Big Science in the Big Data Innovation Ecosystem

David Skinner LBNL and Amber Boehnlein SLAC

The work of designing, engineering, and operating the scientific instruments that allow us to observe and simulate the world in new ways is increasingly collaborative and data-driven. Big Science is one description for how large teams, devices, and data sets come together to produce scientific discovery and innovation. The trajectory toward team-based data-driven science is well documented by way of national strategy statements.[1,2] Big Science is also well known through its big machines with big outcomes[3]. Less-well communicated are the R&D agendas related to how the data from these instruments is acquired, analyzed, and shared. Here we connect the dots between these data-driven science instruments and the broader Big Data Innovation Ecosystem (BDIE).

How massive scientific research data figures into the broader ecosystem is a timely concern. Open Data memos and a general uptick in connected resources leads us to consider how the data flowing from Big Science instruments relates to emerging national-scale data frameworks. Below we identify areas for cross-sector partnership where Big Science has unique requirements and capabilities. The broader BDIE has much to offer Big Science as well.

Without defining what "Big Data" is, we can safely state that a deluge of scientific data presents challenges and opportunities that bear some overlap with broader data trends in the web and commercial contexts. Whereas data *volume* in content distribution scaling (Netflix, PB/year) and service scaling (Google, 450M users) are a well recognized aspect of the BDIE, challenges also arise where the *velocity* of a single data stream and the *complexity* of productive analytics are uniquely challenging. The "serial number zero" detectors and high-throughput techniques found in Big Science instrumentation require streaming data analytics and collaboration around large data sets at unique scales. R&D here informs next generation data science capabilities in the broader technology and workforce arenas.

We suggest there are three areas in which data-driven scientific instruments can, in the next five to ten years, broadly accelerate innovation. Facilities and frameworks for these three areas tackle tough problems that are currently unique to Big Science teams. Sooner than we think, these solutions will emerge as opportunities for technology transfer in manufacturing, engineering, education, and IT.

1) Streaming data analytics: Big Science detectors produce data streams at pace-setting bandwidths and modalities. We require both prompt analytics during operation as well as methodical re-analysis of data offline by large teams requiring advanced algorithms. Making decisions and discoveries from massive complex data streams is a Big Science specialty.

---

[1] http://science.energy.gov/about/

[2] http://www.nsf.gov/nsf/nsfpubs/straplan/mission.htm

[3] http://en.wikipedia.org/wiki/Nobel_Prize

2) High-throughput workflow engines: Big science accelerates the process of scientific discovery and innovation through radical efficiencies in throughput. Found under the hood of the Human Genome Project, Super Nova Factory, and Materials Genome Initiative are high-throughput analytic engines that automate and organize tasks once assigned to individual grad students. Shrinking time-to-knowledge via high-throughput analysis combined with provenance and reproducibility demands present unique challenges.

3) Scalable software for teams: Big Science teams are large and complex, making decisions and discoveries based on large, diverse, and real-time data sets. The software and APIs in Big Science are unique in their scale. Software and policies that incentivize large teams to make effective use of Big Data toward shared curated models may serve as models for tomorrow's IT.

Data-driven scientific instrumentation arrives at these challenges head-on and often in advance of the broader ecosystem. Advancing the state of the art in the above areas or a superset of them is best done in partnership with the full spectrum of stakeholders. In analogy to the shared computing facilities, which emerged from widespread need for scientific computing in the 1970's, today's widespread challenges with scientific data are best served by reliable, efficient, and advanced facilities for scientific data.

Not all data challenges are alike. Across science teams, domains, and agencies the type and schedule of needs to be met will vary widely. An investment in a facilities approach to these challenges that identifies core data grand challenges provides translational leverage between these disparate requirements. The R&D that yields solutions to these challenges also educates tomorrow's workforce at the cutting-edge of data science.

Big Science has a role in pioneering the frontiers of high-bandwidth streaming analysis, high-throughput workflows, and tools for distributed teams to collaboration. A scientific data facility, which addresses the above challenges, could directly attack grand challenges[4] in research, leverage best-of-breed Big Data methods from the commercial world, and help train the next generation data scientist workforce. Interconnection of data from Big Science resources through a facility dedicated to that task is appealing to many stakeholders.

Let us not underestimate the aim and the need for such a super-facility in data. Big Science instruments bring new capabilities and understanding to the world. Discovery and innovation are often found where intercomparison of data from observations and simulations becomes scalable and efficient. A scientific data super-facility approach that promotes composable coordination of existing user facilities and sharing of scientific knowledge at Big Data scales is within reach. Inasmuch as we have recognized the transformative scientific value of intensely concentrated computational power through shared user-facilities, we must now pioneer strategies for intensely concentrated bandwidth and analytics.

[4] National Research Council. Frontiers in Massive Data Analysis. Washington, DC: The National Academies Press, 2013.