Models of Reality

From Data to Science

Prof. Dr. Jochen Kruppa-Scheetz

2025-11-23

Table of contents

Welcome

"Life is difficult." — M. Scott Peck's, The Road Less Traveled

This book has a German predecessor. You might not recognise this book. That's fine. The page can be found at the following link: https://jkruppa.github.io/. However, it is much more unstructured. It is more like a cookbook than this book, which has a continuous storyline.

Preface

It's too early to write a preface. I started writing this book a few weeks ago. From my perspective, it will be a good one, but it will take some time. In the meantime, you are welcome to see what's happening here and what I'm writing. Sometimes things will be in the wrong place. This is a normal part of the writing process. Just so you know, I started writing on 15 November 2025.

text[1]

> *"The first principle is that you must not fool yourself — and you are the easiest person to fool. So you have to be very careful about that." —*[2]

> *"Hic sunt dracones!" — Hic Sunt Dracones on the Hunt–Lenox Globe (eng. "Here be dragons")*

[3]

References

Part I

From science to data by models

Last modified on 22. November 2025 at 19:08:33

> *"I'd like to solve the puzzle." — Wheel of Fortune*

In 2008, Cadiergues and his team won the Ig Nobel Prize at Harvard University in Cambridge for their work with dog and cat fleas[4]. This book tells the story of what had happened if they won the Nobel Prize in Stockholm instead. The Nobel Prize sparked extensive research in many scientific fields. Fleas became all the rage. Lots of data was collected, fleas were measured, and questions were asked that no one had ever asked before. Let's check out this data and see what it tells us about fleas. Come along to Adventure Land and get ready to be amazed.

None of this happened in our reality. At least not in our branch of reality, if you believe Everett[5]. But since this book is also about models of reality, we can question the whole of reality and create a new one that fits.

References

1 What is science?

Last modified on 18. November 2025 at 20:02:34

> *"Reality is negotiable." — Tim Ferriss*

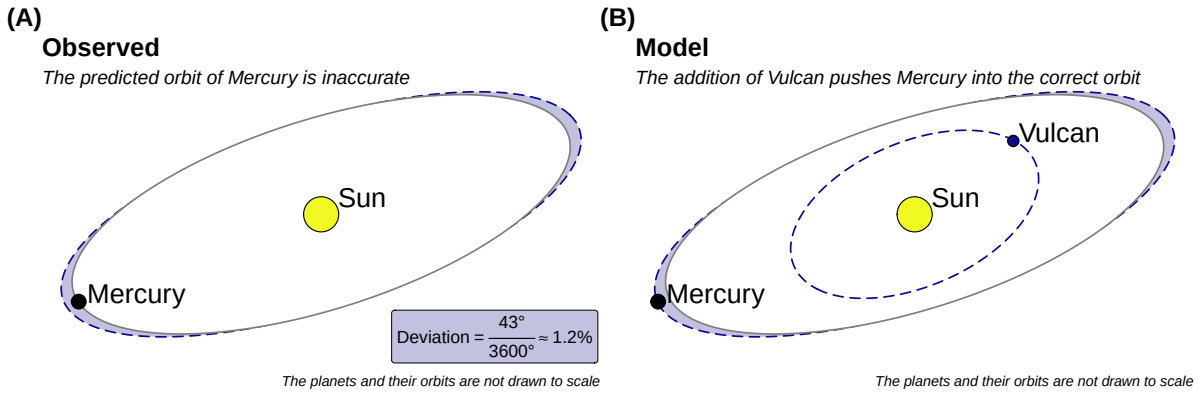

Figure 1.1: foo **(A)** foo **(B)** foo

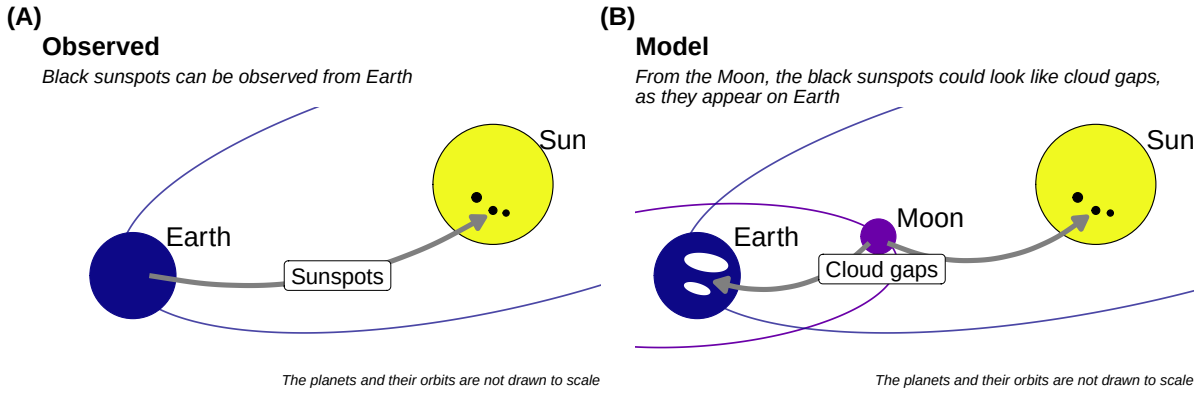

Figure 1.2: foo **(A)** foo **(B)** foo

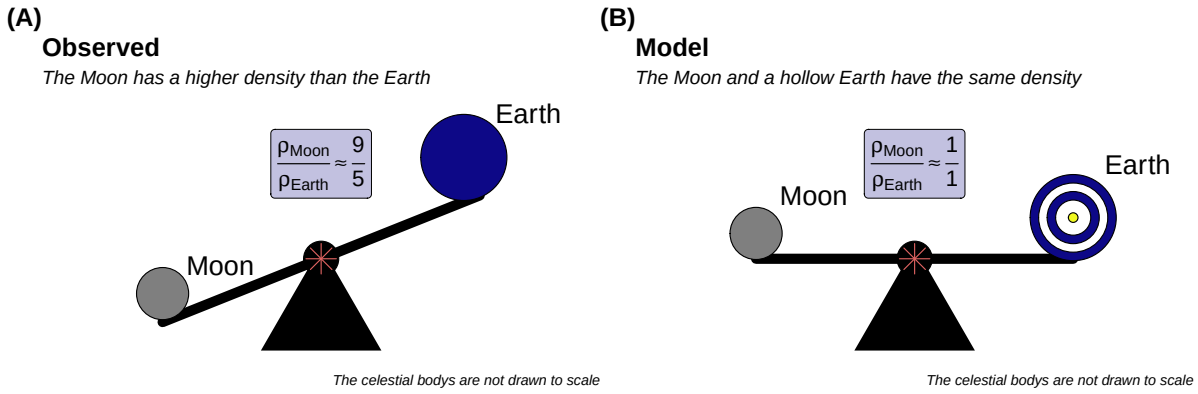

Figure 1.3: foo **(A)** foo **(B)** foo

1.1 General background

> *"Any sufficiently advanced technology is indistinguishable from magic." —*[6]*, Clarke's third law*

