DOI:10.1145/2611567

Exploring the inherent technical challenges in realizing the potential of Big Data.

BY H.V. JAGADISH, JOHANNES GEHRKE, ALEXANDROS LABRINIDIS, YANNIS PAPAKONSTANTINOU, JIGNESH M. PATEL, RAGHU RAMAKRISHNAN, AND CYRUS SHAHABI

Big Data and Its Technical Challenges

IN A BROAD range of application areas, data is being collected at an unprecedented scale. Decisions that previously were based on guesswork, or on painstakingly handcrafted models of reality, can now be made using data-driven mathematical models. Such Big Data analysis now drives nearly every aspect of society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences.

As an example, consider scientific research, which has been revolutionized by Big Data.[1,12] The Sloan Digital Sky Survey[23] has transformed astronomy from a field where taking pictures of the sky was a large part of an astronomer's job to one where the pictures are already in a database, and the astronomer's task is to find interesting objects and phenomena using the database. In the biological sciences, there is now a well-established tradition of depositing scientific data into a public repository, and also of creating public databases for use by other scientists. Furthermore, as technology advances, particularly with the advent of Next Generation Sequencing (NGS), the size and number of experimental datasets available is increasing exponentially.[13]

The growth rate of the output of current NGS methods in terms of the raw sequence data produced by a *single* NGS machine is shown in Figure 1, along with the performance increase for the SPECint CPU benchmark. Clearly, the NGS sequence data growth far outstrips the performance gains offered by Moore's Law for single-threaded applications (here, SPECint). Note the sequence data size in Figure 1 is the output of analyzing the raw images that are actually produced by the NGS instruments. The size of these raw image datasets themselves is so large (many TBs per lab per day) that it is impractical today to even consider storing them. Rather, these images are analyzed on the fly to produce sequence data, which is then retained.

Big Data has the potential to revolutionize much more than just research. Google's work on Google File System and MapReduce, and subsequent open source work on systems like Hadoop, have led to arguably the most extensive development and adoption of Big Data technologies, led by companies focused on the Web, such as Facebook,

» key insights

- **Big Data is revolutionizing all aspects of our lives ranging from enterprises to consumers, from science to government.**
- **Creating value from Big Data is a multi-step process: Acquisition, information extraction and cleaning, data integration, modeling and analysis, and interpretation and deployment. Many discussions of Big Data focus on only one or two steps, ignoring the rest.**
- **Research challenges abound, ranging from heterogeneity of data, inconsistency and incompleteness, timeliness, privacy, visualization, and collaboration, to the tools ecosystem around Big Data.**
- **Many case studies show there are huge rewards waiting for those who use Big Data correctly.**

IMAGE BY AMAZEME GRAPHICS

LinkedIn, Microsoft, Quantcast, Twitter, and Yahoo!. They have become the indispensable foundation for applications ranging from Web search to content recommendation and computational advertising. There have been persuasive cases made for the value of Big Data for healthcare (through home-based continuous monitoring and through integration across providers),[3] urban planning (through fusion of high-fidelity geographical data), intelligent transportation (through analysis and visualization of live and detailed road network data), environmental modeling (through sensor networks ubiquitously collecting data),[4] energy saving (through unveiling patterns of use), smart materials (through the new materials genome initiative[18]), machine translation between natural languages (through analysis of large corpora), education (particularly with online courses),[2] computational social sciences (a new methodology growing fast in popularity because of the dramatically lowered cost of obtaining data),[14] systemic risk analysis in finance (through integrated analysis of a web of contracts to find dependencies between financial entities),[8] homeland security (through analysis of social networks and financial transactions of possible terrorists), computer security (through analysis of logged events, known as Security Information and Event Management, or SIEM), and so on.

In 2010, enterprises and users stored more than 13 exabytes of new data; this is over 50,000 times the data in the Library of Congress. The potential value of global personal location data is estimated to be $700 billion to end users, and it can result in an up to 50% decrease in product development and assembly costs, according to a recent McKinsey report.[17] McKinsey predicts an equally great effect of Big Data in employment, where 140,000–190,000 workers with "deep analytical" experience will be needed in the U.S.; furthermore, 1.5 million managers will need to become data-literate. Not surprisingly, the U.S. President's Council of Advisors on Science and Technology recently issued a report on Networking and IT R&D[22] identified Big Data as a "research frontier" that can "accelerate progress across a broad range of priorities." Even popular news media now appreciates the value of Big Data as evidenced by coverage in the *Economist*,[7] the *New York Times*,[15,16] *National Public Radio*,[19,20] and *Forbes* magazine.[9]

Figure 1. Next-gen sequence data size compared to SPECint.

