

**Raymond D. Aller, MD, and Hal Weiner**

### **What R and Python programming languages bring to the table**

*The following article was contributed by James Harrison, MD, PhD, associate professor and director, Biomedical Informatics Division, University of Virginia, Charlottesville. It is based loosely on part of his CAP Informatics Committee webinar, "Beyond the LIS: open source big data tools for laboratory operational analytics and visualization."*



Dr. Harrison

Broad-based interest in analyzing large and complex data sets has led to the development of analytic software tools that are available at low cost or free of charge. The two most widely used open-source big data analysis tools—R and Python—are proving their mettle in the clinical laboratory.<sup>1-5</sup>

The R and Python programming languages are applicable to a variety of analytic tasks, ranging from standard statistics for small data sets to sophisticated machine learning and predictive modeling. Laboratorians interested in learning how to use either toolkit can take advantage of online tutorials, manuals, videos, and courses; printed manuals; and hands-on workshops and courses. Many of the offerings are available free of charge.

R and Python take a textual programmatic approach to analysis, similar to the commercial software tools SAS, Matlab, and Mathematica. This approach has several advantages, not the least of which is that it is very expressive, allowing users to precisely describe and run analyses and visualizations that might require complex and idiosyncratic graphical user interfaces. Furthermore, once users understand the basic language, they need install and learn only the analysis libraries relevant to them. This decreases the complexity and size of these tools as compared with a comprehensive graphical tool offering similar overall capabilities. R and Python are also broadly available across computing platforms. Therefore, analytic code can be developed and executed in common computing environments and shared between such environments.

While R is a capable analytics tool, the remainder of this article focuses on Python because it is a general-purpose programming language with strong data analytics and the most widely used language for introductory computer science in the United States. As such, laboratorians are likely to find Python useful for many computing tasks, as well as data analysis.

Python is relatively easy to read and write. It typically is not difficult for Python coders to figure out the intent of Python programs they wrote a year previously, even without extensive documentation. Although it was not originally designed for data analysis, Python's popularity and ease of use have led its fans to create and share high-quality statistics, analytics, visualizations, and machine learning libraries of Python code. For these reasons, the language has become widely used in the financial industry, physics, astronomy research, and other analytics-heavy areas. Bank of America, the Los Alamos National Laboratory, Facebook, and Google use Python liberally. Similar Python development has also occurred in other areas of computing, broadening the utility of the language beyond data analysis. For example, several excellent Web frameworks are written in Python, and these frameworks can be used with Python's data-analysis libraries to create Web applications completely in Python that are

accessible to any laboratory user on the network.

Python can be a good alternative to Excel or other spreadsheets commonly used for data analysis in clinical laboratories, particularly for large and complex data sets requiring fast computation, data that are disorganized or messy and require exploration and cleanup, and data that would benefit from sophisticated graphics to aid analysis. Furthermore, Python programs are easy to save, search for by text content, and reuse to process additional data, create new visualizations, or serve as the basis for new work. Python too may be useful for specialized tasks that are not directly supported by general-purpose data-analysis software. For example, open-source Python code libraries are available for processing bioinformatics data; processing and analyzing medical images; processing medical terminology, including SNOMED CT and ICD; parsing HL7 messages; and displaying control charts.

Python's ease of use means that it's relatively painless for users to learn enough of the language to carry out basic spreadsheet tasks. For example, it takes only a few minutes to learn to import a csv or Excel file into Python and calculate basic statistics for one or more columns of numbers. Gaining facility with the basic language to allow simple program design requires more practice, such as several weeks (part-time) of working through example Python code or problem sets from online learning modules or books. Once a budding analyst learns the basics, however, gaining additional expertise becomes less about the programming language than about understanding the topic that is covered by a particular code library—for example, statistics or machine learning.

Most successful Python “data programmers” start with small projects that allow them to create useful code quickly and then expand their knowledge progressively as opportunities arise. In the clinical laboratory, opportunities for using Python abound.

1. Colubri A, Silver T, Fradet T, et al. Transforming clinical data into actionable prognosis models: Machine-learning framework and field-deployable app to predict outcome of Ebola patients. *PLoS Negl Trop Dis*. 2016;10(3):e0004549.
2. Luo Y, Szlovits P, Dighe AS, et al. Using machine learning to predict laboratory test results. *Am J Clin Pathol*. 2016;145(6):778–788.
3. Bach E, Holmes DT. Reqscan: An open source solution for laboratory requisition scanning, archiving and retrieval. *J Pathol Inform*. 2015;6:3.
4. Holmes DT. cp-R, an interface the R programming language for clinical laboratory method comparisons. *Clin Biochem*. 2015;48(3):192–195.
5. Dickerson JA, Schmeling M, Hoofnagle AN, et al. Design and implementation of software for automated quality control and data analysis for a complex LC/MS/MS assay for urine opiates and metabolites. *Clin Chim Acta*. 2013;415:290–294.

## **Philips and PathAI partner on artificial intelligence offerings**

Royal Philips and the artificial intelligence technology company PathAI have teamed up to develop solutions that improve the diagnosis of cancer and other diseases.

“The partnership aims to build deep learning applications in computational pathology, enabling this form of artificial intelligence to be applied to massive pathology data sets to better inform diagnostic and treatment decisions,” according to a press release from Philips. The collaboration will initially focus on developing applications to automatically detect and quantify cancerous lesions in breast cancer tissue.

Philips has been using deep learning—an algorithmic technique that allows computers to analyze vast amounts of data, automatically detect patterns, and make predictions—in its radiology products.

[\*\*Philips\*\*](#), 800-453-6860

## **Hc1.com and 4medica announce collaboration**

Hc1.com, developer of a health care relationship management platform, has partnered with 4medica, a provider of cloud-based data management and clinical data exchange that gives clinicians a unified, real-time view of patient information across disparate care locations.

Under the partnership, 4medica will market and resell Hc1.com’s Lab4 Outreach Analytics solution. The product unlocks siloed lab data to monitor trends in volume, turnaround time, and exceptions, and it allows lab managers to compare business trends across referring physician networks. 4medica will also expedite the integration of LIS data into the Hc1 Healthcare Relationship Cloud, where it is automatically organized into real-time provider profiles powering dashboards that immediately report on issues and opportunities.

[\*\*Hc1.com\*\*](#), 317-219-4722

## **Technidata releases new generation of middleware**

Technidata has introduced the second generation of its TDBactiLink middleware solution and launched a dedicated website, [www.microbiology-middleware.com](http://www.microbiology-middleware.com), to detail the advantages of the middleware for microbiology and bacteriology laboratories.

The product receives electronic test requests from a lab information system, manages receipt of the sample, and resends the results after they have undergone clinical or technical review.

Among the features that have been added to the second generation of <sup>TD</sup>BactiLink are HL7 2.5 LIS communication for orders and results, additional reports for epidemiology, a new worklist with workflow filters, and smart clinical review adapted to bacteriology.

[\*\*Technidata\*\*](#), 770-888-4027

[hr]

*Dr. Aller teaches informatics in the Department of Pathology, University of Southern California, Los Angeles. He can be reached at [raller@usc.edu](mailto:raller@usc.edu). Hal Weiner is president of Weiner Consulting Services, LLC, Eugene, Ore. He can be reached at [hal@weinerconsulting.com](http://hal@weinerconsulting.com).*