PERCON: A PERSONAL DIGITAL LIBRARY FOR HETEROGENEOUS DATA MANAGEMENT AND ANALYSIS

A Dissertation

by

SU INN PARK

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Frank M. Shipman, III
Committee Members,	Richard Furuta
	Ricardo Gutierrez-Osuna
	Steven Smith
Head of Department,	Dilma Da Silva

May 2015

Major Subject: Computer Science

ABSTRACT

Systems are needed to support access to and analysis of larger and more heterogeneous scientific datasets. Users need support in the location, organization, analysis, and interpretation of data to support their current activities with appropriate services and tools. We developed PerCon, a data management and analysis environment, to support such use.

PerCon processes and integrates data gathered via queries to existing data providers to create a personal or a small group digital library of data. Users may then search, browse, visualize, annotate, and organize the data as they proceed with analysis and interpretation. Analysis and interpretation in PerCon takes place in a visual workspace in which multiple data visualizations and annotations are placed into spatial arrangements based on the current task. The system watches for patterns in the user's data selection, exploration, and organization, then through mixed-initiative interaction assists users by suggesting potentially relevant data from unexplored data sources. In order to identify relevant data, PerCon builds up various precomputed feature tables of data objects including their metadata (e.g. similarities, distances) and a user interest model to infer the user interest or specific information need. In particular, probabilistic networks in PerCon model user interactions (i.e. event features) and predict the data type of greatest interest through network training. In turn, the most relevant data objects of interest in the inferred data type are identified through a weighted feature computation then recommended to the user.

PerCon's data location and analysis capabilities were evaluated in a controlled study with 24 users. The study participants were asked to locate and analyze heterogeneous weather and river data with and without the visual workspace and mixed-initiative interaction, respectively. Results indicate that the visual workspace facilitated information representation and aided in the identification of relationships between datasets. The system's suggestions encouraged data exploration, leading participants to identify more evidences of correlation among data streams and more potential interactions among weather and river data.

DEDICATION

To my parents.

ACKNOWLEDGEMENTS

Above all, I would like to express my deepest appreciation to my advisor, Dr. Frank M. Shipman III for his guidance and support during my graduate studies. I have the greatest respect for his personality, research enthusiasm, and work ethics.

I would also like to thank my committee members, Dr. Richard Furuta, Dr. Ricardo Gutierrez-Osuna, and Dr. Steven Smith for their guidance and support in improving the quality of my research.

I also want to thank my friends. Hyun-Chul, Hyundoo, Sangwhan, Kisuk, Eunjoo, and Patricia, they were great help to me during my studies. In addition, I have been privileged enough to work with wonderful group members: Dohyoung, Anna, Arun, Prasanth, James, Sam, Gabe, Josh, and Caio.

Finally, I would like to thank my parents and sisters for their prayer, encouragement, patience, and love.

I give all the glory to God for all that He has done for me.

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGEMENTS v

LIST OF FIGURES viii

LIST OF TABLES x

1. INTRODUCTION 1

2. RELATED WORK 5

2.1 Abstract Modes of Digital Libraries 5
2.2 Domain-Oriented Digital Libraries for Scientific Data 6
2.3 Digital Libraries for Data Management 7
2.4 Digital Libraries for Data Analysis 8
2.5 Interaction between User and System 10
2.6 Recommender System: Relevance Feedback and Filtering 11

3. PROBLEMS AND ISSUES 13

3.1 Visualizations for Large and Various Types of Data 13
3.2 Need for an Integrated Workspace 14
3.3 Digital Library as a Data Platform 15
3.4 User Interest Model and Inference 16
3.5 Interactions for Data Management and Analysis 17

4. PERCON: A PERSONAL DIGITAL LIBRARY FOR HETEROGENEOUS DATA 19

4.1 PerCon Architecture 21
4.2 PerCon Instantiation 24
4.2.1 Database and Repository in Resource Layer 25
4.2.2 Data Processing and Analysis in Middleware Layer 26
4.2.3 PerCon Interface 29
4.2.4 Interoperability and Compatibility 32

4.3 Integrated Visual Workspace 34
4.4 Mixed-Initiative Recommendation 37
4.4.1 Understanding Relationships in Data 38
4.4.2 Recognizing User Interests and Selecting Recommendations 41

5. USER STUDY 44

5.1 Participants 44
5.2 Domain Data for Participant Analysis 45
5.3 Participants Tasks 46
5.4 System Conditions 47
5.5 Procedure 50
5.6 Result Data Collection 51

6. RESULTS 56

6.1 Perceptions of Participants 57
6.2 Participant Work Practices 58
6.2.1 Number of Data Elements Examined 59
6.2.2 Number of Events/Interactions 63
6.2.3 Distribution of Activity 68
6.2.4 Mixed-Initiative Recommendations 75
6.3 Effects of Other Interfaces 80
6.3.1 Effect of Visualization in Repository Browser 80
6.3.2 Resistance to Individual Applications/Representations 82

7. CONCLUSION AND FUTURE WORK 84

REFERENCES 88

APPENDIX A 95

APPENDIX B 99

APPENDIX C 102

APPENDIX D 109

LIST OF FIGURES

Page

Figure 1. Flow of information in PerCon. 20

Figure 2. PerCon's architecture and software components. 22

Figure 3. Example of a Bayesian network designed for user study. 28

Figure 4. PerCon interface 30

Figure 5. Calendar visualization of query results. 32

Figure 6. Example of application objects in the workspace. 35

Figure 7. Multi-datastream synchronized viewer. 37

Figure 8. Mixed-initiative interaction framework. 39

Figure 9. Configuration 1: The visual workspace and the mixed-initiative recommendations are both unavailable. 48

Figure 10. Configuration 2: The visual workspace is unavailable but the mixed-initiative recommendations are available. 48

Figure 11. Configuration 3: The visual workspace is available but the mixed-initiative recommendations are not available 49

Figure 12. Configuration 4: The visual workspace and the mixed-initiative recommendations are both available. 49

Figure 13. Example of the final user-created workspace in Configuration 3. 52

Figure 14. Snippet of event logs recorded in event_interaction.dat. 54

Figure 15. Responses to questions related to workspace. 56

Figure 16. Responses to questions related to recommendations. 57

Figure 17. Average number of data objects examined between two interface modes in each group. 60

Figure 18. Average number of data objects classified or analyzed during the tasks. 62

Figure 19. Average number of events in the repository browser for each data object analyzed. 66

Figure 20. Average number of events in the repository browser per analyzed data object. 67

Figure 21. Ordering of user events in the repository browser and workspace shows distinct patterns of work. 68

Figure 22. Number and percentage of the users' workspace activities of data exploration and data interpretation in configuration 3. 71

Figure 23. Number and percentage of the users' workspace activities of data exploration and data interpretation in configuration 4. 72

Figure 24. Average number of the occurred workspace events for data exploration and data interpretation in configurations 3 and 4. 73

Figure 25. Number and percentage of data objects located. 77

Figure 26. A sequence of suggestion events. 78

Figure 27. A sequence of data creation events in the workspace. 79

Figure 28. Reponses to a question to related to repository browser. 81

LIST OF TABLES

Page

Table 1. Example APIs in a high-level design. 33

Table 2. Evaluation groups and interface modes. 50

Table 3. A list of files which PerCon records. 53

Table 4. Likert-scale questions. 55

Table 5. Number of data classified or analyzed in each group. 61

Table 6. Number of interactions in the repository browser (in bold) and number of data objects analyzed (in parentheses). 65

Table 7. Workspace event types. 70

1. INTRODUCTION

People in scientific communities and industry need help in collecting, managing and interpreting data. A crucial issue for providing useful results from the current data explosion [9] is facilitating interactions with heterogeneous data sources. Advances in sensors along with software for scientific data management/delivery mean more data of more data types and are available than ever before. Additionally, increases in interdisciplinary research put greater demands on scientists to bring together datasets from independent communities to better understand phenomena. As a result, they are often overwhelmed by the amount of data and the related information activities such as data location, exploration, and analysis.

Tools supporting heterogeneous data management and analysis often focus on domain-specific representations and interfaces [9] or create separate silos for data of each data type (e.g. GenBank [5]). Domain-oriented visualizations are used for exploring and locating data. Indexing and classification of data most often occur in relation to predefined structured representations developed for specific domains using database or repository. Subsequently, searching or querying data applies to the domain-specific classified indices. The required preliminary efforts such a particular data processing and integration are also domain-dependent approaches.

*Part of this section is reprinted from the following paper: ©2014 IEEE. Reprinted, with permission, from Su Inn Park and Frank Shipman. PerCon: A personal digital library for heterogeneous data. In Proceedings of IEEE/ACM Joint Conference on Digital Libraries (JCDL), pages 97-106, IEEE, Sept. 2014.

Despite the benefit of domain-oriented digital libraries, there are still challenges for presentation and analysis in heterogeneous data. The domain-oriented digital libraries provide a limited or no general method that can represent other domain data. Indexing, classification, and search in predefined domain-specific data have no or little consideration to comprehend how different data are potentially related. In addition, locating data of interest in large heterogeneous data becomes a potentially time-consuming activity. These problems motivate the research on a heterogeneous data platform and mixed-initiative interaction for data analysis.

We are interested in supporting the collection, management, and interpretation of unanticipated collections of data types – the type of idiosyncratic collections that occur in the formative stages of exploratory research. We liken these personal and small group data collections to personal book and document collections. Thus, our long-term goal is to support the ingestion, management, indexing, and interpretation of ad-hoc collections of data.

As a step towards this vision, we developed a personal or small group digital library system called PerCon (**Per**sonalized and **Con**textual Data Environment) that allows users to process, manipulate, analyze, and interpret diverse and interrelated data. We investigated domain-specific environments including software tools and applications for data processing, visualization and analysis. Beyond a typical digital library, PerCon is a combination of digital library and data platform for heterogeneous data. The goal of combining digital library and data analysis components led us to design and instantiate a layered architecture for managing the interconnections among the many and varied

necessary software components. Along with the architecture, PerCon's main interface was designed consisting of three interface components: a repository browser, a visual workspace, and a suggestion viewer.

