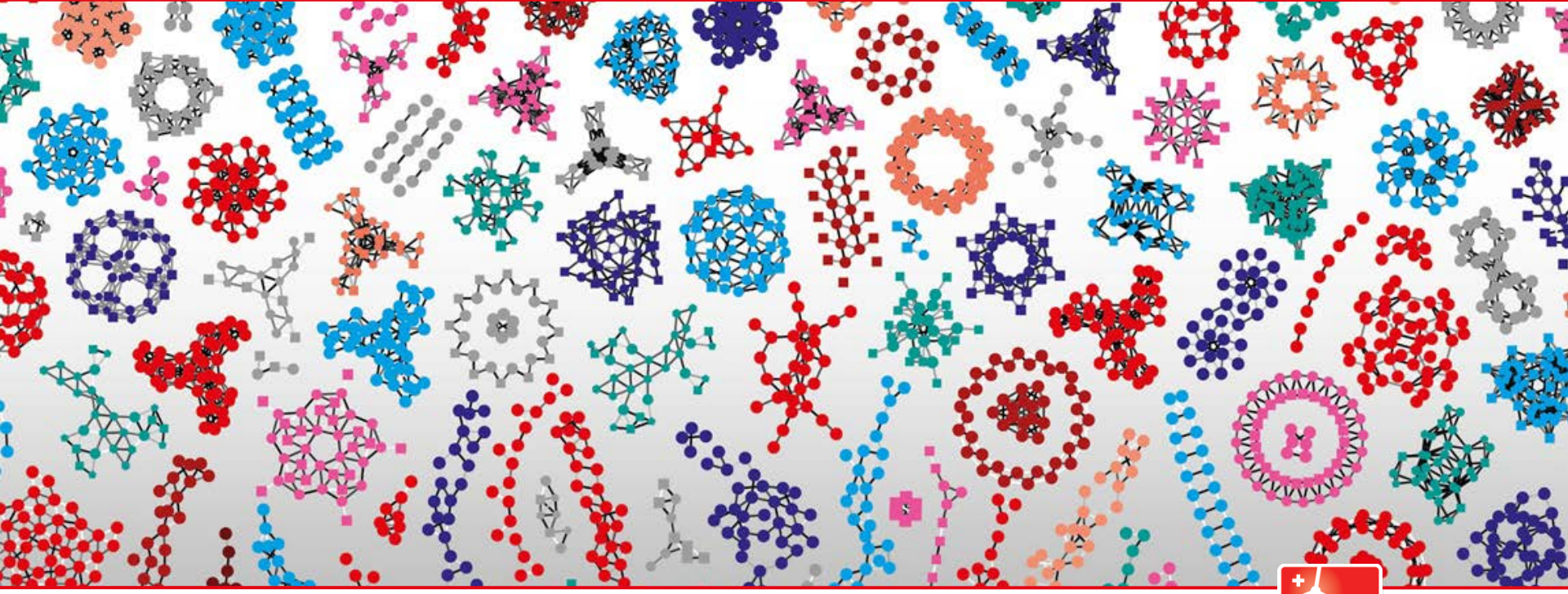


Swiss Institute of Bioinformatics

Advanced R
September 2024

Frédéric Schütz (Frederic.Schutz@sib.swiss)
Frédéric Burdet (Frederic.Burdet@sib.swiss)

Schedule

9:00 - 10:30	
10:30 - 10:45	break
10:45 - 12:15	
12:15 - 13:30	break
13:30 - 15:00	
15:00 - 15:15	break
15:15 - 16:45	
17:00	end of day

Today's Schedule

9:00 - 10:30	
10:30 - 10:45	break
10:45 - 12:15	
12:15 - 13:30	break
13:30 - 15:00	
15:00 - 15:15	break
15:15 - 16:45	
17:15	Apero at the Vortex

Introductions

What do you expect from this course ?

BIOSTATISTICS PLATFORM

Faculty of Biology and Medicine, UNIL

UNIL | Université de Lausanne

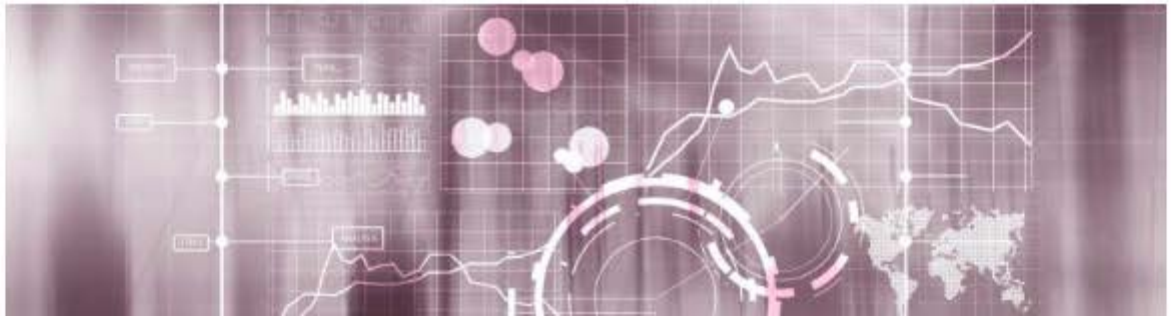
Home | Presentation | Services | Collaborations | Team and contact | More information

https://wp.unil.ch/biostatistics/

HOME

The Biostatistics platform of the Faculty of Biology and Medicine acts as a main entry point for all your questions related to biostatistics and data analysis. If needed, the platform can redirect you to the people with the best competences in order to answer your questions.

Contact: Frederic.Schutz@unil.ch



Are these valid R commands ?
If yes, what do they do ?
If no, why ?

```
c=c(c=c)
c=c(c="c")
```

How would you describe "good programming" ?

