

Thanks a lot for participating!



Artwork by Allison Horst

Bring your own data

Day 4 - Introduction to Data Analysis with R

Selina Baldauf

Freie Universität Berlin - Theoretical Ecology

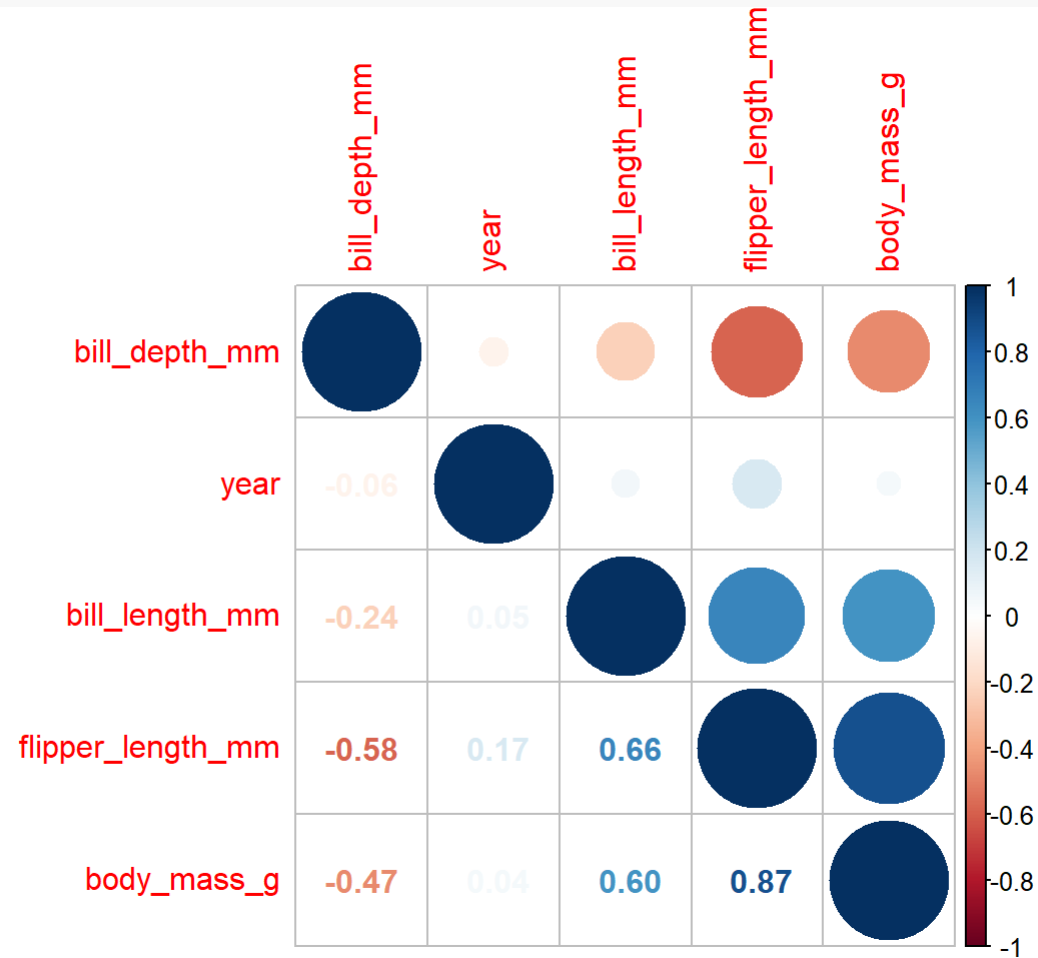
October 17, 2023

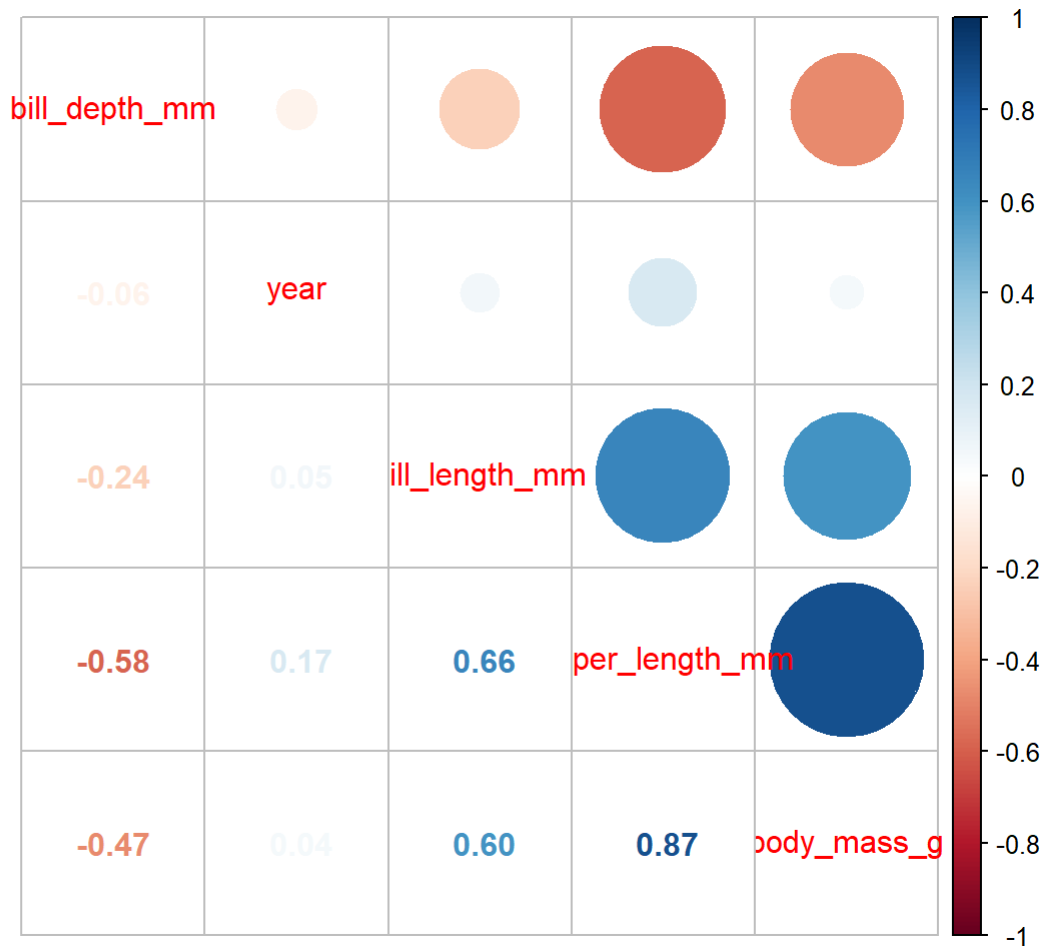
Organization

Schedule of today

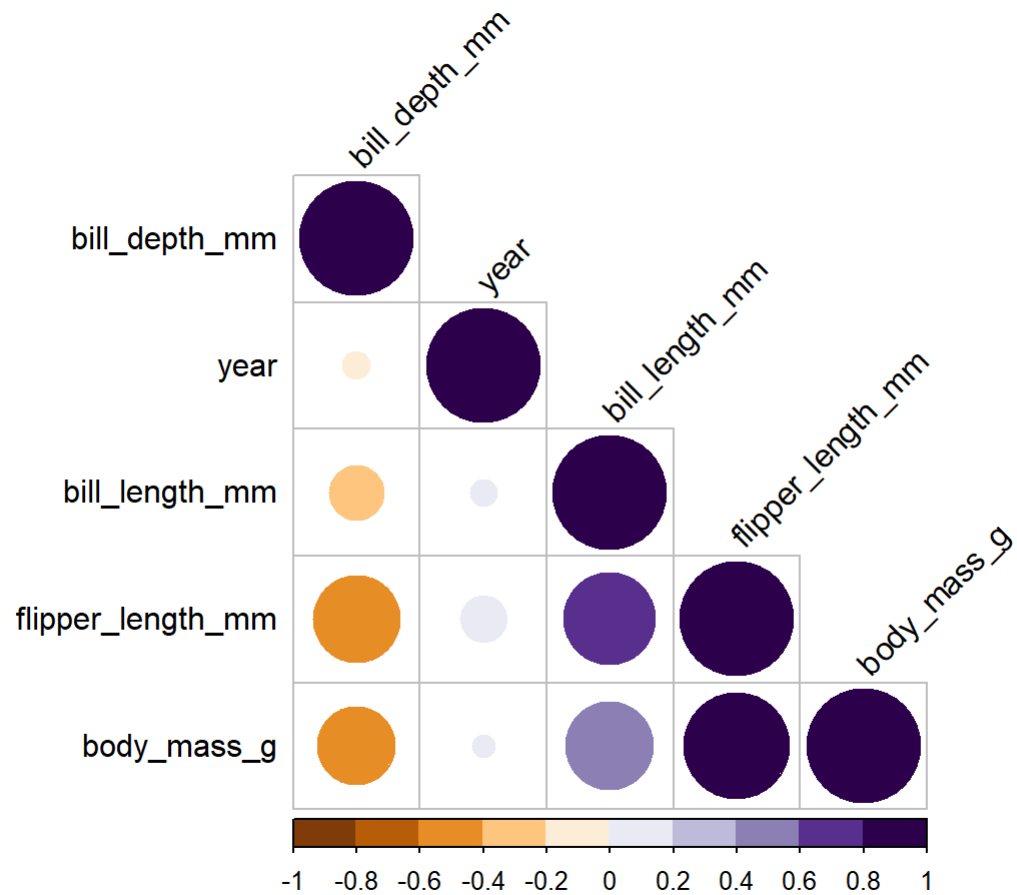
- **Now - 14** (or 14.30 if you are enthusiastic still): Work on the data set(s)
 - Take break(s) as best fits your needs
- **14 (14.30) - 15**: Short feedback round
 - What did you find out about your data set? Plots, summaries, ...
 - Which methods did you use?
 - Did you learn something new?
 - Was there something you struggled with?
 - ...
- **15-16**: Feedback, conclusion