PerCon is unique in that it features a visual workspace and mixed-initiative interaction with the user. The visual workspace includes a model for selecting among multiple applicable data visualizations according to different requirements. Thus, a user can explore and translate data into information visibly in multiple presentations/visualizations (e.g. temporal, thematic, and spatial composition), and discover knowledge from information through data annotation, and spatial organization to express the relationships. Each visual data object in the workspace is composed of two objects; the base data object that is used for user expression and the application object which is used for data visualization. To improve human data analysis ability, PerCon observes user behavior to infer user interests and locates and recommends related data within the current collection. In order to identify relevant data, PerCon builds up feature spaces of data objects including their metadata and preserves records of user activity. It also includes a user interest model to infer the user interest or specific information need. In particular, Bayesian networks in PerCon model user interactions (i.e. event features) and predict the data type of greatest interest through network training. Beyond the system features and capability, PerCon allows us to examine how the visual workspace and mixed-initiative recommendations affect a user's work practices and performance of data exploration, analysis, and interpretation.

Evaluation of PerCon was performed with two central hypotheses associated with the effect of the visual workspace and mixed-initiative recommendation. Twenty four participants were asked to fulfill three tasks under two different system configurations out of four (i.e. with/without the visual workspace and mixed-initiative interaction) with river and weather datasets. Then, all user and system activities recorded were analyzed. The results show user perception, work practices, and effect of the visual workspace and recommendation: more satisfied user experiences, efficient user interactions, and improvements over data analysis by finding more evidences with the workspace and recommendation.

The following section provides an overview of related work. Section 3 addresses problems and issues. Section 4 describes PerCon's architecture, interface, and analysis capabilities. Evaluation of PerCon and findings are discussed in Sections 5 and 6. Finally, Section 7 addresses conclusions and future work.

2. RELATED WORK

Work informing our efforts comes from a variety of subfields: abstract models of digital libraries, digital libraries for scientific data management, analysis and interpretation tools, interactions within digital libraries, and implicit relevance feedback.

2.1 Abstract Modes of Digital Libraries

With early efforts to build digital libraries to provide basic functionalities [35] fundamental abstractions and models/architectures of digital libraries have been well-defined and established.

McCray and Gallagher [34] addressed underlying principles for digital libraries development. Gonçalves *et al*. [18] explored and defined fundamental concepts for digital libraries in their 5S (Streams, Structures, Spaces, Scenarios, and Societies) model. The DELOS Network of Excellence on Digital Libraries [10] introduced a reference model for systematic approaches to digital libraries defining four perspectives (end-user, designer, system administrator, and application developer). In that model, conceptual frameworks are represented in six core domains – content, user, architecture, policy, quality, and functionality. In addition, DSSP [17] provides abstraction for interrelated but disparate data management.

Scientific literature digital libraries, such as CiteSeerχ [32], NDLTD [16], and SAO/NASA Astrophysics Data System [21], are built around service-oriented

*Part of this section is reprinted from the following paper: ©2014 IEEE. Reprinted, with permission, from Su Inn Park and Frank Shipman. PerCon: A personal digital library for heterogeneous data. In Proceedings of IEEE/ACM Joint Conference on Digital Libraries (JCDL), pages 97-106, IEEE, Sept. 2014.

models/architectures. For example, CiteSeerχ comprises three layers: storage, application, and user interface layer. Since each library focus is unique (e.g., semantic web services or search engines via distributed servers/repositories), the models are specialized to functions matching that focus (e.g., crawling, storage). Open-source digital library systems, such as Fedora [38], DSpace [54], and Greenstone [62], are versatile environments for creating and managing collections in various domains. However, they were mainly designed for contents management and acquisition.

2.2 Domain-Oriented Digital Libraries for Scientific Data

Digital libraries for scientific data often take the form of generic holders of unindexed data, modeling and providing access via generic metadata attached to data files or other institutional repository software.

Domain-oriented digital libraries can make use of domain-specific representations to index into datasets. In practice, these libraries often provide siloed access, limiting user queries to a particular type of data. For example, based on data integration, researchers in bioinformatics have developed databases and computational/statistical analysis tools to explore different types of large-scale genome sequencing. Genbank [5] in the U.S is an implementation of huge databases functioning as a type of fine-grained digital library systems. Diverse geography datasets have also been incorporated into domain-oriented data libraries [60]. For instance, the Alexandria Digital Library [50] provides search services from collections of geographically referenced materials. Healthcare informatics digital libraries, such as Microsoft HealthVault [36] are used for managing and sharing personal health records collected

from different devices. The capabilities of these systems are necessary for our vision of digital libraries of personal data but they assume users have the expertise and tools for merging and analyzing the data located. The separation of data access from data analysis also reduces the likelihood that systems provide proactive support across this access/analysis border (e.g. recommendations of new data based on the user's analysis activity).

2.3 Digital Libraries for Data Management

There have been many proposed architectural approaches to digital libraries to manage data. As an amount of the data and the number of data types are increasing, several architecture approaches to (distributed) data store and processing have been researched. Digital libraries such as Massively Parallel Processing (MPP) Databases [14], Hadoop-based digital libraries [65] are examples for big data management including data processing. Additionally, NoSQL-featured digital libraries [11] are another approach to highly scalable data management.

With regards to scientific data management, digital library instances in various research fields have been developed. Based on data integration, researchers on bioinformatics, as a representative data-intensive science, have developed databases and computational/statistical analysis tools to explore large-scale genome sequencing. EMBL-Bank [25] in the U.K. is an instance of huge databases functioning as a type of fine-grained digital library system. Geography datasets have also been incorporated in domain-oriented data libraries. Health (care) informatics digital libraries, such as Intel Healthcare [23] and Microsoft HealthVault [36], are for managing and sharing personal

health records. The primary boundary is that the datasets are composed of interrelated data streams that are collected simultaneously from individual participants. This breadth has motivated the use of a combination of domain-independent ontologies, domain-dependent ontologies, and personal ontologies for representing the potential relationships within datasets. In addition, online digital library, ETANA-DL [43], manages archaeological heterogeneous data.

As the volume and the need to share heterogeneous data have increased, long-term research on heterogeneous data management, such as DataSpace [17] and DataNet [31], has been performed within the database community. As a traditional approach for data management, the data repository and (relational) database have offered well-defined structures and schema to store and access the original data objects as well as computed/filtered datasets including metadata [42]. In addition, many interdisciplinary areas have contributed to managing various types of data. For processing and integrating the heterogeneous data collections, algorithms based on various models, transformations, and filtering methods have been explored [15].

2.4 Digital Libraries for Data Analysis

For data analysis, computational, statistical, visualization-based, and human-computer interaction-based methods has been applied in digital libraries. Primarily, visualization in digital libraries leverages human visual cognitive perception to help to explore data/information and provides insight into data. Visualization also leads to analytical reasoning for data interpretation and analysis in significantly increasing data size and complexity. In particular, associated with diverse data sources, various

scientific data have been processed using computational and statistical methods and then visualized within digital libraries. For example, Bernard *et al*. [6] proposed metadata visualization with respect to content-based similarity to lead the user to find relationships between data items. Ordonez *et al*. [39] developed a visualization of time-series physiological data to recognize significant changes in a patient's condition. As an effort toward visual representation of textual data, Abbasi and Chen [1] applied a linguistic feature-based visualization technique for analysis and categorization. In addition, the textual search results in a digital library were visualized [48]. Rowe *et al*. [44] represented spatial data visually to model 3D complex geometry data via interactive and sketch-based interfaces. Also, Booker *et al*. [8] visualized geo-temporal data by combining a geospatial node map and a timeline view with their GIANT system. Exploratory visualization techniques have been developed to provide interaction with large and complex datasets and to give users broad bandwidth of understanding interpretation [33], [40]. Beyond type-dependent visualization methods, prior studies for classification, techniques, and challenges of visualization were addressed in [27] and [63]. Curdt *et al*. [13] managed and visualized heterogeneous scientific data in multiplicity of interdisciplinary subprojects including agricultural sciences, hydrology, geohydrology, geophysics, meteorology, and geoinformatics. PerCon builds on such capabilities but places their results in an analytic workspace in which users can organize and interpret data objects, much like the visual expression enabled in our group's prior work on the Visual Knowledge Builder (VKB) [45].

Compared to computational/statistical/visualization-based data analysis within digital libraries, data analysis through human-computer interaction is not still largely applied while interaction modeling/framework [7], or interaction-based applications for design [19], task planning [24], configuration [41] have been described to date.

2.5 Interaction between User and System

Interactions between a user and the digital library system fall into three major approaches – interaction of system control, human control, and human-system interactive control. First, system-controlled interaction is employed in automated systems that take full control and guide users to perform a task. Systems, such as Google, ResearchIndex [30], Informedia [59], are representative examples of automated digital libraries for information discovery/retrieval, citations/hyperlinks/reference linking, and metadata aggregation, respectively. However, many automated digital library systems are designed for lowering cost based on computation power rather than enhancing interaction with human users. The second approach for interaction is a direct manipulation, for example, where users take initiative in terms of manipulating, evaluating, and displaying data. Direct manipulation is supported by well-designed interfaces, functions, and data visualization. For example, various systems researched by Shneiderman [48] or other systems, such as DRAGON [26], approached direct user manipulation with user interfaces and visualization techniques. Finally, as a flexible and collaborative interaction, mixed-initiative interaction combines automated services with direct manipulation to provide various supports such as planning, configuration, and diagnosis while a user performs a task. Conventionally, the system agent(s) computes

related data and information according to the user's responses and infers a user's interests. For example, scheduling system, Lookout [22], combinFormation for information discovery and exploratory search [29], AIDE for exploratory data analysis [51] adopt mixed-initiative interaction between users and systems. As such, these interactive systems incrementally access and achieve goals. In the collaborative mixed-initiative interaction, while the user performs a task, PerCon recognizes a user's information needs, searches relevant data, and assists a user's activities.

