In silico research in the era of cloud computing

Joel T Dudley & Atul J Butte

Snapshots of computer systems that are stored and shared 'in the cloud' could make computational analyses more reproducible.

Scientific findings must be reproducible for them to be formally accepted by colleagues, practitioners, policy makers and the lay public. Although it was once thought that computers would improve reproducibility because they yield repeatable results given the same set of inputs, most software tools do not provide mechanisms to package a computational analysis such that it can be easily shared and reproduced. This inability to easily exchange the computational analyses behind published results hinders reproducibility across scientific disciplines[1].

Several innovative solutions to this problem have been proposed. Most efforts have focused on software tools that aim to standardize the creation, representation and sharing of computational 'workflows'[2] that tie several software tools together into a single analysis. These workflow tools provide a way to represent discrete computational tasks (e.g., processing an input data file) as computational modules that can be connected into workflows by linking the output of one module with the input of another (e.g., the output from the input-data-processing module connects to the input of a data normalization module)[3]. Several of these tools allow researchers to build complex computational workflows through drag-and-drop visual interfaces and to share standardized representations of workflows. Examples include GenePattern[4] for analyzing genomic data, the Trident Scientific Workflow Workbench for oceanography (http://www.microsoft.com/mscorp/tc/trident.mspx), Taverna[5] for generalized scientific computation and web-based resources such as myExperiment[6] for sharing workflows. Recently, a software solution has been described[7] that embeds access to computational systems directly into digital representations of scientific papers.

Existing software applications have not become established solutions to the problem of computational reproducibility. This is not because of any technical shortcomings of the software because, in fact, many programs are technically well designed. Rather, the failures are a consequence of human nature and the realities of data-driven science, including the following: first, efforts are not rewarded by the current academic research and funding environment[8,9]; second, commercial software vendors tend to protect their markets through proprietary formats and interfaces[10]; third, investigators naturally tend to want to own and control their research tools; fourth, even the most generalized software will not be able to meet the needs of every researcher in a field; and finally, the need to derive and publish results as quickly as possible precludes the often slower standards-based development path[9].

The consequence is that nonstandardized, research computational pipelines continue to be developed, especially as new types of molecular measurements generate quantities of data that most standardized software packages cannot handle[11]. We can thus conclude that any effort to establish standard protocols that require investigators to adopt a particular software system, creating a real or perceived threat of 'lock in' to software from a single laboratory or commercial vendor, will be met with disregard or even resistance and will likely fail to drastically change the current behavior of researchers. Given these realities, we propose capturing and exchanging computational pipelines using complete digital representations of the entire computing environment needed to execute the pipeline.

Whole system snapshot exchange

In this approach, which we call whole system snapshot exchange (WSSE), the computer system(s) used by researchers to produce experimental results are copied in their entirety, including the operating system, application software and databases, into a single digital image that can be exchanged with other researchers. Using WSSE, researchers would be able to obtain precise replicas of a computational system used to produce the published results and have the ability to restore this system to the precise state of the system when the experimental results were generated. Even subtle variances caused by discrepancies between specific versions of software or programming languages are avoided by the WSSE approach. In principle, the input data used to produce published results could be exchanged along with the pipeline.

Readers may be quick to realize that WSSE could involve the exchange of data files that might range in size from tens of gigabytes to several terabytes or more. Even with the fastest internet connections, it is not feasible to share such large files easily and efficiently. Thus, it is not our intention that WSSE operate desktop-to-desktop, but rather we propose that data and computational pipelines

Joel T. Dudley and Atul J. Butte are in the Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, USA, and at the Lucile Packard Children's Hospital, Palo Alto, California, USA.
e-mail: abutte@stanford.edu

