Big Data Analysis on Puerto Rico Testsite for Exploring Contamination Threats

Xiangyu Li, Leiming Yu
and David Kaeli

Department of Electrical and
Computer Engineering
Northeastern University
Boston, MA, USA
Email:{xili,ylm,kaeli}@ece.neu.edu

Yuanyuan Yao, Poguang Wang
and Roger Giese

Department of Pharmaceutical Sciences and
Barnett Institute, Bouve College
Northeastern University
Boston, MA, USA
Email:yao.yu@husky.neu.edu, {p.wang,r.giese}@neu.edu

Akram Alshawabkeh

Department of Civil and
Environmental Engineering
Northeastern University
Boston, MA, USA
Email:aalsha@neu.edu

***Abstract*—In this paper, we present the use of Principal Component Analysis and customized software, to accelerate the spectral analysis of biological samples. The work is part of the mission of the National Institute of Environmental Health Sciences sponsored Puerto Rico Testsite for Exploring Contamination Threats Center, establishing linkages between environmental pollutants and preterm birth. This paper provides an overview of the data repository developed for the Center, and presents a use case analysis of biological sample data maintained in the database system.**

Keywords–non-targeted analysis; principal component analysis; environmental health.

I. Introduction

Since the early 1980's, the rate of preterm birth has been increasing worldwide [1]. Preterm birth is defined as a birth of an infant before 37 weeks of pregnancy. Preterm-related deaths accounted for 35% of all infant deaths in 2010. The rate of preterm birth in Puerto Rico is 50% higher than the average in the United States. There are a number of potential factors that can increase the probability of preterm birth. There is documented evidence that ties environmental factors to increased rates in preterm birth, as reported in several studies [2][3][4][5][6][7].

In the Puerto Rico Testsite for Exploring Contamination Threats (PROTECT) Center, we are working with a cohort of over 2000 women in northern Puerto Rico (presently 800 of the 2000 have been recruited), as part of a National Institute of Environmental Health Sciences (NIEHS) P42 Center project. We are studying linkages between a large number of potential contributing factors to premature birth. The goal is to establish a link between environmental pollution, particularly Chlorinated Volatile Organic Compounds (CVOCs) and phthalates, and birth outcomes. The project also considers the fate and transport (distribution, transport and transformation) of these pollutants into water supplies in northern Puerto Rico, as well as remediation methods.

This study is highly data driven, collecting and analyzing data from a wide range of sources, including:

- Environmental Samples and Measurements - soil samples, well and tap water samples, historical Environmental Protection Agency (EPA) data, soil samples, Superfund site data,
- Biological Samples - blood, urine, hair and placenta samples, and
- Human Subjects Information - medical history, reproductive health records, product use data surveys, and birth outcomes.

The data collected is carefully cleaned and maintained in fully-indexed relational database system. The PROTECT Database allows environmental health researchers to effectively tie any two entities present in the database together through relationships across two common indices:

1) Human Subject ID, or
2) Geographic Indexing System (GIS) coordinates.

To date, over 400 million data entries have been collected, cleaned and incorporated into the database. The repository includes a comprehensive data dictionary documenting the over 2457 data entities in the system.

In this paper, we provide an overview of our data management system, discuss our data management challenges, and present a compelling use case that evaluates the urine sample data present in the system. As part of PROTECT's research mission, selected chemicals (e.g., phthalates, bisphenol A) are measured in biological samples (i.e., in blood, urine, hair or placenta samples). Looking for the presence of a suspect chemical can be done utilizing a protocol, commonly referred to as *targeted analysis*. As an example of the power of our database system, we present results on our *non-targeted chemical analysis*, utilizing Principle Component Analysis (PCA) [8] to identify suspect chemicals present in the urine samples provided by the expectant mothers in our study.