New methods: Correlation plots

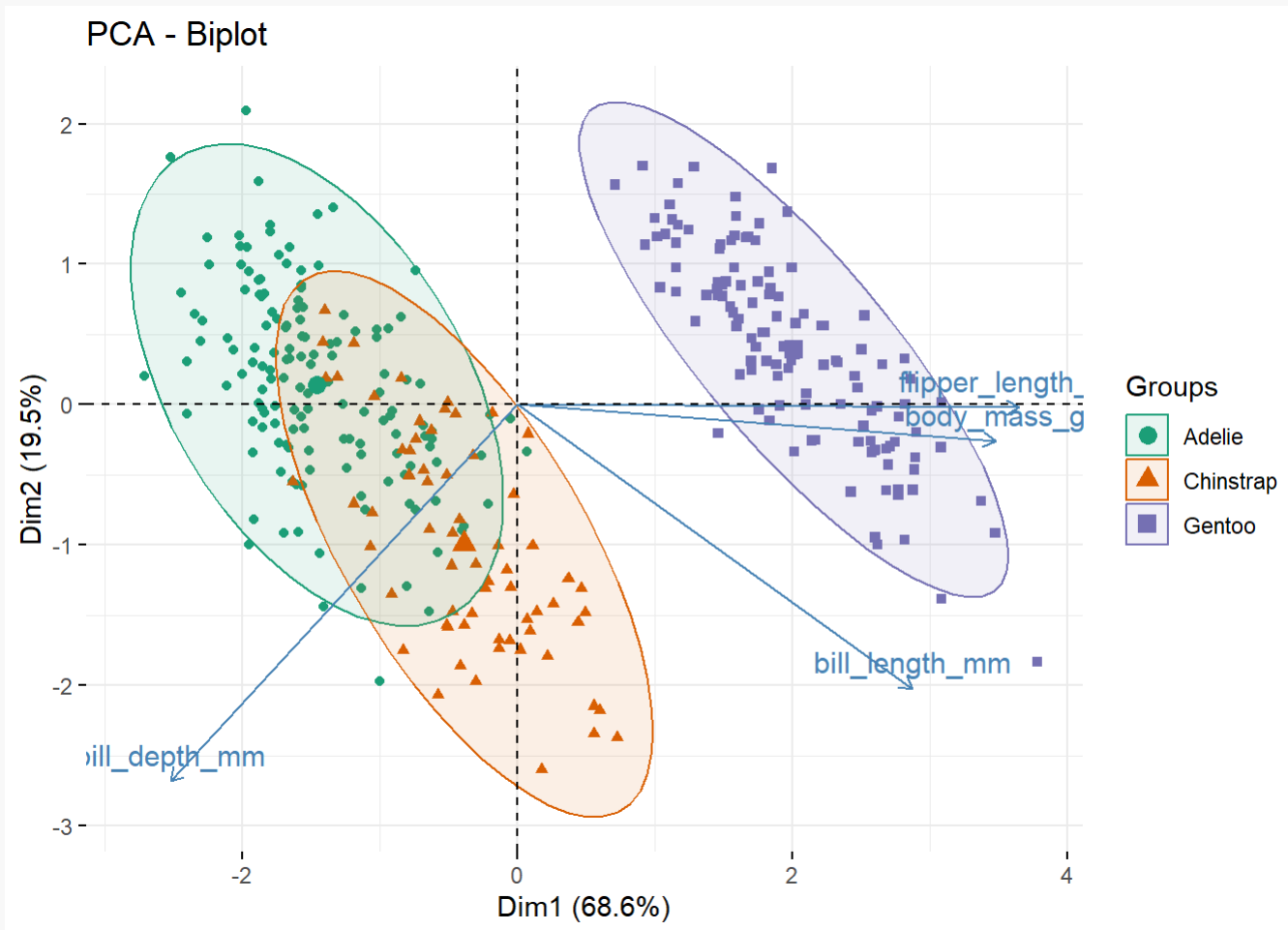




New methods: Correlation plots



New methods: PCAS



New methods: Correlation plots and PCAS

- `corrplot` package for correlation plots
- `factoextra` package for PCA visualization
- Tutorial for PCAs in R, PCA tutorial for penguins
- Correlations and PCAs do not work with `NA` values: use `tidyr::drop_na()` to remove all `NA` values from the data first
- These plots work for: penguins, wine, piecrab (?) dataset

Data set 1: What makes a good wine?

