The Duality of Data Visualization

Visualizing Data for Insight and Communication

Jannik Buhr

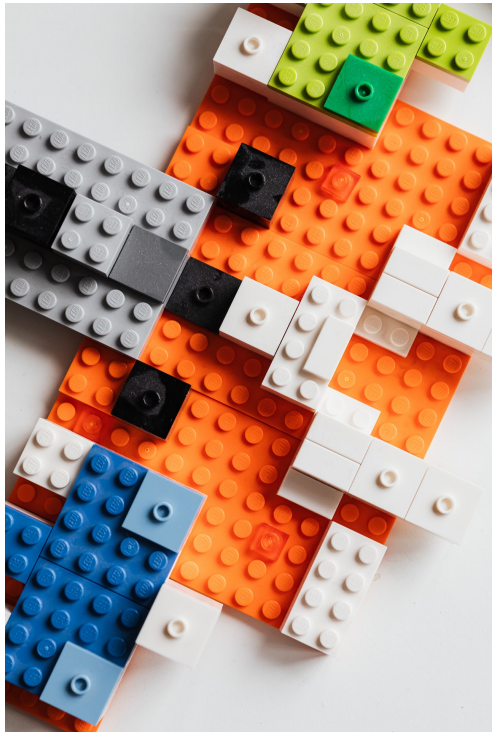
Heidelberg Institute for Theoretical Studies

June 20, 2022

HITS

The Grammar of Graphics

The Grammar of Graphics



- based on “The Grammar of Graphics” by Leland Wilkinson
- popularized by Hadley Wickham with the R package `ggplot2`
- No monolithic plotting functions but instead a set of blocks to combine
- powerful
- iterative
- expressive

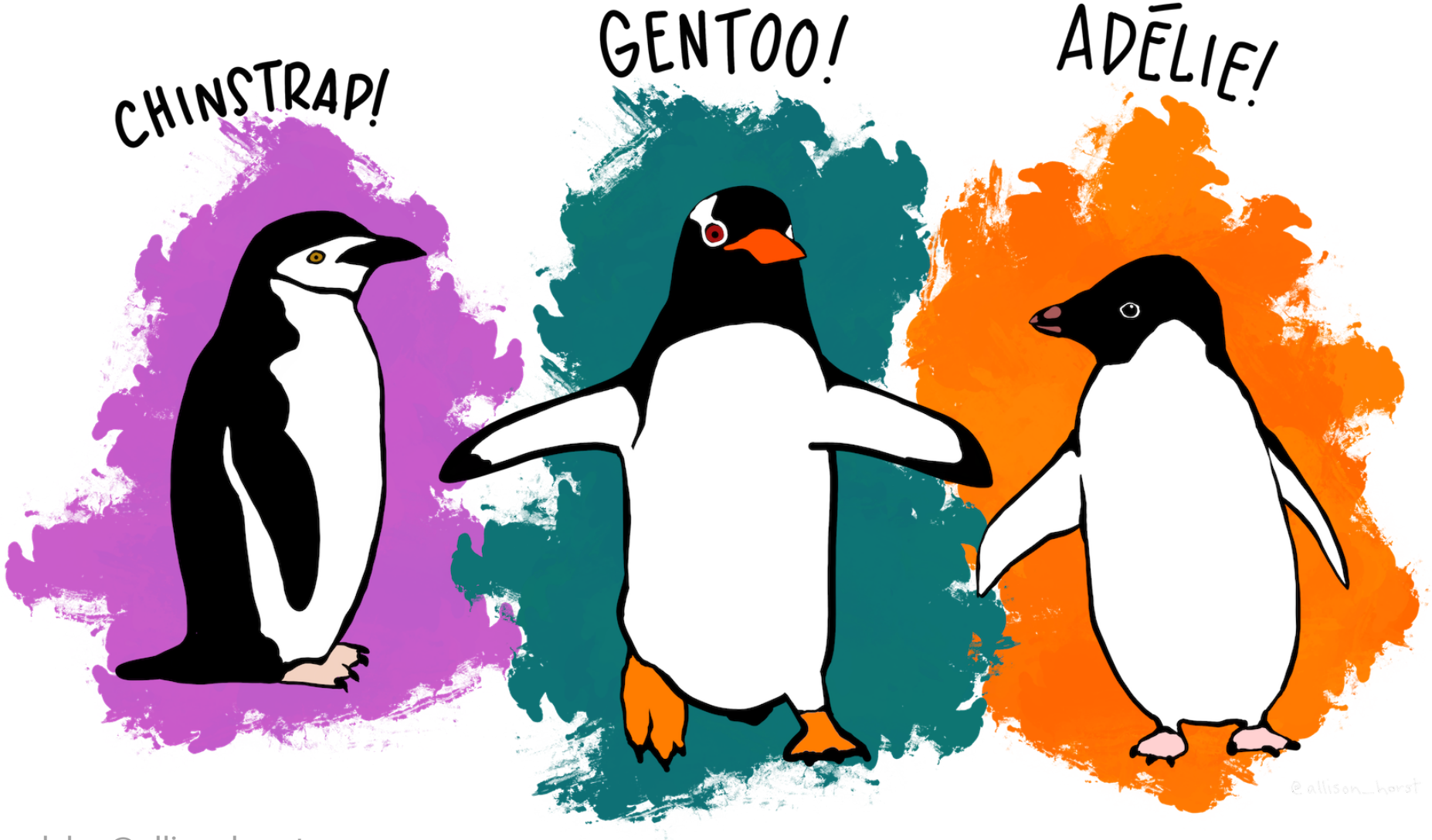
HITS

Let’s go on a trip!

HITS

Palmer Archipelago

Penguins!



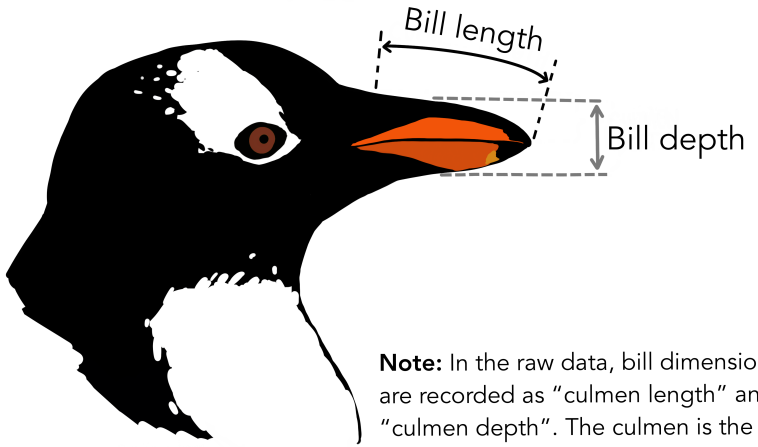
Artwork by @allisonhorst

HITS

Penguins!

```
penguins |> head() |> kable()
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007



Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

Artwork by @allisonhorst

A Grammar of Graphics

