Data Mining and Cyberinfrastructures in Biomedical Informatics

Ryan McGivern

May 7, 2011

1 Introduction

Over the past decade technology has embedded itself deeply in the realm of health care. The proliferation of digitizing patient information from the very first encounter has provided the ability to gain a better understanding of existing knowledge and analyses that could further progress the quality of care. However, there are still many challenges that require resolution before researchers can progress towards new clinical discoveries. This report conjures a discussion of two technologies at the forefront of collaborative research in Biomedical Informatics, namely *Data Mining and Warehousing* and *CyberInfrastructures*. An understanding of these concepts is fundamental to the implementation of a surviving interdisciplinary research community that can continue to investigate health care strategies and management.

One of the most significant technologic additions to health care is the adoption of the *Electronic Medical Record* as a standard. This allows for computable data to be captured in every medical encounter, and pertinent information is now at the fingertips of clinical researchers. Unfortunately, some Health Information System databases still appear to be graveyards in that valuable data is rarely looked at after direct patient care[5]. These new sources of data may be the key to unlocking new strategies for health care and research management. The first part of this discussion will address the nature of biomedical data and challenges in representing such data so that new knowledge may be extracted using sophisticated correlation methods.

Data mining and warehousing is a knowledge discovery technique that has risen in conjunction with the widespread use of relational schemas. Storing data electronically allows the enforcement of structural concepts like entity-relationship models and indexing to facilitate efficient querying of data. Data mining is an extension of this ability in that it provides methods of discovering relationships between data using sophisticated statistical algorithms. Without these techniques, identifying new data correlations within such a massive array of data is not feasible as they are essentially hidden due to the sheer size of the datasets. Furthermore, the disparate nature of information collected in different areas of health care

means that biomedical datasets are intrinsically heterogeneous. This gives rise to challenges related to data interoperability that must be addressed when warehousing such data.

The cyberinfrastructure is a more recent concept that strives to create a virtual environment that caters to all the needs of collaborative research. It is effectively the combination of existing technologic architectures, combined in a way such that it allows important collaboration ideals such as data sharing, computational resource sharing, and complex modeling tools. The idea is that resources that would otherwise be available to only a handful of researchers or research groups are consolidated into a single environment where registered users can assist in and share research methods. This discussion will not provide an in-depth analysis of the technologies used to implement a cyberinfrastructure, but instead analyze at a high level the important components and the opportunities each provides for BMI.

2 The Nature of Biomedical Data

All medical care activities involve in some sense gathering and analyzing data. Whether it is a patient narrative, which will naturally be free text, or a blood pressure reading, which will be a numeric value with some designated level of precision, all of these data are equally important. It has been said that medicine is the single greatest humanitarian art. That said, it is not possible to replace the subjective sense of disease severity that a physician senses in moments during patient interaction[2]. However, in order to best translate this characteristic to the representation of data it is important that all datum have a well-specified order of precision. One can see immediately that the challenge of representing data in a manner most efficient for knowledge discovery in fact begins at the point of data collection, usually during the time of care.

In review, a *medical datum* is any single observation of a patient. While human beings can intuitively make the transition between the unitary view of a single datum point and the associated decomposed information, nothing is intuitive to computational systems. For example, while it would suffice to record a blood pressure reading as 120/80 in an environment where it only matters that the reading is normal, an analytical environment might benefit more from the reading stored as two separate metrics, 120 mm Hg for systolic pressure and 80 mm Hg for diastolic pressure[6]. It is not in the scope of this discussion to analyze medical data acquisition, but it is important to note that it is the first step in ensuring appropriate representation. *Knowledge*, formally, is defined as that which is derived through formal or informal analysis of data. Learned through results of formal studies, heuristics, or research models, knowledge can be organized with data to produce new *information*. A database in the medical realm is a collection of individual patient observations without any summary. Thus, an EHR is at some level simply a database.

There are many advantages to storing medical information in electronic form, but the most relevant application of data mining in health care is related to clinical research support. New knowledge is learned through statistical analysis on aggregated information from a large number of patients, and this is exactly what an EHR can facilitate. However, data warehousing in a large number of hospitals is still generally limited to administrative data

sources and patient charts are rarely stored in clinical data warehouses1. This is something that clinical research groups strive to overcome because access to such data opens up a pleather of research opportunities in investigating new relationships between medical observations. Such data would provide significant support for both prospective and retrospective studies. A *retrospective study* entails the investigation of a hypothesis that was not a subject of the study at the time the data were collected. As an example from Shortliffe[2], suppose a physician notices that patients receiving a common oral medication for diabetes (drug X) seem to be more likely to suffer from post-operative hypotension than do surgical patients receiving other diabetes medications. However, the doctor has based his or her hypothesis on only a few recent observations, and he or she decides to analyze existing hospital records to see if a formal investigation is necessary. Existing patient records in a data warehouse would simplify this process significantly by allowing the physician to run a simple query to obtain data for analysis. In contrast, a *prospective study* is one in which the clinical hypothesis is known in advance, and thus, the research protocol can be designed specifically to collect future data. Both types of clinical research studies can be supported by the successful implementation of a clinical data repository.