Full quote

> *"Reality is negotiable. Outside of science and law, all rules can be bent or broken, and it doesn't require being unethical." — Tim Ferriss*

Idea of hypotheses

Science is guessing and falsification

[7] What is this thing called Science?

[8] What is science

[9] What is science?

[10] Models Demystified: A Practical Guide from Linear Regression to Deep Learning

[11] Statistical Thinking for the 21st Century

[12] The beginning of infinity: Explanations that transform the world

David Deutsch > Quotes

1.2 Theoretical background

1.3 R packages used

1.4 Data

1.5 Alternatives

Further tutorials and R packages on XXX

1.6 Glossary

term what does it mean.

1.7 The meaning of "Models of Reality" in this chapter.

- itemize with max. 5-6 words

1.8 Summary

References

2 What is data?

Last modified on 18. November 2025 at 07:21:12

> *"The limits of my language mean the limits of my world." — Ludwig Wittgenstein*

2.1 General background

2.2 Theoretical background

2.3 R packages used

2.4 Data

```
jump_weight_tbl <- tibble(x = c(0.6, 1, 2.3, 3.5, 5.2, 7.1, 8.4, 9.2, 10),
                          y = 0.15*x^3 - 2.2*x^2 + 8.8*x + 3.2 + rnorm(9, 0, 0.5)) |>
  mutate_all(round, 1) |>
  rename(weight_mg = x, jump_length_cm = y)
```

Table 2.1: foo.

weight_mg	jump_length_cm
0.6	8.5
1.0	9.6
2.3	13.7
3.5	13.2
5.2	11.0
7.1	8.3
8.4	10.9
9.2	14.3
10.0	21.1

Equation 2.1

$$y = 0.15 \cdot x^3 - 2.2 \cdot x^2 + 8.8 \cdot x + 3.2 \tag{2.1}$$

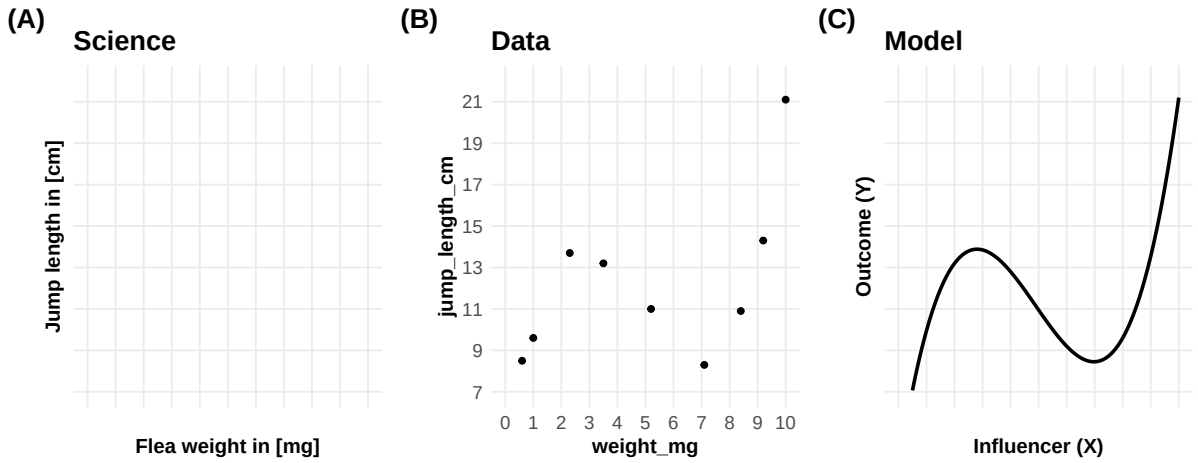

Figure 2.1: foo.

2.5 Alternatives

Further tutorials and R packages on XXX

2.6 Glossary

term what does it mean.