```
1 penguins
```


HITS

A Grammar of Graphics

```
penguins |>
  ggplot(aes(x = bill_length_mm, y = bill_depth_mm))
```

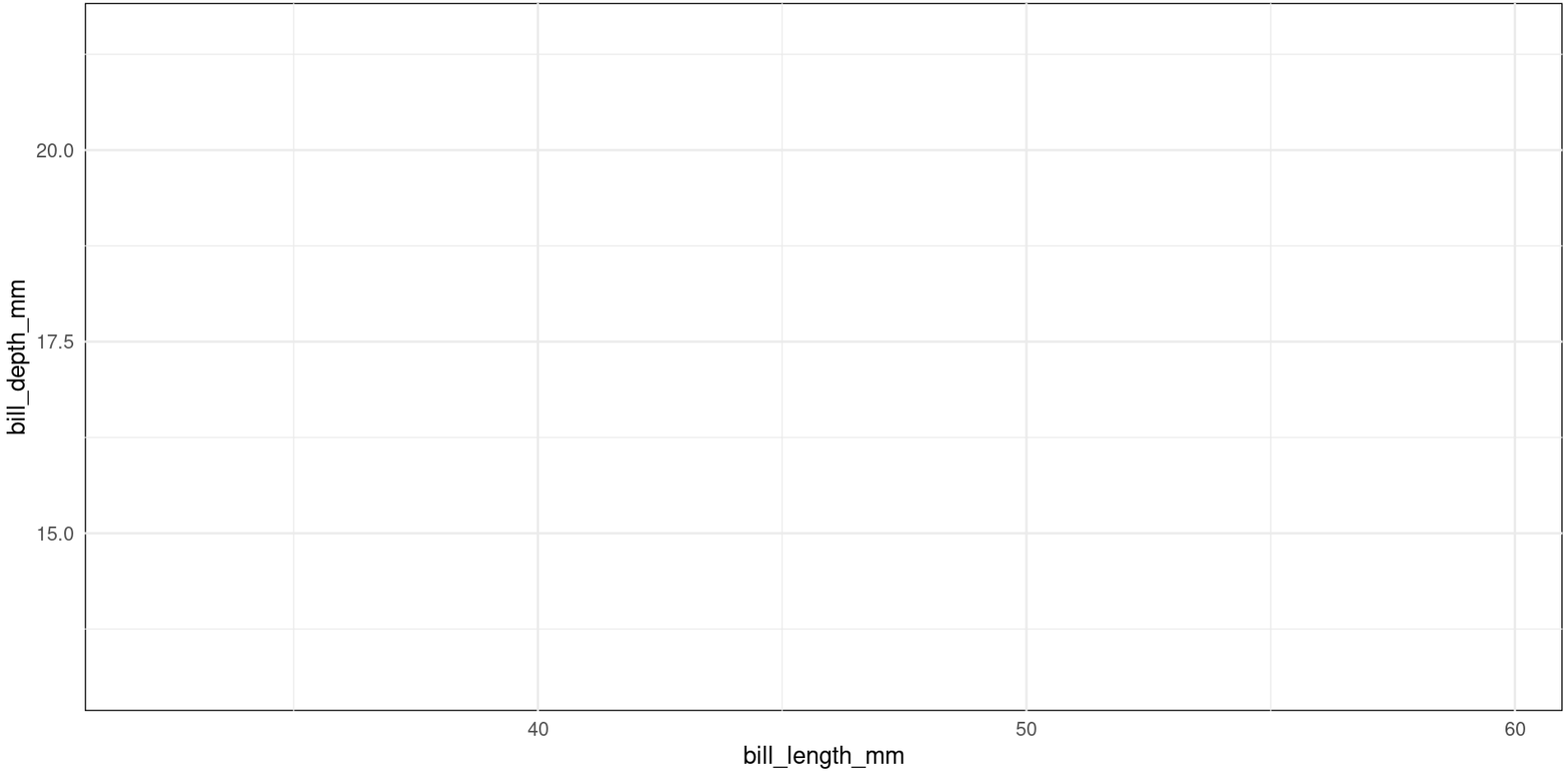
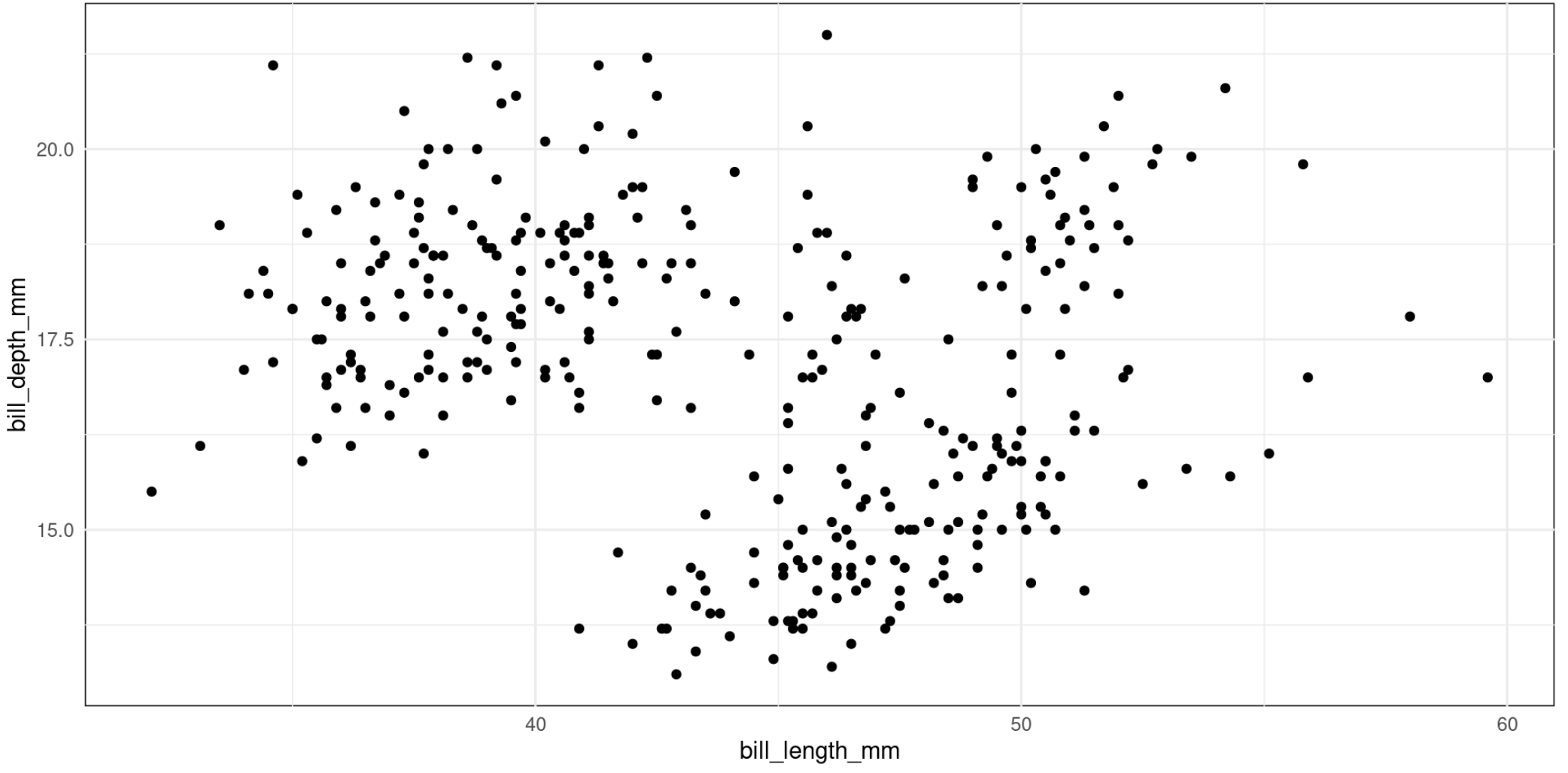


Figure 1: Penguins!

HITS

A Grammar of Graphics

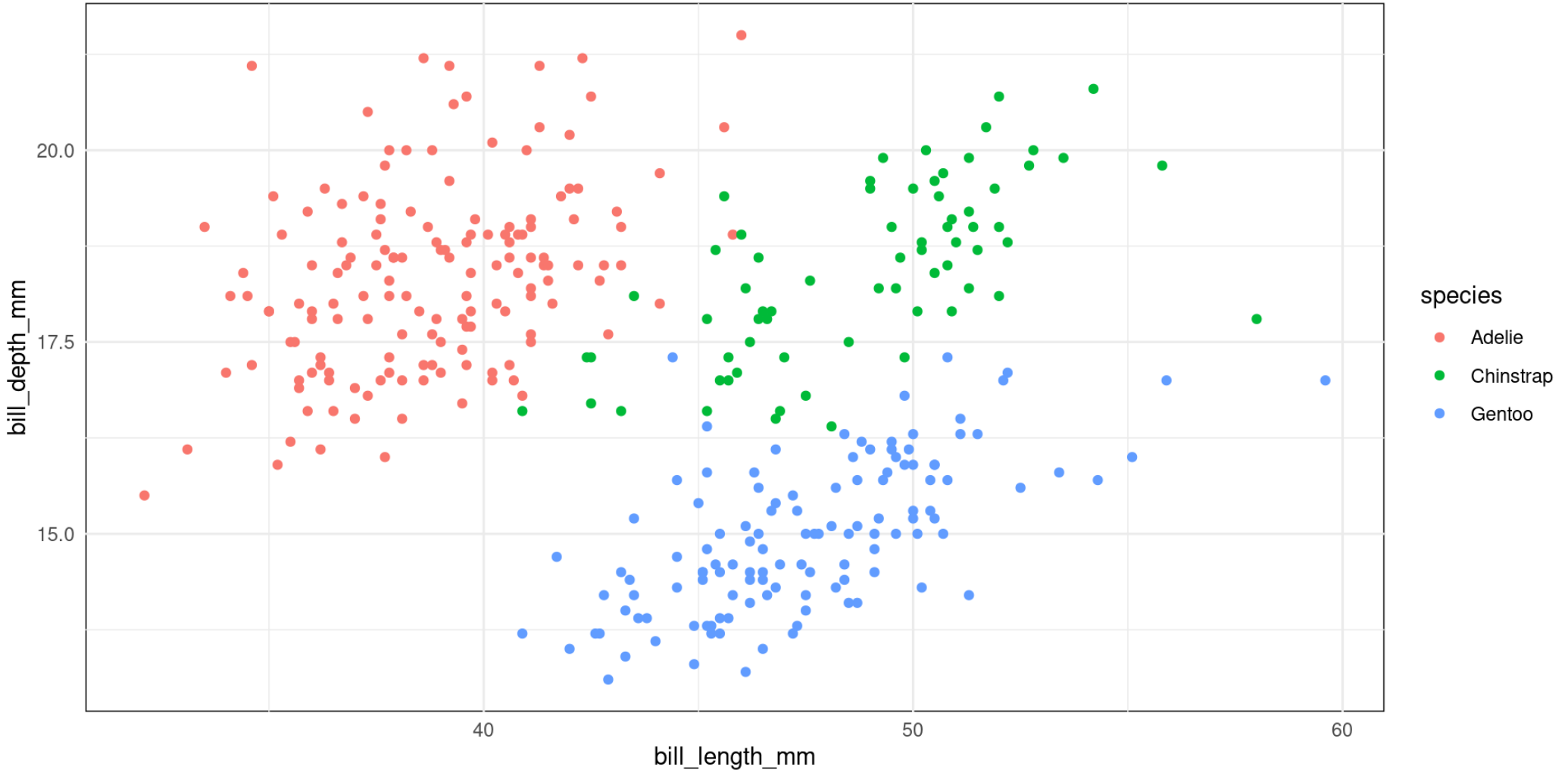
```
penguins |>
  ggplot(aes(x = bill_length_mm, y = bill_depth_mm)) +
  geom_point()
```



HITS

A Grammar of Graphics

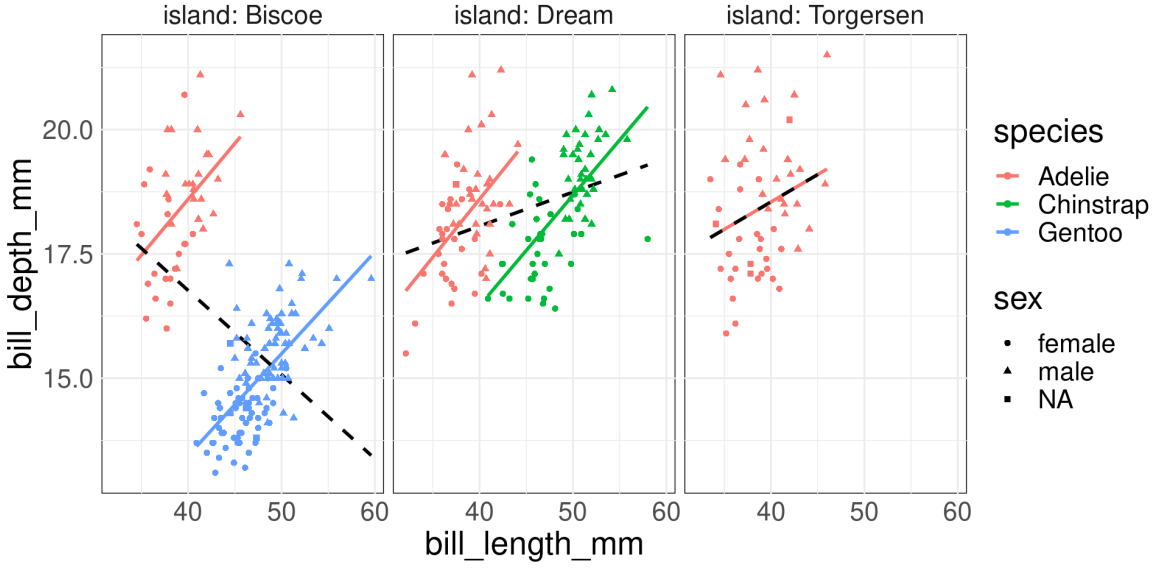
```
penguins |>
  ggplot(aes(x = bill_length_mm, y = bill_depth_mm, color = species)) +
  geom_point()
```



HITS

Elements of a Plot