The rest of this paper is organized as follows. Section II provides an overview of the PROTECT Database System. Section III presents an example of the richness of our data repository. Section IV covers preprocessing steps needed to precondition the data, and performs analysis of the data using Principal Component Analysis. Section V concludes the paper and outlines plans for future capabilities of our system.

II. PROTECT Database

The PROTECT Database system has been built on top of EarthSoft's EQuIS software, and incorporates a Microsoft SQL Server as the database engine. A number of backend tools can work seamlessly with EQuIS, including ArcGIS [9], Surfer [10], and a number of statistical packages. Next, we discuss the different elements of the EQuIS system, which are used to maintain the large and diverse data repository maintained by PROTECT.

A. Data Repository

The Puerto Rico Testsite for Exploring Contamination Threats Center investigates exposure to environmental contamination in Puerto Rico and their causal effects with preterm birth. This program studies the high preterm birth phenomena

ISBN: 978-1-61208-445-9

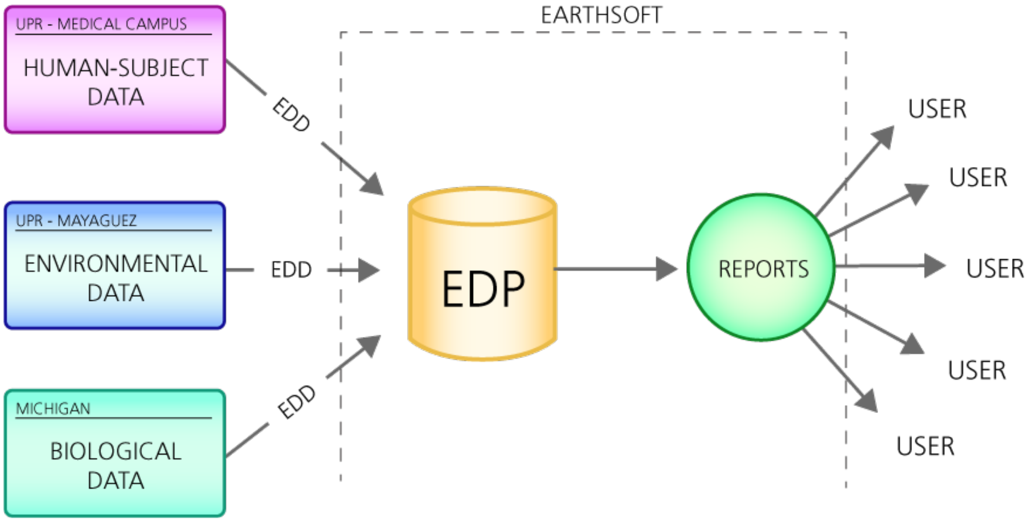

Figure 1. Data flow in the PROTECT database, where the Human Subject, Environmental and Biological Data are exported in Electrical Data Deliverable (EDD) format and verified using EQuIS Data Processor (EDP). After data are cleaned, they could be reported to different users with various permissions.

and transport of hazardous substances in karstic aquifers. In order to develop green remediation strategies to alleviate the exposure, Analytical data from various of sources are collected. Data collection includes Superfund sites, ground water, tap water, expectant mothers and birth outcome. The data repository generated supports a series of analysis activities, such as non-targeted chemical analysis, mechanistic toxicology, and targeted epidemiology. The PROTECT database delivers an efficient framework of *data management and modeling* across different disciplinary research domains.

In the current database system, data from human subjects, environmental sources and biological sampling have been collected and cleaned for further analysis. These entries exceed 400 million data points to date, as shown in Table I. More than 5 billion entries are anticipated to be housed in our system upon completion of the project. Since each data record is potentially related to adverse reproductive outcomes, understanding any correlations present across data sources is necessary. The underlying correlations could unveil important linkages between pollutants and birth outcomes. To find these linkages, machine learning techniques are applied during *data analysis*.

TABLE I. Present PROTECT database contents.