Reproducible research

The Economist, October 2013

THIS WEEK

EDITORIALS

FOOD Fertilizer is the best way to feed Africa right now **p.510**

WORLD VIEW Teach young scientists how to manage their labs **p.511**

VENICE City of Water edges closer to becoming a city under water **p.512**

Must try harder

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.

Science: Branch of knowledge or study dealing with a body of facts or truths systematically arranged. So says the dictionary. But, as most scientists appreciate, the fruits of what is called science are occasionally anything but. Most of the time, when attention focuses

for the first time only when problems in published studies are reported.

In private, scientists who run labs in even the most prestigious universities admit that they have little time to supervise and train all their students. Institutions such as the European Molecular Biology Labora-

Must try harder

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.

The finding resonates with a growing sense of unease among specialist editors on this journal, and not just in the field of oncology. Across the life sciences, handling corrections that have arisen from avoidable errors in manuscripts has become an uncomfortable part of the publishing process.

The evidence is largely anecdotal. So here are the anecdotes: unrelated data panels; missing references; incorrect controls; undeclared cosmetic adjustments to figures; duplications; reserve figures and dummy text included; inaccurate and incomplete methods; and improper use of statistics — the failure to understand the difference between technical replicates and independent experiments, for example.

It is usually the case that original data can be produced, mistakes corrected, and the findings of the corrected research paper still stand. At the very least, however, there is too little attention paid and too many corrections, which reflect unacceptable shoddiness in laboratories that risks damaging trust in the science that they, and others, produce.

The situation throws up many questions. Here are three of them. Who is responsible? Why is it happening? How can it be stopped?

The principal investigators (PIs) of any lab from which the work originates, especially if their names are on the paper, have an absolute and unavoidable responsibility to ensure the quality of the data from their labs, even if the main work is done by experienced postdocs. Officially, postdocs and graduate students are still in training, and it is the PI's job to make sure they are properly trained — in statistics and appropriate image editing, for a start. It is unacceptable for lab heads —who are happy to take the credit for good work — to look at raw data

There is an opportunity here for 'minimum threshold' journals, such as *PLoS ONE* and *Scientific Reports*. Editors and referees cannot be expected to divine when only positive data are included and inconvenient results left out, but journals should encourage online presentation of the complete picture. And scientists should offer it. The complete picture is, after all, what this science of ours strives to provide. ■

Nature 483, 509 (29 March 2012)

Nature editorial: "Must try harder"

Glenn Begley and Lee Ellis analyze the low number of cancer-research studies that have been converted into clinical success, and conclude that **a major factor is the overall poor quality of published preclinical data.**

[…]

The overall impression the article leaves is of ***insufficient thoroughness in the way that too many researchers present their data*.**

Nature 483, 509 (29 March 2012)

Reproducible Research:

“Research is reproducible if it can be reproduced by others”

Reproducible research

Of course, rerunning an experiment will give different results—an observation that gave rise to the development of statistics as a discipline.

Our focus here is “reproducible research” (RR) in the sense of reproducing conclusions from a single experiment based on the measurements from that experiment.

Definition of reproducible research

A **complete description** of the data and the analysis of that data — including computer programs — so the **results can be exactly reproduced** by others.

Amstat News, 1 January 2011

How complicated is it ?

PNAS

RESEARCH ARTICLE | SOCIAL SCIENCES |

Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty

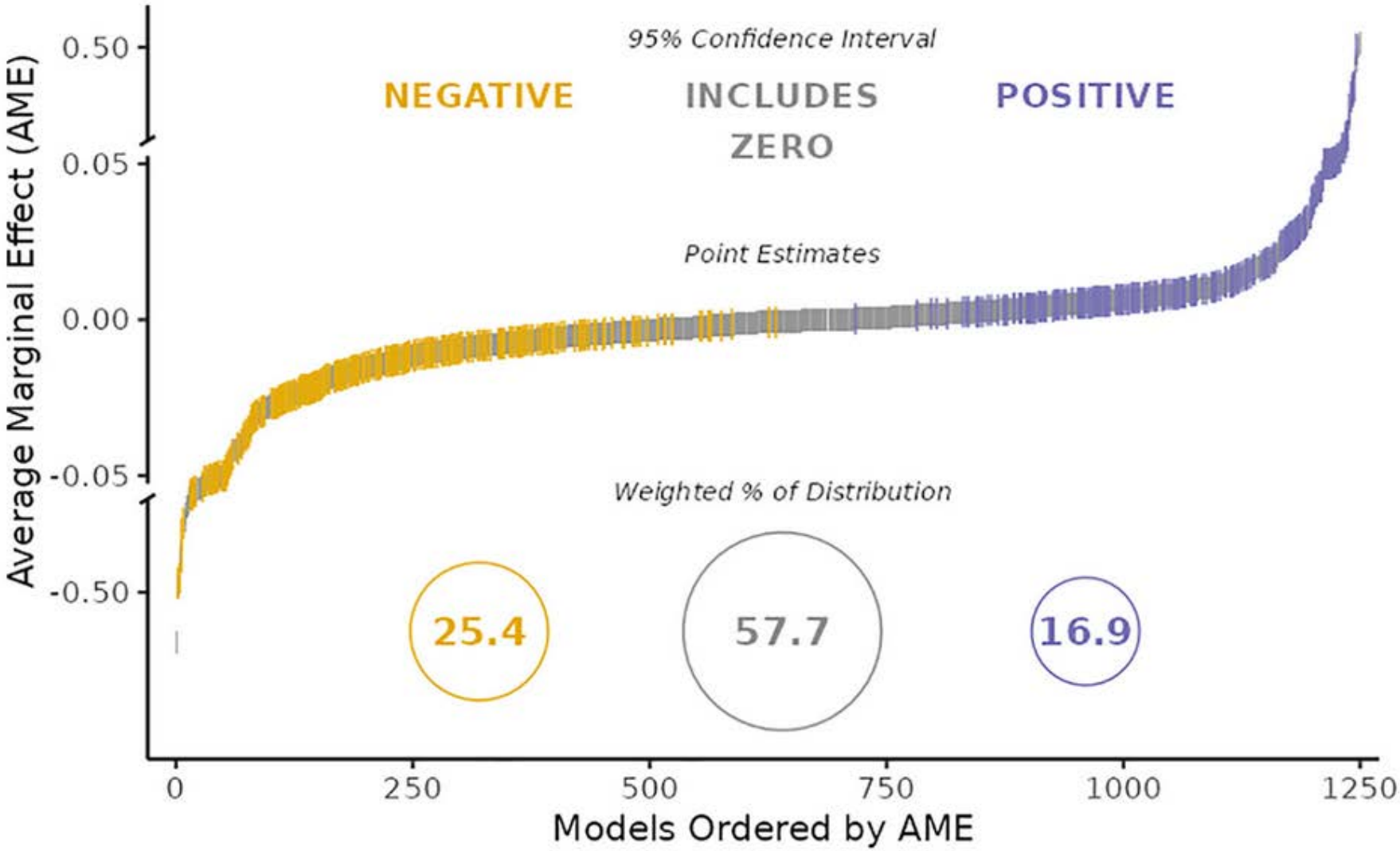
Nate Breznau, Eike Mark Rinke, Alexander Wuttke, +162, and Tomasz Żółtak Authors Info & Affiliations

