FROM EXCEL TO R

KU DATA SCIENCE LABS

WHO ARE WE?

1. Center for Health Data Science (HeaDS) - https://heads.ku.dk/

   SUND Center, which includes a KU data lab

- **Courses & Workshops, Seminars, etc.**
- **Health DS Consultations**
- **Commissioned Research**
- **Matchmaking**
- **Commissioned Supervision**

2. ReNEW NNF Center for Stem Cell Medicine

3. Data Science Laboratory (DSL) - https://datalab.science.ku.dk/

Thilde Terkelsen [1]

Diana Andrejeva [1]

Tugce Karaderi [1]

Henrike Zschach [1]

Helene Wegener [1]

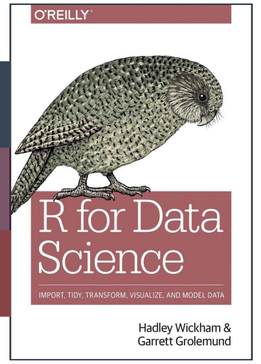
Adrija Kalvisa [2]

THE PRACTICALS

Two days: 9.00-16.00. There will be coffee breaks, we promise

“R for Data Science” - a generally useful book on R, also for this course

The course is build on hands-on presentations & exercises



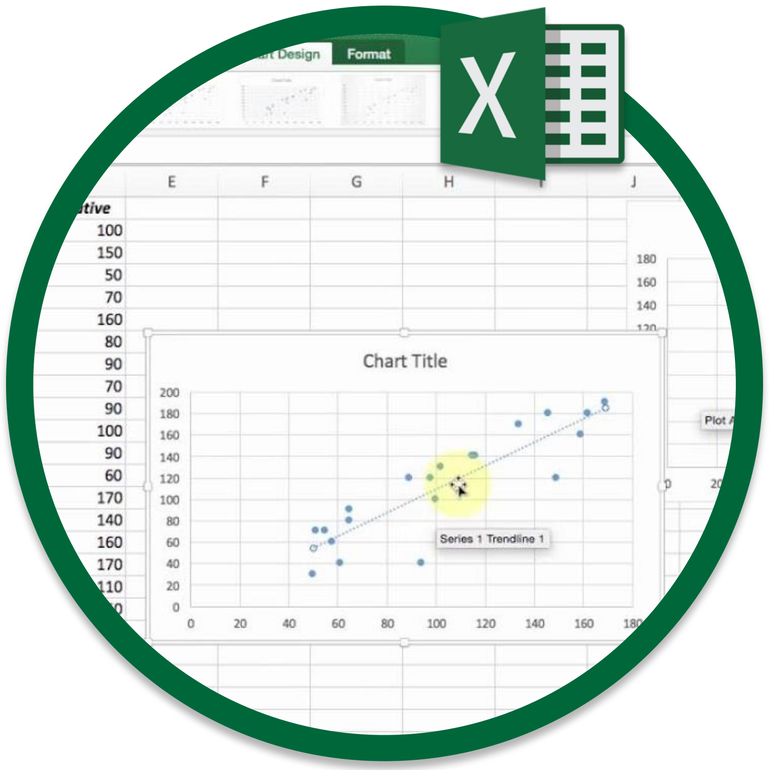
Download and install the newest version of R (https://cran.r-project.org/)

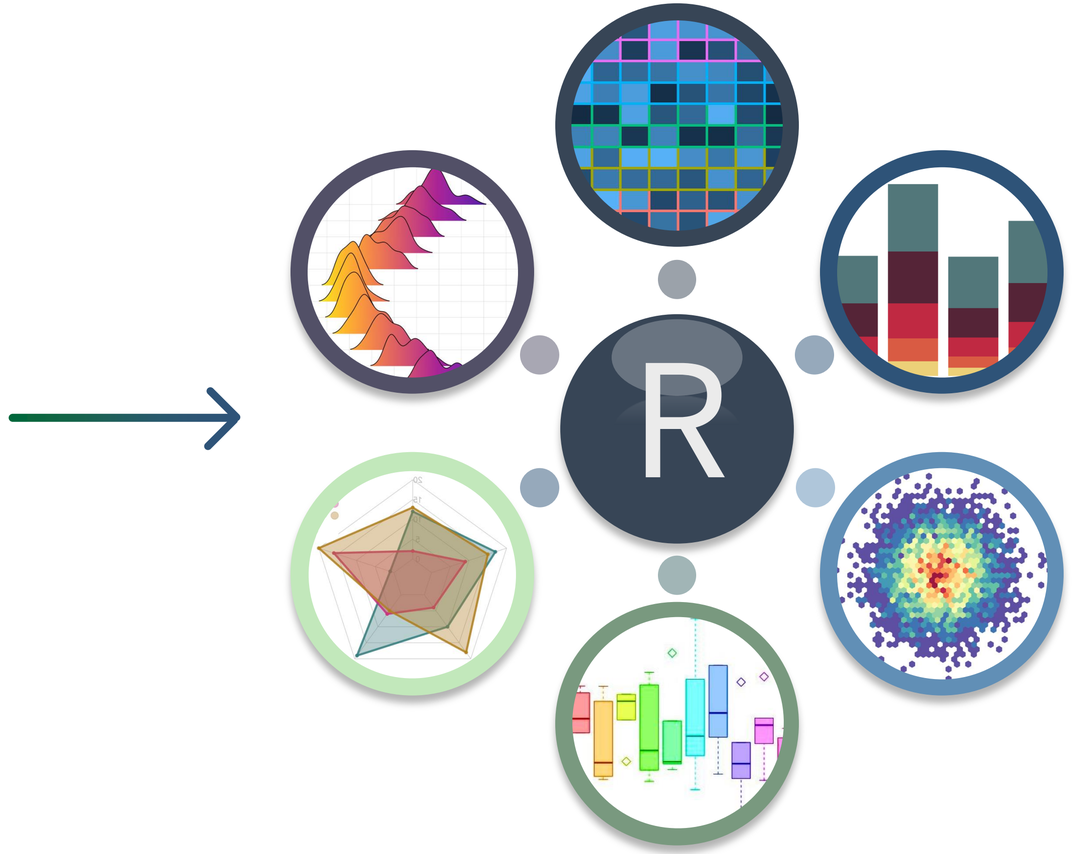
Download and install the newest version of R-studio (https://posit.co/download/rstudio-desktop/)

Download the course material and place it somewhere you can find it again!
https://github.com/Center-for-Health-Data-Science/FromExceltoR

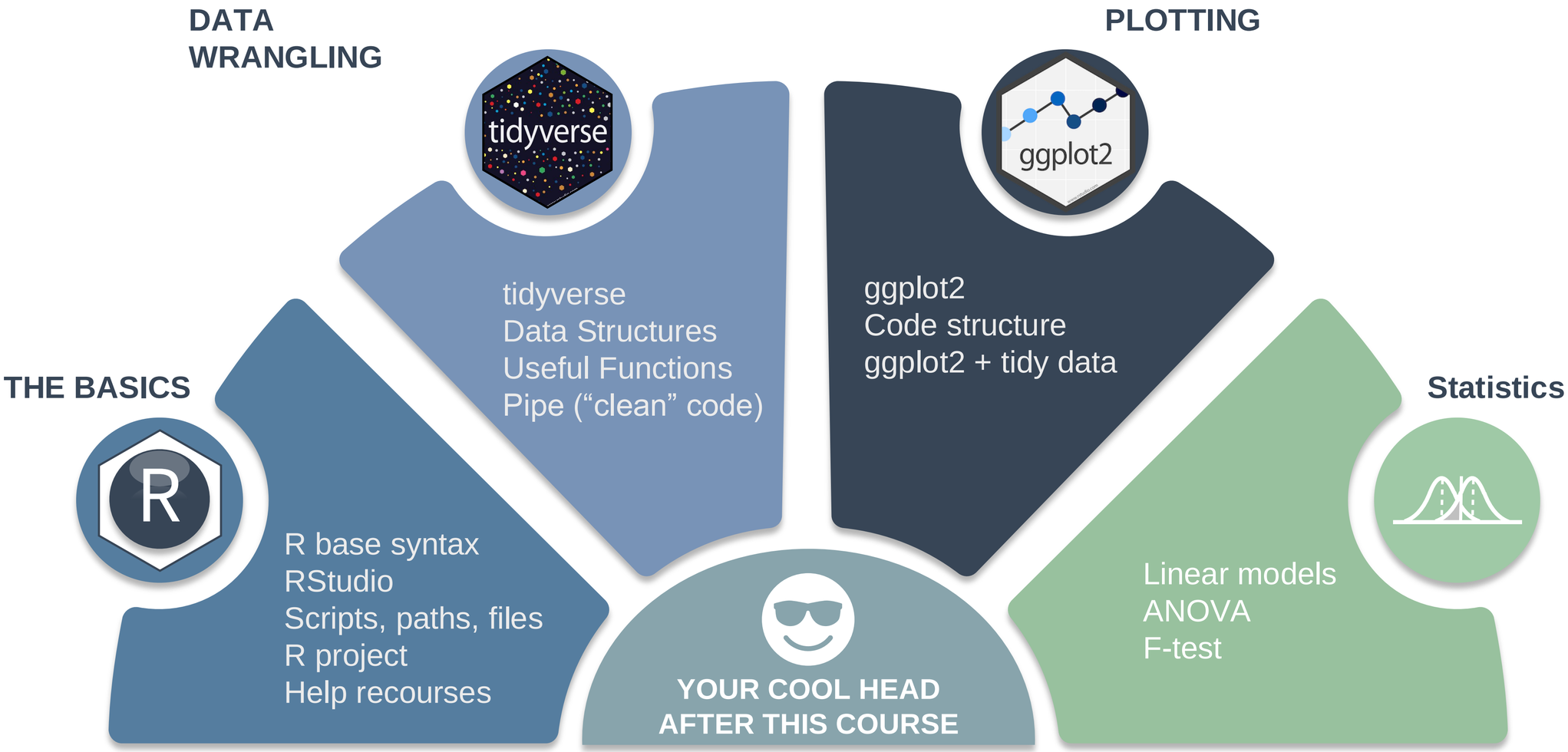
WELCOME TO

FROM EXCEL TO R





WHAT WILL YOU LEARN IN THIS COURSE?



PROGRAM

DAY 1

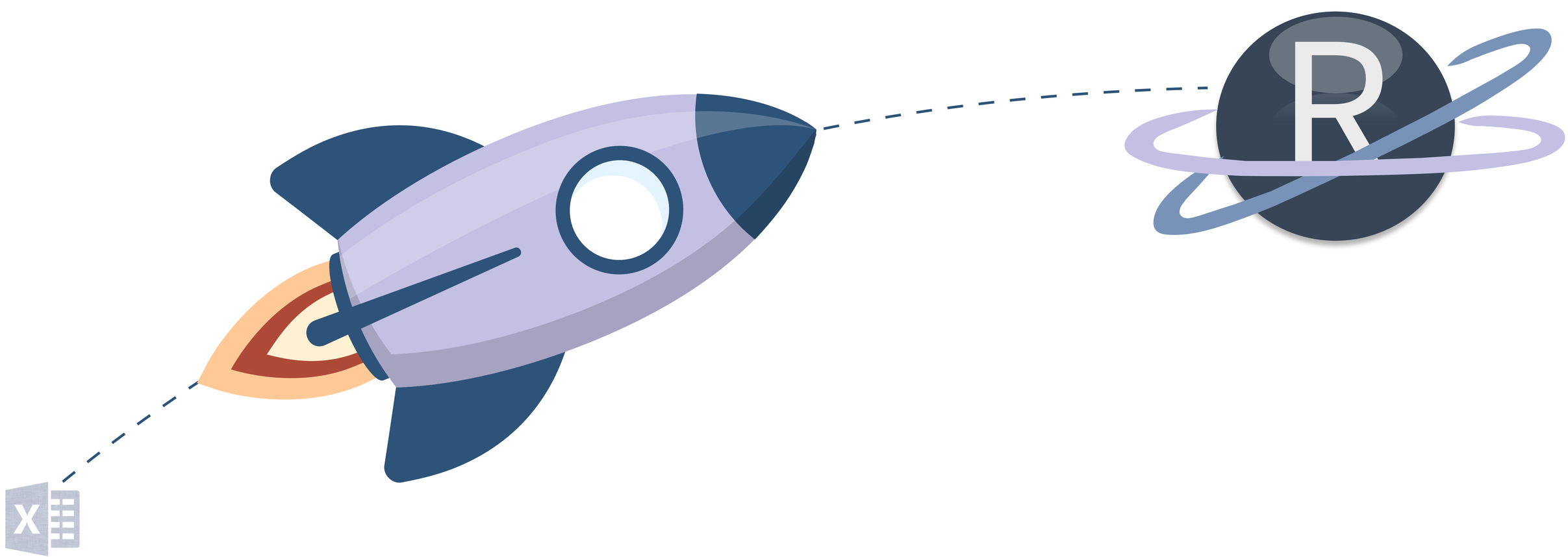
08:30 - Installation Issues & Coffee
09:00 - Introduction to R Basics
09:45 - Rstudio Exercise
10:45 - Break
11:00 - Tidyverse
12:00 - Lunch
13:00 - Tidyverse Exercise
14:15 - Break
14:30 - ggplot2
15:15 - ggplot2 Exercise
16:00 - Q&A + See you tomorrow

