CREST: Center for Aquatic Chemistry & the Environment (CACE)

Subproject 3: Data Analytics for Effects Assessment and Decision Making

Project Summary

During the past three decades, incidents involving pesticides, industrial chemicals, oil, pharmaceuticals, nutrients and metals have attracted worldwide attention and greatly affected environmental conditions (e.g., the Gulf of Mexico Deep-water Horizon Oil spill). These events demonstrate a regional, national and international need for enhanced research on the effects of toxic substances in the environment. The proposed CREST Center for Aquatic Chemistry & the Environment (CACE) at Florida International University (FIU) will transform the institution by integrating discrete campus-wide programs across 10 departments and 4 colleges in fields from environmental chemistry through computer intensive data analysis and visualization, in order to tackle one of the regions most complex challenges: **environmental contamination**. CACE will create innovative opportunities for students, especially encouraging those from underrepresented minorities (URM), to participate in authentic research and foster their development as future STEM professionals. FIU CACE will unify this talented pool of researchers into a cohesive Center that will enhance collaborations, partnerships and synergies. The Center will bridge academic programs that exist across campuses by integrating graduate and undergraduate students into all research subprojects, emphasizing evidence-based educational approaches, technology advances, and analytical chemistry infrastructure, while providing authentic research experiences and solutions. CACE will transform cutting-edge research into technological and science-based solutions for various forms of water contamination using a framework that includes detection/identification, transport and fate in complex ecosystems, and data analytics and visualization. CACE will develop a modeling platform that will enable policy makers and managers to make informed decisions. FIU's CACE will work in collaboration with governmental and private sector partners in S. Florida to develop practical solutions to problems related to water quality in a natural-agricultural-urban setting. This partnership includes the South Florida Management District, the National Park Service, The Miccosukee Tribe of Indians, the Environmental Protection Agency, Everglades National Park, Department of Interior, and others.

Intellectual Merit

FIU CREST CACE will increase opportunities for graduate and undergraduate students, especially encouraging those from URMs, to conduct authentic research while advancing aquatic and environmental chemistry research and data analytics, methodologies, ecological risk assessments. CACE will generate significant new knowledge regarding contaminants and pollutants in aquatic environments, as well as produce innovative new methodologies for detecting and assessing contaminant quantities and impacts, including the use of molecular detection techniques. Using new data analytic approaches for visualization and synthesis of complex data, CACE will provide managers and policy makers, including governmental and private sector partners in S. Florida, real-time, accessible decision tools. The proposed program will advance current efforts on the biological effects, transport, transformation and distribution of contaminants in the environment into new collaborative research areas that investigate the sources and transport of contaminants and pollutants in aquatic systems. The research conducted by the Center will inform the economic, environmental, societal, policy, regulatory, and legal implications of water quality issues.

Broader Impacts

CACE will build on the success of FIU's evidence-based transformation of STEM instructional practices to provide enhanced support for students to pursue and complete STEM graduate degrees, both at FIU or elsewhere. Through an innovative program that spans the graduate school to high school spectrum, CACE will increase the success of students in graduate programs, especially supporting participation of underrepresented students in aquatic chemistry and environment (ACE) fields and future professions. CACE will develop technologies for improving water quality analysis and contaminant detection, as well as translate research findings into actionable information for decision-makers and stakeholders. By providing potential scenarios for understanding the risks, sources, transport and impacts of chemical contaminants that threaten aquatic ecosystems and human wellbeing, CACE can impact global water quality.

Subproject Relevancy Statement

Subproject 3: Data Analytics for Effects Assessment and Decision Making

Technological advances in hardware, storage, and software have significantly increased scientists' ability to find new ways of generating and using data, thus creating new possibilities for conducting research to address the complex challenges of environmental contamination and ecological risk assessment. Conducting scientific research through high-resolution data acquisition, data mining, and visualization enables scientists to better understand the transient nature of aquatic data and pollutant movement across various boundaries.

However, data intensive science requires significant collaboration between environmental and computer scientists. This collaboration is becoming increasingly critical in finding better and more effective ways to research, discover and solve problems. The research conducted by this Subproject will facilitate such collaboration and support CACE researchers at the proposed CREST Center to better detect and understand the sources, transport, transformation and ecosystem responses to contaminants, pollutants and other natural stressors in the aquatic systems of south Florida.

Using a data-intensive approach, CACE researchers will be able to: 1) provide detailed characterization and measurement of the environmental pollutants, 2) improve predictive abilities on effects of pollutants and address future water quality issues, 3) explore, manipulate and visualize data thus collaborate more effectively for risk assessment, 4) conduct literature mining on the nature of contaminants and access relevant environmental information rapidly, and 5) communicate more effectively with decision makers and other stakeholders. **The ultimate goal of this Subproject is to support data-intensive research on aquatic chemistry and the environment by developing transformative and scalable methods for data mining and management, advanced computational modeling, and visualization**. The table below is a summary of how this Subproject relates to the two other Subprojects of the proposed CREST Center.

