A Compendium Platform for Reproducible, R-based Research with a focus on Statistics Education

UseR! 2008 - Patrick Wessa - K.U.Leuven Association, Lessius Dept. of Business Studies

Introduction

- Acknowledgments
- Motivation (based on frustration)
- Reproducible Research and the Compendium:
  - Literature
  - The compendium redefined
  - Proposed solution
- Screenshots
- Conclusions & Future work

http://www.freestatistics.org >> Publications
http://www.wessa.net/download/user2008.pdf

Acknowledgments

- Funding (we accept money):
  - K.U.Leuven Association, OOF 2007/13
  - Donations from private companies
- Contributors:

  Bart Baesens, Eric Bloemen, Eddy Borghers, Christophe Croux, Claude Doom, Dirk Janssens, Christine Lourdon, Koen Milis, Stephan Poelmans, Riko van Dijk, Guido Van Rompuy, Ed van Stee, Larry Weldon, Patrick Wessa (www.freestatistics.org)

My frustration

- Teaching Time Series Analysis
- Exam question:
  Compute (1-B) Y[t] if you know that
  Y[t] = {5, 8, 2, 3, 7, 1, 4}
  BY[t] = Y[t-1]

My frustration

- Teaching Time Series Analysis
- Exam question:
  Compute (1-B) Y[t] if you know that
  Y[t] = {5, 8, 2, 3, 7, 1, 4}
  BY[t] = Y[t-1]

- Result
  - Less than 8% of students got it right.
  - More than 90% of students could prove Wold's decomposition theorem!

Conclusion?

- I am an extremely bad educator.
- I shouldn't have asked that silly question: Students can only reproduce theories – they are no required to understand them!
- ...
- Or maybe there is something wrong with our approach towards statistics education?

A new approach is needed

- Within the pedagogical paradigm of (social) constructivism:
  - Interaction & collaboration (peer review)
  - Experimentation
  - Responsibility (social control)

  => learning & computing technology

  => we need to Free Statistics of irreproducible research

  => www.FreeStatistics.org

Computing Reproducible ~~Research~~ and the Compendium

Green's comment

- Now the methodology is often so complicated and computationally intensive that the standard dissemination vehicle of the 16-page refereed learned journal paper is no longer adequate. ...
  Most statistics papers, as published, no longer satisfy the conventional scientific criterion of reproducibility: could a reasonably competent and adequately equipped reader obtain equivalent results if the experiment or analysis were repeated?

*Source: Peter J. Green

Claerbout'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 that complete set of instructions that generated the figures.

*Source: Jan de Leeuw

Jan de Leeuw's comments*

- First, there is no reason to single out figures. The same ``Principle'' obviously applies to tables, standard errors, and so on. The fact that figures often happen to be easier to reproduce, does not preclude that we should apply the same rule to any form of computer-generated output.
- Second, there is no reason to limit the Claerbout's Principle to published articles. We can make exactly the same statement about our lectures and teaching, certainly in the context of graduate teaching. We must be able to give our students our code and our graphics files, so that they can display and study them on their own computers (and not only on our workstations, or in crowded university labs).
- And third, and perhaps most importantly, it is not clearly defined what a ``software environment'' is. Buckheit and Donoho apply the principle in such a way that everybody who wants to check their results is forced to buy MatLab(R). Not Mathematica(R), Macsyma(R), or S-plus(R). Those you may need to buy for other articles. This violates the Freeware Principle...

*Source: Jan de Leeuw, Reproducible Research: the Bottom Line, 2001, online

Sweave package

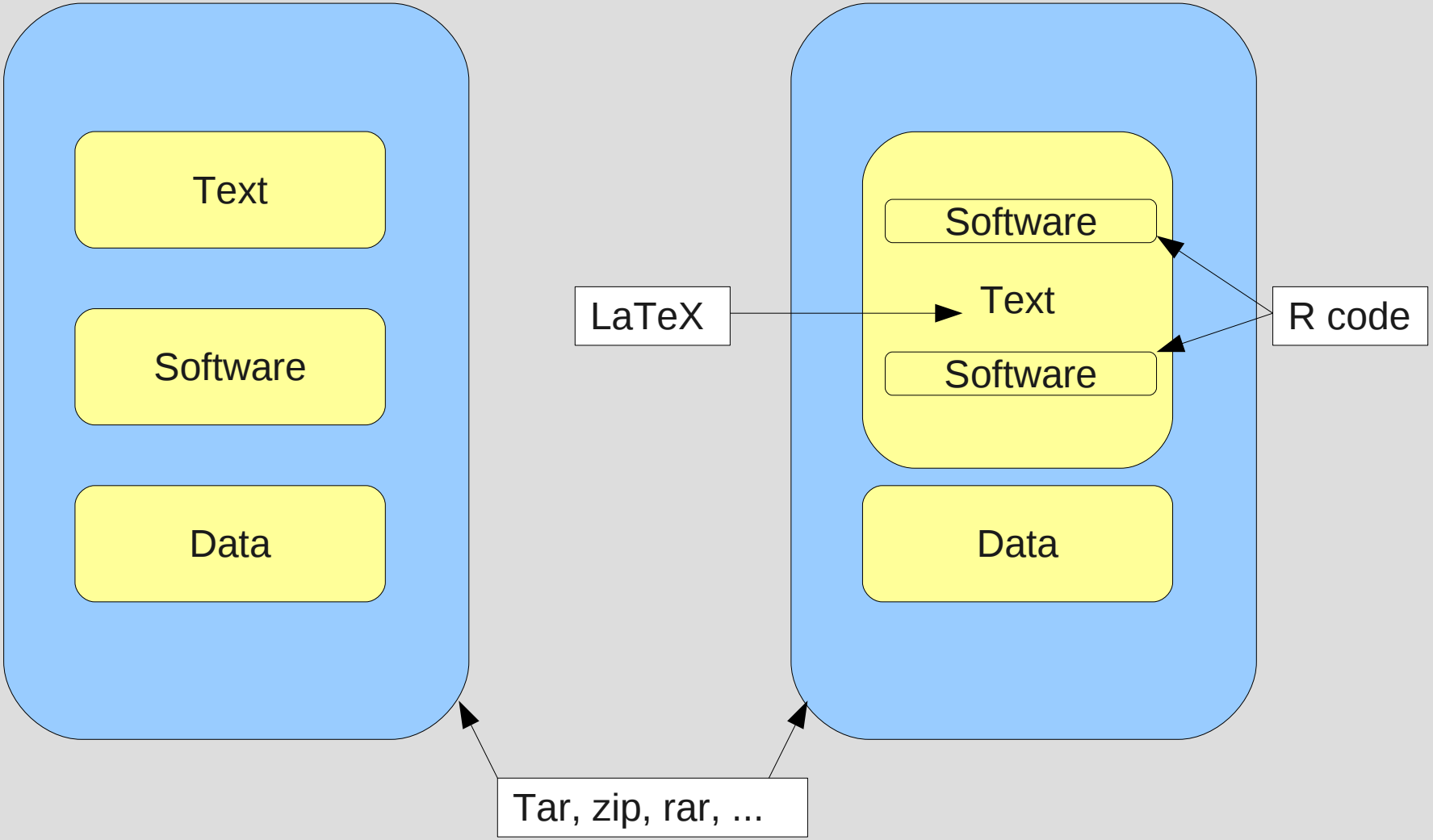
- Excellent solution (in general)
- Somewhat impractical for education because the student:
  - is required to DIE (Download, Install, Execute)
  - must have a working knowledge of LaTeX and R
  - must recreate a working compendium (for each submission)
- Not designed with educational research in mind: there is no way to monitor/measure the actual learning activities