DAY 2

08:30 - Coffee + Optional Q&A
09:00 - Applied Statistics
10:00 - Applied Statistics Exercise
10:45 - Break
12:00 - Lunch
13:00 - Applied Statistics
14:30 - Break
14:45 - Basic Data Analysis Exercise
15:45 - Cool things in R & Course Evaluation
16:00 - Bye-bye

FROM EXCEL TO R

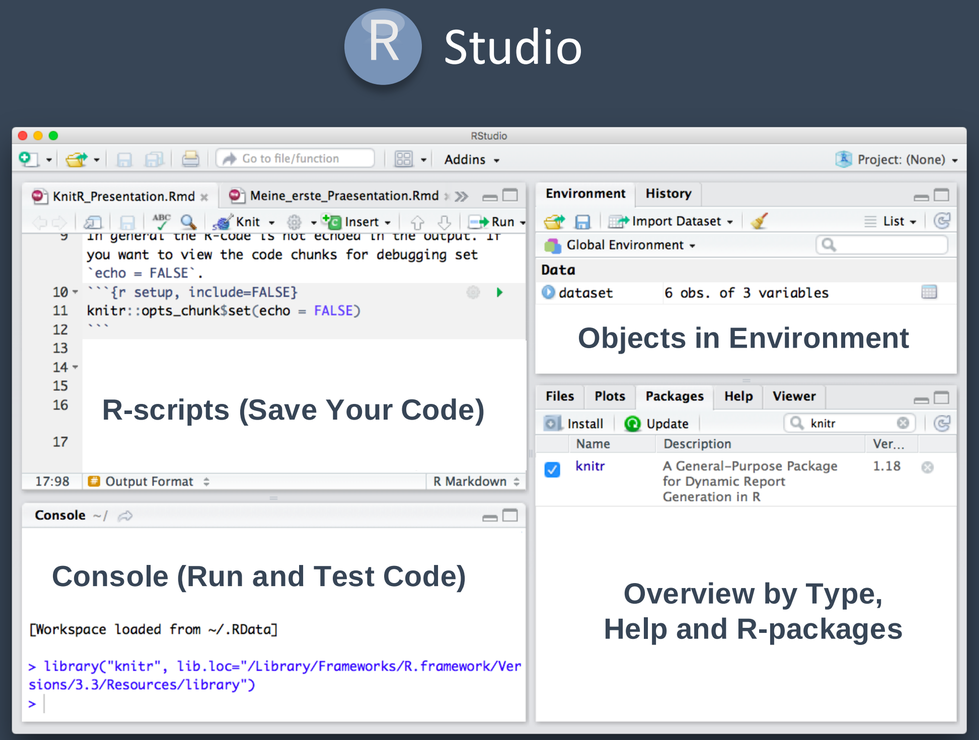
LET'S GET STARTED



R & FRIENDS

Scripting / Programming Language

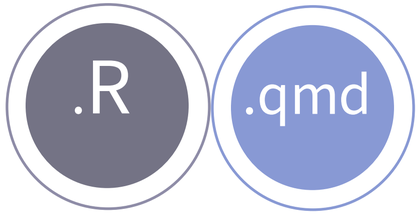
Reports (html, pdf, latex)



R Code Interpreter and Editor

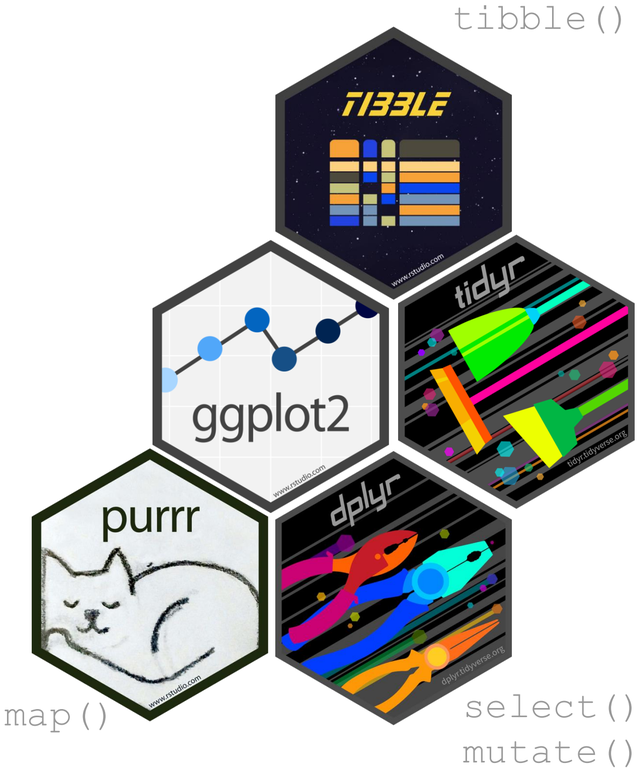
THE ANATOMY OF R

The coding **language**  can be written in the **editor** 

The files you code in are called **scripts** and are saved as 

Packages are libraries of functions.

Functions perform actions based on input arguments and return an output.



ONLINE RESOURCES FOR **R**

https://www.r-project.org/

GET STARTED

https://rseek.org/

https://rstudio.com/resources/cheatsheets/

http://www.cookbook-r.com/

https://www.statmethods.net/r-tutorial/index.html

GRAPHICS

https://www.r-graph-gallery.com/

http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

BOOKS & COURSES

https://www.r-bloggers.com/best-books-to-learn-r-programming/

https://www.datacamp.com/

https://www.codecademy.com/

https://www.coursera.org/

OTHER RESOURCES

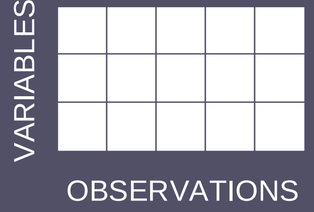
https://github.com/trending/r

https://blog.revolutionanalytics.com/

https://stackoverflow.com/questions/tagged/r

FROM EXCEL TO R

R DATA TYPES & STRUCTURES



DATA STRUCTURES

Vector

Age

Data Frame

Age
Height
BMI

List

Array

Tibbl

DATA TYPES

Numeric

-1.5, 2.7, 3.2

Integer

-2, -1, 0, 3, 5

Character

“A”, “Bat”, “Tom”

Factor

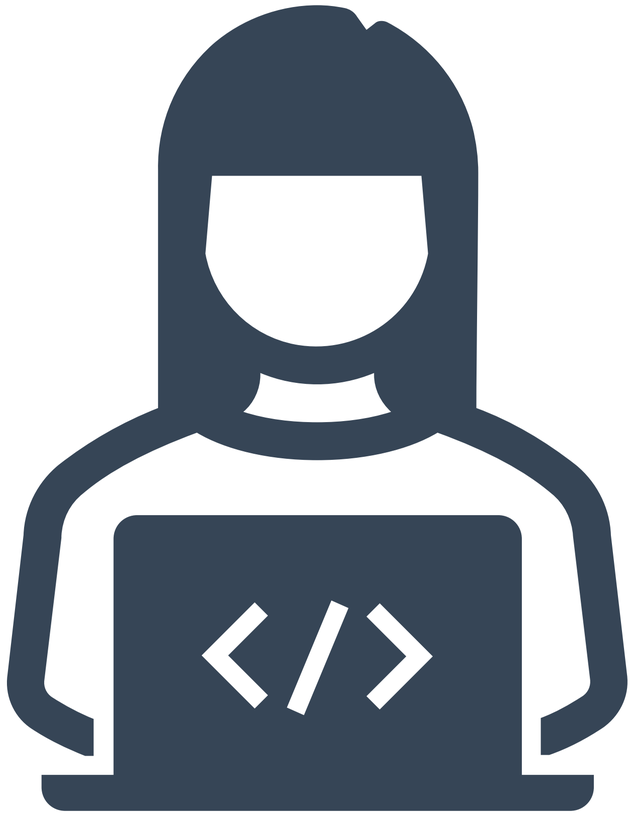
G1, G1, G2, G3

Logical

TRUE, FALSE

FROM EXCEL TO R

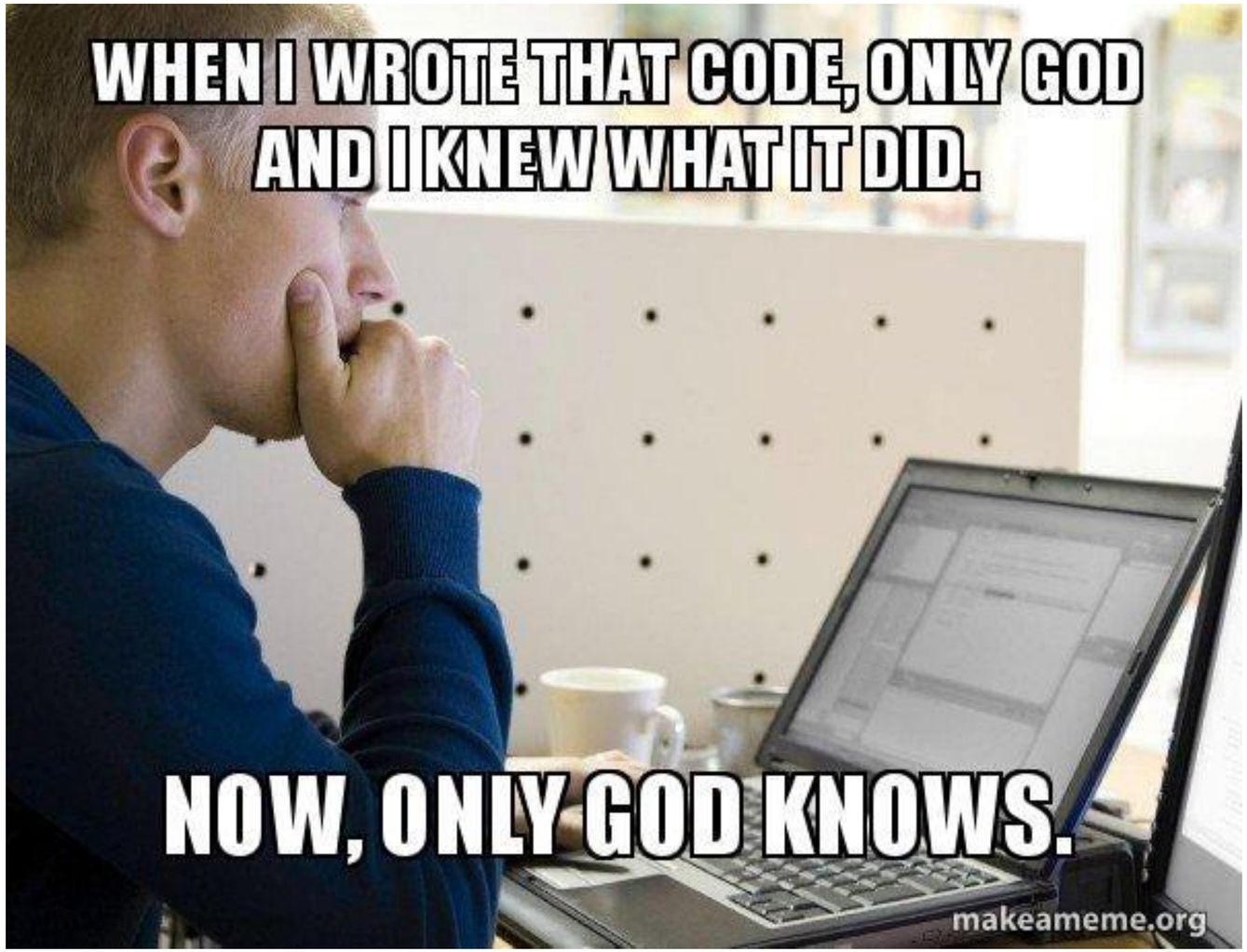
LIVE CODING 1 – INTRO TO R



https://quarto.org

QUARTO STRUCTURE

- Seperate section with headers.
- Do not overfill your code chunks.
- Think about when to use text and comments.