All clinical studies aim to extend or make use of an existing *knowledge base*, which is a collection of facts, heuristics, and complex models used for problem solving. If a knowledge base is structured and implements semantic links, a software implementation might be able to combine this with sample data in order to perform case based problem solving. However, there are still limitations to this concept due mainly to the heterogeneous nature of medical data even within a single institution. It is illusory to conceive of a complete medical dataset that will cater to the needs of all health care providers[2]. This is because medical data can be rather situational in that the data collected is selective based on what is necessary for the treatment performed by the corresponding health care personnel.

Now that weve discussed the nature and certain applications of biomedical data, the next section will continue into data mining techniques and challenges that arise in biomedical data mining.

3 Data Mining and Warehousing in Biomedical Informatics

As mentioned, data mining is a knowledge discovery technique that uses sophisticated statistical methods to identify relationships between data that would otherwise remain hidden beneath the sheer size of the dataset. Generally, data mining is more beneficial when conducted upon larger datasets because naturally more inferences can be made from a greater number of data dimensions. Consequently, most applications of data mining usually require first the implementation of a data warehouse, defined by Han and Kambler[3] to be a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management and decision making. The most prevalent type of data warehousing in health care is the *Clinical Data Repository*. Abbreviated CDR, clinical data

repositories are effectively large-scale relational schemas populated with typed medical data from multiple external sources.

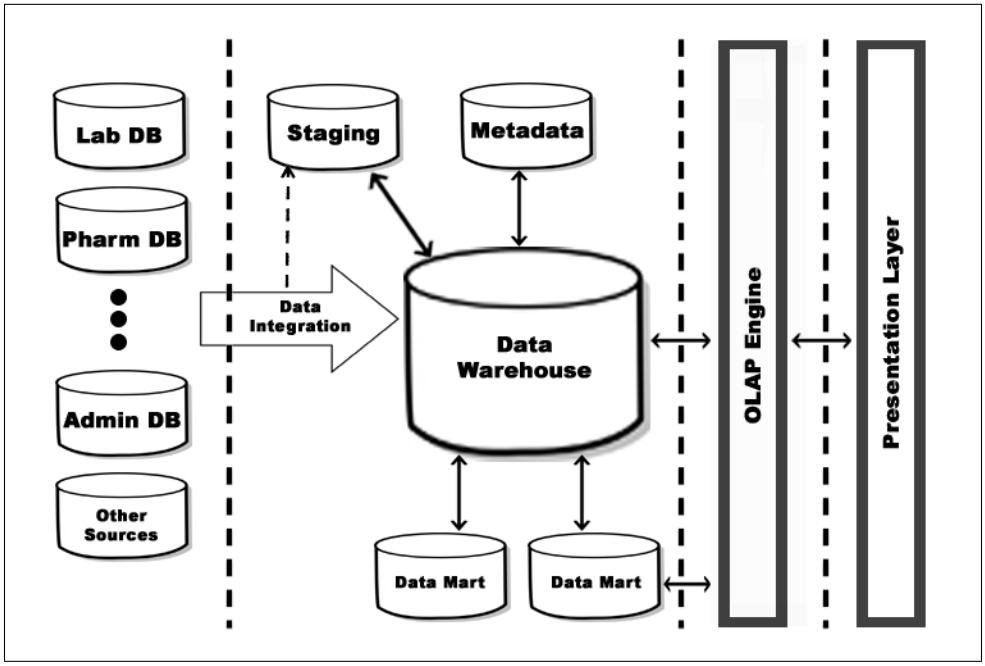

Figure 1: Clinical Data Repository

3.1 Data Mining Architecture

Typically data mining architectures are tiered into four layers: the external data sources, a data storage node, an online analytical processing layer, and a front-end presentation layer as seen in *figure 1.* The external data sources layer consists of all of the operational databases from which the original data came. The data storage layer is where the data warehouse resides and is centric to the whole architecture. It is at this layer that the *metadata* and structural constraints are strictly enforced upon the unified schema in order to maintain data interoperability and facilitate continuous submission. There is a sub-layer that resides between the operational database and data storage layers that is responsible for data exchange, performing tasks such as further data extraction, data transformation, and data refreshes. All of these tasks effectively complete what is referred to as the *data integration* stage, where data undergoes what is commonly called *scrubbing* before being loaded into the data warehouse. The data storage can be extended to implement *data marts*, which provide a storage area separate from but similar to the data warehouse. Here, subsets of data that are tailored to a specific group of users can be stored. This is also an optimal place to cache resultant datasets for reuse. It is also not uncommon to implement a staging node between the external sources and the data warehouse in order to stage the scrubbed data instead of

direct submission. The online analytic processing layer, abbreviated OLAP, is a combination of software implementations that collectively make up the data mining engine. The software modules execute algorithms for different data mining techniques such as association rule mining, classification, prediction, or clustering. This layer also implements a knowledge base that specifies the domain knowledge to govern the evaluation of resultant patterns for knowledge discovery. The front-end presentation layer can theoretically be any interface that validates and accepts a well-specified data mining task and then interprets the results to the user as comprehensively as possible. This is usually a web-based graphical user interface that accepts some job specification and delivers it to the OLAP node for processing and execution. Job submission is generally best designed using a high-level workflow language, as will be discussed in *section 4* on cyberinfrastructure.