```
penguins |>
  ggplot(aes(x = bill_length_mm,
             y = bill_depth_mm,
             color = species)) +
  geom_smooth(method = lm, se = FALSE) +
  geom_smooth(aes(group = island),
                  method = lm,
                  se = FALSE,
                  linetype = 2,
                  color = "black") +
  geom_point(aes(shape = sex)) +
  facet_wrap(~island, labeller=label_both) +
  scale_shape(na.value = 15) +
  theme(text = element_text(size = 20))
```

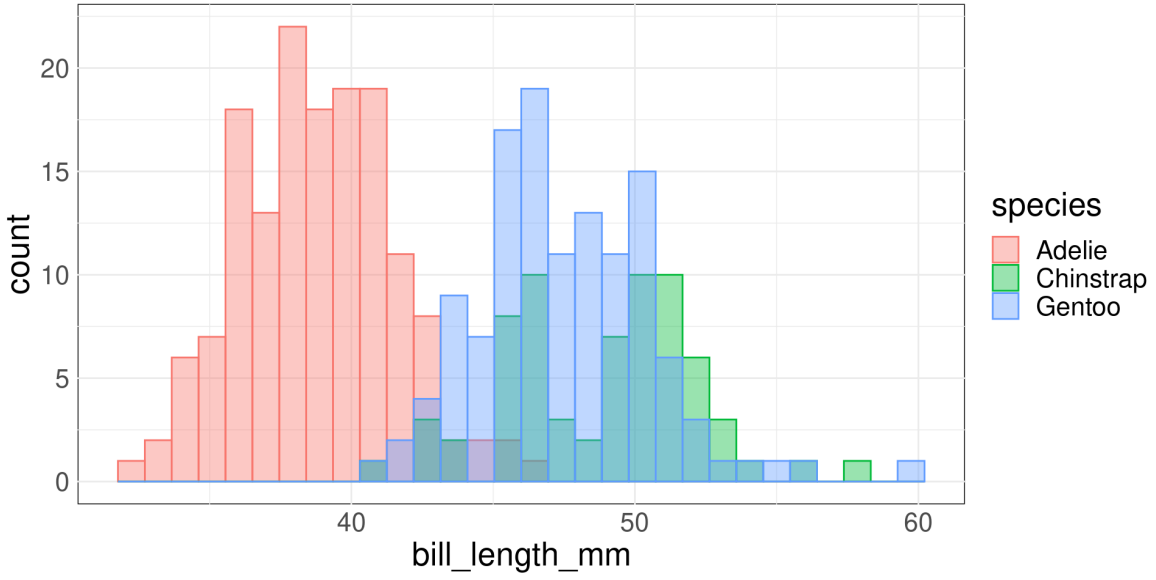


- data and aesthetic mappings
  - x/y(/z)
  - color
  - fill
  - alpha
  - group
  - shape/linetype
  - size/area
- (layers of) geometric objects
- scales
- facet specification
- statistical transformations
- the coordinate system

HITS

Elements of a Plot

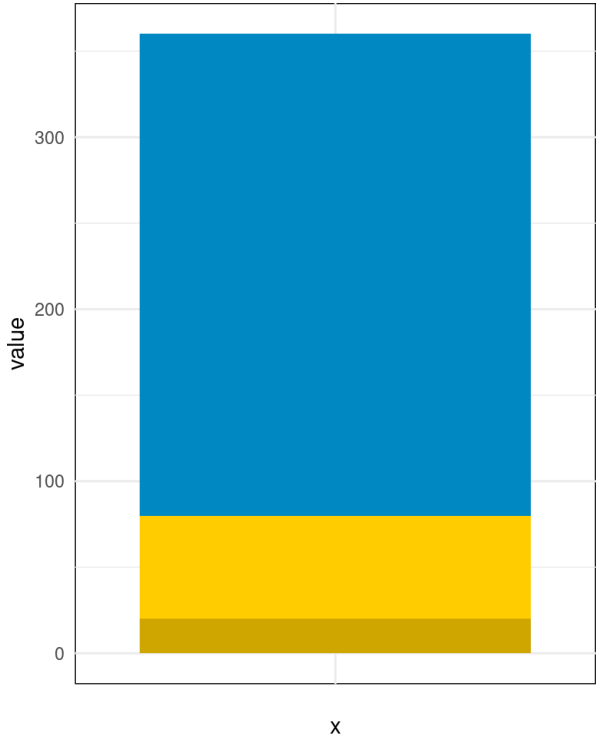
```
penguins |>
  ggplot(aes(x = bill_length_mm,
             color = species,
             fill = species)) +
  geom_histogram(alpha=0.4,
                 position = "identity") +
  theme(text = element_text(size = 20))
```



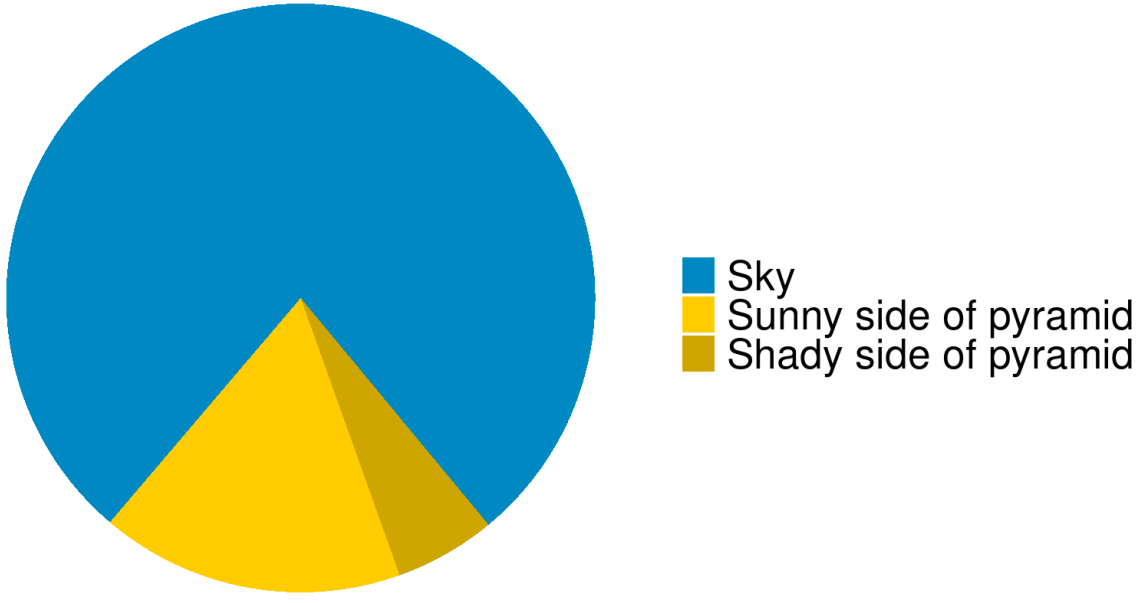
- data and aesthetic mappings
  - x/y(/z)
  - color
  - fill
  - alpha
  - group
  - shape/linetype
  - size/area
- (layers of) geometric objects
- scales
- facet specification
- statistical transformations
- the coordinate system

HITS

Elements of a Plot



