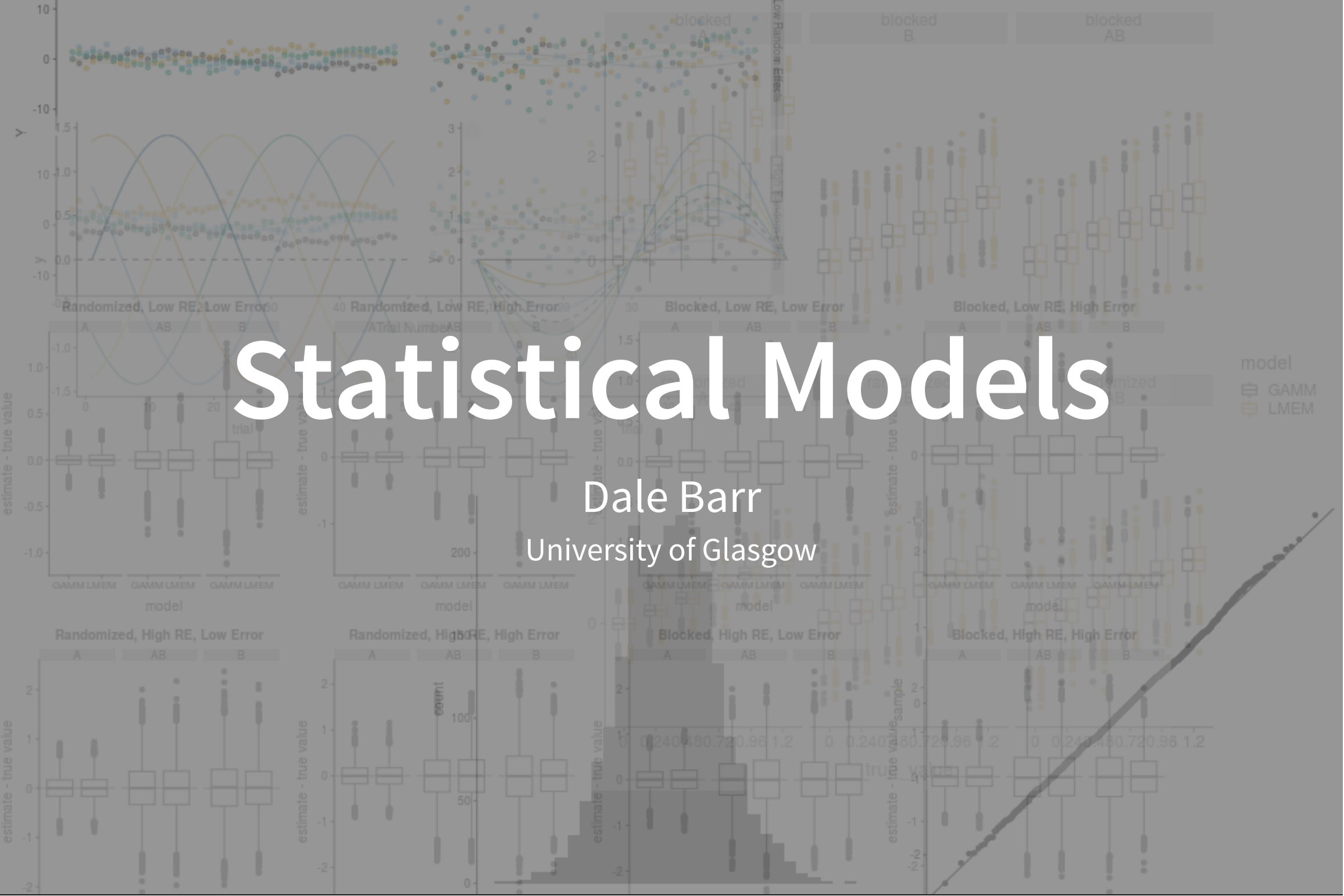
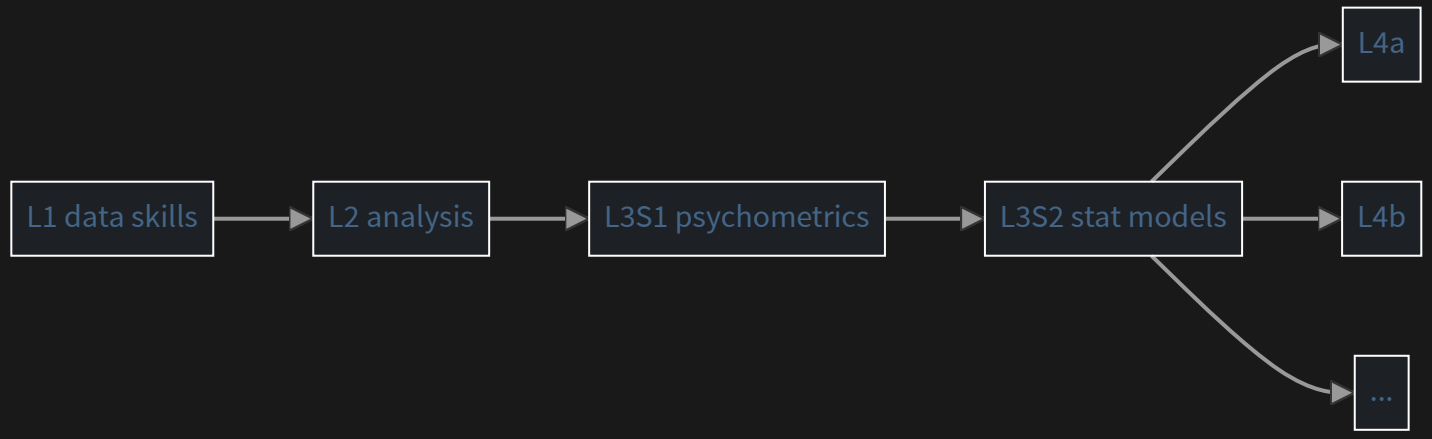