3.2 Online Analytic Processing

OLAP layers return aggregated data in a multidimensional format that can be evaluated and visualized at the presentation layer. User queries result in a selective collection of data that was pulled from the data source, but this does not visualize trend patterns or interesting associations among data. The OLAP engine resolves this through implementing a set of functions for the user to specify summary and comparison techniques. One of the most common summary techniques is the *data cube*, which effectively visualizes results in a multidimensional nature as seen in *figure 2*. The data cube provides what are known as roll-up and drill-down operations, which allow the researcher to control the level of abstraction at each dimension when viewing data.

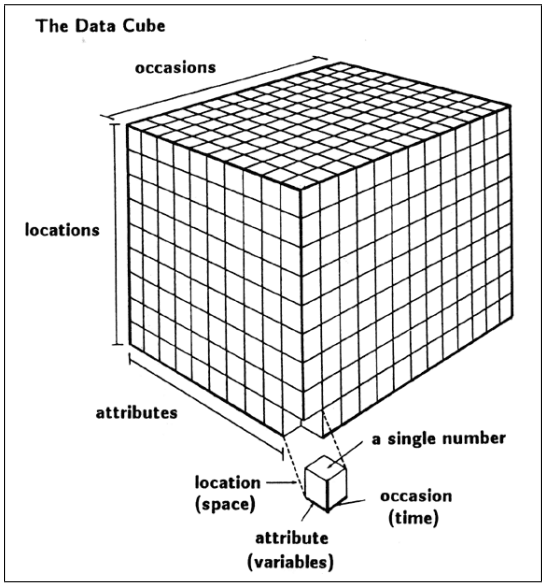

Figure 2: Multidimensional Data Cube

3.3 Data Integration

The data integration phase is at the core of the difficulties that arise from implementing architectures to support clinical research. This is because of the copious amounts of heterogeneous data being pulled into a unified relational schema. Ontologies are used to assist this process in that they allow for the mapping of primary raw data expressions to a well-specified, structured vocabulary. Then, the unified schema is open to a common set of algorithms that facilitate efficient searching and processing. Furthermore, ontologies are inherently hierarchical, meaning that these algorithms can analyze data at varying levels of abstraction. Ultimately, allowing researchers to vary abstraction levels on multiple data dimensions could unveil relationships that are not so obvious, potentially leading to new clinical discoveries. The *Cancer Biomedical Informatics Grid* (caBIG) is a biomedical informatics initiative that aims to integrate all cancer research data. Their method of doing so is by analyzing the whole data life cycle consisting of data acquisition, formatting, processing, and storage. However, the case is not as simple when attempting to cater to translational research, which appears to be crucial to exposing new clinical discoveries. This is because there rarely exists a common architecture for vocabularies among disciplines, making it difficult to consolidate terms under a single system.

3.4 Data Mining Techniques

Classical data mining techniques are classified as either *predictive* or *descriptive*. Descriptive methods mine the data for relationships between different attribute types with as few variables as possible. Predictive techniques iterate through the attributes and classify the data into predefined classes to identify similarity. Each of these techniques provides a way for recognizing patterns in query results. Both methods are ideally transparent to the user, and applications should be allowed to experiment with either at different levels of abstraction as discussed. A more recent technique is using neural networks to identify relationships within datasets. *Neural networks* are a discovery method developed to model the extremely large and complex nature of the human nervous system. Applying this method to data mining is increasingly popular mainly because it provides an efficient method of identifying sparse relationships across wide intervals.

One of the most implemented data mining mechanisms is *Machine learning*, a method that involves the discovery of trends and rules by analyzing data over time. Using a set of historical medical records, this technique has proven to assist in improving medical decision making. An example from Mitchell[12], suppose a research study involving over nine thousand medical records describing pregnant women with two hundred and fifteen distinct attributes. These attributes include age, diabetes, and previous pregnancies among others to describe the evolution of each pregnancy over time. Sample results from this hypothetical study can be seen in *figure 3*.

