

DMMReview

information infrastructure

Science Intelligence



Can a Business Intelligence Approach Enable “Smart” Science?

By Richard Hackathorn

When I first learned to use computers many years ago, I was forced to make a choice. Should I use FORTRAN for scientific processing? Or, should I use COBOL for business processing? The two were mutually exclusive. Back then, my choice was clear: I was into science; besides, COBOL was for wimps unable to program a DO loop.

Since then, we have come a long way in information processing! However, it is amazing how old fallacies died slowly.

This article argues that business and science are converging in terms of technologies, architectures and procedures for information technology. Both have the objective of understanding complex systems in uncertain environments by diverse stakeholders. Both business and science face challenges of scalability, diversity and complexity, along with an explosion of data volumes.

In the future, I believe that a unified information infrastructure can serve both communities, contributing to smarter business and to smarter science.

Defining Science Intelligence

To explore this argument, let's coin a new phrase – science intelligence (SI). SI is close to a mirror image of business intelligence (BI). I believe that the symmetry will lead us to realize that a side of one is the flip side of the other. Both can gain from the lessons learned by those on the other side. There is science to business, and there is business to science.

By SI, I am not implying some spy story where foreign agents steal secrets from a remote scientific laboratory. SI means conducting the activities of science in a more productive manner enabled by advanced informatics. SI is defined as the information infrastructure that enhances the decision making and collabora-

tion for a science community focused on a specific domain.

Over the past several decades, BI has become a successful and mature technology, now considered an essential component of enterprise systems. Without a single view of business reality, a corporation is flying blind. Without actionable business analytics, a corporation is flying dumb. Today, business executives accept both assertions.

Over the same period, SI has also become a successful and mature technology through innovative use of large-scale simulations, remote sensing and advanced visualizations. The face of science has changed from images of white-robed lab technicians with clipboards to cute golf carts bouncing over Martian rocks. The societal importance of science is underscored by daily headlines about global climate, genetic engineering, near-earth asteroids and energy conservation.

Sizing of Science Intelligence

One way to gauge the significance of BI and SI is to estimate the size of their respective information technology (IT) markets.

Sizing the BI market is easy because several analyst firms closely watch this market. According to IDC, the worldwide BI market grew in 2003 by 8.5 percent to \$13.4 billion.¹ They define the BI market as business analytics, which is composed of analytic applications, data warehousing tools, production BI platforms and technical data analysis. This does not include many traditional IT areas, such as networking, hardware, operating systems and transaction processing applications.

It is more difficult to estimate the size of the SI market. Research and development (R&D) in the U.S. is a large industry, comprising approximately 2.6 percent of the gross domestic product (GDP) or approximately \$284 billion in 2003.² Research

performed by industry accounts for 68 percent of this market. The traditional categories of basic research, applied research and development are spread as 19, 24 and 57 percent of the total. The National Science Foundation (NSF) plays a unique role in funding \$5.5 billion in basic research in universities across all fields of science and engineering (excluding the medical sciences). This is approximately 20 percent of all federally supported basic research.⁵

IT support in business is typically in the range of two to five percent of revenues. If we assume the same ratio for science, IT expenditures for R&D could be \$6 to \$30 billion, and this estimate is just within the U.S. More work is obviously needed to refine this sizing; however, the point is that worldwide markets for BI and SI are roughly comparable.

Examples of Science Intelligence

Let's consider an example of a large-scale SI project. The Sloan Digital Sky Survey (SDSS) is the "most ambitious astronomical survey project ever undertaken" involving 25 institutions and more than 200 scientists.⁴ The survey is mapping one-quarter of the entire sky, including more than 100 million objects plus distance measures to one million galaxies and quasars. As of late 2004, SDSS had published the fourth of six data releases (approximately 40TB of raw image data) and is on schedule for the final release in mid 2006. The astronomy data is available to scientists students, and the general public through the SkyServer portal.⁵

Based on the experience with the SkyServer portal, Jim Gray and others have noted several unique requirements for SI.⁶ In this project, database technology has proved to be a useful point of integration among a diverse community of users. Because the database brings the data into a single physical storage, all parties are motivated to agree on a single consistent view (schema) of that data, along with common requirements and documentation. Further, the database infrastructure manages the dissemination of data to various parties.

In recent years, large shared databases have become important resources for many areas of science, allowing hundreds of scientists to share and collaborate over the same data sets. For examples, see the sidebar on large shared scientific databases.