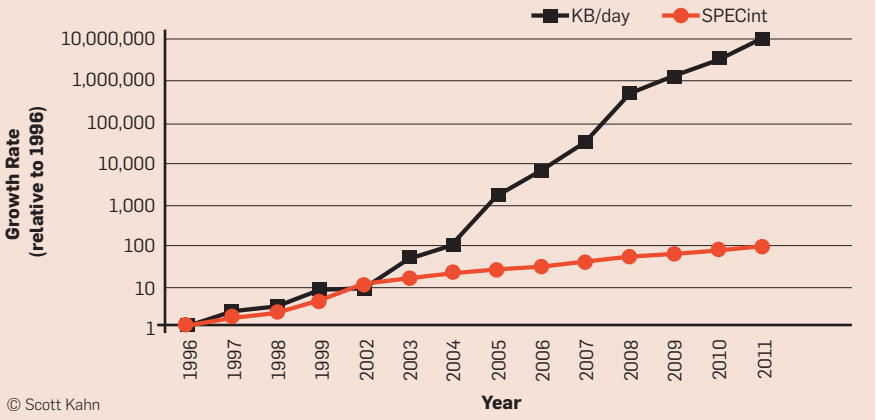
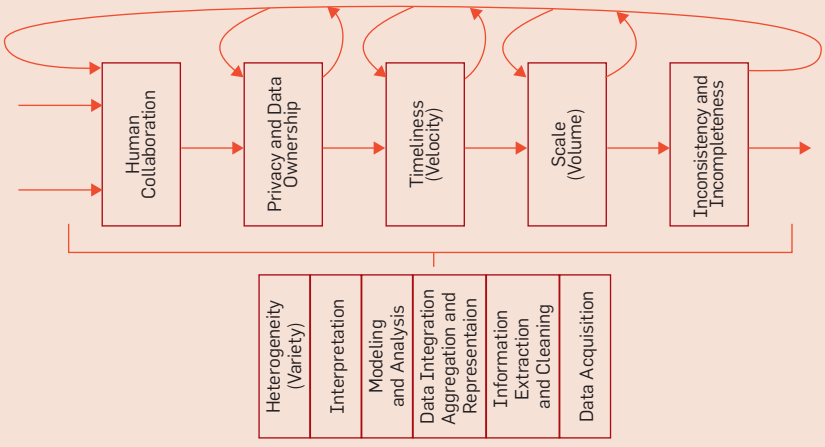


Figure 2. The Big Data analysis pipeline. Major steps in the analysis of Big Data are shown in the top half of the figure. Note the possible feedback loops at all stages. The bottom half of the figure shows Big Data characteristics that make these steps challenging.



While the potential benefits of Big Data are real and significant, and some initial successes have already been achieved (such as the Sloan Digital Sky Survey), there remain many technical challenges that must be addressed to fully realize this potential. The sheer size of the data, of course, is a major challenge, and is the one most easily recognized. However, there are others. Industry analysis companies like to point out there are challenges not just in *Volume*, but also in *Variety* and *Velocity*,[10] and that companies should not focus on just the first of these. Variety refers to heterogeneity of data types, representation, and semantic interpretation. Velocity denotes both the rate at which data arrive and the time frame in which they must be acted upon. While these three are important, this short list fails to include additional important requirements. Several additions have been proposed by various parties, such as *Veracity*. Other concerns, such as privacy and usability, still remain.

The analysis of Big Data is an iterative process, each with its own challenges, that involves many distinct phases as shown in Figure 2. Here, we consider the end-to-end Big Data life cycle.

Phases in the Big Data Life Cycle

Many people unfortunately focus just on the analysis/modeling step—while that step is crucial, it is of little use

without the other phases of the data analysis pipeline. For example, we must approach the question of what data to record from the perspective that data is valuable, potentially in ways we cannot fully anticipate, and develop ways to derive value from data that is imperfectly and incompletely captured. Doing so raises the need to track provenance and to handle uncertainty and error. As another example, when the same information is represented in repetitive and overlapping fashion, it allows us to bring statistical techniques to bear on challenges such as data integration and entity/relationship extraction. This is likely to be a key to successfully leveraging data that is drawn from multiple sources (for example, related experiments reported by different labs, crowdsourced traffic information, data about a given domain such as entertainment, culled from different websites). These topics are crucial to success, and yet rarely mentioned in the same breath as Big Data. Even in the analysis phase, which has received much attention, there are poorly understood complexities in the context of multi-tenanted clusters where several users' programs run concurrently.

