Data in Motion:

A New Paradigm in Research Data Lifecycle Management

Nicholas F. Tsinoremas, Joel Zysman, Christopher Mader, Ben Kirtman and Jay Blaire

Center for Computational Science

University of Miami

The biggest challenges of the modern-day world are far from simple. Whether they involve chronic disease, climate change or failing economies, these problems, their solutions and the process in reaching said solutions are all multi-faceted, with an extremely large array of contributing factors.

No longer do researchers look at a given health problem only in potential biologic or chemical causes. They must consider several layers of genetic and environmental links. Climate change questions envelop historic information from a variety of sources and an increasing number of ongoing environmental observations to derive complex models of weather and climate patterns. Financial market analyses include countless macro influences, but also micro influences that can emanate from new, man-made factors including social media and market manipulation.

Starting with industry, data volume grew and continues to grow exponentially.[i] Reportedly, Wal-Mart processes more than one million customer transactions hourly that translate into databases estimated at more than 2.5 petabytes.[ii] In the same issue of *The Economist* that reported that statistic, it also relayed that a 2008 International Data Corp study projected approximately 1,200 exabytes of digital data would be generated in 2010.[iii] In research academic institutions, science disciplines such as genomics have created our most significant "data deluge." In the 1990s, data set sizes in bioinformatics/genomics ranged from several to tens of gigabytes. In the past five years and with the introduction of new

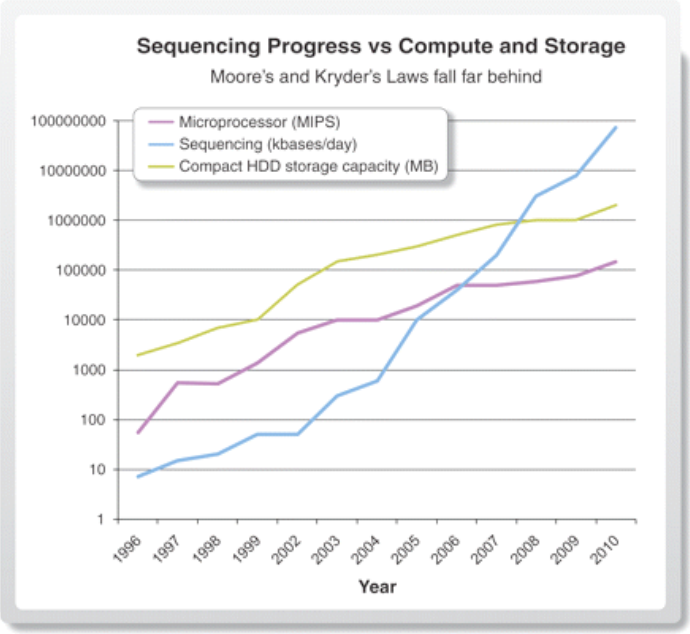

Figure 1. A doubling of sequencing output every nine months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields. (Graph credit: S.D. Kahn, *Science*)

chemistry and instrument technologies a typical Next Generation Sequencing (NGS) instrument generates 4+ terabytes of data in a matter of a few days. Full production levels can reach more than 150 TB/year per instrument (40 weeks x 4 TB). Furthermore, processing these data can produce interim sizes of ten times that amount. But, it is also this very instrumentation that exponentially accelerates researchers' ability to study disease causal genes and thus further medical science more quickly.

Climate science, too, revolves around data from instrumental, paleoclimatic, satellite and model-based sources. And while these data are increasing significantly in quantity and complexity, the discipline faces an additional challenge; the high volume of observations and model output is shared with an extremely diverse collection of user communities that exhibit a very large spectrum in their level of sophistication and that use a very disparate set of tools to interrogate the data. Just one example are the increasing number of resource managers (not just research scientists) who work in fields such as water, public lands, health and marine resources are accessing this data to make informed decisions. As Overstreet et al. point out, these "climate data provide the backbone for billion-dollar decisions. With this gravity comes the responsibility to curate climate data and share it more freely, usefully and readily than ever before."[iv]