https://www.rstudio.com/wp-content/uploads/2016/10/r-cheat-sheet-3.pdf

INTRO CHEAT SHEET

Basics:

```
getwd() # location
install.packages("package_name") # install packages
library(packag_name) # load package
```

Read in data:

```
read.xlsx("name.xlsx"),
read.delim("name.txt", sep ="\t")
read.csv("name.csv", sep=";")
```

GETTING STARTED

Overview:

```
head(df, n=10), tail(df, n=10) # first or last 10 rows
unique(), table(), count() # unique vals, count vals
```

```
view() # view data as table
df$col1 # extract column from dataframe
nrow(df), ncol(df) # number of rows/columns
```

OVERVIEW

Type/Class:

```
class() # get data type/class
is.numeric(), is.character(), is.factor(), is.integer() # ask for data type
as.numeric(), as.factor(), as.matrix(), as.data.frame() # change data type
```

DATA TYPES

Summary statistics:

```
summary() # summary statistics
mean(), median, sd(), sum()
n() # number of rows
```

Plots:

```
plot(x)
plot(x,y) or plot(col1, col2, df) # scatter
hist(x) # histogram
```

STATISTICS & BASE PLOTS

https://rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf

QUARTO CHEAT SHEET

GETTING STARTED

YAML parameter:

```
---
title: My Project Name
output:
      html_document (pdf_document, …)
---
```

Code Chunk:

Source mode:

```
```{r}
# some R code
```
```

Visual mode:

```
{r}
# some R code
```

CHUNK OPTIONS

Code Options:

```
{r echo = FALSE} # don't print code (default is TRUE)
{r eval = FALSE} # don't run code (default is TRUE)
{r error = FALSE} # don't display error message (default TRUE)
(Can also be done with warning and message)
```

Figure Options:

```
fig.align (= 'left', 'right', 'center')
fig.cap (= 'my figure cation')
fig.height (= n), fig.width (= n)
```

TEXT

Source mode:

Header

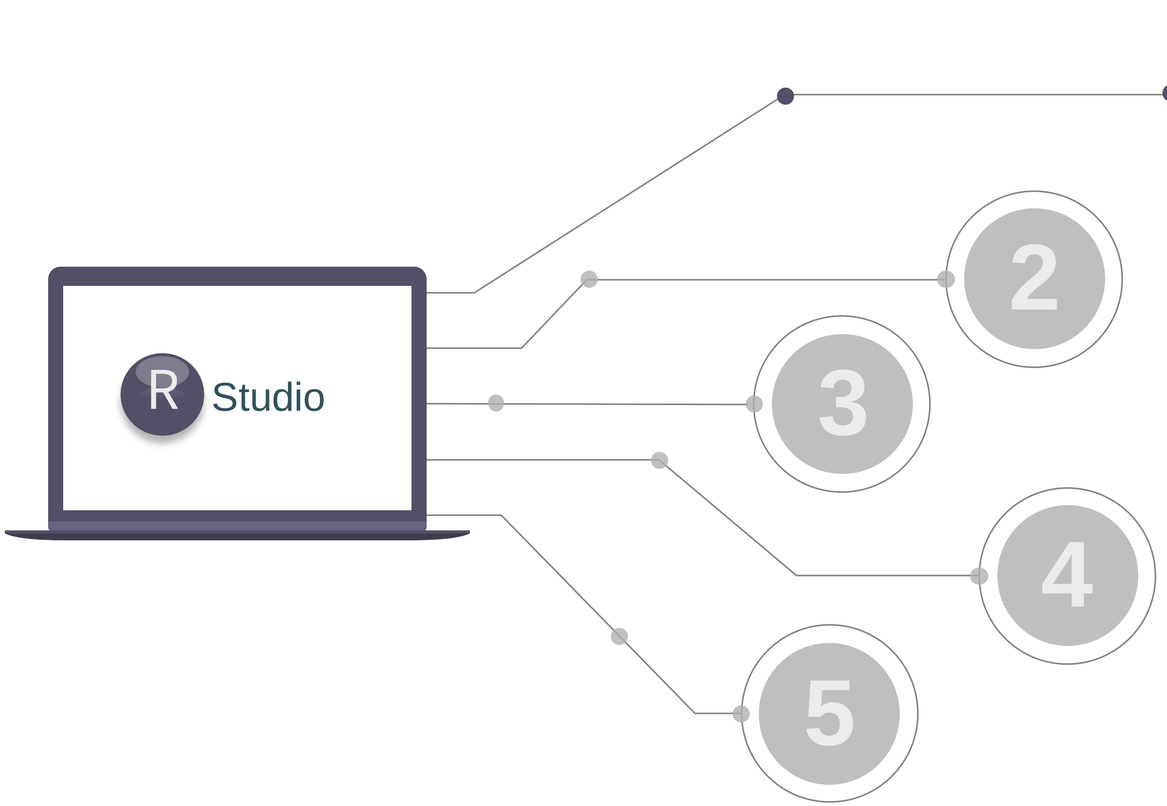
```
Header size ranging from largest (one #)
to smallest (six #):
# my.text, ## my.text, ### my.text, etc.
```

Text

```
*italics*
**bold**
`highlighted`
```

Lists

```
* List item1 (filled dot)
      + sub-item1 (open dot)
1. List item1 (numbered)
      i) sub-item1 (roman)
```



- R Script in Quarto
- Read & Write
- Basic calculus
- Base plot
- R help
- Save & Shutdown

FUNDAMENTALS

EXERCISE **1**

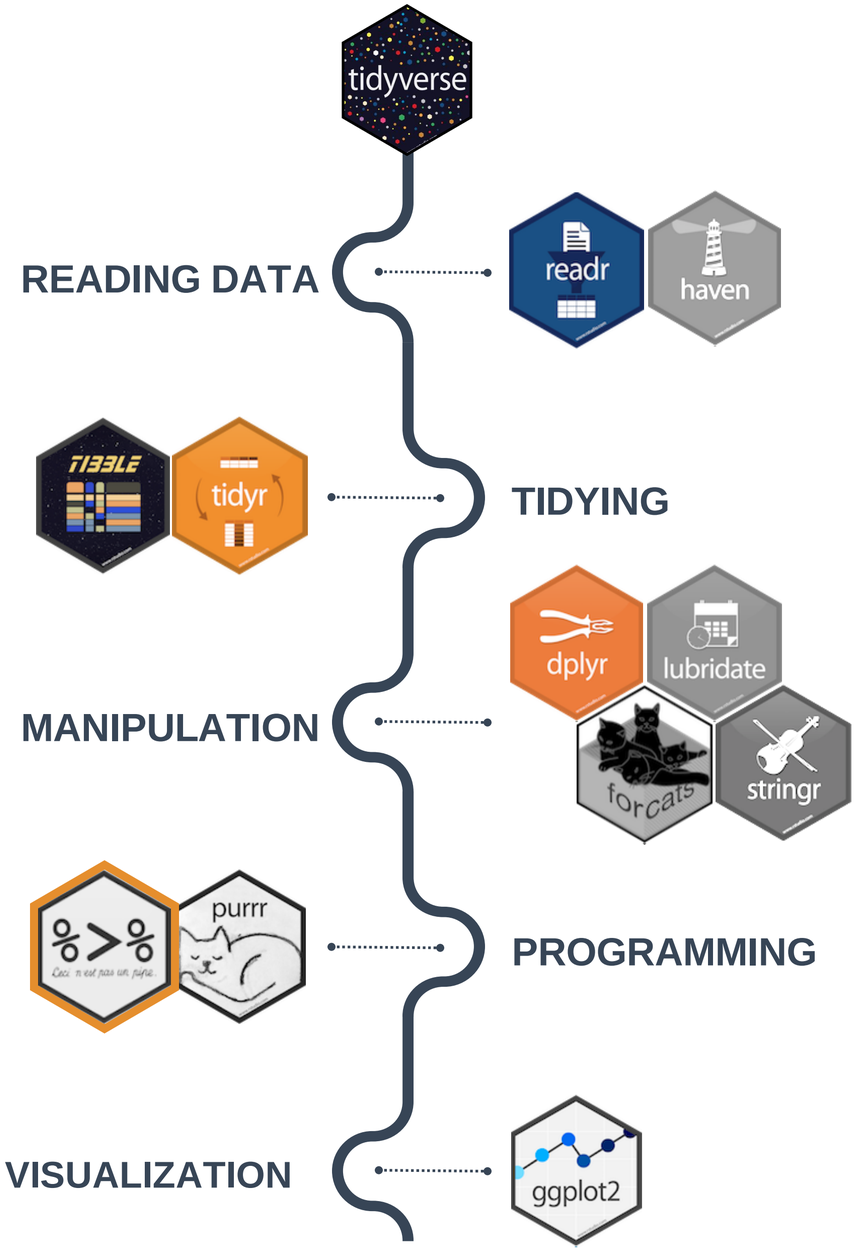
https://www.tidyverse.org/

TIDYVERSE

tidyverse is a collection of R packages for data science

"The packages share an underlying design philosophy, grammar, and data structures." *Wickham and Grolemund*

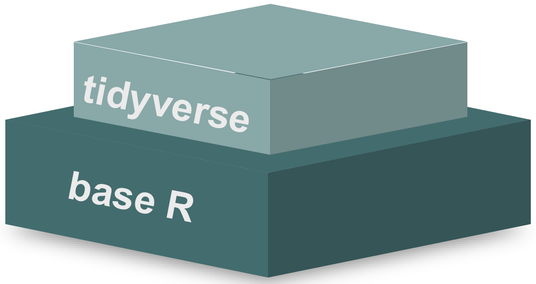
tidyverse is used to "tidy up" your datasets, so they are easy to work with



https://www.tidyverse.org/

CECI N'EST PAS UNE PIPE
%>%

- You do NOT have to “chose” between tidyverse and base R



BENEFITS

- Short & well-organised code
- Tidy datasets, easy to work with
- Great documentation
- Functions with logical names & inputs

CONSIDERATIONSS

- Can be less stable
- “Different syntax”
- Remember what tidyverse is made for!

