Data integration strategies for predictive analytics in precision medicine

Lewis J Frey*,¹

¹Biomedical Informatics Center, Medical University of South Carolina, 135 Cannon Street/Suite 405K, MSC 200, Charleston, SC 29425, USA

*Author for correspondence: frey@musc.edu

With the rapid growth of health-related data including genomic, proteomic, imaging and clinical, the arduous task of data integration can be overwhelmed by the complexity of the environment including data size and diversity. This report examines the role of data integration strategies for big data predictive analytics in precision medicine research. Infrastructure-as-code methodologies will be discussed as a means of integrating and managing data. This includes a discussion on how and when these strategies can be used to lower barriers and address issues of consistency and interoperability within medical research environments. The goal is to support translational research and enable healthcare organizations to integrate and utilize infrastructure to accelerate the adoption of precision medicine.

First draft submitted: 11 April 2018; Accepted for publication: 17 July 2018; Published online: 5 November 2018

Keywords: common data models • data integration • infrastructure • infrastructure-as-code • interoperability • multiomics • precision medicine • predictive analytics • sociotechnical • virtual machines

The onslaught of large genomic and imaging datasets is already upon us and researchers are examining ways of coping with the acquisition, integration, storage, distribution and analysis demands [1]. This includes the use of cloud computing [2]. With whole-genome sequencing of compressed data being 25 gigabytes per participant, the storage requirements accumulate quickly. The same is true for whole-slide imaging. For example, as part of the Moonshot initiative [3] the National Cancer Institute has requested proposals for the construction of three-dimensional tumor atlases. For a small study with 200 participants and tumor volumes of 5 mm each, whole slide images of 5 micron slices at 40 $\times$, with each slide taking up 15 uncompressed gigabytes of storage [4], will consume 3 petabytes (each petabyte is 1 million gigabytes) of storage. Add to that the adoption of deep learning predictive analytics on pathology images [5,6]. Moreover the lack of interoperability between clinical, molecular and imaging data systems is holding back the acceleration of translational and precision medicine research [7–10]. There are promising advancements around interoperability; [11–13] data sharing [14] and findable, accessible, interoperable and reusable principles [15], but true infrastructure-as-code [16–19] with rigor and reproducibility of results [20,21] that communicate seamlessly with electronic health record systems is still on the horizon. The lack of this integration will continue to drag on translational and precision medicine research for the near future.

In the 1980s, the computer science community faced a similar issue of hardware providers building systems that could not communicate with each other at the byte ordering level with some systems ordering data and messages transmitted from the most significant bit (e.g., Sun Microsystems SPARC) while others from the least significant bit (e.g., Intel x86) [22]. It was termed the big-endian and little-endian war [23] in reference to Gulliver's travels [24] and the war between the Blefuscudians and Lilliputians over which end to crack an egg: the big end or the little end. Seldom discussed today because these differences were resolved by having an interoperable network order that allowed both approaches to communicate with each other through transforming data and messages into and out of the agreed upon order. Virtualization provided another level of interoperability in which virtual machines could be run on either big-endian or little-endian architectures, supporting seamless transmission of data and messaging across both architectures. This review proposes that infrastructure-as-code solutions, using common data representations and virtualization, can be used to avoid the healthcare information technology research equivalent of an interoperability big-endian, little-endian war in precision medicine.

Figure 1. Overview of infrastructure-as-code used for data integration. Infrastructure-as-code solutions utilize methods to define coding instructions for provisioning and managing computers across multiple institutions, the cloud and fog. Infrastructure-as-code can be used to strengthen data integration strategies. This is particularly relevant to multiomic approaches in precision medicine predictive analytics. By employing infrastructure-as-code strategies to define computers, adopters have the leverage to deploy as many cloned or similar computers to optimize data analysis, costs and utilization. This includes deploying virtual machines at the place where data reside, to as many partners as appropriate and in or out of the cloud.

