

Four Papers on Contemporary Software Design Strategies for Statistical Methodologists

Vincent Carey and Dianne Cook

Software design impacts much of statistical analysis, and as technology changes, dramatically so in recent years, it is exciting to learn how statistical software is adapting and changing. This leads to the collection of papers published here, written by John Chambers, Duncan Temple Lang, Michael Lawrence, Martin Morgan, Yihui Xie, Heike Hofmann and Xiaoyue Cheng.

John Chambers has been at the forefront of advances in computing for data analysis since the 1970s, and his contributions have been recognized through the 1998 Association for Computing Machinery Software Systems Award. The award statement noted that the S system "forever altered how people analyze, visualize, and manipulate data" and remarked on the role of Chambers' "insight, taste, and effort" in establishing S as an "elegant, widely accepted, and enduring software system". Chambers' contribution to this symposium includes historical background and focuses on the joint roles of object-oriented and functional programming disciplines in fostering effective and extensible data analysis environments.

Duncan Temple Lang is the creator of the Omegahat project [2], and a long-standing member of R core. He has provided transformative energies and software tools for data-analytic computing, emphasizing the opportunities for interoperabilities among diverse components [4]. Temple Lang's contribution to this collection addresses the role of emerging compilation techniques in the evolution of R. His paper centers on the LLVM (formerly "Low-Level Virtual Machine") compiler infrastructure as a basis for transformation of R programs to highly performant and retargetable modules. This work is demonstrated in two packages available through the Omegahat project. The first is *Rllvm*, which connects R and LLVM infrastructures. The second is *RllvmCompile*, which builds on top of *Rllvm* to transform selected R idioms to LLVM-based intermediate representations, and ultimately to highly optimized platform-targeted modules. The strategy is illustrated in a number of practical examples.

Michael Lawrence and Martin Morgan are long-standing core members of the Bioconductor project. Their paper discusses software design for analysis and visualization of genome-scale data, as practiced at Bioconductor[1]. We are in a biological revolution with rapid advances are occurring in our understanding of how living organisms function. For the statistics community R provides the golden standard for analysis: open source, easy to contribute, hot off the press methods, high-level language, and good data plots. To analyze genomic data in R requires data structures and algorithms that accommodate its massive size and contextual structure. Additions to the Bioconductor suite that better facilitate genomic data analysis are described conceptually and illustrated with examples using currently available R packages.

Yihui Xie and Xiaoyue Cheng are major contributors to the Rstudio system, providing highly accessible and extensible methods for interactive graphics and literate statistical computing [7]. Heike Hofmann is Professor of Statistics at Iowa State University, and is a widely recognized pioneer in visual inference with high-dimensional data [6, 3]. The paper by Xie, Hofmann and Cheng describes creating interactive graphics with mutable objects in R. Although R has presented some especially important new ways to make static plots of data it was several steps back from XLispStat [5] for interactive graphics. Rudimentary R graphics are designed from an ancient pen on paper model, which does not easily enable the event loop hooks that are necessary to make plots interactive and dynamic. Several new packages built on new infrastructures in R enabling mutable objects, objects that can be changed inside of functions. How these mutable objects are used to form the pipeline creating multiple linked interactive graphics in the new *cranvvas* package is described by Xie, Hofmann and Cheng.

This collection of papers on statistical computing and visualization methods is partly a product of in-

vited session 267 of the 2012 Joint Statistical Meetings of the ASA, entitled “Contemporary Software Design Strategies for Statistical Methodologists”. Subsequently, related symposia on dynamic programming languages for analysis of large data have been sponsored by NSF: see <https://www.cs.purdue.edu/homes/jv/events/PBD13>, and <http://www.ws13.dynali.org/talks.html>

References

- [1] GENTLEMAN, R., CAREY, V., BATES, D., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., HORNIK, K., HOTHORN, T., HUBER, W., IACUS, S., IRIZARRY, R., LEISCH, F., LI, C., MAECHLER, M., ROSSINI, A., SAWITZKI, G., SMITH, C., SMYTH, G., TIERNEY, L., YANG, J., AND ZHANG, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5, 10 (Jan 2004), R80.
- [2] LANG, D. T. The omegahat environment: New possibilities for statistical computing. *Journal of Computational and Graphical Statistics* 9, 3 (2000), 423–451.
- [3] MAJUMDER, M., HOFMANN, H., AND COOK, D. Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association* 108 (2013), 942–956.
- [4] NOLAN, D., AND LANG, D. T. *XML and Web Technologies for Data Sciences with R*. Springer, 2013.
- [5] TIERNEY, L. *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. Wiley, 1990.
- [6] UNWIN, A., THEUS, M., AND HOFMANN, H. *Graphics of Large Datasets*. Springer, 2006.
- [7] XIE, Y. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, 2013.