base R

```
# think from the inside out

g(f(x,y),z)
```

tidyverse

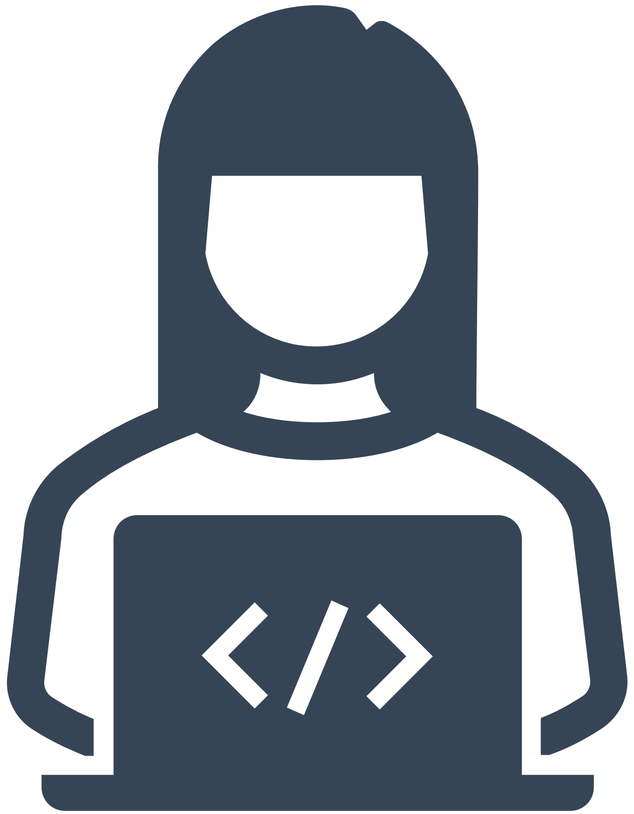
```
# no brain acrobatics

x %>% f(y) %>% g(z)
```

pipe symbol

FROM EXCEL TO R

LIVE CODING 2 - TIDYVERSE



https://www.tidyverse.org/

TIDYVERSE CHEAT SHEET

readr, tidyr, dplyr, …

Read Data (*readr*)

Reading tabular data
There are solutions for multiple data types

```
read_excel()  # using readxl package
read_table()
read_csv()
```

Useful arguments
Skip lines: `read_csv(file, skip=1)`
Read subset: `read_csv(file, n_max=1)`

Data types
readr guesses the types of each column and tells you about it

```
("Parsed with column specifications: …")
```

Workflow

Tidyverse workflow

```
df %>%
  select(col1)

df %>%
  filter(col1 < x)

df %>%
  summarize(n_1 = n(),
            avg_1 = mean(col1),
            sd_1 = sd(col1))
```

Select column

```
df$col1
```

Data Manipulation (*dplyr*)

Summary

```
summarize()
count()
```

Group

```
group_by()
```

Functions will manipulate each group separately and combine results.

Extract and sort observations (rows)

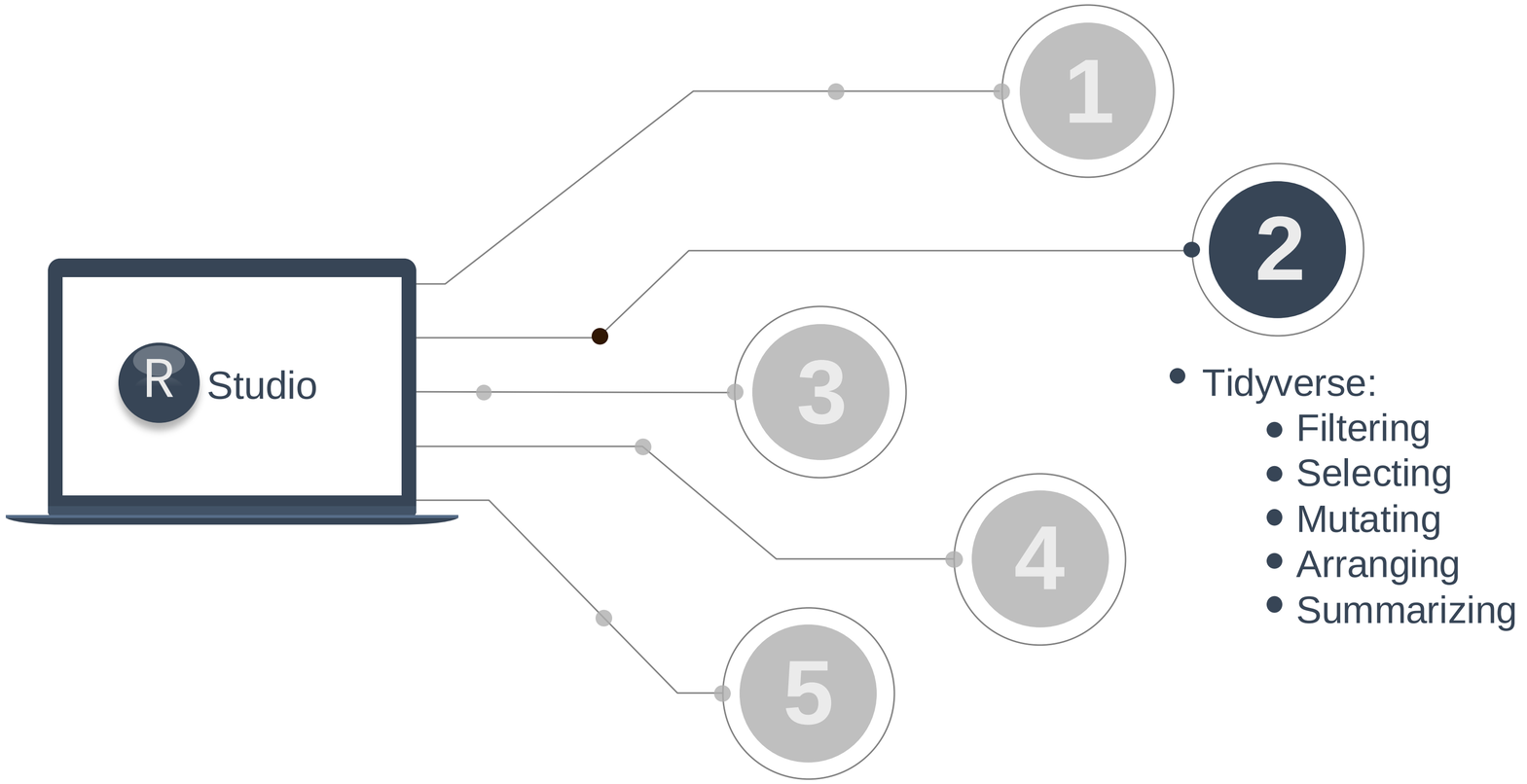
```
filter()  # subset rows by condition
distinct() # subset to unique values
top_n()  # subset by position
arrange() # sort low->high, other way with desc()
```

Manipulate variables (columns)

```
select()  # subset columns
mutate(new_col = f(col1))
mutate(new_col = ifelse(col1 < x, "Yes", "No"))
mutate(new_col = col1 + col2)
```

HELP

R Documentation (e.g. enter **`?dplyr::filter`** and see examples)
Much more info and detailed cheat sheets:
https://brianward1428.medium.com/introduction-to-tidyverse-7b3dbf2337d5
It also helps to google "*tidyverse + whatever_you_want_to_do*"



- Tidyverse:
  - Filtering
  - Selecting
  - Mutating
  - Arranging
  - Summarizing

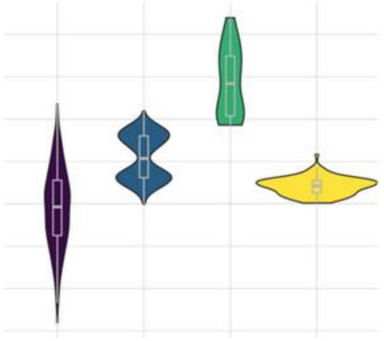
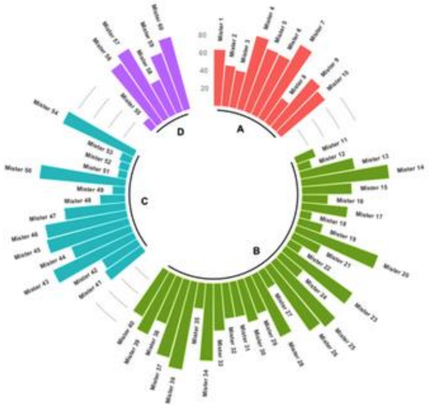
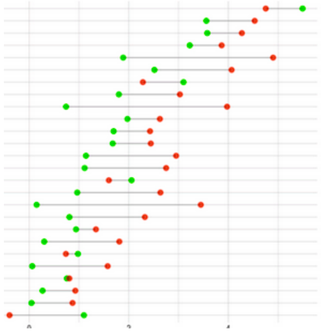
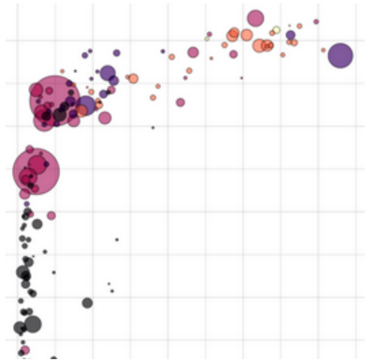
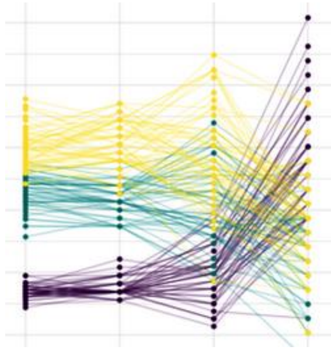
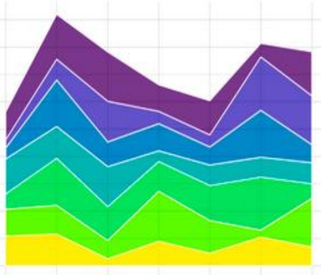
TIDYVERSE

EXERCISE **2**

https://www.r-graph-gallery.com/

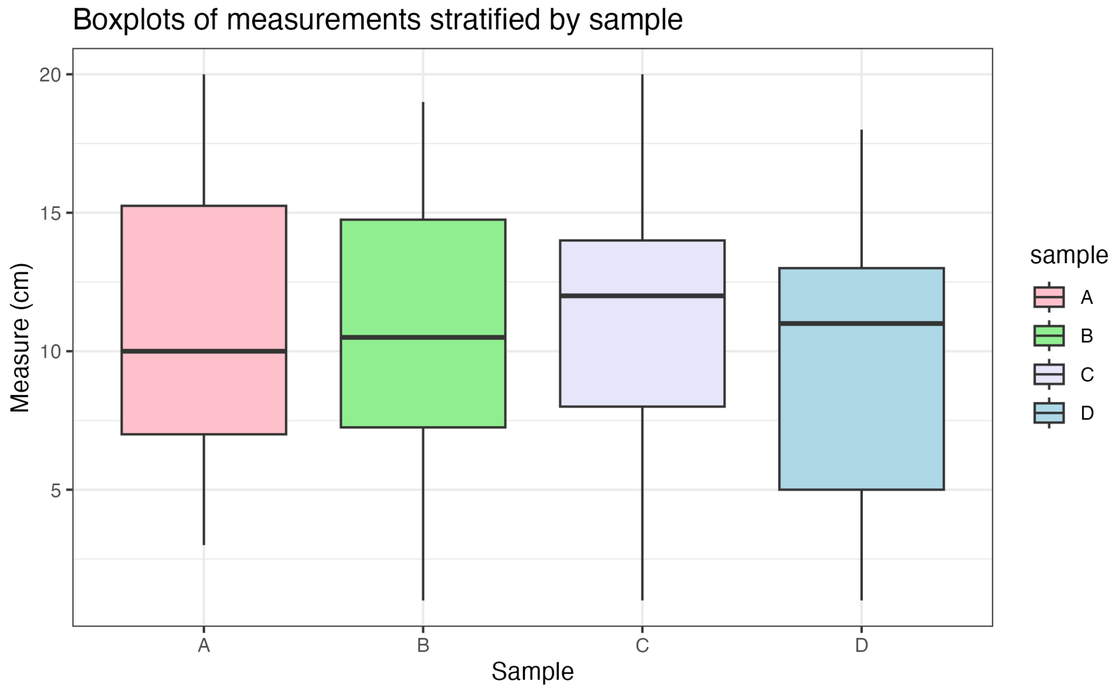
GGPLOT2 - EASY GRAPHICS

- Aesthetically pleasing graphics.
- Well-defined “additive” (+) structure.
- Integrates perfectly with tidy data.
- Great documentation & community

http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html

GGPLOT2 ADDITIVE STRUCTURE



DATASET, SAMPLES & OBSERVATIONS

```
ggplot(df,
       aes(x = Sample,
           y = Measure))
```

DEFINE PLOT TYPE

```
ggplot(df,
       aes(x = Sample,
           y = Measure)) +
  geom_boxplot()
```