The analysis is designed to identify the high risk of emergency Caesarian-sections in pregnant women. The result at the bottom of the figure reflects the data mining outcome, including one of the rules learned from this particular dataset. This rule predicts a 60% risk

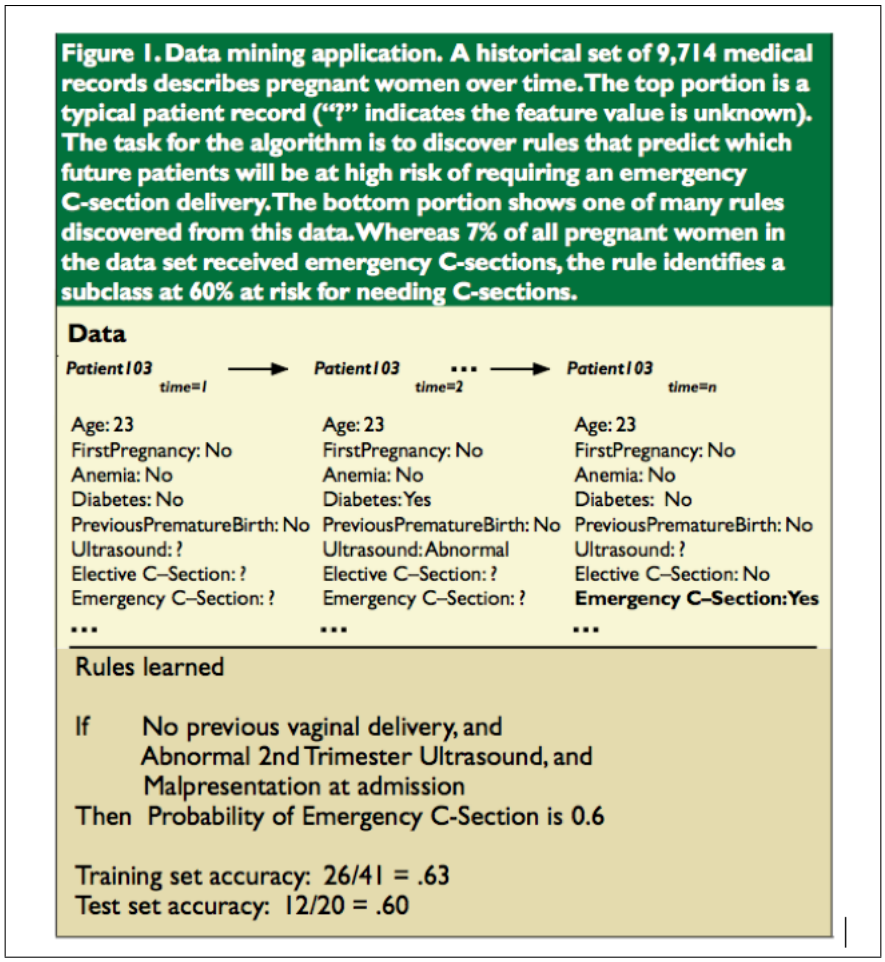

Figure 3: Machine Learning Results

of emergency C-section in women exhibiting this particular combination of age, diabetes, and whether it is the womans first pregnancy, out of the 215 possible attributes. Then observe that this regularity holds over both the rule training data and a separate test data set.

3.5 Applications of Data Mining in BMI

Most of the documented initiatives to explore data mining in biomedical informatics occur at medical institutions related to academia in some way. One study at the University of West Virginia, in cooperation with Virginia Commonwealth University, explored the results of differing data mining techniques on patient data. The data warehouse that was implemented consisted of 667,000 outpatient and inpatient records, which translates to tens of millions of datum for analysis4. A health suite implemented and maintained by IBM, *HealthMiner*, allowed the research group to extend the suite with tools that would conduct predictive and associative analysis techniques. *CliniMiner* is a data mining engine optimized for clin-

ical data, and as an extension of HealthMiner its implementation provided transparency of compliance issues for the research group, particularly maintaining patient privacy. Without having to worry about the de-identification of the subject data, the researchers at UWV were able to focus on experimenting with data mining techniques on biomedical data. The tool used for predictive analysis was a data mining engine titled THOTH.

An earlier study on this topic was conducted at Duke University Medical Centre in collaboration with other departments at Duke University related to biomedical informatics. The research group documented the study in 1997, which explained how they used data mining analyses to investigate factors related to perinatal outcomes. Smaller than the research study done at UVW, this study involved roughly 46,000 electronic patient records, which translated to the data seen in *figure 4*.