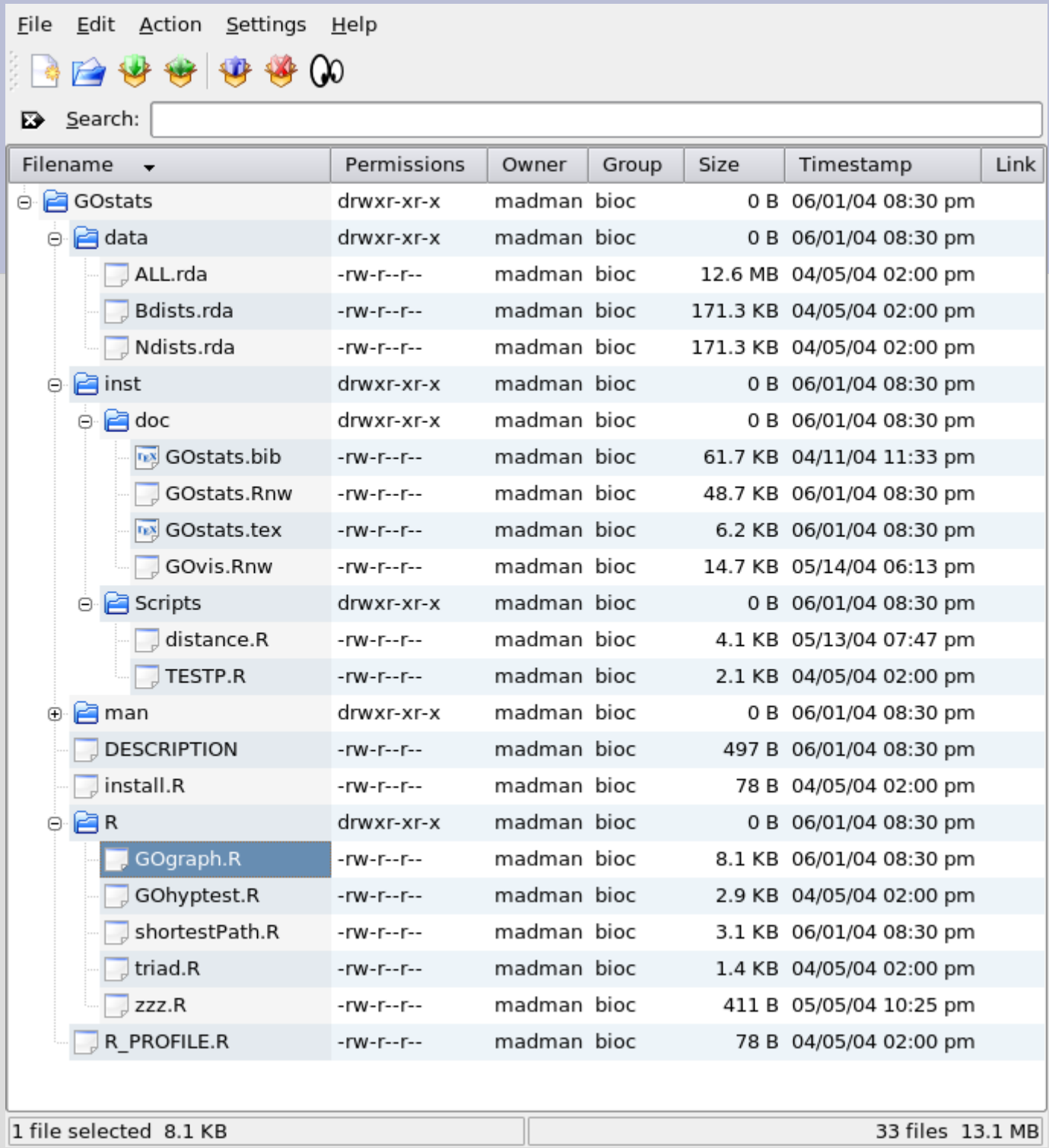
Compendium

- Original definition:

  An electronic collection of Text, Data and Software that allows the reader to reproduce the research that is presented in the document

Compendium


Text
Software
Data
LaTeX
Software
Text
Software
R code
Data
Tar, zip, rar, ...


File Edit Action Settings Help
Search:
Filename Permissions Owner Group Size Timestamp Link
GOstats drwxr-xr-x madman bioc 0 B 06/01/04 08:30 pm
data drwxr-xr-x madman bioc 0 B 06/01/04 08:30 pm
ALL.rda -rw-r--r-- madman bioc 12.6 MB 04/05/04 02:00 pm
Bdists.rda -rw-r--r-- madman bioc 171.3 KB 04/05/04 02:00 pm
Ndists.rda -rw-r--r-- madman bioc 171.3 KB 04/05/04 02:00 pm
inst drwxr-xr-x madman bioc 0 B 06/01/04 08:30 pm
doc drwxr-xr-x madman bioc 0 B 06/01/04 08:30 pm
GOstats.bib -rw-r--r-- madman bioc 61.7 KB 04/11/04 11:33 pm
GOstats.Rnw -rw-r--r-- madman bioc 48.7 KB 06/01/04 08:30 pm
GOstats.tex -rw-r--r-- madman bioc 6.2 KB 06/01/04 08:30 pm
GOvis.Rnw -rw-r--r-- madman bioc 14.7 KB 05/14/04 06:13 pm
Scripts drwxr-xr-x madman bioc 0 B 06/01/04 08:30 pm
distance.R -rw-r--r-- madman bioc 4.1 KB 05/13/04 07:47 pm
TESTP.R -rw-r--r-- madman bioc 2.1 KB 04/05/04 02:00 pm
man drwxr-xr-x madman bioc 0 B 06/01/04 08:30 pm
DESCRIPTION -rw-r--r-- madman bioc 497 B 06/01/04 08:30 pm
install.R -rw-r--r-- madman bioc 78 B 04/05/04 02:00 pm
R drwxr-xr-x madman bioc 0 B 06/01/04 08:30 pm
GOgraph.R -rw-r--r-- madman bioc 8.1 KB 06/01/04 08:30 pm
GOhyptest.R -rw-r--r-- madman bioc 2.9 KB 04/05/04 02:00 pm
shortestPath.R -rw-r--r-- madman bioc 3.1 KB 06/01/04 08:30 pm
triad.R -rw-r--r-- madman bioc 1.4 KB 04/05/04 02:00 pm
zzz.R -rw-r--r-- madman bioc 411 B 05/05/04 10:25 pm
R_PROFILE.R -rw-r--r-- madman bioc 78 B 04/05/04 02:00 pm
1 file selected 8.1 KB
33 files 13.1 MB

