

# Reproducible Computational Research

## A Perspective

Manuel J. A. Eugster

Institut für Statistik  
Ludwig-Maximilians-Universität München

58. Biometrisches Kolloquium,  
Beuth Hochschule für Technik Berlin, 2012

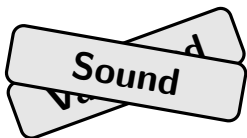


# Reproducible Computational Research A Perspective

Manuel J. A. Eugster

Institut für Statistik  
Ludwig-Maximilians-Universität München

58. Biometrisches Kolloquium,  
Beuth Hochschule für Technik Berlin, 2012



# Computational Research

## A Perspective

Manuel J. A. Eugster

Institut für Statistik  
Ludwig-Maximilians-Universität München

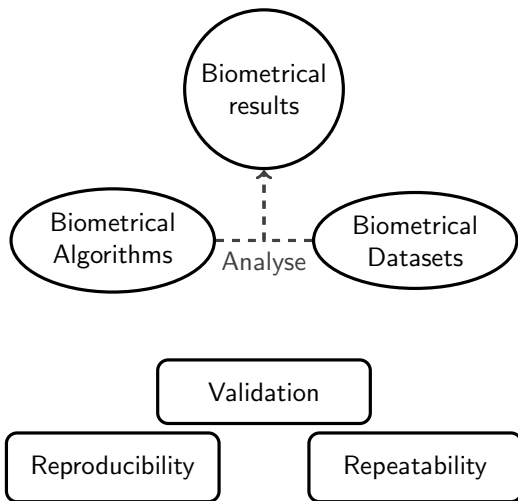
58. Biometrisches Kolloquium,  
Beuth Hochschule für Technik Berlin, 2012



Manuel J. A. Eugster

Institut für Statistik  
Ludwig-Maximilians-Universität München

58. Biometrisches Kolloquium,  
Beuth Hochschule für Technik Berlin, 2012



(\*) Based on Boulesteix (2011)

**Challenges of  
Reproducible and/or Repeatable  
Validation of Algorithms on Datasets**

# Challenges of Reproducible and/or Repeatable Validation of Algorithms on Datasets



**No Answers—Only Experiences  
and (a lot of) Open Questions!**



LUDWIG-  
MAXIMILIANS-  
UNIVERSITÄT  
MÜNCHEN

INSTITUT FÜR STATISTIK



Manuel J. A. Eugster & Torsten Hothorn & Friedrich Leisch

## Exploratory and Inferential Analysis of Benchmark Experiments

Technical Report Number 030, 2008  
Department of Statistics  
University of Munich

<http://www.stat.uni-muenchen.de>








## B. Components of the exemplar benchmark experiment


### B.1. Algorithms


The candidate algorithms are:


**linear discriminant analysis:** encoded as , orange; available through the function `lda` in package `MASS`.

**naive bayes classifier:** encoded as , yellow; available through the function `naiveBayes` in package `e1071`.

**$k$ -nearest neighbour classifier:** encoded as , purple; available through the function `knn` in package `class`. The hyperparameter  $k$  (the number of neighbours) is determined with cross-validation between 1 and  $\sqrt{n}$ ,  $n$  the number of observations.

**classification trees:** encoded as , red; available through the function `rpart` in package `rpart`. The fuled tree is pruned according to the 1-SE rule (e.g., Venables and Ripley, 2002; Hastie et al., 2001).

**support vector machines:** encoded as , blue; available through the function `svm` in package `e1071`. We use the  $C$ -classification machine, which has two hyperparameters  $\gamma$  (the cost of constraints violation) and  $c$  (the kernel parameter). Following Meyer et al. (2003) the best choices are determined with a grid search over the two-dimensional parameter space  $(\gamma, c)$ ,  $\gamma$  ranges from  $2^{-5}$  to  $2^{12}$  and  $c$  from  $2^{-10}$  to  $2^5$ .

**neural networks:** encoded as , green; available through the function `nnet` in package `nnet`. The hyperparameter is the number of hidden units. The best value is searched with cross-validation between 1 and  $\log(n)$ ,  $n$  the number of observations (following Meyer et al., 2003).

