Make your paper figures professionally: Scientific data analysis and visualization in R

Julien Gagneur
Gagneur lab - Computational biology

gagneurlab.in.tum.de

To understand the genetic basis of gene regulation and its implication in diseases

You've been building algos and software processing data…

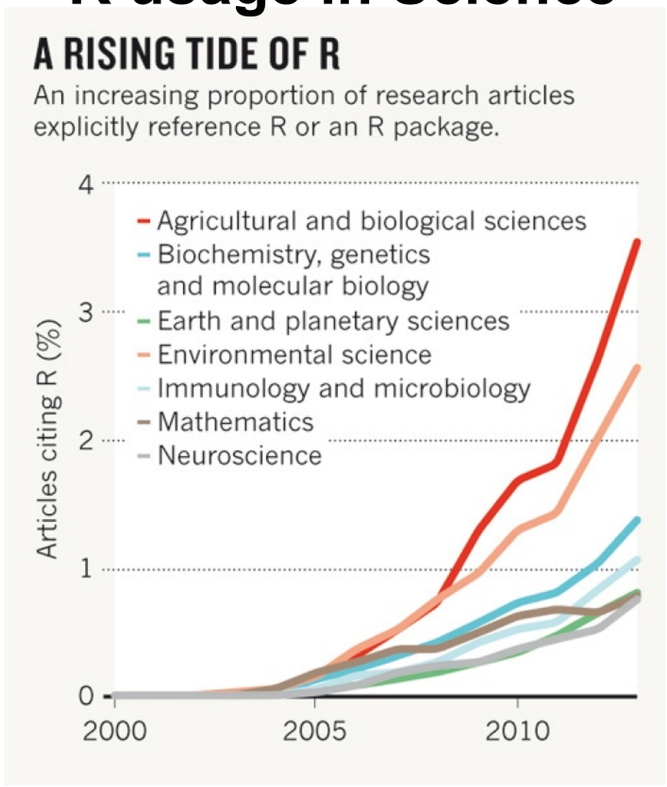
… now sail!

Caution: different skills required!

R: A data analyst-oriented language

- Written by statisticians for statisticians
- GNU, Open source
- Widely used
- Thousands of packages available for specific statistical analysis and application domains.

R usage in Science



https://www.youtube.com/watch?v=TR2bHSJ_eck

R: A data analyst-oriented language

- R is used **interactively** to explore data. The outcome of one analysis session are turned into **scripts**.

- Most valuable time is the human thinking time, not machine computing time. Data analysts spend most of your time in exploring the data and thinking about it, not in coding.

➡ High-level language, loosely typed.
e.g. variables are not declared

➡ Running time can be slow because the language takes care of what you don't need to care of.

Two sides of R programming

1. As **data analyst**
   scripting. Few functions and data structures

2. As **package developer**
   robust code that copes with flexibility of the language.

Our teaching team has experience with both.

This lecture focuses on 1.

Reproducibility: Reports with R Markdown

Analyze. Share. Reproduce.

Your data tells a story. Tell it with R Markdown.
Turn your analyses into high quality documents,
reports, presentations and dashboards.

Get: reading and manipulating data

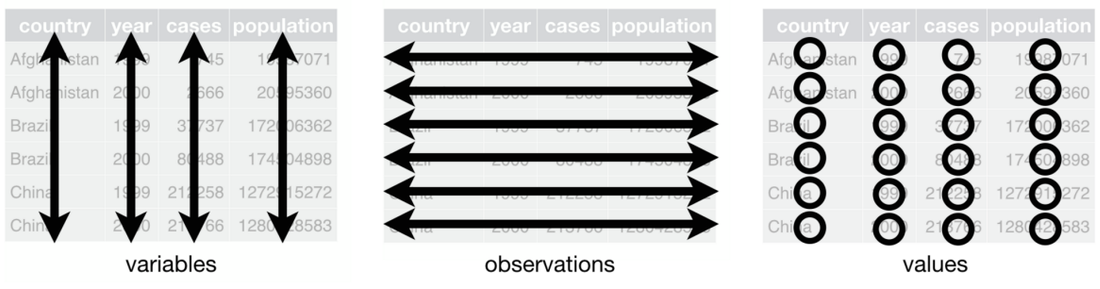
Data import

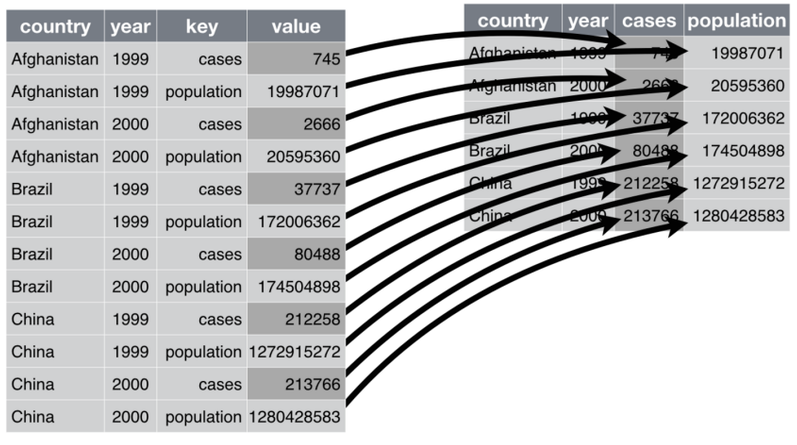
flat files, excel, XML, JSON, relational database