One place to do it all. The most important shift may well be that increasingly, the same data goes through all five stages of the life cycle, and it is no longer acceptable to have silos that address each stage. How do we provide an integrated set of data management and analysis capabilities that support all five stages adequately?

In the rest of this article, we begin by considering the five stages in the Big Data pipeline, along with challenges specific to each stage. We also present a case study (see sidebar) as an example of the issues that arise in the different stages. Here, we discuss the six crosscutting challenges.

Data acquisition. Big Data does not arise in a vacuum: it is a record of some underlying activity of interest. For example, consider our ability to sense and observe the world around us, from the heart rate of an elderly citizen, to the presence of toxins in the air we breathe, to logs of user-activity on a website or event-logs in a software system. Sensors, simulations and scientific experiments can produce large volumes of data today. For example, the planned square kilometer array telescope will produce up to one million terabytes of raw data per day.

Pushing summarization to edge devices. What we can filter and compress is often tied to the intended analysis in intimate ways, and a fixed filtering strategy does not work well. Can we provide flexible complex event processing frameworks that can optimize data acquisition by pushing down permissible filtering and compression criteria based on the user's analysis to edge devices where the data is generated?

Much of this data can be filtered and compressed by orders of magnitude without compromising our ability to reason about the underlying activity of interest. One challenge is to define these "on-line" filters in such a way they do not discard useful information, since the raw data is often too voluminous to even allow the option of storing it all. For example, the data collected by sensors most often are spatially and temporally correlated (such as traffic sensors on the same road segment). Suppose one sensor reading differs substantially from the rest. This is likely to be due to the sensor being faulty, but how can we be sure it is not of real significance?

Furthermore, loading of large datasets is often a challenge, especially when combined with on-line filtering and data reduction, and we need efficient incremental ingestion techniques. These might not be enough for many applications, and effective in-situ processing has to be designed.

Information extraction and cleaning. Frequently, the information collected will not be in a format ready for analysis. For example, consider the collection of electronic health records in a hospital, comprised of transcribed dictations from several physicians, structured data from sensors and measurements (possibly with some associated uncertainty), image data such as X-rays, and videos from probes. We cannot leave the data in this form and still effectively analyze it. Rather, we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis. Doing this correctly and completely is a continuing technical challenge. Such extraction is often highly application-dependent (for example, what you want to pull out of an MRI is very different from what you would pull out of a picture of the stars, or a surveillance photo). Productivity concerns require the emergence of declarative methods to precisely specify information extraction tasks, and then optimizing the execution of these tasks when processing new data.

Most data sources are notoriously unreliable: sensors can be faulty, humans may provide biased opinions, remote websites might be stale, and so on. Understanding and modeling these sources of error is a first step toward developing data cleaning techniques. Unfortunately, much of this is data source and application dependent.

Data integration, aggregation, and representation. Effective large-scale analysis often requires the collection of heterogeneous data from multiple sources. For example, obtaining the 360-degrees health view of a patient (or a population) benefits from integrating and analyzing the medical health record along with Internet-available environmental data and then even with readings from multiple types of meters (for example, glucose meters, heart meters, accelerometers, among others[3]). A set of data transformation and integration tools helps the data analyst to resolve heterogeneities in data structure and semantics. This heterogeneity resolution leads to integrated data that is uniformly interpretable within a community, as they fit its standardization schemes and analysis needs. However, the cost of full integration is often formidable and the analysis needs shift quickly, so recent "pay-as-you-go" integration techniques provide an attractive "relaxation," doing much of this work on the fly in support of ad hoc exploration.

It is notable that the massive availability of data on the Internet, coupled with integration and analysis tools that allow for the production of derived data, lead to yet another kind of data proliferation, which is not only a problem of data volume, but also a problem of tracking the provenance of such derived data (as we will discuss later).