### B.2. Data sets

The benchmark survey is made up of 21 binary classification problems originated from the UCI Machine Learning repository (Asuncion and Newman, 2007):

| Problem      | #Attributes |            | #Samples |            | Class distribution (%) |
|--------------|-------------|------------|----------|------------|------------------------|
|              | nominal     | continuous | complete | incomplete |                        |
| promotergene | A           | 57         | 106      |            | 50.00/50.00            |
| hepatitis    | B           | 13         | 80       | 75         | 20.65/79.35            |
| Sonar        | C           | 60         | 208      |            | 52.27/47.73            |

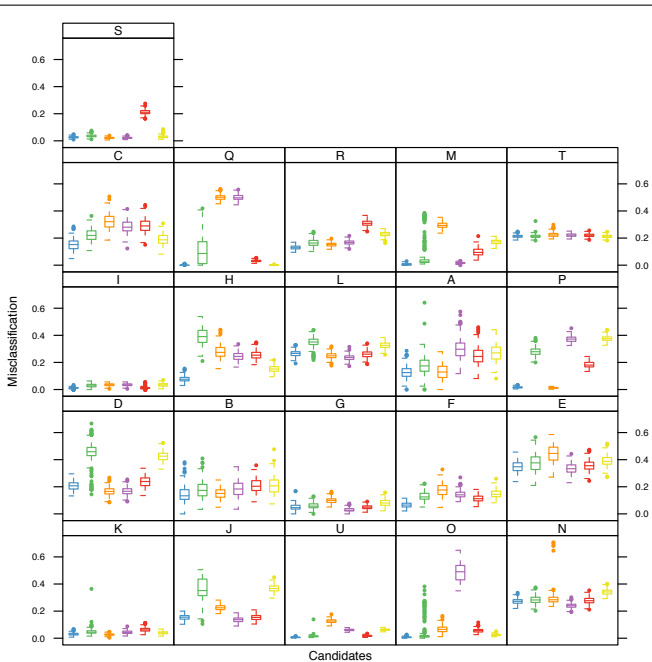


Figure 7: Grouped box plots (Trellis display): the raw performance measures of the candidate

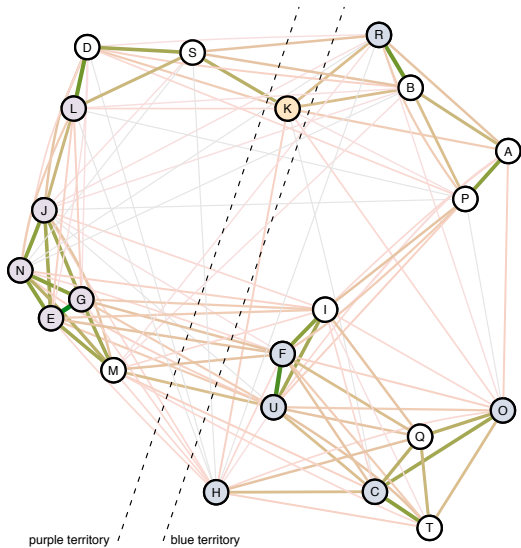


Figure 11: Benchmark survey graph: another representation of the distance matrix. The first 15 distance levels are shown, the color and the with of the edges represent them. Vertices are filled according to the winner algorithm if there is a unique one.

### 4.3. Inference

**Mixed effects model.** With linear mixed effects models, the assumptions on the random effects are  $b_i \sim N(0, \sigma_1^2)$ ,  $b_{ik} \sim N(0, \sigma_2^2)$  and  $\epsilon_{ijk} \sim N(0, \sigma^2)$ . Therefore, we estimate two parameters  $\sigma_1^2$  and  $\sigma_2^2$  for the effect of the data sets and the sampling within the data sets, respectively. Additionally,  $1 + (K - 1) + (M - 1) + (K - 1)(M - 1)$  fixed effects parameters are estimated.