	Data Points (In millions)
Environmental	1.3
Human Subjects	1.5
Biological	400

Before incorporating any information into the database system, a careful data cleaning process is conducted. Each data export file consists of fields for different targeted analysis. For each field, the data type, format and nullability need to be verified. Its corresponding data value should stay within the range of pre-defined scope. Checking the dependencies between fields is also required. A comprehensive cleaning procedure pinpoints any corrupted data and avoids them leaking into the database. This procedure abides by the PROTECT Data Dictionary (containing detailed definitions of the 2457 different data entities in the system).

B. *EQuIS Professional*

In PROTECT database, we perform the automated cleaning by leveraging EarthSoft's EQuIS Professional [11] for the standalone development. Using EQuIS Professional, input data is first placed in an Electronic Data Deliverable (EDD) format, a format that is also supported by Microsoft Excel. The specific format of the EDD is essential for proper data checking. Each EDD entry defines the data type, range, constraints and dependencies of each individual field. The EDD format is customizable and typically includes three or four files: 1) format definition file, 2) custom handler, 3) enumeration file, and 4) reference values. The format definition file holds the definition and mapping for every field. The custom handler provides the detailed rules that apply to each data format. Common operations check for the specific data range, null data format and specific data types allowed/supported for each data field. The enumeration file is optional, and requires the EQuIS Data Processor (EDP) to execute a set of lookup values. The reference value file is needed when users need to check reference values remotely [12].

EDDs are checked according to the constraints defined in the data dictionary. First, the EDD format is verified using the format definition file in XML Schema Definition format. Whenever any conflicts occur, the corresponding fields will be highlighted by leveraging the custom handler coded in visual basic script. Additional error messages could be added to the script to facilitate the debugging process. The other two files are also necessary to make sure each field comply with the listed rules and mapping schemes. Since conflicts still exist in the EDDs, the data input are returned to the submitting project.

Once all errors are resolved, the record can be committed, and the repository is updated. Figure 1 describes the data flow of our database system. The database is frequently backed up and provides the flexibility for users to produce customized reports.

C. *EQuIS Enterprise*

EQuIS Enterprise provides web-based access to the PROTECT data, which suits the distributed development. Online access is critical since the PROTECT team is distributed across Puerto Rico, Massachusetts, Michigan and West Virginia. Instead of managing data locally, EQuIs Enterprise automates the workflow using web-based applications.

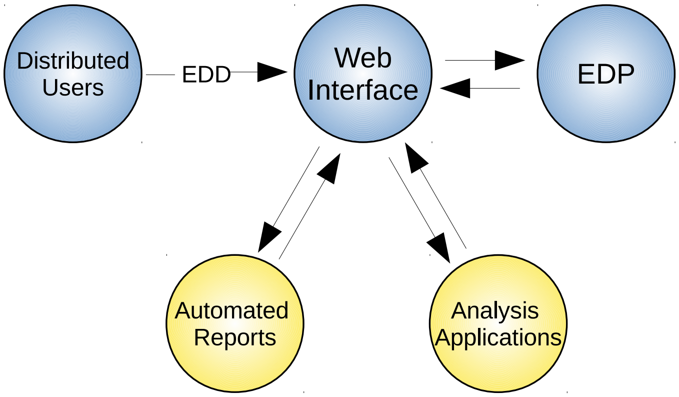

Figure 2. Data flow of EQuIS Enterprise. Distributed users can upload EDDs through the web interface where EDP is used and produce customized report accordingly.

Figure 2 describes the workflow we applied for the PROTECT database frontend. EDDs can be processed through a web interface. Users can receive status notification through File Transfer Protocol (FTP), email or web widgets. The web interface provided by Enterprise can produce standard or customized reports. It also provides the researcher with visualization of their data through Geographic Indexing System tools.

III. Urine Study

Next, we provide a use case of the PROTECT Database system. The goal is not to answer any particular question, but instead to demonstrate the richness and the challenges associated with the research project.