Even for simpler analyses that depend on only one dataset, there usually are many alternative ways of storing the same information, with each alternative incorporating certain trade-offs. Witness, for instance, the tremendous variety in the structure of bioinformatics databases with information about substantially similar entities, such as genes. Database design is today an art, and is carefully executed in the enterprise context by highly paid professionals. We must enable other professionals, such as domain scientists, to create effective data stores, either through devising tools to assist them in the design process or through forgoing the design process completely and developing techniques so datasets can be used effectively in the absence of intelligent database design.

Modeling and analysis. Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Big Data is often noisy, dynamic, heterogeneous, inter-related, and untrustworthy. Nevertheless, even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge. In fact, with suitable statistical care, one can use approximate analyses to get good results without being overwhelmed by the volume.

Interpretation. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results. Usually, this involves examining all the assumptions made and retracing the analysis. Furthermore, there are many possible sources of error: computer systems can have bugs, models almost always have assumptions, and results can be based on erroneous data. For all of these reasons, no responsible user will cede authority to the computer system. Rather, she will try to understand, and verify, the results produced by the computer. The computer system must make it easy for her to do so. This is particularly a challenge with Big Data due to its complexity. There are often crucial assumptions behind the data recorded. Analytical pipelines can involve multiple steps, again with assumptions built in. The recent mortgage-related shock to the financial system dramatically underscored the need for such decision-maker diligence—rather than accept the stated solvency of a financial institution at face value, a decision-maker has to examine critically the many assumptions at multiple stages of analysis. In short, it is rarely enough to provide just the results. Rather, one must provide users with the ability both to interpret analytical results obtained and to repeat the analysis with different assumptions, parameters, or datasets to better support the human thought process and social circumstances.

While the potential benefits of Big Data are real and significant, and some initial successes have already been achieved, there remain many technical challenges that must be addressed to fully realize this potential.

The net result of interpretation is often the formulation of opinions that annotate the base data, essentially closing the pipeline. It is common that such opinions may conflict with each other or may be poorly substantiated by the underlying data. In such cases, communities need to engage in a conflict resolution "editorial" process (the Wikipedia community provides one example of such a process). A novel generation of data workspaces is needed where community participants can annotate base data with interpretation metadata, resolve their disagreements and clean up the dataset, while partially clean and partially consistent data may still be available for inspection.

Challenges in Big Data Analysis

Having described the multiple phases in the Big Data analysis pipeline, we now turn to some common challenges that underlie many, and sometimes all, of these phases, due to the characteristics of Big Data. These are shown as six boxes in the lower part of Figure 2.

Heterogeneity. When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and are poor at understanding nuances. In consequence, data must be carefully structured as a first step in (or prior to) data analysis.

An associated challenge is to automatically generate the right metadata to describe the data recorded. For example, in scientific experiments, considerable detail regarding specific experimental conditions and procedures

Case Study

Since fall 2010, as part of a contract with Los Angeles Metropolitan Transportation Authority (LA-Metro), researchers at the University of Southern California's (USC) Integrated Media Systems Center (IMSC) have been given access to high-resolution spatiotemporal transportation data from the LA County road network. This data arrives at 46 megabytes per minute and over 15 terabytes have been collected so far. IMSC researchers have developed an end-to-end system called *TransDec* (for Transportation Decision-making) to acquire, store, analyze and visualize these datasets (see the accompanying figure). Here, we discuss various components of TransDec corresponding to the Big Data flow depicted in Figure 2.

Acquisition: The current system acquires the following datasets in real time:

- *Traffic loop-detectors:* About 8,900 sensors located on the highways and arterial streets collect traffic parameters such as occupancy, volume, and speed at the rate of one reading/sensor/min.
- *Bus and rail:* Includes information from about 2,036 busses and 35 trains operating in 145 different routes in Los Angeles County. The sensor data contain geospatial location of each bus every two minutes, next-stop information relative to current location, and delay information relative to predefined timetables.
- *Ramp meters and CMS:* 1851 ramp meters regulate the flow of traffic entering into highways according to current traffic conditions, and 160 Changeable Message Signs (CMS) to give travelers information about road conditions such as delays, accidents, and roadwork zones. The update rate of each ramp meter and CMS sensor is 75 seconds.
- *Event:* Detailed free-text format information (for example, number of casualties, ambulance arrival time) about special events such as collisions, traffic hazards, and so on acquired from three different agencies.