2.6 Recommender System: Relevance Feedback and Filtering

Most data search, location and analysis tools leave the users in control, and on their own, during the performance of their task. The users' actions in the system provide evidence of their interests that can be used to generate recommendations; the system adapts to individual user's information needs or interests. For the recommendations, techniques of relevance feedback and filtering have been considered to model user interest and support user-dependent/personalized information delivery.

Relevance feedback takes user's feedback from an initial set of queried results and uses that information (i.e., whether the result set are relevant or not) to improve the retrieval process. The types of relevance feedback are usually distinguished as implicit feedback, explicit feedback, and semi-explicit feedback. Recommendations via implicit relevance feedback have been well studied in the context of information location environments [28]. Conventionally, the system infers a model of the users' interests, based on patterns in metadata or contents, and uses this model to locate related information. For example, the Curious Browser [12] uses observations of mouse usage,

keyboard usage and the time spent viewing documents to generate suggestions and our prior work on the Interest Profile Manager builds a model based on activity in multiple applications [4]. Explicit feedback requires users to proactively indicate their interests or assess the relevance of documents on a scale using score/number or descriptions (e.g. “relevant”, “not relevant”), etc. Thus, explicit relevance feedback is strong and direct evidence of user interest and information need, however, the users are usually reluctant to indicate (e.g. ratings), apart from information seeking activities or given tasks. In addition, user interest or task can be out of explicit feedback due to interest shift and unexpected findings. Furthermore, placing graded relevance on documents or expressing user interests using keywords can be limited to expand in the overwhelming information space.

3. PROBLEMS AND ISSUES

Prior research has considered domain-specific environments (e.g. database, schema and its mapping, code-based algorithms) for data integration rather than considering integrated data environments. Even various software tools and applications that offer computing environments or libraries for data processing, visualization, and analysis are widely used in specific research groups. However, many additional efforts to manage, access, and share diverse data are needed upon demand. Outside of this research and tools specific to some data domain, there are more generic systems for sharing data presentations and interpretations. However, these systems do not have capabilities or services to handle diverse datasets. Herein, the overall problems and issues related to heterogeneous data management and analysis are stated in detail.

3.1 Visualizations for Large and Various Types of Data

To manage and analyze heterogeneous data, domain-specific and cross-domain visualization methods are necessary. Much of the raw and computed data collected in many research areas consist of numerical values associated with measurements in time-series or spatial domain. Without appropriate visual representation, tasks such as understanding waveforms, obtaining desired information, and interpreting data relationship (even in a single data object) may require a long time or even impossible to complete. In particular, due to the nature of data-intensive measurements in many recent research areas, visualization (e.g., time-based plot, map, and graph) needs to be emphasized to explore and analyze data as addressed in Section 2.5. Besides, the non-

numerical values in the dataset (e.g., textual data or multimedia data) often lack the accountability of itself because a spatial situation or a specific context presumes to affect or be related with other data/measurements. Multi-parameter synchronized visualization (e.g. spatial representation with textual data) is required to enable one to find correlations or relationships between heterogeneous datasets.

3.2 Need for an Integrated Workspace

The importance of a shared workspace to support heterogeneous data management and analysis needs to be emphasized. A workspace serves as a virtual repository to produce a user's data and an important framework in digital libraries to translate data into information through appropriate processing, visualization methods, and applications depending on data types. A workspace, where the translation occurs, potentially manages the interrelationship between information objects and includes information evolving records.

Users and the system need a workspace to produce data and to build and represent the corresponding information explicitly in their desired manner. In particular, formalizing the information, which usually takes place in a workspace, at the human level is another challenge to data management. Computers only process formalized information and the formalization allows domain-dependent/independent data management. In addition, as Shipman *et al*. addressed in [46], for systems to support users, formal models or formalities are required to add necessary functionalities as users deal with increasing datasets and information. With regards to users, knowledge which is an understanding of information through human cognitive process is often implicit to

share. By formalizing large scale and complex information explicitly in a workspace, users can (1) improve the understanding of the nature of diverse types of dataset, (2) interpret the given information and discover knowledge in different ways, (3) share findings and understandings, and (4) reuse data. Furthermore, in domains whose knowledge can change rapidly, the formalization in a workspace facilitates knowledge acquisition [47].

With regards to human-computer interaction, a shared workspace is valuable for managing user-system cooperation. Terveen [55] describes the potential for a shared workspace to mediate user-system collaboration. In particular, a shared workspace where a user and the system can interactively and visibly build and manipulate a context enhances data management and analysis. When the user and system can communicate by constructing representations of information such as Critspace [3] in a shared workspace, a complementary approach [56] between human and computer can be achieved for data analysis.

3.3 Digital Library as a Data Platform

We intend to support the collection, management, and interpretation of unanticipated collections of data types. Developing a digital library system to support heterogeneous data management, including data management plans, poses many difficulties. First, managing heterogeneous data requires a comprehensive understanding in order to process and integrate the data collected from different sources. Second, data granularity varies grossly based on data transactions and desired performance. Developing a digital library with coarse-grained and fine-grained data structures and

models involves various schema designs and it also needs to consider schema matching for (meta)data sharing and interchangeability. Third, diverse services and applications should be provided in order to manage and analyze heterogeneous data. For the services and applications, different data processing methods, flexible and scalable infrastructure/architecture for new types of data are required. Finally, due to the inconsistency and lack of standards [37] between heterogeneous datasets, discovering interrelationships is important but difficult. In practice, one digital library instance cannot support any required or necessary visualization applications on demand. To tackle these issues, our digital library should have the ability that serves as a unifying platform to collect, process, organize, analyze, and interpret (un)structured data from any source. Namely, the solution to a data platform with methods is required. Compared to domain-oriented digital libraries, little research on digital libraries as a data platform has been carried out.

3.4 User Interest Model and Inference

Locating and finding (relevant) data of interest or in a specific context can be time-consuming and difficult among a large amount of heterogeneous data. Identifying the user's context and interest/goal for the data location is not simple. In specific, there are many difficulties to tackle: how we define implicit/explicit feedbacks, how a system acquires and correlates them with user interest, how we quantify relationships/similarities between the heterogeneous data in various domains, how a system notifies

Compared to studies on relevance feedback, less attention has been given to mechanisms of how user events and feedback are applied to mixed-initiative interaction, or user interest modeling with heterogeneous data. Heterogeneous data involves various variables associated with relevant or preferred visualizations, activities and attributes in a workspace, and relationships between data sources/objects. Besides similar or relevant data location with proper ranks, the user interest model needs to encompass these variables.

By recognizing user's context and data of interest, a digital library enables recommendation for the user's data analysis activity in collaborative mixed-initiative interaction. As the amount of data being analyzed and the evidence being collected increases and many machine learning algorithms are being developed and sophisticated, the recommendations are appreciated with values.

3.5 Interactions for Data Management and Analysis

While visualization conceptually enables a user to interact with data/information, in the real world, a user's interaction with a system involves his/her use of physical user interfaces for various purposes. Several conventional approaches to user interaction, such as command windows, selectable menu items, and dialogue popups have been employed.

However, even if digital libraries potentially include numerous selectable menu items on the screen and a user has great control over interfaces, one still faces several issues in terms of data management and analysis [49]. In particular, the issues seem more obvious when a system requires users to formalize information in a workspace;

Shipman *et al*. [46] mentioned possible reasons for rejecting formalisms in an interactive system. First, task overload and workload are likely to continuously increase. As researchers deal with more complicated information domains and build more knowledge space for data analysis, they may have a number of subtasks to undertake, issues to monitor, and relations among heterogeneous data to examine. This leads to limitations for researchers to explore all the related datasets as the amount of datasets increases. Second, users have various backgrounds, interests, hypotheses, and tasks regarding the datasets. Individual users have different data/work practices in terms of data exploration and structure [2], even when pursuing for the same purpose. In addition, the relevant data to find and analyze varies depending on the user's goals. There is no way to facilitate flexible and effective data analysis depending on several conditions of users and given tasks. Third, the more complex the system environment becomes, the more time users need to spend to learn the system interface and functions. Difficulty in locating interface or menu of interest can be posed within the system. Even if the user knows every possible interface and has full control over it, the user may not want to perform all of the potential subtasks for some reason. These issues indicate flexible interactions between a user and system for data analysis are necessary and should provide a high degree of collaboration. A collaborative interaction for solving problems or performing tasks allows users to reach goals in an incremental fashion.

4. PERCON: A PERSONAL DIGITAL LIBRARY FOR HETEROGENEOUS DATA

PerCon is more than a typical digital library; the system is a combination of digital library and data platform for heterogeneous data. It integrates data management with data manipulation, presentation, and analysis. The envisioned process for collecting, locating, and making use of data streams in PerCon is shown in Figure 1. The lower left represents the collection and ingestion of heterogeneous data, and the lower right shows the visualization, location, analysis, and interpretation of data through applications. The top right is a database and data repository for the data and the top left shows the feature space used to encode characteristics of the data necessary for its analysis and use as metadata.

PerCon was originally developed as a tool for a local research group to manage and analyze data from an investigation looking for patterns in physiological data from a variety of wearable sensors (e.g. heart rate, galvanic skin response), contextual data from portable devices (e.g. geocode and sound pressure level data), and behavior data (e.g. users' answers to questions). PerCon has since been extended to work with data from external sources and other domains, as is the case in the evaluation reported in Sections 5 and 6.

*Part of this section is reprinted from the following paper: ©2014 IEEE. Reprinted, with permission, from Su Inn Park and Frank Shipman. PerCon: A personal digital library for heterogeneous data. In Proceedings of IEEE/ACM Joint Conference on Digital Libraries (JCDL), pages 97-106, IEEE, Sept. 2014.