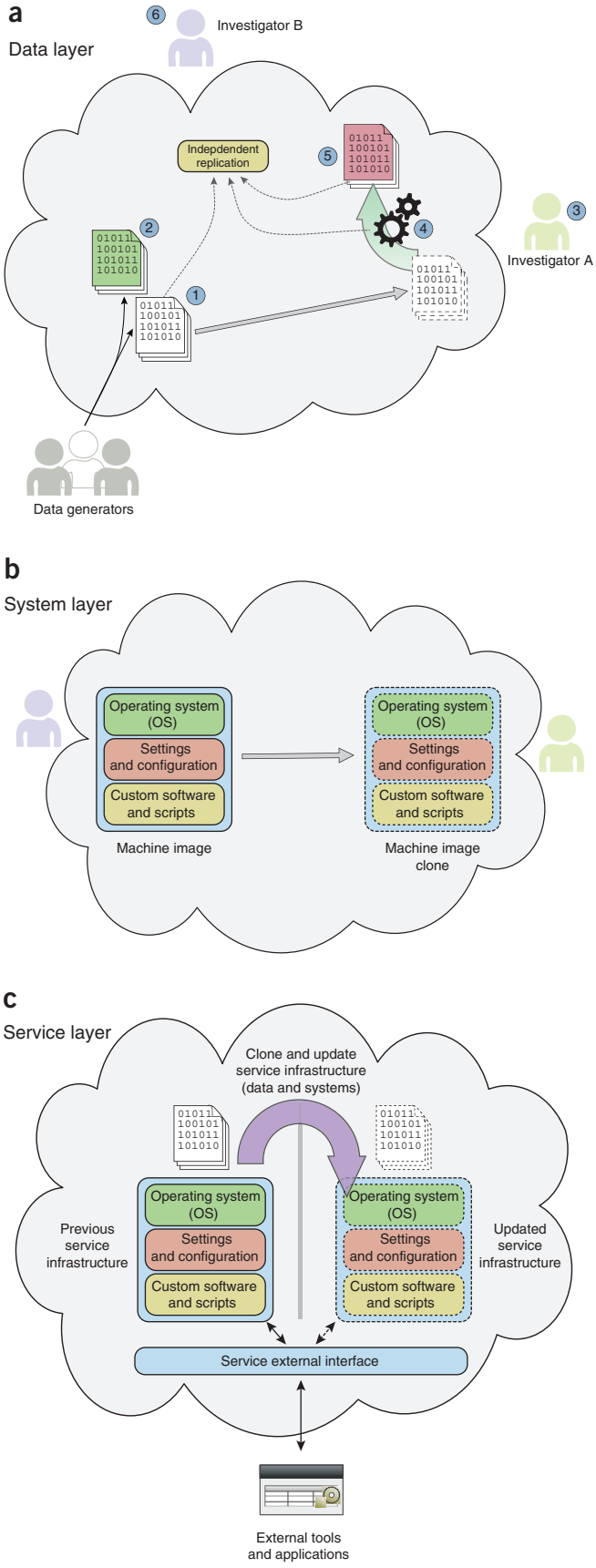

Figure 1 Layers of reproducible computing in the cloud. (**a**) Data layer. Generators of large scientific data sets can publish their data to the cloud (1) and substantial updates to these data sets can exist in parallel without loss or modification of the previous data set (2). Primary investigators can clone entire data sets within the cloud (3) and apply custom scripts or software computations (4) to derive published results (5). An independent investigator can obtain digital replicates of the original primary data set, software and published results within the cloud to replicate a published analysis and compare it with published results (6). (**b**) System layer. Investigators can set up and conduct scientific computations using cloud-based virtual machine images that contain all of the software, configurations and scripts necessary to execute the analysis. The customized machine image can be copied and shared with other investigators within the cloud for replicate analyses. (**c**) Service layer. Instead of replacing an old version of a scientific computing service with a new version of the software, a service that is virtualized in the cloud can be easily replicated and the old version made available alongside the new. Requests made by applications through the external service interface could incorporate a version parameter. This could enable published results that used older versions of the service to be evaluated for reproducibility.

will be exchanged exclusively using cloud computing, defined here as "computing in which dynamically scalable and virtualized resources are provided as a service over the Internet"[12].

In cloud computing, computation and data are 'virtualized', meaning that software and data are not tied to physical computing resources, such as a specific server with hard drives plugged into it. Instead, cloud-computing infrastructures are comprised of large and often geographically disparate clusters of computing hardware that are made to appear as a single, homogeneous computational environment[13]. The virtualization of data in the cloud makes it possible to move or copy 'snapshots' of large data sets from point to point within the cloud at high transfer rates, without the need to associate particular machines or storage drives at either the source or destination of the data transfer. Some cloud providers offer technologies that enable the creation of large databases that can persist as omnipresent resources that can be accessed by any computation in the cloud.

Concerns have been voiced[14] that scientific computing in the cloud could make results less reproducible. One concern is that cloud computing will be a computing 'black box' that obfuscates details needed to accurately interpret the results of computational analyses. Although this is an important concern, we argue that cloud computing could actually

make scientific computing more transparent by making it practical to share entire computational infrastructures so that they can be scrutinized. Without cloud computing, in contrast, sharing research software and data typically requires that they be modified and repackaged for distribution, and whoever receives the software must possess the necessary computing infrastructure.

Cloud-based support for reproducibility

Cloud computing could support reproducibility in several ways, which correspond to different 'layers' for accessing computing resources (**Fig. 1**).