One exploding area that has embraced many aspects of SI is the life sci-

ences, particularly in pharmaceuticals, genetics and healthcare. Major BI vendors, such as IBM, SAS and Microsoft, have devoted considerable attention to leveraging their technology for specific problems in the life sciences (see sidebar on page 47). For instance, IBM offers DiscoveryLink, Discovery Warehouse and Clinical Genomics as tools to integrate diverse scientific information.^{7,8} SAS offers SAS Research Data Management as part of their SAS Scientific Discovery Solutions.^{9,10} Microsoft offers Digital Pharma, which supports drug discovery, drug development, manufacturing/supply chain and sales/marketing.¹¹

Each of these efforts has shown that current approaches and technology of BI are applicable to SI, guided by expertise within the specific science domain. Bill Rapp, IBM CTO for Healthcare and Life Sciences Solutions Development, stated, "We did not have to redo our BI tools for life sciences. We just used them on new types of data. The trick is to capture and format the data for optimal use by the existing tools."¹²

Vision for Science Intelligence

These efforts are evolving toward an integrated information infrastructure that supports a consistent view of reality (CVR), whether for business or for science. Figure 1 illustrates this infrastructure, with the CVR as the central component supporting multiple groups and generating a set of products and services for various consumers.

All information is managed within a common framework with consistent semantic definitions, regardless of whether the data is *materialized* (physically stored), *virtual* (derived upon demand) or *federated* (pulled from other repositories). This data includes the entire value chain of science from raw sensor readings to peer-reviewed journal articles, along with all data derived in the process.

The *observers* are people/equipment/procedures that collect the raw empirical data and ensure its validity, possibly using Laboratory Information Management Systems (LIMS). The *experimenters* are people/equipment/procedures that analyze relationships in the data based on predictions from theory. The *simulators* are people/equipment/procedures that model the data and simulate future situations and verify it with observations. The *publishers* are people/equipment/procedures that take all of this information and package it into products and services for the consumers.

In small-scale science, the same per-

son may play all these roles. However in large-scale science, many people with differing skills and perspectives become involved.

Consumers are people who derive value through science products and services. For large-scale science, consumers are often a diverse group, ranging from peer scientists, funding agencies and government regulators (in certain areas such as medicine). The public is involved in two important ways. *Science public* are people who understand and are interested in the science. They may actually contribute to the science, such as finding a new comet. *General public* are people who do not understand or even care about the science but are impacted by the results. For example, specific predictions about global warming should be of interest to anyone living near an ocean.

Maturing Science Intelligence

A large-scale science project cannot adopt full SI in a single step. It is a long-term ongoing process. The critical success factor is how SI matures over time, especially in terms of the integration of processing and data.

As shown in the upper part of Figure 2, any new domain will start with a large software library of domain-specific solutions.

For instance, the library may consist of FORTRAN routines that correct instrument readings or MathLab pages that simulate ranges of conditions. The common platform may be Linux, along with an assorted array of Perl and shell scripts to glue together the various domain-specific routines. As the domain matures toward SI, a greater emphasis will be placed on a common processing platform that a few domain-specific libraries will augment.

Likewise, in the lower part of Figure 2, the data side of SI will mature similarly. Initially, a large collection of domain-specific data sets will be maintained. Each data set will be inconsistent with the others. Detailed knowledge of the format, collection, calibration and the like will be required to make proper use of this data. As the domain is maturing toward SI, a greater emphasis is being placed on common integrated data, which will eventually evolve toward a consistent view of reality about the science domain.

Challenges of Science Intelligence

The objective of SI is to conduct "smart" science that efficiently uses information resources to understand specific science domains and progress toward useful