Statistical Models

Dale Barr

University of Glasgow

The whole picture



Stat models: A survey course

Focus is on *breadth* rather than *depth* …except correlation & regression, which provide the foundation for most advanced techniques

Aim: basic understanding of how to understand statistical models and basic skills for implementing them in R

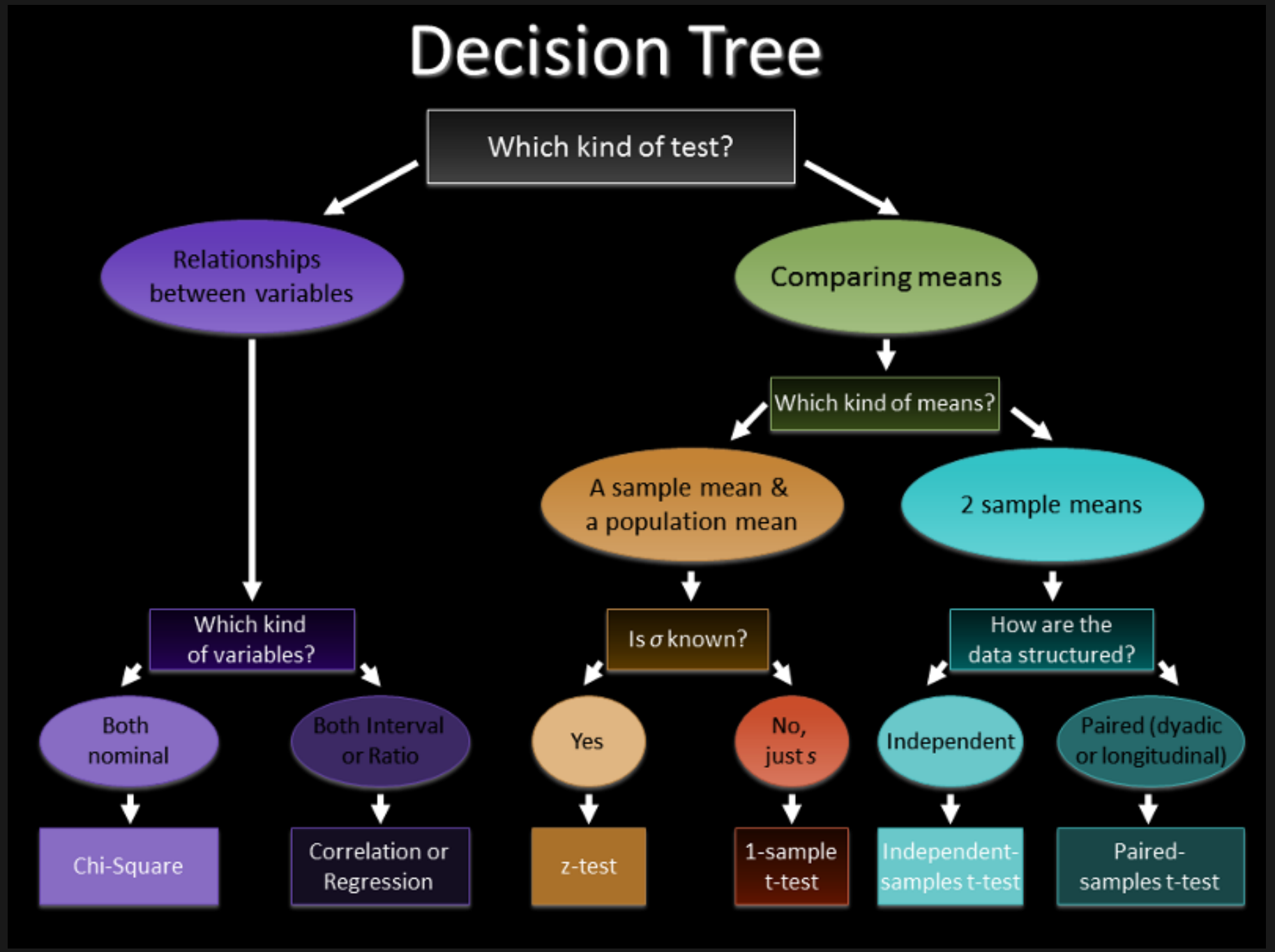
Why models?

Models are just… models

A statistical model is a *simplification* and *idealization* of reality that captures our key assumptions about how the data came about (the *data generating process* or DGP).

A statistical model is a "theory" about data.

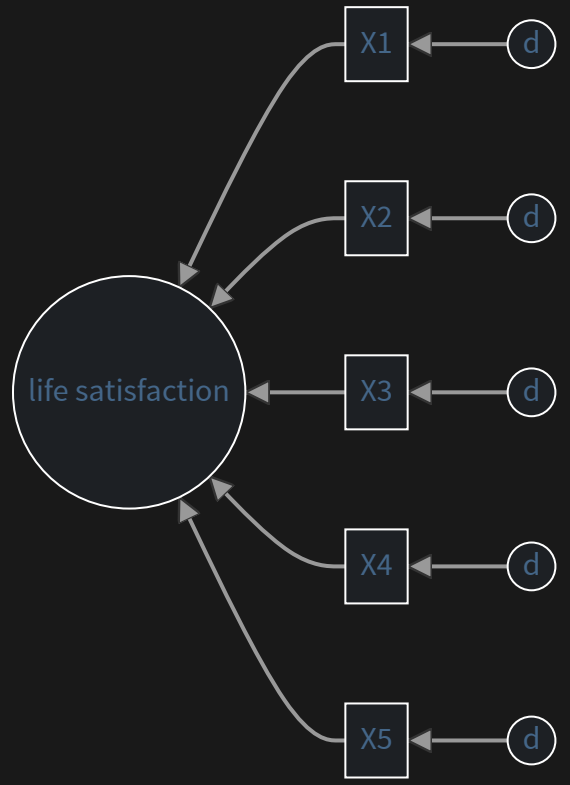
> *"All models are wrong, but some are useful."* - George Box


Decision Tree
Which kind of test?
Relationships between variables
Comparing means
Which kind of means?
A sample mean & a population mean
2 sample means
Which kind of variables?
Is σ known?
How are the data structured?
Both nominal
Both Interval or Ratio
Yes
No, just s
Independent
Paired (dyadic or longitudinal)
Chi-Square
Correlation or Regression
z-test
1-sample t-test
Independent-samples t-test
Paired-samples t-test

Models make things concrete

Models improve measurement

observation = truth + error



Models are flexible

… recipes aren’t

Every study design and every resulting dataset presents unique challenges to the analyst.