Physicochemical properties of wine and quality judgements

```
'data.frame': 1599 obs. of 12 variables:
 $ fixed.acidity      : num 12.7 9.8 6.5 8.6 7.5 7.6 10.1 6.4 6.1 6.7 ...
 $ volatile.acidity   : num 0.6 0.66 0.88 0.52 0.58 0.5 0.935 0.4 0.58 0.46 ...
 $ citric.acid        : num 0.49 0.39 0.03 0.38 0.14 0.29 0.22 NA 0.23 0.24 ...
 $ residual.sugar     : num 2.8 3.2 NA 1.5 2.2 2.3 3.4 1.6 2.5 1.7 ...
 $ chlorides          : num 0.075 0.083 0.079 0.096 0.077 NA 0.105 0.066 0.044 0.077 ...
 $ free.sulfur.dioxide : num 5 21 23 5 27 5 11 5 16 18 ...
 $ total.sulfur.dioxide: num NA 59 47 18 60 NA 86 12 70 34 ...
 $ density            : num 0.999 0.999 0.996 NA 0.996 ...
 $ pH                 : num 3.14 3.37 NA 3.2 3.28 3.32 3.43 3.34 3.46 3.39 ...
 $ sulphates          : num 0.57 0.71 0.5 0.52 0.59 NA 0.64 NA NA 0.6 ...
 $ alcohol            : num 11.4 11.5 11.2 9.4 9.8 11.5 11.3 9.2 12.5 10.6 ...
 $ quality            : int 5 7 4 5 5 6 4 5 6 6 ...
```

Data set 1: What makes a good wine?

Ideas - know methods

- Plot of wine quality against chemical properties
- Plot of distribution of chemical properties
- Summary tables using `dplyr`

Ideas - new methods

- **Correlation plots:** How are the different wine properties correlated with each other?
- **PCA:** How are the wine properties related to each other?



Frederik Vandaele - originally posted to Flickr as Château Pétrus, CC BY 2.0,
<https://commons.wikimedia.org/w/index.php?curid=5145286>

Data set 1: What makes a good wine?

Hints

- Transform the quality column to a factor before plotting: use `dplyr::mutate` and `as.factor()` to transform the column
- Try the `janitor::clean_names()` function



Frederik Vandaele - originally posted to Flickr as Château Pétrus, CC BY 2.0,
<https://commons.wikimedia.org/w/index.php?curid=5145286>

Data set 2: Paralympic games from 1980-2016

Most important variables:

variable	class	description
gender	character	Binary gender
event	character	Event name
medal	character	Medal type
athlete	character	Athlete name (LAST NAME first name
abb	character	Country abbreviation
country	character	Country name
type	character	Type of sport
year	double	year of games

Data set 2: Paralympic games from 1980-2016

Get the data:

```
athletes <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/ti
```

Ideas - know methods

- Create summaries of medal counts for different groups with `dplyr`
- Did the ratio of men/women winning medals change over time?
- Which countries were the most successful ones? Does this differ between sports type?
 - Which types of sports accumulated the most medals?
- Make plots such as:
 - Age distribution of athletes winning gold, silver and bronze
 - Compare the total number of medals over the years between winter and summer Olympics

Data set 2: Paralympic games from 1980-2016

Hints

- To reduce complexity of the data, first filter only the athletes that won a medal
`(!is.na(medal))`

Data set 3: Crab data set

Atlantic marsh fiddler crab (*Minuca pugnax*)

- Crab from Florida is expanding northward due to ocean warming
- Data on 13 marshes across a range of latitude in the USA
- Recording of the size of the crab
- Rather small and good to handle



Image by LTER under CC BY-SA 4.0

Data set 3: Crab data set

```
# A tibble: 6 × 9
  date      latitude site    size air_temp air_temp_sd water_temp water_temp_sd
  <date>      <dbl> <chr> <dbl>   <dbl>      <dbl>      <dbl>      <dbl>
1 2016-07-24      30 GTM    12.4    21.8        6.39      24.5        6.12
2 2016-07-24      30 GTM    14.2    21.8        6.39      24.5        6.12
3 2016-07-24      30 GTM    14.5    21.8        6.39      24.5        6.12
4 2016-07-24      30 GTM    12.9    21.8        6.39      24.5        6.12
5 2016-07-24      30 GTM    12.4    21.8        6.39      24.5        6.12
6 2016-07-24      30 GTM    13.0    21.8        6.39      24.5        6.12
# 1 more variable: name <chr>
```