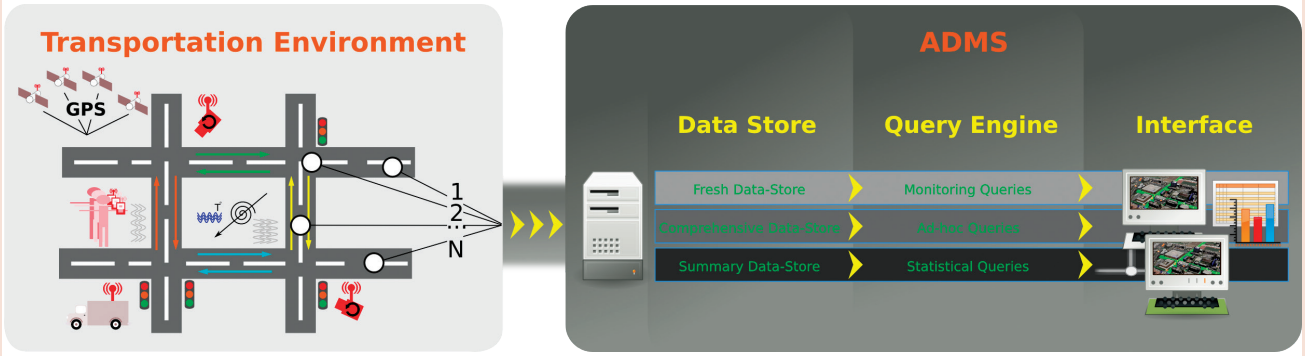
Cleaning: Data-cleaning algorithms remove redundant XML headers, detect and remove redundant sensor readings, and so on in real time using Microsoft's StreamInsight, resulting in reducing the 46MB/minute input data to 25MB/minute. The result is then dumped as simple tables into the Microsoft Azure cloud platform.

Aggregation/Representation: Data are aggregated and indexed into a set of tables in Oracle 11g (indexed in space and time with an R-tree and B-tree). For example, the data are aggregated to create sketches for supporting a predefined set of spatial and temporal queries (for example, average hourly speed of a segment of north-bond I-110).

Analysis: Several machine-learning techniques are applied, to generate accurate traffic patterns/models for various road segments of LA County at different times of the day (for example, rush hour), different days of the week (for example, weekends) and different seasons. Historical accident data is used to classify new accidents to predict clearance time and the length of induced traffic backlog.

Interpretation: Many things can go wrong in a complex system, giving rise to bogus results. For example, the failures of various (independent) system components can go unnoticed, resulting in loss of data. Similarly, the data format was sometimes changed by one organization without informing a downstream organization, resulting in erroneous parsing. To address such problems, several monitoring scripts have been developed, along with mechanisms to obtain user confirmation and correction.

TransDec.



may be required in order to interpret the results correctly. Metadata acquisition systems can minimize the human burden in recording metadata. Recording information about the data at its birth is not useful unless this information can be interpreted and carried along through the data analysis pipeline. This is called data provenance. For example, a processing error at one step can render subsequent analysis useless; with suitable provenance, we can easily identify all subsequent processing that depends on this step. Therefore, we need data systems to carry the provenance of data and its metadata through data analysis pipelines.

Inconsistency and incompleteness. Big Data increasingly includes information provided by increasingly diverse sources, of varying reliability. Uncertainty, errors, and missing values are endemic, and must be managed. On the bright side, the volume and redundancy of Big Data can often be exploited to compensate for missing data, to crosscheck conflicting cases, to validate trustworthy relationships, to disclose inherent clusters, and to uncover hidden relationships and models.

Similar issues emerge in crowdsourcing. While most such errors will be detected and corrected by others in the crowd, we need technologies to facilitate this. As humans, we can look at reviews of a product, some of which are gushing and others negative, and come up with a summary assessment

based on which we can decide whether to buy the product. We need computers to be able to do the equivalent. The issues of uncertainty and error become even more pronounced in a specific type of crowdsourcing called participatory-sensing. In this case, every person with a mobile phone can act as a multi-modal sensor collecting various types of data instantaneously (or example, picture, video, audio, location, time, speed, direction, acceleration). The extra challenge here is the inherent uncertainty of the data collection devices. The fact that collected data is probably spatially and temporally correlated can be exploited to better assess their correctness. When crowdsourced data is obtained for hire, such as with Mechanical Turks, the varying motivations of workers give rise to yet another error model.