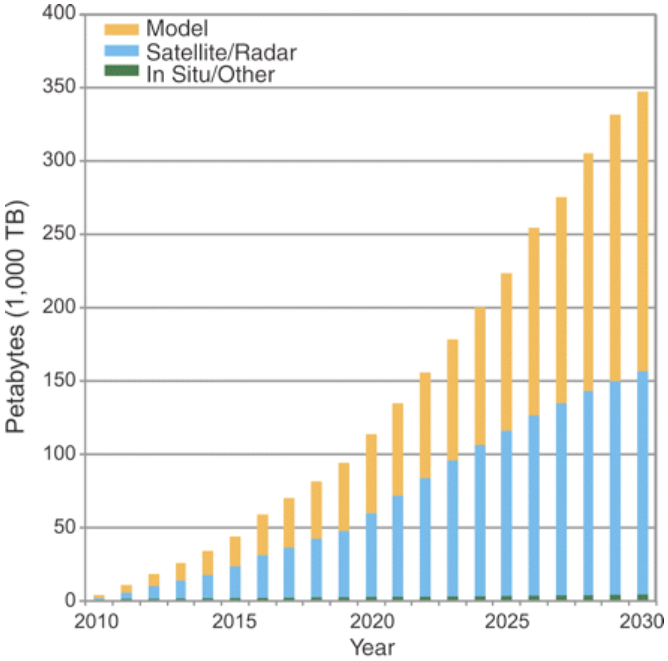

Figure 2. Climate models, remotely sensed data, and in situ instrumental/proxy data are expected to continue the significant increase that has occurred in climate change data. Graph credit: Overpeck et al., *Science*.

Take for example how multiple climate models (order 4-8) are used to make routine seasonal predictions for use by NOAA forecasters.

Typically, these models produce 3500TB of data that need to be analyzed "on-the-fly. The problem of weather and climate prediction from days-to-decades will necessarily involve heterogeneous observational data collected from satellites, ships, planes, buoys, subsurface ocean platforms and land-based stations to "initialize" and validate forecasts. Moreover the predictions systems themselves will involve complex heterogeneous computational models covering a variety of space and time scales, medium (e.g., ocean, atmosphere, land) and physical and biological processes. These state-of-the-art models will represent at high resolution the coupled ocean-atmosphere-land-cyrosphere system and all the dynamical, physical and biological processes relevant on a broad spectrum of time scales. Ultimately analyzing the observations and using them to confront the models and forecasts represents an enormous data volume challenge that also translates into substantial data accessibility and integration issues.

The volume of data generated is tremendous. For example, in terms of the weather and climate modeling systems, Kinter and Taylor estimate the data generated for a single weather or climate prediction numerical experiment (e.g., the IPCC climate change simulations) involving O(10) modeling groups (the current IPCC experiment, AR5, involves over 20 modeling groups) to be on the order of 10 exabytes.[α][v] While online data storage of this magnitude may be available in 2015, it is clear that network capacity or bandwidth is not keeping pace with technological advances in data storage media or with high performance computing throughput. It is impossible to imagine that any

[α]A rough estimate of data storage needs for a "single experiment" performed with the climate models of the future can be made as follows: Such models will have O(102) levels representing the vertical structure in the system and O(108) columns, subsampled before saving at a resolution perhaps a factor of 100 lower, or in some cases run only at a lower, O(10 km), resolution, yielding O(106) saved columns. The models will output O(102) three dimension fields and O(103) two-dimensional fields, representing the prognostic and diagnostic variables that characterize the physical, chemical and biological state of the system. Data will be saved O(103) times per run, whether it is a relatively short weather prediction run or a longer climate simulation run – typically, sampled every half hour for weather prediction, four times per day for seasonal prediction, and monthly for climate simulation. The model integrations will be instantiated O(101- 102) times to represent ensembles that can be used to estimate uncertainty in each of O(102 – 103) cases – e.g., three years of weather prediction cases or O(103) choices of uncertain parameter values in climate prediction cases. Thus, O(1010 – 1011) bytes will be stored for each of O(103) save times in O(104 – 105) runs per experiment suite, which means the global repository of COPES model output data sets will amount to O(1017 – 1019) bytes or O(0.1 to 10) exabytes (1018 bytes) per model per suite of experiments for each of O(10) modeling groups worldwide.

single data center or repository can serve this data from end-to-end; indeed a distributed approach must be adopted and data access, sharing and interrogation must be viewed based on tiered usage (See Figure 3).

