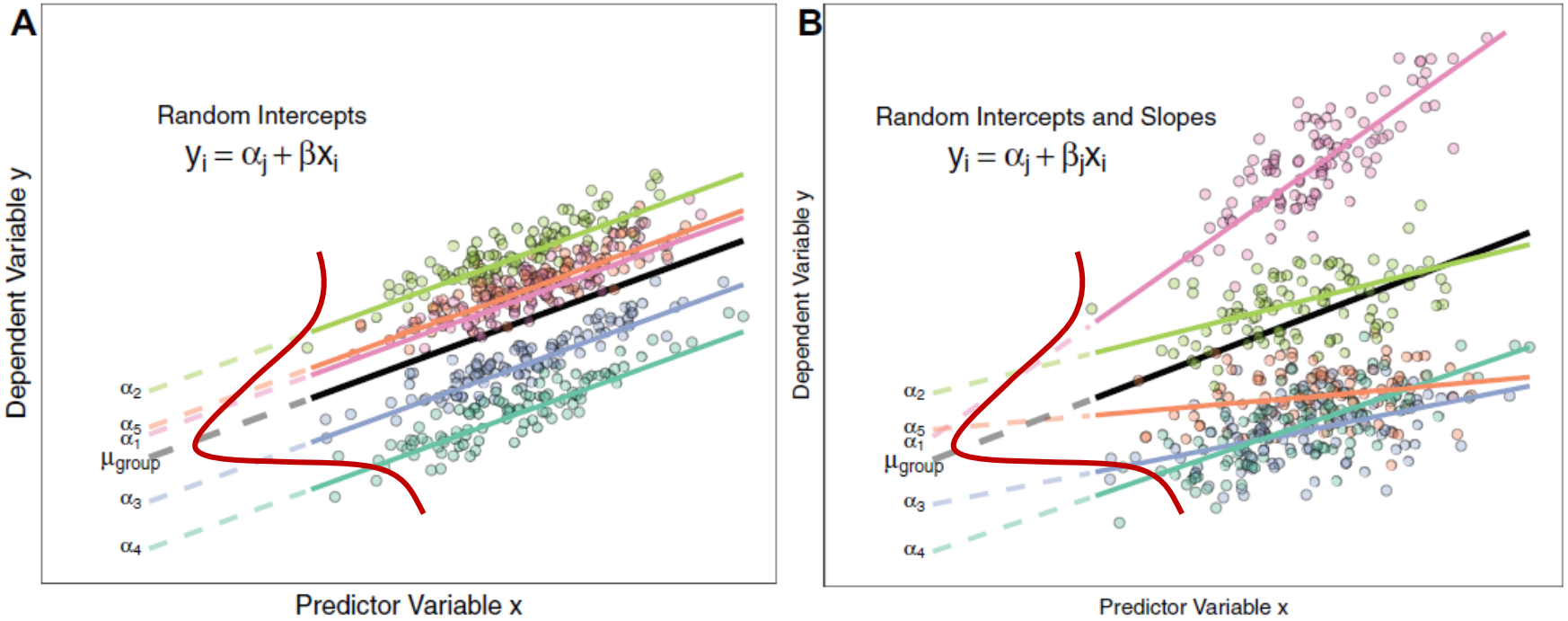
Assumptions

Linear mixed-effects models are linear models. So, just like simple linear regression, we are trying to fit a line (straight or curvy) through a cloud of points…and thus the same set of assumptions apply:

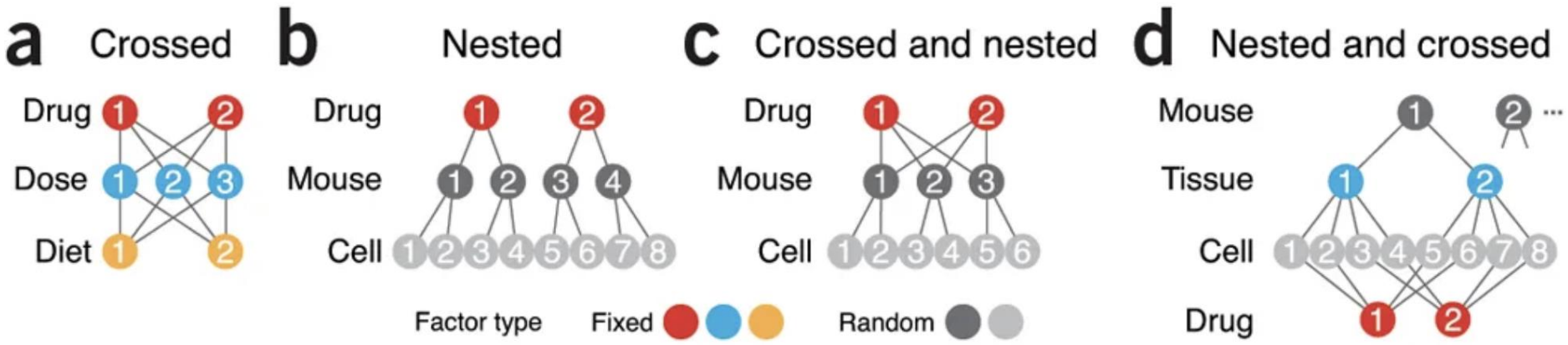
1. The relationship between response and predictor is linear
2. Residuals are independent
3. Residuals are normally distributed
4. Residuals are homoscedastic (i.e., the variance in Y does not increase or decrease as X increases or decreases)

We also make the assumptions that our random effects are normally distributed.