Links between CACE subprojects	**Subproject 1: Advanced Sensing of Environmental Exposure to Anthropogenic Contaminants, Pollutants and Other Natural Stressors**	**Subproject 2: Quantifying the Fate and Transport of Contaminants across Natural, Agricultural and Human Systems**
Subproject 3: Data Analytics for Effects Assessment and Decision Making	• Create novel multi-tiered data analysis architecture, consisting of sensors and cloud/HPC computing systems • Provide mining capability investigation by utilizing associations and correlations among the data to understand the characteristics and to extract semantics and patterns from the data • Provide techniques for managing complex analogue environmental and digital molecular biology digital data • Provide synthesis and analysis of gene function networks • Create data visualization and decision making support tools	• Provide mining capability investigation by utilizing associations and correlations among the data to understand the characteristics and to extract semantics and patterns from the data • Provide data analysis support for quantifying and trend identification of current and historical sources and loading of pollutants, • Develop appropriate visualization tools to examine specific plausible and realistic scenarios for future changes in water and land resource management • Provide capacity for literature mining of biological and visual analytics and visualization algorithms to assist in assessment and strategic decision-making

Subproject 3: Data Analytics for Effects Assessment and Decision Making

1. Introduction

Computational research and tools developed under this Subproject are designed to support the CREST Center's team efforts in identification of source, transport, transformation and ecosystem responses to contaminants, pollutants and other natural stressors. Researching environmental contamination and ecological risk assessment requires investigation of non-chemical, as well as organic and inorganic chemical stressors including **nutrients, contaminants and pollutants** with a multitude of exposure types (e.g., single-slug, intermittent and continuous) with native, exotic and standard test species.

Conducting this research entails collection of large volumes of data from various heterogeneous sources such as data from analytical chemistry techniques and data from biogeochemical cycles used to determine how natural processes affect ecosystems. As the scale and complexity of these data types increase exponentially, it becomes challenging to effectively model the increasing volumes of data, discover useful information, and provide data analytics capability to support effective and accurate assessment and decision-making capability for the scientists and their partners.

To address these challenges, the CREST center will provide a suite of data analytics algorithms, including computation modeling, data mining, and visualization tools. The computation-modeling component provides the computational and system support for diverse data analytics and decision-making tasks based on novel multi-tiered data analysis architecture. As data analysis tasks are computationally intensive, we will address system and architecture issues related to computational requirements for data gathering, data analysis, and decision-making (Liu et al., 2014; Ren and van der Schaar, 2013; Xu et al., 2012; Xu and Shatz, 2003).

Two areas of computing research will be developed. First area will focus on creating novel multi-tiered data analysis architecture, consisting of sensors and cloud/HPC computing systems. Second area will be to support data mining and visualization research components. The novelty of this architecture is in allowing data analysis to be conducted at different granularities and satisfy different timing requirements (from near-real-time, global static data analysis to real-time, local dynamic data analysis).

The data mining component proposes a comprehensive investigation on utilizing associations and correlations among the data to understand the characteristics and to extract semantics and patterns from the data (Chen et al., 2007; Ha et al., 2013; Ha et al., 2015; Lin et al., 2012; T. Meng and M.-L. Shyu, 2013; Shyu et al., 2005; Thompson 2005; Yang et al., 2014). This component will also develop the dimension reduction and information fusion algorithms to address the scalability and multi-source issues (Gehler and Nowozin, 2009; Ha et al., 2013; Liu et al., 2009; Yu and Liu, 2003).

Once large amounts of heterogeneous data are processed by the computation modeling component, it will be ready for easy access to useful information and pattern extraction that can assist quick evaluation and assessment of contaminants transport and fate. This will be support a collaborative effort among various stakeholders including scientists, local government and industry partners to create plausible and realistic scenarios for evaluating the risk and possible course of action for the South Florida Region.

The visualization component focuses on developing visual analytics and visualization algorithms to assist in assessment and strategic decision-making. This research will explore novel ways of displaying information visually, aggregating existing techniques into visual ensembles that are tailored for solving specific problems and providing the interactive means for users to work with these systems in their specific context (Li et al., 2009; Saleem et al., 2007; Zhang et al., 2006). In particular, we would like to create novel visualization and visual analytics methods for displaying complex data types, data aggregates, and analytic concepts at the border between humans and computing.

2. Research Plan

Our research plan is focused on developing new methodologies to support detection and evaluation of trends, analyzing contaminant transport, and creating visualization tools for querying data that allows for early intervention and restoration of the water and the ecosystems of South Florida. This plan will be conducted in four research thrusts as described below.

2.1 Multidimensional Data Analytics of Environmental and Molecular Biology Information

The research conducted by scientists in Subproject 1 requires characterization and measurement of a myriad of stressors associated with urban and agricultural landscapes. Using advanced methodologies they will collect vast amounts of data to determine the environmental exposure from trace analysis of critical pollutants such as nutrients, trace metals, DDT and PCBs to other biologically active compounds such as antibiotics and pharmaceuticals (e.g. endocrine disrupters), mercury, black carbon and fossil fuels (oil). These data will be of two types: 1) environmental data that are analog parameters; and 2) molecular biology information that have a digital signal.