At the data layer, cloud computing enables data sets to be easily stored and shared virtually—that is, without necessarily copying it to another computer. Most importantly, when a data set has been used to derive published results, it can be copied and archived in the cloud. This would be facilitated if large public data repositories were regularly archived into the cloud. One cloud-computing vendor, Amazon Web Services, has already taken the initiative to archive many such public data sets. Their extensive online catalog (http://aws.amazon.com/publicdatasets/) contains large public data sets from various domains of science, including astronomy, biology, chemistry and climatology.

At the system layer, cloud computing allows snapshots of complete computer systems to be exchanged, as proposed in WSSE. This addresses observations that computer systems can substantially confound the reproducibility of analyses[15].

A higher-level layer of scientific computation in the cloud is the service layer, in which computational services are exposed to external applications through some form of programming interface[16]. Examples include the Entrez Utilities from NCBI (http://eutils.ncbi.nlm.nih.gov/) that provide access to data and computational resources from NCBI's repertoire of bioinformatics applications and databases. From the standpoint of reproducibility, there are problems introduced by such a service-oriented approach to scientific computing: the underlying application supporting the computational service may be altered significantly without any apparent change to the public-facing service interface. If the applications and infrastructure supporting such services were migrated to a cloud-computing environment, the application service providers could run and maintain replicate instances of their entire application to maintain access to previous versions without the need to duplicate the hardware infrastructure underlying the service. Alternatively, the component data and systems underlying a computational service could be archived as 'images' in the cloud in a nonactive state, which would provide an economical means of preservation for the service provider while maintaining an efficient route of access to previous versions of a computational service for investigators. This may make it more feasible for authors to cite specific versions of data, systems and services in publications describing their results, and have a means to easily direct colleagues to these resources for reproducibility and reuse.

Reproducibility through preservation

One often-overlooked barrier to reproducibility is the loss or abandonment of grant-funded databases and other computational resources[17]. Therefore, an advantage of cloud computing is that virtual machines and cloud-based data storage can offer a means to sustain bioinformatics projects after funding cuts, project termination or abandonment. Although it may be easy to demonstrate the need to make biomedical databases available after their original purpose has been fulfilled, from a practical standpoint, funding the maintenance of these databases has become a contentious issue for funding agencies, who must balance maintenance with declining budgets and the need to fund new investigators and initiatives. For example, the *Arabidopsis* Information Resource (TAIR), used by thousands of researchers each day, recently fell into disarray after its funding was terminated after 10 years of support from the National Science Foundation[18]. Instead of turning off these servers and losing access to these resources, virtual machine images of the server could have been created and distributed across the Internet and executed on-demand through cloud computing, ensuring that investments in bioinformatics resources such as TAIR could be used for years after discontinuation.

The costs associated with preserving data in a cloud-computing environment are relatively low, making it feasible for research groups and institutions to preserve data through extended gaps in funding, or to fulfill commitments to collaborators or stakeholders well beyond the defunding of the primary database infrastructure. To illustrate, the current rate for storage space in Amazon's cloud service is $0.10 per gigabyte of storage per month, meaning that even a large 1-terabyte data set could be maintained for ~$100 per month. As the per-gigabyte cost of storage is expected to decrease with time, it is likely that a 1-terabyte biomedical database could be preserved in accessible form in the cloud for more than 10 years at a total cost well below $10,000. The cost of this active preservation could be written into the budgets of all proposals for creation and renewal of biomedical databases to address the problem of preservation and access to data beyond proposed funding periods. Moreover, having a working virtual machine server along with open source code could enable distributed teams of investigators to informally support these disbanded projects.

Toward a cloud-based scientific computing commons

As nearly every scientific discipline is becoming data-driven, one of the most enticing benefits of cloud computing is the means to aggregate scientific data sets efficiently and economically. The aggregation and centralization of public scientific data in the cloud, which offers a unified, location-independent platform for data and computation, are steps toward establishing a shared 'virtual commons' for reproducible scientific computing. In this common computing environment, it would be possible to develop and implement cloud-based scientific computational analyses that retrieve public data from centralized, cloud-based master catalogs. This computational analysis could then be shared and distributed within the cloud as part of a standardized system image. Using the system image, the computational analysis could be reproducibly executed because the same source code would be executing within the same system environment, drawing the same data from the master data catalog.