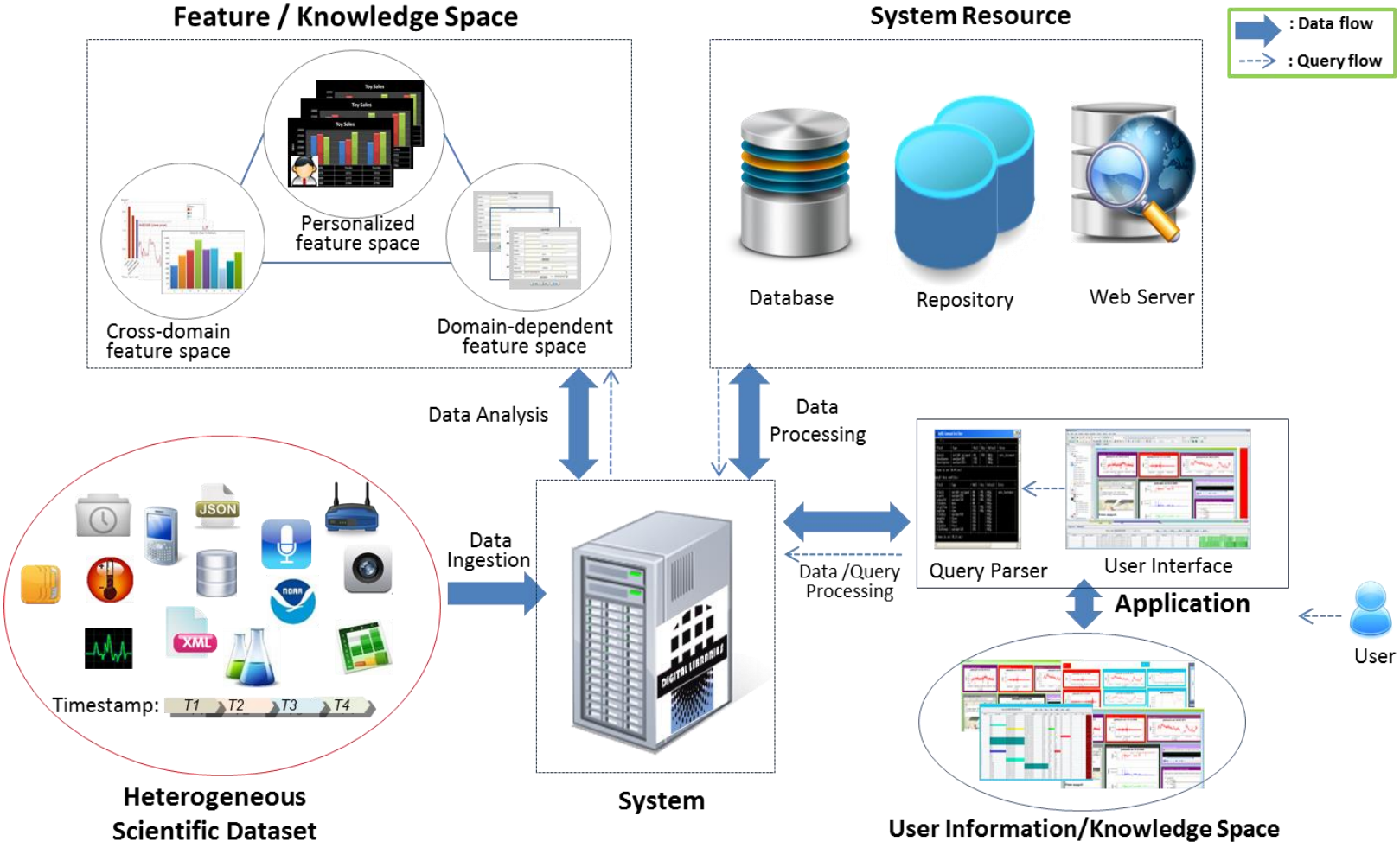


Figure 1. Flow of information in PerCon.

This section describes PerCon's architecture, overall interface, visual workspace, and mixed-initiative recommendation subsystem. The next subsection describes the layered architecture being developed to isolate the PerCon software components.

4.1 PerCon Architecture

The goal of combining digital library and data analysis components led us to develop a layered architecture for managing the interconnections among the many and varied necessary software components. As illustrated in Figure 2, PerCon's architecture for managing the interconnections and interoperations among the diverse software components and data resources consists of three layers: the resource layer, the middleware layer, and the application layer. The figure also describes some of the core capabilities at each layer and the components within it. As is typical for layered software development, software components in a layer are accessed only by software components in the next higher layer.

The resource layer provides functionality to store and preserve the original data objects, computed and filtered datasets, and metadata. This includes maintaining a record of the data provenance of data ingested from other sources (e.g. where it came from and when), data entered from local sources (e.g. project and user IDs, data types, dates) and computed data (e.g. recording what operations were performed on which existing data streams and when). It is implemented via a combination of a local repository and a database. The repository is designed to include loosely connected data sources. It stores and manages the raw and processed data objects, and data relation objects that record similarities and differences between data objects. The database stores

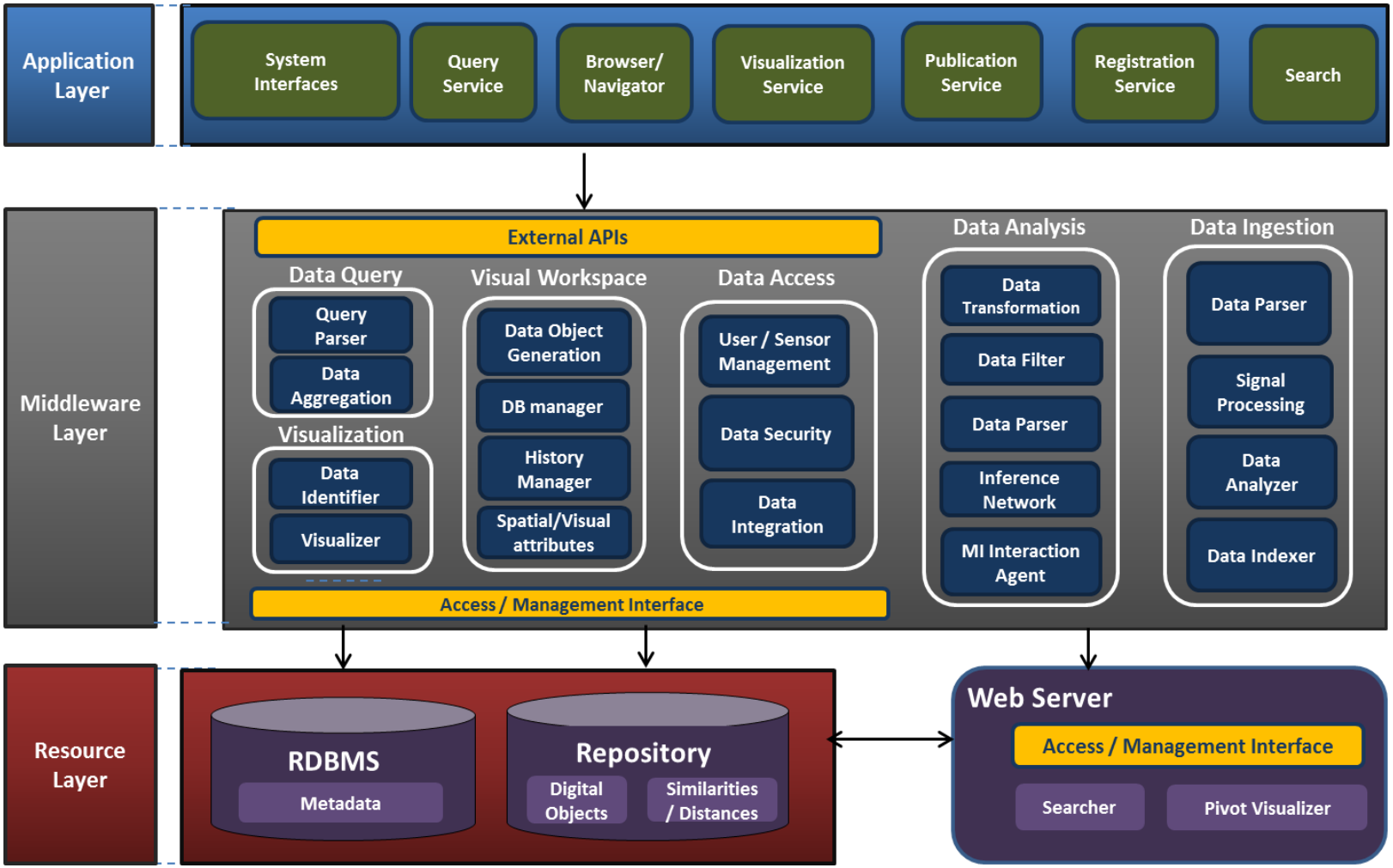


Figure 2. PerCon's architecture and software components.

descriptive metadata, visual thumbnails of the digital objects, and indexes into the data objects stored in the repository. The resource layer also includes a web server to provide network applications access to the data resources in the repository and database. Finally, the resource layer records information about data provenance.

The middleware layer of the architecture includes functionality for data ingestion, data access, automated data analysis, and lower-level components that enable data visualization the data workspace, which is a modified version of VKB [45].

The data ingestion component provides three services: data processing, data integration, and provenance recording. The processing service includes a data reliability check to ensure that the data object content correlates with that expected for the data object type. The processing service also provides algorithms for processing existing data to generate computed data objects and metadata, and to populate the data relations objects in the repository that record correlations and other similarities/distances between data objects. The data integration service populates database tables used to index the data object repository. For example, a table of high and low values per day enables searching the repository based on these content features efficiently. Finally, the provenance recording service extracts metadata regarding the data provenance.

The second functionality of the middleware layer is to enable the application layer components to access and interact with the resource layer's contents through a set of external APIs. A query-processing module parses requests and determines the communication necessary with the resource layer (e.g., database and repository) to fulfill the requests.