Compendium redefined

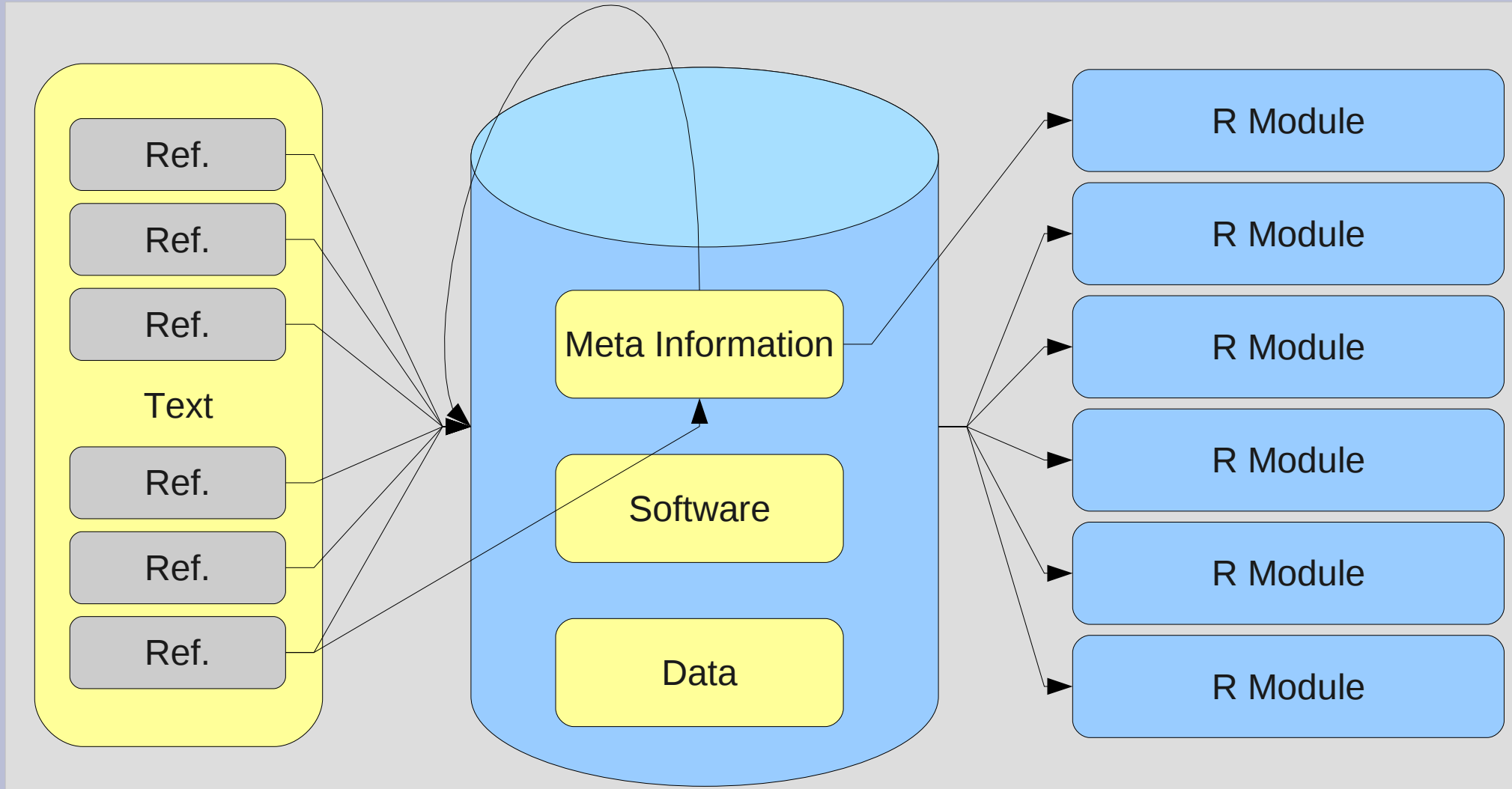
- New definition:

  A document with (open-access) references to (remotely) archived Computations (including Data, Meta-data, and Software) that allow us to reproduce, and reuse the underlying analysis

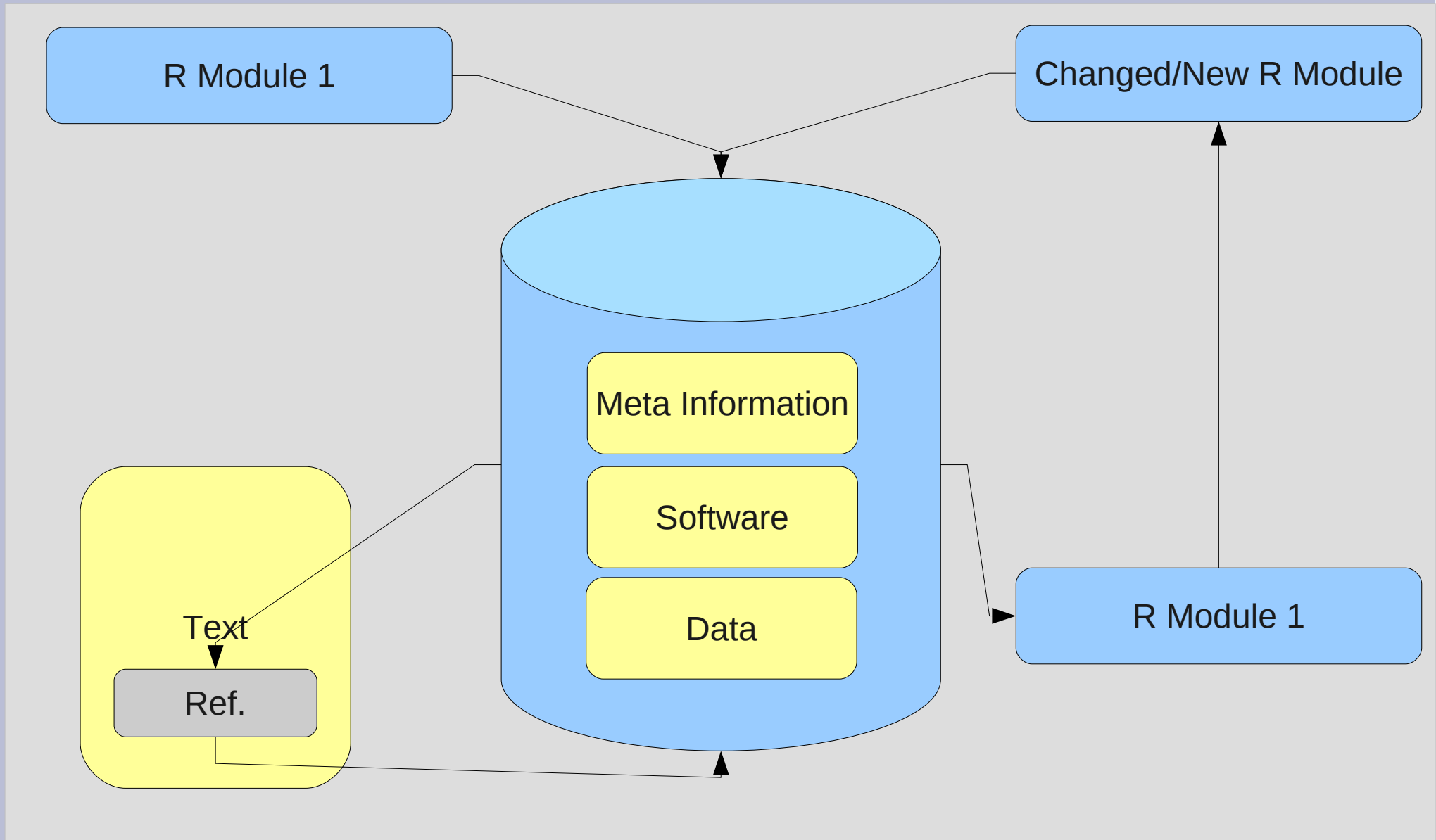
- Complete separation of:
  - text and computing
  - computational result and computing infrastructure

  => the compendium platform is a tool for collaboration, dissemination, and monitoring.

