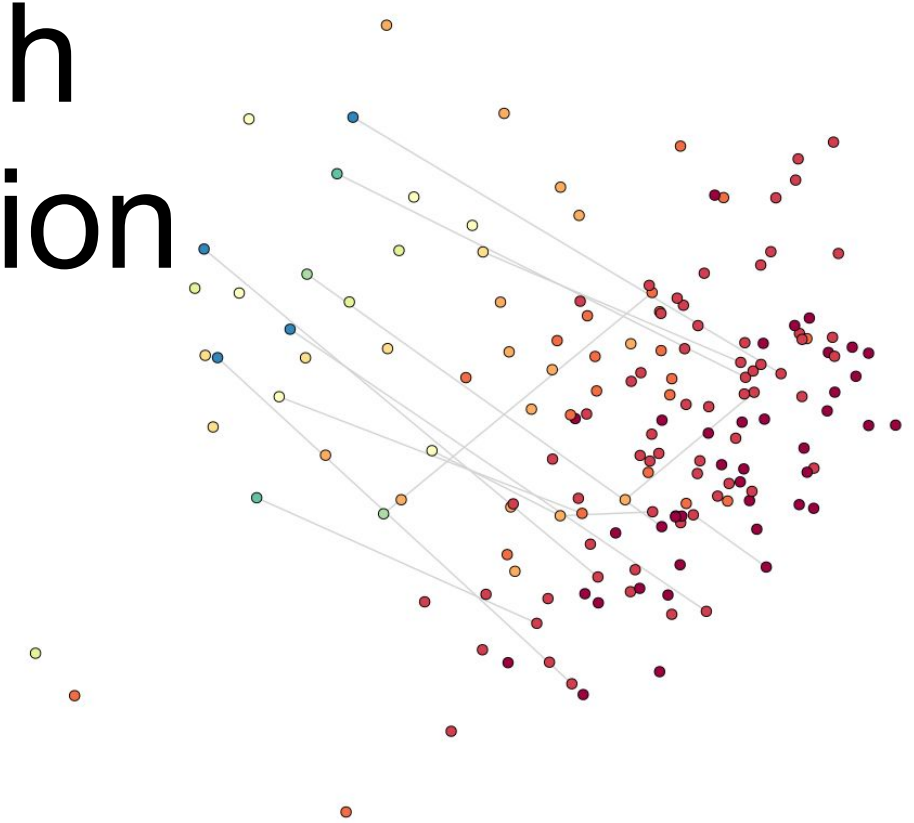
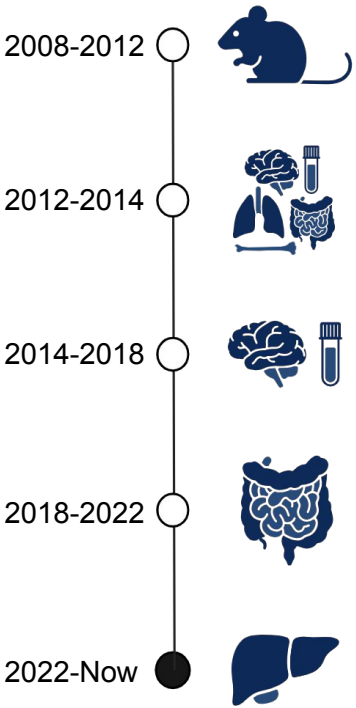
Communicating Biology Through Data Visualization



Rachel Edgar

University Health Network - University of Toronto
Postdoctoral Fellow - MacParland and Bader labs

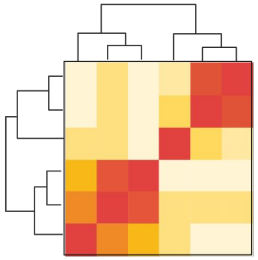
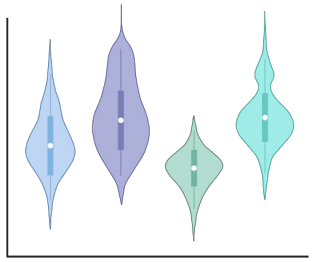
About Me: Rachel Edgar

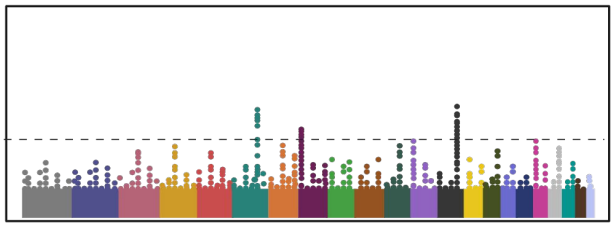
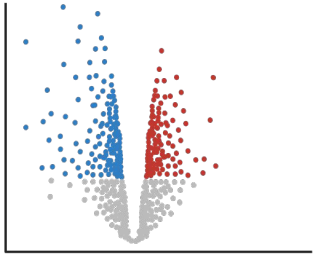
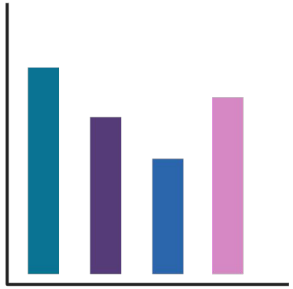
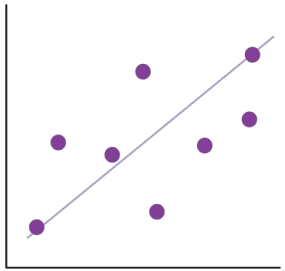
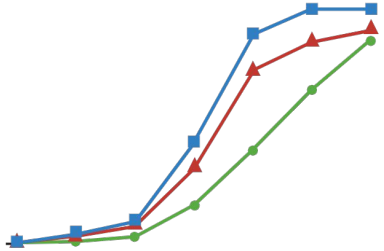


Somatic mutations in the cerebellum of harlequin mice	University of Western Ontario BSc - Kathleen Hill
Meta-analysis of Human DNA Methylation Data	University of British Columbia MSc - Paul Pavlidis
Investigating the use of blood as surrogate for brain tissue in epigenetics	University of British Columbia Research Assistant - Michael Kobor
Multiomics Of The Human Gut Toward Insight Into Childhood Inflammatory Bowel	University of Cambridge - EMBL-EBI PhD - Daniel Zerbino
Single-cell gene expression of human liver	University Health Network Postdoc - Sonya MacParland and Gary Bader

Above all else show the data.