The third functionality of the middleware layer is data visualization, which instantiates data as a visual object in the workspace. The initial visual and spatial attributes of the visual object are assigned depending on the data type and query.

The final functionality of the middleware layer is data analysis. PerCon's analysis framework associates data objects in the resource layer with metadata and visual properties determined by the user and workspace, ranging from the highly interactive to the highly automated.

PerCon's application layer enables end-users and external systems to access the content in the digital library. This layer implements the user interfaces for browsing, searching, and visualizing the contents of the resource layer. A data registration interface enables adding, updating, and deleting data and datasets. The application layer also includes a data publication interface that enables remote access to certain contents of the digital library.

4.2 PerCon Instantiation

The architecture of PerCon presented in section 4.1 is instantiated for use in heterogeneous data. As such, the current PerCon system components provide a platform for various data types and data analyses via mixed-initiative interaction. Here, we describe the instantiation of the database and repository, the data processing and analysis components, the interfaces and services provided, and the API for developing additional services and providing external access.

The system is integrated into Java technology interoperating with other languages/scripts and interfaces of data repository and database. It attempts to provide a data platform as a solution to issues found in prior domain-specific digital libraries.

4.2.1 Database and Repository in Resource Layer

The resource layer of PerCon is instantiated through the combination of a data file repository and a relational database (e.g. MySQL) to maintain metadata, indexes, processed data objects into the data files. The repository includes both raw and computed data files, which can greatly vary in size depending on what type of data is being recorded and the data rate. A local or network storage can consist of repository directories. As is common in such repositories, the relational database maintains relationships between the data files in the repository and the metadata stored in the database. In addition, the relational database maintains indexes into the data files for particular events, words, or values.

Metadata in the database and repository is organized based on a representation of the domain of the research. In Figure 1, this domain representation is found in the Feature/Knowledge Space, consisting of domain-independent (or cross-domain) features, domain-specific features, and personalized features. The feature space also represents the potential relations between data files. Thus, the forms of computation over data files used to generate other data files are represented as a hierarchy. Unlike the original raw data, the domain-dependent, domain-independent, personalized features and relation-representing features are generated in a binary format for efficient data storage and fast access. Classes of functions used to generate computed data files include normalization

functions, filtering functions, and statistical functions, which are supported by the middleware layer. The preservation of the metadata about data collection (e.g. name, date, time, capture device, digital object types, etc.) and metadata describing the relationships between the digital objects combined with the feature/knowledge base enables greater understanding of the raw and computed data files.

The repository of digital objects (i.e. data files) is organized as a hierarchy of file directories, such as study, data type, research participant, and the sensor type. Thus, raw data and the data objects computed based on that raw data are found together. In addition, this can facilitate access control, which relates to issues of data privacy.

Along with the repository and database, as many digital libraries operate on web-based applications such as search engine and pivot viewer, PerCon can communicate with a web server. The search engine library in the web server is provided and the indexed data files are referenced by the search engine.

Finally, to abstract the resource layer, a PerCon configuration file is provided. Simply designating IP or domain address in the configuration (with authorization information if required), PerCon requests a connection then imports information of repository data files and database tables. In addition, by addressing web server address, users are allowed to use web-based applications or tools.

4.2.2 Data Processing and Analysis in Middleware Layer

The middleware layer of PerCon mainly supports ingestion of data into the repository/database, data analysis, query facilities, and visual workspace. As already implied in the previous section, data ingestion begins with a preprocessing step to verify

whether the raw data has been recorded correctly. In particular, the first check is to make sure the data file format is as expected (i.e. the data can be parsed if it is of a type meant to be manipulated). This step consists of validating data files, ensuring that the custom parser can read the data file without error. In addition, the ingestion may invoke a data-specific "sanity check" program to determine if the data has the characteristics expected. The user ingesting the data file is notified of any issues. Part of the ingestion process is to generate computed data files and indexes into the data files. Pre-processing of the raw includes modules for data parsing, filtering, transforming, indexing, and partitioning (based on a logical criterion or physical size). For example, the timestamps of raw data (which are recorded with CPU clocks) are transformed into a user-readable time to the second. It can add the results of a window mean, variance, and normalization to the computed data file. The ingestion process, including all processing steps and generation of processed data files, is dated/recorded for provenance analysis.

Data analysis capabilities of the middleware layer include a set of computation modules associated with the feature space and inference networks for mixed-initiative interaction. In particular, the computation modules enable various distances/similarities analyses in time-domain and frequency-domain (e.g., Fast Fourier Transform). The similarities/distances such as Pearson correlation coefficient, cosine similarity and user-defined metrics for representing relationships are used to find relevant data objects of interest to the user. In turn, to infer the user interest based on user-created events and feedback, Bayesian networks are modeled. Causal relationships between user interests and interactions in the workspace (e.g. events, visual/spatial attributes, applications) can

be modeled by the Bayesian network nodes and edges. Since the inference networks need to reflect the current user interest that can change over time and a given task, the learning process in the Bayesian network is light-weight. Furthermore, when user data is sufficiently accumulated in the resource layer, more personalized user interest models can be obtained through network structure learning. As a result, the computed data files and Bayesian inference networks enable mechanisms of how user events and feedback are applied to mixed-initiative interaction. Figure 3 demonstrates an example of a Bayesian network designed for our study in Section 5. The procedure of mixed-initiative interaction in PerCon is addressed in Section 4.4 in detail.

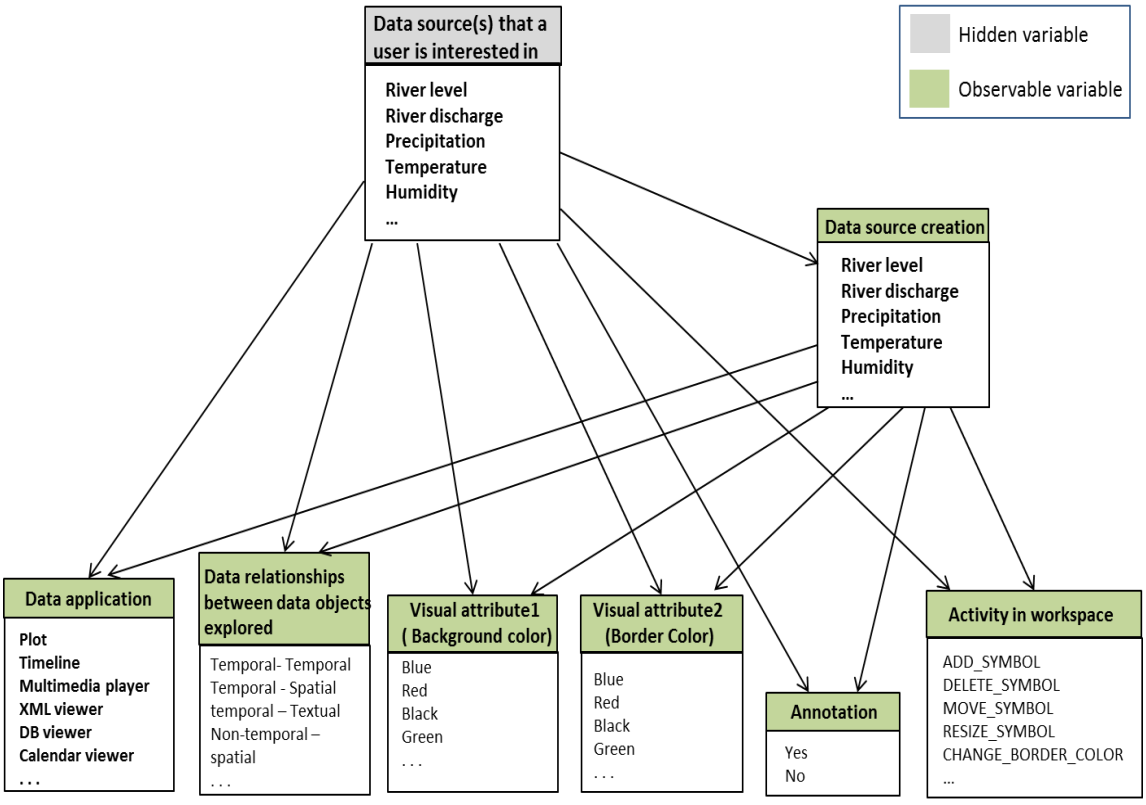


Figure 3. Example of a Bayesian network designed for user study.

Besides, the visual work including data multiple representations/visualizations and history mechanism is another main component in the middleware layer. We delineate the visual workspace in Section 4.3.

4.2.3 PerCon Interface

PerCon's main interface is composed of three interface components: a repository browser, a visual workspace, and a suggestion viewer. Figure 4 shows this interface being used to analyze precipitation and river data.

The repository browsing interface organizes the digital objects (i.e., data files) into a hierarchy. The hierarchy ensures that raw data and the data objects computed from it are found together. Users can filter the view by selecting the types of data or files that are of interest. A property often important when analyzing heterogeneous data is locating concurrent data sources, i.e., temporal overlap in data capture. Users of the repository browser can bring up the list of overlapping data objects from a selected element in the hierarchy.

The hierarchic browser also allows users to preview the data objects in precomputed thumbnails. Previewing was found crucial in early testing of PerCon as it aids the ability of users to rapidly locate initial data pertaining to their task while reducing undesired activity and complexity in the workspace.