A stacked bar chart



`+ coord_polar()`

Figure 2: Coordinate systems

HITS

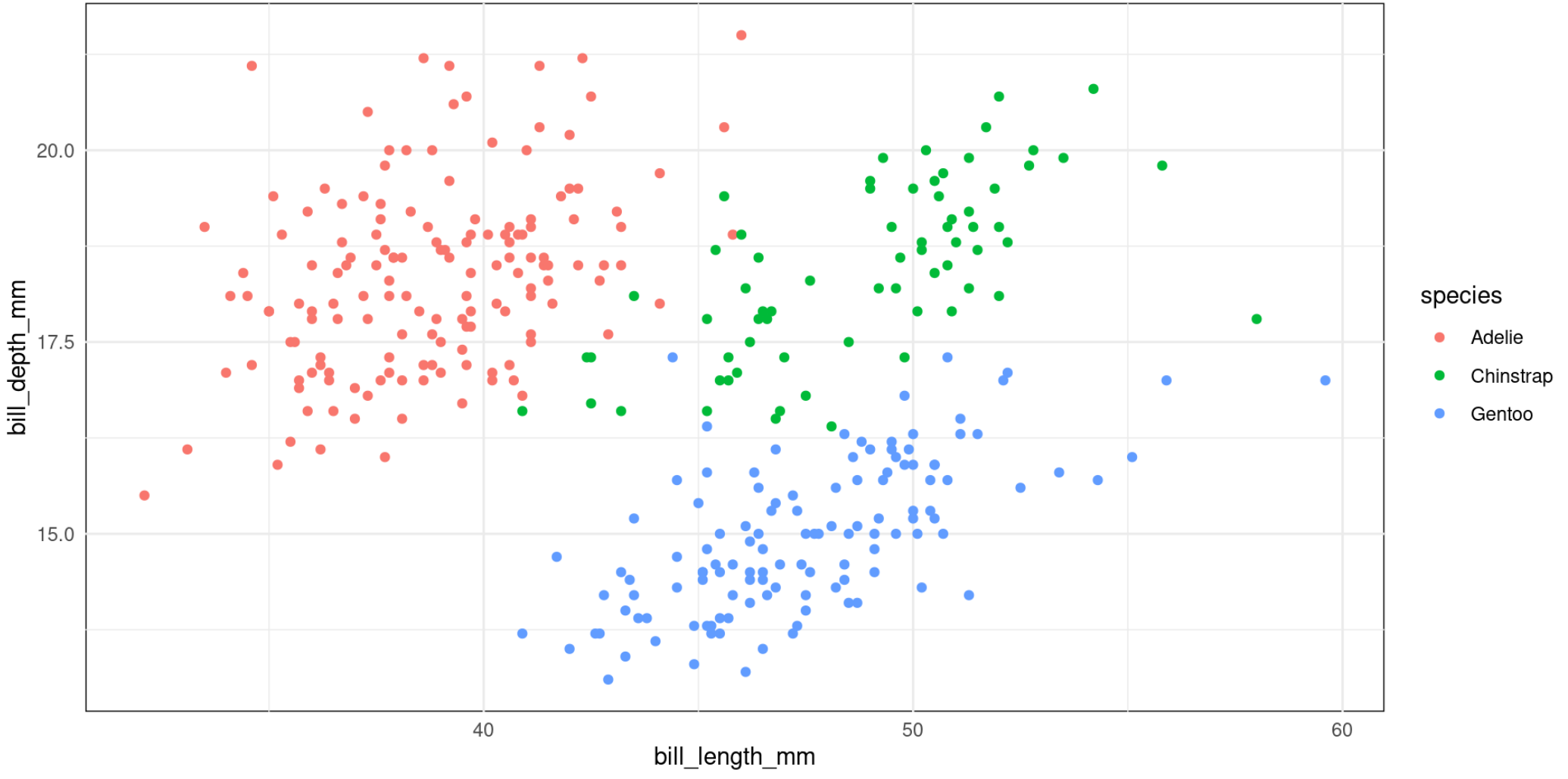
The Duality of Data Visualization

Plotting for yourself vs. plotting to **communicate**.

HITS

Consistent Colors

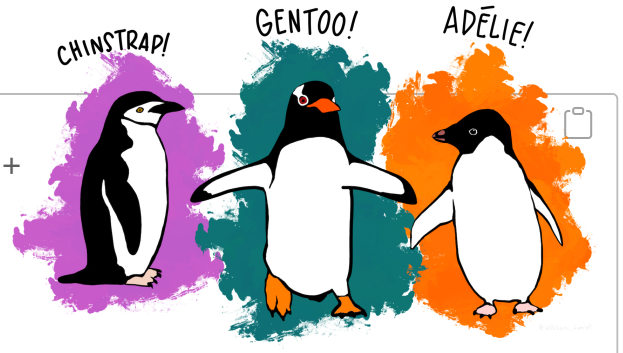
```
1 penguins |>
2   ggplot(aes(x = bill_length_mm, y = bill_depth_mm, color = species)) +
3   geom_point()
```

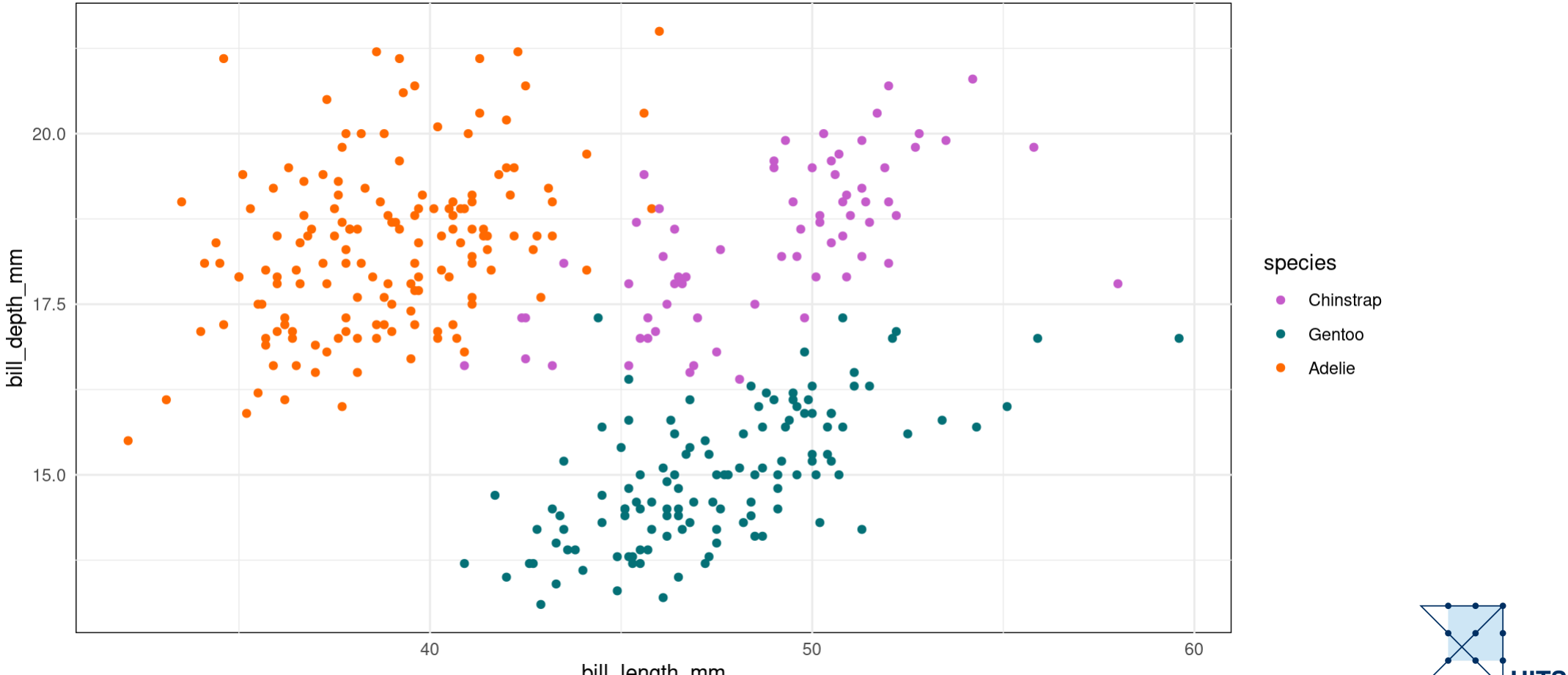


HITS

Consistent Colors

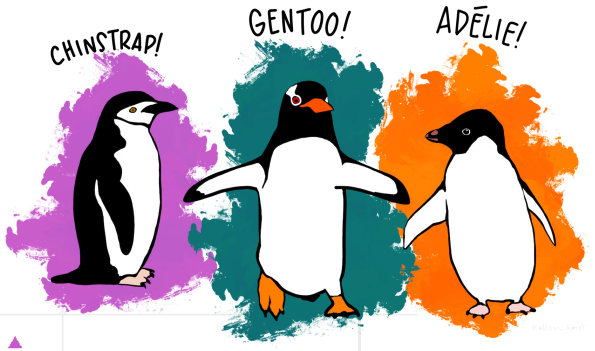
```
penguins |>
  ggplot(aes(x = bill_length_mm, y = bill_depth_mm, color = species)) +
  geom_point() +
  scale_color_manual(values = c(Chinstrap = "#c55bca",
                                Gentoo    = "#047076",
                                Adelie    = "#ff6900"))
```



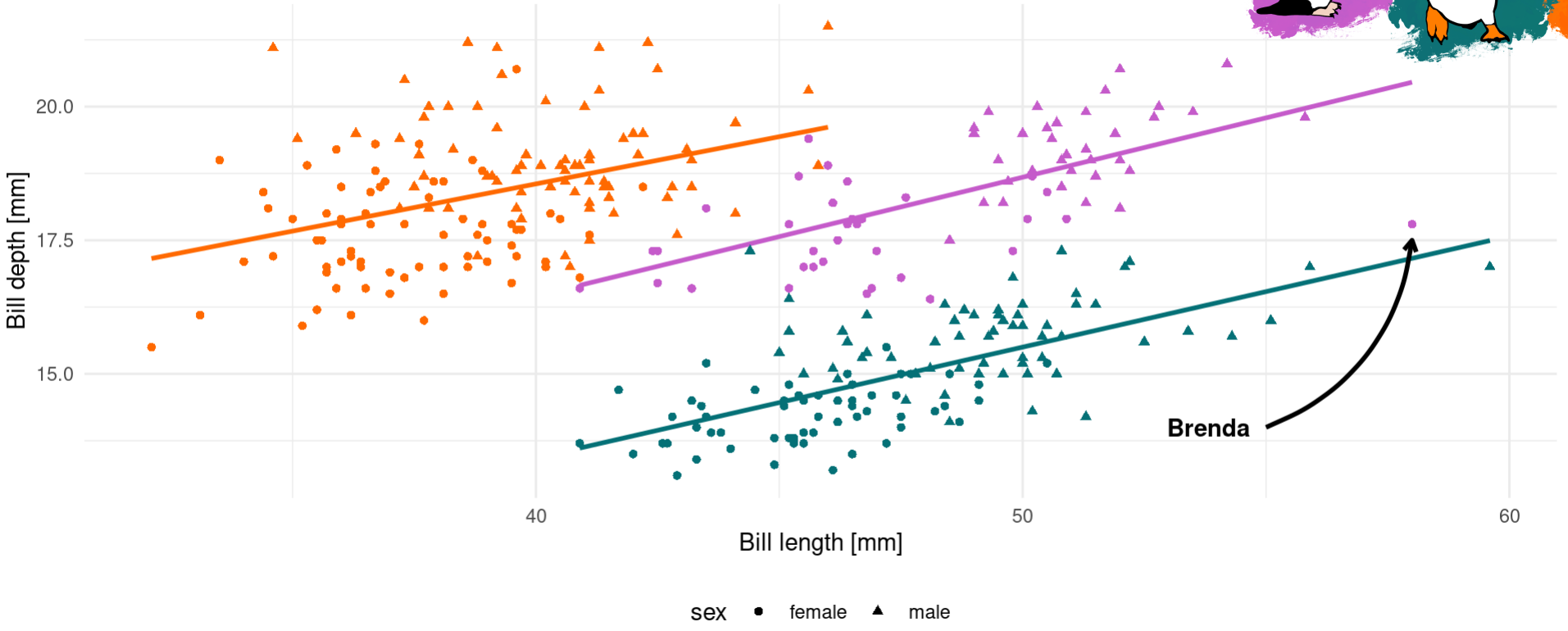


HITS

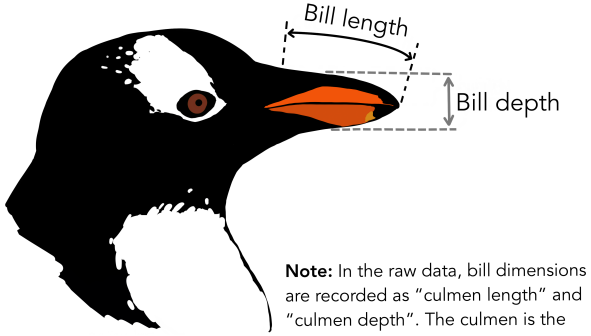
Direct Labels



One of these penguins is not like the others
Species: Chinstrap, Gentoo and Adelie



Datasource: (Horst, Hill, and Gorman 2020)

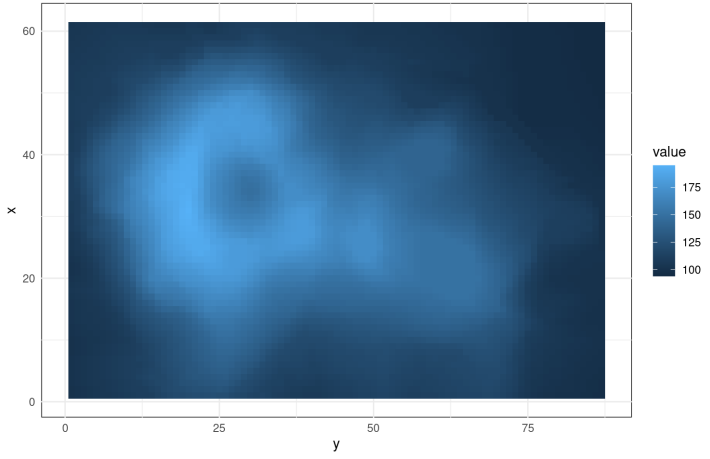


Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the

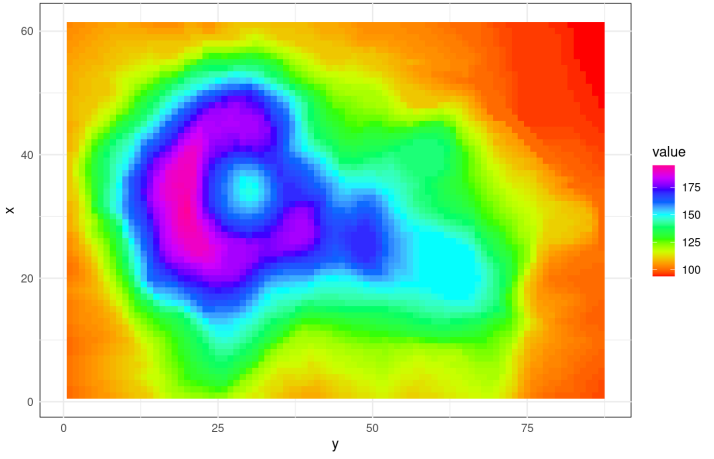
jmbuhr.de/hits-scientific-seminar-datavis

HITS

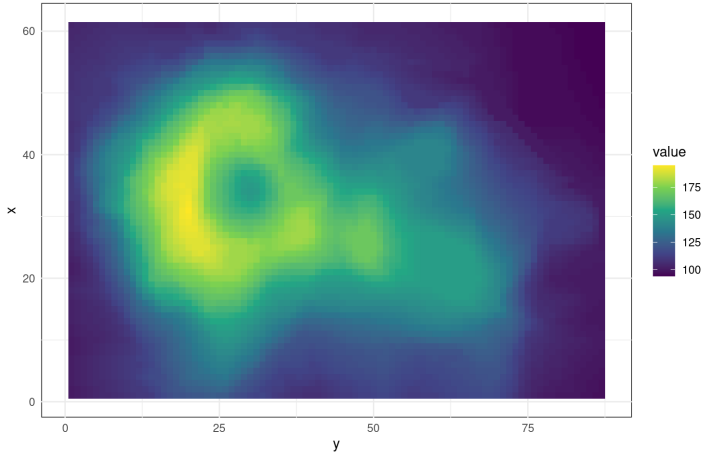
Inclusive and Truthful Colors



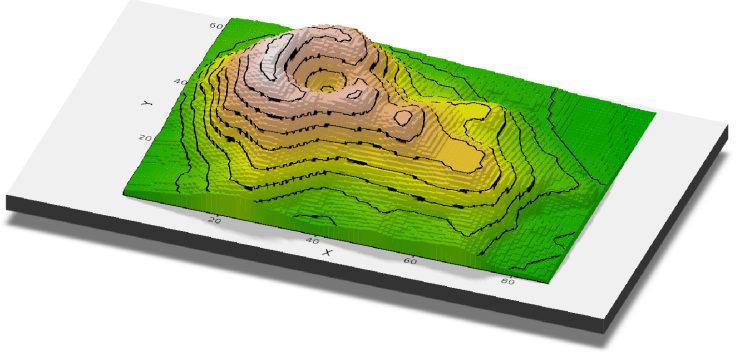