> “If all you have is a hammer, everything looks like a nail”

Models can improve our inferences

…and allow more nuanced questions

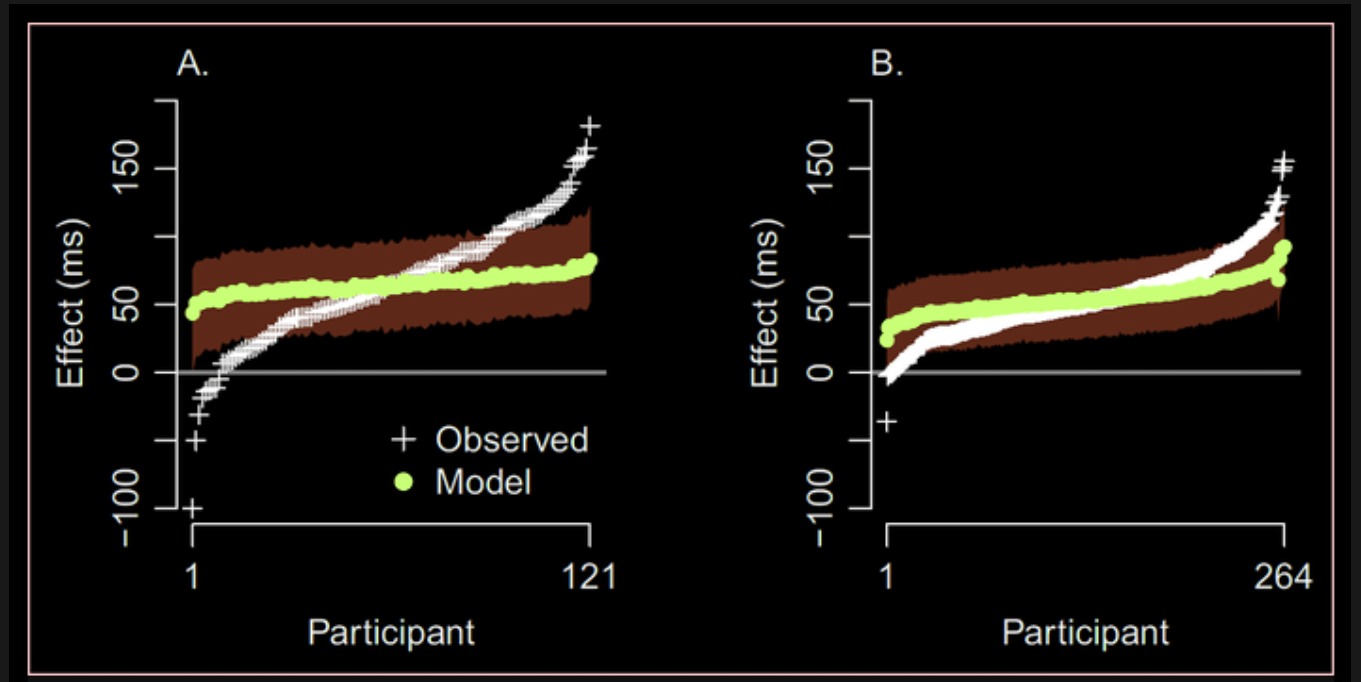


Figure 2

Observed and model-estimated effects. The observed effects are shown as crosses, and the variability of these estimates reflects both trial noise and true variability across people. The model-estimated effects are shown as circles, and they account for trial noise reflecting only true variability across people. **A.** Stroop-effect data from Von Bastian et al. (2015). **B.** Stroop-effect data from Rey-Mermet et al. (2018).