Edited by Douglas Massey, Princeton University, Princeton, NJ; received March 6, 2022; accepted August 22, 2022

October 28, 2022 | 119 (44) e2203150119 | https://doi.org/10.1073/pnas.2203150119

We coordinated 161 researchers in 73 research teams and observed their research decisions as they used the same data to independently test the same prominent social science hypothesis.

Fig. 1.


95% Confidence Interval
NEGATIVE
INCLUDES ZERO
POSITIVE
Point Estimates
Weighted % of Distribution
25.4
57.7
16.9
Average Marginal Effect (AME)
0.50
0.05
0.00
-0.05
-0.50
0
250
500
750
1000
1250
Models Ordered by AME

Submitted to the Annals of Applied Statistics
arXiv: math.PR/0000000

DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

By Keith A. Baggerly* and Kevin R. Coombes†

U.T. M.D. Anderson Cancer Center

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report, we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

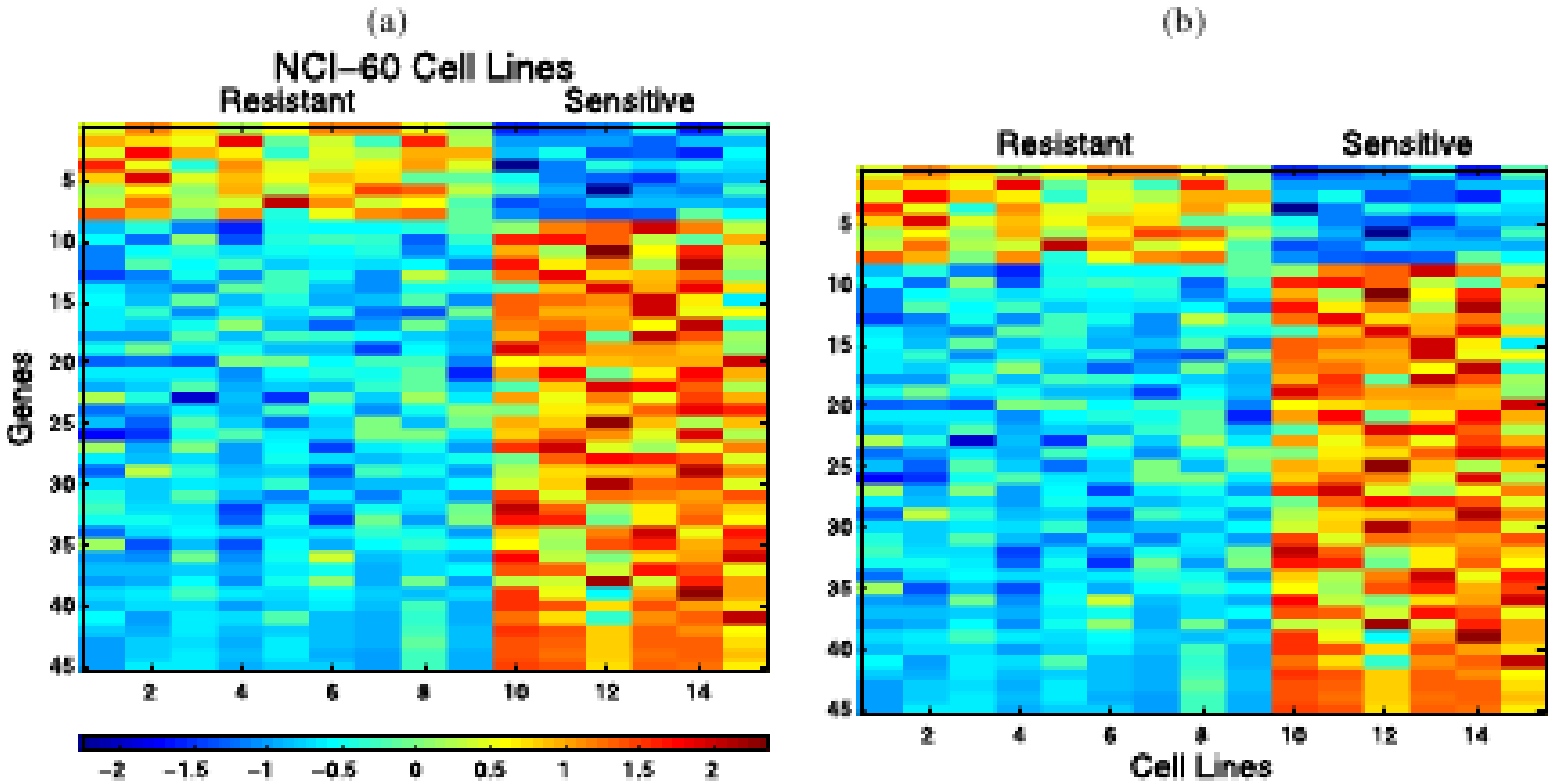


FIG. 4. *Approximations to* (a) *the heatmap initially presented in Figure* 4A *of Augustine et al.* (2009) *for temozolomide, with lines reportedly chosen from the NCI-60 cell line panel, and* (b) *the heatmap presented in Figure* 1 *of Hsu et al.* (2007) *for cisplatin, with cell lines chosen from the* 30-*line panel of Györffy et al.* (2006). *The heatmaps are the same. We have independently generated the cisplatin heatmap using the Györffy et al.* (2006) *data, but the temozolomide heatmap is neither for temozolomide nor from the NCI-60 panel.*

Conclusions from Baggerly and Coombes

"Poor documentation led a report on drug A to include a heatmap for drug B and a gene list for drug C.

These results are based on simple visual inspection and counting, and are not documented further."

Corrections made in the journal led to **further problems**.

Conclusions of Baggerly and Coombes

- The most common problems are simple:
  - confounding in the experimental design
  - mixing up the gene labels (off-by-one errors)
  - mixing up the group labels (sensitive/resistant)

- Most of these mix-ups involve simple switches or offsets.

- These mistakes are easy to make, particularly if working with Excel

- … and/or if working with 0/1 labels instead of names

We know we will make mistakes.
So…

1) We should work in a way that will reduce the number of errors


NO SMOKING
IN LAVATORY
C1160-03-197


SWISS
HB-JCA
HB-JCA


NO SMOKING
IN LAVATORY
C1160-03-197
C1168-01-196

"Regardless of whether smoking is allowed in any other part of the airplane, lavatories must have self-contained, removable ashtrays located conspicuously on or near the entry side of each lavatory door "

US Code of Federal Regulations for airworthiness

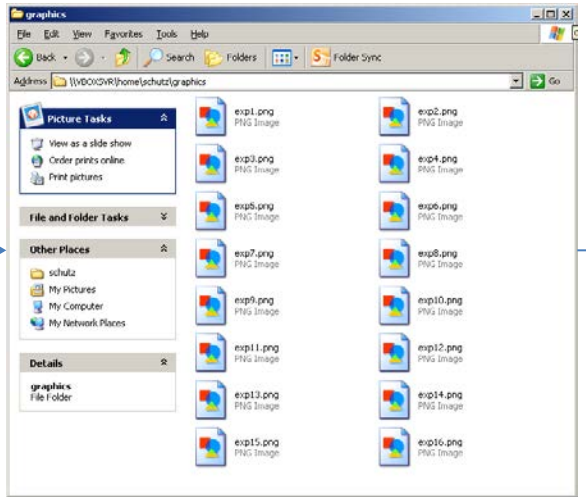
https://www.law.cornell.edu/cfr/text/14/25.853


JK
VARIG
VARIG
BRASIL

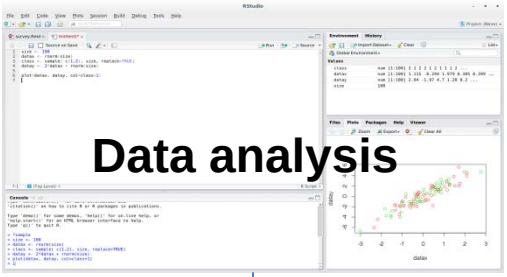
We know we will still make mistakes !

2) We must be able to track what we have done, so we can later find mistakes and correct them

Graphics



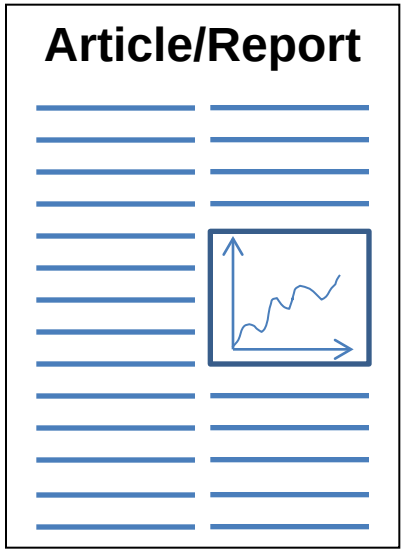
Manually



Data analysis

Stats results

```
H2-D -- CD8
            Df Sum Sq Mean Sq F value   Pr(>F)
Nlrc5        1  33571   33571 343.424 1.71e-14 ***
CIITA        1     77      77   0.785    0.386
Nlrc5:CIITA  1    609     609   6.225    0.021 *
Residuals   21   2053      98
---
H2-D -- CD4
            Df Sum Sq Mean Sq F value   Pr(>F)
Nlrc5        1  31499   31499  90.402 4.65e-09 ***
CIITA        1   3467    3467   9.951  0.00478 **
Nlrc5:CIITA  1    447     447   1.283  0.27009
Residuals   21   7317     348
---
H2-D -- NK
            Df Sum Sq Mean Sq F value   Pr(>F)
Nlrc5        1  18867   18867  70.462 3.78e-08 ***
CIITA        1    386     386   1.441    0.243
Nlrc5:CIITA  1    403     403   1.506    0.233
Residuals   21   5623     268
```