Much of PerCon's interface is a workspace for visualizing and organizing data objects. Users can drag objects from the repository browser into the workspace to generate a new manipulable visualization of the data. More than one visualization can be available for individual data types. The initial visualization is based on the data type and

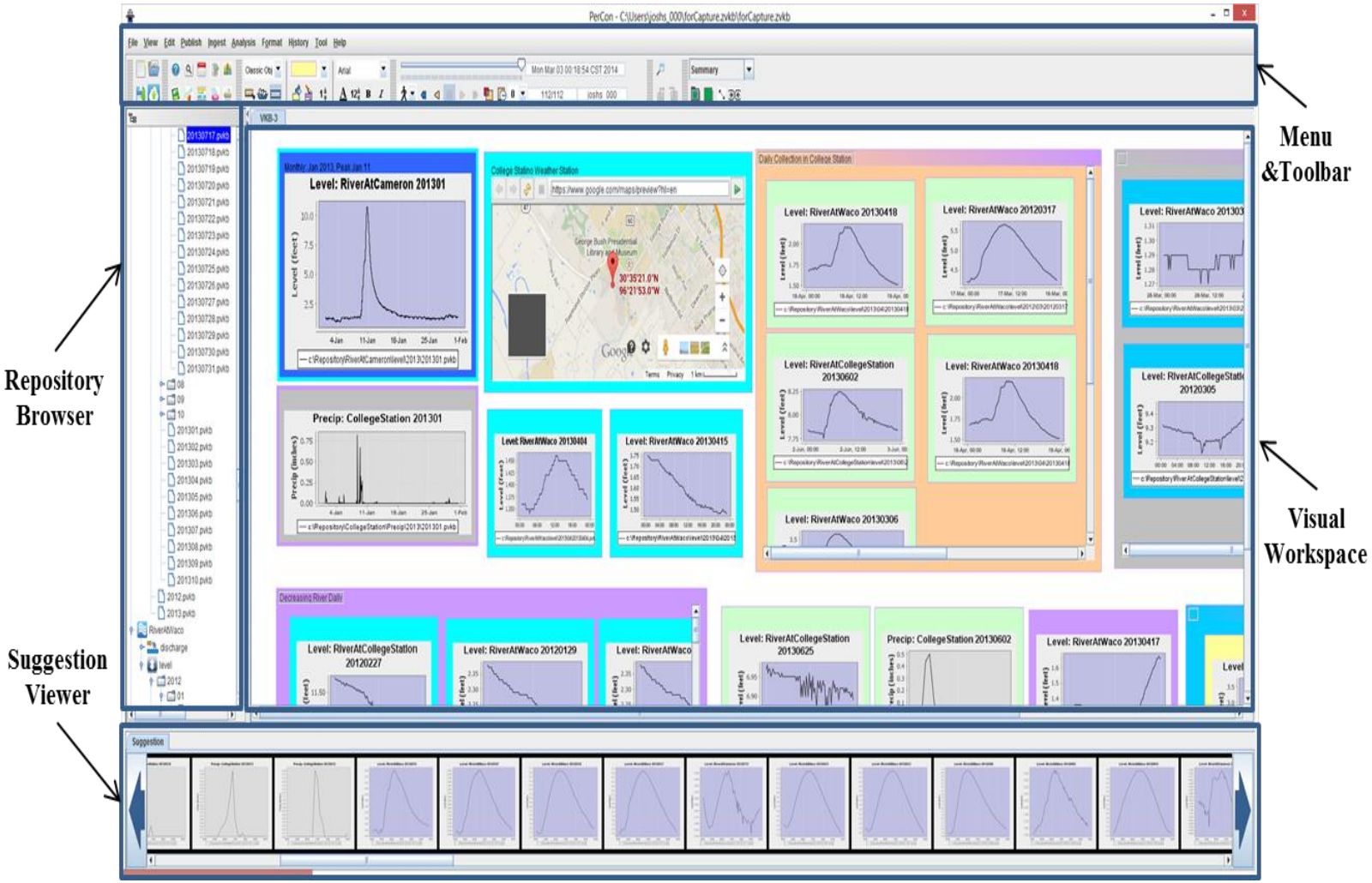


Figure 4. PerCon interface.

query used to locate the data but can be changed by the user. The workspace is described in more detail in the next subsection.

A history mechanism records user actions in the workspace. This history can be replayed to facilitate the comprehension of visualized information and to formalize interrelated knowledge. Based on user activity within the workspace, PerCon infers the user's interests resulting in additional data being presented in the suggestion viewer.

The suggestion viewer displays thumbnails of each recommended data object. The user can drag a suggested data object from the suggestion viewer into the workspace. A selection of the suggested data object from the suggestion viewer highlights a corresponding data object in the repository browser. This allows the user to know the suggested data source location. The suggestion viewer preserves the recommendation history for future reuse. This recommendation history can reflect shifts in user interest or transitions between subtasks.

The displays of the three components of the user interface communicate to ensure presentation consistency. When a data object is selected in one component of the interface, it is indicated in all of the interface components.

In addition to the main persistent interface, PerCon includes a query interface for locating data within the collection. As the data objects increase in number and size, locating data within the main tree view can be challenging. The query interface supports searching for particular types of data in particular date ranges. Figure 5 shows the calendar view of query results. Data that matches the query is shown as a set of labels and visualizations on the days from which the data comes. The data visualizations in this

interface use color to represent the type and content of the data. Each data type maps to its own unique color with different tones of that color used to represent different data values. Data entities in the calendar view can be dragged into the workspace and opened for a more detailed view.

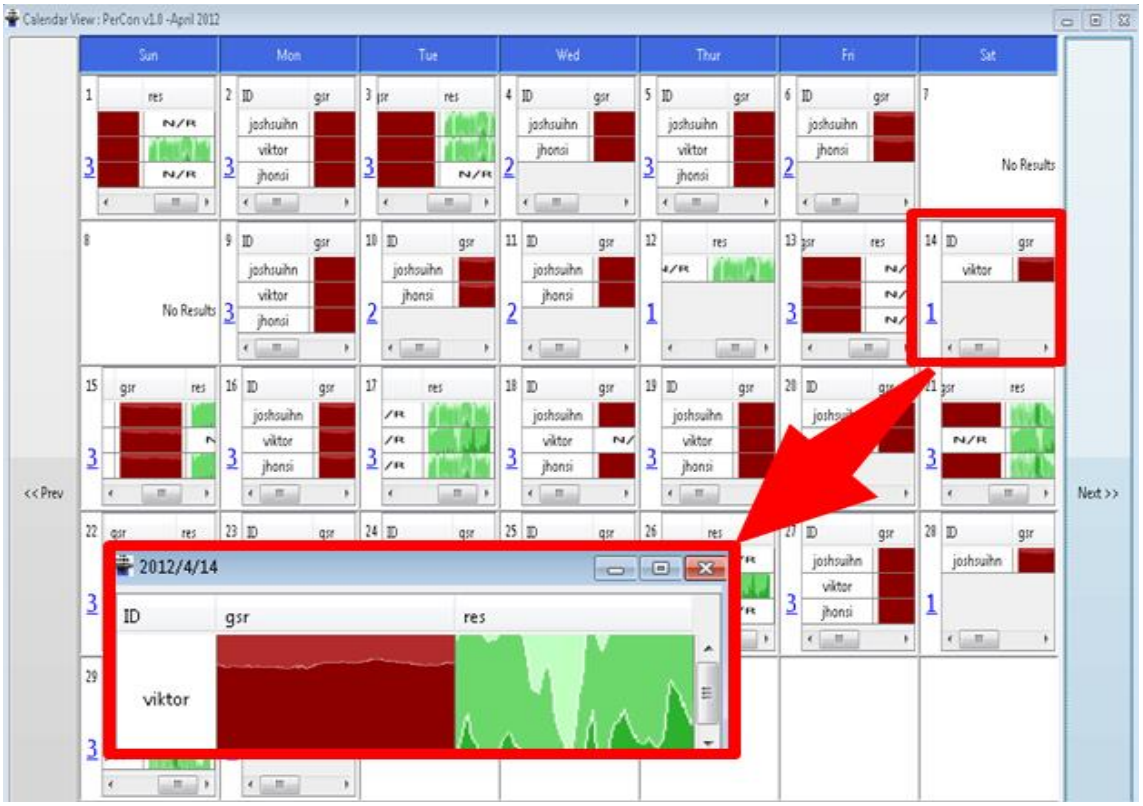


Figure 5. Calendar visualization of query results.

4.2.4 Interoperability and Compatibility

As indicated by the vision of data reuse in e-Science, digital libraries should not only preserve scientific datasets, but also share the resources with other systems. To share the resources, the applications and tools in the digital library system should

provide services such as searching, browsing, digital object manipulation, and analytical visualization. Associated with the possibility of infrastructure consolidation for relieving technical limits and reliable work with other system, operators in the middleware should be addressed. To respond to these demands, PerCon middleware components include application programmer interfaces (APIs). The example APIs for middleware services are shown in Table 1.

Table 1. Example APIs in a high-level design.

Method	API	Description
Browse	getObjectList	lists digital objects and directories from specified directory
Search	findByUserId	searches data/object by user ID
	findByTimePeriod	searches data/object within specific data-created time period
	findBySensorId	searches data/object by sensor id
	findByFunction	searches data/object optional) setting time window
	findByURI	finds data/object by location address
Result Access	getGraph	retrieves result data/object in visualized fashion
	getDocument	retrieves result data/object in specified document types
	getData	retrieves result data/object in specified data type
Data Object creation in the workspace	createNewObject	creates data object in the workspace
	plugin2Visual	adds data application object to base object in the workspace
Data object manipulation	componentResized	resizes data object
	setLocation	sets data object coordinate (top left coordinate)
Mixed-initiative interaction	addEventInteraction	records events to internal database
Exception	showException	returns error codes each of which has the meaning.

4.3 Integrated Visual Workspace

Based on the architecture and instantiation aforementioned, the PerCon offers various capabilities and functionalities for user data management and interpretation via the visual workspace.

PerCon extends the capabilities of our group's prior visual workspaces such as [45] in that it includes a model for selecting among multiple applicable data visualizations according to different requirements. For example, the same stream of quantitative data can be presented as a plot showing the value over time or as a bar chart showing the relative frequency of values in different ranges. Besides the system-generated visualization of the data, each data object includes visual and spatial attributes (e.g. border/background colors, font styles, data object x-y coordinates) that users can manipulate to express interpretations of the data.