The environmental data have identity (parameter), intensity above a threshold (signal or concentration), and potentially an environmental limitation benchmark as well as toxicological indicators. The parameter list could be as large as 100 to 200 items. The molecular biology data will be much larger and will include a "gene identity" (related to a function i.e., gene responsible for metal detoxification); and whether exposure to environmental stressors caused the gene to be under-expressed (-1), over-expressed (+1), or showed no change (0) when compared to the untreated group. These data could rank in the thousands depending on the generating methods. All these data can be represented and expanded on in a matrix such as the one shown below:

Sample ID	Temperature	Concentration	Toxicity effect	Location	Gene Signal	Epigenomic Markers	Other Parameters …
A1	Analog	Analog	Analog	Analog	Digital	Digital	Digital
A2	Analog	Analog	Analog	Analog	Digital	Digital	Digital
…	Analog	Analog	Analog	Analog	Digital	Digital	Digital
B1	Analog	Analog	Analog	Analog	Digital	Digital	Digital
…	Analog	Analog	Analog	Analog	Digital	Digital	Digital
C1	Analog	Analog	Analog	Analog	Digital	Digital	Digital

In Subproject 2, CACE scientists will examine the biogeochemical processes that govern the ultimate fate of these pollutants and their impacts on the environment**.** They will establish three transects that encompass the main transitions between agriculture, urban and natural landscapes. These transects will provide a common platform for detection and measurement for Subproject 1, sample water quality with a common experimental design, and application of advanced modeling techniques to couple flows with contaminants. The generated data from this process will provide a third dimension to the matrix provided above to inform our data analytics research.

Low-rank matrix factorization:

The matrices generated from Subprojects 1 and 2 could involve low-dimensional structures (e.g., sparse or low-rank). In particular, the numbers of samples in these cases are typically far less than the total number of degrees of freedom (i.e. determined by the analog parameters and molecular biology information). The goal of our analysis is to understand the relationships between the samples and the parameters and the relationships among various parameters. Factor analysis (Child, 2006; Mulaik, 1972) or topic modeling (Blei et al., 2003) can be used to establish the relationships. In this project, we will jointly perform factor analysis and topic modeling while exploiting the low-dimensional matrix structure.

Compute the densely connected components:

In genomics, given a protein interaction network, it is often useful to compute the densely connected components as protein interaction modules. In these cases, the input data is the adjacency matrix *A* of an undirected graph with weights in {0;1}. We can formulate the problem as computing maximal cliques, although the rigorous definition of a clique is often unnecessary. We propose to solve the following optimization problem:

$$\max_X X^T AX, \;\; s.t. \; \sum_{i=1}^{n} x_i^{\alpha} = 1; \; x_i > 0$$

where $\alpha \in [1,2]$ is a parameter (Ding et al., 2008). The nonzero entries in the solution vector correspond to the vertices of the densely connected component we are seeking. It can be shown that: (1) a maximal clique is obtained when $\alpha = 1 + \varepsilon,\ 0 < \epsilon < 1$, while setting *Aii* = 1 (this enables us to generalize this approach to bipartite graphs). (2) When $\alpha = 1$, this formulation reduces to the Motzkin and Strauss formulation (Motzkin & Straus, 1965), where *Aii* = 0 is required. (3) As α goes close to 1, the sparsity of the solution increases steadily, reflecting the close relation between L_1 constraints and sparsity. At α = 2, the solution is given by the principal eigenvector of *A*.

2.2 Synthesis and Analysis of Gene Function Networks

Computational methods for gene functional prediction fall into two categories: direct annotation schemes, which infer the function based on the functional annotations of genes in its neighborhood in the network, and module-assisted schemes, which first identify modules of related genes and then annotate each module based on the known functions of its members. In this project, we are interested in computation methods of the second category (i.e., module-based methods). The key step of methods falling in this category is to identify biologically meaningful functional modules. Cluster analysis is a popular methodology for the extraction of function modules from genes and protein interaction networks since it has been observed by biologists that groups of highly interacting proteins could be involved in common biological processes (Spirin and Mirny, 2003).

However, the special characteristics of the data obtained from high-throughput experiments make the clustering task for identifying the function modules very challenging. These challenges include: (a) Poor data quality: The data obtained from the high-throughput experiments are quite noisy and contain many false positives. (b) Specific topological and network properties: The network structure from the data has been observed to have high clustering coefficients and modularity (Yook et al., 2004; Jeong et al., 2001; Ravasz et al., 2002).

A few genes/proteins in the network may have very large degrees while most others only have very few interactions. Clustering algorithms in this context need to pay attention to these topological and network properties. (c) Huge Volume: The datasets obtained from the experiments are of a large volume including tens of thousands of interactions among thousands of proteins even for a unicellular eukaryotic organism. Hence, clustering algorithms need to be fast and scalable. (d) Multi-functional: A gene/protein is often multi-functional and often involved with multiple modules. Hence clustering algorithms need to support “soft assignments”, i.e., assigning a protein into multiple groups.