In our case study, we focus on the biological data that represents a majority of the data maintained in the database, as shown in Table I. The biological data contains samples from urine, blood, hair and placenta. In this example, we will analyze the urine samples present in the biological data, which holds 99% of the total biological data volume. The goal is to demonstrate our ability to perform big data analysis and modeling, which is supported by the PROTECT database system.

To perform non-targeted chemical analysis of the urine samples, we are employing a matrix-assisted laser desorption ionization time of flight/time of flight mass spectrometer (MALDI-TOF/TOF-MS). This instrument can detect many urine metabolites that we are looking for, while also giving us clues to which other chemicals are present.

Each urine sample extract is first separated into 240 droplets by Ultra Performance Liquid Chromatography, based on the analyte polarity Chromatography [13]. Each of the sets of samples is then mixed with a chemical reagent, which is sensitive to the specific laser wavelength in the mass spectrometer. The laser transforms the analyte and reagent in gas phase ions, and the detector registers the analyte (in units of m/z, the mass-to-charge ratio) and the corresponding signal intensity for each droplet. Each observed analyte is then subjected to the fragmentation analysis (TOF/TOF-MS) to check if it belongs to a specific metabolites group, for example, sulfate conjugate.

During the analysis stage, two problems associated with the MALDI-TOF/TOF-MS system (model 5800 from AB SCIEX is used) are encountered. The first problem is that the data exports (t2d) for MarkerView (the proprietary software used on the system) are stored in a binary format, which are not easily decoded in order to analyze metabolite data measured [14]. Hence, our own methodology to decode the binaries before porting them to the databas is developed, as shown in Figure 3. At first, the t2d file is decoded into mzXML format, and then the mzXML file is exported to a text file using ProteoWizard [15][16]. Customized python programs are written to extract the peak and intensity values from the text file and exported them to our database in the EDD format.

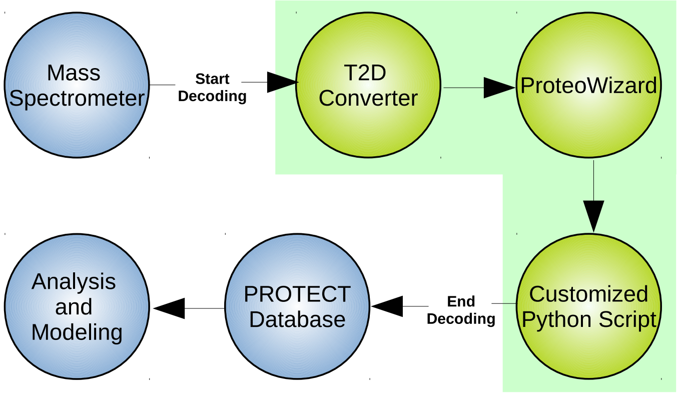

Figure 3. Urine study methodology: 1) process raw data via Mass Spectrometer 2) decode the binary using T2D converter 3) read deciphered binaries using ProteoWizard 4) extract data using Python 5) data cleaning via the database 6) support data analysis and modeling.

The second problem is the limited processing capabilities of the proprietary software that comes with the Mass Spectrometer, MarkerView. To identify metabolomic features present in the data, Principal Component Analysis (PCA) is applied on the data [8]. The software could only compute PCA on a few sets of metabolomic features (5000 highest intensity mass-to-charge features are selected), but it took 20 minutes to compute these on just 6 data sets. The processing spends 10 minutes on input processing and peak picking, and 10 minutes on computing PCA. This throughput becomes a barrier to discovery, especially when the data to be processed requires different scaling and weighting factors. A huge increase in PCA processing time when processing larger data sets is also observed. Exploring different scaling and weighting factors is severely impacted by the limited computation power of MarkerView. To accelerate the processing, our own PCA modeling scheme is developed, which is described in the next section.