2.7 The meaning of "Models of Reality" in this chapter.

- itemize with max. 5-6 words

2.8 Summary

References

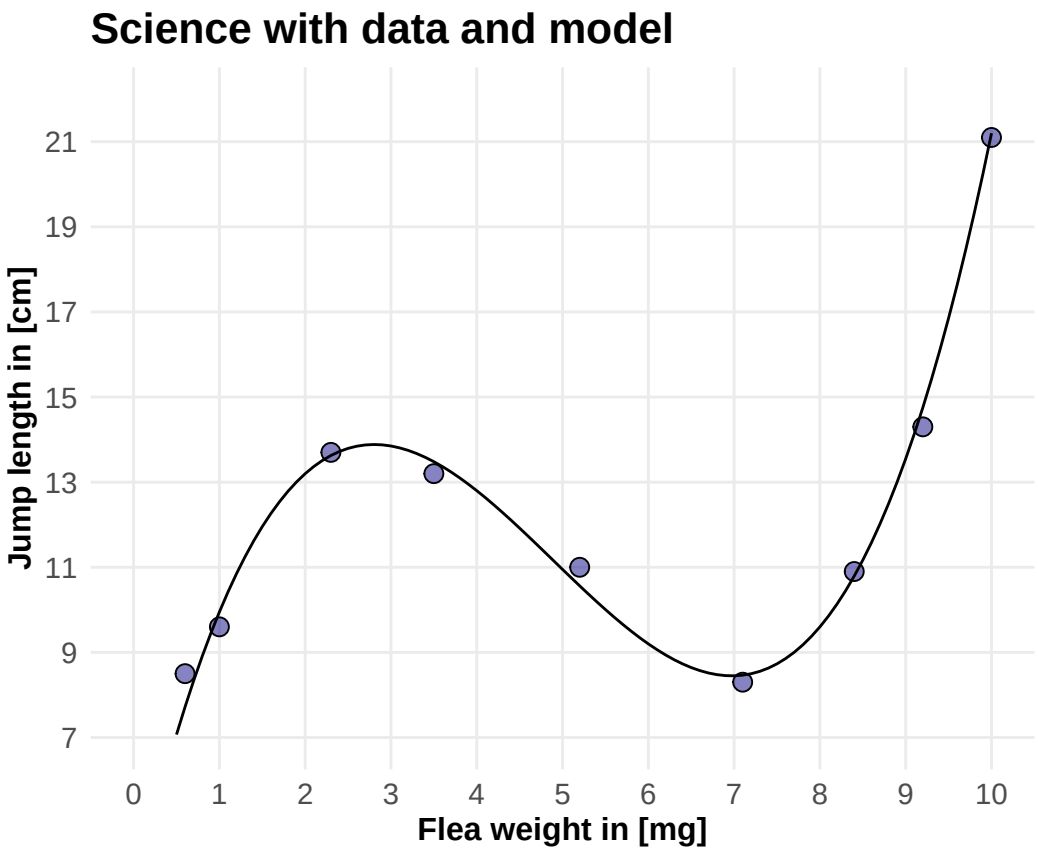

Figure 2.2: foo.

3 What is a model?

Last modified on 22. November 2025 at 19:25:57

> *"A quote." — Dan Meyer*

Imagine you are standing on a globe but still believe it is flat. You are standing on solid ground, with boiling rocks a few kilometers below you. A few kilometers above, you cannot breathe anymore. You are not moving, but the Earth is moving at speeds of up to 828,000 km/h, depending on the reference point. Your reality is a serious misinterpretation.

3.1 General background

Table 3.1: Table of terms used in the statistical modelling. The terms in bold are used here. Depending on the scientific background, the usage of these terms can vary widely.

Symbol	Name Application	Description
y	**outcome**, response, endpoint, dependent variable	*The right-hand side (abbr. RHS) of the model. Describing the values measured in an experiment or study.*

Symbol	Name	Application	Description
x	**influencer**, influential variable, risk factor, fixed effect, independent variable		*The left-hand side (abbr. LHS) of the model. Describing the influential variables in an experiment or study.*
z	**random effect**		*A factor that provides a description of an grouping variable, which is not part of the controlled experimental setting.*
x	**explanator**, explanatory variable	Explanation	*The influencer is used to describe or explain the outcome.*
x	**predictor**, predictive variable	Prediction	*The influencer is used to predict the outcome.*

Symbol	Name	Application	Description
x	**focal explanator**, **focal predictor**, focal variable	Main effect	*In a model with multiple influencers, the focal variable is the variable of primary interest.*
c	**covariate**, covariable	Continuous x	*The influencer is a numeric variable with continuous values.*
f_A	**factor A**, factorial variable, categorical variable	Categorical x	*The influencer is discrete, functioning as a grouping variable, such as an experimental group or a treatment.*
$A.1$ to $A.j$	**levels**, groups, treatment groups	Factor f_A	*The discrete groups included in one factor A.*

A sentence why we use y and not x for mean and other stuff.

[13] What is a statistical model?

[14] What do we mean by a statistical model?

[15] What Is the Purpose of Statistical Modeling?

[16] Where do statistical models come from? Revisiting the problem of specification

[17] Statistical modelling

3.2 Theoretical background

3.3 R packages used

3.4 Data

3.5 Alternatives

Further tutorials and R packages on XXX

3.6 Glossary

term what does it mean.

3.7 The meaning of "Models of Reality" in this chapter.

- itemize with max. 5-6 words

3.8 Summary

References

Part II

Tales of data

Last modified on 22. November 2025 at 19:18:43

> *"X-Factor, the unfathomable. We live in a world where dreams and reality are closely intertwined, where facts often seem like figments of the imagination that we cannot explain. Can you distinguish between truth and lies? Yours, Jonathan Frakes." — German intro of Beyond Belief: Fact or Fiction*

Once upon a time Cadiergues and his team won the Nobel Prize in Stockholm for their work with dog and cat fleas[4]. This sparked a time of increased scientific research on fleas of all types of hosts, with scientists studying them in great detail. A wealth of research questions remained to be addressed, while funding opportunities lay in wait. We now have the great opportunity to use open-source data to revisit all the exciting findings in flea research. What are our main research topics and questions that we want to cover?

Unfortunately, none of this ever happened in our branch of reality. But it might happen in countless other branches[5].

References

4 Body weight of fleas

Last modified on 22. November 2025 at 09:30:36

> *"A quote." — Dan Meyer*

4.1 General background

4.2 Theoretical background

4.3 R packages used

4.4 Data

4.4.1 Linear

```
jump_weight_tbl <- tibble(x = abs(rnorm(21, 5, 4)),
                          y = 10 + 1.2 * x + rnorm(21, 0, 2)) |>
  mutate_all(round, 1) |>
  rename(weight_mg = x, jump_length_cm = y)
```

```
jump_weight_feeding_tbl <- expand_grid(f = c(0, 1),
                                       x = abs(rnorm(21, 5, 4))) |>
  mutate(y = 10 + 1.2 * x + 10 * f + rnorm(42, 0, 2),
         f = factor(f, labels = c("sugar_water", "blood")))  |>
  mutate_if(is.numeric, round, 1) |>
  rename(weight_mg = x, jump_length_cm = y, feeding = f)
```