Currently, approximately 20 percent of raw satellite data is used directly in weather forecast models and for basic research. The relatively low usage rate is a problem that primarily lies in translating the raw satellite data into useful information that can be used in models or in scientific discovery. This low usage is, in part, a data management problem that will only be exacerbated as satellite date volumes grow at exponential rates. The U.S. National Environmental Data and Information Service (NESDIS) estimates its satellite data holding increased by a factor of 20 from 1999-2005 and that by 2015, it will rise to 14,000 TB. In addition to the massive increase in data, requests for information are expanding at nearly the same rate. While clearly an important component, the solution is not simply bigger and faster computers. Data policies, procedures and standards must be developed to ensure that the wide variety of data coming from all sources (not just satellites) can be integrated, assessed and be used by a diverse, expanding research community.

Beyond genomics and climate change, researchers in many, if not all disciplines must consider the implications of growing multi-scale interdisciplinary work and the consequences of data derived from a plethora of sources, in different formats, on different systems and from different parts of the world. Because of this data-centric focus or fourth paradigm of research[vi], researchers cannot afford to ignore computational needs. Grants are in jeopardy if researchers fail to show their data management planning in protocols and proposals. Research universities are investing more and more in multidisciplinary approaches to scientific discovery.

Our institution, the University of Miami, consequently invested in the ongoing integrity of its scientific research by establishing a Center for Computational Science four years ago. In that time, our center has collaborated with experts across the University and around the globe, exploiting

supercomputers that could perform trillions of calculations per second. To be a relevant part of research teams and ensure that scientific advances can proceed in a timely fashion, we have developed a four-tiered data management approach. But effective data management must also incorporate the perspectives of a wide range of people. Librarians, researchers and data management experts must all be at the table, refining these solutions so that data as they even exceed Moore's Law in their growth potential, are effectively managed, so people are aware of available, accessible data, and so ownership and responsibility for maintaining data and their integrity become clearer.

Managing data in motion

The immediate challenges faced by any research informatics organization are those associated with simply storing and moving the ever-increasing volumes of data produced by the modern scientific discovery process. We have progressed beyond the point where upgrading individual parts of the data management ecosystem is enough. Now systems and informatics architects need to evaluate their entire operations and look holistically at data and user needs.

At the University of Miami, we currently operate eight NGS instruments that help researchers identify genes that carry a heightened risk for autism, study the oncogenesis for viral-associated cancers, genetically dissect viruses for future vaccine development, and conduct other work that explores genetic connections to chronic diseases. Already we plan to double the number of NGS equipment in the next few years. These instruments are located in specialized wet lab facilities, far removed (relatively speaking) from the University Data Center. Like most places, we use Ethernet networks for data movement. Given the location of the instruments and the size of data needing to be transferred, we were forced to reengineer not only our campus network, but also our inter-campus backbone as well to accommodate data transfer and the network service interruption during the daily moves of these massive data. A Dense Wavelength Division Multiplexing (DWDM) ring between

campuses offered fast, simple, and dynamic provisioning of network connections for the high-bandwidth services we needed, so we are now able to assign 10Gb wavelengths for different needs. We have isolated the Lab and HPC/Research networks from the rest of the university's traffic to minimize the impact in all directions. But campus networks are only a part of the equation. Serious thought needs to be given to the architecture of the Local Area Network (LAN) as well. Many functional data operations share common traits. Identifying the needs of the consumers of these data and identifying the common features can help optimize a network design for data movement and availability. The traditional hub and spoke design of most data centers may need to shift to other paradigms, such as top-of-rack switches and multiple network access points for different servers. Server connections themselves are another area that needs to be analyzed. While Gigabit Ethernet is now ubiquitous in research facilities, the adoption of a 10GbE (and higher) needs to be analyzed not only at the edge, but within LANs as well.

But computational centers face other technical challenges when storing and processing this data. Data analysis and capacity planning now need to be centered on data consumers, the motion of data, and the utility derived from the data for, say decision support; rather than just size alone. That said, data size is obviously still formidable. Instead of dealing with file-systems ranging in the terabyte range, we must now manage data stores occupying *petabytes* of space, with hundreds if not thousands of clients (and often from significantly varying locations) needing access at any point in time.