- Edward R. Tufte
The Visual Display of Quantitative Information

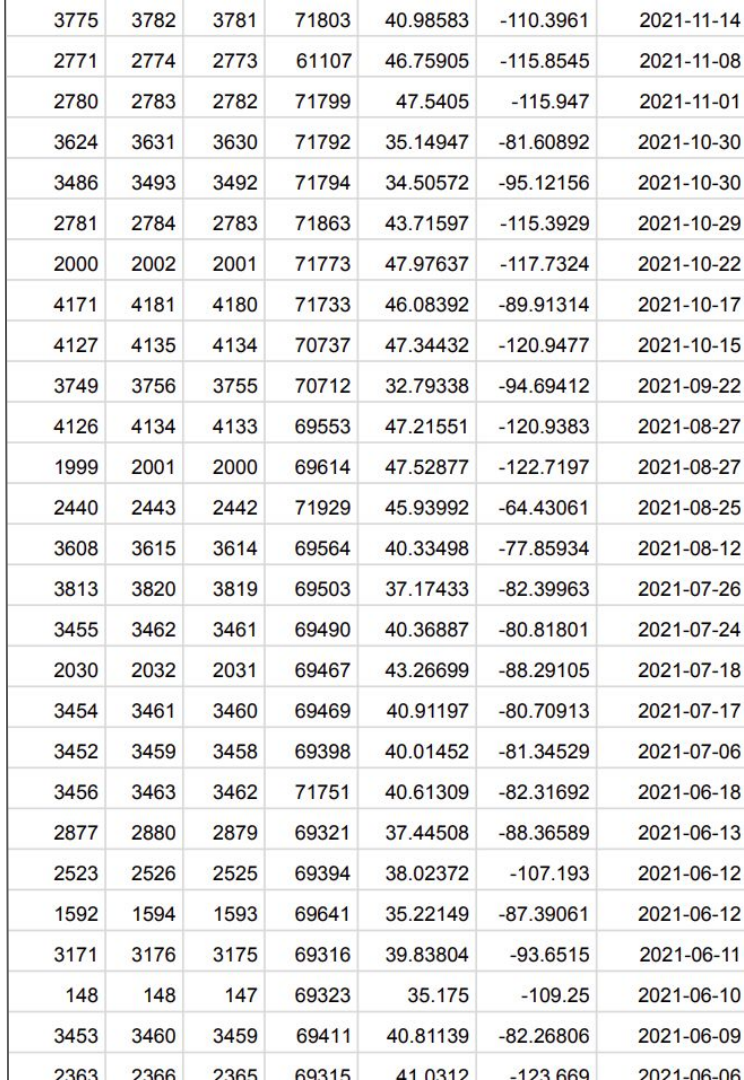



Data visualizations are powerful tools

- Data is often incomprehensible in its raw form

- Visualization allows you to:
  - **Troubleshoot** your analysis
  - **Understand** data for yourself
  - **Generate** hypothesis
  - **Communicate** your new understanding to others



3775	3782	3781	71803	40.98583	-110.3961	2021-11-14
2771	2774	2773	61107	46.75905	-115.8545	2021-11-08
2780	2783	2782	71799	47.5405	-115.947	2021-11-01
3624	3631	3630	71792	35.14947	-81.60892	2021-10-30
3486	3493	3492	71794	34.50572	-95.12156	2021-10-30
2781	2784	2783	71863	43.71597	-115.3929	2021-10-29
2000	2002	2001	71773	47.97637	-117.7324	2021-10-22
4171	4181	4180	71733	46.08392	-89.91314	2021-10-17
4127	4135	4134	70737	47.34432	-120.9477	2021-10-15
3749	3756	3755	70712	32.79338	-94.69412	2021-09-22
4126	4134	4133	69553	47.21551	-120.9383	2021-08-27
1999	2001	2000	69614	47.52877	-122.7197	2021-08-27
2440	2443	2442	71929	45.93992	-64.43061	2021-08-25
3608	3615	3614	69564	40.33498	-77.85934	2021-08-12
3813	3820	3819	69503	37.17433	-82.39963	2021-07-26
3455	3462	3461	69490	40.36887	-80.81801	2021-07-24
2030	2032	2031	69467	43.26699	-88.29105	2021-07-18
3454	3461	3460	69469	40.91197	-80.70913	2021-07-17
3452	3459	3458	69398	40.01452	-81.34529	2021-07-06
3456	3463	3462	71751	40.61309	-82.31692	2021-06-18
2877	2880	2879	69321	37.44508	-88.36589	2021-06-13
2523	2526	2525	69394	38.02372	-107.193	2021-06-12
1592	1594	1593	69641	35.22149	-87.39061	2021-06-12
3171	3176	3175	69316	39.83804	-93.6515	2021-06-11
148	148	147	69323	35.175	-109.25	2021-06-10
3453	3460	3459	69411	40.81139	-82.26806	2021-06-09
2363	2366	2365	69315	41.0312	-123.669	2021-06-06

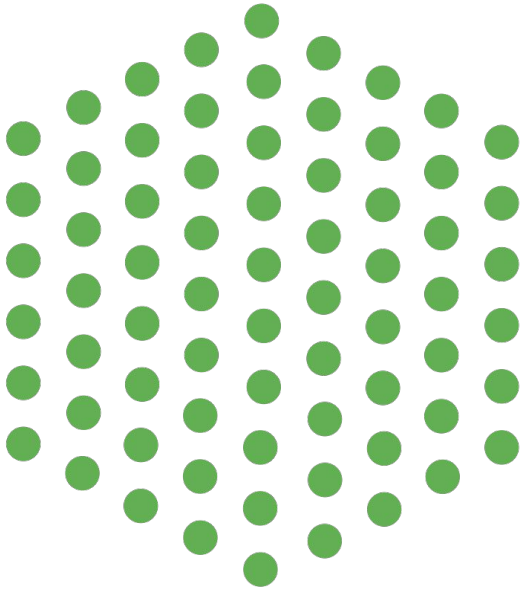
We create figures so that **humans** understand our data

