Reproducible Research

Wieso? Weshalb? Warum? Aber... Und Wie?

Manuel J. A. Eugster

Institut für Statistik
Ludwig-Maximiliams-Universität München

Herbstworkshop, Ruhr-Universität Bochum, 2011

Joint work with Friedrich Leisch and Torsten Hothorn:
Executable papers for the R community: The R^2 platform for reproducible research. Procedia Computer Science, 4:618–626, 2011. Proceedings of the International Conference on Computational Science, ICCS 2011. (Leisch et al., 2011)

Joint work with Anne-Laure Boulesteix:
Seminar: Reproduzierbarkeit. Institut für Statistik, Ludwig-Maximilans-Universität München. Winter term 2011/2012.

(*) Parts of this talk are based on Hothorn (2010) and Boulesteix (2010/2011).

THE VIEWS EXPRESSED IN THIS TALK DO NOT NECESSARILY REPRESENT THE VIEWS OF MY COLLABORATORS!

Reproducible research in computational science

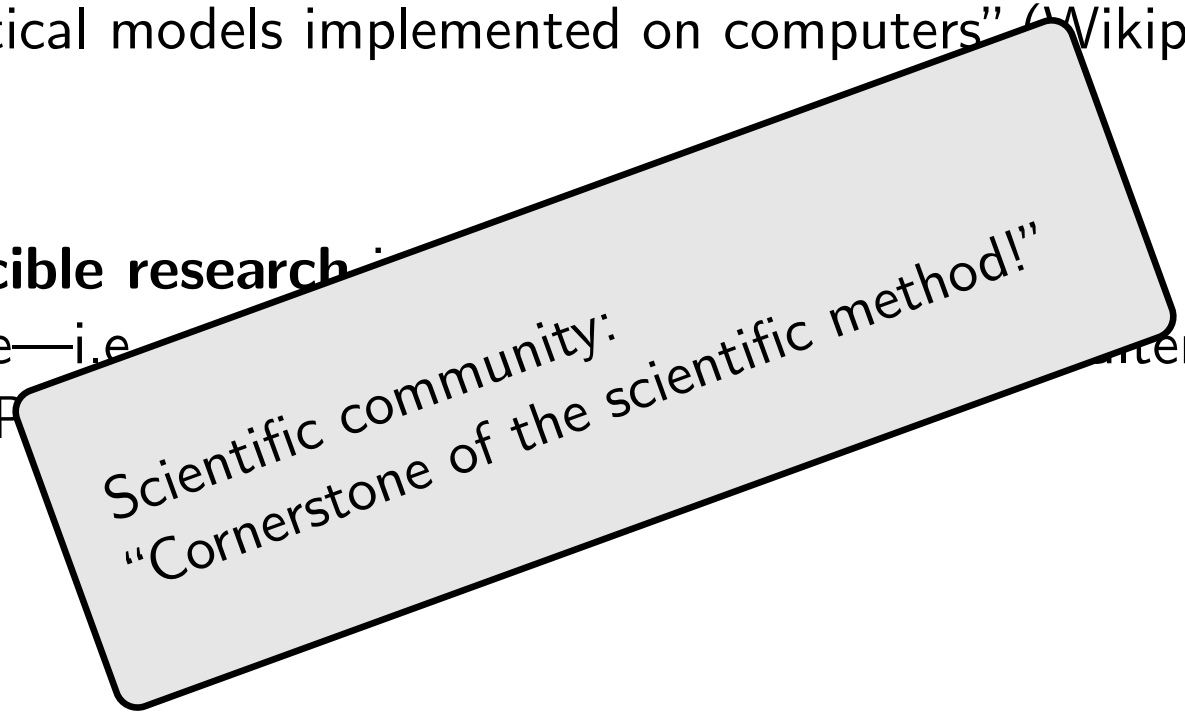
Computational science or scientific computing is "the analysis of mathematical models implemented on computers" (Wikipedia, 2011a).

Reproducible research is the ability to independently recompute—i.e., to verify—findings (and to conduct alternative analyses, Peng et al., 2006).

Reproducible research in computational science

Computational science or scientific computing is "the analysis of mathematical models implemented on computers" (Wikipedia, 2011a).

Reproducible research i
recompute—i.e
analyses, R



teruallo quadratorum, & Canones iidem hîc etiam locum habebunt, vt manifestum est.

QVÆSTIO VIII.

PROPOSITVM quadratum diuidere in duos quadratos. Imperatum sit vt 16. diuidatur in duos quadratos. Ponatur primus 1 Q. Oportet igitur 16 - 1 Q. æquales esse quadrato. Fingo quadratum à numeris quotquot libuerit, cum defectu tot vnitatum quot continet latus ipsius 16. esto à 2 N. - 4. ipse igitur quadratus erit 4 Q. + 16. - 16 N. hæc æquabuntur vnitatibus 16 - 1 Q. Communis adiiciatur vtrimque defectus, & à similibus auferantur similia, fient 5 Q. æquales 16 N. & fit 1 N. $\frac{16}{5}$. Erit igitur alter quadratorum $\frac{256}{25}$. alter verò $\frac{144}{25}$. & vtriusque summa est $\frac{400}{25}$ seu 16. & vterque quadratus est.

ΤΟΝ ἐπιταχθέντα τετράγωνον διελεῖν εἰς δύο τετραγώνους. ἐπιτετάχθω δὴ τὸν ις̄ διῃρῆσθαι εἰς δύο τετραγώνους. καὶ τετάχθω ὁ τοῦ πρώτου δυνάμεως μιᾶς. δεῖ δὴ ἄρα μονάδας ις̄ λείψει δυνάμεως μιᾶς ἴσους εἶναι τετραγώνῳ. πλάσσω τὸν τετράγωνον ἀπὸ ς̄ς̄. ὅσων δή ποτε λείψει τοσούτων μ̄ ὅσων ἐστὶν ἡ τοῦ ις̄ μ̄ πλευρά. ἔστω ς̄ς̄ β̄ λείψει μ̄ δ̄. αὐτὸς ἄρα ὁ τετράγωνος ἔσται δυνάμεων δ̄ μ̄ ις̄ [λείψει ς̄ς̄ ις̄] τουτέστιν ἴσα μονάσι ις̄ λείψει δυνάμεως μιᾶς. κοινὴ προσκείσθω ἡ λεῖψις, καὶ ἀπὸ ὁμοίων ὅμοια. δυνάμεις ἄρα ε̄ ἴσαι ἀριθμοῖς ις̄. καὶ γίνεται ὁ ἀριθμὸς

teruallo quadratorum, & Canones iidem hîc etiam locum habebunt, vt manifestum est.

QVÆSTIO VIII.

PROPOSITVM quadratum diuidere in duos quadratos.

TON ἐπιπεδον τετράγωνον

In the sense of Fermat's Last Theorem:
"I have discovered a truly marvelous *source code* of this, which this *paper* is too narrow to contain."

(*) See Wikipedia (2011b)

Communis adiiciatur vtrimque defectus, & à similibus auferantur similia, fient 5 Q. æquales 16 N. & fit 1 N. $\frac{16}{5}$. Erit igitur alter quadratorum $\frac{256}{25}$. alter verò $\frac{144}{25}$. & vtriusque summa est $\frac{400}{25}$ seu 16. & vterque quadratus est.

ἄρα ὁ τετράγωνος ἔσται δυνάμεων δ μ̄ ις̄ [λείψεως ις̄] παντὶ ἴσα μονάσι ις̄ λείψει δυνάμεως μιᾶς. κοινὴ προσκείσθω ἡ λεῖψις, καὶ ἀπὸ ὁμοίων ὁμοῖα. δυνάμεις ἄρα ε̄ ἴσαι ἀριθμοῖς ις̄. καὶ γίνεται ὁ ἀριθμὸς

Clearbout's principle:

"An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."

(*) Buckheit and Donoho (1995) and de Leeuw (2001)

“Publish data and source code!”

“Publish data and source code!”

Under the hood—truly interdisciplinary field:

Philosophy of science: scientific method, contribution to knowledge

Law: licensing for scientific innovation

Forensics: reproducibility of publications

Computer science: consequences of heterogeneous technology

Implementation: tools and workflows for practical realization

Wieso? Weshalb? Warum?