Rouder & Haaf (2021)

Models enable simulation

- Data simulation allows us to ask “what if?” questions
- Data simulation enables us to estimate power for complex analyses where closed-form solutions are unavailable

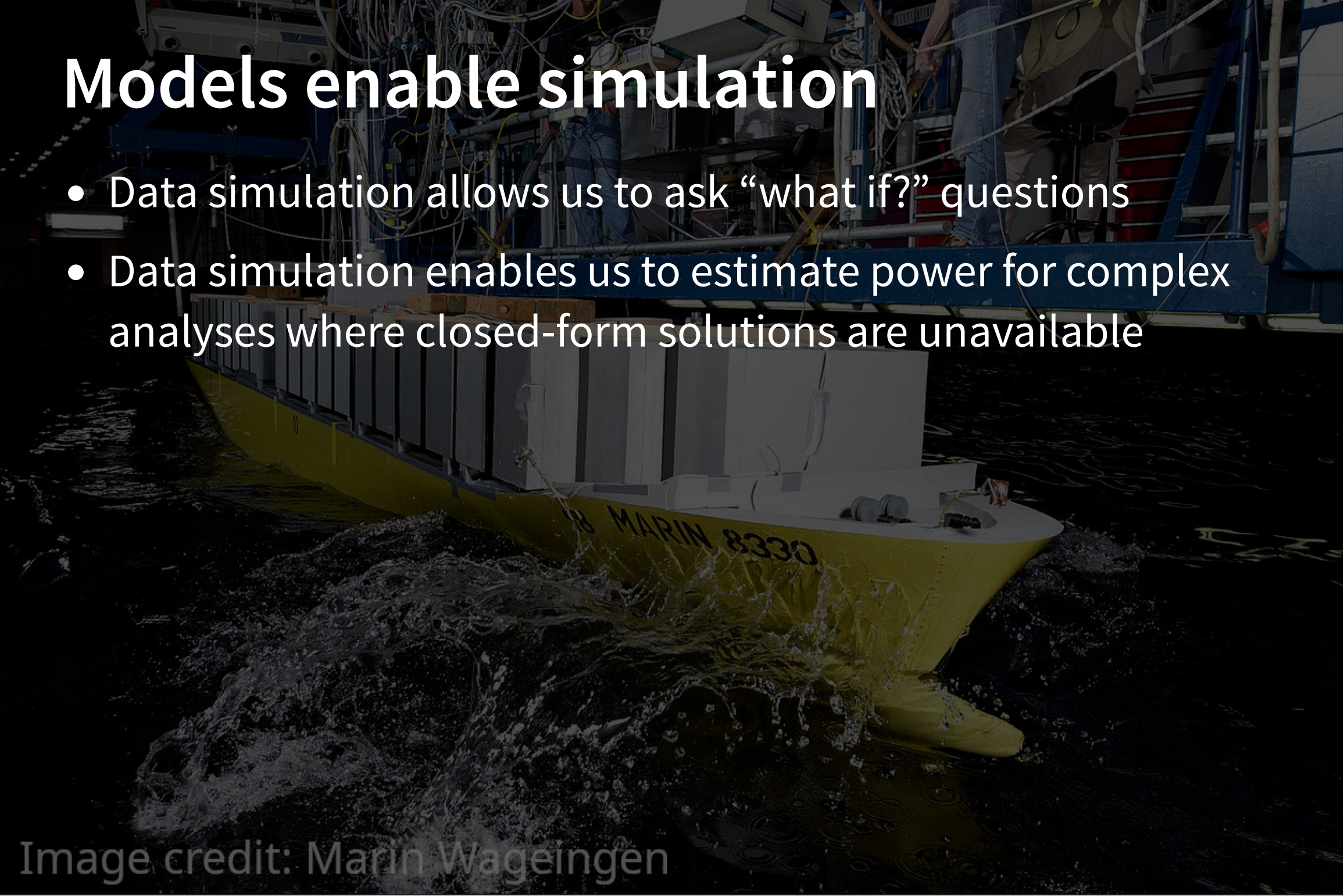


Image credit: Marin Wageingen

Everything is just regression

- t-test
- correlation & regression
- multiple regression
- analysis of variance
- linear mixed-effects modeling

- All are special cases of the General Linear Model (GLM).