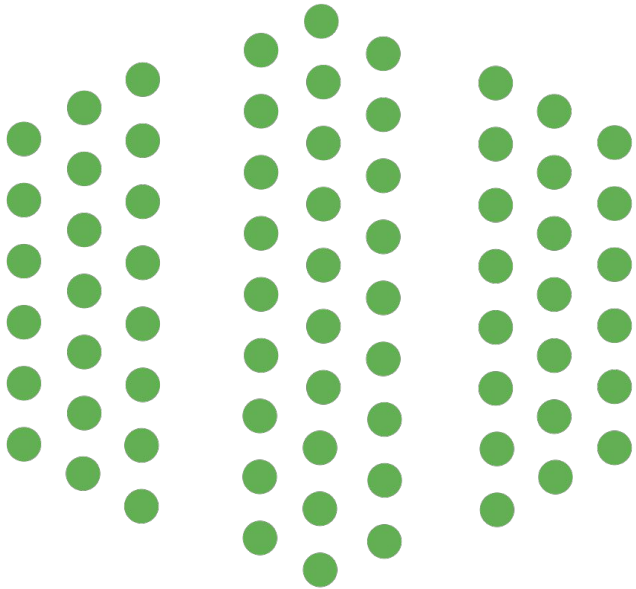
Gestalt principles of grouping describe human perception

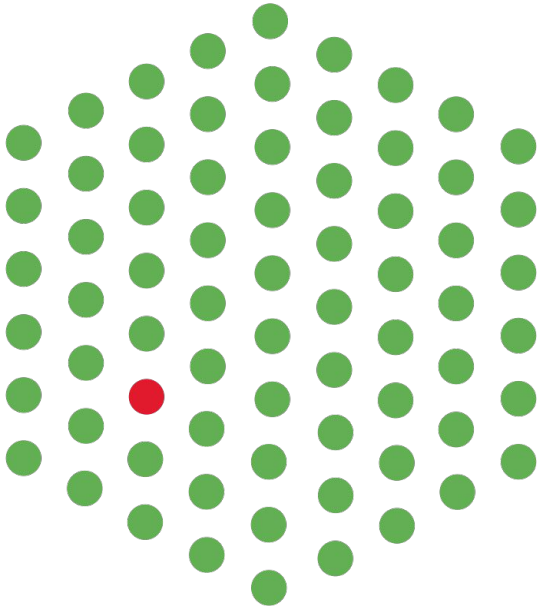
How do humans perceive?

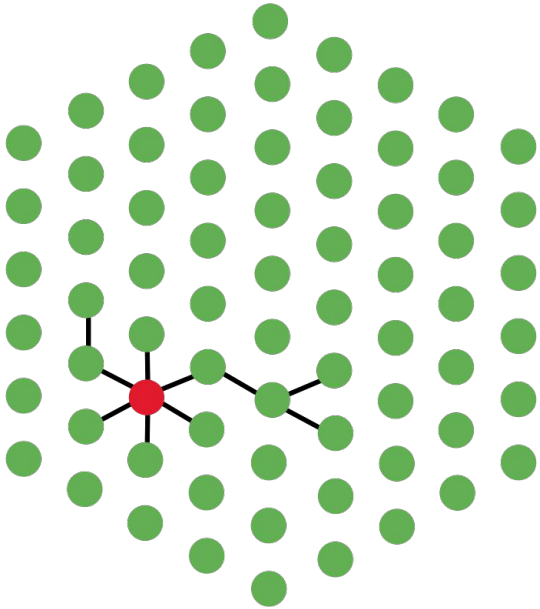
Humans naturally perceive objects as organized patterns and objects - law of *Prägnanz*

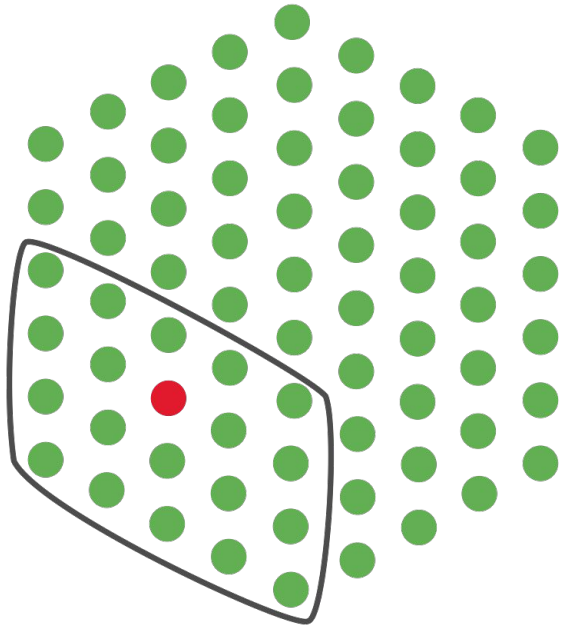
Similarity **Proximity** **Connection** **Enclosure**