Tidy data

A standard way of structuring dataset for statistical analysis

data.table: Efficient in memory storage and operations



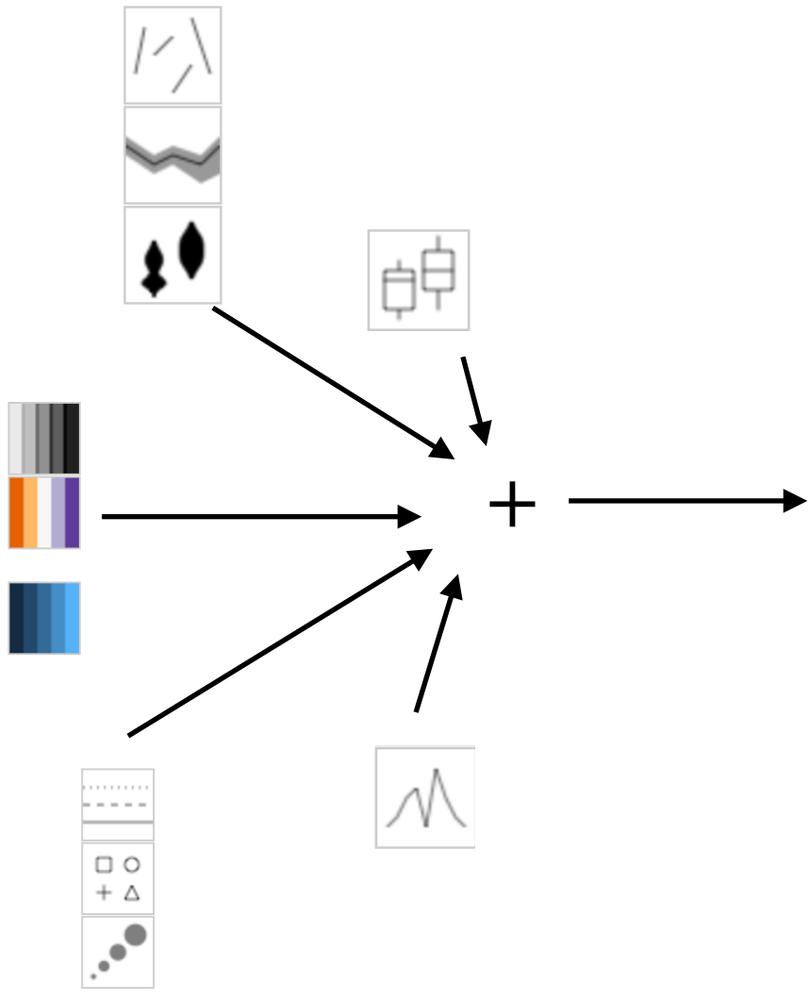


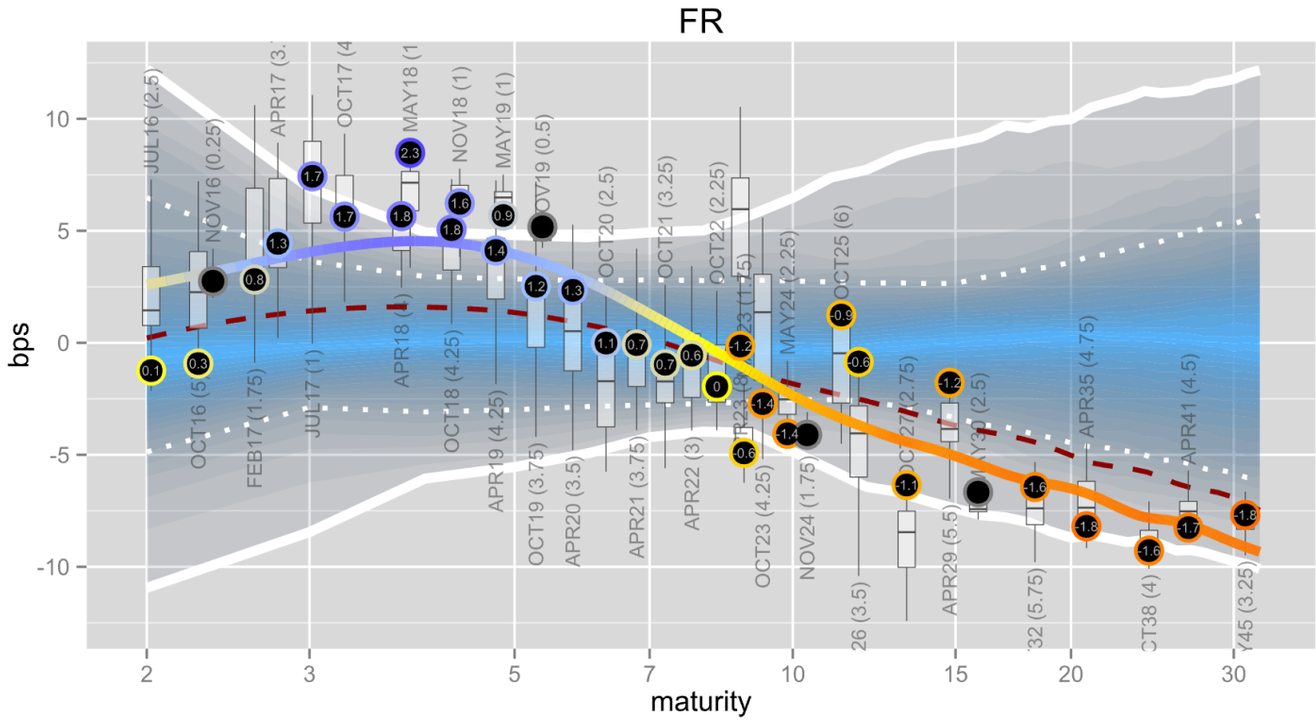
country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Look: Grammar of graphics with ggplot2