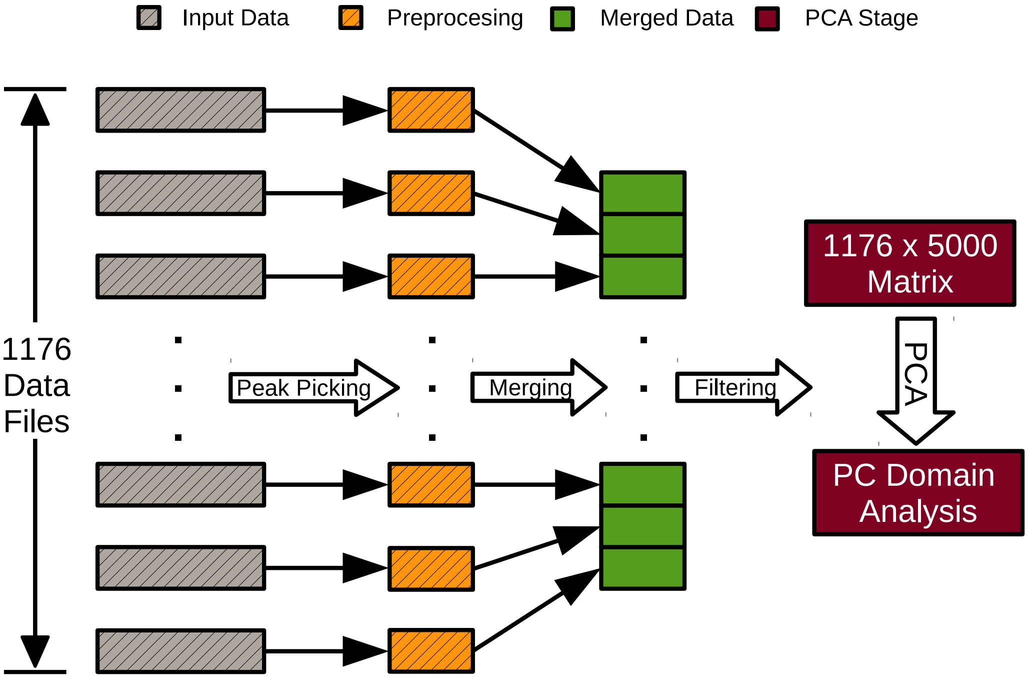

Figure 4. Urine analysis workflow. Peak picking reduces the feature dimension at the preprocessing stage. Filtered peaks are formed into a matrix for PCA.

IV. PCA of Urine Samples

The existing urine samples contain 400 million data points to be characterized, which does not include the associated TOF/TOF-MS data yet. The execution time will become unacceptable if all of them are to be directly analyzed, especially given that more than 80 urine samples are expected to be processed in the future. In addition, exploring all possible features across a large input data set may not lead to a proper classification. There may be too much correlation between the different features. Therefore, PCA is used to represent such a large number of data points with fewer uncorrelated features, while retaining the important variations present in the original dataset. In this section, we discuss how PCA is applied to analyze the patterns present in the urine samples.

A. Principal Component Analysis

PCA is often used to reduce the dimensionality of large multi-variate datasets. It transforms a set of possibly correlated samples into a set of linearly uncorrelated data points called principal components [8]. To be more specific, PCA takes as input a numerical matrix, where the rows of the matrix correspond to different input samples, and where the columns correspond to different dimensions of each sample. Then the input matrix is transformed orthogonally into the principal component domain, where each principal component is a linear combination of the input dimensions.

Principal components are sorted in a decreasing order, which captures the variance present in the input samples. These variances after orthogonal transformation are also known as eigenvalues, that describe the scaling factor of the orientation given a linear transformation. The first few principal components, usually two or three, can represent most of the variations present in the data. Therefore, PCA can significantly reduce the complexity in the data, without eliminating patterns and outliers in the data. Using PCA can greatly simplify and accelerate big data analytics by reducing the feature dimensionality.