The innovative approach to PerCon's visual data object is the separation of the base data object that is used for user expression and the application object which is used for data visualization. The base object model provides a method to add application objects onto the base object without constraints. Since movement, resizing and other event-based workspace interactions are managed by the base data object, only application-specific interactions must be considered when adding a new application object type. The result is a combination of human visual expression via base data objects and data visualization via application objects. Application objects can allow users to investigate the data in detail using appropriate methods (e.g. zooming) within the application object portion of the overall data object. To facilitate creating necessary

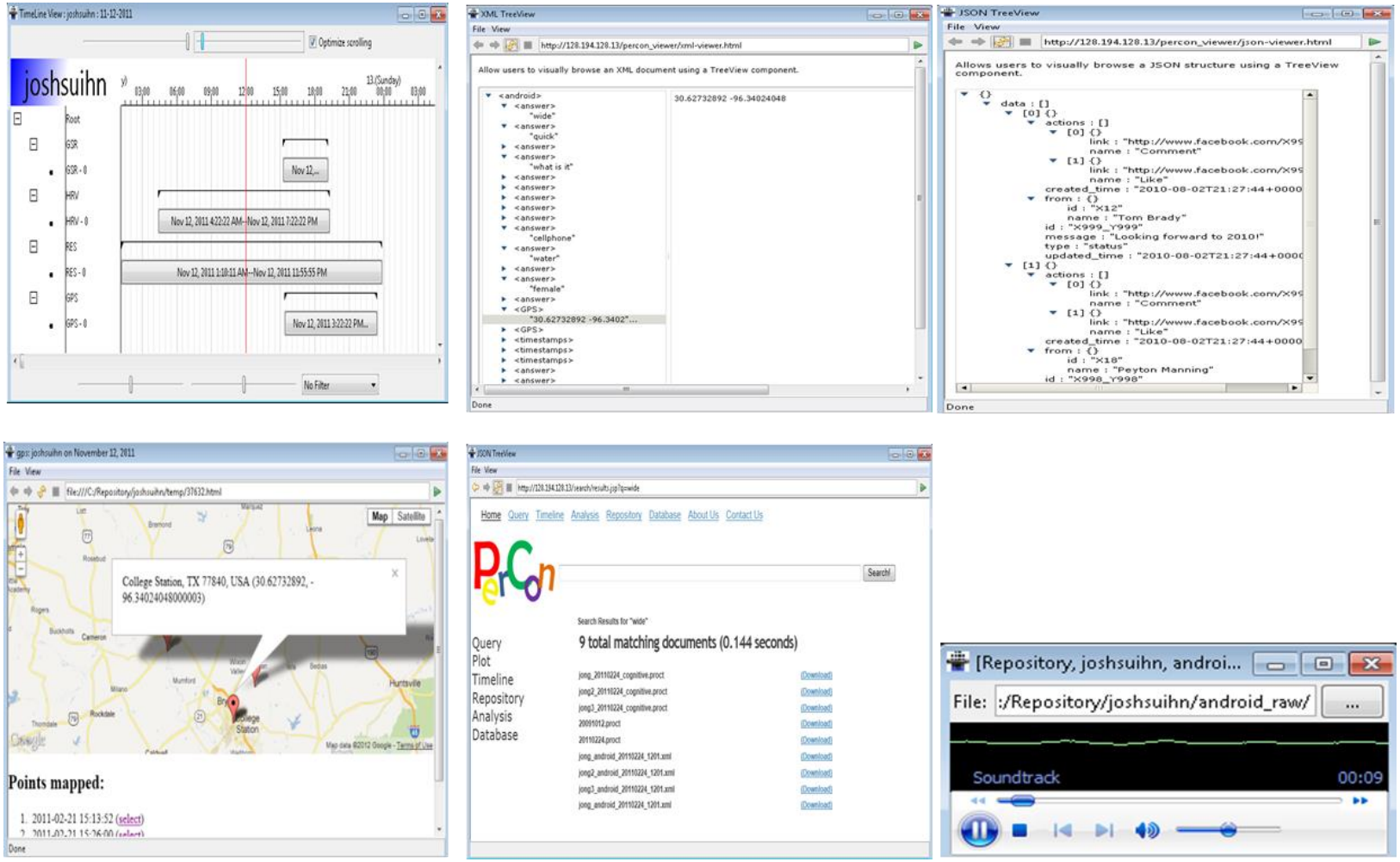


Figure 6. Example of application objects in the workspace.

application objects on the workspace, PerCon provides the external APIs. Other types of data with required management tools and applications can be integrated and accommodated in the workspace with less effort. This design rationale enables fine-grained data manipulations, adaptation of new types of data to visual workspace, and maximizing utilization.

A variety of application objects have been integrated in PerCon. Data editing and plotting tools, a HTML viewer, a multimedia data streaming application, interfaces to database tables, and timelines are available. In many cases, metadata standards and structured languages like XML and JSON, are employed for encoding characteristics of the data objects stored in the repository. Hence, PerCon also includes XML and JSON viewers. Figure 6 shows examples of application objects.

One PerCon-specific application object type is the multi-datastream synchronized viewer (e.g. [61]) shown in Figure 7. This was developed for our original application domain of analyzing physiological and contextual data. Since these datastreams are recorded from different sources in parallel, their correlations/relationships are not seen or detected visually in a single parameter view. An application object that integrated and visualized data from multiple data elements from the repository was desired to help identify patterns/relationships, as well as to form and assess new hypotheses. The multi-datastream synchronized viewer allows users to visually observe interrelated changes regarding the relationship between heterogeneous data.

Interoperating with the history mechanism, the workspace records and stores all interactions in the system internal database. The captured records can be used to revisit and replay workspace activity and are used for mixed-initiative interaction, which is described in the following section.

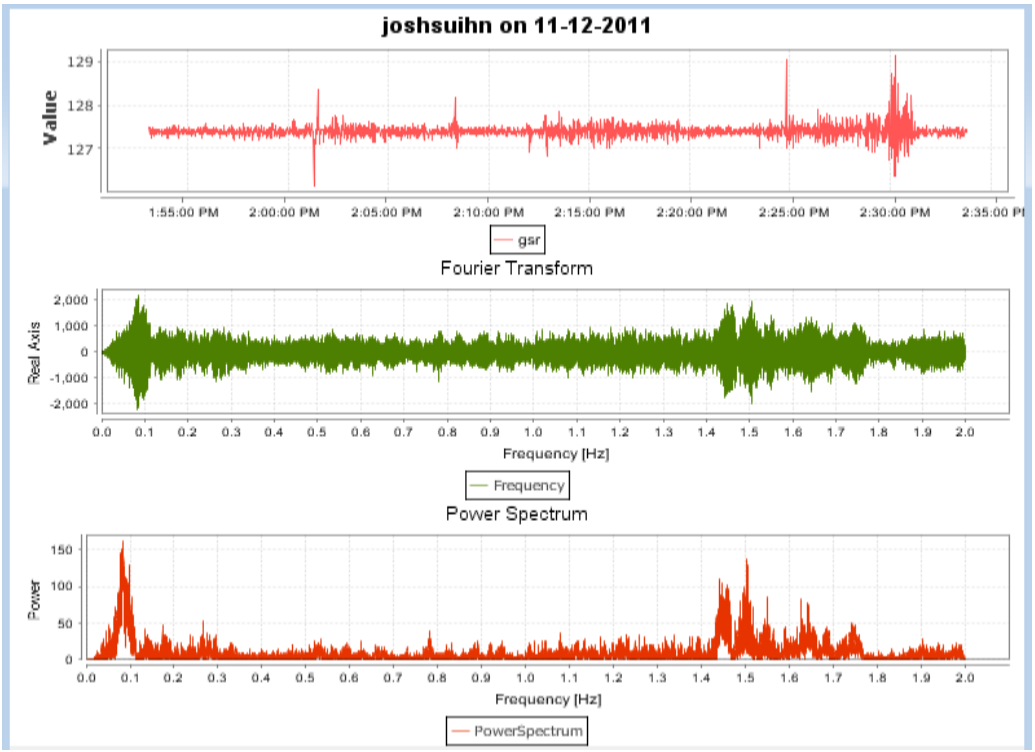


Figure 7. Multi-datastream synchronized viewer.

4.4 Mixed-Initiative Recommendation

Knowing what content is available is a challenge for users of most any library but is particularly difficult in a library containing a large quantity of heterogeneous data. PerCon includes a mixed-initiative interface for recommending data objects to the user

to improve data exploration and analysis in this context. The key functions of the recommender subsystem are (1) understanding relationships within the heterogeneous datasets, (2) recognizing user interests from a record of user activity, and (3) associating the relationships with the user interests. The agent's framework to accomplish these three functions is shown in Figure 8. The following two subsections describe the key procedures in mixed-initiative recommendation.

In an effort to avoid making recommendations that are perceived as random, we initially focused on getting users attention to the data they had not already seen that is most like what they have been recently examining. Future efforts can alter this strategy.