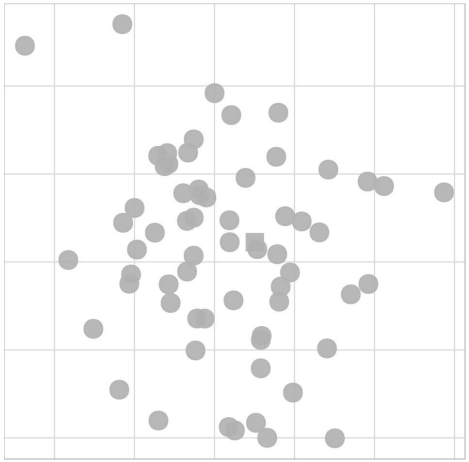


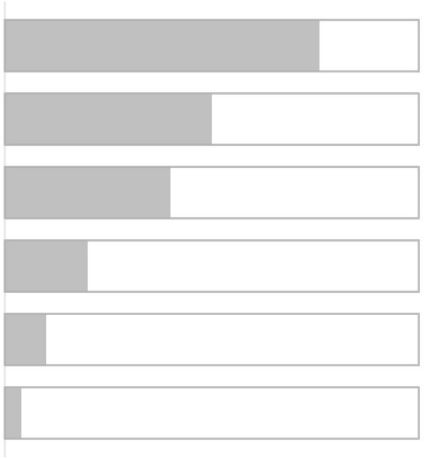




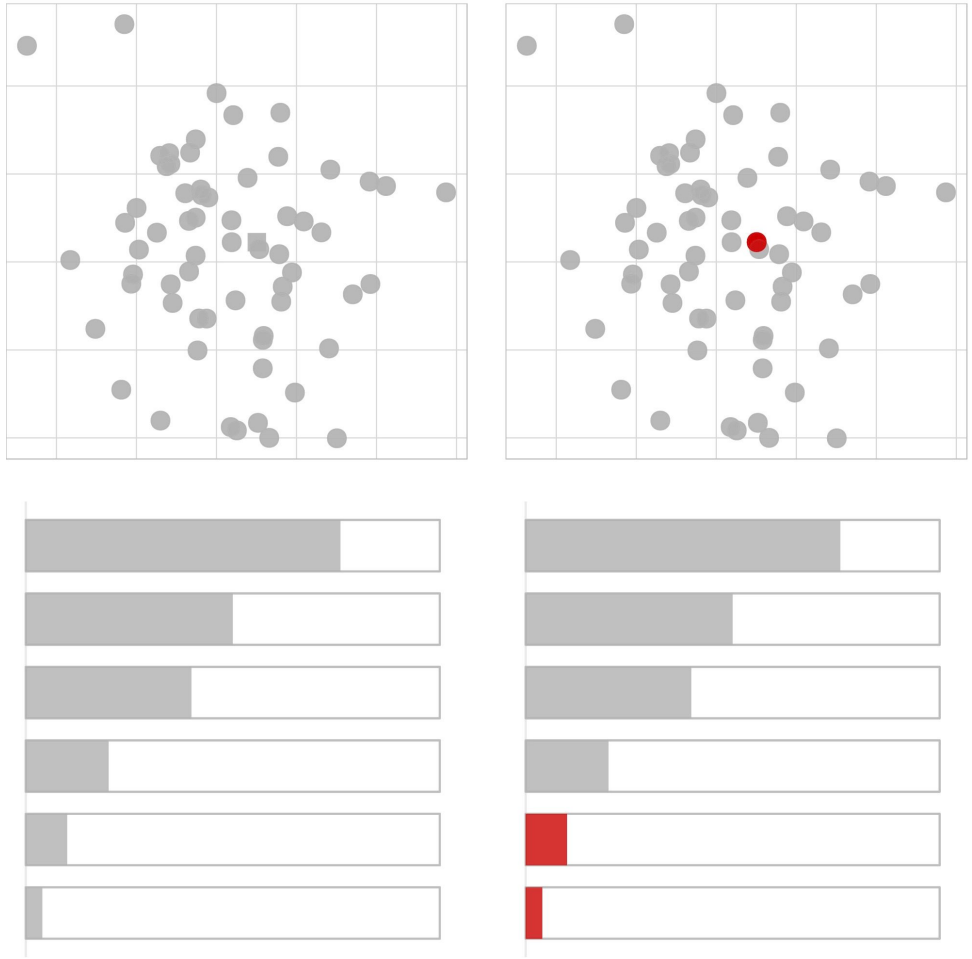




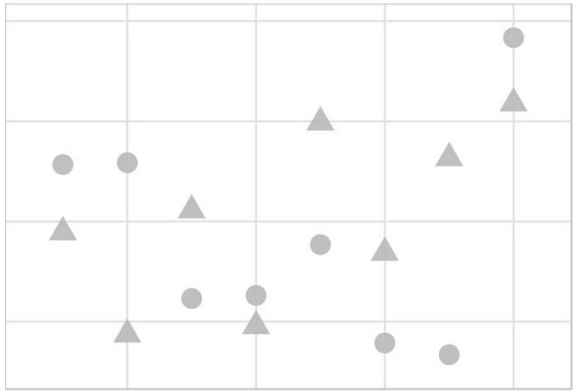




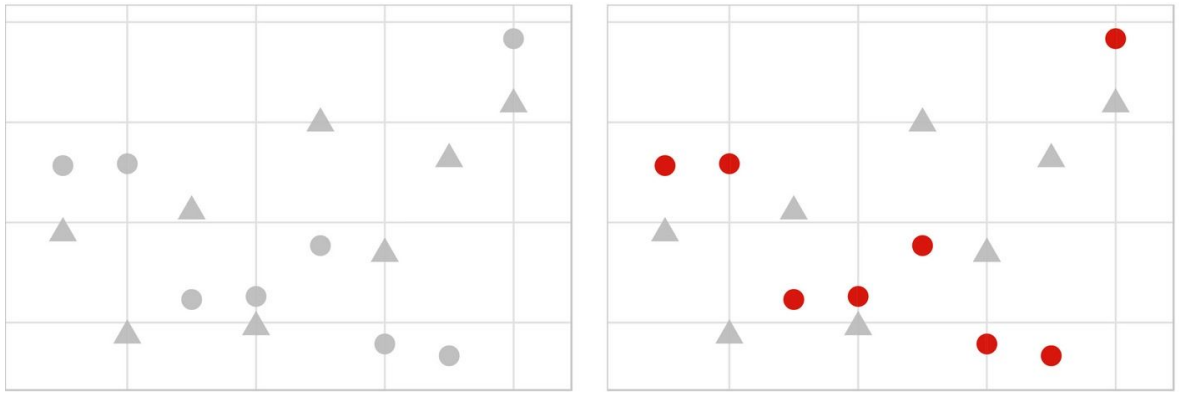
Vandemeulebroecke et al. *CPT* 2019



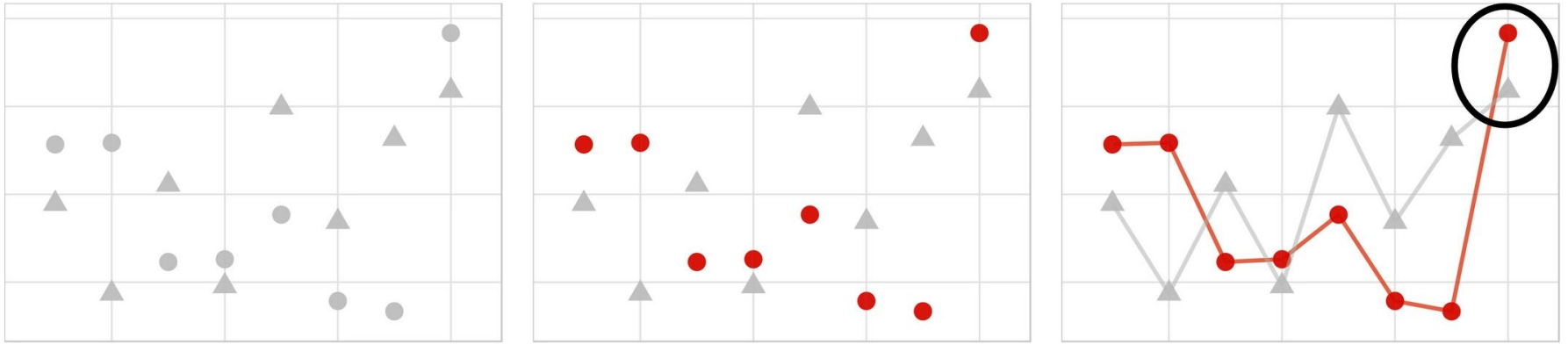
Vandemeulebroecke et al. *CPT* 2019



Vandemeulebroecke et al. *CPT* 2019



Vandemeulebroecke et al. *CPT* 2019

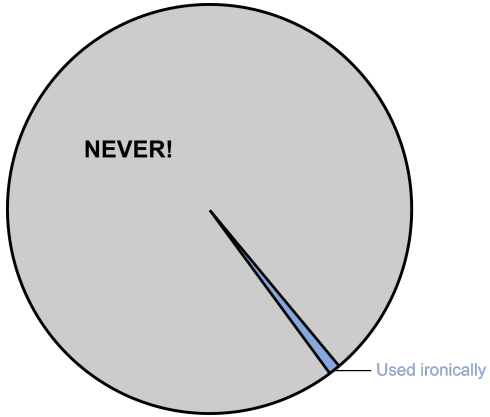


Vandemeulebroecke et al. *CPT* 2019

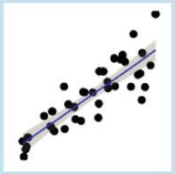
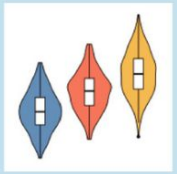
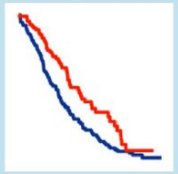
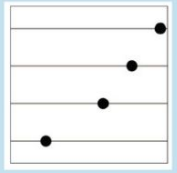
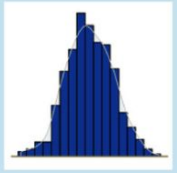
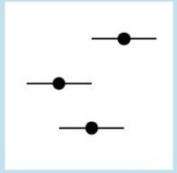
Effective Visual Communication

1. Have a defined purpose
2. Show the data clearly
    a. No pie charts
3. Make the message obvious

When is it OK to use a pie chart?



What is the right type of plot for my data?

Deviation	Correlation	Ranking	Distribution	Evolution	Part-to-whole	Magnitude
Chg. from baseline	Scatter plot	Horizontal bar chart	Boxplot	Kaplan Meier	Stacked bar chart	Vertical bar chart
						
Waterfall	Heat map	Dotplot	Histogram	Line plot	Tree map	Forest plot
						

Vandemeulebroecke et al. *CPT* 2019

Graphs can be just for you or for communication

Analysis graphs