Although the cloud-computing technology required to facilitate straightforward approaches to reproducible computing, such as WSSE, is now widely available, there are a number of measures that could be taken to help and encourage the utilization of cloud-based resources by the broader scientific community. Journals and funding agencies could support this vision by mandating that more types of measurements be deposited into approved cloud-based data catalogs, and funding could be provided for existing consortia to move their valuable data from isolated data silos into centralized cloud-based catalogs. As cloud computing becomes commonplace, we expect technologies to emerge that allow one to move data between cloud-computing providers, to prevent 'lock in'[19]. For example, the open-source Eucalyptus platform (http://www.eucalyptus.com/) enables one to move cloud-based resources out of the Amazon Web Services commercial platform into a private cloud infrastructure.

Table 1 Features of reproducible scientific computing in the cloud

	Traditional challenges	Cloud-computing solutions
Data sharing	• Large data sets difficult to share over standard internet connections; can require substantial technical resources to obtain and store. • Public data sets change frequently. Difficult to archive and share entire data repositories used for analyses.	• Large data sets can be stored as 'omnipresent' resources in the cloud. Easily copied and accessed directly from any point in the cloud. • 'Snapshots' of large public data sets can be rapidly copied, archived and referenced.
Software and applications	• Reproducibility of results often requires replication of the precise software environment (that is, operating system, software and configuration settings) under which the original analysis was conducted. Specific versions of software or programming-language interpreters often required for reproducibility. • Analyses typically conducted by several types of software or scripts executed in a precise sequence across one or several systems as part of an analysis pipeline. Only the individual programs or scripts are usually provided with published results. Substantial technical resources typically required to recreate the pipeline used in the original analysis. • Standard software packages cannot serve all the needs of a scientific domain. Investigators develop nonstandard software and computational pipelines to facilitate computational analysis exceeding the capabilities of common tools.	• Computer systems are virtualized in the cloud, allowing them to be replicated wholesale without concern for the underlying hardware. Snapshots of a fully configured system or group of systems used in analysis can be rapidly archived as digital machine images. System machine images can be copied and shared with others in the cloud, allowing reconstitution of the precise system configuration used for the original analysis. • System images can be preconfigured with common and customized software and tools in a standardized fashion to facilitate common tasks in a scientific domain (e.g., assembly of genome sequences from DNA sequencer data). Preconfigured images can be shared as public resources to promote reproducibility and follow-up studies.
System and technical	• Substantial computational resources might be required to replicate an analysis. Original computational analyses requiring several hundred processors to complete becoming more common. Reproducibility limited to those with requisite computational resources. • Substantial technical support often required to reproduce a computational analysis and to replicate the software and system configuration required by the analysis. Prevents reproducibility by nontechnical investigators lacking substantial IT support.	• Cloud-based computational resources can be scaled up in a dynamic fashion to provide necessary computational resources. Investigators can create large computational clusters on demand and disperse upon analysis completion. • Complete digital representations of a computational pipeline can be shared as machine images along with deployment scripts that can be executed by nontechnical users to reconstitute a complete computational pipeline.
Access and preservation	• Grant-funded software and data repositories often disappear from the public domain after funding is discontinued or the maintainers abandon the project. Leads to loss of access by dependent users and loss of public investment into the resource.	• Software, code and data from grant-funded projects can be archived and provided as publicly accessible resources in the cloud. Economies of scale in the cloud allow for active preservation of grant-funded resources for many years past funding for nominal cost. • Cloud-computing providers already show a willingness to host public scientific data sets at no cost.

Some have already called for the creation of a publicly funded cloud-computing infrastructure[13]; however, we suggest that it would be most prudent to focus funds and effort into building cloud-based scientific computing software on top of existing cloud infrastructures in such a way that they are portable across vendors. Given the economic and competitive pressures facing commercial cloud-computing vendors, as well as their substantial technical and capital resources, it is not clear that a publicly funded cloud infrastructure would offer any significant technical or economic advantages over commercial clouds. Furthermore, we suggest that, as has been observed in many aspects of computing, standards enabling cross-vendor interoperability are likely to emerge as cloud computing becomes more prominent and additional vendors increase competition in the market. Groups such as the Open Cloud Consortium (http://opencloudconsortium.org/) have already been formed to explore this issue. Notably, many of the popular tools for managing and interacting with cloud-based infrastructures, such as Chef (http://www.opscode.com/chef), are designed to be independent of the specifics of any one cloud-computing vendor's infrastructure. Peer-to-peer data distribution technologies, such as BitTorrent which is already being leveraged in the bioinformatics community[20], could be used to store and distribute large biological data sets beyond the confines of any single cloud-computing provider.