Source: [Johnson, D. 2019](#). Fiddler crab body size in salt marshes from Florida to Massachusetts, USA at PIE and VCR

Selina Baldauf // Bring your own data

Data set 3: Crab data set

Ideas - known methods

- Explore Bergmann's rule (organisms are large in higher latitudes)
- t-tests to compare size between locations
- Plot relationship between latitude and size
- Plot distributions of variables



Image by LTER under CC BY-SA 4.0

Data set 4: Ice cover and temperature

Temperature and ice duration on lakes since 19th century

- 2 data sets with measurements of
 - ice start, end and duration on 2 lakes in Wisconsin
 - daily air temperature since 1870
- Explore the effect of climate change on ice cover



Image by LTER under CC BY-SA 4.0

Source ice data: [Magnuson, J.J., S.R. Carpenter, and E.H. Stanley. 2021.](#) North Temperate Lakes LTER: Ice Duration - Madison Lakes Area 1853 - current ver 35. Environmental Data Initiative.

Data set 4: Ice cover and temperature

Ice data:

```
# A tibble: 6 × 5
  lakeid      ice_on      ice_off      ice_duration  year
  <fct>      <date>      <date>      <dbl> <dbl>
1 Lake Mendota NA      1853-04-05      NA 1852
2 Lake Mendota 1853-12-27 NA      NA 1853
3 Lake Mendota 1855-12-18 1856-04-14    118 1855
4 Lake Mendota 1856-12-06 1857-05-06    151 1856
5 Lake Mendota 1857-11-25 1858-03-26    121 1857
6 Lake Mendota 1858-12-08 1859-03-14     96 1858
```

Temperature data:

```
# A tibble: 6 × 3
  sampledate  year ave_air_temp_adjusted
  <date>      <dbl>      <dbl>
1 1870-06-05  1870         20
2 1870-06-06  1870        18.3
3 1870-06-07  1870        17.5
4 1870-06-09  1870        13.3
5 1870-06-10  1870        13.9
6 1870-06-11  1870         15
```

Data set 4: Ice cover and temperature

Ideas - known methods

- How did ice cover duration change over the years?
- How did air temperature change over the years?
 - Summarize mean annual temperature or mean temperature in winter
- How do ice duration on the lakes correlate with temperature (e.g. with mean winter temperature)