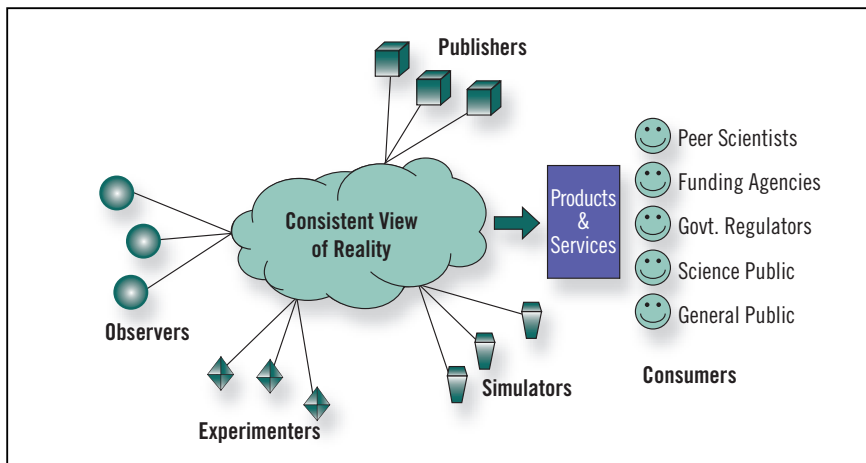


Figure 1: Information Infrastructure of Science

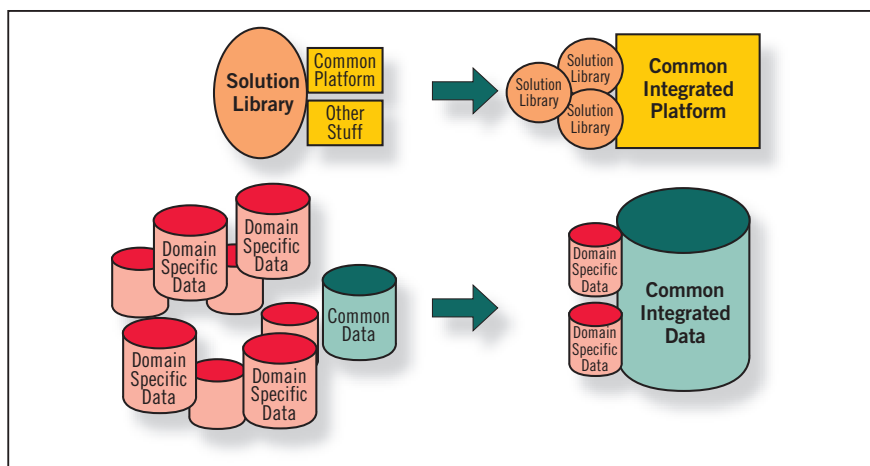


Figure 2: Maturing of Science Intelligence

applications based on that understanding.

In small-scale science, smart science is done intuitively. An experienced scientist manages his/her staff and utilizes sophisticated equipment, resulting in smart science.

The challenges come with large-scale science, involving hundreds of scientists across geographically dispersed organizations and massive complex data requirements. These challenges are as follows:

Data Scalability: In the past, data was recorded on a clipboard, entered onto punch cards and stored in simple files. With small-scale science, current data collection has not fundamentally changed, with the laptop replacing the clipboard. Launching this month, the Mars Reconnaissance Orbiter will generate a continuous data stream of 4 megabits per second from six instruments.¹⁵ Over the past decade, the data used by science has evolved from megabytes to petabytes – nine orders of magnitude! This fundamentally changes the

nature of information processing for science intelligence. We cannot continue to use the same infrastructure employed in the past.

Source Diversity: In small-scale science, each experiment has a couple of sensors. Now, there may be hundreds of different sensors collecting a wide diversity of data, all of which must be verified and calibrated to known standards. For instance, consider the sensors on the Mars Rovers, which are complex robotic instrument platforms.

Analysis Complexity: In small-scale science, the experiment is run, data is analyzed and the article is written. Now, analysis of data may continue for years, involve hundreds of scientists and generate thousands of articles. After the equipment goes silent, the data from the Mars Rovers will occupy the careers of planetary geologists for decades. How will this data and all the derived analyses be managed? The larger problem is the proper reuse of scientific data for secondary research.


Consumer Diversity: In small-scale science, the peer-review journal article is the delivery mechanism. Results are eventually published, and the underlying data is typically forgotten in dusty card decks. With large-scale science, a variety of products and services should radiate to a diverse group of consumers. With the high price for science, these consumers are increasingly demanding products and services that are relevant to them. NASA has set the standard in creating Web sites for specific missions that are informative and timely for everyone in the world.

Sharing and Collaboration: In small-scale science, this only happens in small groups of colleagues. In large-scale science, many people from many institutions located in many countries must become involved. Sharing of information at all levels and collaboration among all parties is necessary to maximize the outcome and obtain the holistic view needed to tackle the inherently more complex scientific questions being explored. We are faced with spectrum of new issues concerning intellectual property rights, access rights, privacy of personal data, confidentiality of corporate data, liability of inaccurate data, compliance with government regulations and many more.

Cultural Issues: There is a huge cultural chasm between traditional science and business, which impacts the synergism of SI and BI. Susan Flood, principal strategist at SAS, remarked, "Performance management with the proper metrics is required to show the business impacts of projects. But, most science people initially don't understand the economic [resources] and scientific benefits gained by such an intrusion on their research."¹⁴ The point is that a balance must be struck between the subjective discovery activities and the objective business goals of any large-scale scientific endeavor.

The key factor for these challenges is defining the data semantics and interrelationships – the ability to integrate disparate data, derived analyses and packaged products into a single, consistent view of science reality that drives results interpretation in context. With enterprise data warehousing of business systems, the management of meta data has always been considered a critical component. SI will take meta data to another level, requiring the latest advances from semantic analysis and service-oriented architectures.

These are tough challenges, but the stakes are high. We are currently engaged in the intrepid journey of understanding

and defining the rules to achieve such a holistic big-picture view. Hopefully, we can engineer the right technology and design the right infrastructure to not only make science intelligence practical, but also deliver the anticipated benefits. As a global society, we must be smart in the use of science, whether it is to discover ways to generate cheap reliable energy, develop new safe drugs or manage the impacts of climate changes. The future of mankind may depend on it. 