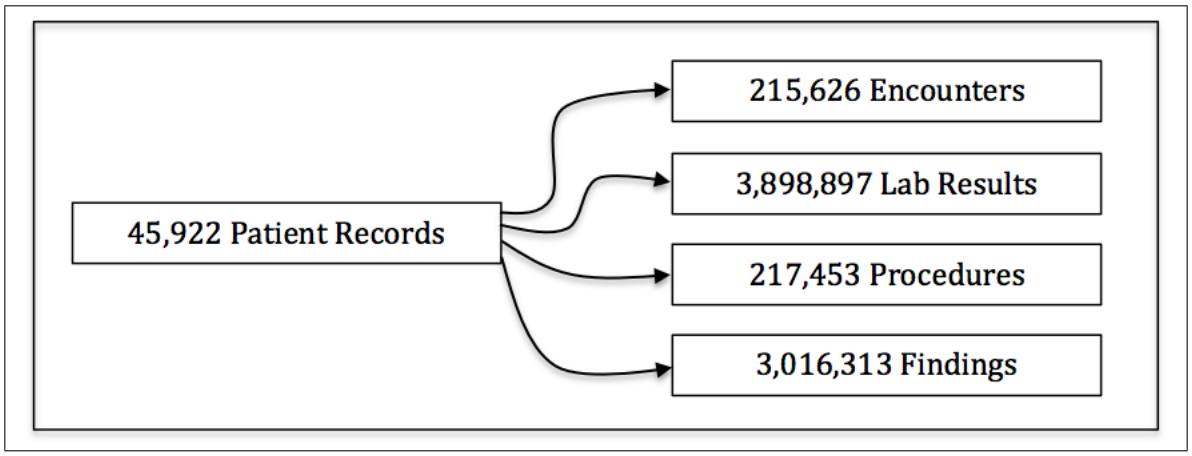

Figure 4: Patient Record Translation

The figure reflects in a nice way the amount of data that can be pulled from electronic patient records. The importance of these data is high, because one cannot discriminate any data points when the research goal is new clinical discoveries. It is then easy to imagine the possibilities that could come from enabling research on such data. Many other implementations of data mining architectures in clinical research exist that are not documented in a way that provides information suitable for research, and this could be for many reasons. Just as research groups are at times hesitant to share data, they are also hesitant to share research methods or provide a complete blueprint of the associated architectures. This is because the research community can be very competitive, where funding is acquired through grants, and research studies are rarely articled before the research goals have been achieved. In contrast, some research groups that are more committed to the greater good even publish theoretical research studies beforehand, stating their intended research and technologies to potentially enlighten other researchers. In any case, the research community is well aware of the potential benefits that could stem from analyses on a large set of patient data. Historical clinical discoveries that have been revealed through analyzing paper records and a few observations have already revealed invaluable knowledge on a pleather of diseases and

conditions. However, allowing sophisticated analysis on large patient data sets over time could shine light on new condition management strategies or lead to a firmer understanding of disease progression.

3.6 Challenges in Mining Biomedical Data

While we have discussed the nature of biomedical data and the difficulties arising through plain heterogeneity, there are a few, possibly more obvious, difficulties worthy of note. Firstly, analysis on large data sets and with what is known as a non-hypothesis driven approach can effectively result in a combinatorial explosion. What this means is that investigating relationships across such a large number of data types can exceed computability when not designed correctly. To overcome a degree of non-reducibility, researchers apply heuristics to lessen the computational intensity of experiments. Furthermore, as mentioned earlier these analyses can involve extremely high dimensionality, making it difficult to discover sparse relationships spread thinly across many dimensions. Another challenge is eliminating the bias imposed by traditional clinical research. When designing an experiment or analysis one must choose the applied heuristics diligently, so as not to limit new discovery.

Difficulties also arise from warehousing biomedical data so that datasets large enough for meaningful research can be conducted. Technology infrastructures for clinical data repositories have become quite well established for clinical trials, but are still separated from electronic medical record systems. However, if this separation was to be minimized the de-identification of data is still a challenging aspect of the CDR concept. Finally, as is reiterated throughout this discussion, the greatest challenge comes through the integration of data from such a multitude of external operational data sources.

3.7 Clinical Trial Cohort Selection

One not so obvious and commonly overlooked advantage of a clinical data repository is the ability to identify an optimal population for clinical trials. Given that patient data is de-identified within the repository, it is not possible to identify individual candidates for trials. However, what it can enable for the researcher is the ability to analyze a set of subjects in order to support the hypothesis and possibly better select optimal patient types for the clinical research study.

4 Cyberinfrastructures in Biomedical Informatics

The cyberinfrastructure is an architectural model that combines a subset of existing technologies to implement a complete research environment. Since it is a model, it is not characterized by the components that make it up, but more so by the functionalities they provide. Furthermore, it is extensible in that if further functionality is necessary but not facilitated by the basic cyberinfrastructure model, the supportive technologies can effectively be plugged in. Cyberinfrastructures have been implemented to support many areas of research and are becoming increasingly popular in the area of biomedical informatics.

4.1 Motivations for cyberinfrastructures in BMI

Computer systems are now more than essential to research. Technology has allowed for the development of complex modeling tools, providing researchers with further ability to interpret results and design new experiments. However, at times these tools are available to only a limited number of researchers and accessibility to these methods is essential to enabling a medical research task force. On top of this, cyberinfrastructures provide an environment that not only facilitates but also encourages collaboration intrinsically. Through implementing architectures to support ideals such as data sharing and a sense of community, research groups can not only make use of other datasets but share research models and methodologies.

As discussed in *section 3*, data integration is an inescapable difficulty in the realm of medical research. Integrating data from multiple external sources can require specialized training in statistics, mathematics, and at time software engineering. While most researchers have extensive training in one or more medical disciplines, medical knowledge does not generally cross into areas. Ideally, there should exist a layer of abstraction over this integration so that researchers can seamlessly integrate data into their analysis without having to worry about the correctness of data combinations.