Clustering algorithms for extracting function modules should address these challenges. Despite the significant progress that has been made in the area, existing clustering methods for extracting function modules are far from satisfactory due to the presence of noisy false positives, specific topological challenges, and the huge amount of data (Asur et al., 2007; Jaimovich et al., 2006). In addition, most of the existing methods do not support the “soft assignment” as they assign each gene/protein into a specific group.

In this project, we propose to develop ensemble-clustering methods for combining multiple, diverse and independent clustering results to improve the quality and robustness of identification. Different base clustering algorithms (e.g., spectral clustering algorithms and graph clustering algorithms) might have their own strengths and limitations. Ensemble clustering offers an appealing framework for taking advantage of the strengths of individual clustering algorithms and for improving the quality of identification.

<u>Base Clustering:</u>

In addition to the conventional clustering algorithms, e.g., network motifs, local cluster growing, graph-theoretic, and hierarchical clustering (King et al., 2004; Brun et al., 2004; Arnau et al., 2005; Dunn et al, 2005; Enright et al., 2002), we will also explore the use of spectral clustering-algorithms with diverse yet informative topological and graph properties (e.g. edge-betweenness and clustering coefficients) as base clustering algorithms. Spectral clustering algorithms have well-motivated objective functions that can easily incorporate the graph properties and can be computed efficiently using mature scientific computing software tools.

A spectral clustering algorithm is obtained by recursively applying a spectral method for graph partitioning (Shi and Malik, 2000; Dhillon et al., 2007). Let *Q* denote the Laplacian matrix of a graph with weights w_{ij} on its edges $<i; j>$: the diagonal elements q_{ii} of Q is the sum of the weights of the edges incident on the vertex i, and for other elements, $q_{ij} = w_{ij}$. The partition problem is first modeled as the minimization of a quadratic program p^TQp over all partition vectors p, whose elements are either 1 or -1. The integer constraints can be relaxed and we can then solve the continuous version of the optimization problem over real vectors with components bounded in the interval [-1;+1]. The solution to the continuous optimization problem is obtained by computing the eigenvalues or singular values of p^TQp. The eigenvector components provide a natural soft assignment since the values in the components reflect the degree of association between the vertices and the clusters.

Ensemble Clustering:

Ensemble clustering, also called aggregation of clustering, refers to the situation in which a number of different clustering results have been obtained for a particular dataset and it is desired to find a single (combined or consensus) clustering which is a better fit in some sense than the existing clustering results (Hu et al., 2006; Gionis et al., 2005). Empirical evidence has suggested that ensemble clustering can improve clustering robustness and discover useful cluster structures even if the data is quite noisy (Topchy et al., 2004).

However, there is a significant drawback in current ensemble clustering approaches (Strehl and Ghosh, 2002; Asur et al., 2007), i.e., all input clustering solutions are treated equally, despite the facts that: (1) different input clustering results could differ significantly, and (2) subsets of input clustering results could be highly correlated. As a result, when collecting a large number of input clustering results, quite often many clustering results could be close (similar) to each other. These would easily skew the final consensus clustering. Hence, simply applying current ensemble clustering for extracting protein function modules is inadequate.

In this Subproject, we propose the weighted ensemble clustering for extracting protein function modules. In (Li et al., 2007), we show that the ensemble-clustering problem can be efficiently solved within the nonnegative matrix factorization framework. Building on our previous work, the weighted ensemble clustering can also be formulated as an optimization problem. In weighed ensemble clustering, different input clustering results weigh differently, i.e., a weight for each input clustering is introduced, but the weights are automatically determined by an optimization process similar to a kernel matrix learning (Lanckriet et al., 2006). It should also be noted that the weights obtained in the weighted ensemble clustering could be useful for selecting input clustering. Clearly, an input clustering with larger weight contributes more to the final consensus clustering.

Heterogeneous Data Integration:

The data from heterogeneous data sources (e.g., gene expression and protein interactions) are useful for inferring gene functions (Bhardwaj and Lu, 2005; Jansen et al., 2002; Tu et al., 2006, Wang et al., 2012). Despite previous efforts in the integration of heterogeneous data, there is still a lot of room for improvements since the information enriched in each biological source has not been fully utilized.

The integration of different types of experimental data into an overall model is a critical and challenging task because of the vast difference in data type, dimension and quality (Shannon et al., 2003; Cline et al. 2007; Camargo and Azuaje, 2007). Two major problems must be addressed in order to integrate the heterogeneous data sources/types and extract the optimal conclusions from the combined data: (a) Data types must be unified or scaled in order to allow comparison and combination. For example, gene expression data is continuous and relative in nature while protein interaction data is pairwise and binary; (b) the data must be weighted or verified in a quantitative and consistent manner.

In this project, we will use the following three approaches for data integration.

1. Feature Integration: This approach enlarges the feature representation to incorporate all data and produces a unified feature space. In particular, continuous data types will be converted into discrete levels and categorical data type will be mapped into similar discrete levels. The data are then transformed into the same feature space and standard computational methods, such as prediction and clustering, can be performed. The advantage of feature integration is that the unified feature representation is often more informative and also allows many different data

mining methods to be applied and systematically compared. One disadvantage is the increased learning complexity and difficulty as the data dimension becomes large.