Infrastructure-as-code overview

Infrastructure-as-code is a method for formally defining coded instructions on how a set of computers should be provisioned and managed. Those instructions are executed so that the infrastructure envisioned in the code is realized across one or more computers. There are multiple applications that can be used as part of an infrastructure-as-code strategy to code and execute data integration and predictive analysis for precision medicine, some of these will be touched upon. A key component of the approach involves virtualization in which a prespecified computer is encoded as a virtual machine and a program (e.g., VMware [25], VirtualBox [26]) called a hypervisor on a target computer can be used to run the virtual machine. A hypervisor can be thought of as a ‘hyper supervisor’ that micromanages every action the virtual machine performs in the target environment. The hypervisor keeps the virtual machine from getting into trouble with the overarching management of the target environment while enabling it to execute its processes. The infrastructure that results is an encapsulated set of processes that can be built and tested locally and then distribute as clones to other users/sites to replicate the analysis. This is especially relevant for big data where it is more cost-effective to conduct analyses where the data reside [27].

Infrastructure-as-code methodology supports a combination of deliberate and emergent strategies [28] in which a virtualized infrastructure can be completely planned and encoded to achieve a desired goal or can emerge through adaptive and agile development in a dynamic interaction with the environment. Deployment of infrastructure can occur strategically using the right resources for a specific project, supporting in-house fog computing [29] as well cloud computing (See Figure 1) [2]. Fog computing can be thought of as a cloud close to the ground or local cloud [29] and will have advantages when locality is important and/or the system is required to be close to the data source. The cloud is typically a distributed data center service (e.g., Amazon Web Services) where you pay for what you use. Cloud computing has advantages for global analysis on longer timescales where the latency of data transmission is not of concern.

The ability to create virtual machines with a common data model and common code base lowers the barrier to interoperability. Machines and processes are interoperable because the distributed systems are identical to each other. The distribution of these machines means that one team can do the heavy lifting of code development to

create an infrastructure-as-code solution and then distribute it to other sites. The partner sites do not need to write the code, but instead only need to be able to execute scripts within the virtual machine. Several examples of infrastructure-as-code will be illustrated using clinical personalized pragmatic predictions of outcomes (C3PO), an open source big data platform [30,31]. The C3PO has been utilized in three very different healthcare research environments: Medical University of South Carolina (MUSC), Veterans Affairs Informatics and Computing Infrastructure (VINCI) environment and Christiana Care Health System in Delaware. Adoption of infrastructure-as-code solutions can address issues of interoperability and enable large scale data analysis for precision medicine. The power of these solutions comes from their ability to support generalizability, agility of deployment, flexibility of integration strategies and replicability of infrastructure.

Generalizability

The C3PO is a good example of infrastructure-as-code because it has been developed with interoperability in mind through using the common data model observational medical outcomes partnership (OMOP) [12] and it has been reused by generalizing it to new use cases. The C3PO was originally developed to examine a novel algorithm for clustering and predicting outcomes for patients with diabetes. Since, it was designed as infrastructure-as-code, we can reuse and expand the system to generalize it for the MUSC Transdisciplinary Collaborative Center to support precision medicine in minority men's health, specifically examining prostate cancer and the effects of chronic stress on patients experienced in rural and urban environments. Ansible [32], Kubernetes [33,34] and Docker [35] were used to stand up the multiple machines needed to run the infrastructure. Our team had direct access to the machines so difficulties with network configuration could be dealt with directly. The C3PO MUSC Transdisciplinary Collaborative Center system ingests clinical data from REDCap [36] for the project and integrates it into the OMOP model in its Spark/Hadoop framework. Since, C3PO was developed so it can generalize to other data types such as genomic and imaging, Spark/Hadoop frameworks [37–40] for genomic and imaging can be integrated in future versions of the system.

The beauty of the infrastructure-as-code solution though, is that it can generalize to hypervisor and bare metal systems that reside in other healthcare settings resulting in an interoperable infrastructure for running analyses. It is not limited to running inside a healthcare environment but can also be run in the cloud through initializing the machines in the cloud environment and letting the infrastructure-as-code configure them. Infrastructure-as-code solutions can generate sets of services and integrate them into the fabric of the environment.