GLM approach

1. Define a mathematical model representing the processes that are assumed to give rise to the data
2. Estimate the parameters of the model
3. Validate the model
4. Interpret the results
5. Transparently report what you did
    - share your code & (anonymized) data

Example: Parent reflexes

Does being the parent of a toddler sharpen your reflexes?

- simple response time to a flashing light
- dependent (response) variable: mean RT for each parent

Simulating data

```r
set.seed(1451) # RNG seed: arbitrary integer value
parents <- rnorm(n = 50, mean = 480, sd = 40)

parents
```

```
 [1] 371.6093 441.0581 431.5599 500.2028 494.0001 559.8576 477.7761 438.0509
 [9] 419.6831 473.9525 567.7300 522.1104 515.5739 478.6110 489.2046 458.0759
[17] 409.6295 442.3012 569.0816 460.1299 536.3534 554.4357 499.2558 437.7825
[25] 394.8439 474.3383 470.9229 495.3417 524.6270 442.9606 473.4193 480.3155
[33] 524.6923 462.7873 425.7251 479.8248 530.8883 453.9530 452.5232 459.4812
[41] 483.1470 503.4610 503.3792 545.1496 477.1833 493.2693 461.2054 467.7529
[49] 524.6115 475.6842
```

Control group

```r
set.seed(1451) # RNG seed: arbitrary integer value
parents <- rnorm(n = 50, mean = 480, sd = 40)

parents
```

```
 [1] 371.6093 441.0581 431.5599 500.2028 494.0001 559.8576 477.7761 438.0509
 [9] 419.6831 473.9525 567.7300 522.1104 515.5739 478.6110 489.2046 458.0759
[17] 409.6295 442.3012 569.0816 460.1299 536.3534 554.4357 499.2558 437.7825
[25] 394.8439 474.3383 470.9229 495.3417 524.6270 442.9606 473.4193 480.3155
[33] 524.6923 462.7873 425.7251 479.8248 530.8883 453.9530 452.5232 459.4812
[41] 483.1470 503.4610 503.3792 545.1496 477.1833 493.2693 461.2054 467.7529
[49] 524.6115 475.6842
```

```r
control <- rnorm(n = 50, mean = 500, sd = 40)

control
```

```
 [1] 532.8189 543.7581 528.3925 509.5635 453.8690 559.6398 540.3500 506.3915
 [9] 504.1397 515.3259 449.9757 406.2392 533.1140 468.4039 486.0484 546.7010
[17] 549.2453 449.2057 477.6283 591.3575 486.1334 510.9294 460.1222 516.6436
[25] 515.0298 494.2704 537.5971 511.3171 496.7403 472.8884 563.3012 501.2946
[33] 564.4603 522.0694 444.4553 512.0510 482.7866 563.2460 521.9919 521.8803
[41] 498.7435 453.5540 509.3144 513.5244 524.6281 470.4128 461.5146 485.6436
[49] 519.5610 453.9396
```

t-test

```r
t.test(parents, control, var.equal = TRUE)
```

```
    Two Sample t-test

data:  parents and control
t = -2.9687, df = 98, p-value = 0.00376
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -40.467066  -8.040887
sample estimates:
mean of x mean of y
 480.5903  504.8443
```

Analysis of variance (ANOVA)

```
dat <- tibble(
  group = rep(c("parent", "control"),
              c(length(parents), length(control))),
  rt = c(parents, control))

dat
```

```
# A tibble: 100 × 2
   group     rt
   <chr>  <dbl>
 1 parent  372.
 2 parent  441.
 3 parent  432.
 4 parent  500.
 5 parent  494.
 6 parent  560.
 7 parent  478.
 8 parent  438.
 9 parent  420.
10 parent  474.
# ℹ 90 more rows
```

```r
summary(aov(rt ~ group, dat))
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
group        1  14706   14706   8.813 0.00376 **
Residuals   98 163535    1669
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$e_i \sim N(0, \sigma^2)$$