COLOR BY GROUP

```
ggplot(df,
       aes(x = Sample,
           y = Measure,
           fill = Sample)) +
  geom_boxplot()
```

TITLE AND LEGEND

```
 … +
labs(title = "Basic boxplot")
```

CUSTOM COLORS

```
 … +
scale_fill_manual(values =
c("pink", "lightgreen",
"lavender", "lightblue"))
```

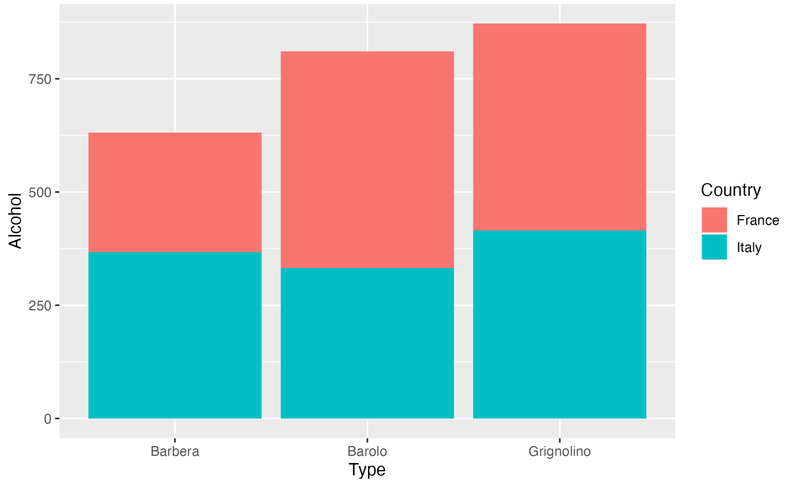
BACKGROUND

```
… +
theme_bw()
```

GGPLOT BASIC STRUCTURE

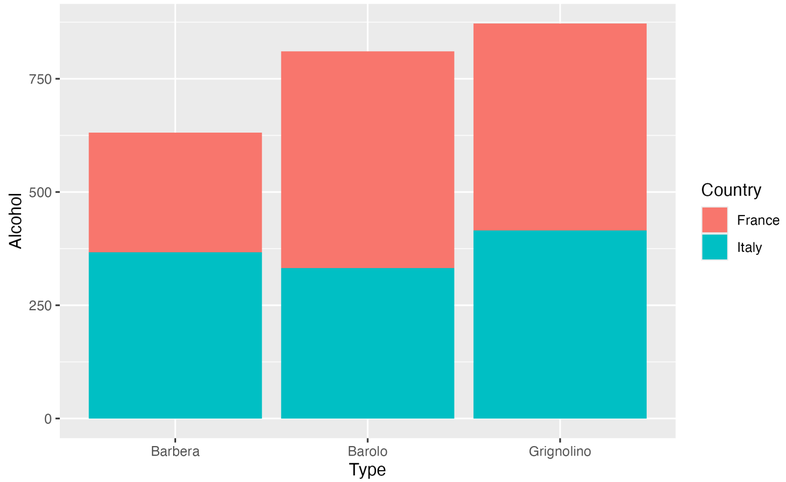
Geoms inherit the parameters from the ggplot they are added to:

```
ggplot(my_wine,
       aes(x = Type,
           y = Alc,
           fill = Country))+
  geom_col()
```



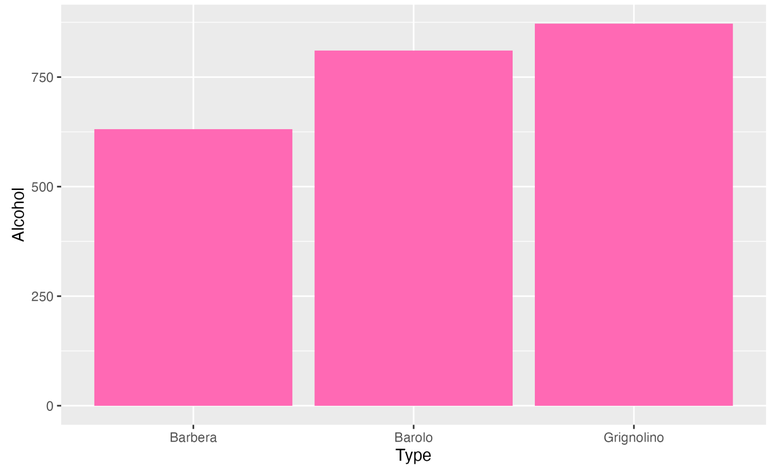
Alternatively, you can specify additional parameters directly in the geom:

```
ggplot(my_wine,
       aes(x = Type,
           y = Alc))+
  geom_col(aes(fill = Country))
```



Things outside the aes are applied to everything!

```
ggplot(my_wine,
       aes(x = Type,
           y = Alc))+
  geom_col(fill = ”hotpink”)
```



https://rstudio.com/wp-content/uploads/2016/11/ggplot2-cheatsheet-2.1.pdf

GGPLOT CHEAT SHEET

Define Plot:

```
ggplot(data = my.data,
       aes(x = x.var,
           y = y.var))
```

Add Plot Type:

```
+ geom_point()# scatter plot
+ geom_line()
+ geom_boxplot()
+ geom_col()
+ geom_density()
+ geom_histogram()
```

GET STARTED

One Color:

```
ggplot(…, aes(…),
color = "green")
```

Color Fill by Group:

```
ggplot(…, aes(…,
fill = z.var))
```

Custom Colors:

```
+ scale_[color/fill]_manual(values = c())
ex: scale_color_manual(values = c("blue", "pink"))
```

COLORS

More Colors:

```
+ scale_fill_grey(start = 0.2, end = 0.8)
+ scale_fill_gradient(low="white", high="red")
```

Labels:

```
+ labs(title = "Title",
       x = "X label",
       y = "Y label")
```

COLOR SCALES & THEMES

Theme:

```
+ theme_bw()
+ theme_minimal()
+ theme_dark()
+ theme_classic()
```

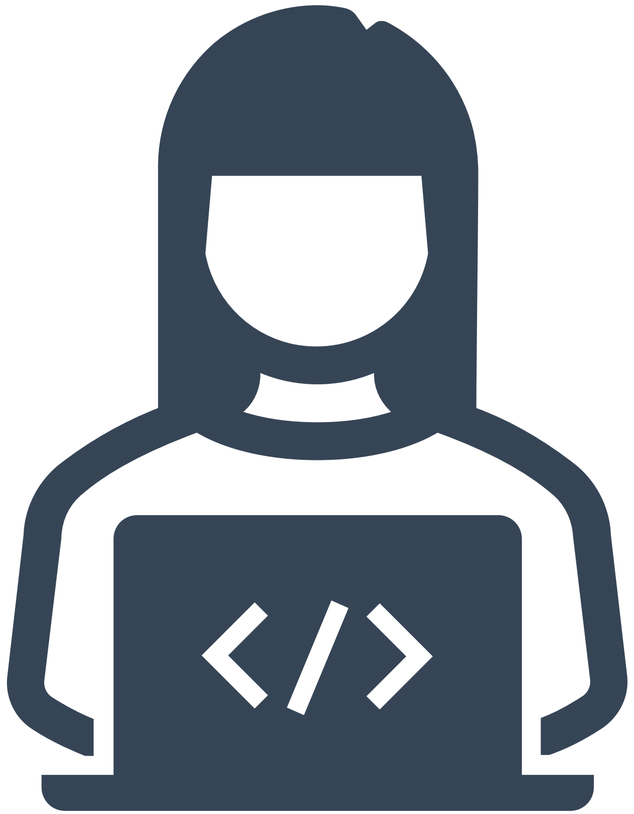
Legend:

```
+ theme(legend.position = *)
* = "none", "top", "bottom", "left", "right"
```

TEXT

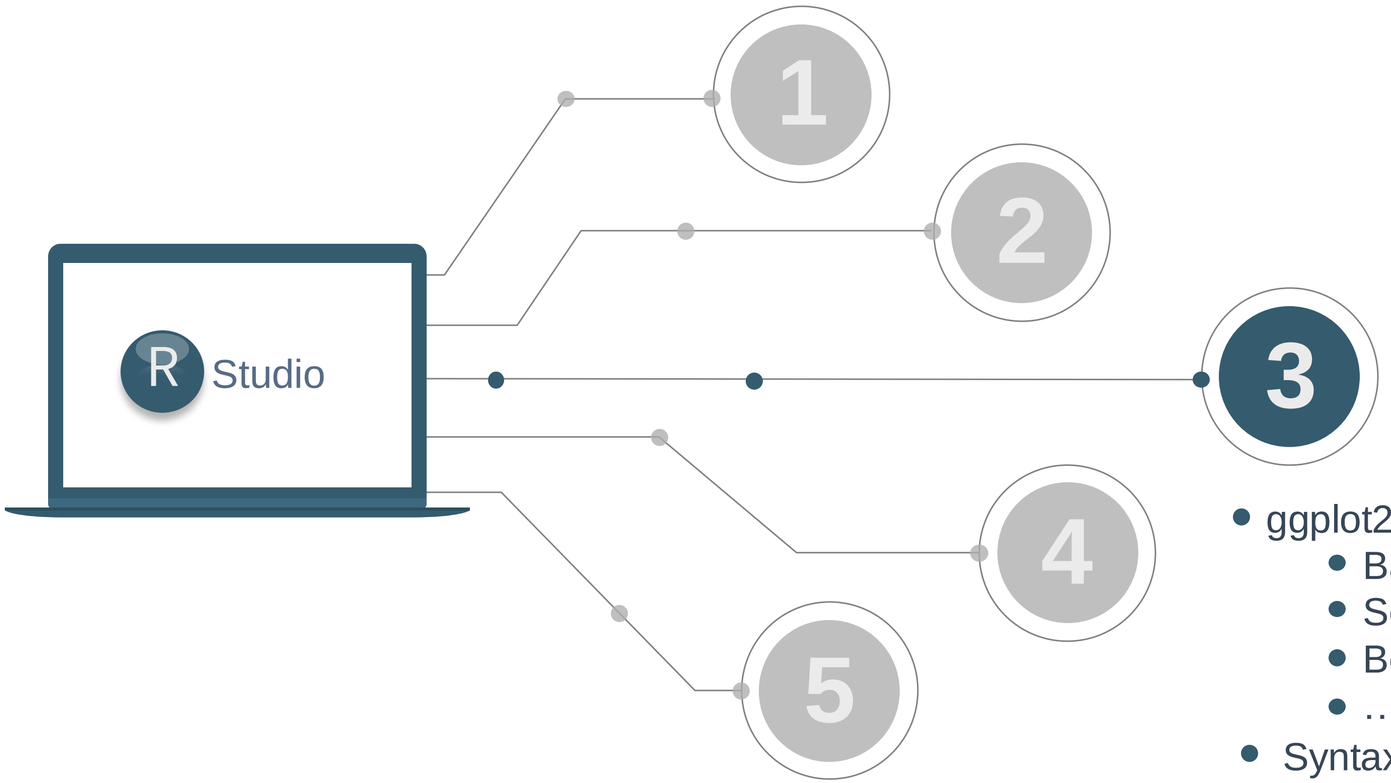
FROM EXCEL TO R

LIVE CODING 3 – GGPLOT2



GGPLOT2

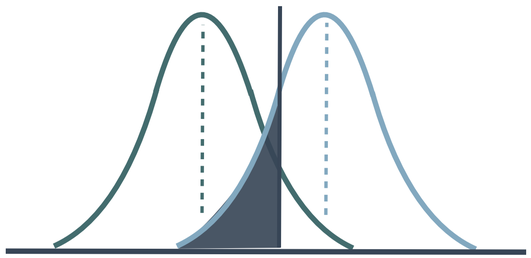
EXERCISE 3

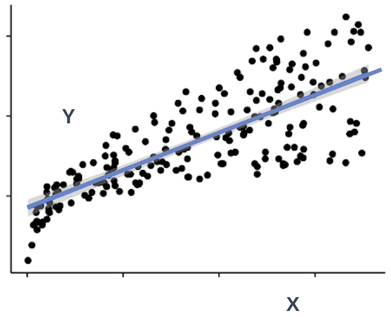


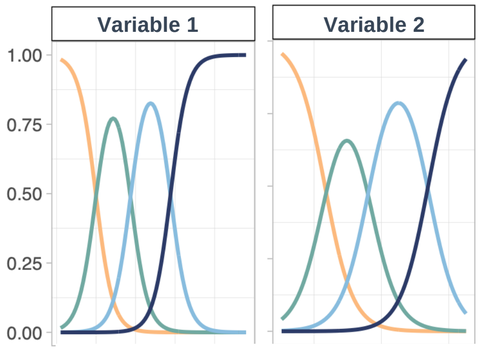
- ggplot2
  - Bar plot
  - Scatter plot
  - Boxplot
  - …
- Syntax
- Aesthetics