Manually

First steps:
keep track of everything you do, and all results you get

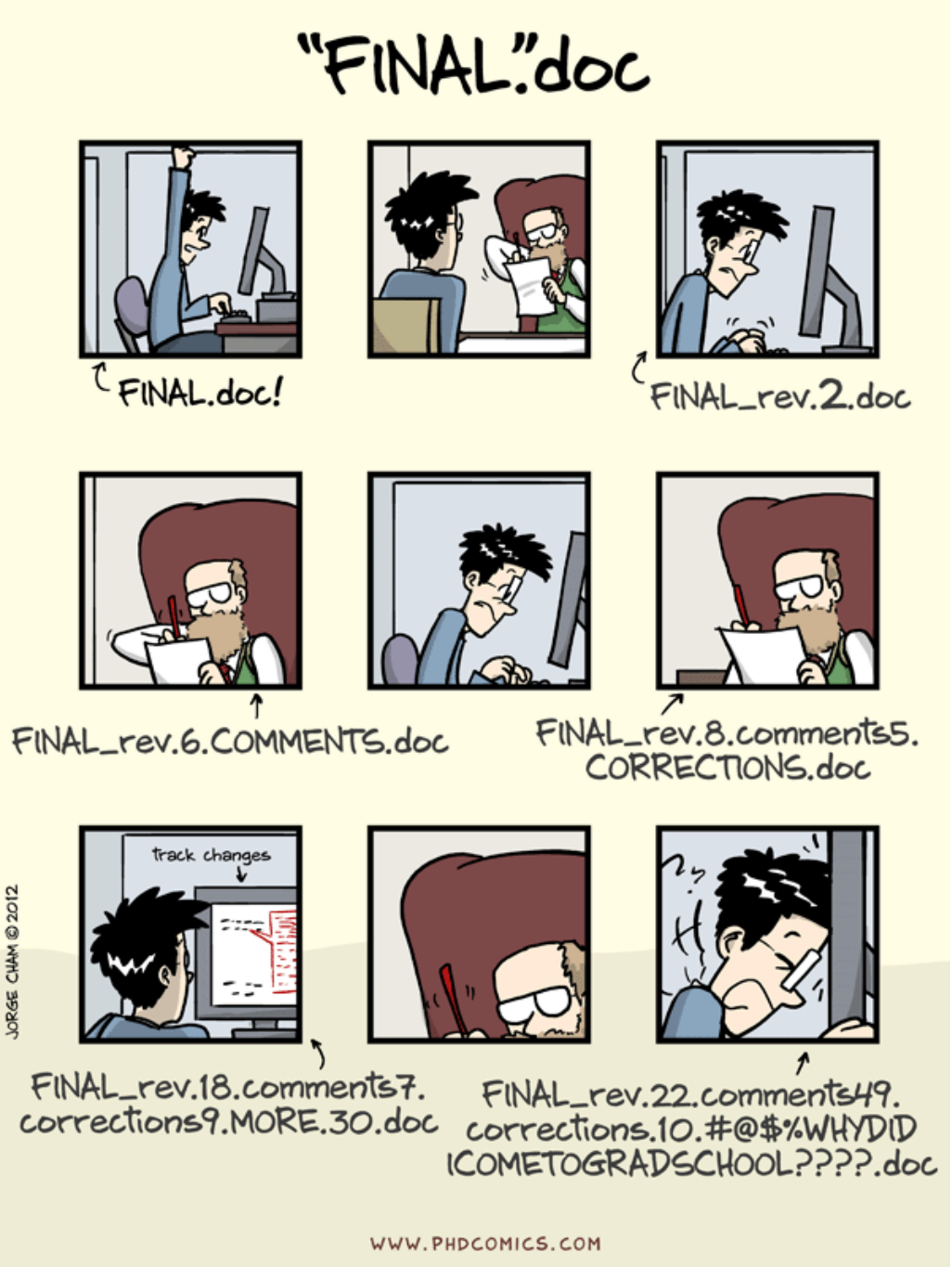
Store information about your sessions

Save your commands in a script.

Ideally, use Subversion or GIT to manage the revisions of this file

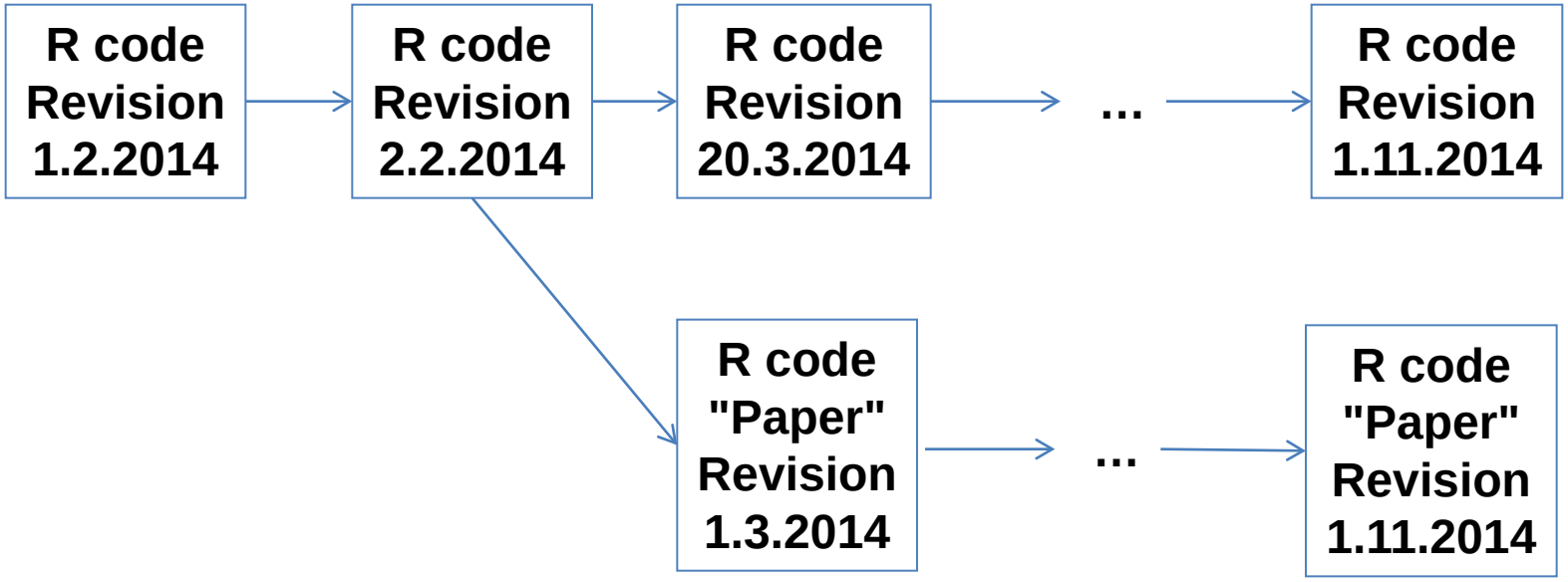
When you run it, save everything it produces (or at least the important results) to a text file.