*Example (cont.).* In case of our exemplar benchmark survey these are 3 + 126 parameters, see Appendix A.2 for the parameters and a model summary. The global test with ANOVA and the F-test rejects the null hypothesis that all algorithms have the same performance on all data sets. Using Tukey contrasts we test pairwise differences and calculate simultaneous confidence intervals. Figure 12 shows the 95% family-wise confidence intervals. The only Non-significant difference is

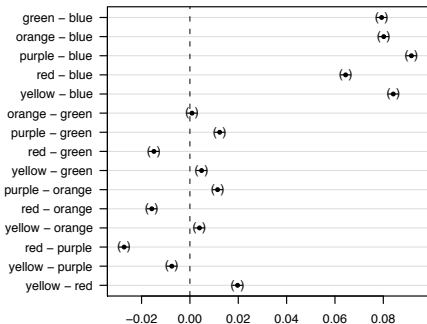


Figure 12: Simultaneous 95% confidence intervals for multiple comparisons of means using Tukey contrast based on the mixed effects model of the example experiment.

between **orange** and **green**). An interesting aspect appears, **blue** is a lot better than all other algorithms. We establish the algorithm order

$$\text{blue} < \text{red} \approx \text{orange} \approx \text{green} < \text{yellow} < \text{purple}.$$

All computations are performed using R (R Development Core Team, 2008), the corresponding R functions are part of an R package for the analysis of benchmark experiments which is currently under development and will be released on CRAN in due course. Preliminary versions of the functions and all data used in this article are available from <http://www.statistik.lmu.de/~eugster/>.

## 2. Design of benchmark experiments

Following Hothorn et al. (2005), we set up a benchmark experiment according to their *real world situation*. Given is a data set  $\mathcal{L} = \{z_1, \dots, z_m\}$ . We draw  $B$  learning samples using some resampling method, e.g. sampling with replacement (bootstrapping):

$$\begin{aligned}\mathcal{L}^1 &= \{z_1^1, \dots, z_n^1\} \\ &\vdots \\ \mathcal{L}^B &= \{z_1^B, \dots, z_n^B\}\end{aligned}$$

Another possibility is cross-validation. Furthermore we assume that there are  $K > 1$  candidate algorithms  $a_k$  ( $k = 1, \dots, K$ ) available for the solution of the underlying problem. For each algorithm  $a_k$  the function  $a_k(\cdot | \mathcal{L}^b)$  is the fitted model based on the sample  $\mathcal{L}^b$ . This function itself has a distribution  $\mathcal{A}_k$  as it is a random variable depending on  $\mathcal{L}^b$ :

$$a_k(\cdot | \mathcal{L}^b) \sim \mathcal{A}_k(\mathcal{L}), k = 1, \dots, K$$

The performance of the candidate algorithm  $a_k$  when provided with the training data  $\mathcal{L}^b$  is measured by a scalar function  $p$ :

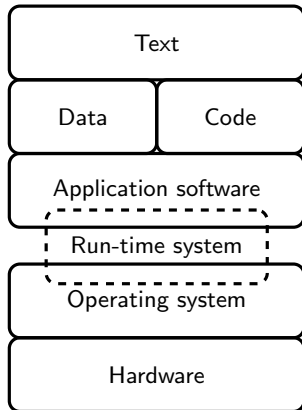
$$p_{kb} = p(a_k, \mathcal{L}^b) \sim \mathcal{P}_k = \mathcal{P}_k(\mathcal{L})$$

The  $p_{kb}$  are samples drawn from the distribution  $\mathcal{P}_k(\mathcal{L})$  of the performance measure of the algorithm  $k$  on the data set  $\mathcal{L}$ .

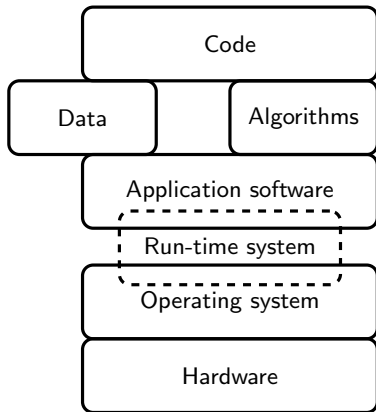
In this paper, we illustrate the analysis of benchmark experiments by means of supervised learning problems. The observations  $z$  are of the form  $z = (y, x)$  where  $y$  denotes the response variable and  $x$  describes a vector of input variables. The aim of this learning task is to construct a learner  $\hat{y} = a_k(x | \mathcal{L}^b)$  which, based on the input variables, provides us with information about the unknown response. The discrepancy between the true response  $y$  and the predicted response  $\hat{y}$  for one observation  $z$  is measured by a scalar loss function  $L(y, \hat{y})$ . The above introduced performance measure  $p$  is in this case defined by some functional  $\mu$  of the distribution of the loss function:

$$p_{kb} = p(a_k, \mathcal{L}^b) = \mu(L(y, a_k(x | \mathcal{L}^b))) \sim \mathcal{P}_k(\mathcal{L})$$