(a) Volcano dataset, default colors



(b) Rainbow colors



(c) Viridis colors

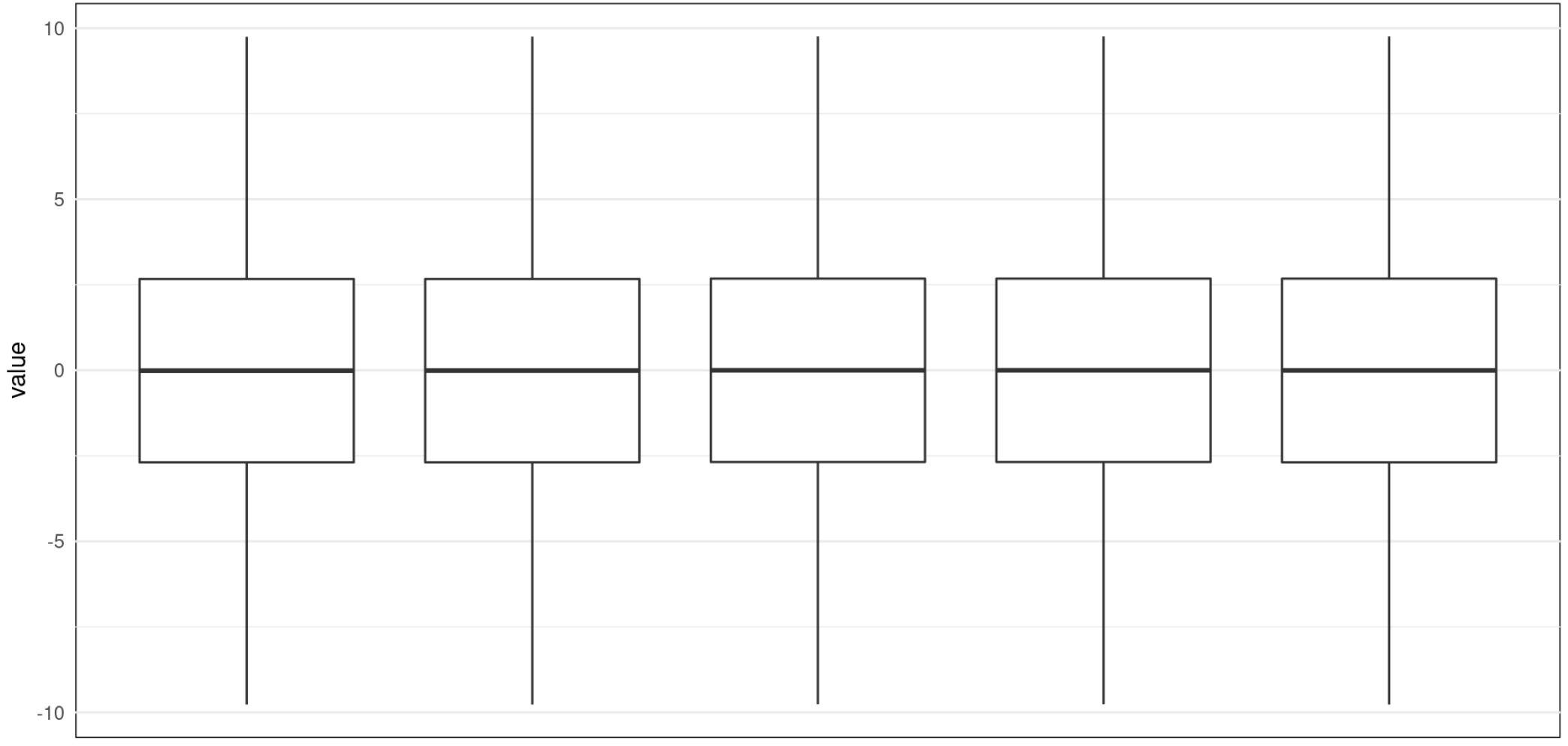


(d) Terrain colors

HITS

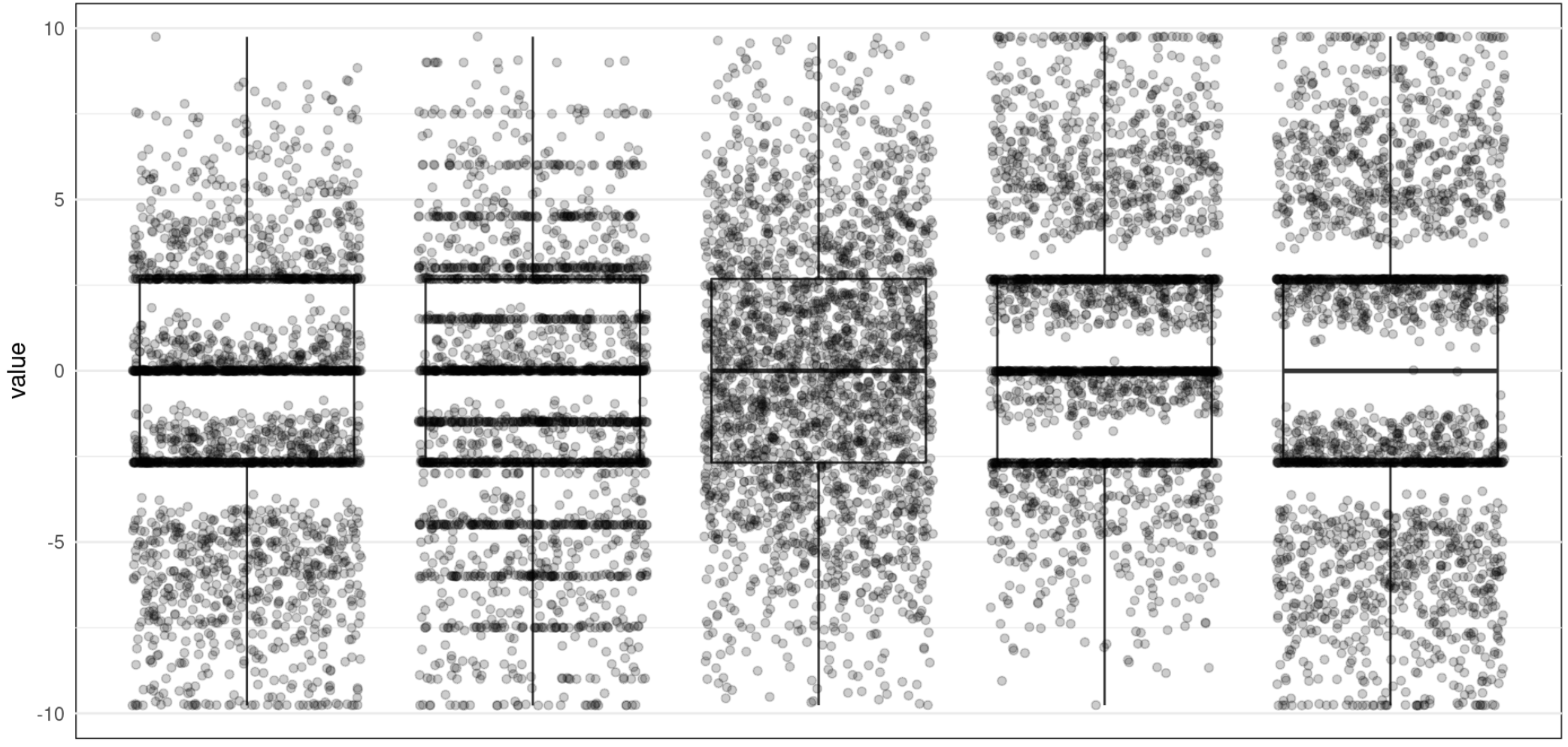
Figure 3: Your choice of color palettes matters.

Simplify, but Don’t Oversimplify



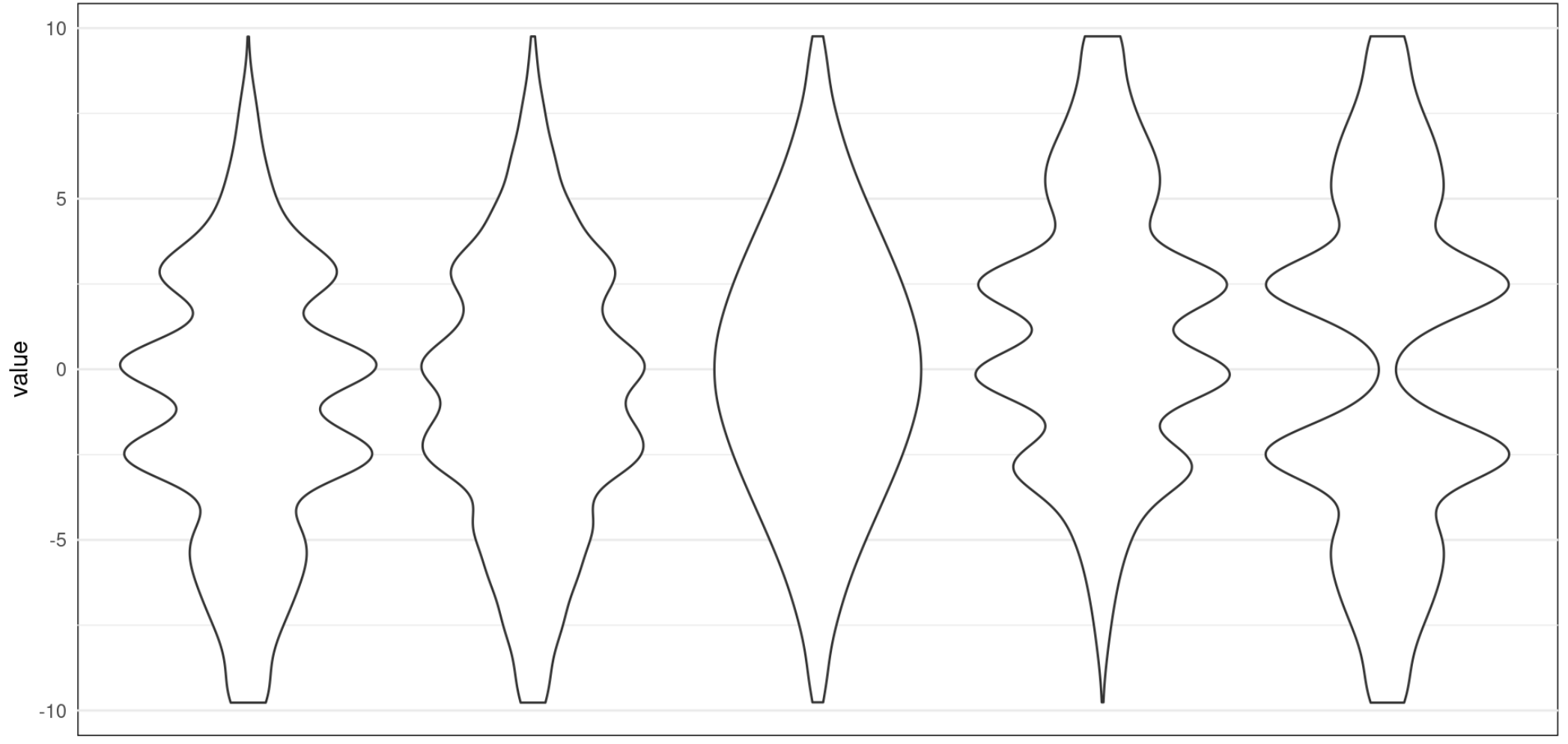
HITS

Simplify, but Don’t Oversimplify



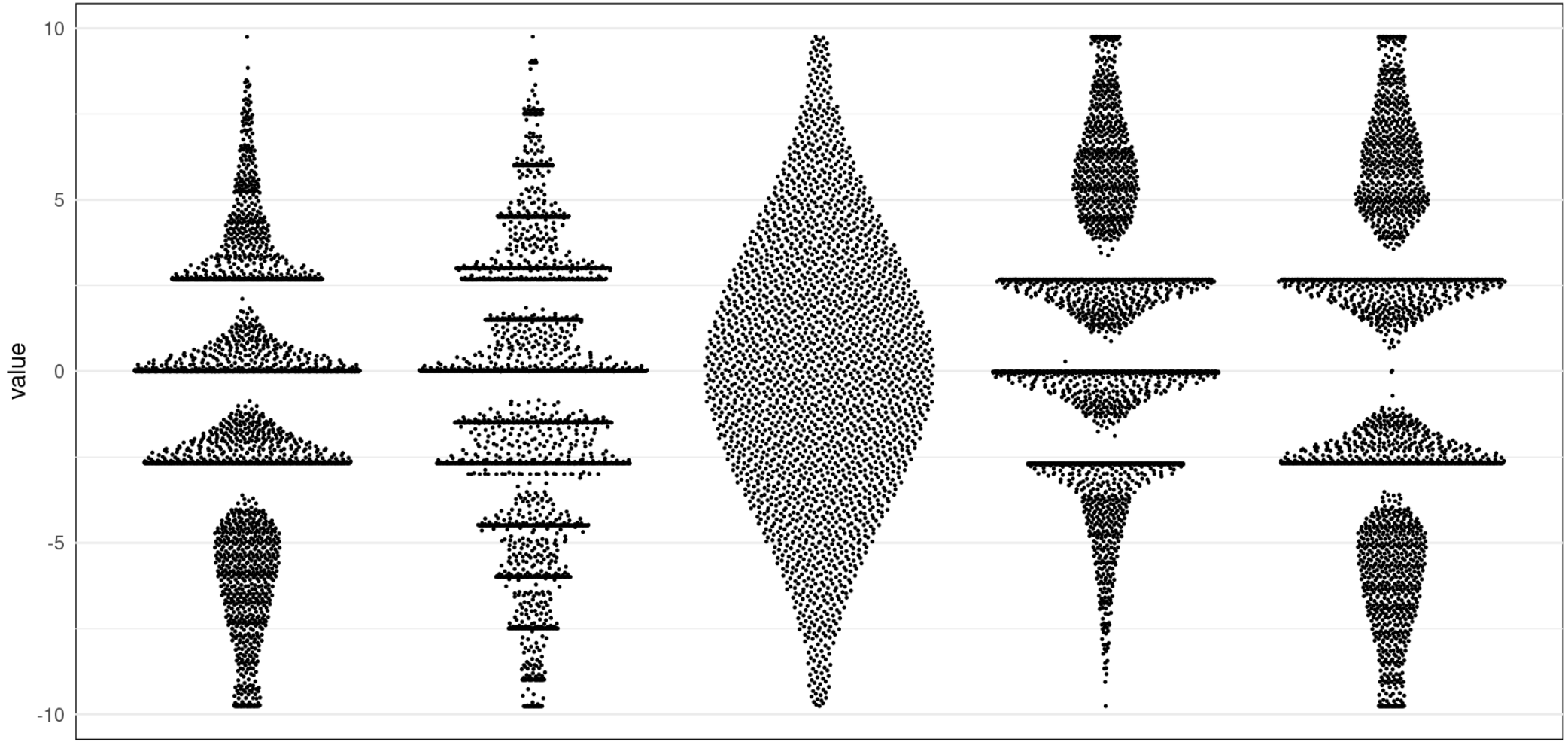
HITS

Simplify, but Don’t Oversimplify



HITS

Simplify, but Don’t Oversimplify



HITS

Simplify, but Don't Oversimplify

"Artwork by @Allison_horst" (2020)

HITS

Tidy Data

> “Happy families are all alike;
> every unhappy family is unhappy in its own way”
> — Leo Tolstoy

Tidy Data

- by Hadley Wickham (Wickham 2014)
- derived from Codd's Third Normal Form (Codd 1972)
- allows the direct mapping between aesthetics and variables

Every row is a penguin

Every column is a property of penguins

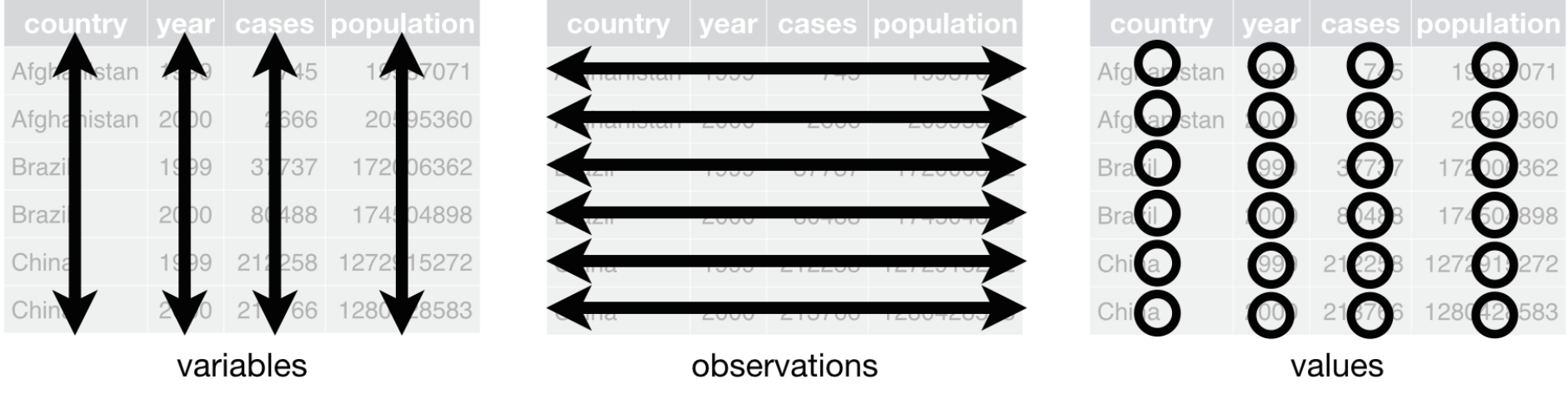


Figure from Grolemund and Wickham (2017)

HITS

Tidy Data

- not always the most space effective
- not being in a tidy format is not necessarily bad
- but tidy works well with ggplot
- and allows you to think about your data more effectively
- similarities: feature matrix / vector in machine learning are in a tidy format!
- knowing the rules allows you to break them with purpose

Break the Rules