At the beginning of the session, use sessionInfo() to record the software used.



http://www.phdcomics.com/comics/archive.php?comicid=1531

If needed, use a tool like **Subversion** or **GIT** or subversion to manage different versions/ revisions/branches of your files.



Dynamic documents with knitR

- Based on the idea of **literate programming**
- Combine program code and explanation/ documentation in same document (Donald Knuth, 1984)
- Documents in which the information is always up-to-date
- Writing report step by step while processing the data, in the same file

- Integrate your results in a report: *write the R code directly with the text, and later integrate the results directly into the text.*

Use knitr in order to prepare reports

Allows you to integrate your results in a report.

Write the R code directly with the text, and later integrate the results directly into the text.

Home Objects Options Hooks Patterns Demos

knitr

Elegant, flexible and fast dynamic report generation with R

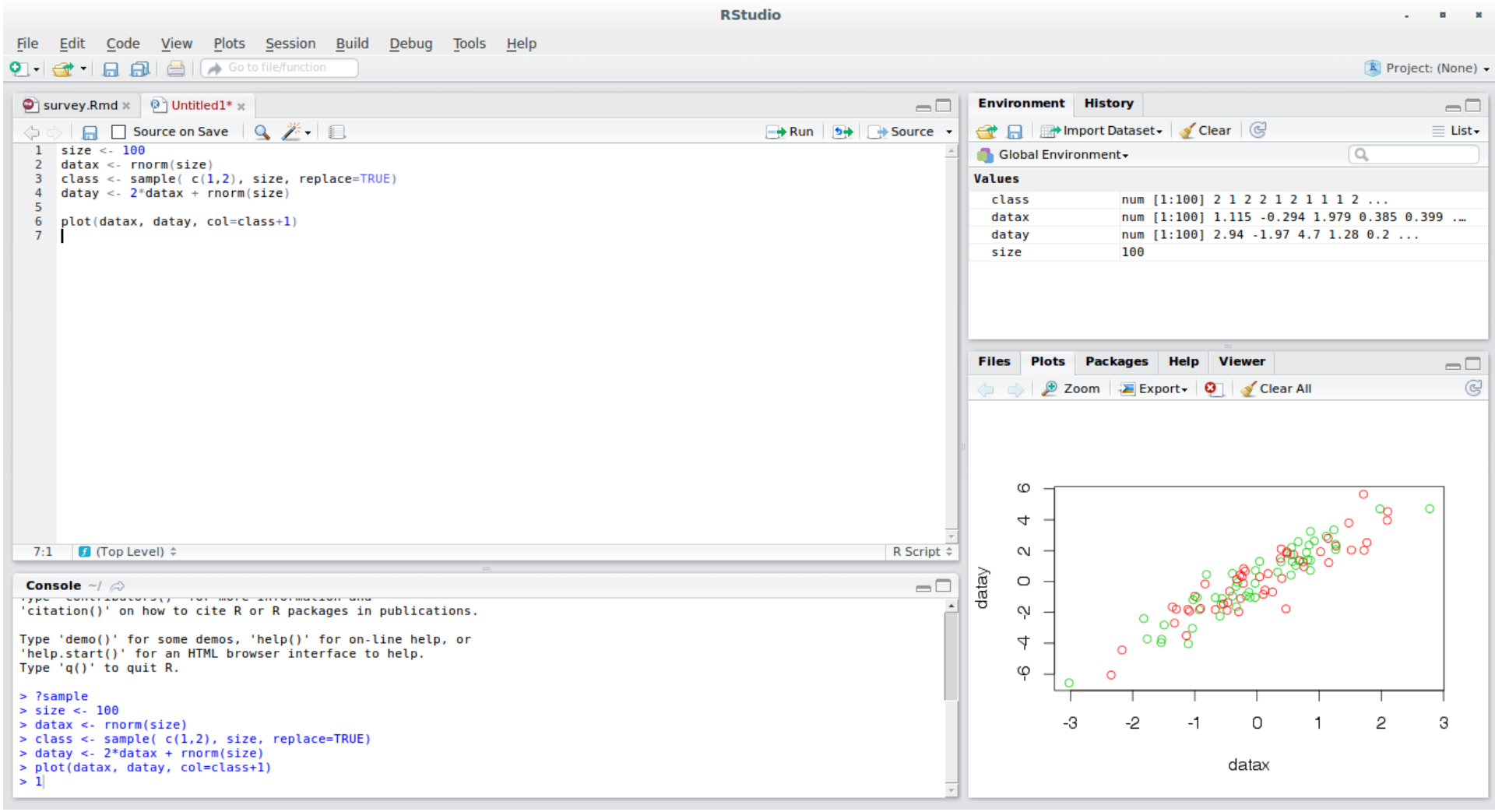
Overview

The **knitr** package was designed to be a transparent engine for dynamic report generation with R, solve some long-standing problems in Sweave, and combine features in other add-on packages into one package (**knitr** ≈ Sweave + cacheSweave + pgfSweave + weaver + `animation::saveLatex` + `R2HTML::RweaveHTML` + `highlight::HighlightWeaveLatex` + 0.2 * brew + 0.1 * SweaveListingUtils + more).