- Just for you
- See patterns
- Help design next questions or decide on better visualizations

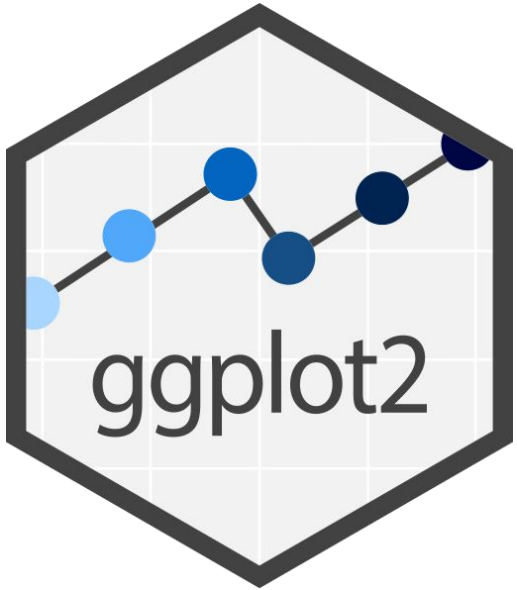
Communication graphs

- Make you point as clearly (and simply) as possible
- More is **not** better
- These will be refined versions of analysis graphs

How to get inspired?

- It can help to draw the plot you are imaging on paper
- Not imagining anything? Shamelessly steal and copy!
  - Try to recreate plots you see (in papers etc) for a dataset you have
  - I often google image search a description of what I am imaging to see similar examples and then try to remake these

Now you have a plot in mind how do you make it?



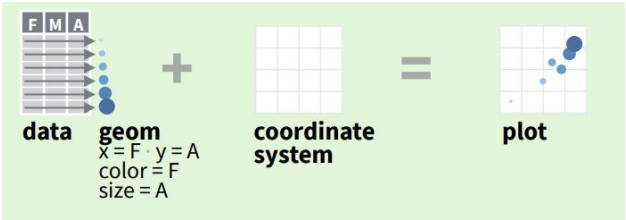
The grammar of graphics - ggplot2

data: a data frame: quantitative, categorical; local or database query
aesthetic mapping of variables into visual properties: size, color, x, y
geometric objects (“geom”): points, lines, areas, arrows, …

rstudio.github.io/cheatsheets/data-visualization.pdf

The grammar of graphics - ggplot2

data: a data frame: quantitative, categorical; local or database query
aesthetic mapping of variables into visual properties: size, color, x, y
geometric objects (“geom”): points, lines, areas, arrows, …



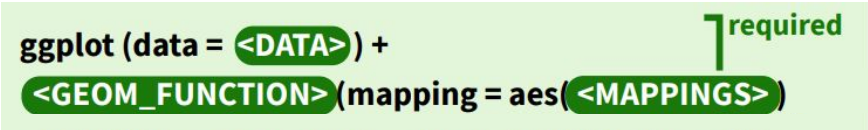
rstudio.github.io/cheatsheets/data-visualization.pdf