```r
summary(lm(rt ~ group, dat))
```

```
Call:
lm(formula = rt ~ group, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-108.98  -26.78   -0.49   23.04   88.49

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  504.844      5.777  87.388  < 2e-16 ***
groupparent  -24.254      8.170  -2.969  0.00376 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Single- vs Multi-level data

sub	A	Y
1	A1	876
2	A1	623
3	A1	856
4	A2	755
5	A2	764
6	A2	933

sub	stim	A	Y
1	A	A1	817
1	B	A1	756
1	C	A1	786
2	A	A2	768
2	B	A2	769
2	C	A2	745

Our operational definition of multi-level data: Data where there are multiple observations per sampling unit (e.g., participant) on the dependent variable

Issues with multi-level data

- GLMs assume independence of observations, but observations within a cluster (participant) are not independent
- Any sources of non-independence must be modeled (we'll learn this later!) or aggregated away
- Typical consequence of failing to do so: High false positives

technique	t-test	ANOVA	regression
Categorical IVs	✓	✓	✓
Continuous DVs	✓	✓	✓
Continuous IVs	-	?	✓
Multi-level data	?	?	✓
Categorical DVs	-	-	✓
Unbalanced data	-	-	✓
>1 sampling unit	-	-	✓

Four functions to rule them all

1. Is the data single- or multi-level?
2. Is the response continuous or discrete?
3. How are the observations distributed?

structure	response	distribution	R function
single	cont	normal	`base::lm()`
single	cont/disc	various	`base::glm()`
multi	cont	normal	`lme4::lmer()`
multi	cont/disc	various	`lme4::glmer()`

Part 1: Regression & multilevel models

lecture	topic
1	introduction
2	correlation & regression
3	multiple regression
4	interactions
5	multilevel models

Part 2: Multivariate models & psychometrics

lecture	topic
6	introduction to multivariate data
7	path analysis
8	mediation models
9	confirmatory factor analysis
10	structural equation modeling

How it will go

Course materials available on Moodle:

- Online book
- Lecture slides & recordings
- Weekly formative exercises
  - download R Markdown ‘stub’ file and fill in with code

Complete the formative exercises *after* attending each lecture.

Read the corresponding book chapter before *before* attempting the formative assignment.

Formative assignments

- Download the assignment files from Moodle*
- Fill your answers into the code chunks provided
- Check for errors (knitting and validation)
- Complete and submit the plain R Markdown file on Moodle before the due date
- Compare your answers with the solution

* Look for your assignment files under ‘feedback files’. One will be an R Markdown file (.Rmd). Others will be associated data files.

Help and discussion

- My student drop-in hours:
  - 58/60 Hillhead Street, Room 557
  - Tuesdays 2:00-3:30
- Microsoft Teams channel

Communication is important!

NB: If you have a question it is better to ask it on the Teams channel than by email/DM so that others can benefit

Assessment

Two one-hour, timed, online assessments held during reading week and the week of the final lecture (dates on Moodle).

- First assessment: chapters/lectures 1-5
- Second assessment: chapters/lectures 6-9

Question format:

- Multiple choice / true-false questions
- Variations on the formative exercises

Hot tip!

Tip

Summative assessments will involve the same basic workflow as the formative assessments. You are responsible for mastering this workflow. Completing the full workflow (including submission) for all the formative assessments is the best way to avoid nasty surprises.

You are also responsible for making sure you have the necessary technical resources including software ready to go in time for the assessment.

The following types of excuses will not find a sympathetic ear:

- “I couldn’t find the stub file / couldn’t find where to upload the R Markdown Script”
- “I uploaded the wrong file”
- “Tidyverse was taking forever to install and so I ran out of time”

For next week

- Get your workstation in order
  - R version 4.1.0 or higher
  - R packages for analysis:
    - `tidyverse`, `lme4`, `psych`, `corrr`, `lavaan`
  - R packages for working with R markdown
    - `rmarkdown`, `knitr`
- Attempt formative assignment 1