In *Science*'s special issue on data, Kahn discusses the challenge of storing and working with data, using the example of the 1000 Genomes Project (www.1000genomes.org), noting that despite the data's cloud storage, downloads even at a well-connected North American location can range between 7 and >20 days.[vii]

A Tiered Approach

One way we have addressed this issue is to look at how data are used within different analysis/modeling and production pipelines. After observing that a core data set from observation or experimentation may undergo many transformations during its "lifecycle," we determined that taking a multi-tiered approach to storage would deliver the best price/performance compromise. Prior to archiving, data's movement can vary in its level of activity, which gives us an opportunity to reconsider storage possibilities, to make movement more efficient when data are most active and minimize storage demands and costs when they are less active. We recognized that the system must be flexible so data were not imprisoned by its storage, and at all times, data must be secured both physically and logically. Consequently, we implemented and recommend a tiered structure as follows:

Figure 3

Tier 1	High-speed storage designed for pure processing and highly parallel data manipulation (Data in motion)
Tier 2	Mid-range storage designed for data presentation and mid-range parallel data manipulation (Data in motion)
Tier 3	Deep storage designed for long-term storage, presentation of data, and single-thread data manipulation (Data still in motion)
Archive	Near-line or off-line storage of past data (Data at rest, not really at rest but rather not accessed as frequently)

As can be seen from Figure 3 most data until archived are considered in motion and can move easily between all three layers. Data within the archive component can still be moved to faster tiers but with higher latency than motion between the other tiers. The size of each tier needs to be customized at each facility, but we have found that the ratio between tiers occurs roughly at an order of magnitude between them. Having data pools of this size requires new patterns of design in file-systems and archive solutions. We have utilized different forms of parallel file-systems for Tiers 1-3 (depending on data patterns and access requirements) and traditional archive methods for archive

storage. But with increasing archive sizes, new approaches will be needed as that size nears multi-petabyte ranges and even higher in the future.

Data in motion means that data exist in a number of states from their generation to their eventual archiving. Normally, the scientific data life cycle can be described in the following process:

- Data acquisition
- Management and storage
- Processing/modeling
- Post-processing analysis and data mining
- Integration
- Decision support and knowledge generation and preservation
- Archiving

How these data are treated during these various stages is often different. Information and knowledge of how the data are understood and used is critically important. There is often a large difference in the nature of these data during developmental stages. Following the NGS genomics example, the 4-5 terabytes of raw data produced by a single experiment need to go though an intense pipeline of several dozen steps including multiple quality assurance/quality control processes, assembly of all data, mapping them on to a given genome, to be able to call and understand the differences at the nucleotide level (polymorphism) and to conduct studies at the population level.

In climate science, a newly emerging project known as the National Multi-Model Ensemble (NMME) has shown that thinking in terms of one central data storage center, too, no longer meets the needs of today's data sets because of their enormity and their broad spectrum of uses. NMME is an effort to cull six major climate modeling efforts in response to a U.S. National Academies recommendation for a U.S. national approach to intraseasonal, seasonal and interannual climate prediction. The sharing or distributing of both real-time forecasts and the retrospective or historical forecasts to develop decision support application tools clearly has required a networked or distributed approach, sophisticated sub-setting tools and on-demand processing and visualization. While this

approach may not be transparent to users because they must visit multiple distribution sites, acquire inhomogeneous data that lacks sufficient standards and uniformity for easy use in application models, its need is well recognized, and leveragable software and hardware tools (e.g., THREADS servers) are evolving. Likewise, as DOIs are increasingly used to identify and document data sets, we develop an archive that is searchable and usable and quite possibly an archive best suited for maintenance via academic and other research libraries.

Another important issue has to do with how long to maintain data from validity, usefulness, and economic points of view. Also, who is to decide this is an important question that will need to be reviewed periodically? For example, with many climate model simulations the data have a relatively short shelf life, whereas climate observations need to be preserved for generations. With NGS, too, we work closely with research personnel to determine what data are primary to retain and further analyze. This type of definition is absolutely critical from both a technological and financial perspective. From a technological perspective, the correct storage architecture must accommodate the processing needs of this data while in motion. From a financial perspective, defining a minimum acceptable performance profile is critical to make sure that the right cost storage is used in the right case. In our case, using the Tier definitions in Figure 3 have enabled us to stage the data to the appropriate levels. We support roughly 300TB Tier-1 storage, which costs roughly $2,000/TB, so it is important to use this space wisely. By staging data to Tier-2 storage, which costs roughly $600 - $700/TB, we can keep far more data online for researchers at a much more reasonable cost. By extending this to Tier-3 ($300/TB) we can keep data sets online much longer than we could by using only one tier. Given the latest efforts by NIH and NSF to implement consistent data management plans and programs across all grants, keeping data online and usable for as long as possible will be very important. It is critical for systems architects to work not only with research personnel but also with IT administration to look for long-term solutions within existing IT strategies.