```
penguins |>
  select(species, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g) |>
  head() |>
  kable()
```

species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_
Adelie	39.1	18.7	181	375
Adelie	39.5	17.4	186	380
Adelie	40.3	18.0	195	325
Adelie	NA	NA	NA	NA
Adelie	36.7	19.3	193	345
Adelie	39.3	20.6	190	365

Break the Rules

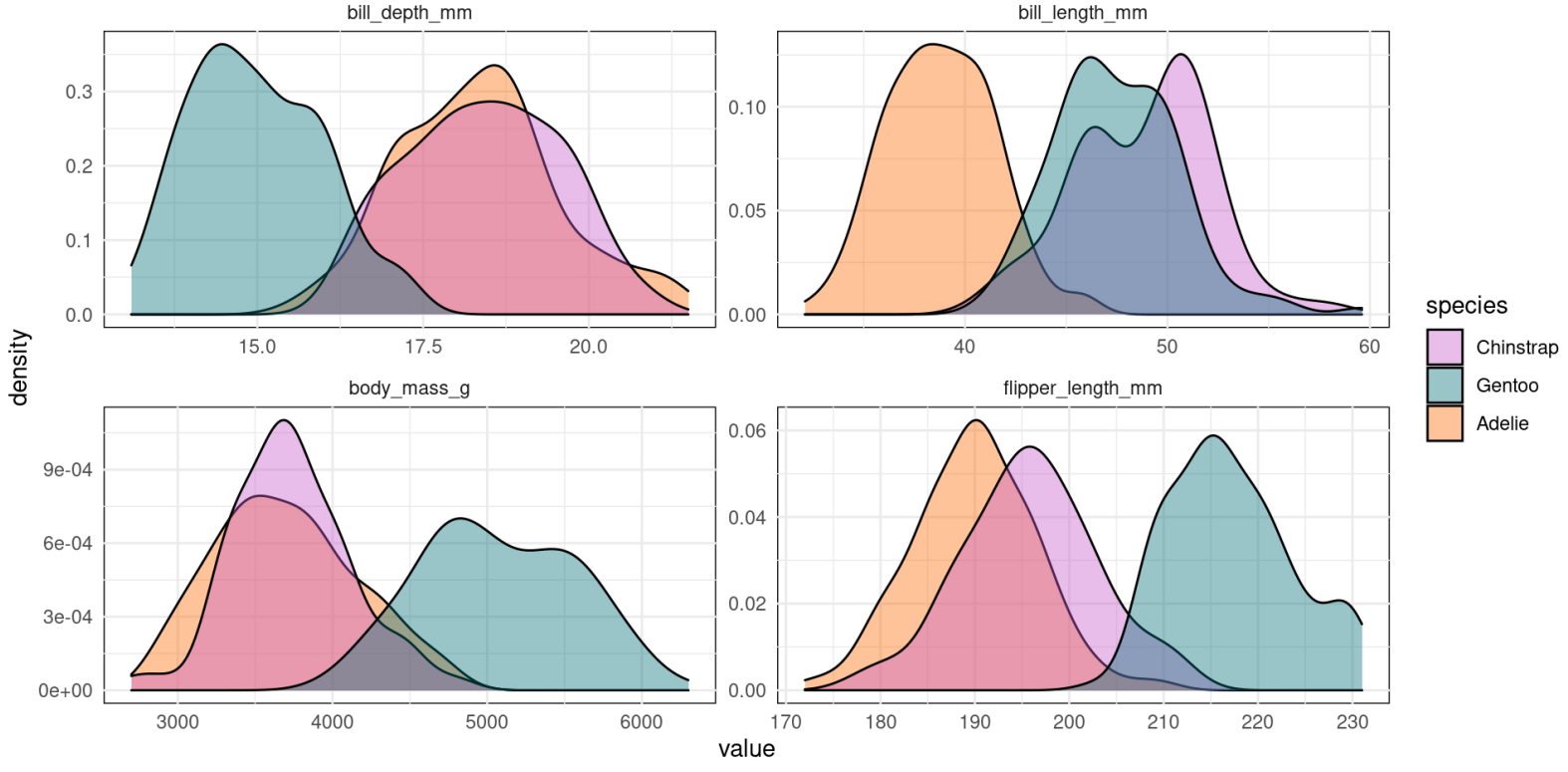
```
penguins |>
  select(species, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g) |>
  pivot_longer(-species) |>
  head() |>
  kable()
```

species	name	value
Adelie	bill_length_mm	39.1
Adelie	bill_depth_mm	18.7
Adelie	flipper_length_mm	181.0
Adelie	body_mass_g	3750.0
Adelie	bill_length_mm	39.5
Adelie	bill_depth_mm	17.4

Every penguin is now **spread across 4 rows**.

Break the Rules