Assumptions on random effects



Nested vs. crossed random effects



There are amendments to this paper

THIS MONTH

POINTS OF SIGNIFICANCE

Nested designs

For studies with hierarchical noise sources, use a nested analysis of variance approach.

Many studies are affected by random-noise sources that naturally fall into a hierarchy, such as the biological variation among animals, tissues and cells, or technical variation such as measurement error. With a nested approach, the variation introduced at each hierarchy layer is assessed relative to the layer below it. We can use the relative noise contribution of each layer to optimally allocate experimental resources using nested analysis of variance (ANOVA), which gener-

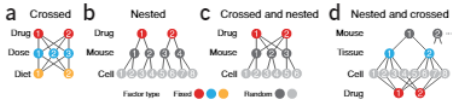


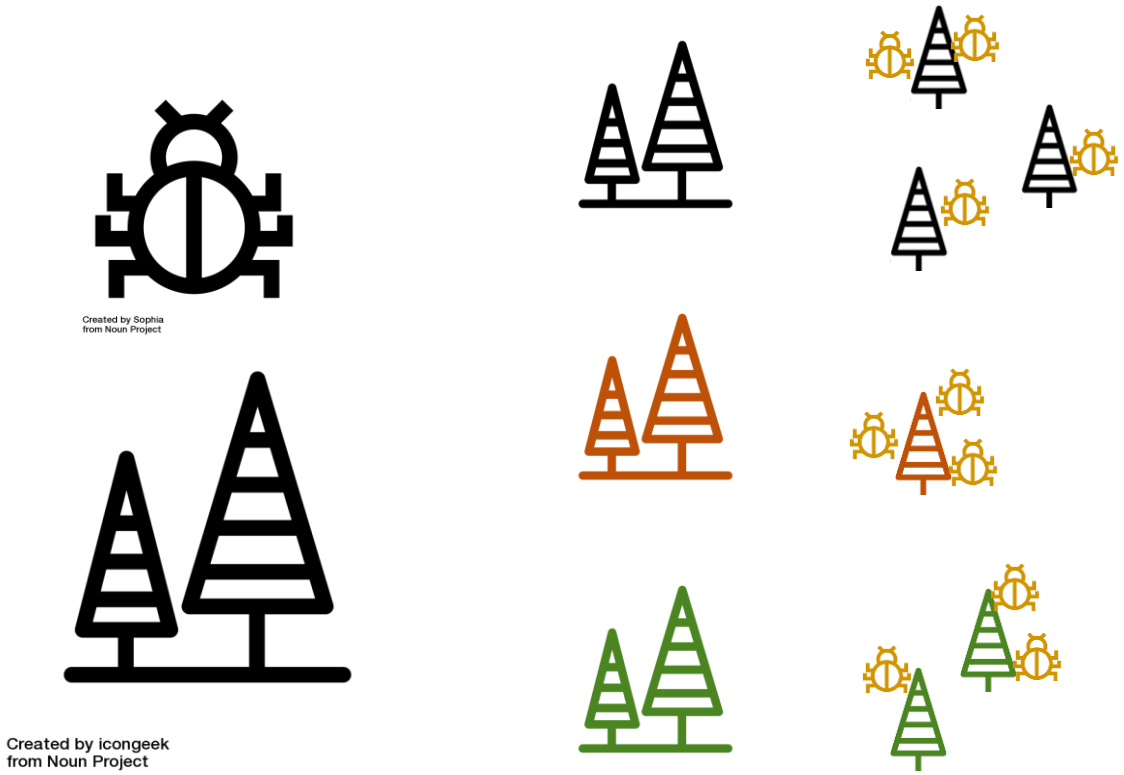
Figure 2 | Factors may be crossed or nested. (**a**) A crossed design examines every combination of levels for each fixed factor. (**b**) Nested design can progressively subreplicate a fixed factor with nested levels of a random factor that are unique to the level within which they are nested. (**c**) If a random factor can be reused for different levels of the treatment, it can be crossed with the treatment and modeled as a block. (**d**) A split plot design in which the fixed effects (tissue, drug) are crossed (each combination of tissue and drug are tested) but themselves nested within replicates.

factor on all mice and cells. If mice can be reused, we can cross them with the drug and use them as a random blocking factor[2] (**Fig. 2c**).