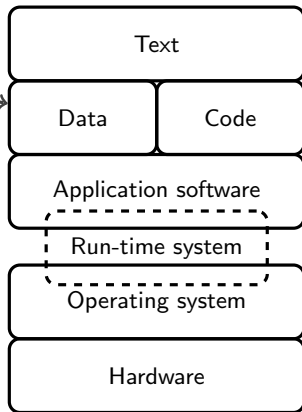
## Analysis/Manuscript



## Benchmark experiment

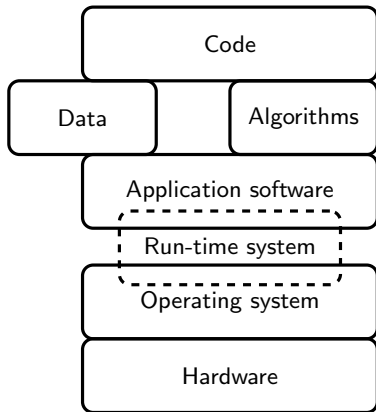


## Analysis/Manuscript



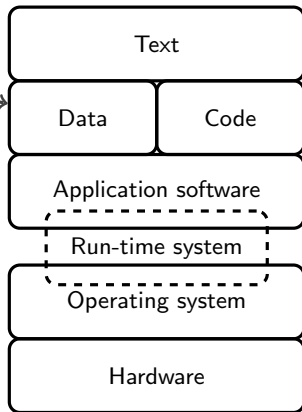
result of

## Benchmark experiment



- self-made parallelization
- nine days running time

## Analysis/Manuscript

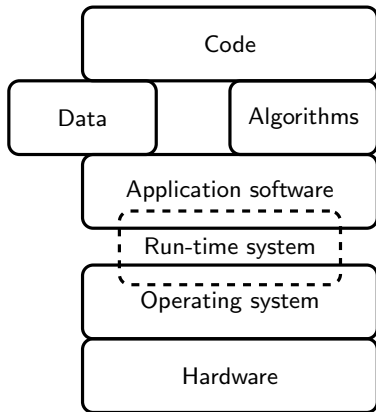


- published R package
- Sweave

result of



## Benchmark experiment

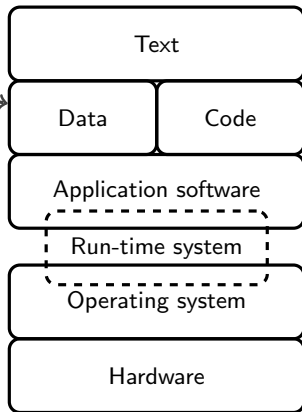


- self-made parallelization
- nine days running time

**interesting**

## Analysis/Manuscript

result of



- published R package
- Sweave

**easy**

# Challenges

## **Ince et al. (2012):**

*"The vagaries of hardware, software and natural language will always ensure that exact reproducibility remains uncertain, but withholding code increases the chances that efforts to reproduce results will fail."*

## Local (for me) reproducible validation:

- Clear separation between benchmark experiment and analysis/manuscript.
- Clear separation between setup and execution:
  - Setup:** data preparation, algorithm definition, experiment design, resampling strategy, etc.
  - Execution:** parallel computation, cloud computing, etc.

## Global (for others) reproducible validation:

- Publish data sets and source code:
  - ▶ Data set repositories
  - ▶ Licence strategies for data and code:
    - Open data and open source initiatives
    - Reproducible Research Standard by Stodden (2009) for Code (GNU GPL or BSD), Media (CC BY), Data (Science Commons Database Protocol)
- Use already published data sets:  
e.g., UCI Machine Learning Repository (Asuncion and Newman, 2007), ArrayExpress (Parkinson et al., 2010)
- Use “standardized” benchmarking/validation software:  
e.g., benchmark (Eugster, 2012), caret (Kuhn, 2012), mlr (Bischl, 2012)

## **Repeatable validation:**

- See validation/benchmarking as “real” experiments!
- Follow the rules of experimental designs: “data sets as patients and algorithms as treatments”

Google scholar Hedonic prices and the demand for clean air Meine Zitate

Suche [Erweiterte Scholar-Suche](#)

Web-Suche  Seiten auf Deutsch

In Artikeln suchen mit Zitaten aus: [Harrison: Hedonic housing prices and the demand for clean air](#)

Scholar Jederzeit Zitate einschließen E-Mail-Aktuell erstellen Ergebnisse 1 - 10 von ungefähr 791. (0,14 Sek.)