B. Input Data

In this preliminary case, urine sample data sets collected from 6 project participants are analyzed, where each sample includes 196 urine mass spectra, resulting in 1176 urine sample files in total. In each of those urine sample files, 130K mass to charge ratio (analyte molecular weight) and associated intensities are recorded. Accurate mass to charge ratio, less than 10 ppm with internal calibration, combined with metabolome database search and the associated TOF/TOF analysis could potentially reveal the chemicals present in the urine sample. In order to detect patterns in these urine mass spectra, we need to cluster the urine samples in a 130K-dimensional space, where each dimension represents the mass of a potential chemical. However, the data dimensionality is too large for us to analyze. On the other hand the majority of data points are baseline measurements. Hence, we filter out a lot of the data before applying PCA. We describe this process in following sections.

C. Pre-processing Stage

When the mass spectrum data are decoded in the MarkerView's binaries, the 130K measurements are decoded in each file. Each measurement is separated by 0.007 *Da* (the unified atomic mass unit), which indicates the mass on an atomic or molecular scale.

The data is first pre-processed where only peaks are picked. A peak represents an analyte with the local maximum intensity value above the preselected signal to noise ratio threshold (20 in our current analysis) within a mass range . The peak-picking process eliminates the noise and baseline points that have too low intensities. The *MALDIquant* package for Quantitative

Analysis of Mass Spectrometry Data in R (a free software programming language for statistical computing) is applied to facilitate the process [17][18]. As illustrated in Figure 4, the peak-picking process normally reduces the size of the data from 130K data points to 300 data points. After the peak-picking process, the number of analytes is reduced from the original number of 153M to 353K. These selected peaks are then merged into a single data file and screened in the filtering stage.

D. Filtering Stage

After completing the previous stage, the 300 peaks from each input file are merged into one single peak list file. The peak width is defined as 0.01 Da, in order to filter out small experimental variation for different sample runs, where the same analyte from different data files are assigned to slightly different mass-to-charge value. As a result, analyte peaks within 0.01 *Da* are all assigned to the same mass-to-charge value, which is associated with the highest intensity within this mass range. The remaining analytes are sorted again by its intensity so that only the top 5000 chemicals with the highest intensities are kept for PCA processing. Python is used to implement the filtering stage. A sparse matrix is generated in the filtering stage, as shown in Figure 4, where a row corresponds to a spectrum data file ID and a column correspond to an analyte. A non-zero element *(i,j)* represents the intensity of an analyte *j* from spectrum data file *i*, while an empty cell *(m,n)* shows that file *m* does not contribute that particular analyte or intensity to the matrix.

E. PCA Stage

The filtering stage generates a 1176x5000 sparse matrix, where an analyte in a certain column could be identified in a urine spectrum file specified by the row number. PCA then is applied to transform the matrix into principal components. Each principal component is a linear combination of the 5000 variables in that row. In most dimensionality reduction problems, the first and second principal components usually capture 70-90% of the variance in the data. Therefore the first two principal components are considered. *scikit-learn*, a machine learning toolbox in Python, is used to carry out PCA analysis on the filtered data [19]. We discuss our preliminary clustering results of the chemicals in next section.

F. Clustering Results Analysis

Among the current 6 urine samples, 5 samples are from Puerto Rico and 1 sample is from Boston. Due to security reasons, each human subject is identified using the particular naming scheme, location plus study id.

The 1176 urine spectra defined in 5000 mass-to-charge feature space is projected onto a two-dimensional plane, where the first two principal components (PC1 and PC2) are used for the x-axis and the y-axis, respectively. The resulting projection is presented in Figure 5, showing some clustering on the right side of the plot. This clustered region is associated with the early and later droplets from the Ultra Performance Liquid Chromatography, which contain common background noise. On the left side of the plot, where higher PC1 variance appears, the red dots (PR3_10_14) and blue dots (PR7027) are grouped in clusters. It implies a possible environmental and metabolomic difference.