References:

1. IDC Market Analysis: Worldwide Business Analytics Software 2004-2008 Forecast, Report #31857, 2004.
2. National Science Foundation, InfoBrief: US R&D Projected to Have Grown Marginally in 2003, NSF 04-307, February 2004. <http://www.nsf.gov/sbe/srs/infbrief/nsf04307/start.htm>.
3. NSF At A Glance, <http://www.nsf.gov/about/>.
4. <http://www.sdss.org/>.
5. <http://cas.sdss.org/>.
6. Gray, Jim and Alex Szalay. "Where the Rubber Meets the Sky: Bridging the Gap between Databases and Science." Microsoft Research Report MSR-TR-2004-110, October 2004. <http://arxiv.org/ftp/cs/papers/0502/0502011.pdf>.
7. <http://www.research.ibm.com/journal/sj/402/haas.html>.

8. <http://www-1.ibm.com/industries/healthcare/doc/content/solution/976257305.html>.
9. <http://www.sas.com/industry/pharma/rdm/>.
10. <http://www.sas.com/industry/pharma/sds/>.
11. <http://www.microsoft.com/Industry/Healthcare/LSvision.msp>.
12. Interview on May 10, 2005.
13. MRO Communications with Earth - <http://marsprogram.jpl.nasa.gov/mro/mission/comm.html>.
14. Interview on May 10, 2005.
15. National Science Board, Science and Engineering Indicators, January 12 2002. <http://www.nsf.gov/sbe/srs/seind02/>. Examples of shared databases were taken from Chapter 8, Significance of Information Technology.

*Dr. Richard Hackathorn is president and founder of Bolder Technology, Inc., an established consultancy in Boulder, Colorado. Hackathorn has more than 30 years of experience in the IT industry and is a well-known technology innovator and international educator, conducting professional seminars in 18 countries. He has written three textbooks entitled **Enterprise Database Connectivity, Using the Data Warehouse** (with W.H. Inmon) and **Web Farming for the Data Warehouse**. He earned his B.S. degree from the California Institute of Technology and his M.S. and Ph.D. degrees from the University of California, Irvine. To contact, send e-mail to richardh@bolder.com.*

Large Shared Scientific Databases¹⁵

- GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>) is the National Institute of Health's annotated collection of publicly available DNA sequences. As of June 2001, GenBank contained approximately 12.9 billion base pairs from 12.2 million sequence records. The number of nucleotide base pairs in its database has doubled approximately every 14 months. As part of a global collaboration, GenBank exchanges data daily with European and Japanese gene banks.
- The Protein Data Bank (<http://www.rcsb.org/pdb/>) is the worldwide repository for the processing and distribution of three-dimensional biological macromolecular structure data.
- The European Space Agency (ESA) Microgravity Database (<http://spaceflight.esa.int/eea/>) gives scientists access to information regarding all microgravity experiments carried out on ESA and National Aeronautics and Space Administration missions by European scientists since the 1960s.
- The Tsunami Database (<http://www.ngdc.noaa.gov/seg/hazard/tsu.shtml>) provides information on tsunami events from 49 B.C. to the present in the Mediterranean and Caribbean Seas and the Atlantic, Indian and Pacific Oceans. It contains information on the source and effects of each tsunami.
- The Earth Resources Observation Systems Data Center (<http://edcwww.cr.usgs.gov/>) houses the National Satellite Land Remote Sensing Data Archive, a comprehensive, permanent record of the planet's land surface derived from almost 40 years of satellite remote sensing. By 2005, the total holdings will reach approximately 2.4 petabytes of data.

Examples of BI Vendor Offerings in Life Sciences

IBM offers DiscoveryLink, Discovery Warehouse and Clinical Genomics as tools to integrate diverse scientific information. Using these products from IBM, the Mayo Clinic has integrated 4.4 million patient records containing highly diverse data formats and protected patient confidentiality.

SAS offers SAS Research Data Management as part of their SAS Scientific Discovery Solutions. Through a partnership with RTI International, SAS supports the data management for the Models of Infectious Disease Agent Study (MIDAS) of the National Institute of Health (NIH). This is an integral component of the overall NIH biodefense plan for terrorism using infectious diseases.

Microsoft offers Digital Pharma that supports drug discovery, drug development, manufacturing/supply chain and sales/marketing. The integrated Cancer Care Unit of Roche's Diagnostics Division had major problems sharing data among 50 R&D employees scattered worldwide. Microsoft, along with their partner Avinci, developed a solution using Microsoft Office SharePoint Server, incorporating complex and reliable security restrictions on data sharing.

Additional information on each of these three examples is available in the Web version of this article at www.dmreview.com/article_sub.cfm?articleID=1032139.