Replicability of infrastructure

The strength of infrastructure-as-code is its ability to manage software systems across institutional boundaries while taking into consideration the particular needs in different institutions. An example of this concept is interacting with the VINCI system using components of the C3PO system. The VINCI environment has a vastly different set of needs compared with MUSC or Christiana Care. In large environments like VINCI that are resource rich with 24 million patients in their data warehouse and database systems that support thousands of researchers, the specifications for deploying infrastructure-as-code are far different than other environments. For security purposes, they do not permit code that can generate virtual machines in their environment, instead they provide virtual environments that researchers connect with and are monitored through data access protocols. Thus, the Veterans Affairs (VA) cannot be thought of as a tabula rasa and must be connected to along approved lines of communication. They are better viewed as a service provider with which your resources interact.

Within our VINCI research environment, we encoded similar services inside the VINCI environment that were replicated in identical cloned virtual machines in a local environment to reproduce the C3PO system in terms of application interfaces and available command sequences. This approach enabled us to troubleshoot on local machines issues that occurred within the VA system without direct access to their data systems (not permitted). When commands were returning errors for our service calls, we could replicate the error on our cloned and interoperable system that existed outside the VA and diagnose the issue. The approach of developing in a local environment with full access in order to replicate conditions encountered in a global environment with restricted access allowed for the construction of more robust system.

Agility of deployment

Infrastructure-as-code can speed deployment for cross institutional/researcher data integration and sharing.

For example, to connect eight children's hospitals across a national network, a research consortium [41] encoded a secure platform within VirtualBox and VMW are virtual machines that supported secure data communication protocols using globus [11,42,43]. We found that deployment could be done in a short period of time, within hours the virtual machine, database and secure connection could be established at the external site. Additionally, there was a reduction in time and errors for loading and curating the data load across all the sites. The challenge that composed 99% of the effort for the project was building social acceptance and political readiness at the hospitals involved in the project. The acceptance was accomplished through the trust network of the research consortium and required clinical champions at each hospital to spearhead the effort. Political readiness included meeting with Chief Information Officers and/or technology leadership at each institution. Only after the social and political concerns were addressed, did each site move forward. The national debate mirrors these issues on a larger scale where the replication of science involves aspects of political consensus in conjunctions with the goals of science to increase knowledge and insight along with pressures to publish novel findings [44]. The use of virtualized globus-based technologies deployed in children's hospitals across the USA demonstrated the proof of principle of multisite deployment within complex sociotechnical systems using an integrated security model for transferring data to support research to improve care for critically ill or injured children [11].

The ease of deployment, the global synchronization of a network data model and the reduction in data curation processes supported the concept of using infrastructure-as-code virtual machines to reduce barriers to interoperability, but the technology for a robust infrastructure-as-code strategy was still nascent when it was deployed in 2011. The wave of container deployments since Docker publicly released its platform in 2013 has signaled the arrival of technology that can speed an agile deployment and consistency of interoperable systems through cloned virtualization. As technology advances so to can the infrastructure-as-code repository. The C3PO is an example of this (i.e., the first version used scripted code and web services). The second version was deployed via Puppet [45] and Vagrant [46] scripts. The latest versions use Ansible, Kubernetes and Docker.

Flexibility of integration strategies

Infrastructure-as-code should be examined as part of the strategy to achieve the goals of efficiency and agility by flexibly managing the machines provisioned in a cloud to run analysis on clinical and molecular data [47]. It can be used for deployment and reproducible adoption of technology. The use of infrastructure-as-code in healthcare enables a flexible expansion of virtualization combined with public and private cloud and fog computing [29,48] encompassing services that spread out within and across institutional boundaries. Pipelines and code can be documented and deployed within cloud-based environments and connected with the data residing there to perform reproducible analysis. Pipelines can scale with the size of the data through on demand generation of data analysis nodes for completing analysis within a specified timeframe. Such an approach will enable the integration of large imaging and genomics data with clinical data resources in cloud environments. A number of healthcare environments are using cloud solutions to scale analysis and to improve the consistency of their data integration processes. Two examples are: University of California San Diego health system's multiyear migration to move their data to Epic in the cloud and Beth Israel Deaconess Medical Center's adoption of Amazon's cloud for storing 7 petabytes of their electronic health record data [49]. The hope of these institutions and others is that the move to the cloud will result in the creation of operational efficiencies and agility.