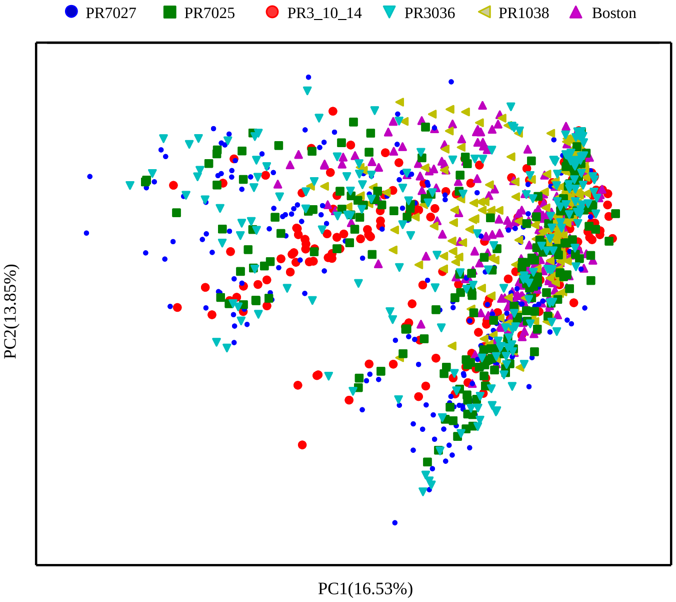

Figure 5. PCA results on six urine samples using PC1 and PC2.

Our current results show that by combining PC1 and PC2, only 30% of the total variation is captured without applying any weighting or scaling. However, if the distribution of these 1176 mass spectra are visulized using other principal components, the variance captured is only smaller with the spectra distributions being similar. Non-negligible differences can be observed using alternative principal components. As additional urine samples are included, more insights are to be learned from this statistical analysis. Our automation process has been able to accelerate urine analysis from 20 mins to 12 seconds by leveraging *MALDIquant* and *scikit-learn*. A 100x speedup has been effectively achieved by this analysis framework, which will enable more extensive analysis of our data sets as the number of participants in our study group grows.

V. Conclusion and Future Work

This paper provided an overview of the data repository developed for the PROTECT Center, and presented a use case analysis of biological sample data maintained in the database system.

Establishing linkages between environmental pollutants and preterm birth can produce important health benefits for both expectant mothers and their babies. In the PROTECT Center, we are collecting detailed information from expectant mothers during their entire 9 month pregnancy, as well as after delivery. The data collected in this study encompasses a wide variety of data types, ranging from soil and water composition to birth outcomes. We expect to be managing billions of data points over the next few years.

In the PROTECT database system, we have developed an efficient framework to handle efficient big data cleaning and entry. We have built our architecture on top of EQuIS Professional to handle the data cleaning and provide online and secure access through EQuIS Enterprise.

To demonstrate the utility of our data, as well as describe our challenges when working with such large data, we present results of a preliminary urine sample study. A sophisticated decoding scheme has been proposed to extract the spectra information from mass spectrometer measurements. With the proposed PCA tools, we are capable of reducing the analysis time by 100-fold, while maintaining the same accuracy as the proprietary software. Since a customized toolset has been developed, we can directly interface the metabolite database to apply data weights. Given that there are approximately 50 associated Mass Spectrometry (MS^2) spectra for each original mass spectrum, the actual data set size we will be working with is much bigger than 1176x136K. From these MS^2 spectra, we can learn more about the detected chemicals. This additional information can then be fed into our PCA analysis, to allow this huge data set to become more manageable.

Even though a significant speedup has been achieved, there is still room for further acceleration. Currently, we are working on a standalone CPU optimization using the open-source software. However, we plan to also leverage the benefits of parallel accelerators, such as Graphics Processing Units. Prior work utilizing accelerators has provided 9x to 130x speedup on a variety of data analysis domains [20]. We are developing a GPU-accelerated PCA implementation to further accelerate our analysis tools.