Image by LTER under CC BY-SA 4.0

Hints

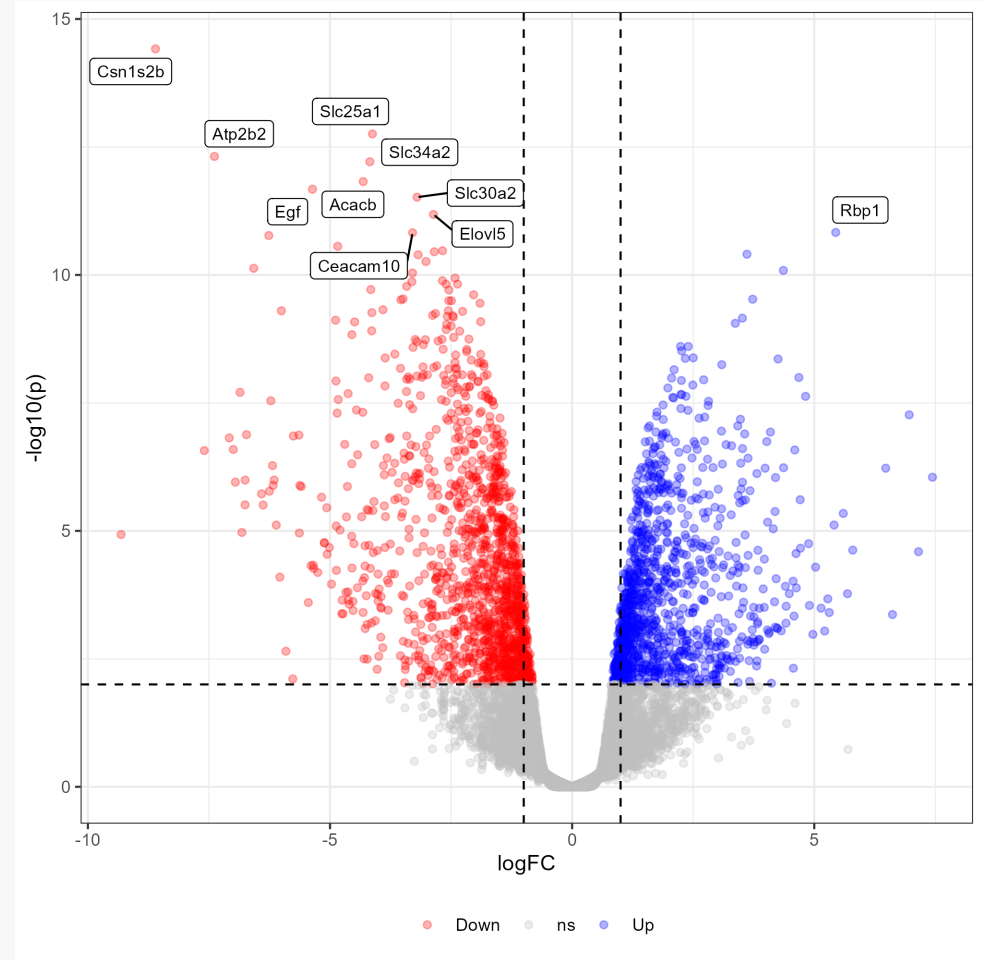
Data set 5: RNAseq data

- Data from [FU et al. 2015, Nature Cell Biology](#)
- Data found via [Tutorial on heat maps](#) using this data
- 3 csv files:
 - `heatmap_genes.csv`: A list of the names of interesting genes to look at (Genes used in Figure 6b in paper)
 - `DE_results.csv`: Gene expression in luminal cells in pregnant versus lactating mice
 - logFC, AveExpr, t, p-value
 - Also contains non-significantly expressed genes
 - `normalized_counts`: Normalized counts for genes for the different samples

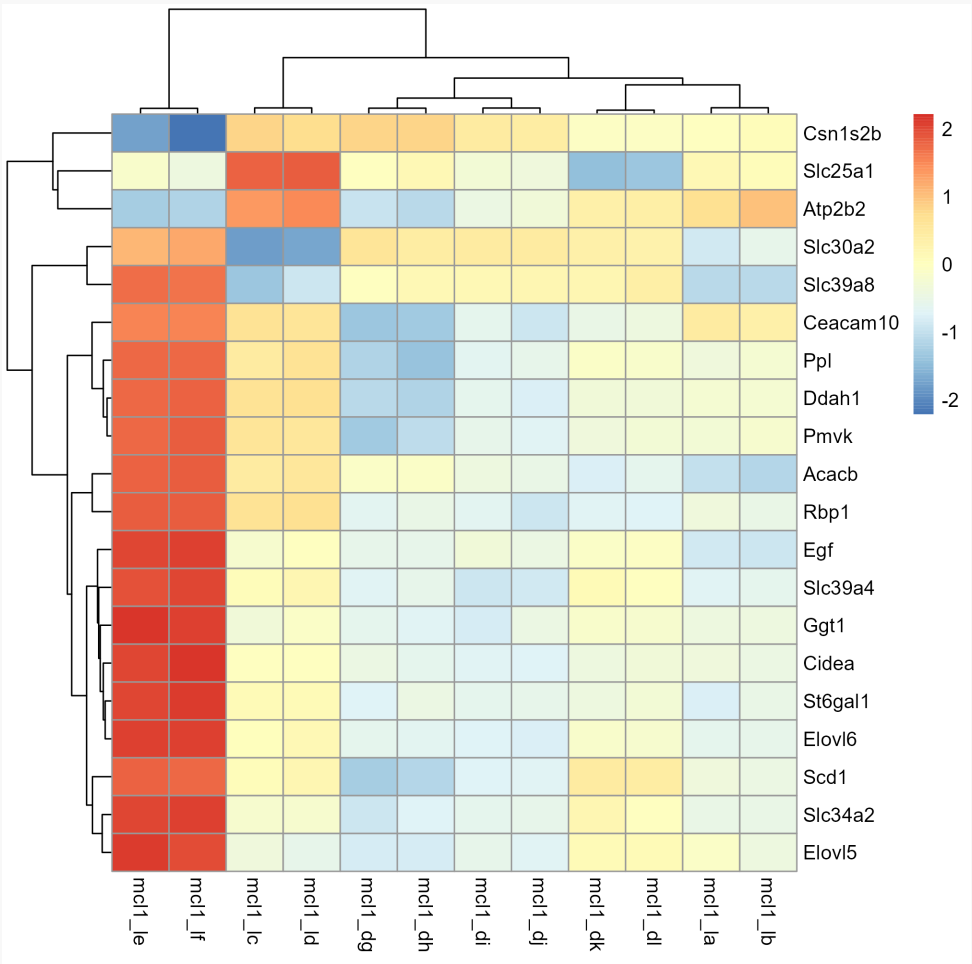
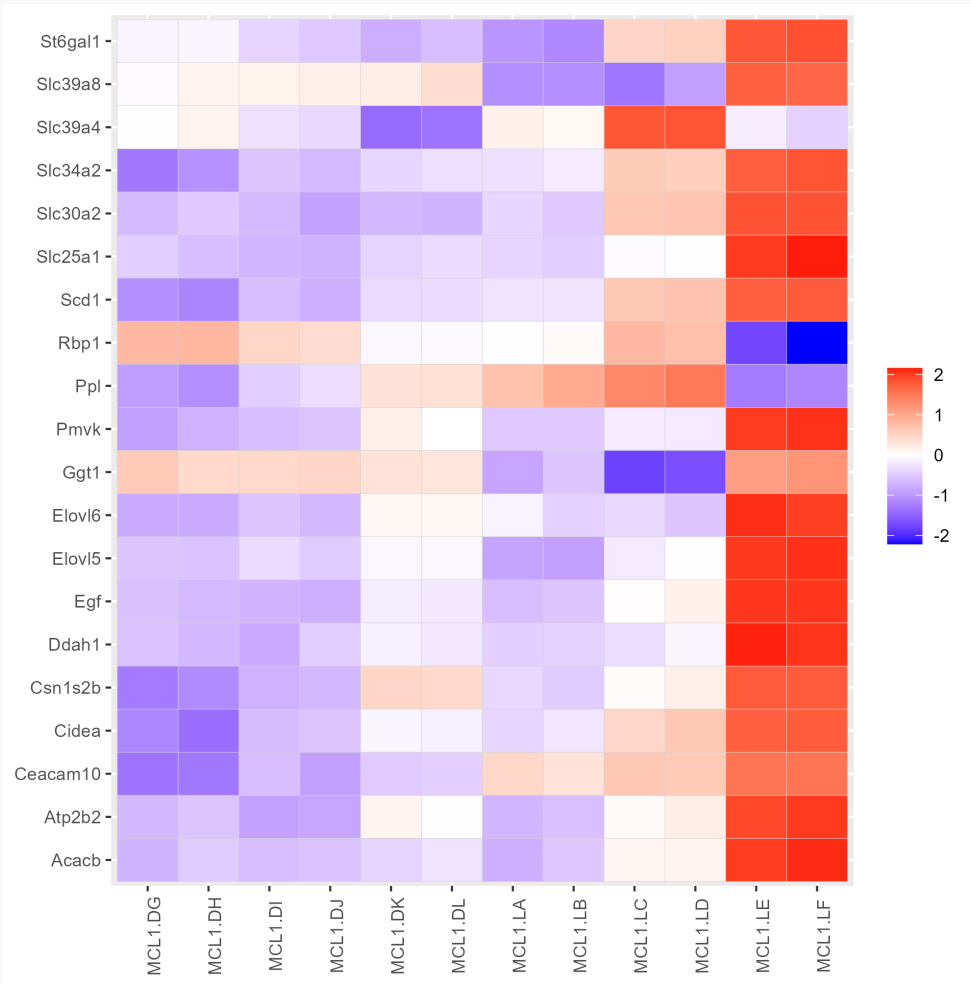
Data set 5: RNAseq data