One of the issues in integrating data is establishing the data sharing agreements to transfer data from one healthcare environment to another. The wrangling involved can add time to the project and the complexity grows with the number of independent data sources included in the analysis. The preanalytical variation in omics datasets is another issue that requires a robust metadata documentation process so that data can be consistently analyzed when they are shared [50]. The variability of the measurement modalities in multiomics projects have become complex to the point of necessitating the development of standardized taxonomies to manage the measurements applied to biological systems and ultimately precision medicine [51]. A question that repeatedly comes to mind is why not perform the same analysis without transferring the data to a central site (i.e., distributed analysis). In our data rich future, transferring data to each research project's centralized repository will become cost prohibitive assuming the exponential growth in size given larger numbers of patients records including genomic and other large data files (e.g., proteomic, imaging) [27]. In a small pilot with Christiana Care, we examined a tabula rasa example to demonstrate distributed analysis using infrastructure-as-code.

Specifically, we operationalized the scenario where we deployed to an external research site (Christiana Care) that provided a tabula rasa server with no direct access to the data, instead staff assistance was provided. To

lowered the barrier of resources that a site is required to have in order to do the analysis, we assumed a minimal infrastructure that can be deployed into small windows servers. The Christiana Care team was not familiar with the C3PO code base. They execute scripts to automate the system configuration, including loading Christian Care data into an OMOP format. For security purposes, the Christian Care team isolated the virtual system, but allowed analysis within it behind their firewall. The C3PO infrastructure-as-code environment included components of the Spark/Hadoop [52–54] system along with RStudio analysis all within a vagrant and puppet scripting framework published publicly on a github repository. The identical cloned C3PO systems, one at Christiana Care and another at MUSC [55] proved invaluable for troubleshooting the environment. For the analysis, they provided a person who could access the data and run queries based on specific inclusion and exclusion criteria to generated file with a specified file format. The approach allowed us to replicate analysis originally performed at MUSC and produced an integration of the summary results without the need to centralize the data or share patient data outside the institution.

Limitations

Mechanisms are needed for maintenance and coordination of components of infrastructure-as-code systems, along with documenting the infrastructure and managing deployment. Questions of security linger with some approaches (e.g., Docker requires administrative privileges on machines), but the difficulties of improving security are not insurmountable and active development of the community has resulted in systems with more security features through the use of nested containers and scoped privileges [56].

Conclusion

Precision medicine will involve combinations of data on a scale not previously experienced in clinical care. Whether it is the explosion of genomic data with whole genomic sequences or imaging of multiple whole slide pathology images the level of storage and computation will put pressure on our existing infrastructure. Turning to the cloud as an alternative to data centers that are ill prepared to manage the volume of data is a viable alternative. The expertise to manage the cloud infrastructure is still needed and will mean a refocusing of information technology personnel to learn the core skills needed to maintain cloud infrastructure nodes that provide secure access to clinical and molecular data. Understanding the complexities of clinical data does not go away when the bits and bytes now reside on a distributed data center. Complexity remains and there is a need to manage that growing complexity, with ever growing volumes of data. The infrastructure-as-code methodology is proposed as a means of managing this growing complexity of the technology infrastructure along with information management. The ultimate goal is achieving more efficient and interoperable systems while enabling innovation in a quickly changing medical space. The simplification of communicating about medical data through the use of a common data model in conjunction with virtual machines for running analysis on the data is one area where the complexity can be reduced. The use of agile development pipelines of infrastructure-as-code in this emerging paradigm will help us reach our goals.