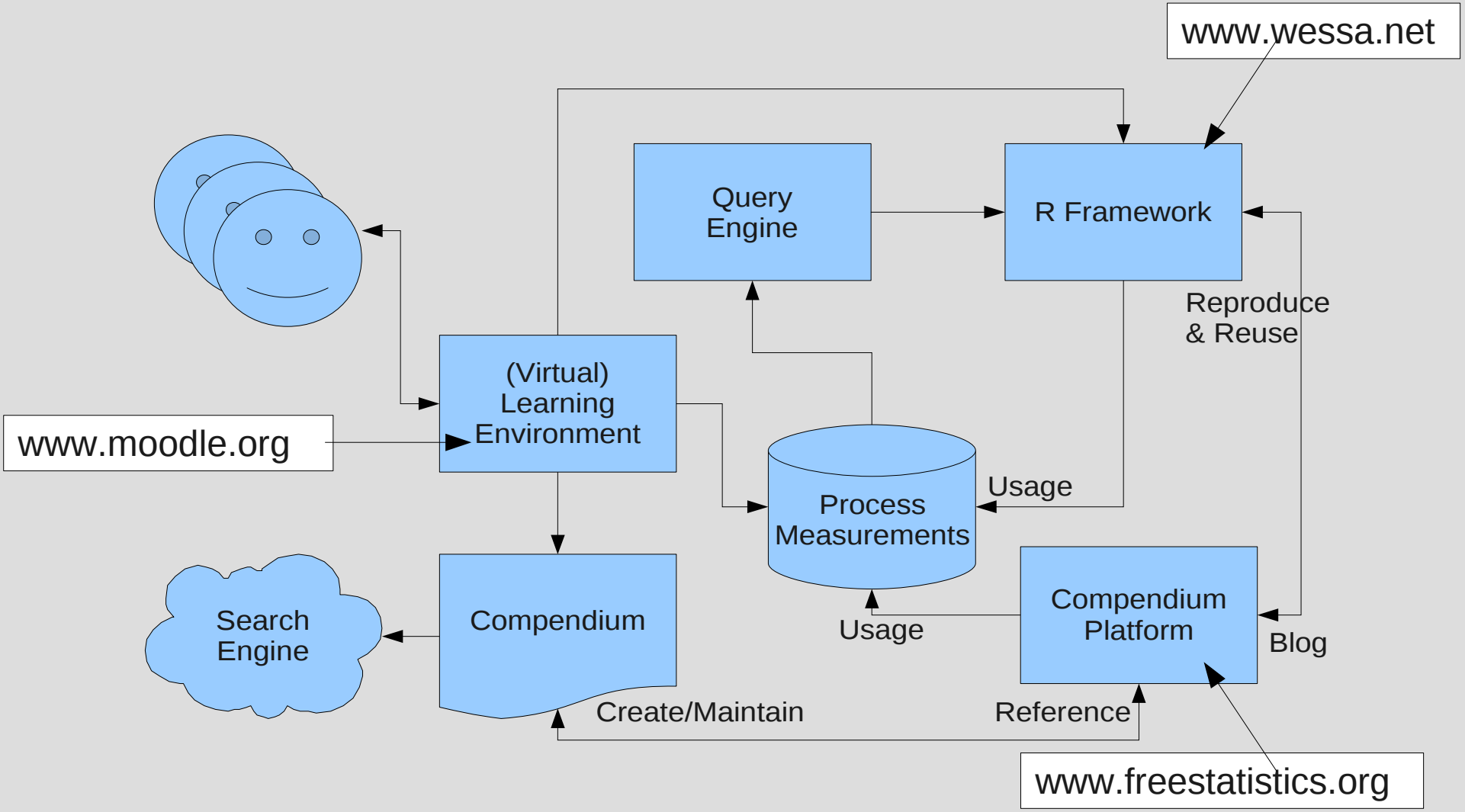
Computations Database



Compendium Dynamics



Learning System or Educational Laboratory?

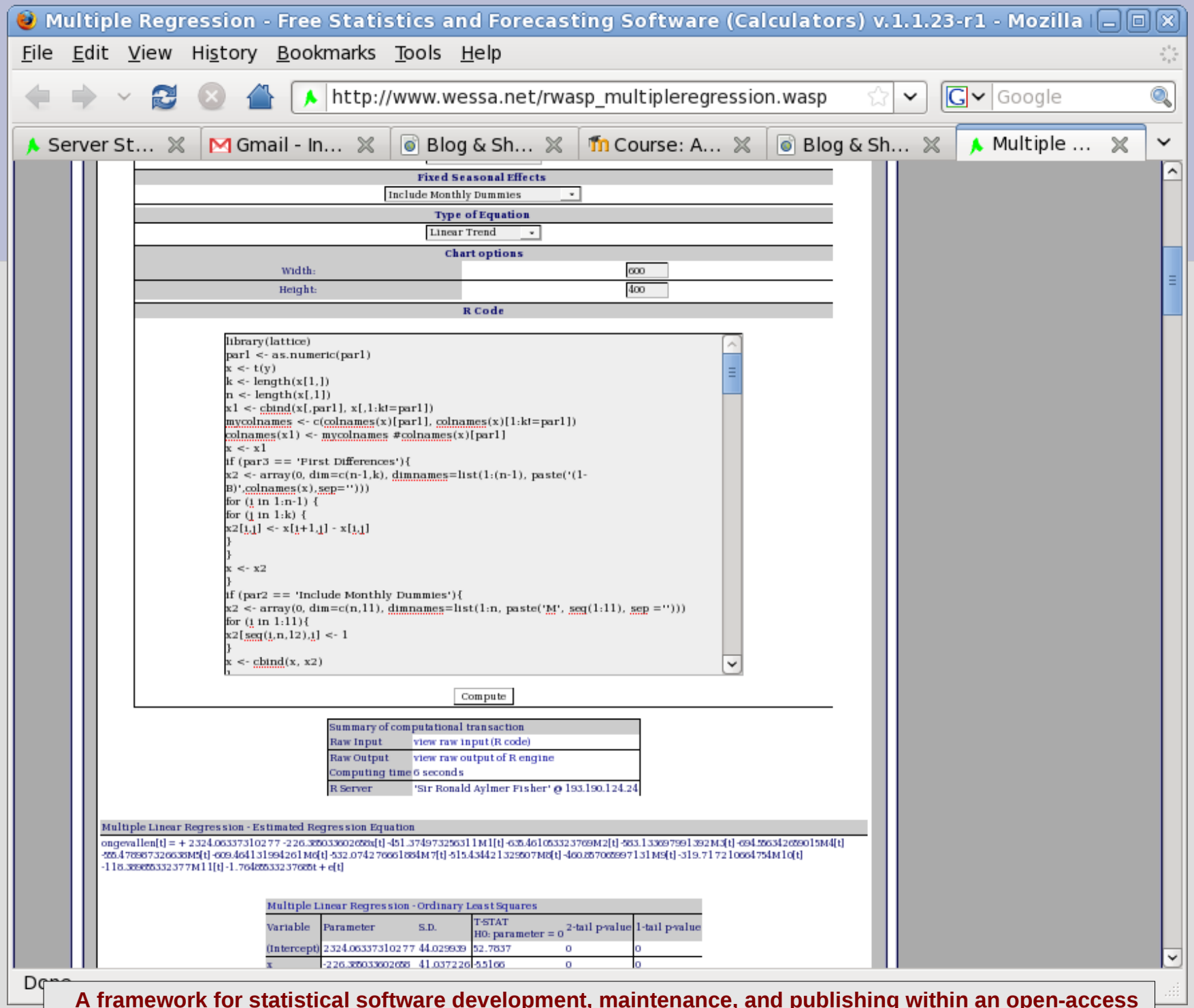


Examples of Compendia

http://www.wessa.net/download/tutorial.pdf
(Descriptive Statistics – Central Tendency)
http://www.wessa.net/download/tutorial1.pdf
(Time Series Analysis - Introduction)