2. Semantic Integration: This approach keeps data intact in their separate original form. Computational methods are applied to each dataset separately. Results on different datasets are then combined by either voting (Carter et al., 2001), Bayesian averaging (Bishop, 2006), or the hierarchical expert system approach (Jordan and Jacobs, 2004). This approach seems to work reasonably well. One advantage of semantic integration is that it can implicitly learn the correlation structure between different sets of features (Li and Ogihara, 2005).

3. Intermediate Integration: This approach can be viewed as a compromise between the feature-level integration and the semantic integration. The data is kept in their original form and they are integrated at the similarity computation or the Kernel level (Lanckriet et al., 2006). For example, for protein p_i and p_j, their total pairwise similarity or affinity is $S_{ij} = A_{ij} + B_{ij}$, where A_{ij} is computed from gene expression profiles and B_{ij} is obtained from protein interaction. Standard computational methods can then be applied once the total similarity is computed.

We will carefully compare these integration methods in this project and explore their trade-offs through the design of suitable experiments.

2.3 Literature Mining and Curating Biological and Environmental Information

The biological and environmental literature databases provide knowledge warehouses to cross-reference experimental and analytical results with previously known biological facts, theories, and results. They can also be used to identify function commonalities of genes. In this Subproject, we propose to incorporate expert genetic knowledge for function discovery, instead of relying on purely empirical methods. Each domain expert only knows a few objects well. Hence relating the measurements in observation data with existing knowledge is a key part for data analysis. Mining on observation data alone may not be able to reveal the biological information and the pollutants impact. On the other hand, references on the literature will provide additional information. We will facilitate using text literature as a guide for detection, identification, and effect of chemical stressors in the ecosystem.

There are many literature databases publicly available. One good source is KEGG (Kyoto Encyclopedia of Genes and Genomes). This is a particularly high-quality data source, as it is curated by a knowledgeable team based on reported information in the scientific literature and is continuously updated. Text mining techniques can be applied to provide descriptive information from the literature.

There are two steps when performing text mining on a set of literature documents: (1) Document Pre-processing; (2) identification of text summaries with observation data. Document pre-process includes stripping unwanted characters/markup (e.g, HTML tags, punctuation, numbers, etc.), removing common stop words (e.g. a, the, it, etc.) and stemming keywords into “root” words etc. developing a synonyms list. Note that there are many words and phrases that refer to the same entity, hence a synonyms list will also be developed in pre-processing. After document preprocessing, step (2) is to thus provide meaningful textual summaries with the information that domain experts may be interested in. Therefore, techniques to perform text summarization will be studied (Mittal et al., 2000; Lin & Hovy, 2002).

In this project, we will investigate keyword search based algorithm and sentence extraction for literature summarization (Kankar et al., 2002; Masys et al., 2001; Jurafsky and Martin, 2008).

After we get the literature description, the remaining question is how we should combine the literature information with observation information. There are some existing approaches such as MedMiner (Tanabe et al., 1999) and PubGene (Jenssen et al., 2001). Medminer first performs clustering on observation data and then interprets textually while PubGene first performs clustering on textual data and then interprets numerically). These two techniques will be studied initially. However, both types of approaches ignore the correlation structure between different sources. We discuss the integration of different data sources in a separate section.

2.4 Data Visualization and Decision Making Support

Visualization of the dynamic variables affecting the hydrology, fate and contaminant transport in ecosystems of south Florida can be critical in understanding, identifying and acting upon valuable information produced by Subprojects 1 and 2. First, visualization can help the scientists in both groups to understand and interpret patterns in the data (Goldstein et al., 1994; Ward, 1994). For example, a scatter plot can help to identify patterns of significance from a large amount of monitoring data. To display categorical data in the matrices, the categorical values need to be mapped into numerical values. One challenge is to choose an effective mapping, as a random order may not be effective, as it tends to spread the data across the visual space. We will combine clustering and dimensionality reduction techniques for better visualization.

Second, by developing event relationship networks, a graphical representation of event correlation (Burns et al., 2001), we will enable the domain experts to easily review and understand information. Formally, an event relationship network is a directed graph where the vertices are events and the edges indicate the dependence relationships among the events. In addition, it also serves as a concise representation of the domain knowledge. We propose to develop techniques to construct, validate and complete event relationship networks using the discovered temporal patterns.

Finally visual data analysis, facilitated by interactive interfaces, enables the detection and validation of expected results while enabling unexpected discoveries in science (Hansen et al., 2008) The scientists in Subproject 1 will develop scenarios for evaluating the impact of water and land resources management decisions on the hydrology to determine the transport of contaminates and eventual ecological vulnerabilities. Our research and development of visualization tools, will support these scientists and decision makers to explore “what if” scenarios, define hypotheses, and examine data using multiple perspectives and assumptions on fate and transport of the contaminants (Hansen et al., 2008). Utilizing our research and tools they can identify coherent patterns and assess the reliability of their assumptions.