Local and global improvements of publishing source code and data

Local improvements

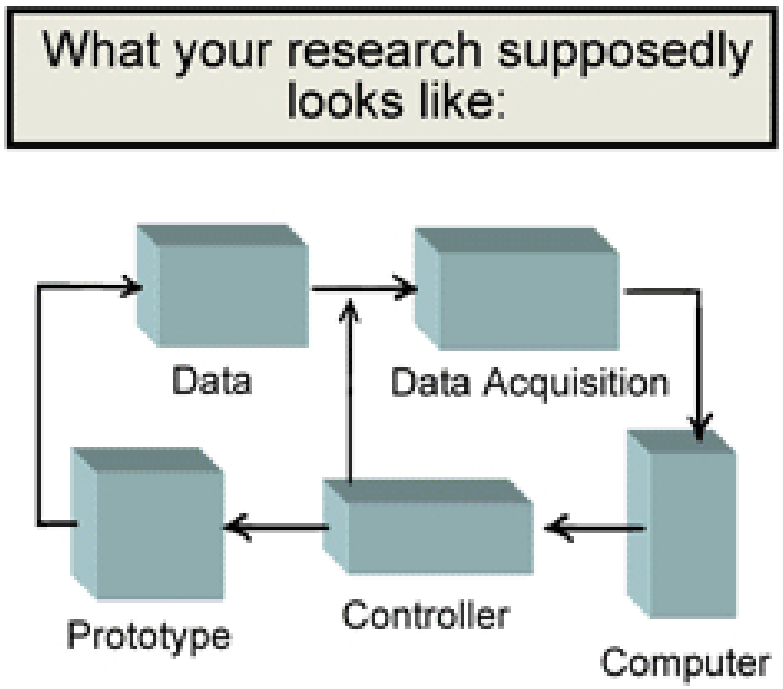


Figure 1. Experimental Diagram

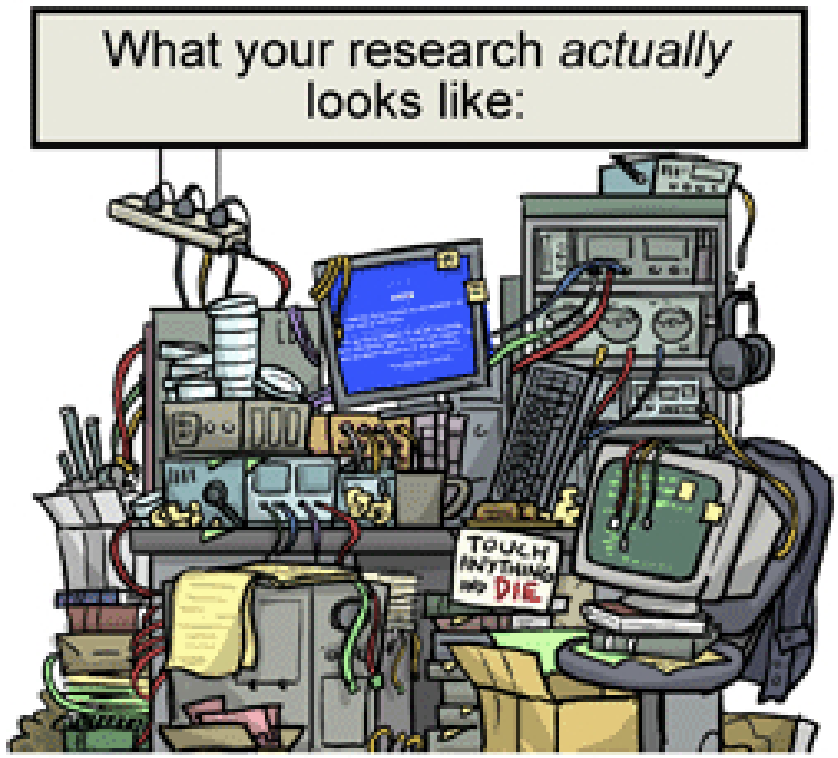


Figure 2. Experimental Mess

Global improvements

"We reproduced two analyzes in principle and six partially or with some discrepancies; ten could not be reproduced. The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis."

(*) Ioannidis et al. (2009) on microarray gene expression analysis

THE TIMES
THE SUNDAY TIMES

Archive Article | Please enjoy this article from The Times & The Sunday Times archives.

From The Sunday Times

January 18, 2009

Wealthy men give women more orgasms

Jonathan Leake, Science and Environment Editor

Scientists have found that the pleasure women get from making love is directly linked to the size of their partner's bank balance.

They found that the wealthier a man is, the more frequently his partner has orgasms.

"Women's orgasm frequency increases with the income of their partner," said Dr Thomas Pollet, the Newcastle University psychologist behind the research.

(*) Nettle and Pollet (2009), and Herberich et al. (2010)

Nettle and Pollet (2009):

"Partner wealth predicts self-reported orgasm frequency in a sample of Chinese women".

The study is based on the freely available Chinese Health and Family Life Survey.

The main conclusion is drawn from a proportional odds model linking the self-reported orgasm frequency of women with male partners to sociodemographic and wealth variables of the couple.

Reproducing Nettle and Pollet (2009):

The paper is actually reproducible because

- the data are publically available,
- the data preprocessing is well-described in the manuscript, and
- the software used to fit the model and perform AIC-based model selection is cited (SPSS).

However, Esther Herberich and Torsten Hothorn failed to reproduce the analysis in R.

It turned out that SPSS 15.0 did not exclude a model-specific constant in the multinomial log-likelihood before comparing models differing in the covariates.

Herberich, Hothorn, Nettle, and Pollet (2010):

When calculating the AIC in a correct manner, the women's education is most strongly (positively) related to the response.

Herberich, Hothorn, Nettle, and Pollet (2010):

When calculating the AIC in a correct manner, the women's education is most strongly (positively) related to the response.

BERLINER KURIER News

Bookmarken Drucken Artikel versenden

ORGASMUS-STUDIE

Kluge Frauen kommen öfter

Klischee vom geilen Dummerchen haben Münchner Uni-Forscher widerlegt

ANESTHESIA & ANALGESIA

100 Pine Street, Suite 230, San Francisco, CA 94111
Phone: (415) 777-2750, Fax: (415) 777-2803

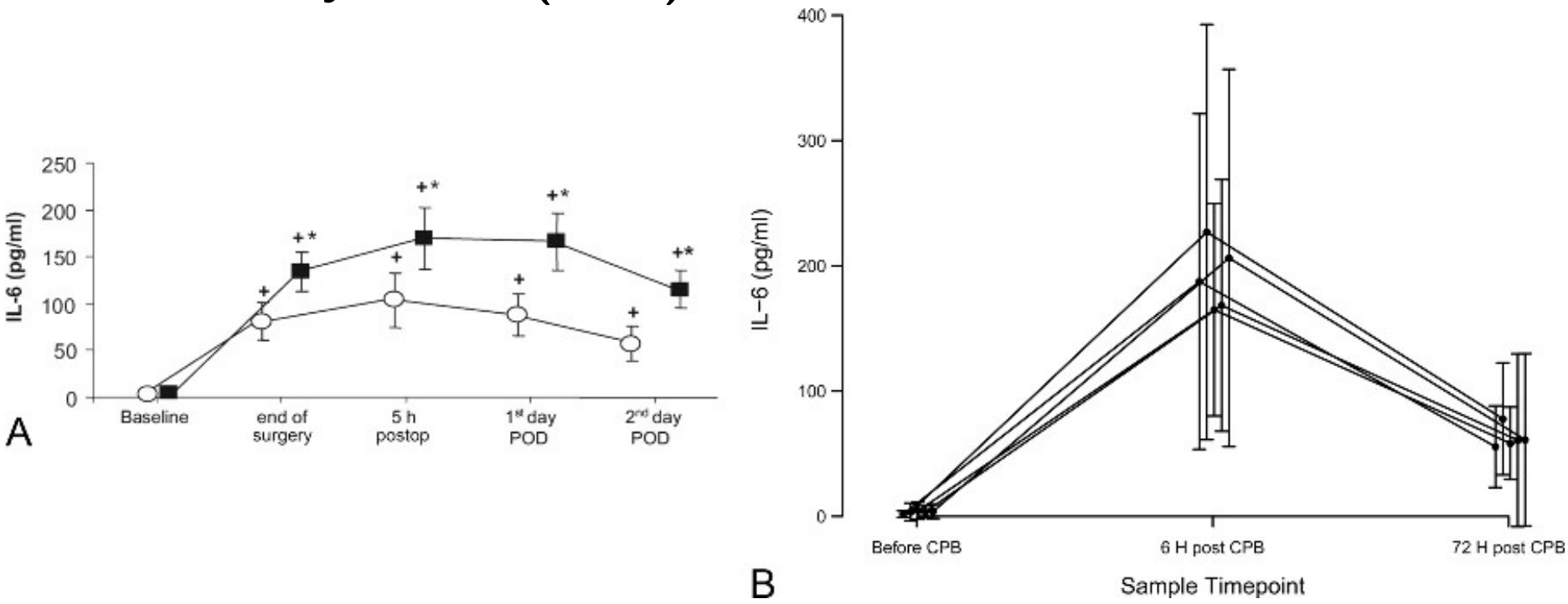
Steven L. Shafer, MD
Editor-in-Chief

February 25, 2011

To our readers:

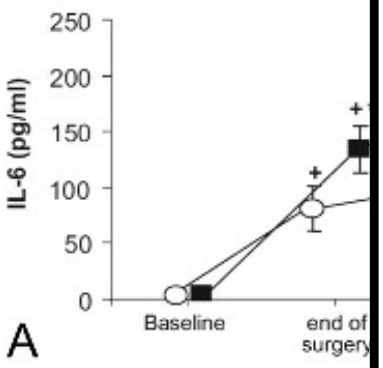
In 2009 Dr. Joachim Boldt published a manuscript in *Anesthesia & Analgesia* comparing albumin and hydroxyethyl starch priming cardiopulmonary bypass.[1] The study was retracted in December 2010 for lack of IRB approval.[2] A subsequent investigation by Klinikum Ludwigshafen determined that the study was fabricated.[3]

Forensics by Shafer (2011):

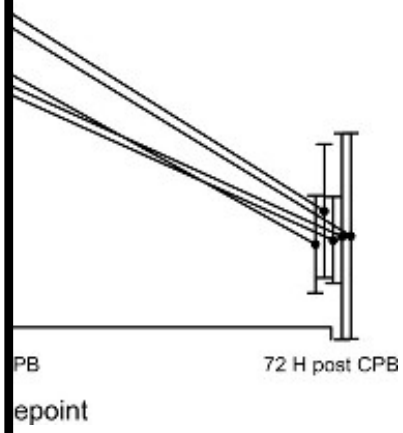


Three readers contacted the journal to question the small standard deviations of the interleukin IL-6 concentrations reported by Boldt (A). For comparison, B shows a similar study.

Forensics b



A



Three reade ll standard deviations c d by Boldt (A). For co

Table 1. Findings from LÄK-RLP, October 25, 2010

Statement in article	Finding by LÄK-RLP
Fifty consecutive patients undergoing elective coronary artery bypass grafting were studied after approval of the IRB.	For the respective study, there does not exist an approval by the respective IRB.
And after receiving individual written informed consent.	For the respective study, there does not exist written informed consent.
The patients were prospectively randomized into one of the two groups by a computer-generated list and sealed envelopes.	For the respective study, there does not exist a prospective randomization process into two groups by a computer-generated list and sealed envelopes.
A questionnaire was sent to the patients' primary physicians to receive information on patients' serum creatinine, renal failure requiring renal replacement therapy, and mortality approximately 60 days after hospital discharge.	For the respective study, there does not exist a written questionnaire sent to the patients' primary physicians approximately 60 days after hospital discharge.

LÄK-RLP = Landesärztekammer Rheinland-Pfalz.