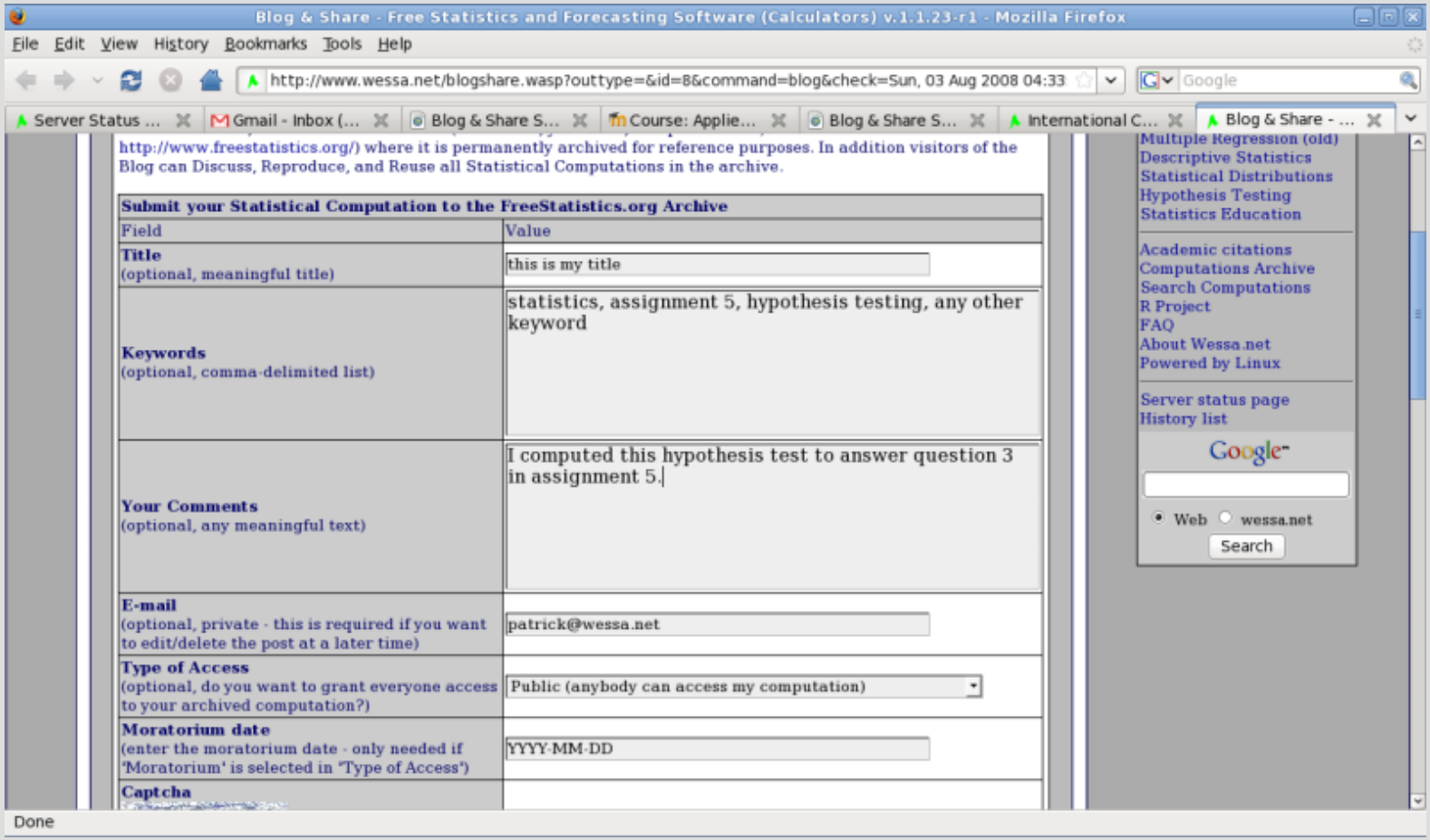
Note: both documents are “work in progress”
Please, send corrections & suggestions to
patrick@wessa.net

Screenshots



A framework for statistical software development, maintenance, and publishing within an open-access business model, 2008, Computational Statistics

Computations are “blogged” (not archived)



workshop - OpenOffice.org Writer

File Edit View Insert Format Table Tools Window Help

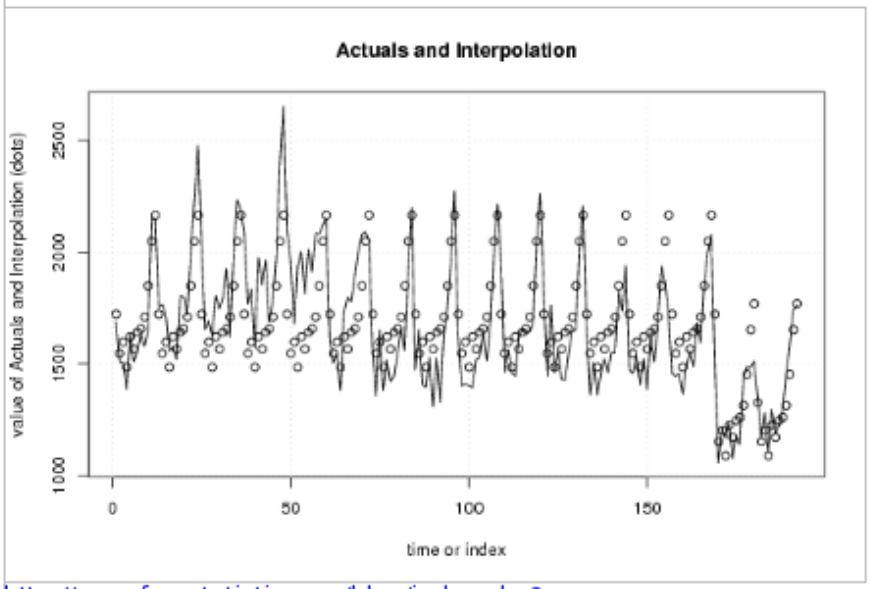
Equation Expand Config

Default Garamond 12

Question 2: Investigate the prediction errors of the model that you used in question 1. Are the underlying regression assumptions satisfied?

De Adjusted R-squared is gelijk aan 0,6412. Dit wil zeggen dat we 64% van de wijzigingen van het aantal verkeersslachtoffers kunnen verklaren. Het resterende gedeelte (36%) kunnen we niet verklaren aan de hand van ons model maar zijn bijvoorbeeld te wijten aan uitzonderlijke weersomstandigheden.

Het volgende waar we heen zullen kijken is de Interpolation Plot. De stippellijn op deze grafiek geeft het werkelijke aantal slachtoffers weer. De volle lijn geeft het aantal verkeersslachtoffers weer die voorspeld zijn door het model.