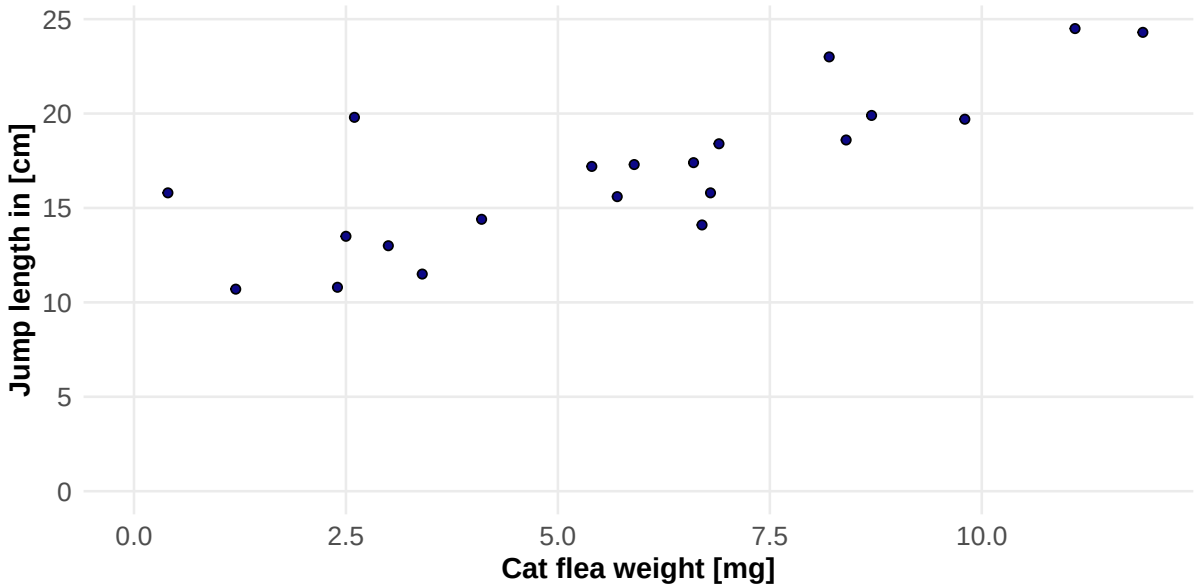

Figure 4.1: foo **(A)** foo **(B)** foo

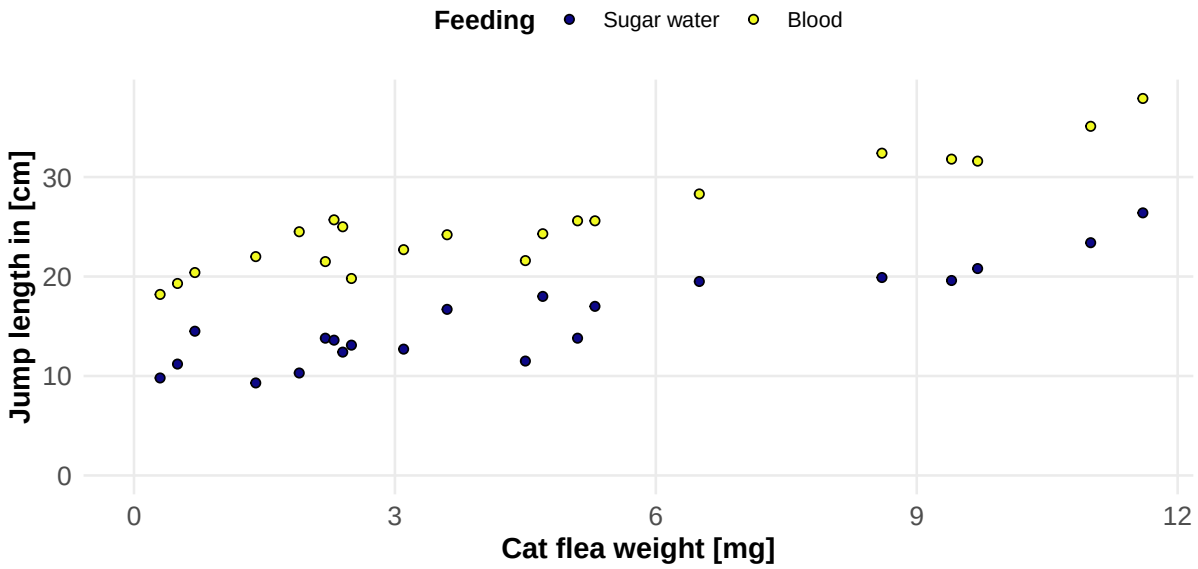

Figure 4.2: foo **(A)** foo **(B)** foo

4.4.2 Non-linear

```
jump_weight_non_linear_tbl <- tibble(x = abs(rnorm(32, 3, 3)),
                                     y = 0.15*x^3 - 2.2*x^2 + 8.8*x + 3.2 + rnorm(32, 0, 1))
  rename(weight_mg = x, jump_length_cm = y)
```