Reality check "CSDA, Volume 56, Issue 3":

Number of 23 papers with simulation studies and/or examples giving direct access to data or code (0.5 if code either for simulation or example):

Simulation	Example	Data	Code
21	19	6.5	2.5

(*) For numbers on *Bioinformatics* see Hothorn and Leisch (2011).

Aber...

Discussion on publishing source code and data

- **“A precise description does it as well!”**

- **"A precise description does it as well!"**

A trivial example:

"A multivariate logistic regression model was built based on the predictors sex, age, and tumor localisation and the response variable 'relapse yes/no'. A forward selection was applied with entry threshold $= 0.05$."

- **"A precise description does it as well!"**

A trivial example:

"A multivariate logistic regression model was built based on the predictors sex, age, and tumor localisation and the response variable 'relapse yes/no'. A forward selection was applied with entry threshold $= 0.05$."

- How was the coding of sex and localisation?
- How did they handle missing values?
- How did they compute the p-value for the forward selection procedure? Wald? LR?

(*) See, for example, Donoho (2010)

- **“My code is crap.”**
- **“With the next Software update, it doesn’t work anymore.”**
- **“No one will spend hours to check my code.”**
- **“People will find errors.”**
- **“Statistical analysis is only a small part of the scientific approach.”**

- **“My code is crap.”**
- **“With the next Software update, it doesn’t work anymore.”**
- **“No one will spend hours to check my code.”**
- **“People will find errors.”**
- **“Statistical analysis is only a small part of the scientific approach.”**

Probably true for all points; however, not really an excuse to do unreproducible statistical analysis. It is always better to have a code than no code at all.

(*) See, for example, Barnes (2010)

- **“Replication is the important thing.”**

- **“Replication is the important thing.”**

Drummond (2009):

“[...] replicability is not reproducibility. Replicability requires changes; reproducibility avoids them.”

True, but verifying the findings based on the same data is the minimum standard and the basis for more complex verification.

(*) Note that terms are changed to match with the remaining publications

- **“I don’t have the authorization to publish the source code and the data.”**

- **“I don’t have the authorization to publish the source code and the data.”**

  Is this then “the game” of scientific research anymore? Or industrial research?

- **“But it’s my intellectual property!”**

- **"But it's my intellectual property!"**

Stodden (2009) defines a Reproducible Research Standard:

Source code: GNU GPL or (modified) BSD license

Media: Creative commons attribution license (CC BY)

Data: Science Commons Database Protocol

(*) See Free Software Foundation, Creative Commons, Science Commons

- **“You need to know the substantive context.”**

- **"You need to know the substantive context."**

Keiding (2010a) and Keiding (2010b):

"The statistician needs to understand how data were generated and selected in order to produce relevant analyses."

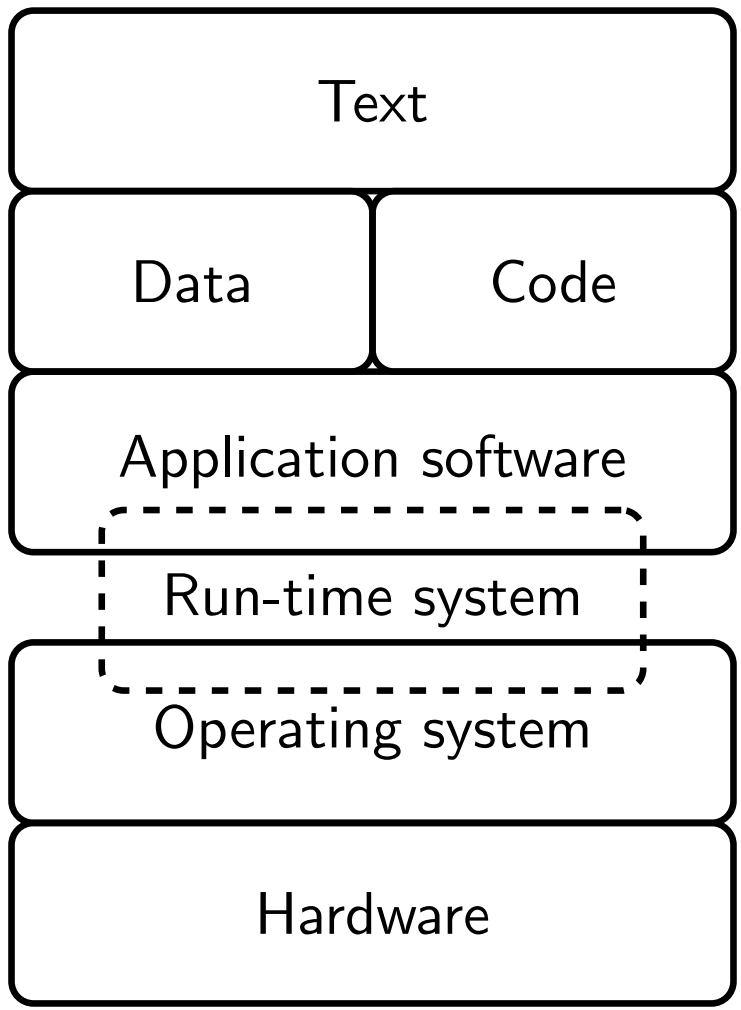
"[...], there at least has to be sufficient information to make it realistic for another interdisciplinary group of researchers to understand the substantive context and the strengths and weaknesses of the data."

"Availability of naked datasets may well be counterproductive."

- **“I used a Mac...”**

- **“I used a Mac...”**

Reproducibility in the view of heterogeneous technology:



The pitfalls of verifying floating-point computations:

"An important factor throughout the discussion is that it is not the hardware platform that matters in itself, but its combination with the software context, including the compiler, libraries, and possible run-time environment."

(*) Monniaux (2008)

The dot or scalar product:

$$a = (a_1, a_2, \ldots, a_n)'$$

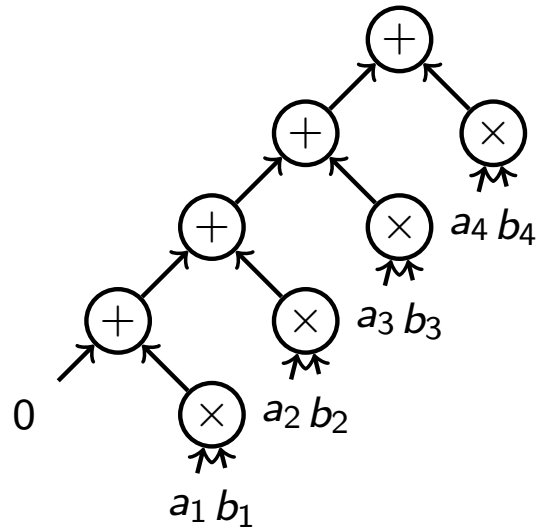
$$b = (b_1, b_2, \ldots, b_n)'$$

$$a \cdot b = \sum_{i=1}^{n} = a_1 \times b_1 + a_2 \times b_2 + \cdots + a_n \times b_b$$

We suppose (in an arbitrary mathematical software) that a function `dotprod(a, b)` returns the correct mathematical dot product.

(*) Example by Whitehead and Fit-Florea (2011)

Serial method:



$t = 0$

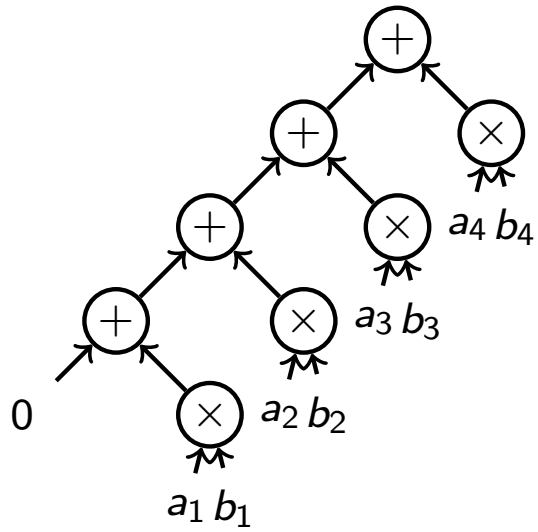
for i from 1 to 4 :

$p = \mathrm{rn}(a_i \times b_i)$

$t = \mathrm{rn}(t + p)$

return t

Serial method:



$t = 0$

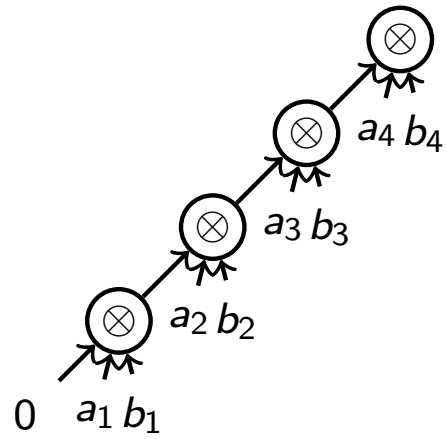
for i from 1 to 4 :

$\quad p = \mathrm{rn}(a_i \times b_i)$

$\quad t = \mathrm{rn}(t + p)$

return t

FMA method:



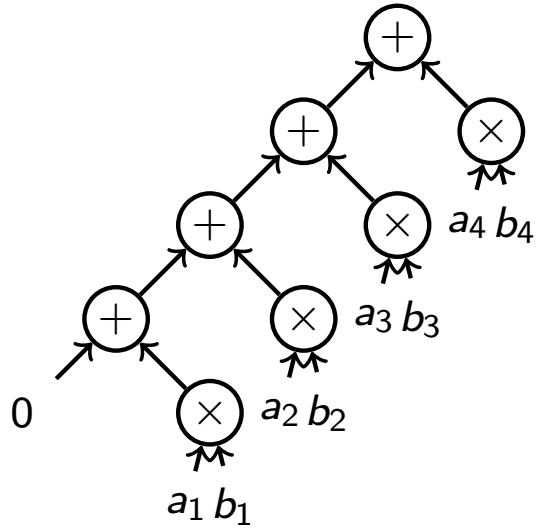
$t = 0$

for i from 1 to 4 :

$\quad t = \mathrm{rn}(a_i \times b_i + t)$

return t

Serial method:



$t = 0$

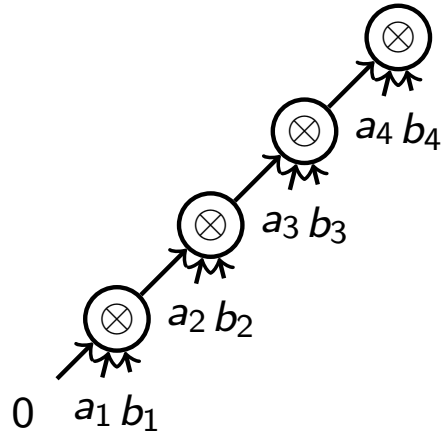
for i from 1 to 4 :

$\quad p = \mathrm{rn}(a_i \times b_i)$

$\quad t = \mathrm{rn}(t + p)$

return t

FMA method:



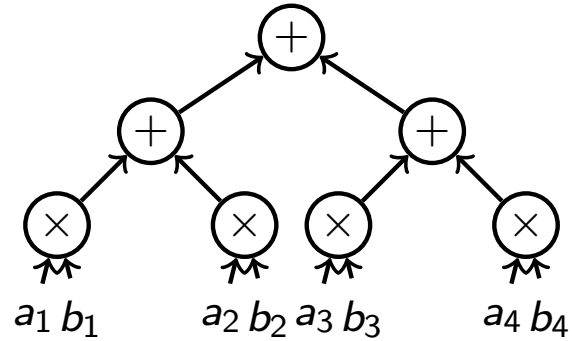
$t = 0$

for i from 1 to 4 :

$\quad t = \mathrm{rn}(a_i \times b_i + t)$

return t

Parallel method:



$p_1 = \mathrm{rn}(a_1 \times b_1)$

$p_2 = \mathrm{rn}(a_2 \times b_2)$

$p_3 = \mathrm{rn}(a_3 \times b_3)$

$p_4 = \mathrm{rn}(a_4 \times b_4)$

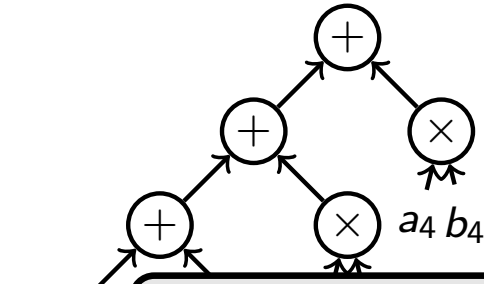
$s_l = \mathrm{rn}(p_1 + p_2)$

$s_r = \mathrm{rn}(p_3 + p_4)$

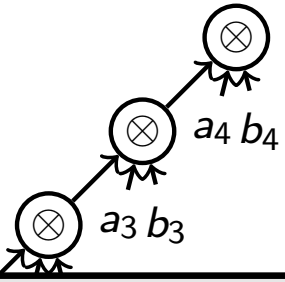
$t = \mathrm{rn}(s_l + s_r)$

return t

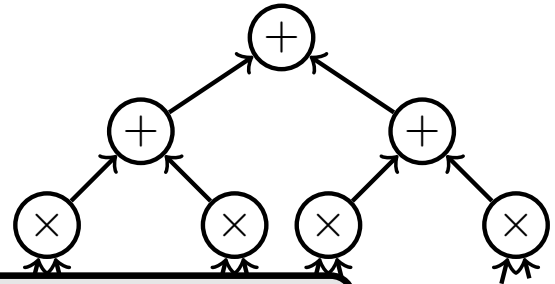
Serial method:



FMA method:



Parallel method:



Method	Result	`float` value
exact	0.0559587528435 . . .	$0x3D65350158\ldots$
serial	0.0559588074	$0x3D653510$
FMA	0.0559587515	$0x3D653501$
parallel	0.0559587478	$0x3D653500$

0

$t =$ (

for i

p =

t = rn($t + p$)

return t

b_1)

b_2)

b_3)

b_4)

p_2)

p_4)

$t = \mathrm{rn}(s_l + s_r)$

return t

- **“It requires so much time.”**

- **“It requires so much time.”**

  Yes! And motivation, patience, good organization, etc.

- **"It requires so much time."**

Y

"How to be a Highly Cited Author in the Mathematical Sciences" by Donoho (2002):

"I am a statistician, so when ISI contacted me with the news that I was a 'Highly Cited Author' [...] I looked at my list of 10 highly cited papers [...]"

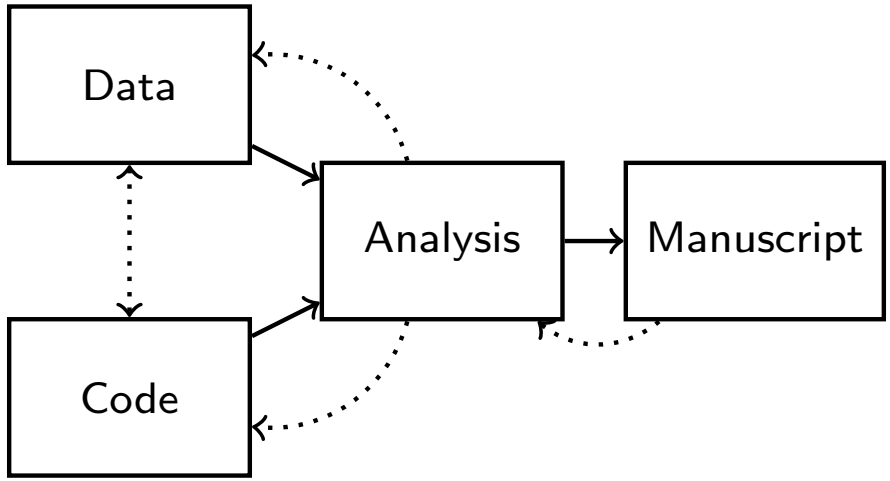
"In our most-cited papers, we developed methodology for wavelet-based noise removal which was implemented in MATLAB [...] was available for free download over the Internet [...]"

Wie?

Warning: my attempt to make my research reproducible

“Choose the right tools!”

(*) Image from Volkswagen Commercial Vehicles: Choose the right tool 1



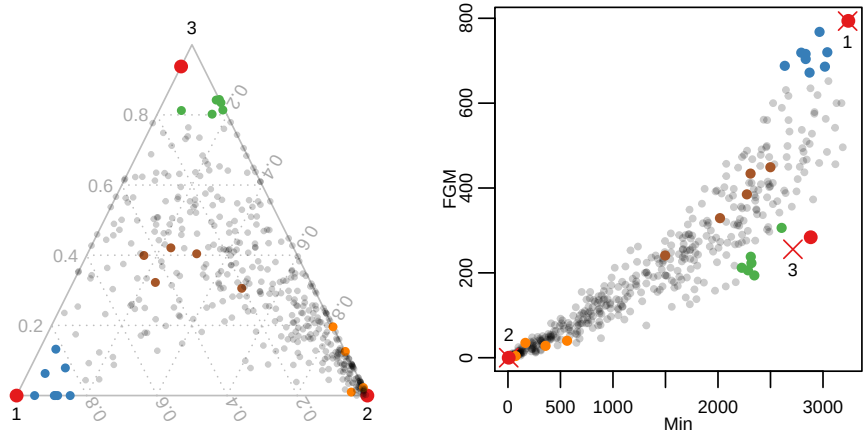
Local:

1. Programming environment
2. Document preparation system (literate programming)
3. Version control

Global:

1. Mainstream & Long-Term Software
2. Distribution system
3. (Open)

(*) See, for example, Koenker and Zeileis (2009)



(a) (b)

Figure 2: (a) Visualization of the α coefficients using a ternary plot and (b) the data set in case of the $k = 3$ archetypes solution. The red dots are the archetypes' nearest players; dots colored with blue, orange, and green are players where Archetype 1, 2, and 3 contribute more than 0.8.

problem (Formula 1) define how much each archetype contributes to the approximation of each individual observation (as convex combination). This allows the assignment of the observations to their nearest archetypes and, consequently, the identification of the most archetypal observation(s). Figure 2 shows the corresponding ternary plot of the α coefficients for the above $k = 3$ archetypes solution. The three players (red points) nearest to the respective archetypes (red crosses) are:

	Name	Team	Role	Min	FGM	$\alpha_{\cdot 1}$	$\alpha_{\cdot 2}$	$\alpha_{\cdot 3}$
Archetype 1	Kevin Durant	OKL	SF	3241	794	1.00	0.00	0.00
Archetype 2	Dwayne Jones	PHO	C	7	0	0.00	1.00	0.00
Archetype 3	Jason Kidd	DAL	PG	2883	284	0.06	0.00	0.94

Archetype 1 and 3 have well-defined nearest observations; Archetype 2, on the contrary, has a set of nearest observations and the concrete player identification should be considered as a "random" selection from the set of similar players.

We have identified Archetype 1 as the "*good*" archetype in this data setting—on this account. Kevin Durant can be considered as the best scorer. To find other good scorers.

(*) Page 5 of Archetypal Athletes, Eugster (2011)

Programming environment

"The programming language should support the user in turning theory into software that reflects how we think about the underlying method conceptually."

"To assure reproducibility and reusability by other authors, the structural features of a language should facilitate (and not suppress) the ability to build on innovations of prior authors."