In our current PCA analysis of the urine samples, the first two dominant principal components captured 30% of the variation of all the variables. As part of our ongoing research, we would like to consider additional principal components to cover more feature variations.

Acknowledgment

This work is supported in part by Award Number P42ES017198 from the National Institute of Environmental Health Sciences. In addition, we would like to thank EarthSoft for their generous support.

References

[1] H. Blencowe et al., “National regional and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications,” The Lancet, vol. 379, 2012, pp. 2162–2171.

[2] J. D. Meeker et al., “Urinary phthalate metabolites in relation to preterm birth in mexico city,” Environ. Health Perspect., vol. 117, 2009, pp. 1587–1592.

[3] D. Cantonwine et al., “Bisphenol a exposure in Mexico City and Risk of prematurity: a pilot nested case control study,” Environ. Health, vol. 9, 2010, pp. 62–68.

[4] A. P. Mucha et al., “Abstract: Association between pbde exposure and preterm birth,” in 10 Annual Workshop on Brominated Flame Retardants, Victoria, BC Canada, 2008, p. 42.

[5] K. Tsukimori et al., “Long-term effects of polychlorinated biphenyls and dioxins on pregnancy outcomes in women affected by the yusho incident,” Eniron. Health. Perspect., vol. 116, 2008, pp. 626–630.

[6] P. Z. Ruckart, F. J. Bove, and M. Maslia, “Evaluation of contaminated drinking water and preterm birth, small for gestational age, and birth weight at Marine Corps Base Camp Lejeune, North Carolina: a cross-sectional study,” Environ. Health, vol. 13, 2014, pp. 1–10.

[7] J. D. Meeker, “Exposure to environmental endocrine disruptors and child development,” Arch. Pediatr. Adolesc. Med., vol. 166, 2012, pp. 952–958.

[8] I. Jolliffe, Principal component analysis. Wiley Online Library, 2005.

[9] J. McCoy, K. Johnston, and E. systems research institute, Using ArcGIS spatial analyst: GIS by ESRI. Environmental Systems Research Institute, 2001.

[10] D. Keckler, “Surfer for windows-users guide.-golden software,” Inc., Golden, CO, 1995.

[11] EarthSoft, “EQuIS Professional,” http://www.earthsoft.com/products/professional/, 2015 (accessed March 1, 2015).

[12] EarthSoft, “EarthSoft: Standalone EQuIS Data Processor (EDP) User Guide,” http://www.dec.ny.gov/docs/remediation_hudson_pdf/edpuserguide.pdf, 2008.

[13] R. Plumb et al., “Ultra-performance liquid chromatography coupled to quadrupole-orthogonal time-of-flight mass spectrometry,” Rapid Communications in Mass Spectrometry, vol. 18, no. 19, 2004, pp. 2331–2337.

[14] M. Sugimoto, M. Kawakami, M. Robert, T. Soga, and M. Tomita, “Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis,” Curr. Bioinform., vol. 7, 2012, pp. 96–4108.

[15] Y. Gao, “T2D converter,” http://www.pepchem.org/, 2013 (accessed March 1, 2015).

[16] D. Kessner, M. Chambers, R. Burke, D. Agus, and P. Mallick, “Proteowizard: open source software for rapid proteomics tools development,” Bioinformatics, vol. 24, no. 21, 2008, pp. 2534–2536.

[17] S. Gibb and K. Strimmer, “Maldiquant: a versatile r package for the analysis of mass spectrometry data,” Bioinformatics, vol. 28, no. 17, 2012, pp. 2270–2271.

[18] J. Tuimala and A. Kallio, “R, programming language,” Encyclopedia of Systems Biology, 2013, pp. 1809–1811.

[19] F. Pedregosa et al., “Scikit-learn: Machine learning in python,” The Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.

[20] J. Nickolls and W. J. Dally, “The gpu computing era,” IEEE micro, vol. 30, no. 2, 2010, pp. 56–69.

ISBN: 978-1-61208-445-9