Future perspective

In working toward a future where healthcare organizations integrate and utilize infrastructure to accelerate translational research and the adoption of precision medicine, we will need to create seamless interfaces where you will not know if you are on the cloud, in the fog, virtualized or running on bare metal. The infrastructure staff will be trained to manage the infrastructure-as-code repository and serve up the services needed for clinicians and researchers. The solution will support distributed analysis that increases privacy and security of patient records while enhancing the ability to advance discovery. The seamlessness and ease of use of the environment will foster collaborations among translational researchers and clinicians treating patients and the shift to personalized and precision medicine will flow from the availability of resources to analyze and visualize the data. The future looks bright for precision medicine, especially when the field realizes the power and accessibility of the infrastructure they need to achieve their vision is already here and waiting to be distributed.

Financial & competing interests disclosure

The work was supported in part by NIH grants (grant numbers 1R01GM108346-01, U54-MD010706, and U54-GM104941) and Health Equity and Rural Outreach Innovation Center grant (grant number CIN 13-418). The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

This review points in the direction of infrastructure-as-code that includes standardized models and virtualized systems encoded in such a way as to perform replicable analysis on the integrated data. In making programmatic decision about the type of solution that is needed, the first thing to consider is the type of organization and the sociotechnical environment you are deploying within and what are the barriers to effective interoperable systems.

Benefits of infrastructure-as-code

Generalizability

- Allows for building and reusing the infrastructure across multiple projects

Replicability

- Supports local and distributed testing of infrastructure to harden system.

Agility

- Enables fast deployment and evolution of the system in the environment.

Flexibility

- Provides the ability to choose among multiple integration strategies and architectures.

Infrastructure

- Supports secure services that are integral to visualize data in distributed environments.

Cloud & fog

- Distributed and local data centers combined with infrastructure management will be needed to handle the influx of precision medicine data.

The missed opportunities to make new discoveries using the plethora of data being generated in medical environments will only grow with ever larger datasets that are squirreled away without being integrated into translational systems. The barriers to accomplishing such integration are challenging but not insurmountable and with the impetus to move data to the cloud, there is an opportunity to seize the moment and ensure that such an undertaking involves integrating molecular and imaging data in the bargain. In tackling the issues of interoperability and usability of accessible data, the field will move the adoption of precision medicine forward by great strides.

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Stephens ZD, Lee SY, Faghri F *et al.* Big data: astronomical or genomics? *PLoS Biol.* 13(7), e1002195 (2015).
2. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. *Nat. Rev. Genet.* 19, 208–219 (2018).
3. National Cancer Institute Cancer Moonshot Blue Ribbon Panel report (2016) www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel
4. Chlipala E, Elin J, Eichhorn O *et al.* Archival and retrieval in digital pathology systems. Digital Pathology Association, WI, USA, 1–10 (2011).
5. Saltz J, Gupta R, Hou L *et al.* Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* 23(1), 181–193 (2018).
6. Bejnordi BE, Veta M, van Diest PJ *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 318(22), 2199–2210 (2017).
7. Frey LJ, Lenert L, Lopez-Campos G. EHR big data deep phenotyping: contribution of the IMIA genomic medicine working group. *Yearb. Med. Inform.* 9(1), 206–211 (2014).
8. Frey LJ, Bernstam EV, Denny JC. Precision medicine informatics. *J. Am. Med. Inform. Assoc.* 23(4), 668–670 (2016).
9. Pritchard DE, Moeckel F, Villa MS, Housman LT, McCarty CA, McLeod HL. Strategies for integrating personalized medicine into healthcare practice. *Per. Med.* 14(2), 141–152 (2017).
10. Denny JC, Van Driest SL, Wei W-Q, Roden DM. The influence of big (clinical) data and genomics on precision medicine and drug development. *Clin. Pharmacol. Ther.* 103(3), 409–418 (2018).
11. Frey LJ, Sward KA, Newth CJL *et al.* Virtualization of open-source secure web services to support data exchange in a pediatric critical care research network. *J. Am. Med. Inform. Assoc.* 22(6), 1271–1276 (2015).
- Provides a use case of virtualization reducing interoperability barriers among children's hospitals.
12. FitzHenry F, Brannen J, Denton JN *et al.* Transforming the National Department of Veterans Affairs Data Warehouse to the OMOP Common Data Model. *AMIA* 1471 (2015).
13. Rosenbloom ST, Carroll RJ, Warner JL, Matheny ME, Denny JC. Representing knowledge consistently across health systems. *Yearb. Med. Inform.* 26(01), 139–147 (2017).
14. Hripcsak G, Duke JD, Shah NH *et al.* Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud. Health Technol. Inform.* 216, 574–578 (2015).