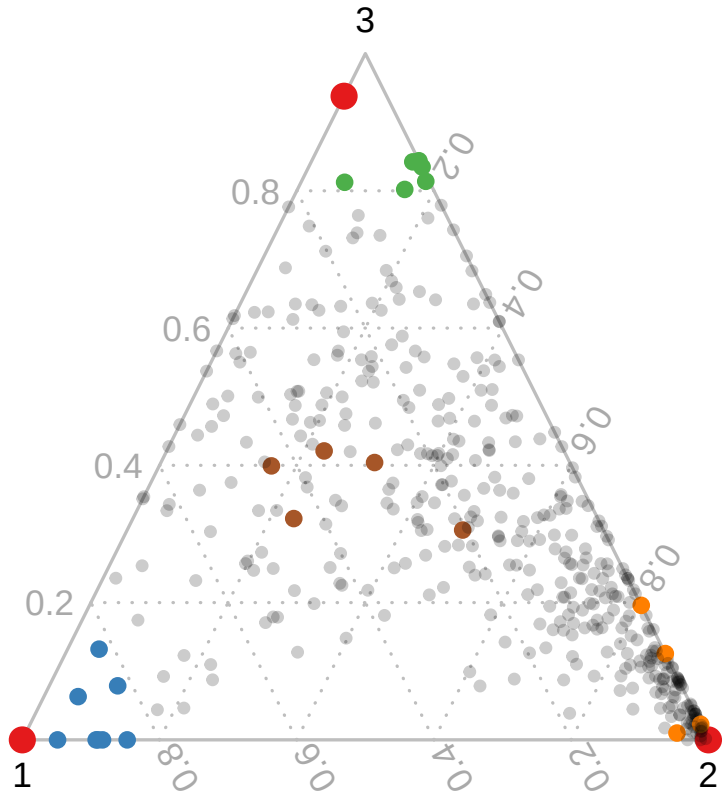
(*) Koenker and Zeileis (2009)

The R Project for Statistical Computing:

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques, and is highly extensible.

CRAN with 3420 add-on packages, daily checked.

(*) R Development Core Team (2011)




```
> library("RColorBrewer")
> library("vcd")
> col_pal <- brewer.pal(7, "Set1")
> ternaryplot(coef(a3, "alphas"),
+             col = col_pal, [...])
```

Document preparation system
Literate programming

Knuth (1986) proposed in *Literate Programming* the combination of a programming language and a documentation language:

> *from "instructing a computer what to do" to "explaining a human beeing what we want a computer to do"*

Literate programming enables to interleave code and documentation chunks: `weave`-ing creates the manuscript, `tangle`-ing the source code.

Sweave: R & LaTeX (Leisch, 2002)

odfWeave: R & OpenOffice (Kuhn, 2010)

StatWeave: R, SAS, Stata, ... & LaTeX, OpenOffice (Lenth, 2011)

Matweave: Matlab, Octave & LaTeX (Lawrence, 2011)

α coefficients for the above $k = 3$ archetypes solution. The three players (red points) nearest to the respective archetypes (red crosses) are:

	Name	Team	Role	Min	FGM	$\alpha_{\cdot 1}$	$\alpha_{\cdot 2}$	$\alpha_{\cdot 3}$
Archetype 1	Kevin Durant	OKL	SF	3241	794	1.00	0.00	0.00
Archetype 2	Dwayne Jones	PHO	C	7	0	0.00	1.00	0.00
Archetype 3	Jason Kidd	DAL	PG	2883	284	0.06	0.00	0.94

Archetype 1 and 3 have well-defined nearest observations; Archetype 2, on the contrary, has a set of nearest observations and the concrete player identification should be

archeplayers.Rnw:

```
The three players (red points) nearest to the respective archetypes
(red crosses) are:
\begin{center}

<<results=tex>>=
## Archetypal players:
atypes <- apply(coef(a3, "alphas"), 2, which.max)

atypes_coef <- coef(a3, "alphas")[atypes, ]
colnames(atypes_coef) <- sprintf("$\\alpha_{\\cdot%s}$", 1:3)

atypes_dat <- dat[atypes, ]
atypes_dat <- cbind(atypes_dat, atypes_coef)
rownames(atypes_dat) <- sprintf("Archetype %s", 1:3)

print(xtable(atypes_dat), floating = FALSE,
      sanitize.colnames.function = identity)
@

\end{center}
Archetype~1 and 3 have well-defined nearest observations;
```

Sweave("archeplayers.Rnw") $\Rightarrow$ archeplayers.tex:

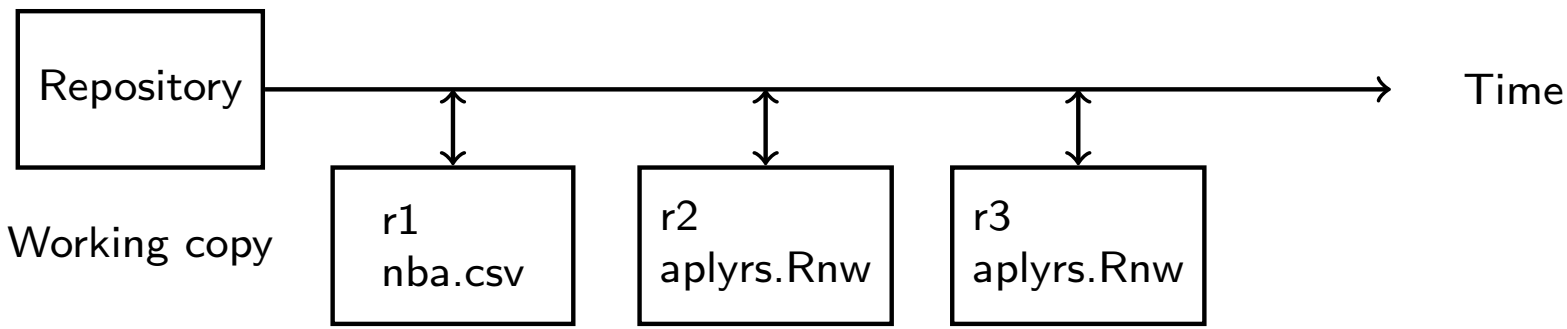
```
The three players (red points) nearest to the respective archetypes
(red crosses) are:
\begin{center}

% latex table generated in R 2.13.1 by xtable 1.5-6 package
% Tue Sep 13 10:27:53 2011
\begin{tabular}{rlllrrrrr}
  \hline
  & Name & Team & Role & Min & FGM & $\alpha_{\cdot1}$ & $\alpha_{\cdot
  \hline
  Archetype 1 & Kevin Durant & OKL & SF & 3241 & 794 & 1.00 & 0.00 & 0.
  Archetype 2 & Dwayne Jones & PHO & C &   7 &   0 & 0.00 & 1.00 & 0.00
  Archetype 3 & Jason Kidd & DAL & PG & 2883 & 284 & 0.06 & 0.00 & 0.94
  \hline
\end{tabular}

\end{center}
Archetype~1 and 3 have well-defined nearest observations;
```

Stangle("archeplayers.Rnw") $\Rightarrow$ archeplayers.R:

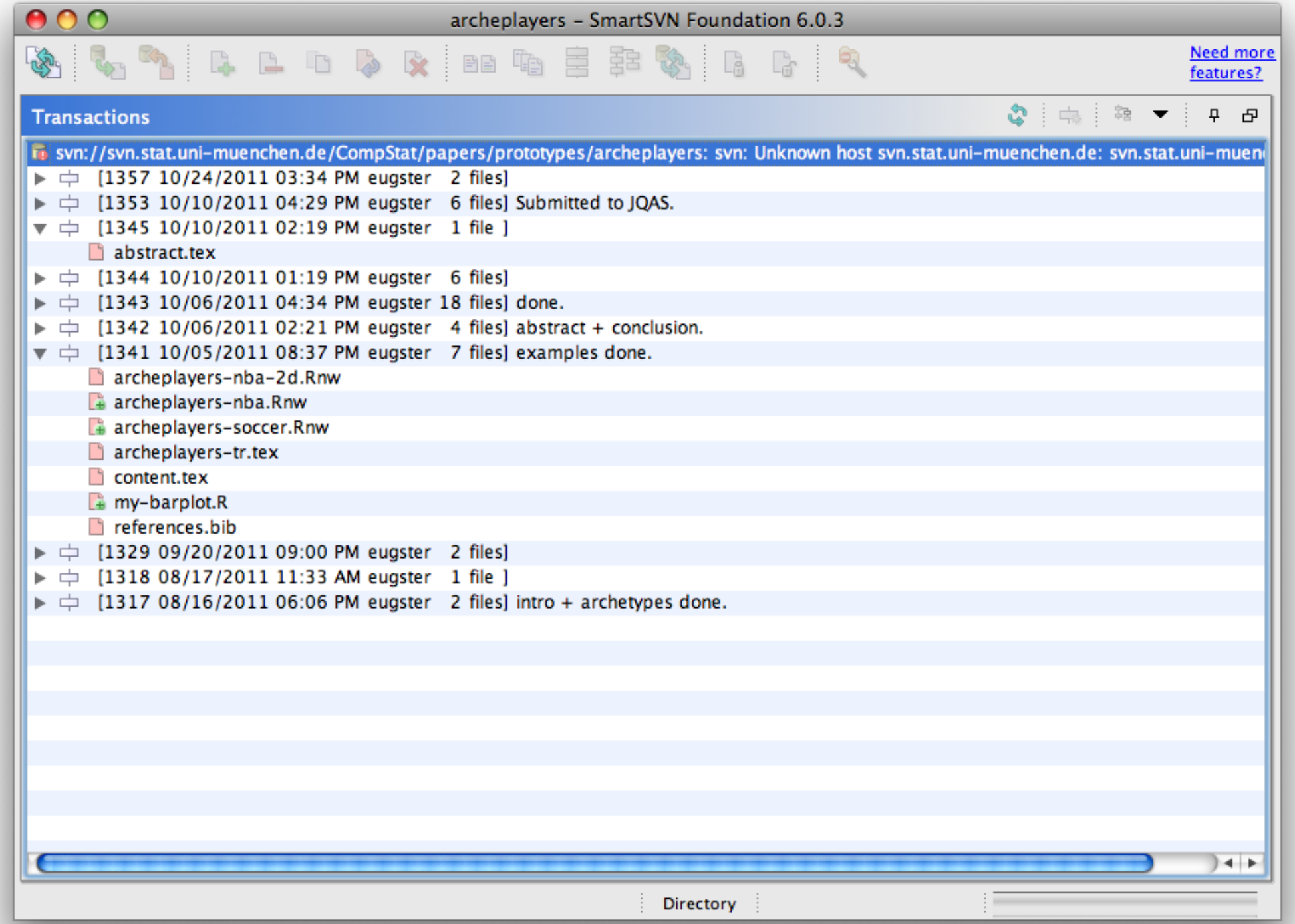
```
## Archetypal players:
atypes <- apply(coef(a3, "alphas"), 2, which.max)

atypes_coef <- coef(a3, "alphas")[atypes, ]
colnames(atypes_coef) <- sprintf("$\\alpha_{\\cdot%s}$", 1:3)

atypes_dat <- dat[atypes, ]
atypes_dat <- cbind(atypes_dat, atypes_coef)
rownames(atypes_dat) <- sprintf("Archetype %s", 1:3)

print(xtable(atypes_dat), floating = FALSE,
      sanitize.colnames.function = identity)
```

Version control

Version control is the management of changes to data, programs, documents, and other information stored as computer files. Well-known systems are Subversion and Git.