Flexible plotting using a set of independent complements that can be composed in many different ways





http://stackoverflow.com/questions/24828341

Conclude: Drawing robust conclusions

Not covered in the CEDOSIA module. See our master module "Data analysis and visualization in R"

Data exploration leads to **hypotheses** made on the data:

"Does smoking significantly associates with increased lung cancer?"

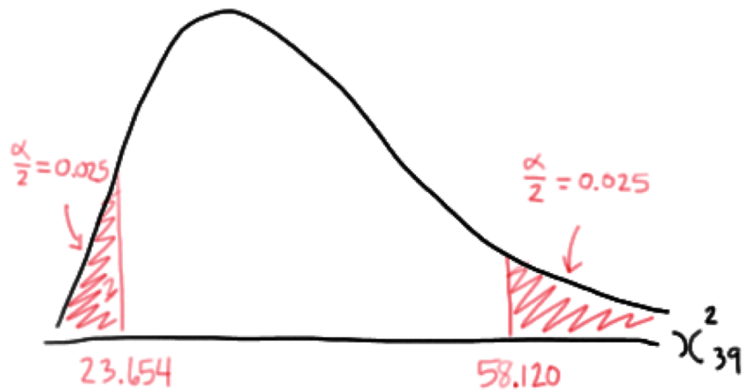
"Does smoking significantly associates with higher academic grades…

…. when I control for student age?

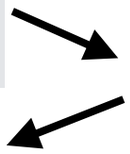
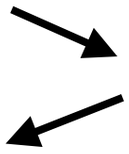
➡ **Statistical testing** (empirical and theoretical)

Classical tests: T-test, Wilcoxon, Fisher test

Generic resampling approaches



Syllabus

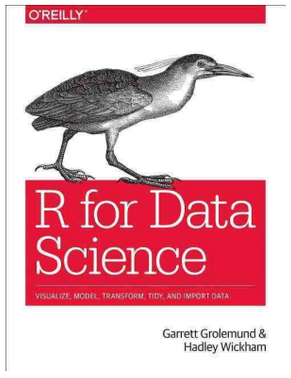
13.05.19	R basics (optional)	
20.05.19	Introduction	
	2’ flash presentations	
	Grammar of graphics	
	Data table	
	Tidy data I	 @home: prepare a dataset and code producing one plot
27.05.19	Tidy data II	
	Plot types	
	Advanced plots	
	Improve one of your own plots (in groups)	 @home: prepare a 10’ presentation & code producing plots
03.06.19	10' presentation group I	
	10' presentation group II	

Evaluation

Presence (signatures)

Honest attempt at last day presentation.

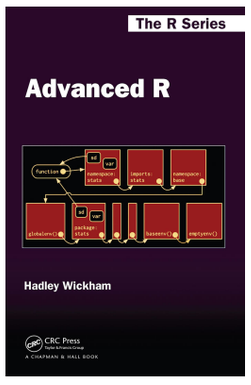
Recommended reading



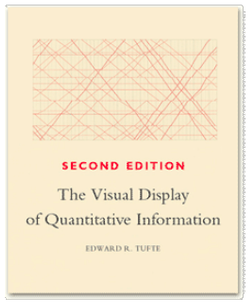
R for Data Science *, by Garrett Grolemund and Hadley Wickham

ModernDive

Modern Dive *, by Chester Ismay and Albert Y. Kim
(under construction)



Advanced R *, by Hadley Wickham



The Visual Display of Quantitative Information, Edward Tufte

* have a free online web version

Conclusion

- Data analysis is complementary to statistical methods and software development
- Get, look, (and conclude). Statistics and visualisation are both necessary.
- Reproducible analyses (Scripted reports).