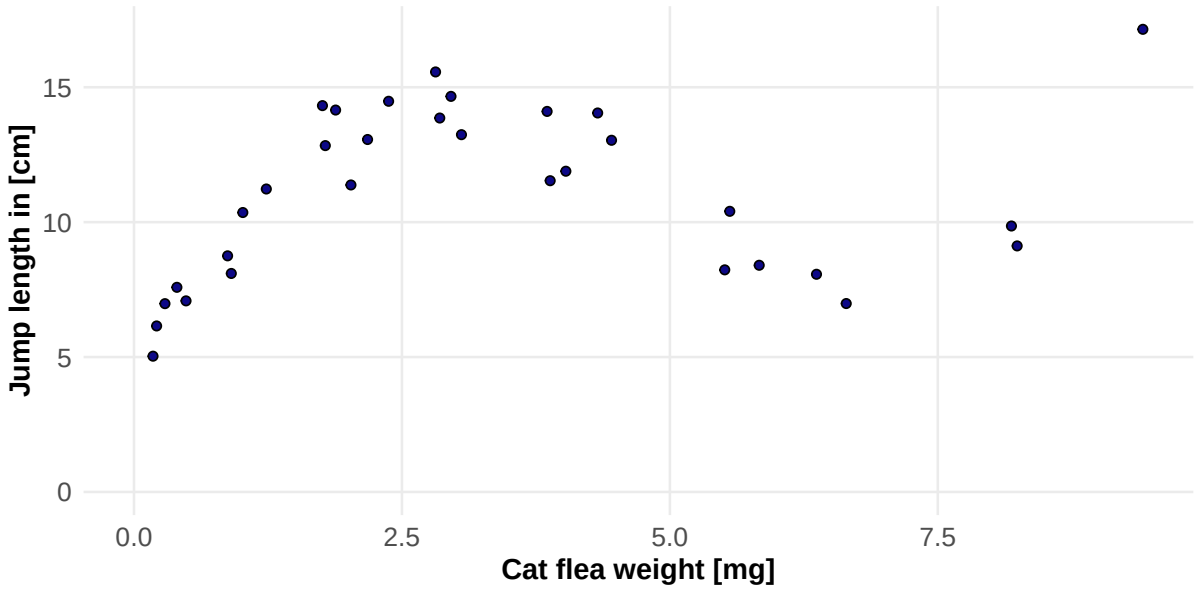

Figure 4.3: foo **(A)** foo **(B)** foo

```
jump_weight_non_linear_tbl <- expand_grid(f = c(0, 1),
                                          x = abs(rnorm(42, 2.5, 2.5))) |>
  mutate(y = 25 - 4*f - 21 * exp(-0.2 * x^2 - 1.1 * f) + rnorm(84, 0, 1),
         f = factor(f, labels = c("sugar_water", "blood")))  |>
  mutate_if(is.numeric, round, 1) |>
  rename(weight_mg = x, jump_length_cm = y, feeding = f)
```

4.5 Alternatives

Further tutorials and R packages on XXX

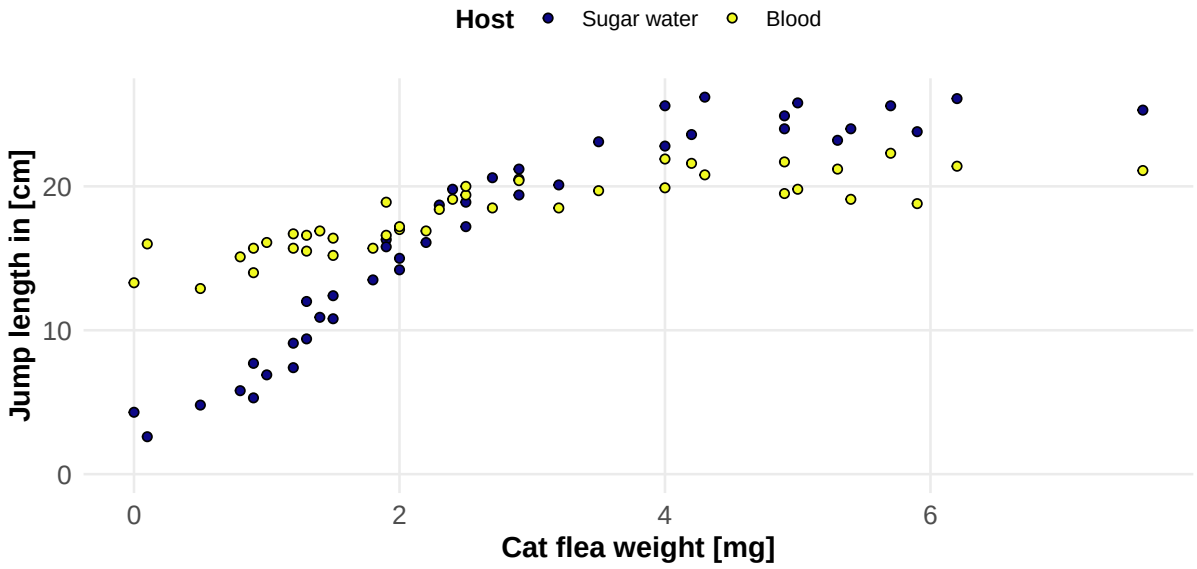

Figure 4.4: foo **(A)** foo **(B)** foo

4.6 Glossary

term what does it mean.

4.7 The meaning of "Models of Reality" in this chapter.

- itemize with max. 5-6 words

4.8 Summary

References

5 Fleas of animals

Last modified on 21. November 2025 at 15:21:35

> *"A quote." — Dan Meyer*

5.1 General background

What is a data grid?

5.2 Theoretical background

5.3 R packages used

```
pacman::p_load(tidyverse, conflicted)
```

18

5.4 Data

Small data and grid

5.4.1 Jump performances of fleas

4

C. felis felis jump was $19.9 \pm 9.1 cm$ with a range from 2 to $48 cm$

C. canis jump was longer $30.4 \pm 9.1 cm$ with a range from 3 to $50 cm$

```r
jump_flea_grid <- expand_grid(host = c("cat", "dog")) |>
  mutate(mean = c(19.9, 30.4),
         sd = c(9.1))
```

```r
jump_flea_tbl <- jump_flea_grid |>
  rowwise() |>
  mutate(jump_length_cm = lst(rnorm(7, mean, sd))) |>
  unnest(cols = jump_length_cm) |>
  mutate(host = as_factor(host))
```

```r
jump_flea_tbl |>
  group_by(host) |>
  summarise(mean(jump_length_cm),
            sd(jump_length_cm)) |>
  mutate_if(is.numeric, round, 2)
```

```
# A tibble: 2 x 3
  host  `mean(jump_length_cm)` `sd(jump_length_cm)`
  <fct>                  <dbl>                 <dbl>
1 cat                     21.2                  7.3
2 dog                     33.0                  6.57
```