(*) See Subversion and Git


archeplayers – SmartSVN Foundation 6.0.3
Need more features?
Transactions
svn://svn.stat.uni-muenchen.de/CompStat/papers/prototypes/archeplayers: svn: Unknown host svn.stat.uni-muenchen.de: svn.stat.uni-muen
[1357 10/24/2011 03:34 PM eugster 2 files]
[1353 10/10/2011 04:29 PM eugster 6 files] Submitted to JQAS.
[1345 10/10/2011 02:19 PM eugster 1 file]
abstract.tex
[1344 10/10/2011 01:19 PM eugster 6 files]
[1343 10/06/2011 04:34 PM eugster 18 files] done.
[1342 10/06/2011 02:21 PM eugster 4 files] abstract + conclusion.
[1341 10/05/2011 08:37 PM eugster 7 files] examples done.
archeplayers-nba-2d.Rnw
archeplayers-nba.Rnw
archeplayers-soccer.Rnw
archeplayers-tr.tex
content.tex
my-barplot.R
references.bib
[1329 09/20/2011 09:00 PM eugster 2 files]
[1318 08/17/2011 11:33 AM eugster 1 file]
[1317 08/16/2011 06:06 PM eugster 2 files] intro + archetypes done.
Directory

Distribution system

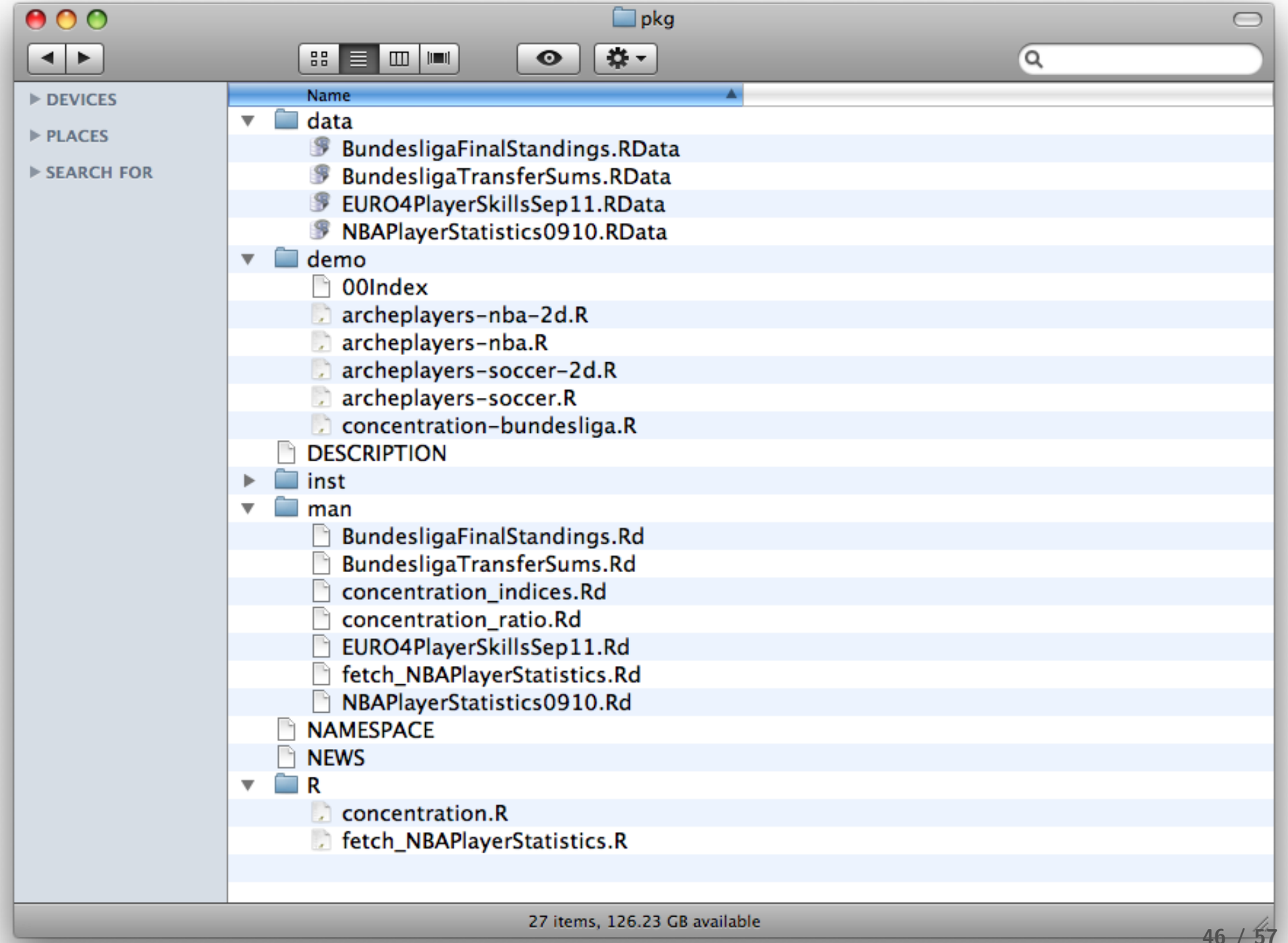
Manuscripts are published in a standardized way—publish source code, data, etc. using a standardized distribution system as well.

R add-on packages and CRAN:

Packages provide a mechanism for loading optional code, data, and attached documentation as needed.

Is is a standardized directory hierarchy with some mandatory files.
A series of checks are available to ensure the technical correctness.

CRAN publishes daily checked packages.


pkg
DEVICES
PLACES
SEARCH FOR
Name
data
BundesligaFinalStandings.RData
BundesligaTransferSums.RData
EURO4PlayerSkillsSep11.RData
NBAPlayerStatistics0910.RData
demo
00Index
archeplayers-nba-2d.R
archeplayers-nba.R
archeplayers-soccer-2d.R
archeplayers-soccer.R
concentration-bundesliga.R
DESCRIPTION
inst
man
BundesligaFinalStandings.Rd
BundesligaTransferSums.Rd
concentration_indices.Rd
concentration_ratio.Rd
EURO4PlayerSkillsSep11.Rd
fetch_NBAPlayerStatistics.Rd
NBAPlayerStatistics0910.Rd
NAMESPACE
NEWS
R
concentration.R
fetch_NBAPlayerStatistics.R
27 items, 126.23 GB available

archetypes (using the α coefficients). The two examples—basketball and soccer—shows that this is an appropriate approach; the estimated archetypal athletes definitely are consistent with the general opinion.

Computational details

All computations and graphics have been done using the statistical software R 2.13.1 [R Development Core Team, 2011], the archetypes package [Eugster, 2010], and the SportsAnalytics package [Eugster, 2011]. R itself and all packages used are freely available under the terms of the General Public License from the Comprehensive R Archive Network at http://CRAN.R-project.org/.

Data sets and source codes for replicating our analyses are available in the SportsAnalytics package. An individual analysis is executed via (replace `***` with `nba-2d`, `nba` and `soccer`):

```
R> demo("archeplayers-***", package = "SportsAnalytics")
```

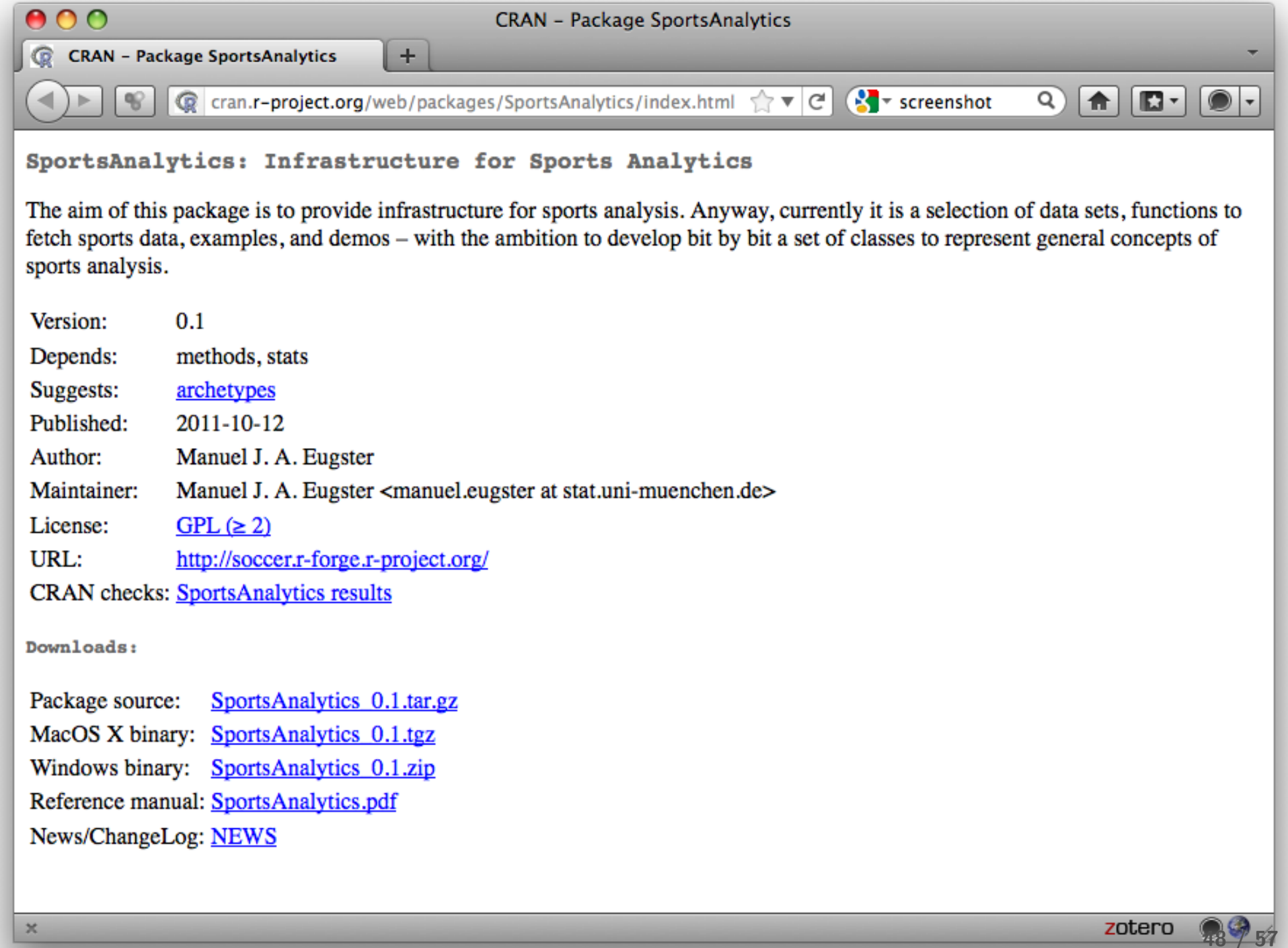
The source code file for a demo is accessible via:

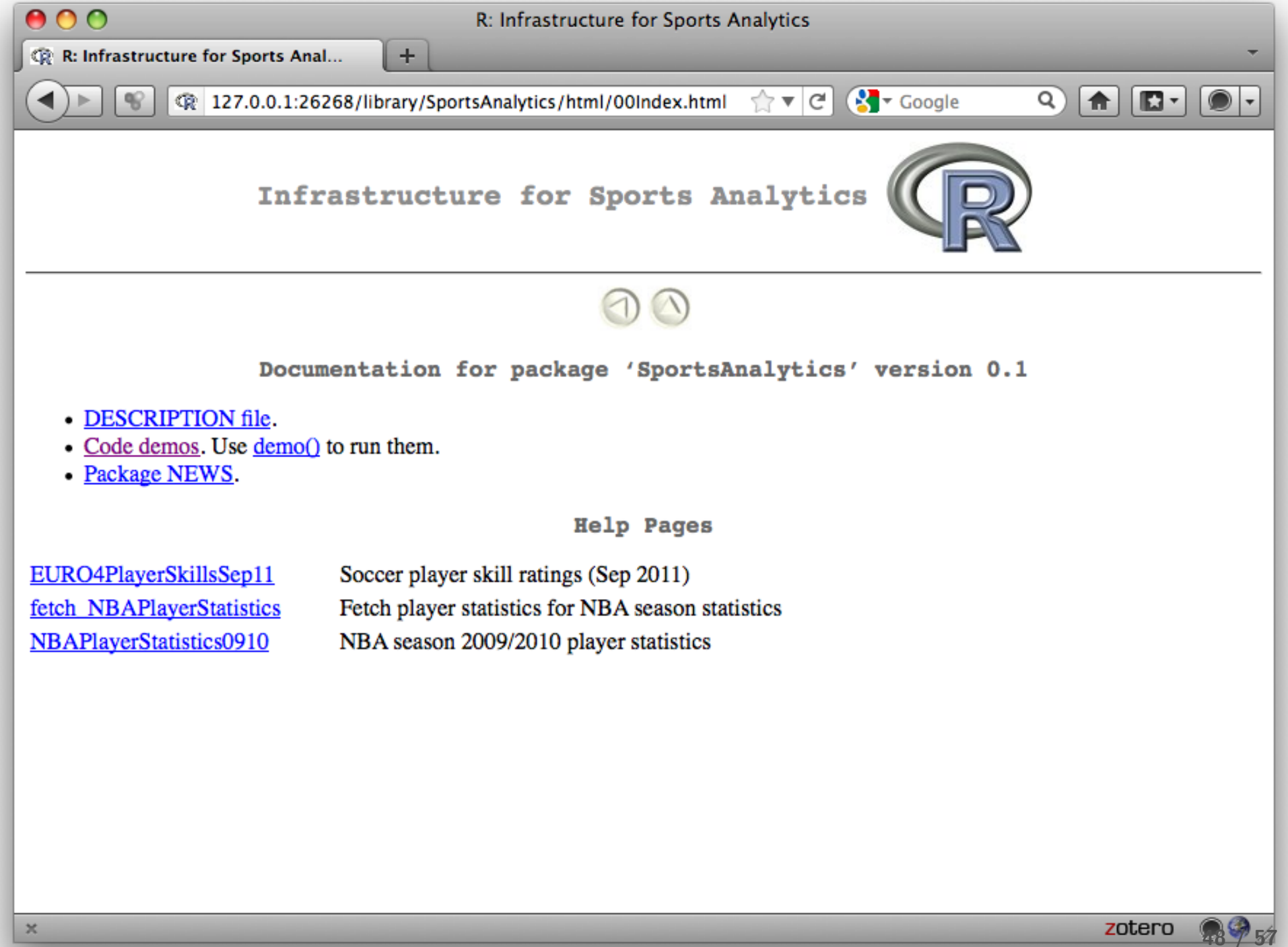
```
R> edit(file = system.file("demo", "archeplayers-***.R",
+                          package = "SportsAnalytics"))
```

References

Christian Bauckhage and Christian Thurau. Making archetypal analysis practical. In

(*) Page 14 of Archetypal Athletes, Eugster (2011)


CRAN – Package SportsAnalytics
CRAN – Package SportsAnalytics
cran.r-project.org/web/packages/SportsAnalytics/index.html
screenshot
SportsAnalytics: Infrastructure for Sports Analytics
The aim of this package is to provide infrastructure for sports analysis. Anyway, currently it is a selection of data sets, functions to fetch sports data, examples, and demos – with the ambition to develop bit by bit a set of classes to represent general concepts of sports analysis.
Version: 0.1
Depends: methods, stats
Suggests: archetypes
Published: 2011-10-12
Author: Manuel J. A. Eugster
Maintainer: Manuel J. A. Eugster <manuel.eugster at stat.uni-muenchen.de>
License: GPL (≥ 2)
URL: http://soccer.r-forge.r-project.org/
CRAN checks: SportsAnalytics results
Downloads:
Package source: SportsAnalytics_0.1.tar.gz
MacOS X binary: SportsAnalytics_0.1.tgz
Windows binary: SportsAnalytics_0.1.zip
Reference manual: SportsAnalytics.pdf
News/ChangeLog: NEWS
zotero


R: Infrastructure for Sports Analytics
R: Infrastructure for Sports Anal...
127.0.0.1:26268/library/SportsAnalytics/html/00Index.html
Google
Infrastructure for Sports Analytics
Documentation for package 'SportsAnalytics' version 0.1
DESCRIPTION file.
Code demos. Use demo() to run them.
Package NEWS.
Help Pages
EURO4PlayerSkillsSep11 Soccer player skill ratings (Sep 2011)
fetch_NBAPlayerStatistics Fetch player statistics for NBA season statistics
NBAPlayerStatistics0910 NBA season 2009/2010 player statistics
zotero

R: Demos in package 'SportsAnalytics'

R: Demos in package 'SportsAn...

127.0.0.1:26268/library/SportsAnalytics/demo

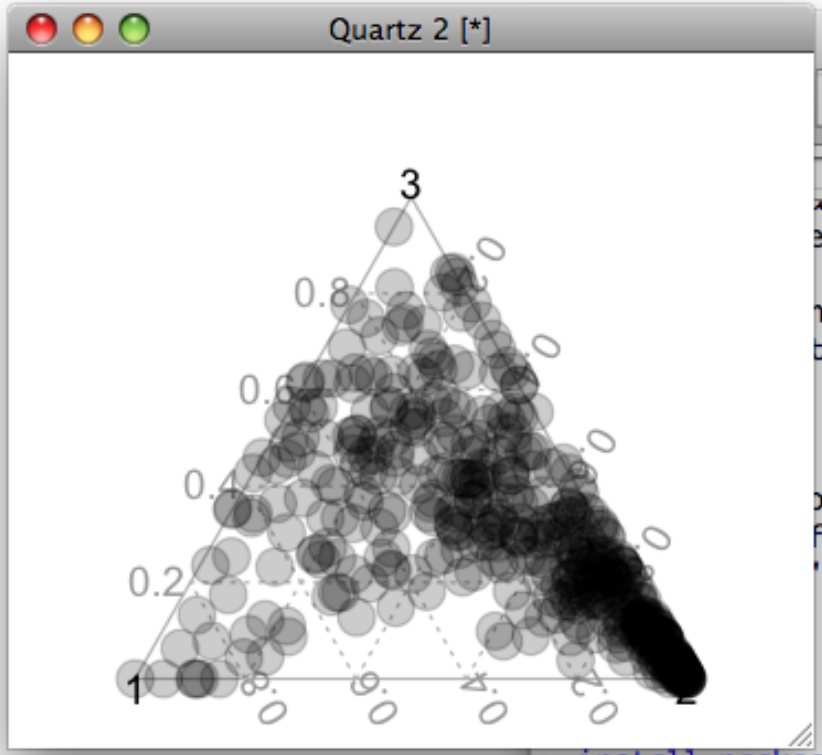
Google

Demos in package 'SportsAnalytics'

Demos in package 'SportsAnalytics'

archeplayers-nba-2d	Two-dimensional archetypal basketball players; illustrative example from manuscript "Archetypal athletes" by Eugster (2011)
archeplayers-nba	Archetypal basketball players based on player statistics; analysis from the manuscript "Archetypal athletes" by Eugster (2011)
archeplayers-soccer-2d	Two-dimensional archetypal soccer players; illustrative example from talk "On the power of modern statistical methodology in soccer analysis – Archetypal soccer players" by Eugster, Abedieh, Schnell, and Augustin (2011)
archeplayers-soccer	Archetypal soccer players based on skill ratings; analysis from the manuscript "Archetypal athletes" by Eugster (2011)

zotero



R Console

Help Search

```
ense()' or 'licence()' für Details dazu.

nschaftsprojekt mit vielen Beitragenden.
tributors()' für mehr Information und 'citation()',
 wie R oder R packages in Publikationen zitiert werden


o()' für einige Demos, 'help()' für on-line Hilfe, oder
für eine HTML Browserschnittstelle zur Hilfe.
', um R zu verlassen.

 (5751) i386-apple-darwin9.8.0]

> install.packages("SportsAnalytics")
> demo("archeplayers-nba-2d", package = "SportsAnalytics")


    demo(archeplayers-nba-2d)
    ---- ~~~~~~~~~~~~~~~~~~~~

Type  <Return>     to start :

> #' @demo Two-dimensional archetypal basketball players; illustrative
> #'   example from manuscript "Archetypal athletes" by Eugster (2011)
>
> library("SportsAnalytics")

> library("archetypes")
Loading required package: modeltools
Loading required package: stats4
```

Also!