Even after error correction has been applied, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge. Recent work on managing and querying probabilistic and conflicting data suggests one way to make progress.

Scale. Of course, the first thing anyone thinks of with Big Data is its size. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's Law. But there is a fundamental shift under way now: data volume is increasing faster than CPU speeds and other compute resources.

Due to power constraints, clock speeds have largely stalled and processors are being built with increasing numbers of cores. In short, one has to deal with parallelism within a single node. Unfortunately, parallel data processing techniques that were applied in the past for processing data across nodes do not directly apply for intra-node parallelism, since the architecture looks very different. For example, there are many more hardware resources such as processor caches and processor memory channels that are shared across cores in a single node.

Another dramatic shift under way is the move toward cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters. This level of sharing of resources on expensive and large clusters stresses grid and cluster computing techniques from the past, and requires new ways of determining how to run and execute data processing jobs so we can meet the goals of each workload cost-effectively, and to deal with system failures, which occur more frequently as we operate on larger and larger systems.

This leads to a need for global optimization across multiple users' programs, even those doing complex machine learning tasks. Reliance on user-driven program optimizations is likely to lead to poor cluster utilization, since users are unaware of other users' programs, through virtualization. System-driven holistic optimization requires programs to be sufficiently transparent, for example, as in relational database systems, where declarative query languages are designed with this in mind. In fact, if users are to compose and build complex analytical pipelines over Big Data, it is essential they have appropriate high-level primitives to specify their needs.

In addition to the technical reasons for further developing declarative approaches to Big Data analysis, there is a strong business imperative as well. Organizations typically will outsource Big Data processing, or many aspects of it. Declarative specifications are required to enable meaningful and enforceable service level agreements, since the point of outsourcing is to specify precisely what task will be performed without going into details of how to do it.

Timeliness. As data grow in volume, we need real-time techniques to summarize and filter what is to be stored, since in many instances it is not economically viable to store the raw data. This gives rise to the acquisition rate challenge described earlier, and a timeliness challenge we describe next. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed—potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination. The fundamental challenge is to provide interactive response times to complex queries at scale over high-volume event streams.

Another common pattern is to find elements in a very large dataset that meet a specified criterion. In the course of data analysis, this sort of search is likely to occur repeatedly. Scanning the entire dataset to find suitable elements is obviously impractical. Rather, index structures are created in advance to find qualifying elements quickly. For example, consider a traffic management system with information regarding thousands of vehicles and local hot spots on roadways. The system may need to predict potential congestion points along a route chosen by a user, and suggest alternatives. Doing so requires evaluating multiple spatial proximity queries working with the trajectories of moving objects. We need to devise new index structures to support a wide variety of such criteria.

Privacy and data ownership. The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what data can be revealed in different contexts. For other data, regulations, particularly in the U.S., are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy effectively is both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of Big Data.

Consider, for example, data gleaned from location-based services, which require a user to share his/her location with the service provider. There are obvious privacy concerns, which are not addressed by hiding the user's identity alone without hiding her location. An attacker or a (potentially malicious) location-based server can infer the identity of the query source from its (subsequent) location information. For example, a user may leave "a trail of packet crumbs" that can be associated with a certain residence or office location, and thereby used to determine the user's identity. Several

other types of surprisingly private information such as health issues (for example, presence in a cancer treatment center) or religious preferences (for example, presence in a church) can also be revealed by just observing anonymous users' movement and usage patterns over time. In general, it has been shown there is a close correlation between people's identities and their movement patterns.[11] But with location-based services, the location of the user is needed for a successful data access or data collection, so doing this right is challenging.

Another issue is that many online services today require us to share private information (think of Facebook applications), but beyond record-level access control we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing in an intuitive, but effective way. In addition, real data are not static but get larger and change over time; none of the prevailing techniques results in any useful content being released in this scenario.

Privacy is but one aspect of data ownership. In general, as the value of data is increasingly recognized, the value of the data owned by an organization becomes a central strategic consideration. Organizations are concerned with how to leverage this data, while retaining their unique data advantage, and questions such as how to share or sell data without losing control are becoming important. These questions are not unlike the Digital Rights Management (DRM) issues faced by the music industry as distribution shifted from sales of physical media such as CDs to digital purchases; we need effective and flexible *Data DRM* approaches.