The grammar of graphics - ggplot2

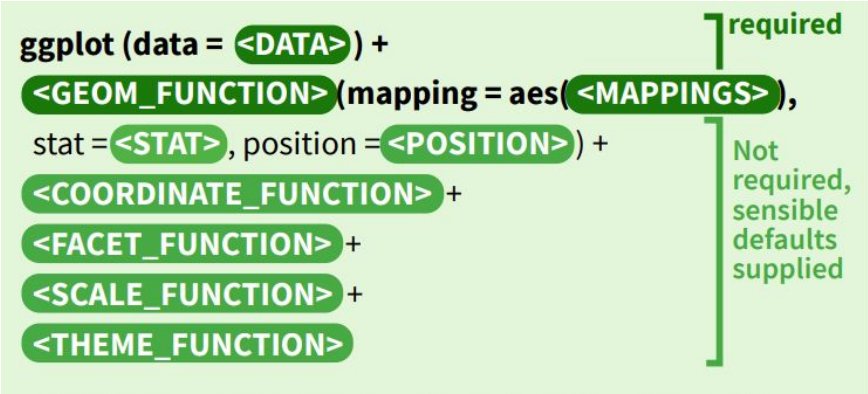
Aesthetics **aes**() make data visible:

x, y: variables
colour: colours the lines of geometries
fill: fill geometries
group: groups based on the data
shape: defines the shape (point, triangles)
linetype: defines the type of line (solid, dashed)
size: define sizes of elements
alpha: changes the transparency

The grammar of graphics - basic structure



The grammar of graphics - basic structure



Build ggplot figures in
layers

Build ggplot figures in **layers**

Example data: Penguins!

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	41.8	19.4	198	4450	male	2008
Adelie	Biscoe	37.8	18.3	174	3400	female	2007
Gentoo	Biscoe	43.5	14.2	220	4700	female	2008
Chinstrap	Dream	45.4	18.7	188	3525	female	2007
...	...	...	...	...	...	...	...

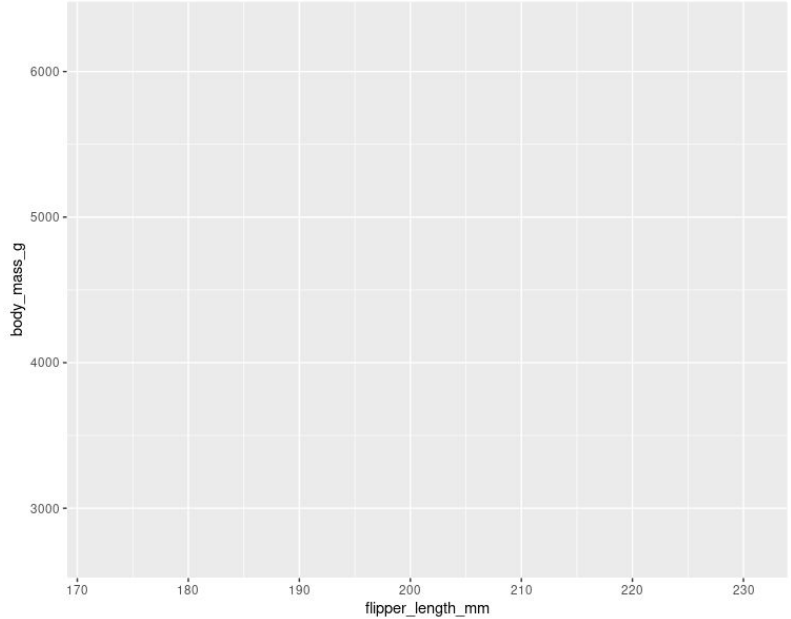
Build ggplot figures in **layers**

```
ggplot(
  data = penguins
)
```


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	41.8	19.4	198	4450	male	2008
Adelie	Biscoe	37.8	18.3	174	3400	female	2007
Gentoo	Biscoe	43.5	14.2	220	4700	female	2008
Chinstrap	Dream	45.4	18.7	188	3525	female	2007
...	...	...	...	...	...	...	...

Build ggplot figures in **layers**

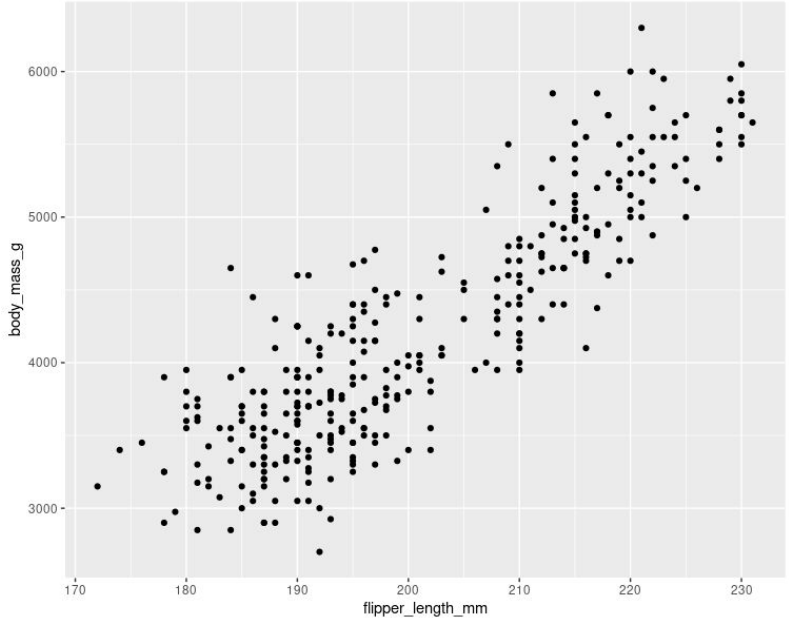
```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
)
```



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	41.8	19.4	198	4450	male	2008
Adelie	Biscoe	37.8	18.3	174	3400	female	2007
Gentoo	Biscoe	43.5	14.2	220	4700	female	2008
Chinstrap	Dream	45.4	18.7	188	3525	female	2007
...	...	...	...	...	...	...	...