Ideas:

- Create a heatmap of the top 20 most significant genes (see plot in the tutorial)
- Create a heatmap of the interesting genes (see Fig. 6 in the paper)
- Create a volcano plot of the data similar to the one [here](#)



Data set 5: RNAseq data



Data set 5: RNAseq data

Some tips:

Data cleaning:

- Read in the data and then use the `janitor::clean_names` function to make the column headers nicer
- Join `DE_results` and `normalized_counts` by their shared columns
- Use `select` to remove columns you don't need for analysis to get a better overview
- Filter only significant genes ([tutorial](#)) defines them as `p_value < 0.01 & abs(logFC) > 0.58`

Data analysis:

- Heatmap with ggplot or with `pheatmap::pheatmap()`
 - `pheatmap` takes a matrix as input (use `as_matrix` on tibble to transform)
- scale the counts -> have a look at the `scale` function
 - `pheatmap` can scale but with ggplot you have to scale before plotting

Some general tips

- First make a plan:
 - What do you want to achieve and what are the steps?
 - Try to think in technical terms
- Start with something small, e.g. reading in the data and bringing it into the right format.
- Google
- If you get stuck, ask in the chat or stop by in General
- Have a look at the [package cheat sheets](#)

Now you

Working with real research data

Meet in your group (if you want)

Work on your data set

Take breaks as you need and **be back** at 2 p.m.