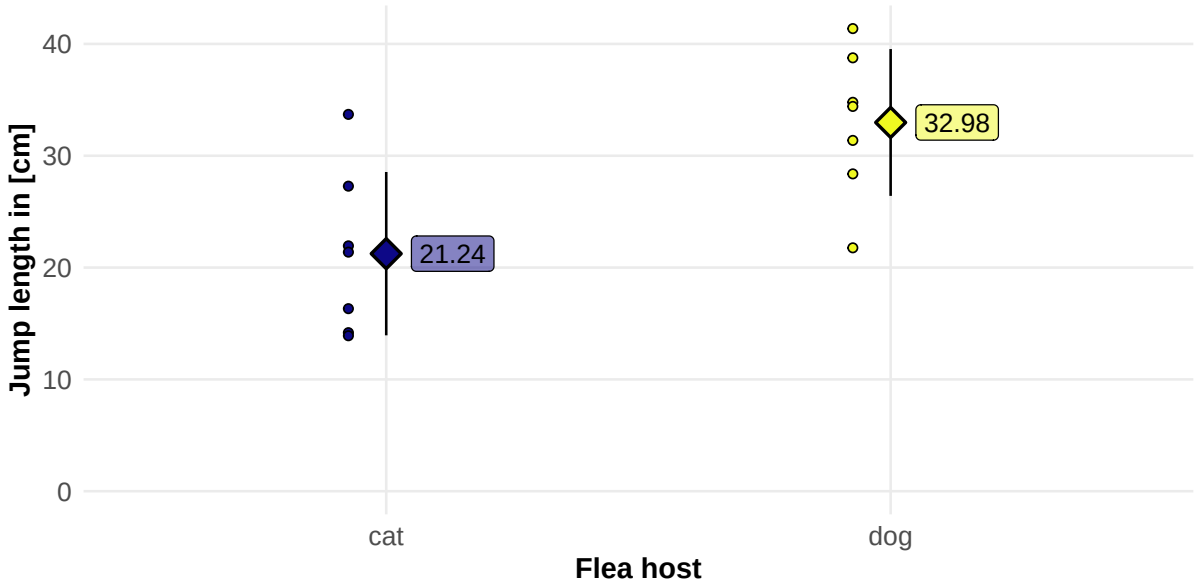

Figure 5.1: foo **(A)** foo **(B)** foo

```
jump_animals_grid <- expand_grid(host = c("cat", "fox", "rat", "dog")) |>
  mutate(mean = c(19.9, 35.2, 15.2, 30.4),
         sd = c(9.1, 10.3, 4.6, 9.1))
```

```
jump_animals_tbl <- jump_animals_grid |>
  rowwise() |>
  mutate(jump_length_cm = lst(rnorm(7, mean, sd))) |>
  unnest(cols = jump_length_cm) |>
  mutate(host = as_factor(host))
```

```
jump_animals_tbl |>
  group_by(host) |>
  summarise(mean(jump_length_cm),
            sd(jump_length_cm)) |>
  mutate_if(is.numeric, round, 2)
```

```
# A tibble: 4 x 3
  host  `mean(jump_length_cm)` `sd(jump_length_cm)`
  <fct>                  <dbl>                <dbl>
1 cat                     21.2                 9.36
2 fox                     34.6                10.2
3 rat                     13.4                 2.51
4 dog                     34.0                17.4
```

5.4.2 Measurements on animal fleas

- Jump length in [cm] called `jump_length`
- Number of hairs on each leg called `count_leg_left` and `count_leg_right`
- Ratings of each flea, as listed in the catalog of the Fédération Internationale de la Beauté des Puces (FIBP) called `rating` on a Likert scale from 1 to 5, with 5 being the strongest expression.
- The infection status with the flea cold is called `infected`, with a value of 0/1 for no/yes.

```
tibble(jump_length_cm = rnorm(7, 5, 1),
       count_right_leg = rpois(7, 4),
       count_left_leg = rpois(7, 4),
       count_leg = (count_left_leg + count_right_leg)/2,
       rating = sample(1:5, 7, replace = TRUE, prob = c(0.1, 0.2, 0.4, 0.2, 0.1)),
       infectd = rbinom(7, prob = 0.5, size = 1))
```

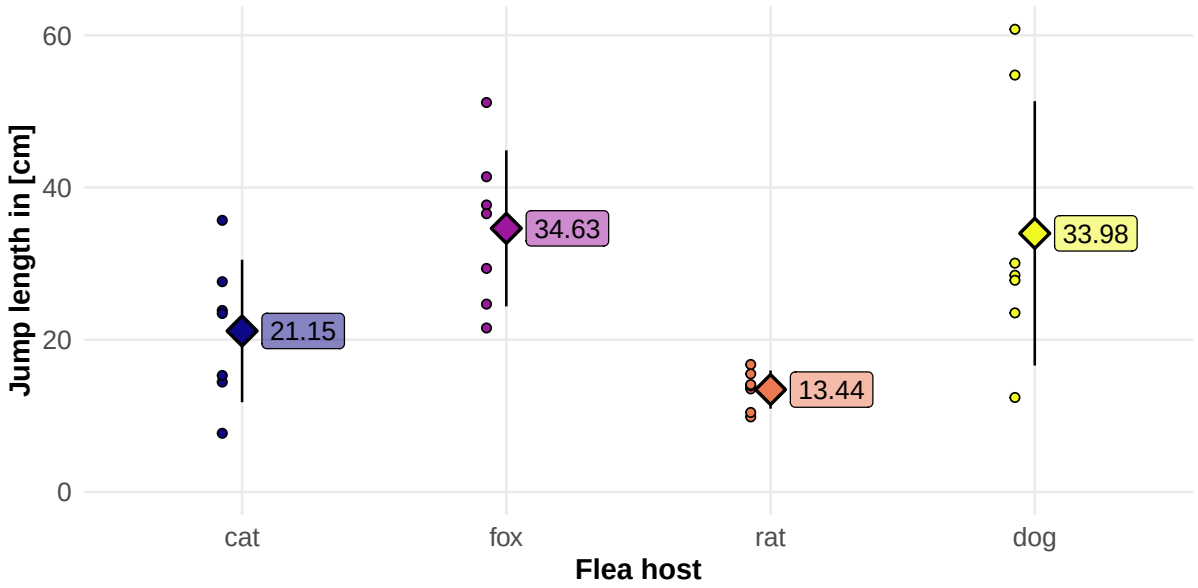

Figure 5.2: foo **(A)** foo **(B)** foo

```
# A tibble: 7 x 6
  jump_length_cm count_right_leg count_left_leg count_leg rating infectd
           <dbl>           <int>          <int>     <dbl>  <int>   <int>
1           6.17               4              8       6        3       1
2           5.03               5              5       5        4       0
3           6.51               6              1       3.5      3       0
4           4.38               2              1       1.5      3       1
5           2.62               0              3       1.5      3       0
6           6.64               3              3       3        2       0
7           3.82               3              3       3        3       0
```

5.4.3 Fleas in urban and rural habitats

Why so complex?

