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

Manuel J. A. Eugster

Institut für Statistik
Ludwig-Maximilians-Universität München

Herbstworkshop, Ruhr-Universität Bochum, 2011
Joint work with Friedrich Leisch and Torsten Hothorn:

Joint work with Anne-Laure Boulesteix:

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

Scientific community:
“Cornerstone of the scientific method!”
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)
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)
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

What your research supposedly looks like:

- Data
- Data Acquisition
- Prototype
- Controller
- Computer

Figure 1. Experimental Diagram

What your research actually looks like:

- Computer
- Controller
- Data Acquisition
- Data

Figure 2. Experimental Mess

(Touch anything and DIE)
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
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.
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. The study was retracted in December 2010 for lack of IRB approval. A subsequent investigation by Klinikum Ludwigshafen determined that the study was fabricated. Landesärztekammer Rheinland-Pfalz (LÄK-RLP), the State Medical Association of Rheinland-Pfalz, Germany, serves as the Institutional Review Board (IRB) for clinical research at Klinikum Ludwigshafen, where Dr. Boldt’s recent research was conducted. Following the retraction LÄK-RLP undertook a careful evaluation of the status of IRB approval for research conducted by Dr. Boldt. Today, on behalf of the editors of multiple journals concerned about the possibility of unethical research, I received from LÄK-RLP a determination, to the best of their ability, of the status of IRB approval for 101 articles published by Dr. Boldt. Table 1 lists the 22 articles published in *Anesthesia & Analgesia* for which LÄK was not able to verify IRB approval. These articles are hereby retracted. The full list of 89 articles that for which LÄK could not verify IRB approval has been circulated to the editors of the affected journals. Early next week, after those editors have had an opportunity to review the list and notify their readers, we will post the full list of 89 articles for which LÄK was unable to verify IRB approval. We will also list the 11 articles for which IRB approval has been verified, and the 2 articles for which IRB approval was deemed by LÄK to not be necessary. The retraction of an article for lack of IRB approval means that the research was unethical, and that IRB approval for the research was misrepresented in the article. It does not mean that the research results are fraudulent. That is a different question. Klinikum Ludwigshafen has commissioned an investigating committee to systematically verify the findings presented in Dr. Boldt’s articles against patient and laboratory records to determine the authenticity of the research results stated in the article. That is a slow and painstaking process. I will keep our readers informed of the findings of the investigating committee at Klinikum Ludwigshafen. It is possible that articles retracted for lack of IRB approval may be retracted again (!) for research fraud to provide a full accounting of Dr. Boldt’s research.
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.
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.

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

<table>
<thead>
<tr>
<th>Statement in article</th>
<th>Finding by LÄK-RLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fifty consecutive patients undergoing elective coronary artery bypass grafting were studied after approval of the IRB.</td>
<td>For the respective study, there does not exist an approval by the respective IRB.</td>
</tr>
<tr>
<td>And after receiving individual written informed consent.</td>
<td>For the respective study, there does not exist written informed consent.</td>
</tr>
<tr>
<td>The patients were prospectively randomized into one of the two groups by a computer-generated list and sealed envelopes.</td>
<td>For the respective study, there does not exist a prospective randomization process into two groups by a computer-generated list and sealed envelopes.</td>
</tr>
<tr>
<td>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.</td>
<td>For the respective study, there does not exist a written questionnaire sent to the patients’ primary physicians approximately 60 days after hospital discharge.</td>
</tr>
</tbody>
</table>

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):

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Example</th>
<th>Data</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>19</td>
<td>6.5</td>
<td>2.5</td>
</tr>
</tbody>
</table>

(*) 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 unrepeatable 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...”
Reproducibility in the view of heterogeneous technology:

- Text
- Data
- Code
- Application software
- Run-time system
- Operating system
- Hardware

“I used a Mac...”
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:

\[ a = (a_1, a_2, \ldots, a_n)' \]

\[ b = (b_1, b_2, \ldots, b_n)' \]

\[ a \cdot b = \sum_{i=1}^{n} a_i \times b_i = a_1 \times b_1 + a_2 \times b_2 + \cdots + a_n \times b_n \]

We suppose (in an arbitrary mathematical software) that a function \texttt{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 :

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

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

return $t$
Serial method:

\[ t = 0 \]
\[ \text{for } i \text{ from 1 to 4 :} \]
\[ p = \text{rn}(a_i \times b_i) \]
\[ t = \text{rn}(t + p) \]
\[ \text{return } t \]

FMA method:

\[ t = 0 \]
\[ \text{for } i \text{ from 1 to 4 :} \]
\[ t = \text{rn}(a_i \times b_i + t) \]
\[ \text{return } t \]
Serial method:

\[ t = 0 \]
for \( i \) from 1 to 4 :
\[ p = \text{rn}(a_i \times b_i) \]
\[ t = \text{rn}(t + p) \]
return \( t \)

FMA method:

\[ t = 0 \]
for \( i \) from 1 to 4 :
\[ t = \text{rn}(a_i \times b_i + t) \]
return \( t \)