It is therefore the mission of cyberinfrastructure implementations to provide a complete and geographically distributed research community. The environment should mainly provide data sharing through a system of data warehouses, computational resource sharing through a distributed computing grid, and collaboration tools to promote research management and the sharing of research methods. The idea is that through collaboration, and the seamless integration of many disciplines, research can head in new directions to extend the data-knowledge spectrum. In biomedical informatics this kind of effort is what could lead to further understanding the information that constitutes the substance of medicine.

4.2 Cyberinfrastructure Components

A basic cyberinfrastructure is composed of four tiers, each contributing an important research ideal to the model as a whole[8]. At the core of the architecture is the *data storage layer*, as seen in *figure 5*. This layer is responsible for providing a series of interconnected data repositories that will hold the data pertinent to research studies. It facilitates data storage,

integration, and retrieval remotely. Usually the data can be browsed using web-based front ends in order to allow the researcher to familiarize with the schema. At times this layer is extended to implement the automatic acquisition of data from external sources, as well as direct submission from researchers. Finally, the data storage layer allows for the pulling of subject data, or the placement of resultant data into allocated repositories for private or semi-private analysis. This layer is almost fundamental to a cyberinfrastructure in that if does not exist the system cannot provide one of its most important aspects, data sharing.

The *computational infrastructure* is implemented using a distributed computational grid to cluster geographically separated systems over the web. Using grid technologies researchers gain the ability to make use of all registered computational resources as they identify themselves as idle, effectively creating a virtual supercomputing node. This component is important because analyses on biomedical data can quickly become hardware and software intensive. Research methods that involve tasks such as image analysis can be greatly expedited by first optimizing the code for parallel execution, and then handing off segments to be executed on available resources throughout the grid. Results are then returned and re-compiled for interpretation and analysis by the original researcher. At the same time, fairness is inherently enforced through the fact that a resource is only used externally when it has been appropriately flagged as idle locally.

The *communication infrastructure* is best viewed at two levels, as it is responsible for more than one concept. At the low level, it is simply responsible for providing connectivity between all system nodes with acceptable bandwidth. That is to say, each layer should not only be able to communicate with the others, but at speeds that enhances the user experience as much as possible and creates the illusion that all resources are local. At a high level, this layer maintains *syntactic* and *semantic* harmony throughout data, similar to the data integration aspect discussed in *section 2*. Suppose a research method requires data from different repositories. The syntactic connectivity involves ensuring a common format for data organization. On the other hand, semantic connectivity involves checking concepts reflected in the data share a common terminology. Semantic connectivity is usually achieved through the implementation of ontology. Both of these high-level connectivity concepts cooperatively solve the data interoperability problem.

The *human infrastructure* is essential to providing the collaborative aspect of the research environment. Ultimately, it must facilitate the sociology of science and create a sense of community. Researchers then have the ability to share research protocols, analysis techniques, and obviously data sets. In addition to this, the human infrastructure is responsible for maintaining optimal user experience. Ideally, a researcher or research group should be able to define an experiment at a high level, by describing datasets, relationships, and confidence intervals. This is usually done using a high-level workflow language tailored to clinical research needs. The other components are then responsible for accepting this definition, performing optimization and execution, transforming the data, and returning the results for interpretation. What this constructs is an environment where researchers can obtain in-depth results from a high level description.