Build ggplot figures in **layers**

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point()
```

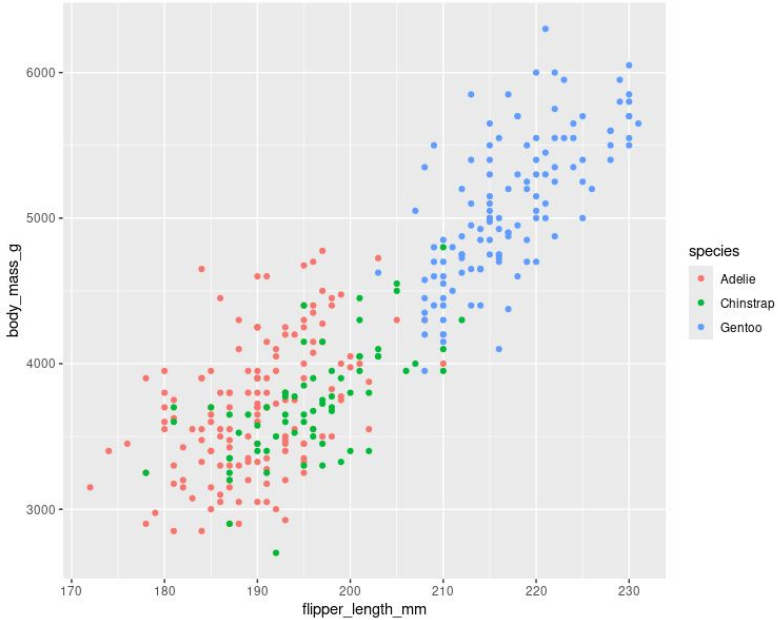


species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	41.8	19.4	198	4450	male	2008
Adelie	Biscoe	37.8	18.3	174	3400	female	2007
Gentoo	Biscoe	43.5	14.2	220	4700	female	2008
Chinstrap	Dream	45.4	18.7	188	3525	female	2007
...	...	...	...	...	...	...	...

r4ds.hadley.nz/data-visualize.html

Build ggplot figures in **layers**

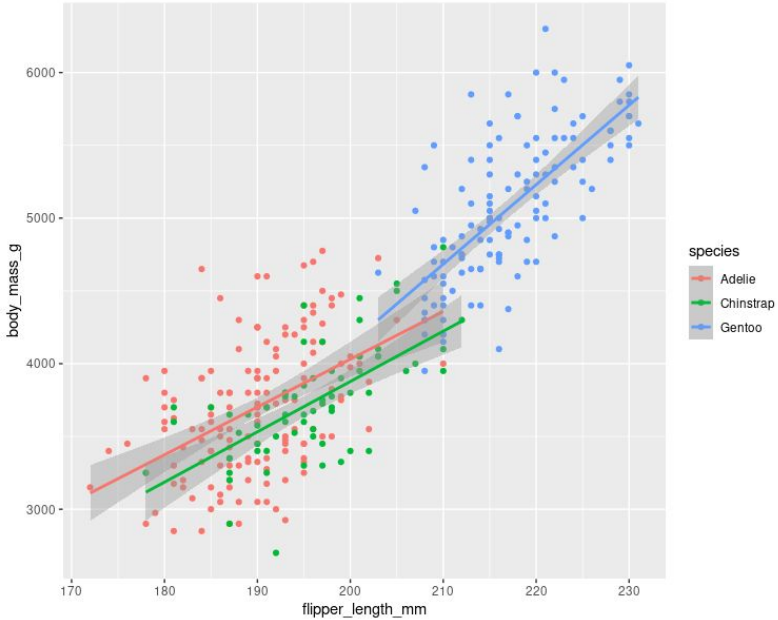
```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g,
                color = species)
) +
  geom_point()
```



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	41.8	19.4	198	4450	male	2008
Adelie	Biscoe	37.8	18.3	174	3400	female	2007
Gentoo	Biscoe	43.5	14.2	220	4700	female	2008
Chinstrap	Dream	45.4	18.7	188	3525	female	2007
...	...	...	...	...	...	...	...

Build ggplot figures in **layers**

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g,
                color = species)
) +
  geom_point() +
  geom_smooth(method = "lm")
```

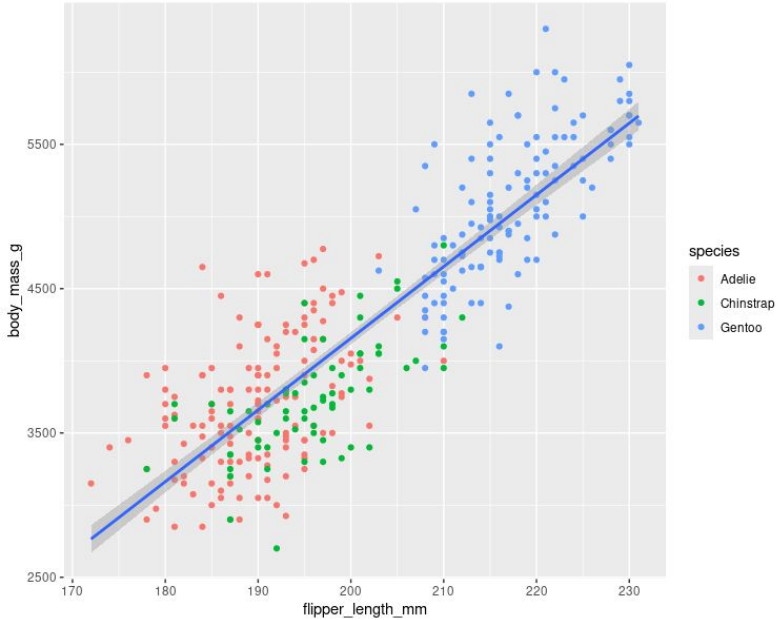


species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	41.8	19.4	198	4450	male	2008
Adelie	Biscoe	37.8	18.3	174	3400	female	2007
Gentoo	Biscoe	43.5	14.2	220	4700	female	2008
Chinstrap	Dream	45.4	18.7	188	3525	female	2007
...	...	...	...	...	...	...	...