Parallel method:

\[ p_1 = \text{rn}(a_1 \times b_1) \]
\[ p_2 = \text{rn}(a_2 \times b_2) \]
\[ p_3 = \text{rn}(a_3 \times b_3) \]
\[ p_4 = \text{rn}(a_4 \times b_4) \]
\[ s_l = \text{rn}(p_1 + p_2) \]
\[ s_r = \text{rn}(p_3 + p_4) \]
\[ t = \text{rn}(s_l + s_r) \]
return \( t \)
Serial method:  

\[
\begin{align*}
  t &= 0 \\
  \text{for } i \text{ from 1 to 4 :} \\
  p_i &= \text{rn}(a_i \times b_i) \\
  t &= \text{rn}(t + p_i) \\
\end{align*}
\]

return \( t \)

FMA method:  

\[
\begin{align*}
  t &= 0 \\
  \text{for } i \text{ from 1 to 4 :} \\
  t &= \text{rn}(a_i \times b_i + t) \\
\end{align*}
\]

return \( t \)

Parallel method:  

\[
\begin{align*}
  p_1 &= \text{rn}(a_1 \times b_1) \\
  p_2 &= \text{rn}(a_2 \times b_2) \\
  p_3 &= \text{rn}(a_3 \times b_3) \\
  p_4 &= \text{rn}(a_4 \times b_4) \\
  s_l &= \text{rn}(p_1 + p_2) \\
  s_r &= \text{rn}(p_3 + p_4) \\
  t &= \text{rn}(s_l + s_r) \\
\end{align*}
\]

return \( t \)

<table>
<thead>
<tr>
<th>Method</th>
<th>Result</th>
<th>float value</th>
</tr>
</thead>
<tbody>
<tr>
<td>exact</td>
<td>0.0559587528435...</td>
<td>0x3D65350158...</td>
</tr>
<tr>
<td>serial</td>
<td>0.0559588074</td>
<td>0x3D653510</td>
</tr>
<tr>
<td>FMA</td>
<td>0.0559587515</td>
<td>0x3D653501</td>
</tr>
<tr>
<td>parallel</td>
<td>0.0559587478</td>
<td>0x3D653500</td>
</tr>
</tbody>
</table>
• “It requires so much time.”
• “It requires so much time.”

Yes! And motivation, patience, good organization, etc.
• “It requires so much time.”

“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)
Figure 2: (a) Visualization of the $\alpha$ 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.

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Team</th>
<th>Role</th>
<th>Min</th>
<th>FGM</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archetype 1</td>
<td>Kevin Durant</td>
<td>OKL</td>
<td>SF</td>
<td>3241</td>
<td>794</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Archetype 2</td>
<td>Dwayne Jones</td>
<td>PHO</td>
<td>C</td>
<td>7</td>
<td>0</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Archetype 3</td>
<td>Jason Kidd</td>
<td>DAL</td>
<td>PG</td>
<td>2883</td>
<td>284</td>
<td>0.06</td>
<td>0.00</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Team</th>
<th>Role</th>
<th>Min</th>
<th>FGM</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archetype 1</td>
<td>Kevin Durant</td>
<td>OKL</td>
<td>SF</td>
<td>3241</td>
<td>794</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Archetype 2</td>
<td>Dwayne Jones</td>
<td>PHO</td>
<td>C</td>
<td>7</td>
<td>0</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Archetype 3</td>
<td>Jason Kidd</td>
<td>DAL</td>
<td>PG</td>
<td>2883</td>
<td>284</td>
<td>0.06</td>
<td>0.00</td>
</tr>
</tbody>
</table>

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...
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;
The three players (red points) nearest to the respective archetypes (red crosses) are:

\begin{center}
\begin{tabular}{rlllrrrrr}
\hline
& Name & Team & Role & Min & FGM & $\alpha_{\cdot 1}$ & $\alpha_{\cdot 2}$ & $\alpha_{\cdot 3}$ \\
\hline
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 \\
\hline
\end{tabular}
\end{center}

Archetype 1 and 3 have well-defined nearest observations;
Stangle("archeplayers.Rnw") ⇒ archeplayers.R:

```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
Transactions

svn://svn.stat.uni-muenchen.de/CompStat/papers/prototypes/archeplayers: svn: Unknown host svn.stat.uni-muenchen.de: svn.stat.uni-muenchen.de

[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]
[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.
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.
The present paper applies the statistical method archetypal analysis to sports data. This enables to compute outstanding—positively and/or negatively—athletes, i.e., the archetypal athletes. Statements like "Dirk Nowitzki is the best basketball player. No, it's Kevin Durant!" can be then discussed completely data-driven and with a well-defined and reproducible amount of subjectivity. The proposed way is (1) to estimate the archetypes, i.e., the archetypal athletes, then (2) to identify the athletes as different types of "good" and "bad" athletes, and finally (3) to set all athletes in relation to the archetypes (using the $\alpha$ 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 Sports-Analytics 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
R> demo("archeplayers-***", package = "SportsAnalytics")
```

The source code file for a demo is accessible via:

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

### References

Christian Bauckhage and Christian Thurau. Making archetypal analysis practical. In
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
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
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)
> 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:

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.
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


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


References II


References III


References IV


Images

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


**page 49:** Volkswagen Commercial Vehicles: Choose the right tool, 1; via [http://adsoftheworld.com/media/print/volkswagen_commercial_vehicles_choose_the_right_tool_1](http://adsoftheworld.com/media/print/volkswagen_commercial_vehicles_choose_the_right_tool_1)