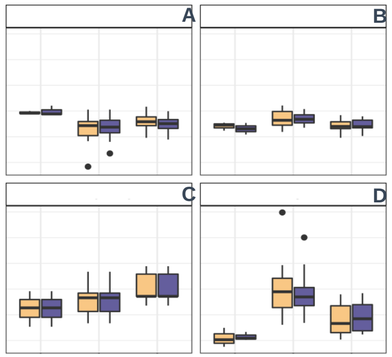
FROM EXCEL TO R

TUGCE STATS FROM HERE









http://www.sthda.com/english/

R - A STATISTICAL SCRIPTING LANGUAGE

MODEL FUNCTIONS

```
lm(), glm()
lmer(), glmer(),
nls(), …
```

EMMEANS PACKAGE

```
emmeans(),
pairs(), cld()
```

APPLY TO MODEL

```
summary(), anova(),
confint(), predict(),
drop1(), update(),
step(), …
```

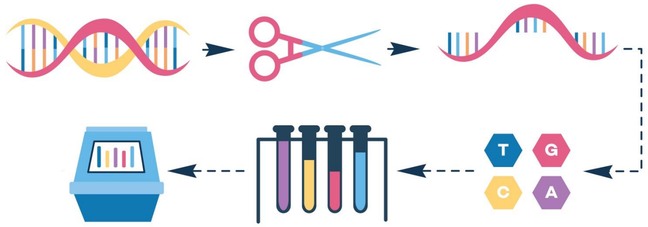
MORE FUNCTIONS

```
t.test(), cor(),
cor.test(), aov(),
quantile(),
p.adjust(),
rank(), …
```

Let’s use R in a statistical analysis

YOU

The researcher with R skills!

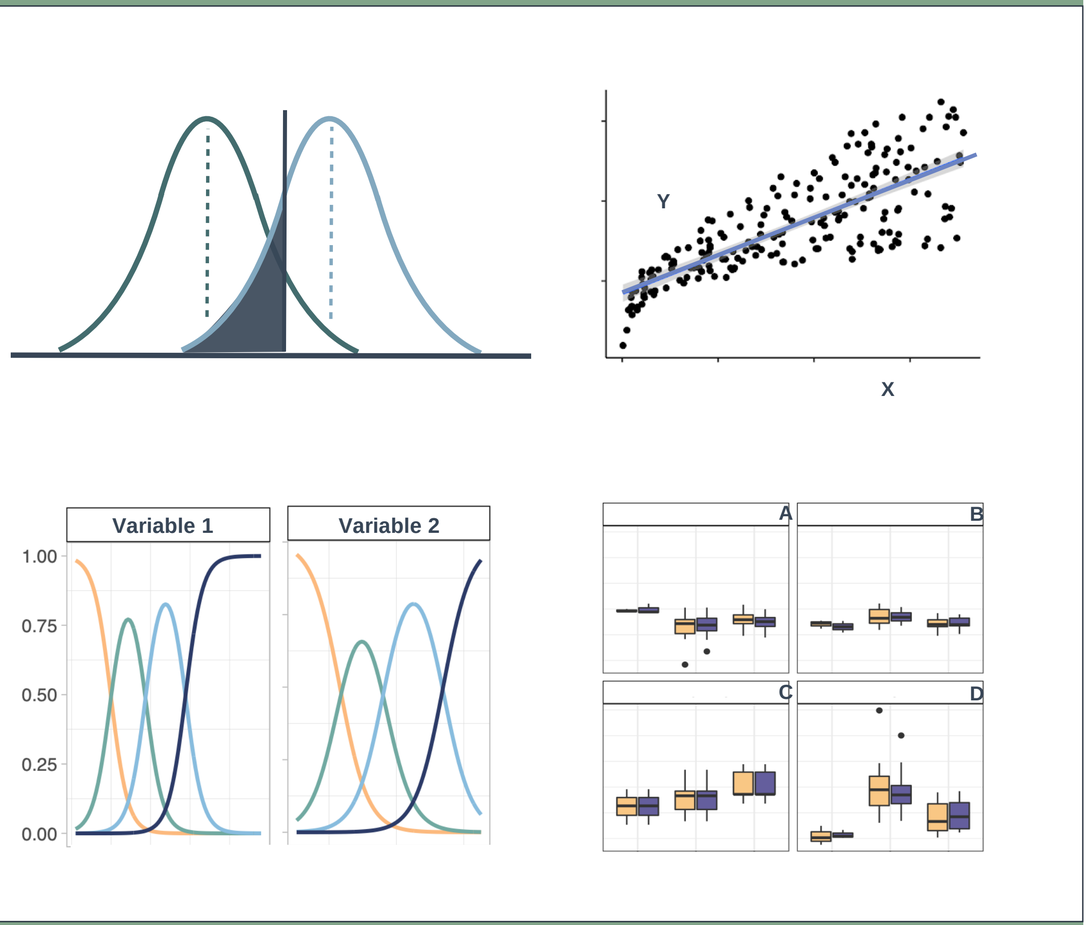


YOUR DATA

Gene expression data, i.e. how much a gene is used to produce a protein. Multiple genes measured in one experiment.

YOUR DISEASE OF INTEREST

Psoriasis, an immune-mediated skin disorder.



http://www.sthda.com/english/

R - A STATISTICAL SCRIPTING LANGUAGE

During this session:

• *Cooperatively* discuss and share ideas about the data

• Apply steps of basic statistical analysis for hypothesis testing consistent with the given data

• Suggest conclusions based on your analysis, regarding the association between psoriasis and gene expression levels

https://rstudio.com/resources/cheatsheets/ for various relevant cheat sheets.
Other example: https://www.dummies.com/programming/r/statistical-analysis-with-r-for-dummies-cheat-sheet/

STATS CHEAT SHEET

GET STARTED

Import Data:

```
read_excel("my.data.xlsx")
```

Overview of Data:

```
summary(my.data)
nrow(my.data)
length(my.data)
names(my.data)
```

REGRESSION MODELS

Linear:

```
lm(y~x, data=my.data)
confint(model)
```

Logistic:

```
glm(y~x,
data=my.data)
```

Linear Mixed:

```
lmer(y~x + (1|z),
data=my.data)
```

Check Model:

```
summary(model)
par(mfrow=c(2,2))
plot(model)
```

TESTS/COMPARISONS

ANOVA:

```
anova(model2, model1)
```

F-Test:

```
drop1(model, test="F")
```

Emmeans:

```
emmeans(model,~x)
pairs(emmeans(model,~x))
```

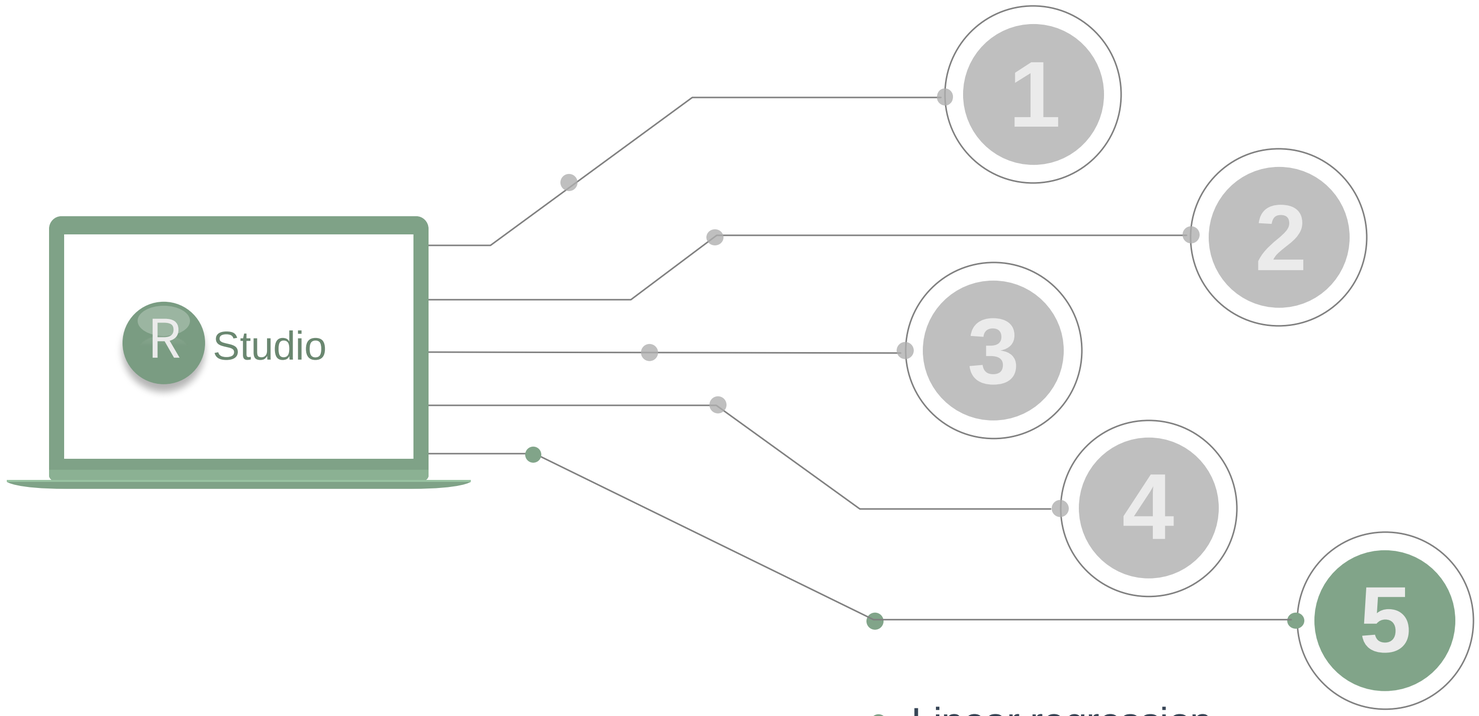
VARIABLES

Check Type:

```
table(my.data$x)
is.numeric(my.data$x)
is.factor(my.data$x)
```

Change Type:

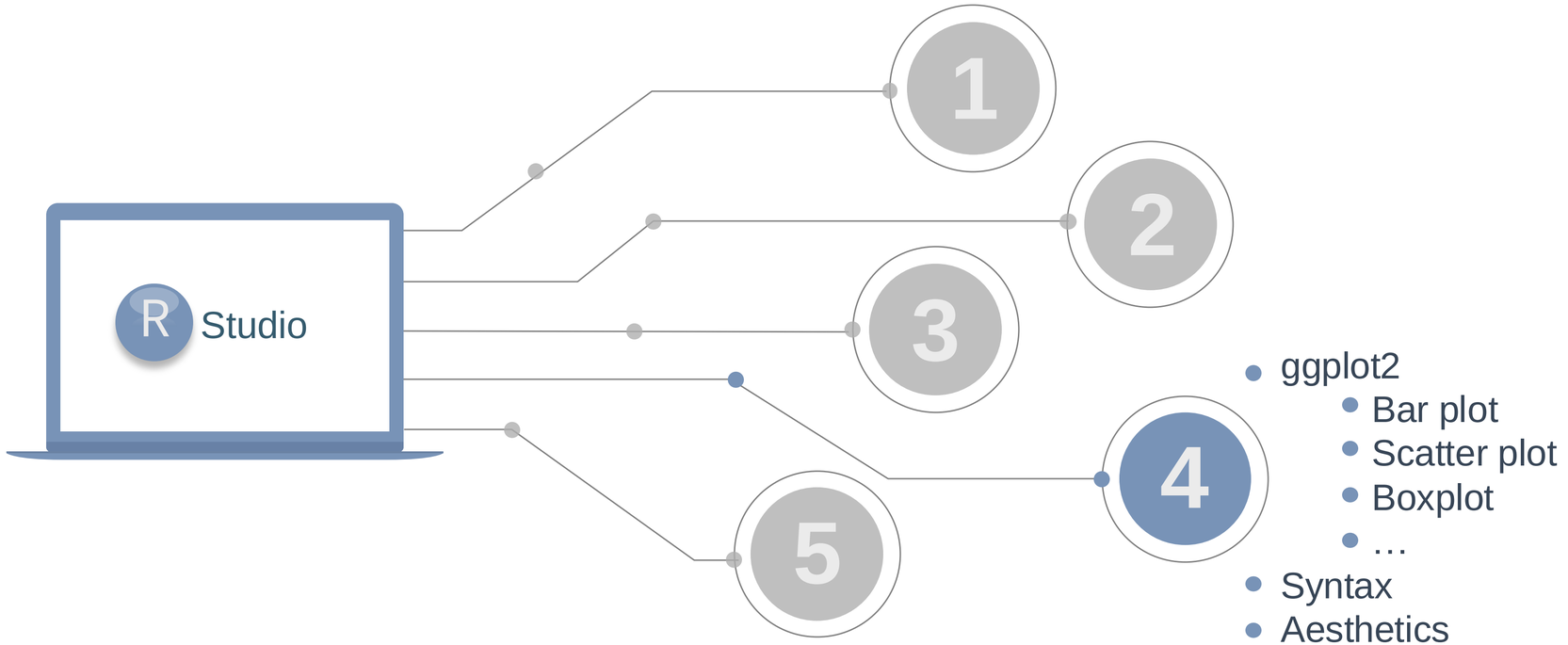
```
my.data <- mutate(my.data, x = factor(x))
my.data$z <- as.numeric(my.data$z)
```



- Linear regression
- Summary Statistics
- ANOVA
- Logistic regression
- Clustering
- Correlation

Statistics in R

EXERCISE **5**



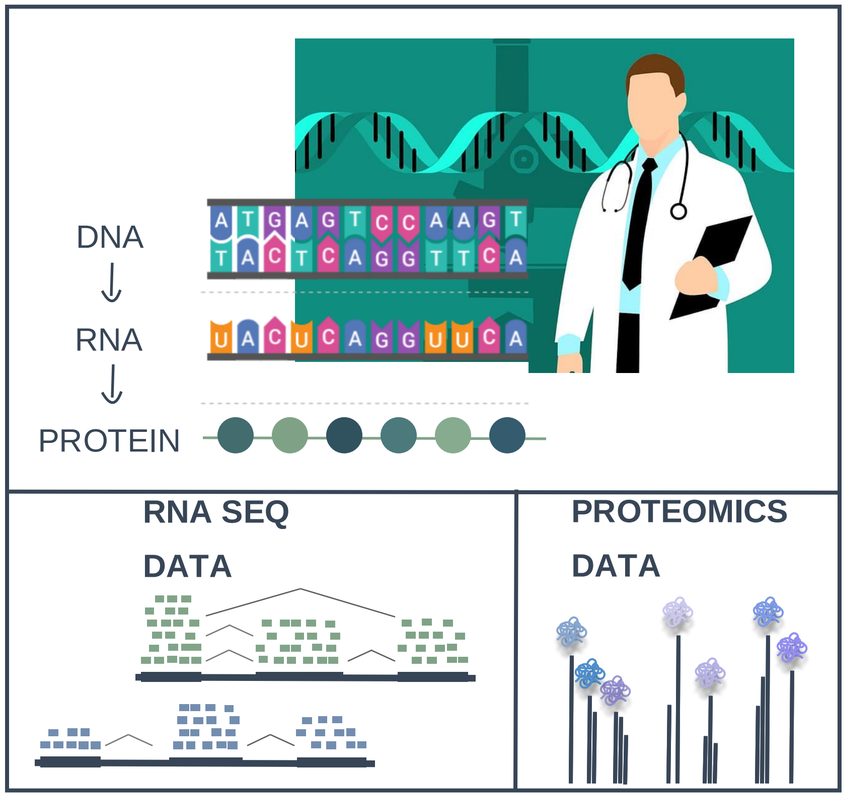
GGPLOT2

EXERCISE **4**

https://www.r-graph-gallery.com/

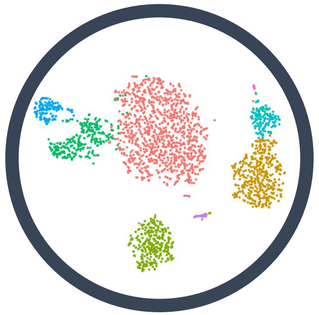
BIOINFORMATICS IN **R**

HIGH THROUGHPUT DATA

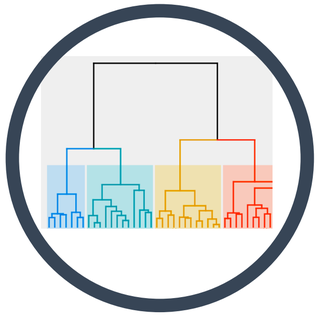


BIOINFORMATIC ANALYSIS

DIMENSIONALITY REDUCTION

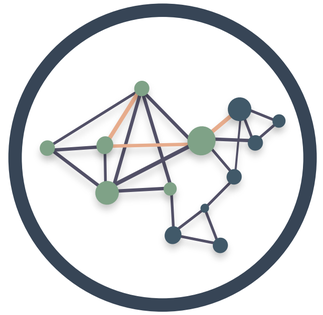


CLUSTERING



MACHINE LEARNING

NETWORK ANALYSIS



THE TOP OF THE R ICEBERG

STATISTICAL ANALYSIS

Statistical models (linear, generalized, mixed, …)

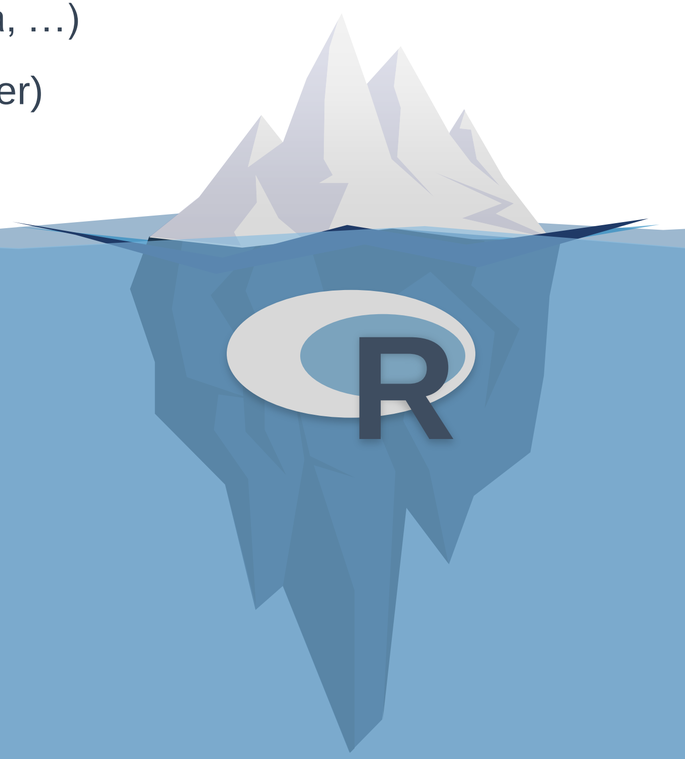
Statistical tests (t-test, chisq, anova, …)

Survival analysis (Cox, Kaplan meier)

DATA MANGEMENT

EASY PLOTTING

BIOINFORMATIC ANALYSIS



WANT MO-R-E?

The Section for Biostatistics offers a number statistics-oriented R courses:

https://publichealth.ku.dk/about-the-department/biostat/

Spring

Basic Statistics for Health Researchers (Danish course)
ECTS: 9,0

Epidemiological methods in medical research
ECTS: 7

Advanced topics in health research B
ECTS: 2,8

Statistical methods in bioinformatics
ECTS: 3,5

Statistical analysis of survival data
ECTS: 4,9

Programming and statistical modelling in R
ECTS: 1,6

Bayesian methods in biomedical research
ECTS: 2,4

Psychometric validation of patient reported outcome measures
ECTS: 2,5

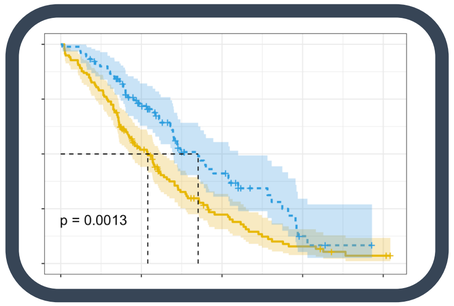
Introduction to validation of patient reported outcome measures
Causal inference I
ECTS: 2,5

Use of the statistical software R
ECTS: 2,1

* These are screenshots. Go to the website and scroll down to ‘Teaching’

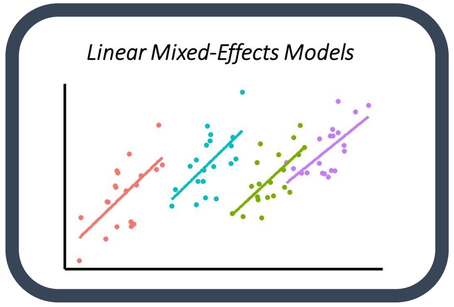
TEASER

STATISTICS in **R**



Survival Analysis

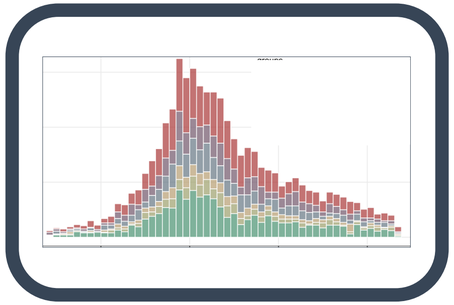
survival: https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r/
survminer: https://cran.r-project.org/web/packages/survminer/survminer.pdf
(https://rpkgs.datanovia.com/survminer/)



Mixed-Effects Models

lme4: https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf

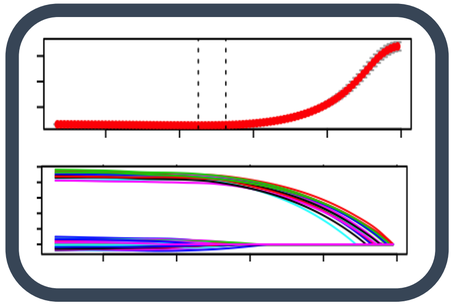
https://cran.microsoft.com/snapshot/2017-08-01/web/packages/sjPlot/vignettes/sjplmer.html
glmmTMB: https://cran.r-project.org/web/packages/glmmTMB/index.html



Epidemiological Analysis

Epi: https://cran.r-project.org/web/packages/Epi/index.html

pubh: https://rviews.rstudio.com/2020/03/05/covid-19-epidemiology-with-r/
https://cran.r-project.org/web/packages/incidence/vignettes/customize_plot.html
https://rviews.rstudio.com/2020/03/05/covid-19-epidemiology-with-r/



Elastic-Net Regression

(R

glmnet: https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

elasticnet: https://cran.r-project.org/web/packages/elasticnet/elasticnet.pdf

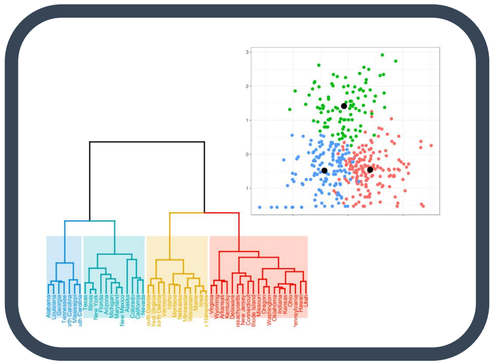
https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net