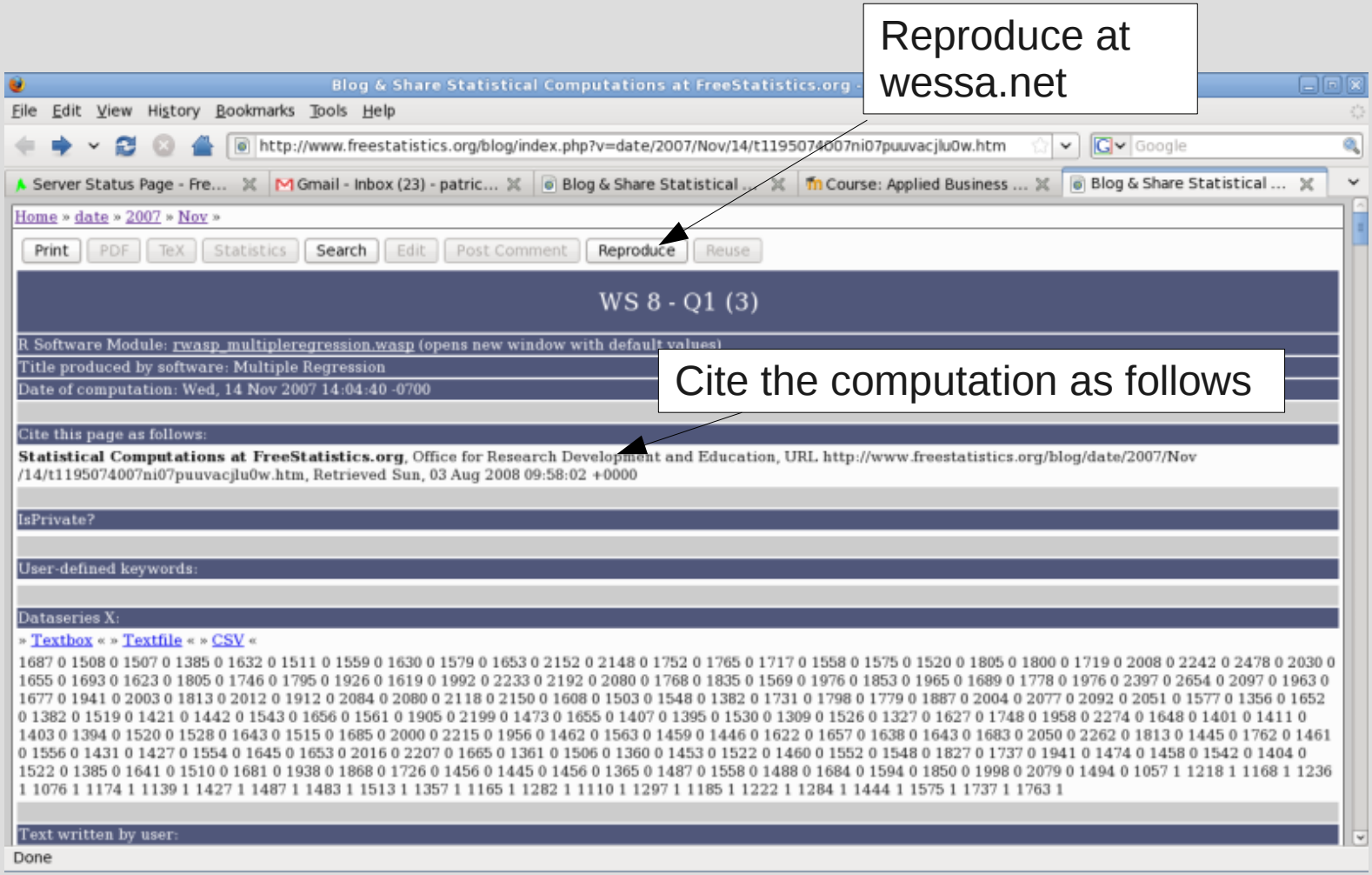


http://www.freestatistics.org/blog/index.php?v=date/2007/Nov/14/t1195074007ni07puuvacjlu0w.htm

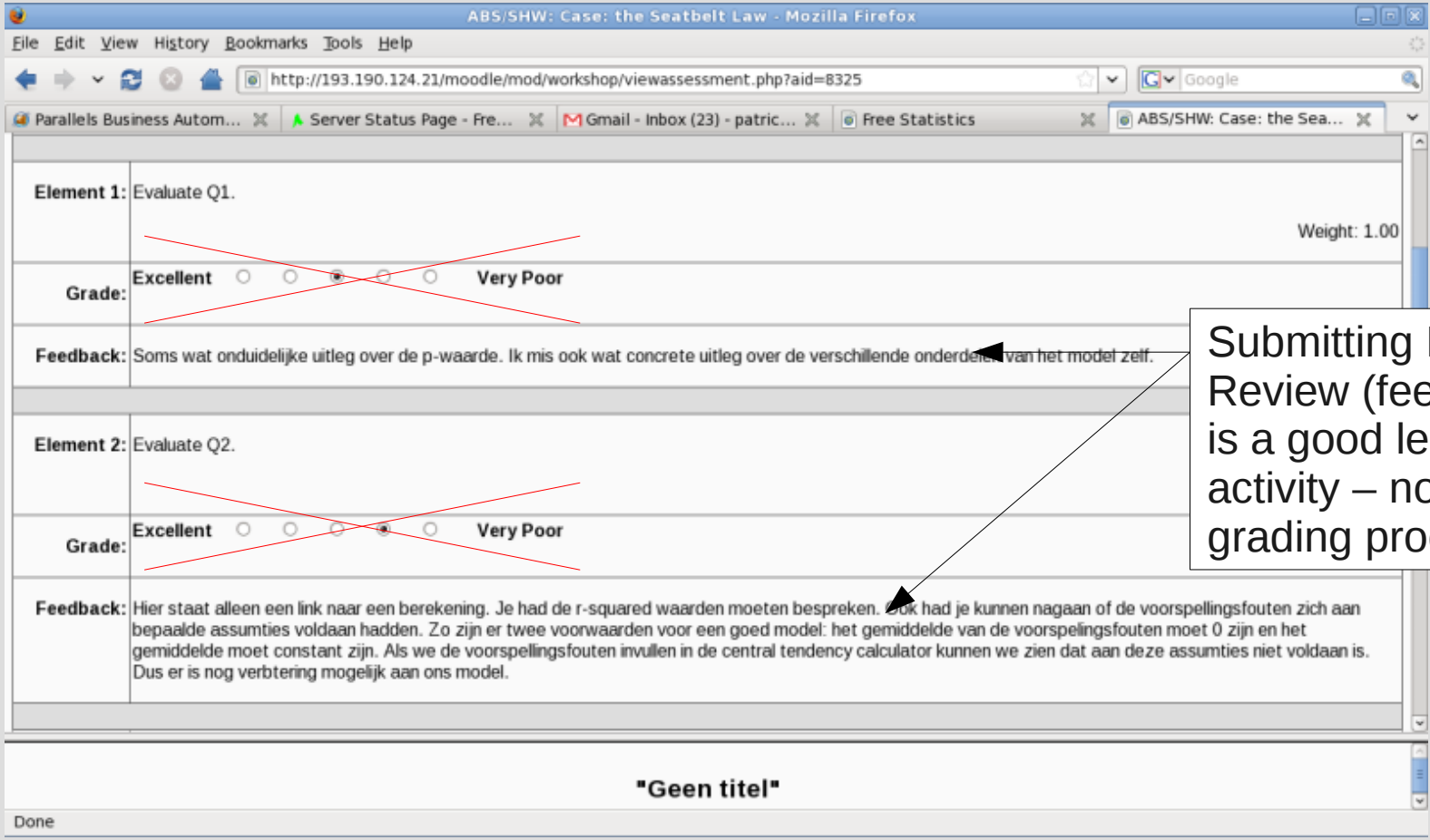
We zien ook hier een dalende trend op lange termijn. Op het einde zien we een structurele breuk, dit is het effect van de seatbelt law (-395). Er treedt ook een

Learning Statistics based on the Compendium and Reproducible Computing, Proceedings of the International Conference on Education and Information Technology (ICEIT'08), Berkeley, San Francisco, USA