Start doing reproducible research because ...

My two cents:

My two cents:

1. Publications which are not reproducible are useless. Source code which computes results shown in papers and the used data sets must be available.

My two cents:

1. Publications which are not reproducible are useless. Source code which computes results shown in papers and the used data sets must be available.
2. Publications which deal with closed data sets have to be "quasi-reproducible" by providing the analyses and artificial data sets (at best with similar characteristics like the closed data sets).

My two cents:

1. Publications which are not reproducible are useless. Source code which computes results shown in papers and the used data sets must be available.
2. Publications which deal with closed data sets have to be "quasi-reproducible" by providing the analyses and artificial data sets (at best with similar characteristics like the closed data sets).
3. Providing some source code and, for example, a binary data file on an author's website is not "reproducible". Similar to publications, source code and data sets must be available "forever" and in a standardized way.

My two cents:

1. Publications which are not reproducible are useless. Source code which computes results shown in papers and the used data sets must be available.
2. Publications which deal with closed data sets have to be "quasi-reproducible" by providing the analyses and artificial data sets (at best with similar characteristics like the closed data sets).
3. Providing some source code and, for example, a binary data file on an author's website is not "reproducible". Similar to publications, source code and data sets must be available "forever" and in a standardized way.
4. A second execution of the analyses on the authors' computers is not "reproducibility". Analyses must be reproducible on an unbiased open platform.

Lawrence (2010):

- It's about habits, not rules.
- It's about good practice: like spell checking.
- It's about courtesy to other researchers.
- It's about keeping track of collaborators work.
- It's about making research reproducible.
- It's something you should all be doing.

References I

Nick Barnes. Publish your computer code: It is good enough. *Nature*, 467:753, 2010. doi: 10.1038/467753a. URL http://www.nature.com/news/2010/101013/full/467753a.html.

Anne-Laure Boulesteix. Reproducible research in epidemiology. Lecture at Medizinische Fakultät, Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie, Ludwig-Maximilians-Universtität München, Winter term, 2010/2011.

Jonathan Buckheit and David Donoho. WaveLab and reproducible research. Statistics Department, Stanford University, CA, USA, 1995. URL http://www-stat.stanford.edu/~donoho/.

Creative Commons. CC BY. http://creativecommons.org/.

Jan de Leeuw. Reproducible research: the bottom line. Statistics Program, Unniversity of California, Los Angeles, CA, USA, 2001. URL http://preprints.stat.ucla.edu/.

David Donoho. How to be a highly cited author in the mathematical sciences. Essay available from http://www.in-cites.com/scientists/DrDavidDonoho.html, 2002.

David L. Donoho. An invitation to reproducible computational research. *Biostatistics*, 11(3):385–388, 2010. doi: 10.1093/biostatistics/kxq028. URL http://biostatistics.oxfordjournals.org/content/11/3/385.full.

Chris Drummond. Replicability is not reproducibility: Nor is it good science. In *Proceedings of the Twenty-Sixth International Conference on Machine Learning: Workshop on Evaluation Methods for Machine Learning IV*, 2009. URL http://www.site.uottawa.ca/~cdrummon/pubs/ICMLws09.pdf.

Manuel J. A. Eugster. Archetypal athletes. Technical Report 113, Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, 2011. URL http://epub.ub.uni-muenchen.de/12336/.

Free Software Foundation. GNU General Public License. http://www.fsf.org/.

Git. Git. http://git-scm.com/.

References II

Esther Herberich, Torsten Hothorn, Daniel Nettle, and Thomas V. Pollet. A re-evaluation of the statistical model in pollet and nettle 2009. *Evolution and Human Behavior*, 31(2):150–151, 2010. doi: 10.1016/j.evolhumbehav.2009.12.003.

Torsten Hothorn. Reproducible statistical analyzes today. Presented at "Validation in Statistics and Machine Learning", Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany, 2010.

Torsten Hothorn and Friedrich Leisch. Case studies in reproducibility. *Briefings in Bioinformatics*, 12(3):288–300, 2011. doi: 10.1093/bib/bbq084. URL http://bib.oxfordjournals.org/content/12/3/288.abstract.

John P. A. Ioannidis, David B. Allison, Catherine A. Ball, Issa Coulibaly, Xiangqin Cui, Aedin C. Culhane, Mario Falchi, Cesare Furlanello, Laurence Game, Giuseppe Jurman, Jon Mangion, Tapan Mehta, Michael Nitzberg, Grier P. Page, Enrico Petretto, and Vera van Noort. Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41(2):149–155, 2009. doi: 10.1038/ng.295. URL http://www.nature.com/ng/journal/v41/n2/full/ng.295.html.

Niels Keiding. Reproducible research and the substantive context. *Biostatistics*, 11(3):376–378, 2010a. doi: 10.1093/biostatistics/kxq033. URL http://biostatistics.oxfordjournals.org/content/11/3/376.full.

Niels Keiding. Reproducible research and the substantive context. Presented at "Validation in Statistics and Machine Learning", Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany, 2010b.

Donald Knuth. *Literate Programming*. 1986.

Roger Koenker and Achim Zeileis. On reproducible econometric research. *Journal of Applied Econometrics*, 24(5): 833–847, 2009. doi: 10.1002/jae.1083. URL http://eeecon.uibk.ac.at/~zeileis/papers/Koenker+Zeileis-2009.pdf.

Max Kuhn. *odfWeave: Sweave Processing of Open Document Format (ODF) files*, 2010. URL http://CRAN.R-project.org/package=odfWeave. R package version 0.7.17.

References III

Neil Lawrence. *MATweave: Integrating MATLAB Software with LaTeX Documents*, 2011. URL `http://staffwww.dcs.shef.ac.uk/people/N.Lawrence/matweave.html`.

Neil D. Lawrence. Making implementations available for the research community. Presented at "Validation in Statistics and Machine Learning", Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany, 2010.

Friedrich Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. URL `http://www.stat.uni-muenchen.de/~leisch/Sweave`. ISBN 3-7908-1517-9.

Friedrich Leisch, Manuel J. A. Eugster, and Torsten Hothorn. Executable papers for the R community: The Rsquared platform for reproducible research. *Procedia Computer Science*, 4:618–626, 2011. ISSN 1877-0509. doi: 10.1016/j.procs.2011.04.065. Proceedings of the International Conference on Computational Science, ICCS 2011; Finalist at the Executable Paper Grand Challenge, prototype available at `http://rsquared.stat.uni-muenchen.de/`.

Russell V. Lenth. *StatWeave*, 2011. URL `http://www.stat.uiowa.edu/~rlenth/StatWeave/`.

David Monniaux. The pitfalls of verifying floating-point computations. *ACM Transactions on Programming Languages and Systems*, 30, 2008. doi: 10.1145/1353445.1353446. URL `http://hal.archives-ouvertes.fr/docs/00/28/14/29/PDF/floating-point-article.pdf`.

Daniel Nettle and Thomas V. Pollet. Partner wealth predicts self-reported orgasm frequency in a sample of chinese women. *Evolution and Human Behavior*, 30(2):146–151, 2009. doi: 10.1016/j.evolhumbehav.2008.11.002.

Roger D. Peng, Francesca Dominici, and Scott L. Zeger. Reproducible epidemiologic research. *American Journal of Epidemiology*, 163(9):783–789, 2006. doi: 10.1093/aje/kwj093. URL `http://aje.oxfordjournals.org/content/163/9/783.abstract`.

References IV

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Science Commons. Open Access Data Protocol. http://sciencecommons.org/projects/publishing/open-access-data-protocol/.

Steven L. Shafer. Shadow of doubt. *Anesthesia & Analgesia*, 112(3), 2011. doi: 10.1213/ANE.0b013e31820ad3b7. URL http://www.anesthesia-analgesia.org/content/112/3/498.full.

Victoria Stodden. Enabling reproducible research: Open licensing for scientific innovation. *International Journal of Communications Law and Policy*, 13, 2009. URL http://www.stanford.edu/~vcs/papers/ERROLSI03092009.pdf.

Apache Subversion. Apache subversion. http://subversion.apache.org/.

Nathan Whitehead and Alex Fit-Florea. Precision & performance: Floating point and ieee 754 compliance for nvidia gpus. Technical white paper by NVIDA, 2011. URL http://developer.download.nvidia.com/assets/cuda/files/NVIDIA-CUDA-Floating-Point.pdf.

Wikipedia. Computational science. Wikipedia entry, http://en.wikipedia.org/wiki/Computational_science, visited on November 08, 2011a.

Wikipedia. Fermat's last theorem. Wikipedia entry, http://en.wikipedia.org/wiki/Fermat%27s_Last_Theorem, visited on November 08, 2011b.

Images

page 6: Page 85 of Diophantus' Arithmetica (with problem II.VIII); via
`http://en.wikipedia.org/wiki/Fermat%27s_Last_Theorem`

page 12: PHD Comics: Reserach Diagram/Research Reality; `http://www.phdcomics.com/comics.php?f=961`

page 49: Volkswagen Commercial Vehicles: Choose the right tool, 1; via
`http://adsoftheworld.com/media/print/volkswagen_commercial_vehicles_choose_the_right_tool_1`

page 71: Hilti TE 1000-AVR; via `http://forum.bauforum24.biz/forum/index.php?showtopic=35548`