We will develop tools to visualize information including condition indexes, ecosystem maps, and genomic responses maps. We will design and develop techniques to bridge the gap between the application and intelligent techniques. Specifically, we will develop an interactive tool that can present patterns visually, in a way that is intuitive and easily understandable for the users. In addition, we will use the discovered patterns to evaluate and validate the relationships among samples and observations.

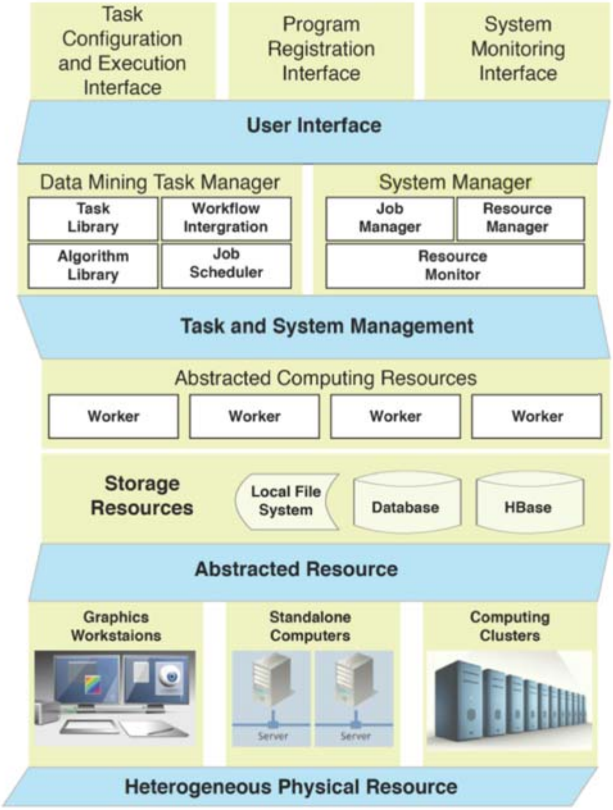

Figure 1: FIU-Miner System Architecture

Software framework for synthesizing decision recommendations:

We propose to develop a framework for synthesizing decision recommendations to aid decision makers. This framework will leverage FIU-Miner: A Fast, Integrated, and User-Friendly System for Data Mining in Distributed Environment to develop system components (Zeng et al., 2013). FIU-Miner allows users to rapidly configure a complex data analysis task without writing a single line of code. It also helps users conveniently import and integrate different analysis programs. Figure 1 shows the system architecture of FIU-Miner.

3. Broader Impacts

This Subproject addresses one of the significant scientific and engineering challenges by enabling a diverse group of environmental scientists to understand and make sense of "big data" ecological information. By developing new computational methodologies to support detection of pollutants, evaluation of trends, analyzing contaminant transport, and visualization of transient aquatic data, scientists and stakeholders can engage in early intervention and restoration of water in order to help the public live safely in their environment. The Subproject's literature mining research not only will help scientists in the other Subprojects more readily link their findings to those made by others, but also will facilitate the work of other environmental scientists in using text literature as tool for their research.

In addition, the computing research findings of FIU’s CREST Center will be applicable to other scientific fields, as we will develop a novel multi-tiered data analysis architecture through data mining and visualization research. Developing this research is significantly important, as science is becoming increasingly data intensive with heavy reliance on using large datasets and visualization for discovery and problem solving. The developed software tools by this Subproject will be released to the open source community for further development and dissemination.

Subproject 3 References

Arnau, V., S. Mars, I. Marin. Iterative cluster analysis of protein interaction data. *Bioinformatics*, 21(3):364.378, 2005.

Asur, S., D. Ucar, S. Parthasarathy. An ensemble framework for clustering protein-protein interaction networks. In *Proceedings of the 15th Annual International Conference on Intelligent Systems (ISMB)*, 2007.

Bhardwaj, N., H. Lu. Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics*, 21(11):2730.2738, 2005.

Bishop, C.M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Blei, D.M., A.Y. Ng, M.I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, Vo. 3, PP. 993-1022, Jan. 2003.

Brohee, S., J. van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7:488, 2006.

Brun, C., C. Herrmann, A. Gunoche. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics*, 5:95, 2004.

Burns, L., J. L. Hellerstein, S. Ma, C.-S Perng, D.A. Rabenhorst, D. Tayler. A systematic approach to discovering correlation rules for event management. In International Symposium on Integrated Network Management, 2001.

Camargo, A. F. Azuaje. Linking gene expression and functional network data in human heart failure. *PLoS ONE*, 2(12):e1347, 2007.

Carter, R., I. Dubchak, S. Holbrook. A computational approach to identify genes for functional rnas in genomic sequences. *Nucl. Acids Res*, 29:3928.3938, 2001.

Chen, M., S.-C. Chen, M.-L. Shyu. Hierarchical temporal association mining for video event detection in video databases. In Proceedings of the Second IEEE International Workshop on Multimedia Databases and Data Management, in conjunction with IEEE International Conference on Data Engineering, pp. 137-145, Istanbul, Turkey, April 15, 2007.

Child, D., The essentials of factor analysis, 3rd ed. New York, NY：Continuum Intl Pub Group, 2006.