Snapshot of “Blogged” Computation

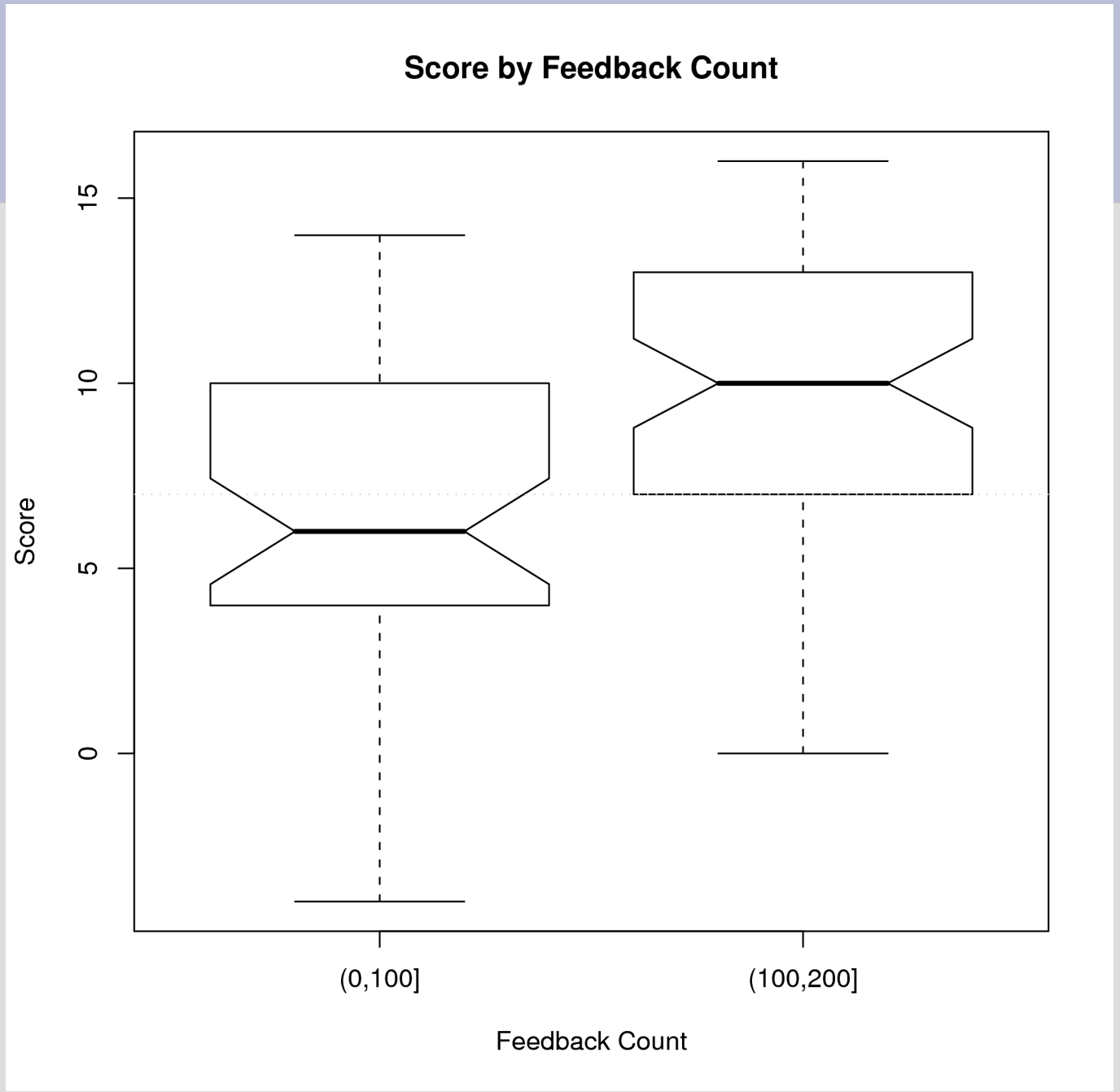


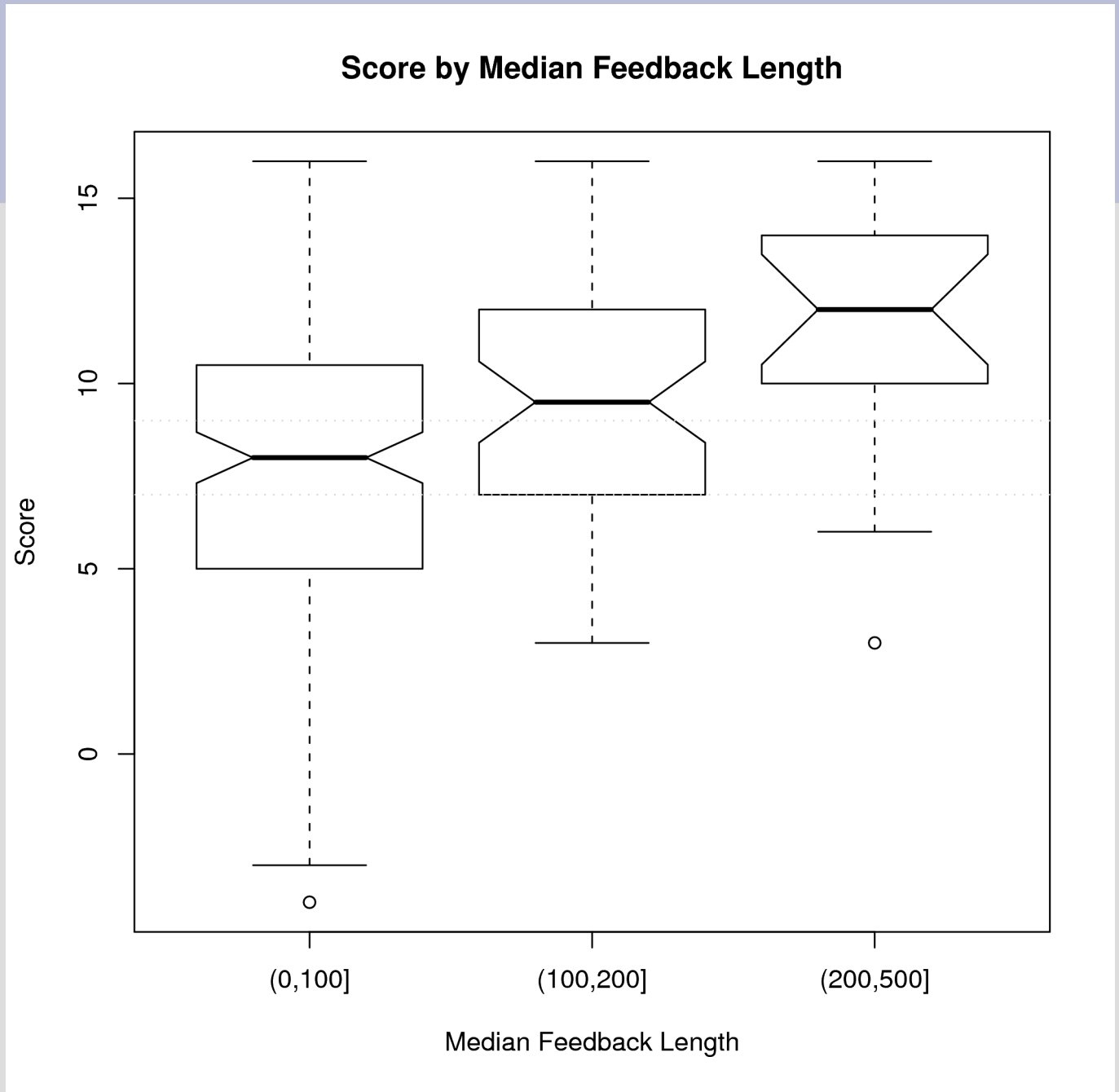
Feedback (Peer Review)

