BIO AC Data Working Group (Apr-Aug 2012)

Jonas S Almeida (chair), Brett Tyler (co-chair), Carol Brewer, Hopi Hoekstra, Gaetano Montelione, Jose Onuchic. Executive Secretary: Melissa Cragin.

Executive Summary

The data workgroup was charged with studying and discussing Data Science in the context of NSF's BIO initiatives. This is an area driven by extremely fast advances in Web Technologies such as Semantic Web and Cloud computing that have the potential to empower individual life science researchers with resources not available even to large organization just a few years ago. In contrast, the articulation of those resources with the wide availability of commodity/social computing infrastructure for data sharing and team science is much less widely understood and rarely informs policy making or calls for proposals in BIO. This tension is compounded by life sciences-specific issues, such as the variable granularity of data at different scales and uneven visibility/accessibility of data generated by BIO initiatives in the past. A number of recommendations are made with a focus on a) closer involvement of the research communities that generate and consume the data and b) opportunities to experiment with the novel computational infrastructures to handle real world biological data.

Sharing: interoperable linked data[1].

A large potential value exists for wide-scale sharing of biological data in an interoperable format. Sharing enables large-scale secondary analysis of the data that can reveal emergent features of the underlying systems[2]. It also enables a more innovative and transformative analysis of primary data, and invites an unprecedented level of participation in the scientific enterprise. This is a level of integration, and engagement, that scientific communities, and the public at large, have come to expect as a basic feature of an era where social computing infrastructure is a commodity resource. Standards for describing and annotating data and associated metadata are essential for capturing the biological implications in an interoperable format, but such standards are still widely unavailable or inadequate for a wide diversity of data types, including proteomics data, metabolomics data, phenotypic data, mass spectrometry data, biological image data, NMR-related data, ecological data, behavioral data.

Recommendation: Individual communities should be funded to define priorities and procedures for sustainable data archiving and sharing, and for establishing community-specific data standards that are required to make their data widely biologically interoperable as well as technically interoperable.

Infrastructure: integrative Web 3.0[3].

The combination of improved logistics of cloud computing and comprehensive linking enabled by the semantic web's Resource Description Framework (RDF) provide a **distributed**

---

[1] Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., et al. (2012). Toward interoperable bioscience data. Nature genetics, 44(2), 121–6. See also map at http://bit.ly/webofdata.

[2] Committee on a New Biology for the 21st Century, National Research Council of The National Academies (2009) "A New Biology For The 21st Century" National Academies Press, Washington, D.C.

[3] Hendler, J. (2009). Web 3.0 Emerging. Computer (Vol. 42, pp. 111-113). IEEE Computer Society. doi:10.1109/MC.2009.30. See also http://www.data.gov/semantic.

infrastructure for wide-scale data sharing and distributed analysis. This has the potential to efface the logical barriers between large centralized infrastructures and specialized "beyond the data deluge"[4] repositories tailored for data acquisition. It also could decouple infrastructure configuration from the inevitable, and desirable, antagonism between data generation and its annotation by further analysis, while enabling traceable provenance and reproducibility[5]. One barrier to realizing the potential of RDF is lack of familiarity with RDF among biology communities and resources for converting legacy data into RDF formats. Another is that cloud computing platforms are still maturing as a flexible, cost-effective interoperable means for storing, moving, sharing and analyzing data.

Recommendation: Resources should be provided to improve communication between information scientists and biology communities, and identify cost-effective, easy-to-use cloud solutions, by funding demonstration projects. These should explicitly encourage public-private partnerships that address long term solutions for archiving public biological data.

Broader Impacts: the rise of the social machines[6].

Data produced by BIO funded initiatives have an uneven record of availability for subsequent use (secondary analysis, which includes the contextualization of primary analysis). Hopeful examples include iPlant, recognized as an innovator in data science, but in contrast with enduring concerns with LTER data. The onset of NEON last year created an urgent need and a also fantastic opportunity to make the most of data intensive approaches[7] to such complex biological systems. The potential scientific, societal and economic impact of initiatives like NEON cannot be overstated and are critically associated with the long term availability and discoverability of the data generated. This is also a challenge that points to education as the ultimate interface for the achievements of BigData science: reaching beyond traditional Academic environments all the way to K-12 and citizen science through participatory (gamified?) initiatives. Education and outreach are areas where NSF BIO is particularly effective and respected. As a consequence, there are ample opportunities, resources and expertise to amplify the impact of societal intervention of this nature.

Recommendation: Consider adding data driven components to existing outreach and education programs - a particular strength of NSF BIO.

Funding models: promises and prizes.

It is reasonable to expect that scientific creativity will remain associated with diverse teams of researchers with reliable access to the necessary resources[8]. The sustainability of data resources beyond a project's funding period is therefore a growing concern.

Recommendation: Extend ongoing experimentation with funding models to the data generation components.

---

[4] Bell, G., Hey, T., & Szalay, A. (2009). Computer science. Beyond the data deluge. Science (New York, N.Y.), 323(5919), 1297–8. doi:10.1126/science.1170411

[5] Peng, R. D. (2011). Reproducible research in computational science. Science (New York, N.Y.), 334(6060), 1226–7.

[6] Berners-Lee, T., Hendler, J. (2010). From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. Artificial Intelligence, 174(2), 156–161.

[7] Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 24(2), 8–12.

[8] Heinze, T., Shapira, P., Rogers, J. D., & Senker, J. M. (2009). Organizational and institutional influences on creativity in scientific research. Research Policy, 38(4), 610–623.