How technologies and their costs evolve will impact the data preservation strategy. At some point generation costs vs. storage costs may shift for even very large data sets and produce changes in preservation strategies. At the University of Miami, for example, our high performance NAS servers present data to our different computational clusters. High-performance storage is defined in this case as being able to provide in excess of 100,000 I/OPs, and 40Gb/sec of bandwidth. This performance is required so data files can be read and written to by over 1,500 simultaneous cluster nodes at any given time. In addition, high-performance storage (primarily SAN storage) also is used for large RDBMS, requiring applications, especially when thousands of transactions and queries must occur simultaneously.

Once the primary analysis of data has taken place, the data sets are moved from the high-performance file-systems to lower performing ones. We do this for technical reason as well as financial. Tier 2 infrastructure is designed specifically for secondary analysis and ease-of-user access. This access can be provided by several mechanisms, depending on organizational standards. We use CIFS for remote access to the data from clients of all types (Linux, Windows, MacOS X). Researchers are able to perform different forms of secondary analysis this way. The data can also be presented to web sites and informatics applications using protocols like WebDAV, SFTP as well as CIFS.

Once a project or data set is no longer being actively developed, but still needs to be "viewed" and occasionally analyzed, we move it to Tier 3 storage, which has a lower performance profile than Tier 2. While Tier 3 does not have the performance profiles of the faster gear, it is remarkably dense. This allows us to store, search and even visualize large data sets more economically than that of the other tiers. Access controls at this level are extremely fine grained. At this tier we utilize parallel file-systems across commodity hardware. While parallel file-systems are typically used for distributed access to data, we use them for redundancy and scalability, as well as for fine-grained access control.

At the end of the day, the data's value is clearly in its productive use. Data access requires careful

consideration at the earliest stages of designing data management architecture, and planning for flexibility to accommodate a broad base of uses is key.

What's used to analyze the data is as important as how it's stored. Commodity clusters are now used at most sites because they have become easy to install and operate. However, design is still often neglected. The traditional design of commodity clusters (and in fact supercomputers) has focused on numbers of cores and Floating Point Operations per Second (FLOPS). This design trend is slowly starting to move to a more balanced approach with I/O being recognized as a high priority along with processing. However, this trend needs to be adopted more widely so that data processing can keep up with the data deluge. While most vendors are happy to provide statistics about processing, few have gone as far with I/O operations. As customers of these vendors, we need to be cognizant that clusters and supercomputers can no longer be viewed in isolation but as a normal part of the data lifecycle. In fact, simply buying more and larger network storage systems is not the answer. Network/interconnect speeds have not kept pace with the tsunami of data. Tiering data offers important flexibility and efficiency so that research and discovery can continue in a timely fashion.

The final piece to a data intense computational ecosystem is the consideration of the physical data center to house the equipment. Data centers need to be designed and priced to accommodate the advanced technologies required for data processing and presentation. Some aspects to be considered are:

Figure 4

Colocation Facilities: Unless data center design and operations are a core competency at an institution, collocation should be considered.
Power requirements: Due to the compact nature and scale, this equipment is power hungry. Modern facilities should be equipped to handle much higher densities of equipment. Power ranges of up to 20kw per rack are common for solutions.
Cooling: Hand in hand with power requirements are cooling. Airflow studies in data centers are critical to maintaining equipment. Local airflow using liquid cooling options (doors, CPUs, etc.) is becoming more popular.
Connectivity: Connectivity is critical to any data-intensive operation. Data must be presented to many clients (both computer and human), which necessitates high speed redundant access to

systems.

Data ownership and responsibility

If you believe that the world is polarized into two groups – those who own "mode of production" and those who use said "mode"– then it should come as no surprise that data ownership is complex and a subject of much debate. The fourth paradigm is rooted in data production, which is the driving force in modern science.