Keep an eye on your **group** and the **general chat**

Sharing

In 1-2 mins:

- What was the highlight of your analysis?
 - Your favorite plot
 - Some cool code
 - A problem that you finally solved
 - Something new you learned
- What was difficult?
- If you want: Share a screenshot in the chat or share your screen

Feedback

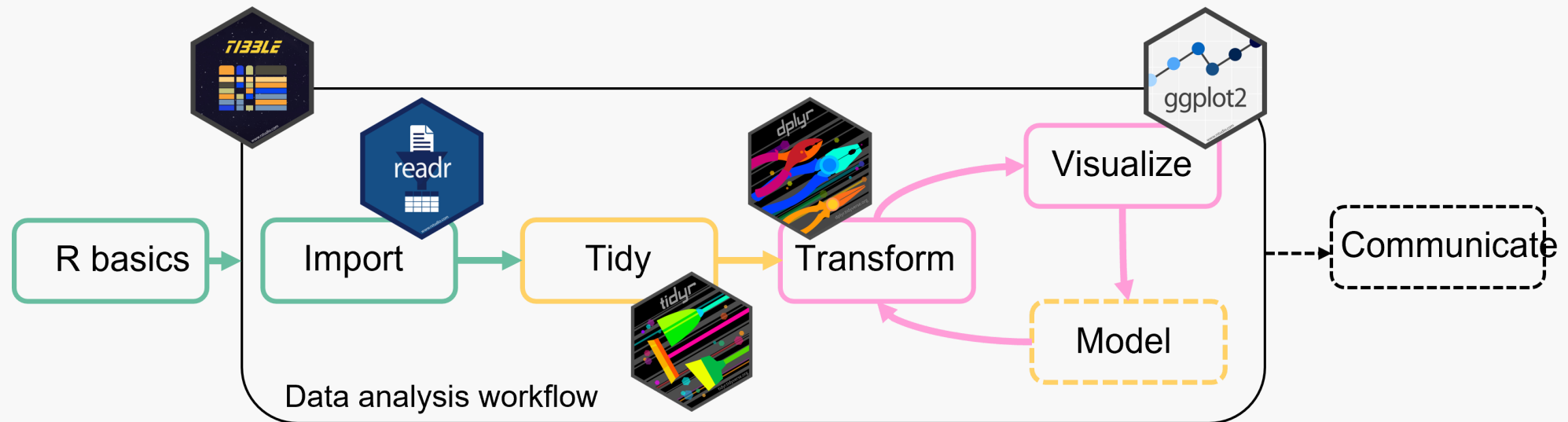
Please take 10 mins to complete the feedback survey for the Graduate center (don't use Internet Explorer)

<https://votingo.cedis.fu-berlin.de/PCNLP3>

Feedback

- Any other feedback or comments from your side?

Conclusion

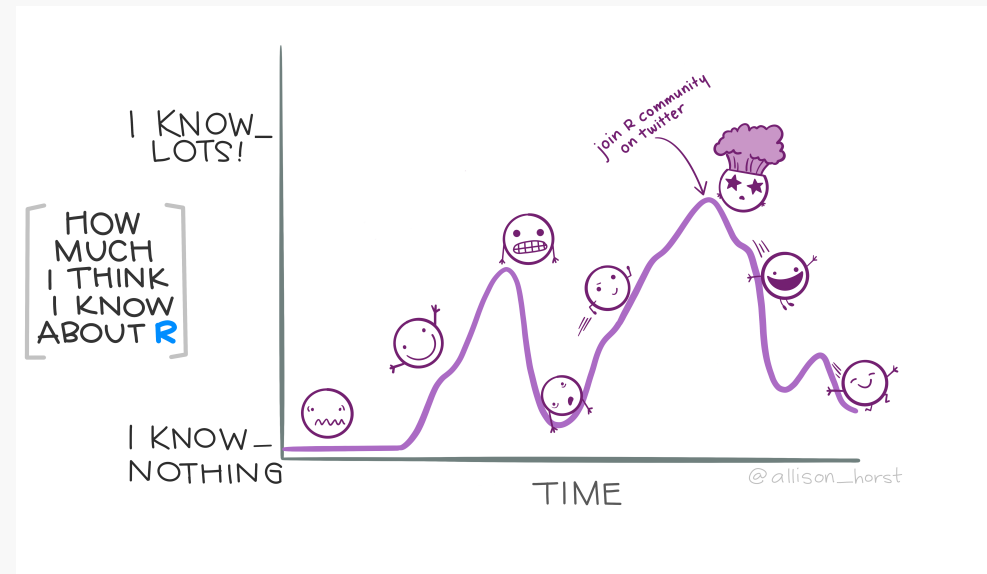


We learned a lot of stuff!

Conclusion

How to continue from here?

- Learning by doing!
- Have a look at some [online resources](#), I recommend the R for Data Science book by Hadley Wickham
- If you like plotting: Consider participating in the [tidytuesday](#)
- [FU statistical consulting](#) for questions regarding statistical methods
- [R Consulting by me](#)
- [Tools and Tips lecture](#)



Artwork by [Allison Horst](#)

The End