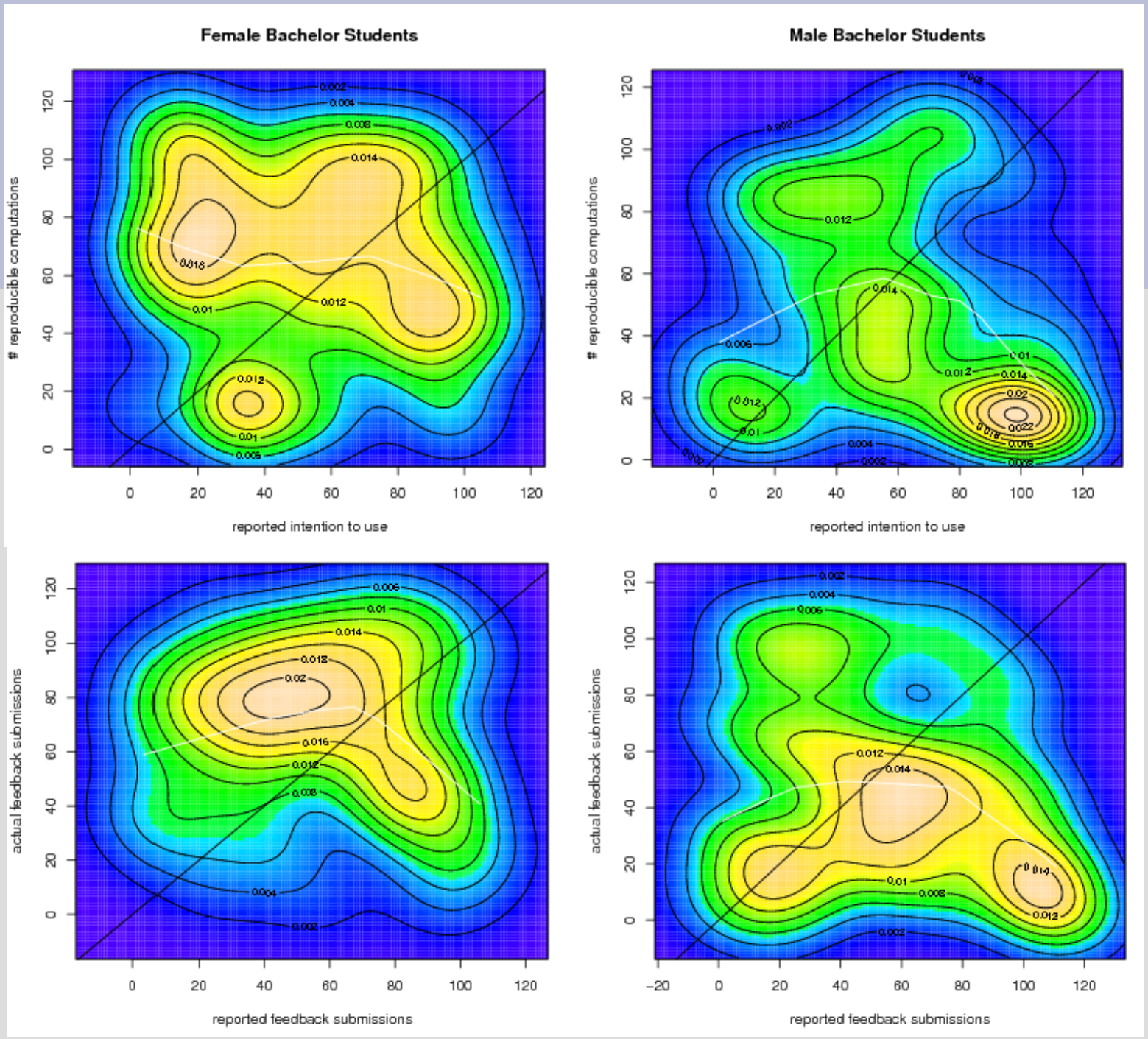


Submitting Peer Review (feedback) is a good learning activity – not a good grading procedure

How Reproducible Research Leads to Non-Rote Learning Within a Socially Constructivist E-Learning Environment, Proceedings of the 7th European Conference on e-Learning (ECEL'08), Cyprus


Score by Feedback Count
Score
0
5
10
15
(0,100]
(100,200]
Feedback Count


Score by Median Feedback Length
Score
0
5
10
15
(0,100]
(100,200]
(200,500]
Median Feedback Length



Measurement and Control of Statistics Learning Processes based on Constructivist Feedback and Reproducible Computing, Proceedings of the 3rd International Conference on Virtual Learning (ICVL '08), Romania

http://www.wessa.net/rwasp_icvl2008.wasp

Conclusions & Future work

- Reproducible Computing can be made easy (for students)
- RC improves statistics learning
- RC allows us to research learning activities (based on actual – not reported – data)

- New features (social interaction, collaboration)
- RC for scientists
- RC for scientific publishing

Some References

- **J. Buckheit and D. L. Donoho**. Wavelab and reproducible research. In A. Antoniadis, editor, Wavelets and Statistics. Springer-Verlag, 1995.
- **Peter J. Green**. Diversities of gifts, but the same spirit. The Statistician, pages 423–438, 2003.
- **T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M.L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander**. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286:531–537, 1999.
- **David L. Donoho, Xiaoming Huo**, BeamLab and Reproducible Research, International Journal of Wavelets, Multiresolution and Information Processing, 2004
- **Roger D. Peng, Francesca Dominici, and Scott L. Zeger**, Reproducible Epidemiologic Research, American Journal of Epidemiology, 2006
- **R. Gentleman**, Reproducible Research: A Bioinformatics Case Study, Bioconductor
- **R. Gentleman**, Applying Reproducible Research in Scientific Discovery, BioSilico, 2005
- **Jan de Leeuw**, Reproducible Research: the Bottom Line, 2001, online

Some References

- **Roger Koenker, Achim Zeileis**, Reproducible Econometric Research (A Critical Review of the State of the Art), Department of Statistics and Mathematics Wirtschaftsuniversität Wien, Research Report Series, Report 60, November 2007
- **Robert Gentleman, Duncan Temple Lang**, Statistical Analyses and Reproducible Research, http://www.bepress.com/bioconductor/paper2
- **Schwab, M., Karrenbach, N. and Claerbout, J.** Making scientific computations reproducible, Computing in Science & Engineering, 2 (6), pp. 61-67, 2000.
- **Robert Gentleman**, Some Perspectives on Statistical Computing, online
- **Leisch, F.**, “Sweave and beyond: Computations on text documents”, Proceedings of the 3rd International Workshop on Distributed Statistical Computing, 2003, Vienna, Austria, ISSN 1609-395

Some more references

- Wessa P., E. van Stee (2008), The Xycoon Stock Market Game: Virtual Learning Environment of Real-Life Laboratory, Proceedings of the International Conference of Education, Research and Innovation (ICERI 2008), *submitted*
- Poelmans S., P. Wessa, K. Milis, E. Bloemen, C. Doom (2008), Usability and Acceptance of E-Learning in Statistics Education, based on the Compendium Platform, Proceedings of the International Conference of Education, Research and Innovation (ICERI 2008), *submitted*
- Poelmans S., P. Wessa, K. Milis, E. Bloemen, C. Doom (2008), The Impact of Gender on the Acceptance of Virtual Learning Environments, Proceedings of the International Conference of Education, Research and Innovation (ICERI 2008), *submitted*
- Wessa P. (2008), Let us free statistics of irreproducible research, Statistics Seminar at Simon Fraser University, Vancouver, Canada
- Wessa P. (2008), How to Research the Effectiveness of Constructivist Statistics Education? An Approach based on Reproducible Computing, Applied Statistics 2008, to be submitted to Advances in Methodology and Statistics
- Wessa P. (2008), A framework for statistical software development, maintenance, and publishing within an open-access business model, Computational Statistics
- Wessa P. (2008), Learning Statistics based on the Compendium and Reproducible Computing, Proceedings of the International Conference on Education and Information Technology (ICEIT'08), Berkeley, San Francisco, USA
- Wessa P. (2008), How Reproducible Research Leads to Non-Rote Learning Within a Socially Constructivist E-Learning Environment, Proceedings of the 7th European Conference on e-Learning (ECEL'08), Cyprus
- Wessa P. (2008), Measurement and Control of Statistics Learning Processes based on Constructivist Feedback and Reproducible Computing, Proceedings of the 3rd International Conference on Virtual Learning (ICVL '08), Romania
- Wessa P. (2008), A Compendium of Reproducible Research about Descriptive Statistics and Linear Regression, URL http://www.wessa.net/download/tutorial.pdf
- Wessa P. (2008), A Compendium of Reproducible Research about Time Series Analysis, URL http://www.wessa.net/download/tutorial1.pdf

All documents will be available at http://www.freestatistics.org/index.php?action=10 in the near future.