[19]

```
jump_habitat_grid <- expand_grid(host = 1:3, site = 1:2) |>
  mutate(mean_host = c(19.9, 19.9, 30.4, 30.4, 15.2, 15.2),
         mean_site = c(5, 0, 5, 0, 5, -5),
         mean = mean_host + mean_site,
         sd = c(9.1, 9.1, 9.1, 9.1, 4.6, 4.6))
jump_habitat_grid
```

```
# A tibble: 6 x 6
   host  site mean_host mean_site  mean    sd
  <int> <int>     <dbl>     <dbl> <dbl> <dbl>
1     1     1      19.9         5  24.9   9.1
2     1     2      19.9         0  19.9   9.1
3     2     1      30.4         5  35.4   9.1
4     2     2      30.4         0  30.4   9.1
5     3     1      15.2         5  20.2   4.6
6     3     2      15.2        -5  10.2   4.6
```

```
jump_habitat_raw_tbl <- jump_habitat_grid |>
  rowwise() |>
  mutate(jump_length_cm = lst(rnorm(7, mean, sd))) |>
  unnest(cols = jump_length_cm)
```

```
jump_habitat_tbl <- jump_habitat_raw_tbl |>
  select(host, site, jump_length_cm) |>
  mutate(host = factor(host, labels = c("cat", "dog", "rat")),
         site = factor(site, labels = c("urban", "rural"))) |>
  mutate_if(is.numeric, round, 2)
```

```
jump_habitat_tbl |>
  group_by(host, site) |>
  summarise(mean(jump_length_cm),
            sd(jump_length_cm)) |>
  mutate_if(is.numeric, round, 2)
```

```
# A tibble: 6 x 4
# Groups:   host [3]
  host  site  `mean(jump_length_cm)` `sd(jump_length_cm)`
  <fct> <fct>                 <dbl>                <dbl>
1 cat   urban                  21.8                 9.26
2 cat   rural                  23.6                 7.23
3 dog   urban                  35.2                 9.1
4 dog   rural                  27.9                12.8
5 rat   urban                  19.2                 1.85
6 rat   rural                  12.8                 3.74
```

5.5 Data availability

The data is available as txt-Files under https://github.com/jkruppa/biodatascience.

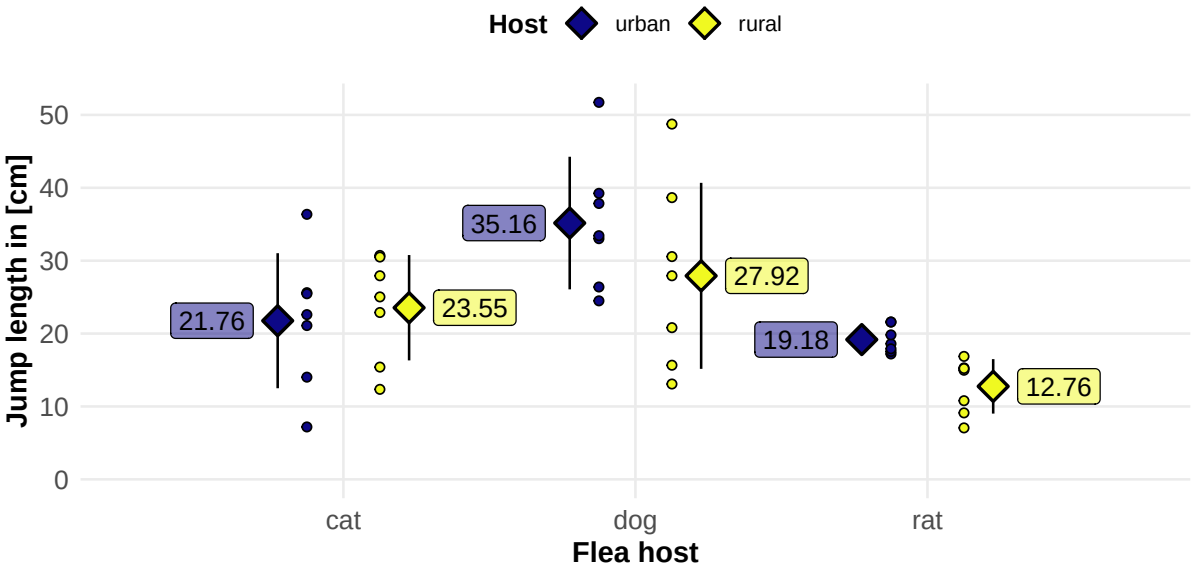

Figure 5.3: foo **(A)** foo **(B)** foo

5.6 Alternatives

Further tutorials and R packages on XXX

5.6.1 Why linear is a bad idea?

```
jump_flea_tbl <- tibble(host = rep(0:1, each = 7),
                        jump_length = 19.9 + 10.5 * host + rnorm(14, 0, 9.1)) |>
  mutate(host = factor(host, labels = c("cat", "dog")))
```

5.6.2 Why is a `tibble()` bad idea?

```
jump_animals_tbl <- tibble(cat = rnorm(n = 7, mean = 19.9, sd = 9.1),
                           fox = rnorm(n = 7, mean = 35.2, sd = 10.3),
                           rat = rnorm(n = 7, mean = 15.2, sd = 4.6),
                           dog = rnorm(n = 7, mean = 30.4, sd = 9.1)) |>
  pivot_longer(cols = cat:dog, values_to = "jump_length", names_to = "host") |>
  mutate(host = as_factor(host))
```

5.7 Glossary

term what does it mean.

5.8 The meaning of "Models of Reality" in this chapter.

- itemize with max. 5-6 words

5.9 Summary

References

6 Eating habits of fleas

Last modified on 20. November 2025 at 10:17:21

> *"A quote." — Dan Meyer*

6.1 General background

Based on standard methods in flea research and experimental entomology, as well as a similar published experiment, the following three types of food for adult cat fleas are possible, representing different nutritional conditions:

Blood: This is the natural and optimal food source for adult fleas.Often, defibrinated or anticoagulated animal blood (e.g. bovine blood, rabbit blood) is used for this purpose, which is offered in special in vitro feeding systems (e.g. through a membrane).Expectation: Fleas that are optimally nourished should show the greatest jumping distance.