- **Describes the efforts of the OHDSI Consortium and the approaches that they have developed to perform interoperable analysis in healthcare environments.**
- 15. Wilkinson MD, Dumontier M, Aalbersberg IJ *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 160018 (2016).
- 16. Vanbrabant B, Joosen W. Configuration management as a multicloud enabler. In: *Proceedings of the 2nd International Workshop on CrossCloud Systems*. ACM, NY, USA, 1:1– 1:3 (2014).
- 17. Scheuner J, Cito J, Leitner P, Gall H. Cloud WorkBench: benchmarking IaaS providers based on infrastructure-as-code. In: *Proceedings of the 24th International Conference on World Wide Web*. ACM, NY, USA, 239–242 (2015).
- 18. Cito J, Leitner P, Fritz T, Gall HC. The making of cloud applications: an empirical study on software development for the cloud. In: *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*. ACM, NY, USA, 393–403 (2015).
- 19. Bessani A, Brandt J, Bux M *et al.* BiobankCloud: a platform for the secure storage, sharing and processing of large biomedical datasets. In: *Biomedical Data Management and Graph Online Querying*. Springer International Publishing, Cham, Switzerland, 89–105 (2016).
- 20. Piccolo SR, Frampton MB. Tools and techniques for computational reproducibility. *Gigascience* 5(1), 30 (2016).
- 21. Tatlow PJ, Piccolo SR. A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. *Sci. Rep.* 6, 39259 (2016).
- 22. Tanenbaum AS, Van Steen M. Distributed systems: principles and paradigms. Prentice-Hall, NJ, USA (2007).
- 23. Cohen D. On Holy Wars and a Plea for Peace. *Computer* 14(10), 48–54 (1981).
- **A computer scientist's historical view of the big-endian, little-endian war.**
- 24. Swift J. Travels into several remote nations of the world four parts. By Lemuel Gulliver, First a Surgeon and then captain of several ships. Benjamin Motte, London. (1726).
- **A satirical classic highlighting the hubris of emphasizing petty differences.**
- 25. Waldspurger CA. Memory resource management in VMware ESX server. *Oper. Syst. Rev.* 36(SI), 181–194 (2002).
- 26. Watson J. VirtualBox: bits and bytes masquerading as machines. *Linux J.* 166, 1 (2008). <http://dl.acm.org/citation.cfm?id=1344209.1344210>
- 27. Stockham N, Wall DP. Open access economics in the big genomics era. In: *AMIA 2018 Informatics Summit*. Tatonetti N, Pathak J, McIntosh L (Eds.). American Medical Informatics Association, MD, USA (2018).
- 28. Mintzberg H, Waters JA. Of strategies, deliberate and emergent. *Strat. Mgmt. J.* 6(3), 257–272 (1985).
- 29. Bonomi F, Milito R, Zhu J, Addepalli S. Fog computing and its role in the internet of things. In: *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*. ACM, NY, USA, 13–16 (2012).
- 30. Frey L, Lenert L, Duvall S *et al.* Flexible machine learning (ML-flex) in the Veterans Affairs Clinical Personalized Predictions of Outcomes (Clinical3PO) System. In: *FLAIRS Conference*. Markov Z, Russell I (Eds.) 704–705 (2016).
- 31. Clinical3PO. Clinical3PO/platform. GitHub. <https://github.com/Clinical3PO/Platform>
- 32. Hall D. Ansible configuration management. Packt Publishing Ltd, Birmingham, UK (2013).
- 33. Burns B, Grant B, Oppenheimer D, Brewer E, Wilkes J. Borg, Omega, and Kubernetes. *Commun. ACM* 59(5), 50–57 (2016).
- 34. Brewer EA. Kubernetes and the path to cloud native. In: *Proceedings of the Sixth ACM Symposium on Cloud Computing*. ACM, NY, USA, 167–167 (2015).
- 35. Merkel D. Docker: Lightweight Linux containers for consistent development and deployment. *Linux J.* 2014(239) (2014).
- 36. Harris PA. Research electronic data capture (REDCap): planning, collecting and managing data for clinical and translational research. *BMC Bioinformatics* 13(12), A15 (2012).
- 37. McKenna A, Hanna M, Banks E *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20(9), 1297–1303 (2010).
- 38. Wang F, Aji A, Vo H. High performance spatial queries for spatial big data: from medical imaging to GIS. *SIGSPATIAL Special* 6(3), 11–18 (2015).
- 39. Mushtaq H, Al-Ars Z. Cluster-based apache spark implementation of the GATK DNA analysis pipeline. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, NY, USA, 1471–1477 (2015).
- 40. Fjukstad B, Bongo LA. A review of scalable bioinformatics pipelines. *Data Sci. Eng.* 2(3), 245–251 (2017).
- 41. Willson DF, Dean JM, Newth C *et al.* Collaborative pediatric critical care research network (CPCCRN). *Pediatr. Crit. Care Med.* 7(4), 301–307 (2006).
- 42. Foster I. Globus Toolkit Version 4: software for service-oriented systems. In: *Network and Parallel Computing*. Springer, Berlin Heidelberg, Heidelberg, Germany, 2–13 (2005).
- 43. Foster I. Globus online: accelerating and democratizing science through cloud-based services. *IEEE Internet Comput.* 15(3), 70–73 (2011).