http://yihui.name/knitr/

- R
- `knitr` R package
- Editor (preferably with some support for R) configured to provide support for **knitr**
  see http://yihui.name/knitr/demo/editors/
- TeX Live (optional)
- pandoc
- learn from demos and examples:
  - `http://yihui.name/knitr/`
  - `http://rpubs.com`

RStudio



http://www.rstudio.com/products/rstudio/download/

- Write .Rnw files, and generate PDF reports using LaTeX
- keep general structure of standard LATEX document:

```
\documentclass{...}
\usepackage{...}
\begin{document}
...
\end{document}
```

- Use the same LATEX packages/configurations as usual
- Add R chunks in the LaTeX code

If LaTeX is too scary, consider:

- LYX: `http://www.lyx.org/`
- markdown:
  - `http://www.rstudio.com/ide/docs/authoring/using_markdown`
  - `https://github.com/adam-p/markdown-here/wiki/`

Markdown is a simple plain text format that allows you to specify the layout of a document, and which can easily be converted to different formats afterwards.

R Markdown combines the core syntax of markdown (easy-to-write plain text format) with embedded R code chunks that are run so their output can be included in the final document.

R Markdown v2 (http://rmarkdown.rstudio.com/)

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function Project: (None)

rmarkdown.Rmd

Knit HTML Run Chunks

```
---
title: "RMarkdown example"
author: "Frédéric Schütz"
date: "19/11/2015"
output: html_document
---

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

# Introduction

Markdown allows you to create sections and subsections, to emphasize text by formatting it in *italic* or **bold**.

## Lists

You can also easily create lists:

* Item 1
* Item 2
    + Item 2a
```

2:9 (Top Level) R Markdown

Console

~/Documents/sib/reproducible_research/rmarkdown.html

rmarkdown.html Open in Browser Publish Find

RMarkdown example

Frédéric Schütz

19/11/2015

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

Introduction

Markdown allows you to create sections and subsections, to emphasize text by formatting it in *italic* or **bold**.

Lists

You can also easily create lists:

- Item 1
- Item 2
    - Item 2a

1. Ordered list
2. Another item
    - A subitem

or even tables:

First Header	Second Header
Content Cell	Content Cell
Content Cell	Content Cell

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function Project: (None)

rmarkdown.Rmd

Knit HTML Run Chunks

```
33
34
35 # Integrating R code
36
37 When you click the **Knit** button a document will be
   generated that includes both content as well as the
   output of any embedded R code chunks within the
   document. You can embed an R code chunk like this:
38
39 ```{r}
40 summary(cars)
41 ```
42
43 You can also add R code directly in the text; for
   example, the dataset cars contains `r nrow(cars)` rows.
44
45 You can also embed plots, for example:
46
47 ```{r, echo=FALSE}
48 plot(cars)
49 ```
50
51 Note that the `echo = FALSE` parameter was added to the
   code chunk to prevent printing of the R code that
   generated the plot
```

43:41 (Top Level) R Markdown

Console

~/Documents/sib/reproducible_research/rmarkdown.html

rmarkdown.html Open in Browser Publish Find

Integrating R code

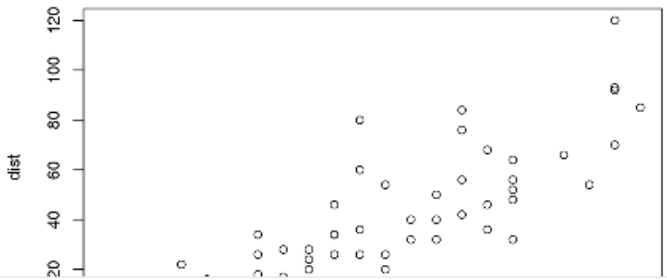
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

You can also add R code directly in the text; for example, the dataset cars contains 50 rows.

You can also embed plots, for example:



R Markdown example

```
---
title: "Untitled"
author: "Frédéric Schütz"
date: "23/01/2015"
output: html_document
---

This is an R Markdown document. Markdown is a simple
formatting syntax for authoring HTML, PDF, and MS Word
documents. For more details on using R Markdown see
<http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be
generated that includes both content as well as the output of
any embedded R code chunks within the document. You can embed
an R code chunk like this:
```