[ZITATION] **Classification and regression trees** [\[PDF\] von uk.y.edu](#)

[L. Breiman](#) - 1984 - Chapman & Hall/CRC

[Zitiert durch: 18829](#) - [Ähnliche Artikel](#) - [Bibliothekssuche](#) - [Alle 13 Versionen](#)

[BUCH] **Regression diagnostics**

[DA Belsley](#); E Kuh... - 1980 - Wiley Online Library

Copyright © 1980, 2004 by John Wiley & Sons, Inc. All rights reserved. Published by John Wiley & Sons, Inc., Hoboken, New Jersey. Published simultaneously in Canada. ... No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any ...

[Zitiert durch: 7181](#) - [Ähnliche Artikel](#) - [Bibliothekssuche](#) - [Alle 16 Versionen](#)

[ZITATION] **Econometric analysis of panel data**

[BH Baltagi](#) - 2005 - Wiley

[Zitiert durch: 7073](#) - [Ähnliche Artikel](#) - [Bibliothekssuche](#) - [Alle 16 Versionen](#)

[BUCH] **Advances in kernel methods: support vector learning** [\[PDF\] von tugraz.at](#)

[CJC Burges](#) - 1999 - books.google.com

The Support Vector Machine is a powerful new learning algorithm for solving a variety of learning and function estimation problems, such as pattern recognition, regression estimation, and operator inversion. The impetus for this collection was a workshop on ...

[Zitiert durch: 2988](#) - [Ähnliche Artikel](#) - [Bibliothekssuche](#) - [Alle 13 Versionen](#)

[BUCH] **Discriminant analysis and statistical pattern recognition**

[GJ McLachlan](#)... - 1992 - Wiley Online Library

This book has been electronically reproduced from digital information stored at John Wiley & Sons, Inc. We are pleased that the use of this new technology will enable us to keep works

## Repeatable validation:

- See validation/benchmarking as “real” experiments!
- Follow the rules of experimental designs: “data sets as patients and algorithms as treatments”
- ? **Future:** Validation/Benchmarking platform for randomized experiments (e.g., random assignment of data sets within a problem domain, of candidate algorithms, etc.)

# References I

- A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. URL <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- Bernd Bischl. mlr: Machine learning in R, 2012. URL <http://mlr.r-forge.r-project.org>. R package.
- Anne-Laure Boulesteix. Editorial special issue: Validation in bioinformatics and molecular medicine. *Briefings in Bioinformatics*, 12(3):187–188, 2011. doi: 10.1093/bib/bbr027. URL <http://bib.oxfordjournals.org/content/12/3/187.short>.
- Manuel J. A. Eugster. benchmark: Benchmark experiment toolbox, 2012. URL <http://cran.r-project.org/package=benchmark>. R package.
- Manuel J. A. Eugster, Torsten Hothorn, and Friedrich Leisch. Exploratory and inferential analysis of benchmark experiments. Technical Report 30, Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, 2008. URL <http://epub.ub.uni-muenchen.de/4134/>.
- Darrel C. Ince, Leslie Hatton, and John Graham-Cumming. The case for open computer programs. *Nature*, 482: 485–488, 2012. doi: 10.1038/nature10836. URL <http://www.nature.com/nature/journal/v482/n7386/full/nature10836.html>.
- Max Kuhn. caret, 2012. URL <http://cran.r-project.org/package=caret>. R package.
- Parkinson et al. Arrayexpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucl. Acids Res.*, DOI: 10.1093/nar/gkq1040. Pubmed ID 21071405, 2010. URL <http://www.ebi.ac.uk/arrayexpress>.
- Soeren Sonnenburg. mldata.org, 2012. URL <http://mldata.org/>.
- 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/ERRROLSI03092009.pdf>.

## Images:

**page 6:** Image from <http://esignsofcancer.blogspot.com/2011/09/warning-signs-of-cancer-in.html>.