44. Sarewitz D. Reproducibility will not cure what ails science. *Nature* 525(7568), 159 (2015).
- **A perspective piece focused on the complexity of reproducibility in politically, socially and technically charged environments.**
45. Loope J. Managing infrastructure with puppet: configuration management at scale. O'Reilly Media, Inc. CA, USA, (2011).
46. Hashimoto M. Vagrant: up and running: create and manage virtualized development environments. O'Reilly Media, Inc. CA, USA, (2011).
47. Madduri RK, Dave P, Sulakhe D, Lacinski L, Liu B, Foster IT. Experiences in building a next-generation sequencing analysis service using Galaxy, Globus Online and Amazon Web Service. In: *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*. ACM, NY, USA, 34:1–34:3 (2013).
48. Shi W, Cao J, Zhang Q, Li Y, Xu L. Edge computing: vision and challenges. *IEEE* 3(5), 637–646 (2016).
49. Davis J. UC Health goes live on shared, cloud-based Epic EHR. *Healthcare IT News*(2017). www.healthcareitnews.com/
50. Lee J-E, Kim Y-Y. Impact of preanalytical variations in blood-derived biospecimens on omics studies: toward precision biobanking? *OMICS* 21(9), 499–508 (2017).
51. Pirih N, Kunej T. Toward a taxonomy for multiomics science?: terminology development for whole genome study approaches by omics technology and hierarchy. *OMICS* 21(1), 1–16 (2017).
- **An article on the complexity of multiomics science that provides a clean taxonomy for organizing the research space.**
52. Borthakur D, Gray J, Sarma JS *et al.* Apache Hadoop goes realtime at facebook. In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*. ACM, NY, USA, 1071–1080 (2011).
53. Shvachko K, Kuang H, Radia S, Chansler R. The Hadoop distributed file system. In: *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*. NJ, USA, 1–10 (2010).
54. Shanahan JG, Dai L. Large-scale distributed data science using Apache Spark. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, NY, USA, 2323–2324 (2015).
55. Frey L, Mauldin P, Obeid J, Moran W, Weintraub W. Clinical personalized pragmatic predictions of outcomes (C3PO) protocols for data integration and analysis. In: *Sixth Biennial National IDeA Symposium of Biomedical Research Excellence (NISBRE)*. NIGMS, DC, USA, (2016).
56. Azab A. Enabling docker containers for high-performance and many-task computing. In: *Cloud Engineering (IC2E), 2017 IEEE International Conference*. IEEE, NJ, USA, 279–285 (2017).