R Markdown example (continued)

```
```{r}
summary(cars)
```

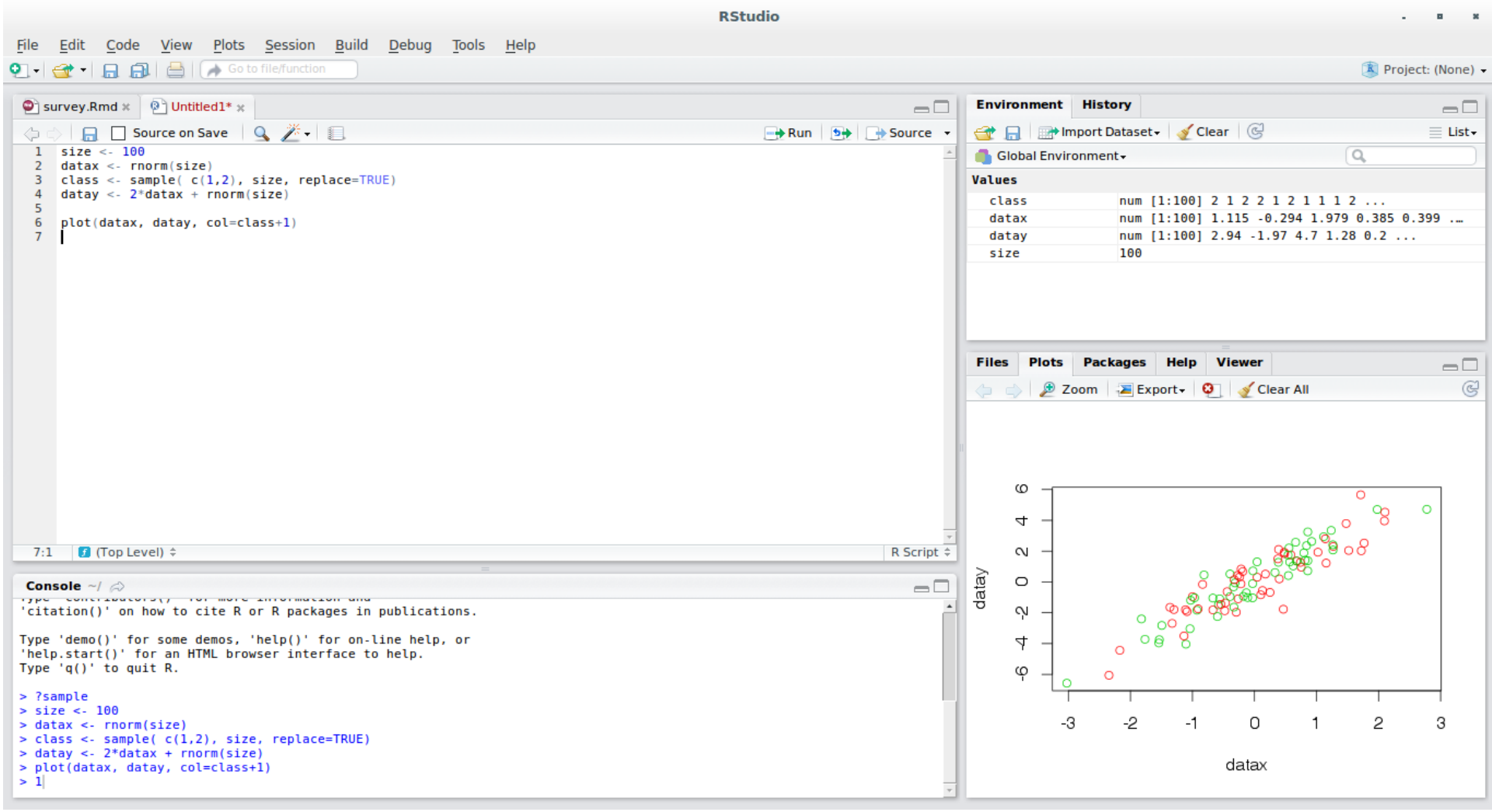

You can also embed plots, for example:


```{r, echo=FALSE}
plot(cars)
```


Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
```

- Emphasis: `*italic*` `**bold**`
  `_italic_` `__bold__`
- Headers
  ```
  # Header 1
  ## Header 2
  ### Header 3
  ```
- Unordered List:
  ```
  * Item 1
  * Item 2
      + Item 2a
      + Item 2b
  ```
- Ordered list:
  ```
  1. Item 1
  2. Item 2
  3. Item 3
      + Item 3a
      + Item 3b
  ```

RStudio



http://www.rstudio.com/products/rstudio/download/

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function

Project: (None)

report.Rmd*

Knit Word

Knit HTML
Knit PDF
Knit Word

Run Chunks

```
---
title: "Statistics analysis for paper Ludigsa et al."
author: "Frédéric Schütz"
date: "4 November 2014"
output: word_document
---

# Figure 1a (new)

The p-values are obtained using a two-sample (two-sided) t-test. Adjustment for
multiple testing is done using a Bonferroni correction over 12 samples.

```{r echo=FALSE}

dotable <- function( datafile ) {

  data <- read.table(datafile, fill=TRUE, header=FALSE,
                     sep="\t", as.is=TRUE)

  results <- NULL

  for (group in unique(data$V1)) {

    # Create section here !

    for (group2 in unique(data$V2)) {
      for (genotype in c("Nlrc5", "Rfx5")){

        thisdata <- data[ grep( genotype, data$V3 ), ]
```

8:1 (Top Level) R Markdown

Console ~/sibtmp/guerda_greta/

```
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Environment History

To Console To Source

```
date "4 November 2014"
```

Files Plots Packages Help Viewer

Zoom Export Clear All
```

R Code chunks

- R code placed in *chunks* will be evaluated and printed

  ```{r}
  summary(cars$dist)
  summary(cars$speed)
  ```

- Inline R Code

  `There were `r nrow(cars)` cars studied`

- Links: use a plain http address or add a link to a phrase:

  `http://example.com`

  `[linked phrase](http://example.com)`

- Images on the web or local files in the same directory:

  `![alt text](http://example.com/logo.png)`

  `![alt text](figures/img.png)`

Why use knitr ?

- all-in-one: analysis, documenting, formatting, reporting
- no annoying and error-prone copy-pasting
- modifying input data or code: changes are directly reflected in report
- easy to display underlying code in report when needed
- split code in chunks, but can still access all previously defined
- variables (single R session)
- flexible: code externalization, child documents, caching,...

R notebooks

Exercises

- Using Rstudio, start a new .Rmd (R Markdown file).
- Look at the template that was provided, change the R code
- Create an HTML, a Word and a PDF file from this Markdown code
- Make sure you know how to do at least the following: sections, lists, create a table from R, insert a graphic from R.
- Note: you may need to install a TeX distribution to generate PDF; you can also generate a Word or Excel document, and print/convert them to PDF if required
- Make sure to include information about the current R session (R version, packages loaded) in the final document
- Adapt an R script of your choice (ideally one you would use in your work) in a Markdown report