Martin Krzywinski, Naomi Altman & Paul Blainey

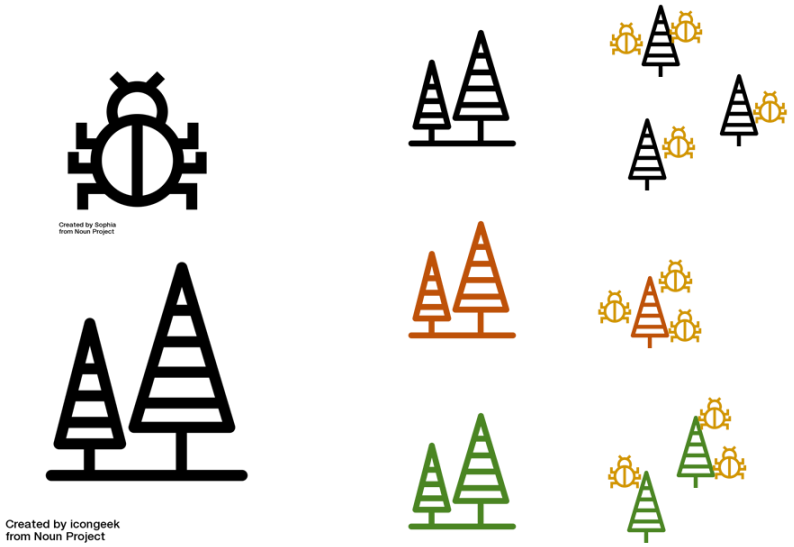
1. Blainey, P., Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 879–880 (2014).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 699–700 (2014).
3. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 597–598 (2014).

Nested random effects

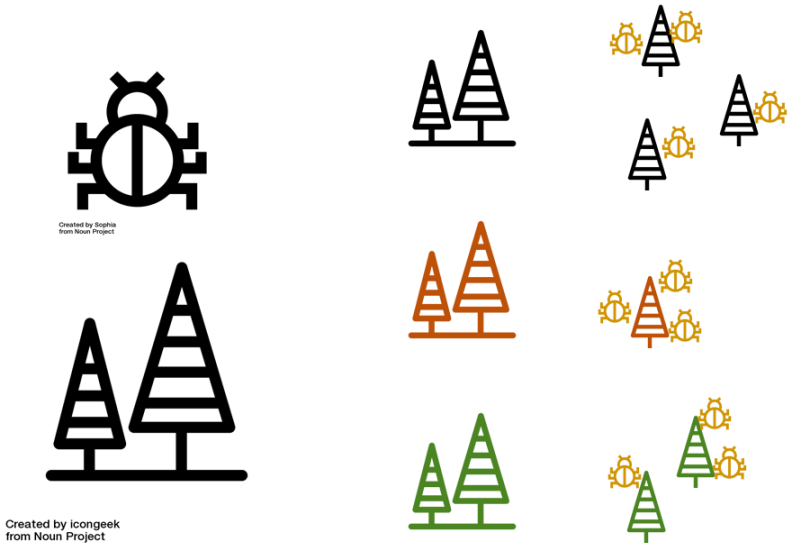


fit1 <- lmer(insect size ~ treatment + (1|site/tree/branch), data=df)

Nested and crossed random effects



Year 1



Year 2

fit1 <- lmer(insect size ~ treatment + (1|site/tree/branch) + **(1|year)**, data=df)

Word of caution on data management

Let's pretend we are modeling insect size at the branch level

site	tree	unique_ID
1	1	1_1
1	2	1_2
2	1	2_1
2	2	2_2
3	1	3_1
3	2	3_2

site	tree	unique_ID
1	A	1_A
1	B	1_B
2	C	2_C
2	D	2_D
3	E	3_E
3	F	3_F

```
fit1 <- lmer(insect_size ~ foliage_monoterpenes + (1|site/tree), data=df)
```

≠

```
fit1 <- lmer(insect_size ~ foliage_monoterpenes + (1|site) + (1|tree), data=df)
```

Word of caution on data management

Let's pretend we are modeling insect size at the branch level

site	tree	unique_ID
1	1	1_1
1	2	1_2
2	1	2_1
2	2	2_2
3	1	3_1
3	2	3_2

site	tree	unique_ID
1	A	1_A
1	B	1_B
2	C	2_C
2	D	2_D
3	E	3_E
3	F	3_F

```
fit1 <- lmer(insect_size ~ foliage_monoterpenes + (1|site/tree), data=df)
```

=

```
fit1 <- lmer(insect_size ~ foliage_monoterpenes + (1|site) + (1|tree), data=df)
```

```
df$unique_ID <- paste(df$site, df$tree, sep="_")
```

site	tree	unique_ID
1	1	1_1
1	2	1_2
2	1	2_1
2	2	2_2
3	1	3_1
3	2	3_2