The human perspective: Visualization and collaboration. For Big Data to fully reach its potential, we need to consider scale not just for the system but also from the perspective of *humans*. We have to make sure the end points—humans—can properly "absorb" the results of the analysis and not get lost in a sea of data. For example, ranking and recommendation algorithms can help identify the most interesting data for a user, taking into account his/her preferences. However, especially when these techniques are being used for scientific discovery and exploration, special care must be taken to not imprison end users in a "filter bubble"[21] of only data similar to what they have already seen in the past—many interesting discoveries come from detecting and explaining outliers.

If users are to compose and build complex analytical pipelines over Big Data, it is essential they have appropriate high-level primitives to specify their needs.

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a difficult time finding. For example, CAPTCHAs exploit precisely this fact to tell human Web users apart from computer programs. Ideally, analytics for Big Data will not be all computational—rather it will be designed explicitly to have a human in the loop. The new subfield of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. There is similar value to human input at all stages of the analysis pipeline.

In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system must accept this distributed expert input, and support their collaboration. Technically, this requires us to consider sharing more than raw datasets; we must also consider how to enable sharing algorithms and artifacts such as experimental results (for example, obtained by applying an algorithm with specific parameter values to a given snapshot of an evolving dataset).

Systems with a rich palette of visualizations, which can be quickly and declaratively created, become important in conveying to the users the results of the queries in ways that are best understood in the particular domain and are at the right level of detail. Whereas early business intelligence systems' users were content with tabular presentations, today's analysts need to pack and present results in powerful visualizations that assist interpretation, and support user collaboration. Furthermore, with a few clicks the user

should be able to drill down into each piece of data she sees and understands its provenance. This is particularly important since there is a growing number of people who have data and wish to analyze it.

Big Data Collaboratories. As many communities begin to rely on cloud-based data management and large shared data repositories become key resources, the potential value of collaboration using shared data goes up significantly. How do we permit users to create data analyses that combine their data with shared data and (selectively) allow other users to re-run, refine, and redistribute these analytic artifacts, which could range from single queries to entire modeling and scoring workflows? This requires us to address a number of issues (for example, provenance, access control, or workflows) but holds great potential for increased collaboration, and raising the level of transparency in collaborative work (imagine being able to re-run all the analysis reported in a paper using the same data and code used by the authors and being able to refine and publish the results!).

A popular new method of harnessing human ingenuity to solve problems is through crowdsourcing. Wikipedia, the online encyclopedia, is perhaps the best-known example of crowdsourced data. Social approaches to Big Data analysis hold great promise. As we make a broad range of data-centric artifacts sharable, we open the door to social mechanisms such as rating of artifacts, leader-boards (for example, transparent comparison of the effectiveness of several algorithms on the same datasets), and induced reputations of algorithms and experts.

Conclusion

We have entered an era of Big Data. Many sectors of our economy are now moving to a data-driven decision making model where the core business relies on analysis of large and diverse volumes of data that are continually being produced. This data-driven world has the potential to improve the efficiencies of enterprises and improve the quality of our lives. However, there are a number of challenges that must be addressed to allow us to exploit the full potential of Big Data. This article highlighted key technical challenges that must be addressed, and acknowledge there are other challenges, such as economic, social, and political, that are not covered in this article but must also be addressed. Not all of the technical challenges discussed here arise in all application scenarios. But many do. Also, the solutions to a challenge may not be the same in all situations. But again, there often are enough similarities to support cross-learning. As such, the broad range of challenges described here make good topics for research across many areas of computer science. We have collected some suggestions for further reading at http://db.cs.pitt.edu/bigdata/resources. These are a few dozen papers we have chosen on account of their coverage and importance, rather than a comprehensive bibliography, which would comprise thousands of papers.

Acknowledgment

This article is based on a white paper[5] authored by many prominent researchers, whose contributions we acknowledge. Thanks to Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwar Dayal, Michael Franklin, Laura Haas, Alon Halevy, Sam Madden, Kenneth Ross, Dan Suciu, Shiv Vaithyanathan, and Jennifer Widom.

H.V.J. was funded in part by NSF grants IIS 1017296, IIS 1017149, and IIS 1250880. A.L. was funded in part by NSF IIS-0746696, NSFOIA-1028162, and NSF CBET-1250171. Y.P. was funded in part by NSF grants IIS-1117527, SHB-1237174, DC-0910820, and an Informatica research award. J.M.P. was funded in part by NSF grants III-0963993, IIS-1250886, IIS-1110948, CNS-1218432, and by gift donations from Google, Johnson Controls, Microsoft, Symantec, and Oracle. C.S. was funded in part by NSF grant IIS-1115153, a contract with LA Metro, and unrestricted cash gifts from Microsoft and Oracle.