r4ds.hadley.nz/data-visualize.html

Build ggplot figures in **layers**

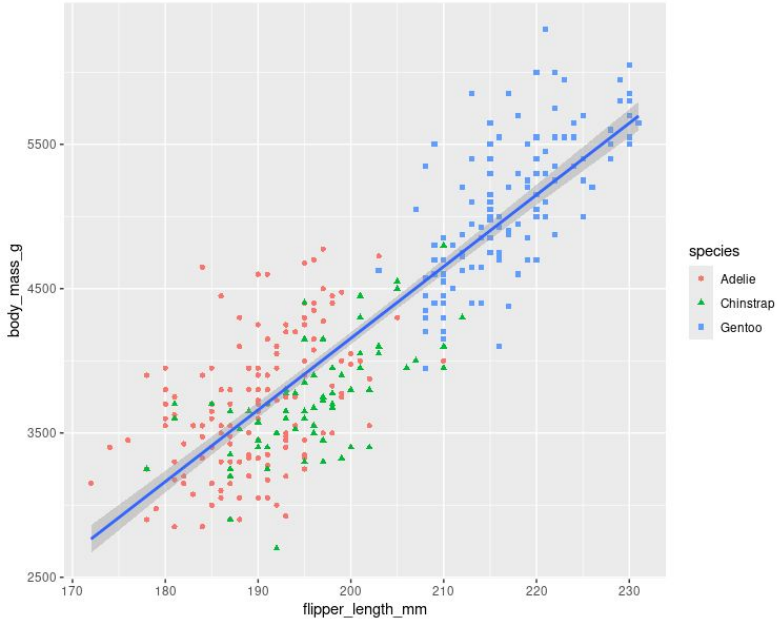
```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point(aes(color = species)) +
  geom_smooth(method = "lm")
```



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	41.8	19.4	198	4450	male	2008
Adelie	Biscoe	37.8	18.3	174	3400	female	2007
Gentoo	Biscoe	43.5	14.2	220	4700	female	2008
Chinstrap	Dream	45.4	18.7	188	3525	female	2007
...	...	...	...	...	...	...	...

Build ggplot figures in **layers**

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point(aes(color = species, shape = species)) +
  geom_smooth(method = "lm")
```

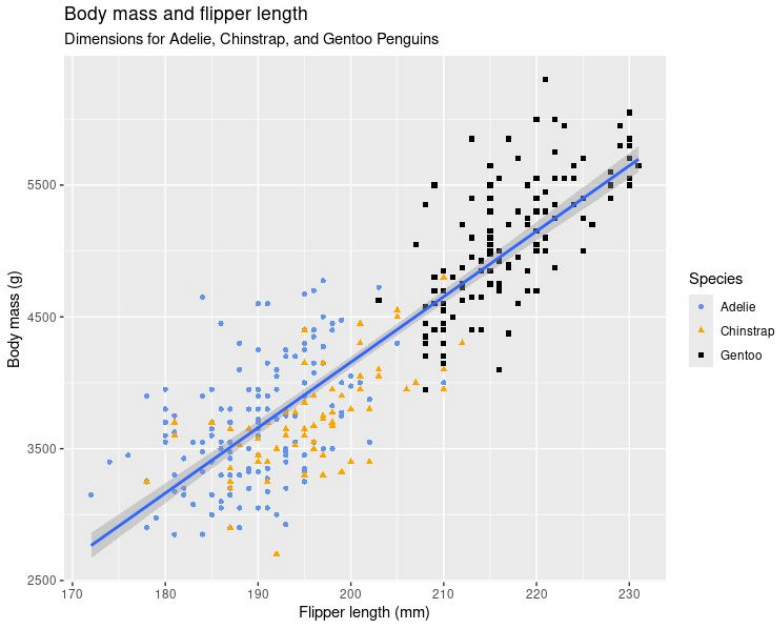


species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	41.8	19.4	198	4450	male	2008
Adelie	Biscoe	37.8	18.3	174	3400	female	2007
Gentoo	Biscoe	43.5	14.2	220	4700	female	2008
Chinstrap	Dream	45.4	18.7	188	3525	female	2007
...	...	...	...	...	...	...	...

r4ds.hadley.nz/data-visualize.html

Build ggplot figures in **layers**

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point(aes(color = species, shape = species)) +
  geom_smooth(method = "lm") +
  labs(
    title = "Body mass and flipper length",
    subtitle = "Dimensions for Adelie, Chinstrap, and Gentoo
Penguins",
    x = "Flipper length (mm)", y = "Body mass (g)",
    color = "Species", shape = "Species")+
scale_color_manual(values=c("cornflowerblue","orange","black"))
```



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	41.8	19.4	198	4450	male	2008
Adelie	Biscoe	37.8	18.3	174	3400	female	2007
Gentoo	Biscoe	43.5	14.2	220	4700	female	2008
Chinstrap	Dream	45.4	18.7	188	3525	female	2007
...	...	...	...	...	...	...	...

Workshop overview and Example Plots

redgar598.github.io/COSS_data_vis_R

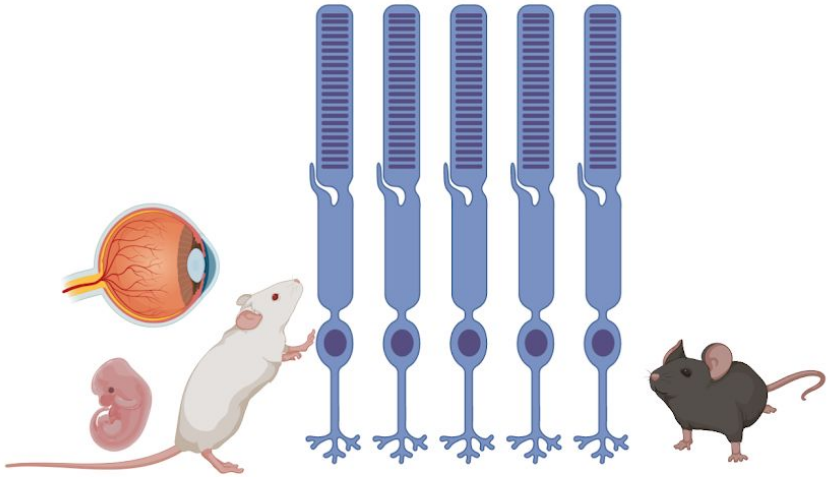
How did
set-up go?

Data We Will Be Using in ggplot Introduction

Gene expression data from mouse photoreceptors

Samples from different developmental stages (**E16,P2,P6,P10 and 4 weeks**)

Two mouse lines, a wildtype (**wt**) and knockouts for rod cell specific transcription factor (**NrlKO**)



Plots Made in Workshop