Cline, M.S., M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, et al. Integration of biological networks and gene expression data using cytoscape. *Nat Protoc.*, 2(10):2366.2382, 2007.

Dhillon, I.S., Y. Guan, S. Kulis. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(11):1944.1957, 2007.

Ding, C., T. Li, M.I. Jordan. Nonnegative Matrix Factorization for Combinatorial Optimization: Spectral Clustering, Graph Matching, and Clique Finding. In Proceedings of 2008 IEEE International Conference on Data Mining (ICDM 2008), Pages 183-192, 2008.

Dunn, R., F. Dudbridge, C.M. Sanderson. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, 6:39, 2005.

Enright, A.J., S.V. Dongen, C.A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7):1575.1584, 2002.

GehlerP., S. Nowozin, “On Feature Combination for Multiclass Object Classification,” IEEE 12th International Conference on Computer Vision, 2009.

Gionis, A., H. Mannila, P. Tsaparas. Clustering aggregation. In *ICDE*, pages 341.352, 2005.

Goldstein, J., S.F. Roth, J. Kolojejchick, J. Mattis. A framework for knowledge-based, interactive data exploration. Journal of visual languages and computing, 5:339–363, 1994.

Ha, H.-Y., F. C. Fleites, S.-C. Chen, “Content-Based Multimedia Retrieval Using Feature Correlation Clustering and Fusion,” International Journal of Multimedia Data Engineering and Management (IJMDEM), Volume 4, No. 2, pp. 46-64, 2013.

Ha, H.-Y., S.-C. Chen, M. Chen, “FC-MST: Feature Correlation Maximum Spanning Tree for Multimedia Concept Classification,” Ninth IEEE International Conference on Semantic Computing, Anaheim, California, USA, pp. 276-283, February 7-9, 2015.

Hansen, C., C. Johnson, V. Pascuuchi, S. Claudio (2009). Visualization for Data-Intensive Science in The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, Redmond, Washington

Hu, X., I. Yoo, X. Zhang, P. Nanavati, D. Das. Wavelet transformation and cluster ensemble for gene expression analysis. *International Journal of Bioinformatics Research and Application*, 1(4):447.460, 2006

Jaimovich, A., G. Elidan, H. Margalit, N. Friedman. (2006). Towards an integrated protein-protein interaction network: a relational Markov network approach. Journal of Computational Biology. 13(2):145-64.

Jansen, R., D. Greenbaum, M. Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res.*, 12:37.46, 2002.

Jenssen, T.-K., A. Lagreid, J. Komorowski, E. Hovig (2001), A Literature network of human genes for high-throughput analysis of gene expression. Nature Genetics, 2001 28: 21-28.

Jeong, H., S.P. Mason, A.L. A.L. Barabasi, Z.N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41.42, 2001.

Jordan, M.I., R.A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181.214, 1994.

Jurafsky, D., J. H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson & Prentice Hall, 2008.

Kankar, P., S. Adak, A. Sarkar, K. Murari G. Sharma (2002), Medmesh Summarizer: Text Mining For Gene Clusters. In Proceedings of the Second SIAM International Conference on Data Mining, 2002.

King, A.D., N. Przulj, I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013.3020, 2004.

Lanckriet, G.R., N. Cristianini, P.L. Bartlett, L.E. Ghaoui, M.I. Jordan. Learning the kernel matrix with semi-de_nite programming. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 323.330, 2006.

Li, T., C. Ding, M. I. Jordan. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Proceedings of 2007 IEEE International Conference on Data Mining (ICDM 2007)*, 2007.

Li, T., M. Ogihara (2005). Semi-supervised learning from different information sources. Knowl. Inf. Syst. 7(3): 289-309.

Li, Y., K. Chatterjee, S.-C. Chen, K. Zhang, “A 3-D Traffic Animation System with Storm Surge Response,” IEEE International Symposium on Multimedia, San Diego, California, USA, pp. 257-262, December 14-16, 2009.

Lin, C.-Y., E. Hovy. From single to multi-document summarization: a prototype system and its evaluation. In ACL 02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002.

Lin, L., M.-L. Shyu, S.-C. Chen, “Association Rule Mining with a Correlation-based Interestingness Measure for Video Semantic Concept Detection," International Journal of Information and Decision Sciences*,* Vol. 4, Nos. 2/3, pp. 199-216, 2012.

Liu, D., S. Hua, Z. Ou, “IR and Visible-light Face Recognition using Canonical Correlation Analysis,” Journal of Computational Information Systems, 5(1), pp. 291-297, 2009.

Liu, J., Y. Liu, Z. Du, T. Li, “GPU-Assisted Hybrid Network Traffic Model,” 2014 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation, Denver, Colorado, May 18-21, 2014.

Masys, D.R., J.B. Welsh, J.L. Fink, M. Gribskov, I. Klacansky J. Corbell (2001), Use Of Keywords Hierarchies To Interpret Gene Expression Patterns. Bioinformatics, 17(4): 319-326.