Any opinions, findings, conclusions or recommendations expressed in this article are solely those of its authors.

References

1. Computing Community Consortium. Advancing Discovery in Science and Engineering. Spring 2011.
2. Computing Community Consortium. Advancing Personalized Education. Spring 2011.
3. Computing Community Consortium. Smart Health and Wellbeing. Spring 2011.
4. Computing Community Consortium. A Sustainable Future. Summer 2011.
5. Computer Research Association. Challenges and Opportunities with Big Data. Community white paper available at http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf
6. Dobbie, W. and Fryer, Jr. R.G. Getting Beneath the Veil of Effective Schools: Evidence from New York City. NBER Working Paper No. 17632. Issued Dec. 2011.
7. *Economist.* Drowning in numbers: Digital data will flood the planet—and help us understand it better. (Nov 18, 2011); http://www.economist.com/blogs/dailychart/2011/11/big-data-0
8. Flood, M., Jagadish, H.V., Kyle, A., Olken, F. and Raschid, L. Using data for systemic financial risk management. In *Proc. 5th Biennial Conf. Innovative Data Systems Research* (Jan. 2011).
9. *Forbes.* Data-driven: Improving business and society through data. (Feb. 10, 2012); http://www.forbes.com/special-report/data-driven.html
10. Gartner Group. Pattern-Based Strategy: Getting Value from Big Data. (July 2011 press release); http://www.gartner.com/it/page.jsp?id=1731916
11. González, M.C., Hidalgo, C.A. and Barabási, A-L. Understanding individual human mobility patterns. *Nature 453*, (June 5, 2008), 779–782.
12. Hey, T., Tansley, S. and Tolle, K., eds. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, 2009.
13. Kahn, S.D. On the future of genomic data. *Science 331*, 6018 (Feb. 11, 2011), 728–729.
14. Lazar, D. et al. Computational social science. *Science 323*, 5915 (Feb. 6, 2009), 721–723.
15. Lohr, A. The age of Big Data. *New York Times* (Feb. 11, 2012); http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html
16. Lohr, S. How Big Data became so big. *New York Times* (Aug. 11, 2012); http://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html
17. Manyika, J. et al. Big Data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. May 2011.
18. National Science and Technology Council. *Materials Genome Initiative for Global Competitiveness.* June 2011.
19. Noguchi, Y. Following the Breadcrumbs to Big Data Gold. National Public Radio (Nov. 29, 2011); http://www.npr.org/2011/11/29/142521910/the-digital-breadcrumbs-that-lead-to-big-data
20. Noguchi, Y. The Search for Analysts to Make Sense of Big Data. National Public Radio, (Nov. 30, 2011); http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data
21. Pariser, E. *The Filter Bubble: What the Internet Is Hiding From You.* Penguin Press, May 2011.
22. PCAST Report. *Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology* (Dec. 2010); http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf
23. SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems (Jan. 2008); http://www.sdss3.org/collaboration/description.pdf/

H.V. Jagadish (jag@umich.edu) is the Bernard A Galler Collegiate Professor of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor.

Johannes Gehrke (johannes@cs.cornell.edu) is the Tisch University Professor of the Department of Computer Sciences a Cornell University, Ithaca, NY.

Alexandros Labrinidis (labrinid@cs.pitt.edu) is an associate professor in the Department of Computer Science at the University of Pittsburgh and co-director of the Advanced Data Management Technologies Laboratory.

Yannis Papakonstantinou (yannis@cs.ucsd.edu) is a Professor of Computer Science and Engineering at the University of California, San Diego.

Jignesh M. Patel (jignesh@cs.wisc.edu) is a professor of computer science at the University of Wisconsin, Madison.

Raghu Ramakrishnan (raghu@microsoft.com) is a Technical Fellow and CTO of Information Services at Microsoft, Redmond, WA.

Cyrus Shahabi (shahabi@usc.edu) is a professor of computer science and electrical engineering and the director of the Information Laboratory at the University of Southern California as well as director of the NSF's Integrated Media Systems Center.