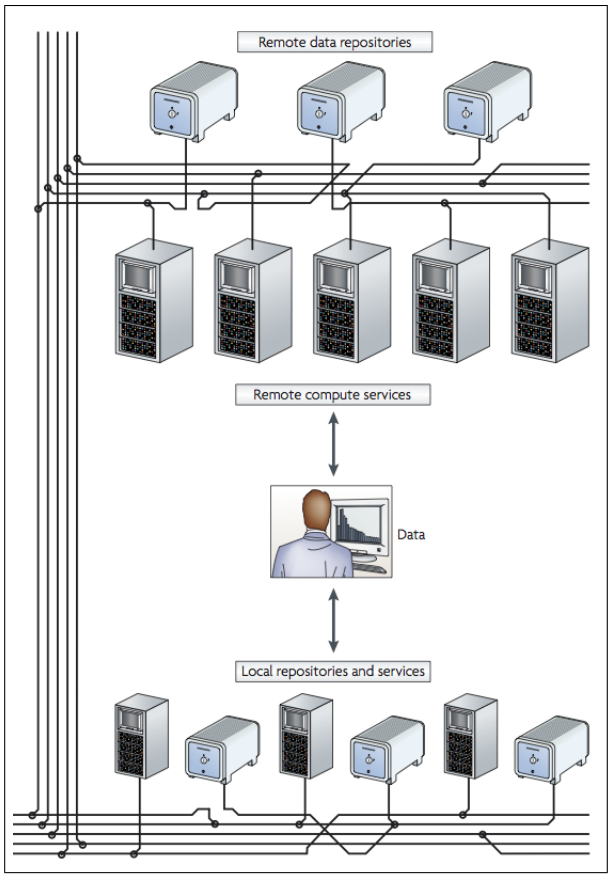

Figure 5: Cyberinfrastrcture

4.3 Existing Cyberinfrastructures

Many of the commonly used research tools out there can be effectively be translated to the cyberinfrastructures model, but they lack components to provide a complete research environment. The most common form is an online database of some form or another. These essentially enable researchers to share some data, but while there is some sense of a human infrastructure there it is certainly not complete. Popular online databases include the *European Molecular Biology Lab* and the *Protein Data Bank* (PDB). Ultimately, these architectures lack the components to facilitate collaboration and interdisciplinary research, which are served naturally by a more complete cyberinfrastructure. Furthermore, online databases are commonly centralized resources overseen and controlled by the owning research group. This can lead to political issues related to bias and permissive issues that hinder a continuous workflow. From a technological standpoint, online databases tend to very data centric in that most of the computational resources are dedicated solely to data access. This may serve the needs of a simple online database well, but it cannot facilitate the computational resource sharing to reduce intensive computations.

Another instance of cyberinfrastructure is the *community annotation hub*, commonly referred to as a wiki. These open up a central repository for direct contribution and annotation from the respective research community. Biomedical informatics wikis have quickly established themselves as some of the most sophisticated community annotation hubs despite being a relatively recent super discipline. BMI is essentially made up of the combination of numerous sub-disciplines, and this induces annotation from varying research communities. San Diego State University has developed what they call the *SDSU Gene Wiki*, which constructs a common environment for the community to cooperatively annotate gene function. However, this architecture still does not implement a complete research environment. One could see though, how this could be a valuable addition to the human infrastructure of a more complete cyberinfrastructure. The community annotation hub could theoretically be supported by the hardware and software of the suggested cyberinfrastructure model. Should some aspect not be available through this model, the extensibility of a cyberinfrastructure will accommodate the necessary additions.

4.4 Data Sharing in Cyberinfrastructures

Similarly to data warehousing architectures, it is still very difficult to interconnect data sources of disparate information classes. This proves challenging even if datasets are related by a subset of attribute types or were designed to achieve the same task in two different places. Even though two datasets related to the same discipline might share a significant number of similarities, research groups commonly use different technologies and data representations. Throughout this discussion one can see this reiterating theme of difficult data integration. This is because it cannot be stressed enough that biomedical data is in its very nature heterogeneous, due mainly to the massive amount of data types involved. On top of this, data is captured differently throughout patient care simply because it is used differently.

There are also many political issues involved with overcoming the data sharing challenge. This is because not only is it difficult for an institution to share their data, it is even more difficult to argue a business case to do so, and there are many reasons for this. Firstly, if the data of some particular owner is available for sharing then it is also open for evaluation by competing researchers. Institutions may not want their treatments, or incorrect treatments, evaluated by others. Furthermore, research groups tend to get a sense of proprietary ownership over their data. When a lot of time, money, and labor have been put into collecting valid data for research, it can be difficult to justify lending that data to other research groups. Some research is even dedicated to identifying who is the owner of medical data in varying circumstances. If the medical data is simply the collection of patient observations, then some speculate as to what kind of ownership the patient has over his or her data. Therefore, some institutions feel or ague that the data is not theirs to share. Others are skeptical as to how the community would react to their health provider releasing their information to outsiders. All of these factors, among others, play into the political or socio-economic obstacles that prevent the steady up rise of collaborative translational research. This shows that it is more than the lack of sufficient technological infrastructures getting in the way of clinical research. However, the implementation of a cyberinfrastructure sets the stage for

a consolidated research environment where policies and fairness can be easily enforced and regulated.

4.5 Data Interoperability in Cyberinfrastructures

Despite the socio-economic factors preventing the prevalence of sharing biomedical data, we of course need a technologic design to facilitate this should data actually be shared. One data interoperability solution used in cyberinfrastructures is the implementation of web services. Web services provide a common architecture for heterogeneous data and services to interoperate, and allow the researcher to call on these services with their input data as necessary. Common web service implementations consist of a *Web Service Description Language* (WSDL) and a transfer mechanism such as the *Simple Object Access Protocol* (SOAP). SOAP is not the only protocol for web service communication but its extensive support and ease of development has brought it to the forefront of web service implementations. Another prominent service protocol is *JavaScript Object Notation* (JSON), which is generally more common in simpler web-based applications. Researchers will never actually use these services directly, but they rely heavily on analysis methods like visualization engines that run on top of these.

Web services also assists in constructing the distributed computational grid. *Globus* is a set of open source libraries that facilitates the coordination of resource sharing. It achieves this through providing mechanisms for announcing the availability of a computer resource, discovering that resource, and invoking that resource. Globus has established itself as the industry heavyweight for web services in many knowledge domains, and is not only used by the Cancer Biomedical Informatics Grid (caBIG), but also the Biomedical Informatics Research Network (BIRN). However, it is important to note that it is not the only product out there. *BioMoby* is a more lightweight implementation that provides most of the functionalities provided by Globus. It is slightly more preferred by programmers due to easier development. What makes BioMoby more lightweight is that it does not implement authentication and authorization like Globus. For this reason, developers either prefer it because they can develop an authentication scheme more adaptable to a currently implemented scheme, or reject it because Globus can manage this transparently. One application of BioMoby is the PlaNet Consortium, a set of plant databases connecting distributed plant genome data.