```
penguins |>
  select(species, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g) |>
  pivot_longer(-species) |>
  ggplot(aes(value, fill = species)) +
  geom_density(alpha = 0.4) +
  facet_wrap(~name, scales="free") +
  scale_fill_manual(values = c(Chinstrap = "#c55bca",
                               Gentoo    = "#047076",
                               Adelie    = "#ff6900"))
```



HITS

Other Implementations of Grammars

Python: plotnine

```
import warnings
warnings.simplefilter("ignore")
from plotnine import *

print(
ggplot(r.penguins, aes('bill_length_mm', 'bill_depth_mm', color='species'))
 + geom_point()
);
```

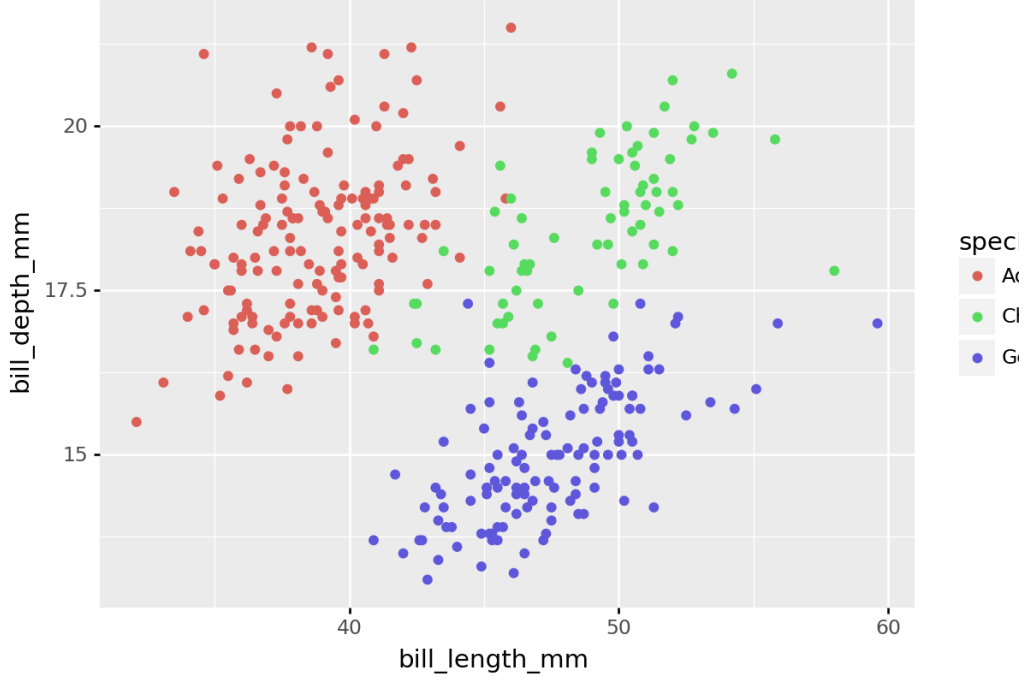


Figure 4: A plotnine plot.

HITS

Python: Seaborn (nextgen)