Sugar water: Often serves as a 'control feed' or as a feed that provides energy (sugar) but lacks essential nutrients (such as proteins from blood).Expectation: Jumping distance could be reduced due to the lack of blood and thus the proteins important for reproduction, which could impair physiological fitness.

Ketchup - a nutrient-poor or unsuitable food: This option is used to simulate poor, incomplete or stressful nutritional conditions.In a similar documented experiment, ketchup (a combination of sugar, vinegar and minimal other substances, but no blood) was used as the third feed.Expectation: It can be assumed that fleas will show the shortest jumping distance under these conditions, as they lack both essential nutrients and the necessary energy.

6.2 Theoretical background

6.3 R packages used

6.4 Data

6.5 Alternatives

Further tutorials and R packages on XXX

6.6 Glossary

term what does it mean.

6.7 The meaning of "Models of Reality" in this chapter.

- itemize with max. 5-6 words

6.8 Summary

References

Part III

Speaking to data

Last modified on 25. October 2025 at 20:17:25

> *"What problem have you solved, ever, that was worth solving where you knew all the given information in advance? No problem worth solving is like that. In the real world, you have a surplus of information and you have to filter it, or you don't have sufficient information and you have to go find some."* — *Dan Meyer in Math class needs a makeover*

Here comes the preface text

7 Programing in the 21st century

Last modified on 20. November 2025 at 14:02:31

> *"A quote." — Dan Meyer*

7.1 General background

20

18

21

22

7.2 Theoretical background

7.3 R packages used

7.4 Data

7.5 Alternatives

Further tutorials and R packages on XXX

7.6 Glossary

term what does it mean.

7.7 The meaning of “Models of Reality” in this chapter.

- itemize with max. 5-6 words

7.8 Summary

References

Part IV

Template Preface

Last modified on 25. October 2025 at 20:17:25

> *"What problem have you solved, ever, that was worth solving where you knew all the given information in advance? No problem worth solving is like that. In the real world, you have a surplus of information and you have to filter it, or you don't have sufficient information and you have to go find some." — Dan Meyer in Math class needs a makeover*

Here comes the preface text

8 Template chapter

Last modified on 17. November 2025 at 20:02:35

> *"A quote." — Dan Meyer*

8.1 General background

8.2 Theoretical background

8.3 R packages used

8.4 Data

8.5 Alternatives

Further tutorials and R packages on XXX

8.6 Glossary

term what does it mean.

8.7 The meaning of "Models of Reality" in this chapter.

- itemize with max. 5-6 words

8.8 Summary

References

A Why does it look like this?

Last modified on 20. November 2025 at 10:01:22

> *"A quote." — Dan Meyer*

A.1 Used R packages

This is a printout of the `init.R` file, which can be found on GitHub

```
pacman::p_load(tidyverse, ggforce, viridis, knitr, patchwork)
```

A.2 Used theme of the visualitations

```
theme_book <- function() {
  theme_minimal() +
    theme(panel.grid.minor = element_blank(),
          plot.title = element_text(size = 16, face = "bold"),
          plot.subtitle = element_text(size = 12, face = "italic"),
          plot.caption = element_text(face = "italic"),
          axis.title = element_text(size = 12, face = "bold"),
          axis.text = element_text(size = 11),
          legend.title = element_text(face = "bold"))
}
```

A.3 Used color palettes

```
col_pal <- \(n, alpha) plasma(n = n, alpha = alpha)
c6_pal <- col_pal(6, 1)
```

References

[1] Smolin L. *Einstein's Unfinished Revolution: The Search for What Lies Beyond the Quantum*. Penguin; 2019.

[2] Feynman RP, Leighton R. *" Surely You're Joking, Mr. Feynman!": Adventures of a Curious Character*. Random House; 1992.

[3] Feynman RP. Cargo cult science. In: *The Art and Science of Analog Circuit Design*. Elsevier; 1998:55-61.

[4] Cadiergues MC, Joubert C, Franc M. A comparison of jump performances of the dog flea, ctenocephalides canis (curtis, 1826) and the cat flea, ctenocephalides felis felis (bouché, 1835). *Veterinary parasitology*. 2000;92(3):239-241.

[5] Everett III H. 'Relative state' formulation of quantum mechanics. *Reviews of modern physics*. 1957;29(3):454.

[6] Clarke AC. Clarke's third law on UFO's. *Science*. 1968;159(3812):255-255.

[7] Chalmers AF. *What Is This Thing Called Science?* Hackett Publishing; 2013.

[8] Feynman R. What is science. Published online 1966.

[9] Campbell NR. What is science? Published online 1952.

[10] Clark M, Berry S. *Models Demystified: A Practical Guide from Linear Regression to Deep Learning*. CRC Press; 2025.

[11] Cox D, Efron B. Statistical thinking for 21st century scientists. *Science advances*. 2017;3(6):e1700768.

[12] Deutsch D. *The Beginning of Infinity: Explanations That Transform the World*. penguin uK; 2011.

[13] McCullagh P. What is a statistical model? *The Annals of Statistics*. 2002;30(5):1225-1310.

[14] Appleton DR. What do we mean by a statistical model? *Statistics in medicine*. 1995;14(2):185-197.

[15] Hand D. What is the purpose of statistical modeling? *Harvard Data Science Review*. 2019;1(1).

[16] Spanos A. Where do statistical models come from? Revisiting the problem of specification. *Lecture Notes-Monograph Series*. Published online 2006:98-119.

[17] Gilchrist W. *Statistical Modelling*. Wiley Chichester; 1984.

[18] Wickham H, Çetinkaya-Rundel M, Grolemund G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. " O'Reilly Media, Inc."; 2023.

[19] Wickham H, Henry L. *Purrr: Functional Programming Tools.*; 2025. https://purrr.tidyverse.org/

[20] Dormann CF, Ellison AM. *Statistics by Simulation: A Synthetic Data Approach*. Princeton University Press; 2025.

[21] Wickham H. *Conflicted: An Alternative Conflict Resolution Strategy.*; 2023. https://conflicted.r-lib.org/

[22] Rinker T, Kurkiewicz D. *Pacman: Package Management Tool.*; 2012. https://github.com/trinker/pacman