In the domain of biomedicine, several efforts toward the development and distribution of cloud-based tools, systems and other resources have recently emerged[21]. Several groups have created standardized machine images with software tools and configuration settings optimized for biomedical research. The most ambitious among these so far is the J. Craig Venter Institute (Rockville, MD, USA) Cloud Bio-Linux project (http://www.jcvi.org/cms/research/projects/jcvi-cloud-biolinux/), which aims to provide a comprehensive and coherent system capable of a broad range of bioinformatics functionality. These preconfigured machine images are made publicly available to the research community, providing individual investigators with a standardized, tuned platform for cloud-based computational analyses. Other efforts have gone into the creation of more comprehensive multipart systems for facilitating biocomputing in the cloud[22,23]. A significant effort in this area is the Galaxy project[24], which provides a platform for large-scale genomic analysis.

Although we suggest that the WSSE approach is a pragmatic and substantial first step toward enabling reproducible scientific computing in the cloud, we acknowledge that it does not address all aspects hindering reproducibility. Foremost, software licensing constraints could prevent the use of WSSE. Many commercial software applications are restricted by licenses that constrain the number of software application instances that can run simultaneously or that restrict execution to a number of processors. These constraints might prevent an investigator from sharing some or all of the system or software components used to produce published results, limiting the effectiveness of WSSE. We also recognize that there is a need for continued research and development into reproducibility enhancements at levels above WSSE. For example, problems in data organization, systematic provenance tracking, standardization and annotation are not solved by WSSE.

Nonetheless, we suggest that WSSE can serve to initiate a value-driven movement toward general reproducible scientific data and computing into the cloud (**Table 1**), and that solutions to problems of reproducibility not solved by WSSE—many of which already exist in some form—will follow WSSE into the cloud toward the realization of a comprehensive platform for reproducible computing enabled by the technical innovations of cloud computing.

ACKNOWLEDGMENTS

J.T.D. is supported by the National Library of Medicine Biomedical Informatics Training Grant (T15 LM007033) and A.J.B. is supported by the National Institute for General Medical Sciences (R01 GM079719). We thank D. Singh for valuable discussion regarding cloud-computing technology.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Gentleman, R. *Stat. Appl. Genet. Mol. Biol.* **4**, Article 2 (2005).
2. Gil, Y. *et al. Computer* **40**, 24–32 (2007).
3. Barker, A. & van Hemert, J. in *Parallel Processing and Applied Mathematics* (eds. Wyrzykowski, R., Dongarra, J., Karczewski, K. & Wasniewski, J.) 746–753 (Springer; 2008).
4. Reich, M. *et al. Nat. Genet.* **38**, 500–501 (2006).
5. Hull, D. *et al. Nucleic Acids Res.* **34**, W729–W732 (2006).
6. De Roure, D., Goble, C. & Stevens, R. *Future Gener. Comput. Syst.* **25**, 561–567 (2009).
7. Mesirov, J.P. *Science* **327**, 415–416 (2010).
8. Ball, C.A., Sherlock, G. & Brazma, A. *Nat. Biotechnol.* **22**, 1179–1183 (2004).
9. Brooksbank, C. & Quackenbush, J. *OMICS* **10**, 94–99 (2006).
10. Wiley, H.S. & Michaels, G.S. *Nat. Biotechnol.* **22**, 1037–1038 (2004).
11. Lynch, C. *Nature* **455**, 28–29 (2008).
12. Bateman, A. & Wood, M. *Bioinformatics* **25**, 1475 (2009).
13. Nelson, M.R. *Science* **324**, 1656–1657 (2009).
14. Osterweil, L.J., Clarke, L.A. & Ellison, A.M. *Science* **325**, 1622 (2009).
15. Schwab, M., Karrenbach, M. & Claerbout, J. *Comput. Sci. Eng.* **2**, 61–67 (2000).
16. Wagener, J., Spjuth, O., Willighagen, E.L. & Wikberg, J.E. *BMC Bioinformatics* **10**, 279 (2009).
17. Merali, Z. & Giles, J. *Nature* **435**, 1010–1011 (2005).
18. Anonymous. *Nature* **462**, 252 (2009).
19. Gu, Y. & Grossman, R.L. *Eng. Sci.* **367**, 2429–2445 (2009).
20. Langille, M.G. & Eisen, J.A. *PLoS ONE* **5**, e10071 (2010).
21. Stein, L.D. *Genome Biol.* **11**, 207 (2010).
22. Langmead, B., Schatz, M.C., Lin, J., Pop, M. & Salzberg, S.L. *Genome Biol.* **10**, R134 (2009).
23. Schatz, M.C. *Bioinformatics* **25**, 1363–1369 (2009).
24. Giardine, B. *et al. Genome Res.* **15**, 1451–1455 (2005).