4.6 Ontologies in BMI Cyberinfrastructures

Web services allow for the interoperation and exchange between heterogeneous data and services. But this in no way enforces data semantics, which is crucial to the correct integration of data. Similar to data mining architectures, cyberinfrastructures use ontologies to ensure an unambiguous standard for data. Cyberinfrastructures in biomedical informatics generally implement ontologies using the *Web Ontology Language* (OWL). This ontology specification is common through many domains, even beyond the scope of medicine.

4.7 Applications of Cyberinfrastructures in BMI

The largest and best documented cyberinfrastructure in biomedical informatics is overseen by the *Biomedical Informatics Research Network*. They have developed a robust software installation and deployment system to provide implementation of a BIRN endpoint. This endpoint provides researchers with services such as hosting data, access to computational resources, access to shared datasets through a web portal, analysis and visualization tools, and the ability to publish resultant datasets in the BIRN data repository. This endpoint is commonly referred to as a BIRN rack, and each costs roughly $20,000. The Biomedical Informatics Research Network makes use of Globus to implement the grid architecture, and have developed an in-house ontology called *BIRNLex*. A smaller cyberinfrastructure implementation is the *Cancer Biomedical Informatics Grid*. Their goal is to provide a common information platform to support the diverse clinical and basic research of the US National Cancer Institute. This is a noble challenge, because as we have discussed the inherent heterogeneity of all medical data, data types related to cancer studies is particularly diverse.

4.8 The Future of Cyberinfrastructures

The use of cyberinfrastructures is growing rapidly in many different research domains. Some researchers speculate that they will soon be as essential to research as plain computer systems are today. Grid computing is also becoming increasingly more efficient, facilitating widespread use of sophisticated analysis tools through resource sharing. The weaknesses of current cyberinfrastructures are related to cross-discipline collaboration. Implementations throughout certain disciplines have established a consistent grid architecture, but all of these are essentially isolated from each other. What is crucial to new clinical discoveries is the integration and communication between different disciplines so as to investigate relationships in datasets not commonly studied in accordance with each other.

Current research in the area of cyberinfrastructures is related to adopting semantic web technologies. One problematic issue with web services is the strong distinction between data and operations on that data, which unfortunately poorly serves the data interoperability problem. For example, if a researcher wants to make user of a service he or she must identify the appropriate service, format the input data, invoke that service, and then unpack and interpret the resultant data. The semantic web offers an alternative approach to this, where there is no data transformation services but only pieces of information and relationships between them. For this reason, sematic web technologies are much more tolerant of diverse data models.

5 Discussion

Hopefully this discussion has provided some insight into the importance of data mining and cyberinfrastructures in biomedical informatics. The prevalence of technology in many domains has lead to the development of new research techniques, and the medical realm presents some of the greatest opportunities out of all of them. While data mining provides

new methodologies for knowledge discovery, cyberinfrastructures incorporate all of the valuable assets essential to a complete research environment. However, there are still many challenges behind the scenes of both of these architectures, mainly related to data interoperability. A lot of current research serves to analyze different solutions to this problem, so that researchers can be enabled with all the tools necessary for revealing new clinical discoveries and extending the data-knowledge spectrum of medicine.

References

[1] Vasa Curcin. Towards a scientific workflow methodology for primary care database studies. *Statistical Methods in Medical Research*, 19:378–393, 2010.

[2] J. Han and M. Kambler. *Data Mining Concepts and Techniques*. Morgan Kaufman, 2001.

[3] Jonathan C. Prather M.S. Medical data mining: Knowledge discovery in a clinical data warehouse. *AMIA*, pages 101–104, 1997.

[4] Irene M. Mullins. Data mining and clinical data repositories: insights from a 667,000 patient data set. *Computers in Biology and Medicine*, 36:1351–1357, 2006.

[5] H. U. Prokosch and T. Ganslandt. Perspectives for medical informatics. *Medical Informatics*, pages 38–44, 2009.

[6] Johnson SB and Chatziantoniou D. Extended sql for manipulating clinical warehouse data. *Proc AMIA Symp*, page 819823, 1999.

[7] Edward H. Shortliffe. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. Springer, 2006.

[8] Barry Smith and Michael Ashburner. The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1355, 2007.

[9] Lincoln D. Stein. Towards a cyberinfrastructure for the biological sciences. *Nature*, 9:678–285, 2008.

[10] Bito Y., Kero R., Matsuo H., Shintani Y., and Silver M. Interactively visualizing data warehouses. *Healthcare Information Management*, 15(2):133142, 2001.