```
1 import warnings
2 warnings.simplefilter("ignore")
3 import seaborn.objects as so
4
5 (
6 so.Plot(r.penguins, "bill_length_mm", "bill_depth_mm", color="species")
7     .add(so.Scatter(fillalpha=1))
8 ).show();
```

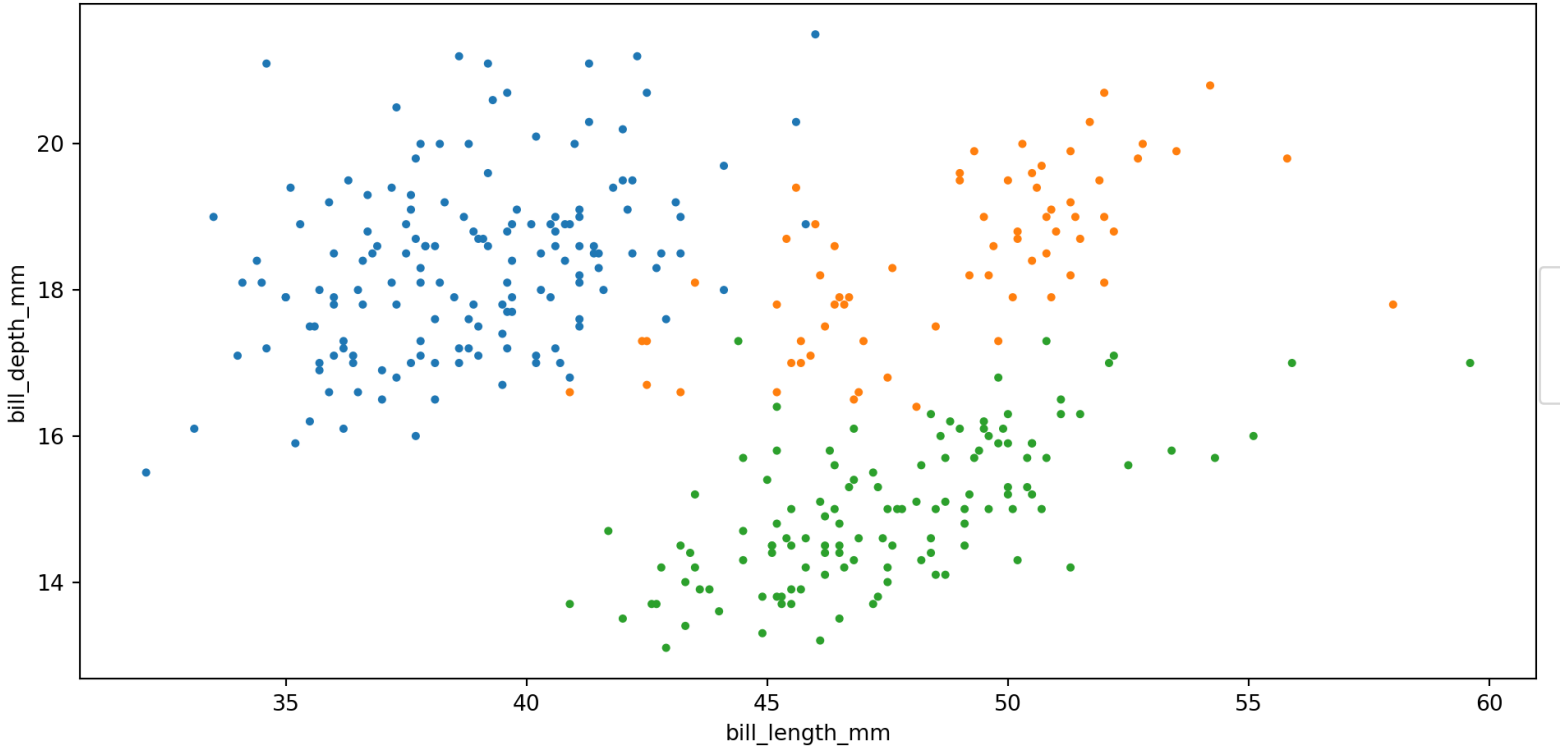
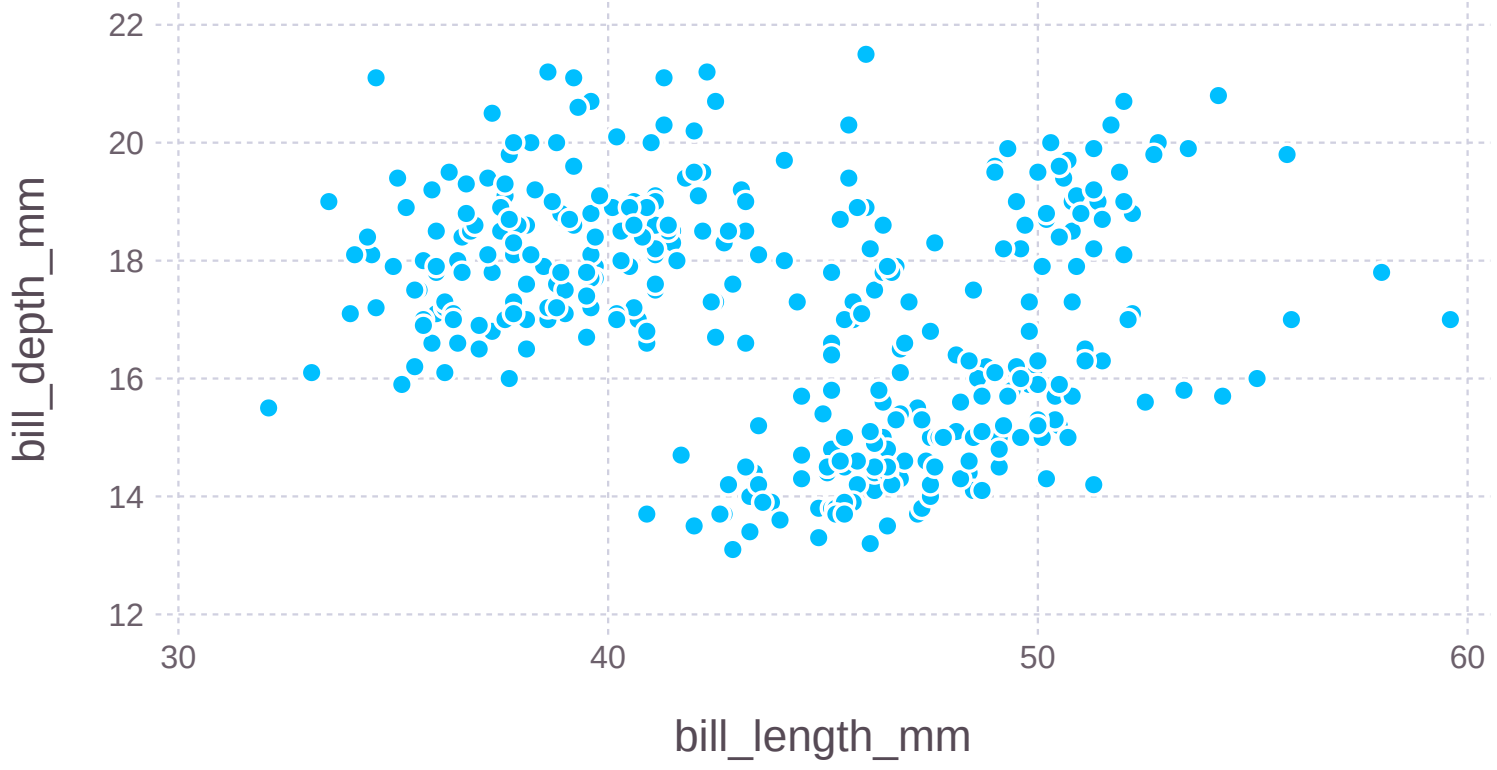


Figure 5: A seaborn plot.

HITS

Julia: Gadfly

```julia
using Gadfly, PalmerPenguins, DataFrames
penguins = DataFrame(PalmerPenguins.load());
p = plot(penguins, x=:bill_length_mm, y=:bill_depth_mm, Geom.point);
img = SVG("img/julia_plot.svg", 14cm, 8cm)
draw(img, p)
```



Gadfly plot

HITS

The greater context

The greater context

> “Somehow all plotting libraries converge into some object based language.”
> — @tsuname

Objects are just **implementation details**

Jannik’s predictions for the future

Jannik’s predictions for the future

HITS

Jannik's predictions for the future

- modal systems \(\gt\) monolithic structures
- composition \(\gt\) inheritance
  - `python`: method chaining

```
1 baseplot(data).add(Scatter()).add(...)
```

  - R: operator overloading (nearly `|>`)

```
1 ggplot(data) + geom_scatter() + theme()
```

  - `haskell`: function composition

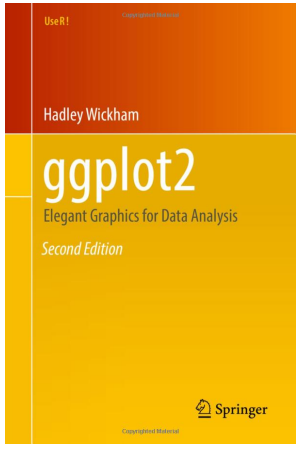
```
1 addScatter :: Plot -> Plot
2 addTheme . addScatter $ basePlot data
3 -- read: addTheme after  addScatter applied to ...
```

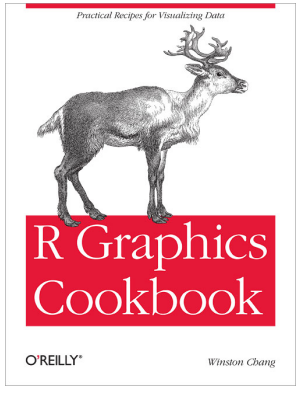
- declarative \(\gt\) imperative

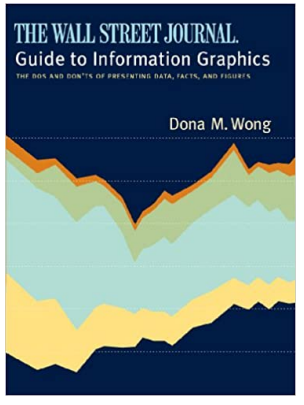
Resources

Books and other resources

- ggplot2: Elegant Graphics for Data Analysis (Pedersen, n.d.)
- R Graphics Cookbook (Chang, n.d.)
- “How Charts Lie” – Alberto Cairo (Cairo 2019)
- “Guide to Information Graphics” – Dona M. Wong (Wong 2013)
- Tidy Data (Wickham 2014)







Thank You!

jmbuhr.de/hits-scientific-seminar-datavis

HITS

Thank You!

Want to learn more?

See you on the otter slide.

jmbuhr

jmbuhr.de

jannikbuhr

jmbuhr

Slides:
github.com/jmbuhr/hits-scientific-seminar-datavis

HITS

References

Allaire, J. J., Charles Teague, Carlos Scheidegger, Yihui Xie, and Christophe Dervieux. 2022. *Quarto* (version 0.3). https://doi.org/10.5281/zenodo.5960048.

"Artwork by @Allison_horst." 2020. https://github.com/allisonhorst/stats-illustrations.

Cairo, Alberto. 2019. *How Charts Lie: Getting Smarter about Visual Information*. Illustrated edition. New York: W. W. Norton & Company.

Chang, Winston. n.d. *R Graphics Cookbook, 2nd Edition*. https://r-graphics.org/.

Codd, E F. 1972. "Further Normalization of the Data Base Relational Model," 34.

Grolemund, Garrett, and Hadley Wickham. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 1st edition. Sebastopol, CA: O'Reilly Media.

Horst, Allison, Alison Hill, and Kristen Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. Manual.

Pedersen, and Thomas Lin, Danielle Navarro. n.d. *Welcome | Ggplot2*. https://ggplot2-book.org/index.html.

Wickham, Hadley. 2010. "A Layered Grammar of Graphics." *Journal of Computational and Graphical Statistics* 19 (1): 3–28. https://doi.org/10.1198/jcgs.2009.07098.

———. 2014. "Tidy Data." *Journal of Statistical Software* 59 (1): 1–23. https://doi.org/10.18637/jss.v059.i10.

Wilkinson, Leland, D. Wills, D. Rope, A. Norton, and R. Dubbs. 2005. *The Grammar of Graphics*. 2nd edition. New York: Springer.

Wong, Dona M. 2013. *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures*. Reprint edition. New York: W. W. Norton & Company.