TEASER

Machine Learning

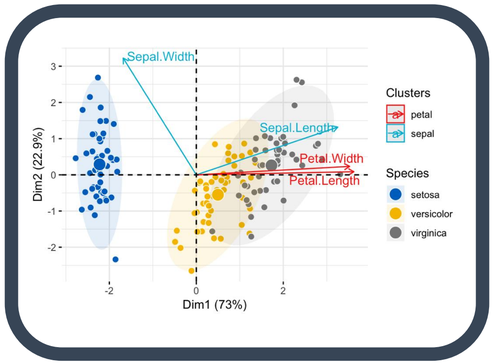
https://lgatto.github.io/IntroMachineLearningWithR/an-introduction-to-machine-learning-with-r.html

Clustering



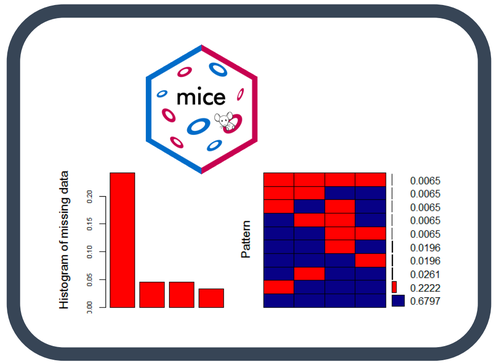
https://statsandr.com/blog/clustering-analysis-k-means-and-hierarchical-clustering-by-hand-and-in-r/

Feature Selection: PCA



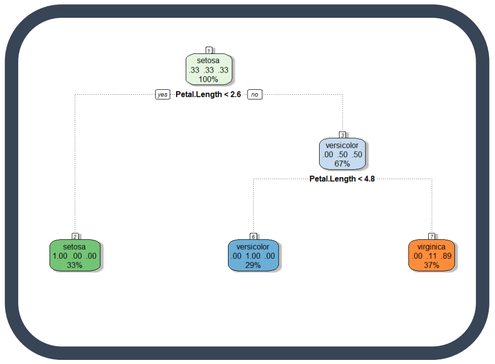
https://bioconductor.org/packages/release/bioc/vignettes/PCAtools/inst/doc/PCAtools.html

Missing Data



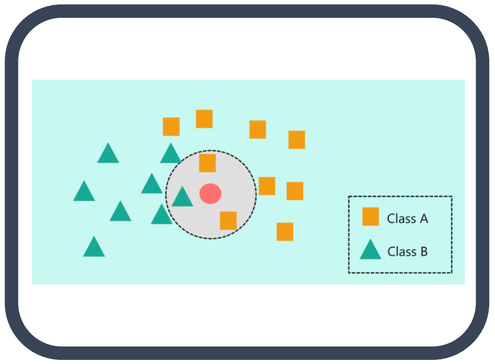
https://amices.org/mice/
https://datascienceplus.com/imputing-missing-data-with-r-mice-package/

Random Forest



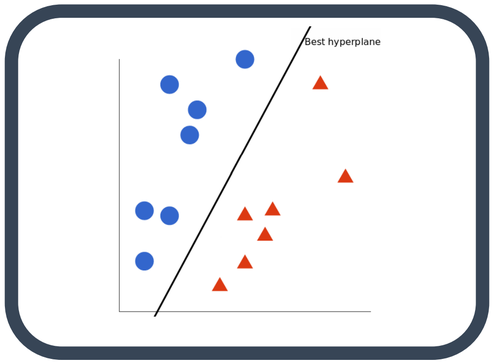
https://www.blopig.com/blog/2017/04/a-very-basic-introduction-to-random-forests-using-r/

kNN



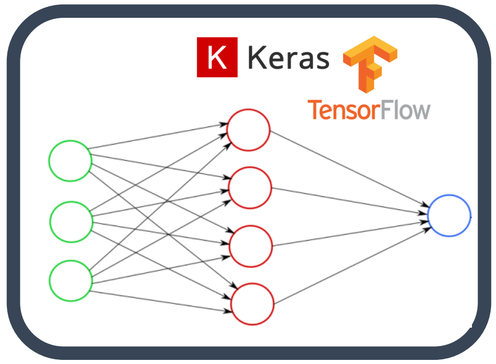
https://www.edureka.co/blog/knn-algorithm-in-r/

SVM



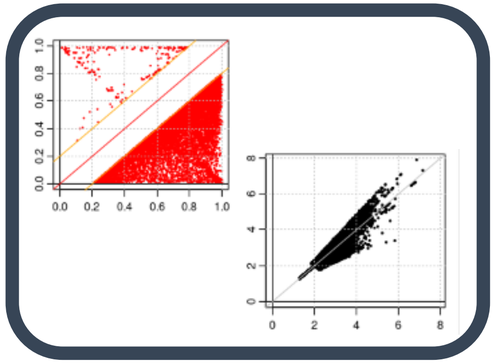
https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf

Neural Networks



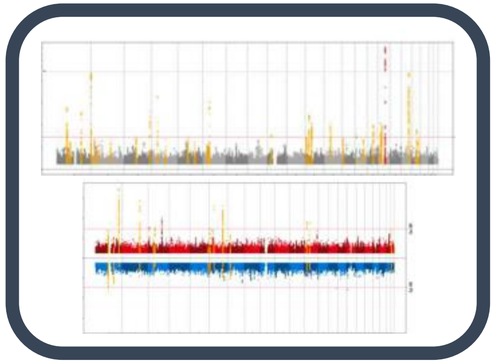
https://keras.rstudio.com/
https://tensorflow.rstudio.com/

GWAS - QC & Data Harmonization



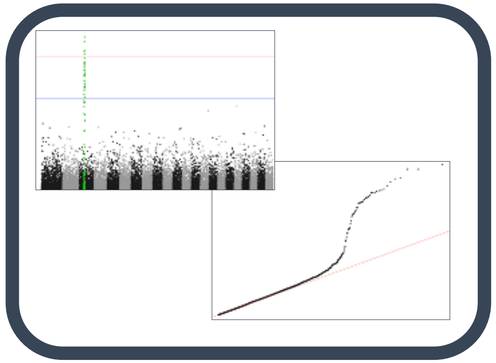
EasyQC: https://www.uni-regensburg.de/medizin/epidemiologie-praeventivmedizin/genetische-epidemiologie/software/

GWAS Data Management & Plots



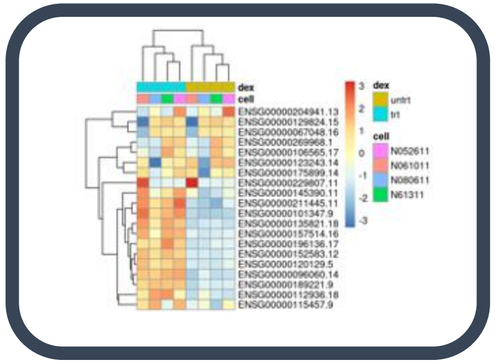
EasyStrata: https://www.uni-regensburg.de/medizin/epidemiologie-praeventivmedizin/genetische-epidemiologie/software/

More Plotting…



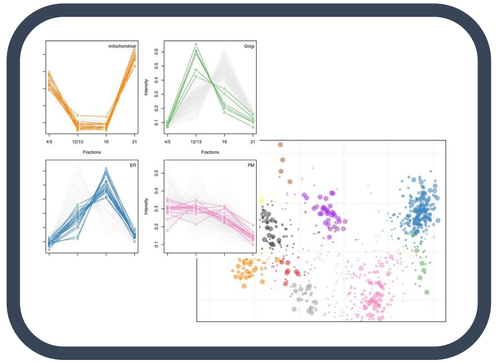
Manhattan and QQ plots: https://cran.r-project.org/web/packages/qqman/vignettes/qqman.html

Gene Expression Analysis



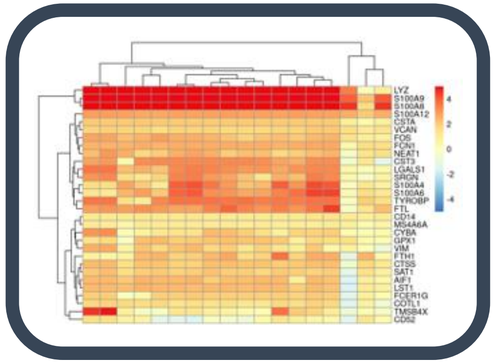
DESeq2, limma, EdgeR, etc.: http://www.bioconductor.org/packages/release/BiocViews.html#___RNASeq

Proteomics Analysis



RforProteomics: http://www.bioconductor.org/packages/release/BiocViews.html#___Proteomics RforProteomics.html

Single-Cell RNASeq



https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf

TEASER

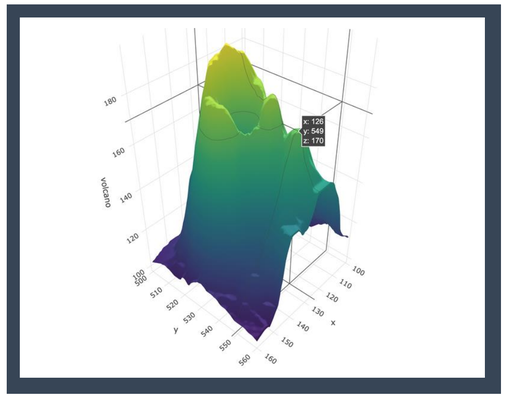
Omics Data

http://www.bioconductor.org/packages/release/BiocViews.html

FROM EXCEL TO R

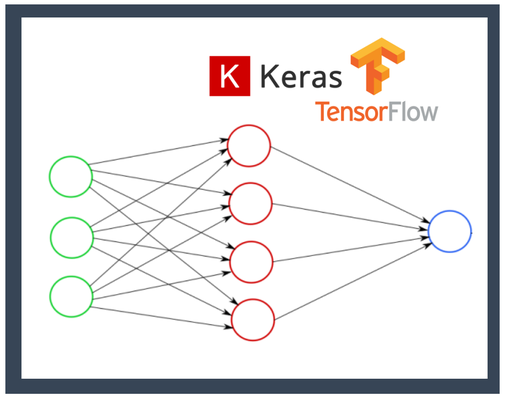
COOL STUFF IN R

PLOTTING IN 3D



https://plotly-r.com/d-charts.html

DEEP LEARNING



https://keras.rstudio.com/
https://tensorflow.rstudio.com/

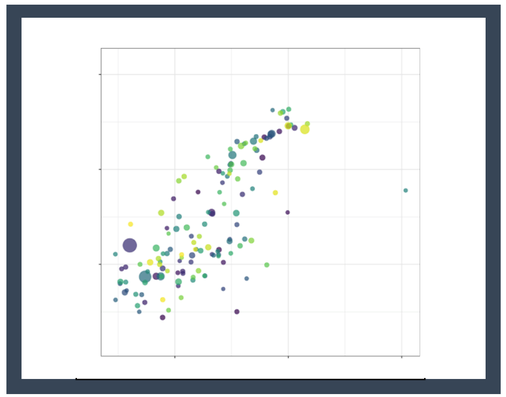
BAYESIAN STATISTICS

https://mc-stan.org/users/interfaces/rstan

WEBPAGE WITH R SHINY

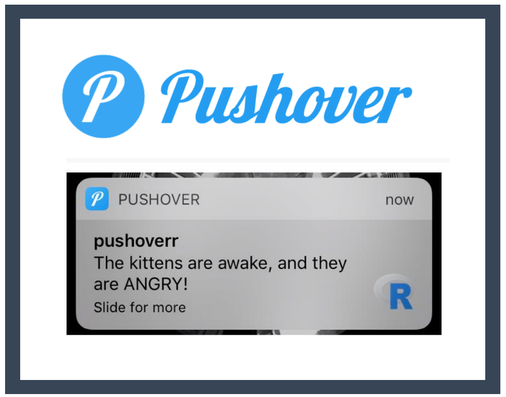
https://shiny.rstudio.com/

INTERACTIVE PLOTS



https://gganimate.com/articles/gganimate.html

MAIL AND MESSAGES



https://github.com/briandconnelly/pushoverr

THANK YOU FOR LISTENING