Data in motion, in particular, create layers of responsibility for data security and ownership. Who is responsible for what? Who has rights to access and use data produced in different parts of discovery? Data collections often have multiple individuals or groups involved in acquisition, generation, organization, curation, interpretation and use of these collections. Assembly of these collections may take place over time spans that can range from days to years to generations and possibly longer. How to "manage" these data with respect to these diverse interests and over these time spans presents complex challenges including technical, interpretive, legal, ethical, and economic considerations at the very least.

The most formidable challenges are likely to be organizational. What are the appropriate roles of central organizations including individual and teams of scientists, disciplinary societies, universities, libraries, government and private laboratories, for managing these data? Where should lines of responsibility be drawn? Any comprehensive organization-wide approach will almost certainly require a substantial commitment of institutional funds to establish both required governance and technical infrastructure to address these management issues in a thorough fashion.

Returning to the NGS genomics example, it's very likely that any sequencing group would be operated as a core facility, and would centrally process samples from different departments within the

same organization as well as, possibly, samples from external collaborating organizations. In the case of clinical samples (e.g., tissue samples from patients) being used for research purposes, a whole set of regulatory requirements (HIPAA and others) must be adhered to when handling these data. Patients must consent to the use of their samples for research purposes and also be able to rescind consent at anytime, requiring the removal of their data from any future or ongoing studies. Any study involving patients will require a study protocol to be approved by an Institutional Review Board (IRB). This protocol must list who is allowed to see and use the data, as well as for what purpose.

Robert Shelton makes the case that informed consent, if done properly and efficiently via electronic means, can "enhance patient participation in research, expand access to data and biological samples, reduce the costs and time associated with the recruitment of patients for clinical trials, and accelerate the discovery of new treatments."[viii] However, privacy considerations can pose important challenges for researchers and limit how they manage the data.

If the protocol is amended, investigators and key personnel may be added or removed. All of this information needs to be effectively communicated between the governing organizations, the investigators and the research informatics groups. Derivative data sets must be tracked and secured, and appropriate access privileges updated and maintained for the life of these data sets. From a purely technical perspective it is certainly possible to secure the data. However, to truly meet combined regulatory and research requirements it is essential that the organizations conducting the research have the requisite governance structure and policies to meet these requirements effectively. To do this, the management structure design and organizational polices should be developed in close collaboration with research informatics organizations who will be required to help enforce these policies.

In August 2010, *The New York Times'* Gina Kolata reported on the "success" of collaborative Alzheimer's disease research with the National Institutes of Health, FDA, academia, industry and

nonprofit organizations, where data ownership was somewhat blurred. "The key to the Alzheimer's project was an agreement as ambitious as its goal: not just to raise money, not just to do research on a vast scale, but also to share all the data, making every single finding public immediately, available to anyone with a computer anywhere in the world," she wrote. "No one would own the data. No one would submit patent applications though private companies would ultimately profit from any drugs or imaging tests developed as a result of the effort."[ix] And while it's not clear how the data was managed in this instance, it is indicative of the way researchers are trying to come to terms with ownership issues and data sharing in an assortment of ways.

Additionally, in health care science, researchers are exploring trial participant data ownership, which can raise as many questions as it addresses, as one looks to resolve storage and management issues. Terry and Terry explore the benefits of crowdsourcing opportunities as a result of this take on data ownership and how it can expand the opportunity for findings beyond traditional research as well as dispense with normal HIPAA considerations.[x] However, the question remains of how and who will manage the data.

It is very clear that these issues are extremely complex and challenging for any organization. It requires developing policies and a common understanding to best accommodate such critical needs. It also requires dialog with multiple Institutional organizations across groups such as academic departments, administrative organizations, compliance organizations, IT and, of course, investigators.

From Data to Knowledge

Restricting access to data to only those scientists directly engaged in a research project limits the scope of legitimate scientific enquiry and the potential for research to influence policy and practice. No individual scientist who collects or collates data has all the possible analytic methods, expertise and time to extract key public health messages from research or routine data sets.

– Alan Lopez, School of Population Health, University of Queensland, Australia[xi]

Ultimately, the goal must be to enable the ubiquitous use of data from any part of the collection,

gathered or generated at any point in time, to be made available to researchers to discover new knowledge while still observing all regulatory and security requirements. Today's underlying assumption toward effective scientific research is that data sharing increases understanding, impact and usefulness.