Meng, T., M.-L. Shyu, “Concept-Concept Association Information Integration and Multi-Model Collaboration for Multimedia Semantic Concept Detection,” International Journal of Information Systems Frontiers*,* pp. 1-13, April 2013.

Mittal, V., J. Carbonell Goldstein, J. M. Kantrowitz. Multi-document summa- rization by sentenceextraction. In NAACL-ANLP 2000 Workshop on Automatic summarization, 2000.

Motzkin, T.S., E.G. Straus. Maxima for graphs and a new proof of a theorem of turan. Canad. J. Math., 17:533-540, 1965.

S. Mulaik, S., The foundations of factor analysis. McGraw-Hill New York, 1972.

Ravasz, E.E., A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551.1555, 2002.

Ren, S., M. van der Schaar, “Efficient Resource Provisioning and Rate Allocation for Stream Mining in a Community Cloud,” IEEE Transactions on Multimedia, vol. 15, no. 4, pp. 723-734, Jun. 2013.

Saleem, K., S.-C. Chen, K. Zhang, “Animating Tree Branch Breaking and Flying Effects for a 3D Interactive Visualization System for Hurricanes and Storm Surge Flooding,” the Third IEEE International Workshop on Multimedia Information Processing and Retrieval, in conjunction with IEEE International Symposium on Multimedia, pp. 335-340, Taichung, Taiwan, R.O.C. , December 10-12, 2007.

Shannon, P., A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498.2504, 2003.

Shi, J., J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888.905, 2000.

Shyu, M.-L., I.P. Kuruppu-Appuhamilage, S.-C. Chen, L. Chang. Handling missing values via decomposition of the conditioned set. In Proceedings of the 2005 IEEE International Conference on Information Reuse and Integration, pp. 199-204, August 15-17, 2005, Las Vegas, Nevada, USA.

Spirin, V., L.A. Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, 100:12123.12128, 2003.

Strehl, A., J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research (JMLR)*, 3:583.617, December 2002.

Tanabe, L., L.H. Smith, J.K. Lee, U. Scherf, L. Hunter, J.N. Weinstein (1999). MedMiner: An internet tool for filtering and organizing biomedical information, with application to gene expression profiling. BioTechniques. 1999; 27: 1210-1217.

Thompson, B., “Canonical Correlation Analysis,” Encyclopedia of statistics in behavioral science, 2005.

Topchy, A.P., M. Law, A.K. Jain, A.L. Fred. Analysis of consensus partition in cluster ensemble. In *Proceedings of International Conference on Data Mining*, pages 225.232, 2004.

Tu, K., H. Yu, Y.X. Li. Combining gene expression profiles and protein-protein interaction data to infer gene functions. *Journal of Biotechnology*, 124:475.485, 2006.

Wang, D., M. Ogihara, E. Zeng, T. Li. Combining Gene Expression Profiles and Protein-Protein Interactions for Identifying Functional Modules, In Proceedings of 11th International Conference on Machine Learning and Applications (ICMLA 2012), pages 114-119, 2012.

Ward, M.O, Xmdvtool: Integrating multiple methods for visualizing multivariate data. In Proceedings of Visualization, 1994.

Xu, H., S. M. Shatz, “A Framework for Model-based Design of Agent-oriented Software,” IEEE Transactions on Software Engineering, pp. 15–30, 2003.

Xu, Y., D. Arteaga, M. Zhao, Y. Liu, R. Figueiredo, S. Seelam, “vPFS: Bandwidth Virtualization of Parallel Storage Systems,” the 28th IEEE Conference on Massive Data Storage, April 2012.

Yang, Y., H.-Y. Ha, F. C. Fleites, S.-C. Chen, “A Multimedia Semantic Retrieval Mobile System Based on Hidden Coherent Feature Groups,” IEEE Multimedia, Volume 21, Number 1, pp. 36-46, January-March, 2014.

Yook, S.H., Z.N. Oltvai, A.L. Barabasi. Functional and topological characterization of protein interaction networks. *Proteomics*, 2004.

Yu, L., H. Liu, “Feature Selection for High-dimensional Data: A Fast Correlation-based Filter Solution,” Twentieth International Conference on Machine Learning, 2003.

Zeng, C., Y. Jiang, L. Zheng, J. Li, L. Li, H. Li, C. Shen, W. Zhou, T. Li, B. Duan, M. Lei, P. Wang. “FIU-Miner: A fast, integrated, and user-friendly system for data mining in distributed environment”. In Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD'13), pages 1506-1509, 2013.

Zhang, K., S.-C. Chen, P. Singh, K. Saleem, N. Zhao, “A 3D Visualization System for Hurricane Storm Surge Flooding,” IEEE Computer Graphics and Applications, Vol. 26, Issue 1, pp. 18-25, Jan.-Feb. 2006.

Zhang, Y., T. Li, “DClusterE: A Framework for Evaluating and Understanding Document Clustering Using Visualization**,”** ACM Transactions on Intelligent Systems and Technology**, 3**(2): 24, 2012.

Zheng, L., T. Li, C. Ding, “Hierarchical Ensemble Clustering,” IEEE International Conference on Data Mining, pp. 1199-1204, 2010.