The generation and organization of data sets by researchers working in discipline specific, or even interdisciplinary studies often cannot imagine how these data might be used in the future, especially by researchers in remote fields of study. Researchers in these remote fields may not be aware of the existence of data well known in the "generating" field, and may be completely unaware of the applicability of these data to the study at hand.

Additionally, as researchers, we are so focused on our own goals that we overlook the opportunity to add perspective. Computer scientists raise and address issues in an academic vacuum, meeting at symposia most often with other computer scientists. Researchers tend only to discuss data sharing issues with others in their field. Librarians, who likely hold the ideal curator's perspective, are often an afterthought as we explore how to manage this data deluge of today and plan for the future in a way that improves research capability.

Making an entire collection visible in a meaningful way, while still respecting all relevant security requirements, will require developing new software tools. These tools will very likely incorporate semantic, text and data mining, and computational linguistics technologies (e.g., Natural Language Processing) to build easily searchable and accessible catalogs and indices of these collections. For example, a "Smart Collaborator Tool" might be developed that could "understand" the study at hand and discover data previously generated by research studies in other fields that should be considered as informative by the current study. Indeed, currently funded projects are developing underlying technologies to create these tools, but fortunately for those of us interested in this subject, much interesting work remains to apply these technologies and produce novel systems to help optimize

using collections as resources for the ongoing discovery of knowledge.

Conclusion

As Jim Gray noted in *The Fourth Paradigm* and in many of his speaking engagements, data being captured 24/7 complemented by new world computer models are likely to reside forever "in a live, substantially publicly accessible, curated state for the purposes of continued analysis."[xii] Comparing it to paper-based storage, its access to anyone who can bring insight to the data has the potential to exponentially advance science and other data-driven fields.

We put forward the notion that in addition to the unprecedented increase in data volume in science and engineering, we also must consider the constant data movement required to extract information and ultimately knowledge. In summary, this paper presents a number of issues for consideration, but those are just a subset of the complex set of interrelated challenges of an integrated data management system. There will naturally be multiple approaches to meeting these challenges, yet it is more than worthwhile to produce examples of best practices as they emerge.

References

[i] Kahn, S.D. (2011). "On the Future of Genomic Data," *Science*. **6018**, 728-729. Retrieved from http://www.sciencemag.org/content/331/6018/728.full#xref-ref-15-1

[ii] Cukier, K. (25 Feb 2010). "Data, data everywhere," *The Economist*. Retrieved from http://www.economist.com/node/15557443

[iii] "All too much," *The Economist.* (25 Feb 2010). Retrieved from http://www.economist.com/node/15557421

[iv] Overpeck, J.T. (2011). "Climate Data Challenges in the 21st Century," *Science*. **331**, 700-702. Retrieved from http://www.sciencemag.org/content/331/6018/700.full

[v] Kinter III, J.L., Taylor, K.E. (06 October 2005). "Data Issues for WCRP Weather and Climate Modeling," White paper based on presentation at First Session of the WCRP Modeling Panel Exeter, UK. Retrieved from http://wcrp.ipsl.jussieu.fr/Documents/WMP/WMP1_KinterTaylor.pdf

[vi] Gray, J. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Kindle Edition. Ed. Tony Hey, Stewart Tansley, and Kristin Tolle. Redmond. Washington: Microsoft Research.

[vii] Kahn, S.D. 728-729.

[viii] Shelton, R. H. (2011). "Electronic Consent Channels: Preserving patient privacy without handcuffing researchers," *Science Translational Medicine*. **69**, 69. Retrieved from http://sagecongress.org/downloads/PrivateAccess_Shelton.pdf

[ix] Kolata, G. (12 August 2010). "Sharing of data leads to progress on Alzheimer's," *The New York Times*. Retrieved from http://www.nytimes.com/2010/08/13/health/research/13alzheimer.html

[x] Terry S.F., Terry, P.F. (2011) "Power to the People: Participant ownership of clinical trial data," *Science Translational Medicine*. **69**, 69. Retrieved from http://stm.sciencemag.org/content/3/69/69cm3.full
[xi] Lopez, A.D. (2010). "Sharing data for public health: where is the vision?" *Bulletin of the World Health Organization*. **88**, 467. Retrieved from http://www.scielosp.org/scielo.php?pid=S0042-96862010000600018&script=sci_arttext&tlng=en
[xii] Gray, J. Location 92, XIV.