4.4.1 Understanding Relationships in Data

To understand relationships within the heterogeneous data collection, the agent generates precomputed tables of data object similarities. These tables are stored in the repository. Because different notions of similarity are important for data selection in different tasks, the processing framework computes five similarities/distances with values from 0 to 1 for all combinations of data objects of the same data type:

- Pearson correlation coefficient – standard measure of similarity of values over time in data elements;
- cosine similarity – breaks the data elements' timestream into segments and aggregates data in each segment into a single value; the two sequences of values are compared as vectors;

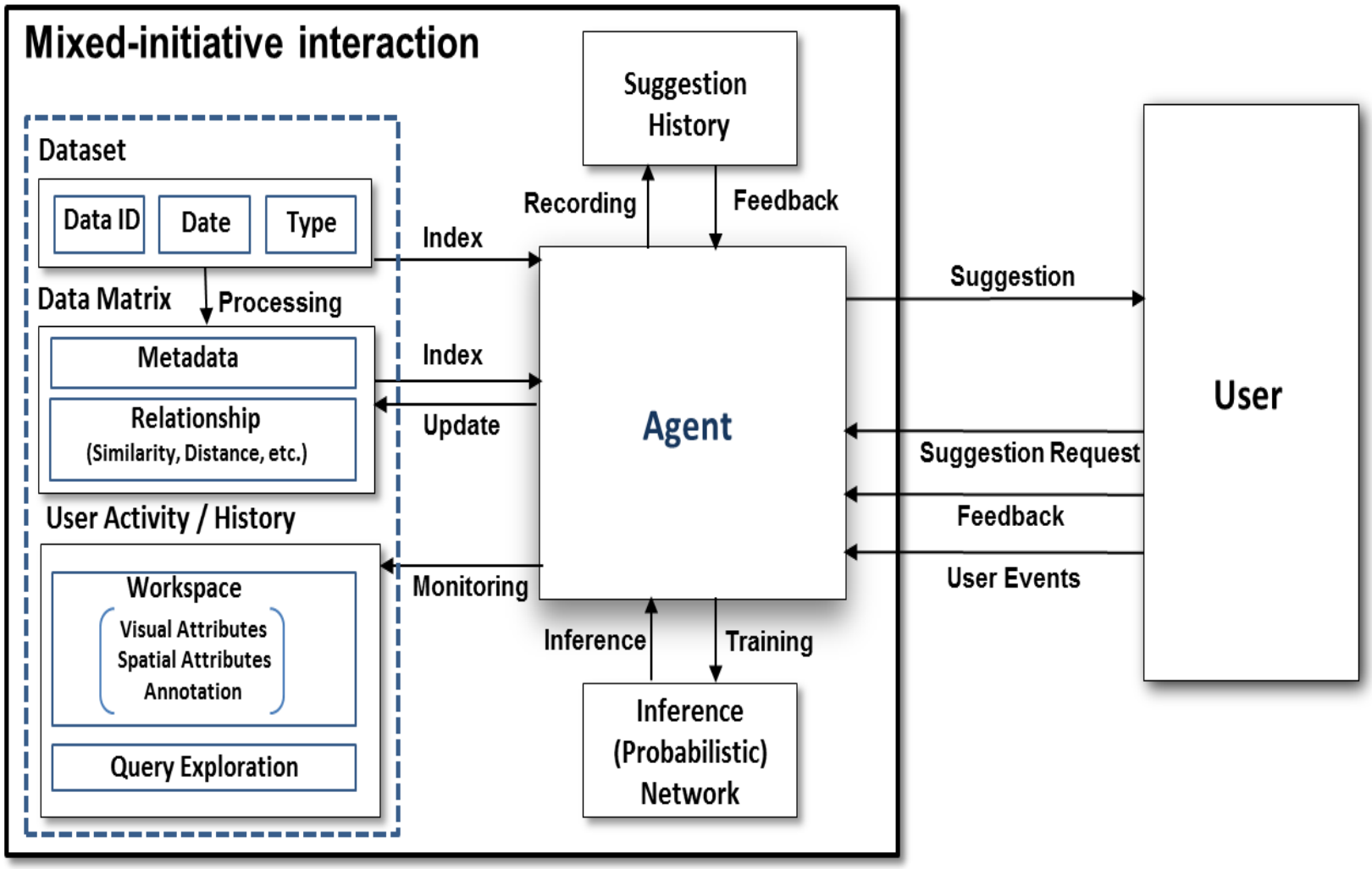


Figure 8. Mixed-initiative interaction framework.

- temporal similarity – how close together in time are the two data elements, value is 1 if the data elements overlap and is 0 if the time gap exceeds a preset threshold;
- trendline similarity – computes the Euclidean distance between the two-dimensional tuples composed of (the slope of linear regression of data readings, the sum of values of readings); and
- mean/variance similarity – difference between the mean/variances of the two data elements.

Before the similarities can be computed, interpolation and smoothing are necessary for comparing data elements with different sampling rates and/or missing values.

Overall similarity of two data elements is a weighted sum of the five similarities above. The initial weights for the five similarity metrics are set heuristically based on experience with the system. These weights are modified based on user interaction with recommendations by increasing the weight for similarity metrics that correlate with accepted recommendations.

The similarity assessment mechanism is extensible. Our initial measure of overall similarity was much simpler, using a combination of the mean values and variances, but the resulting recommendations were quite bad. Iterative testing of the mechanism led to the set of similarity metrics described. Additional metrics can be included in the overall assessment if experience indicates the need to include new features.

4.4.2 Recognizing User Interests and Selecting Recommendations

PerCon records user activity in the workspace as a sequence of events (e.g. dragging data elements from the repository browser into the workspace to create a new data object, resizing the data object, etc.). In addition, each event type has been heuristically assigned an evidence weight (e.g. data object creation is 0.1, resizing a data object is 0.01, etc.). Using the event log and the list of weights, PerCon generates a model of user interest as follows:

1. Each event in the workspace is added to a table recording the event type and features of the data object involved in the event (e.g., data type, application object type, color1, color2, annotation) and the evidence weight for the event is added to an evidence tally.
2. When the evidence tally exceeds a threshold, the table is used to train a Bayesian network composed of nodes modeling the event features in order to predict the data type of greatest interest.
3. The data objects in the workspace of the predicted data type are ranked based on the table of activity and those above a threshold are selected as reference data objects.
4. Similarity between the data elements of the predicted data type not already in the workspace and the reference data objects is computed based on the weighted sum of the five similarity metrics.

5. The five most similar objects that are not already in the history of recommendations are added to the suggestion viewer as thumbnails and the evidence tally is reset to zero.

Rather than building a model around a fixed notion of feature importance, the probabilistic model aims to capture the unforeseeable characteristics of practical action [53]. As users proceed through a data analysis task, the sets of data that are their focus, the applications they use to view the data, and the interpretive coding and annotation they apply in the workspace are likely to change [57], [64]. By including all these event and data object features in the Bayesian network, recommendations can take into account such shifts in behavior.

A last component to the recommendation system is controlling the frequency of recommendations. As described in the steps above, the accumulation of activity into the evidence tally is one method for controlling recommendations. Users have to perform enough actions in the workspace to (re)fill the evidence tally before a recommendation is made. In order to build recommendations that represent different lengths of user activity, the system includes shorter-term and longer-term evidence tallies. This means there can occasionally be multiple recommendations simultaneously or quite near together.

In addition to inferring user interests, the mixed-initiative interface allows a user to express his or her interests explicitly and to ask for recommendations. This is important because user activities in the workspace may not reflect user interests in cases when significant effort is required to examine data to locate data of interest, when there are sudden task shifts, and when the user performs multiple tasks in parallel. Also, a user

may want to explore data different from that which was recently examined. Users can select a workspace data object and request related data to request recommendations. Subsequently, the explicitly expressed interest is included in the agent's user interest model and thus affects future system-generated recommendations.

5. USER STUDY

PerCon is a large software environment and the result of many different design hypotheses. We conducted a user study to observe data analysis in PerCon focusing on two central hypotheses: (1) H1: the visual workspace helps a user to manage data and to translate data into knowledge about the domain, and (2) H2: the mixed-initiative recommendations improve a user's ability to explore and analyze data. The participants were asked to perform several specific tasks with/without the visual workspace and with/without the mixed-initiative interaction.

5.1 Participants

PerCon is meant to support people who need to analyze heterogeneous datasets. Thus, our population of convenience, the students and researchers at a large academic institution, is representative of the target population. The twenty-four participants included one undergraduate, four Masters, sixteen PhD students, and three postdoctoral researchers. The participants ranged in age from 24 to 36 and represented a variety of disciplines: computer science, computer engineering, electrical engineering, soil hydrology, biomedical engineering, industrial engineering, and management information systems.

*Part of this section is reprinted from the following paper: ©2014 IEEE. Reprinted, with permission, from Su Inn Park and Frank Shipman. PerCon: A personal digital library for heterogeneous data. In Proceedings of IEEE/ACM Joint Conference on Digital Libraries (JCDL), pages 97-106, IEEE, Sept. 2014.

5.2 Domain Data for Participant Analysis

Data for the study was selected to include relations that are intuitive in nature but complex in detail. In particular, we selected weather and river data along two different geographic regions of a river system to provide two comparable sets of tasks with which to examine the effects of alternate configurations of PerCon.

Since weather and river data are recorded hourly or daily over decades, they provide a voluminous dataset. Weather data includes many variables that exhibit a great deal of spatial and temporal correlations with one another. Many variables in the weather data also have an observable impact on environmental variables. River stream level is a representative environmental variable significantly affected by weather conditions. Depending on geographic relationships and intermediate reservoirs and dams, river stream levels exhibit a strong relationship between upstream and downstream values. Data was collected from two public repositories: the National Oceanic and Atmospheric Administration [58] and the Brazos River Authority in Texas [52].

Two years of weather and river data (from 2011 to 2013) were ingested into PerCon for this study. The weather data consists of five elements: temperature, precipitation, relative humidity, wind speed, and wet bulb temperature. The river data includes two data elements: the river level and discharge recorded. Data was collected from six different locations in Texas along the Brazos River and its tributaries. One data set included the data from College Station, Waco, and Temple while the other included data from South Bend, Seymour, and Fort Griffin.

The value for each data element was ingested for each hour, resulting in about 25 Mbytes of raw data and 229 Mbytes of computed similarity data for each of the two tasks. To facilitate access to the data, a directory of each data element provides participants access to annual, monthly, and daily segments of data.

5.3 Participants Tasks

Participants were asked to perform three tasks with each of the two datasets. These tasks motivated participants to carry out a complete cycle of data exploration, manipulation, management, analysis, and interpretation. Throughout the tasks, we provided the participants with a basic approach and methodology; each task included step-by-step procedures. To observe and discover how weather and river data are correlated, one possible approach is to classify and organize data objects based on individual changes, trends, and patterns in the data.

Task 1: Participants had 20 minutes to organize and classify river level and precipitation data according to common trends, quantities, durations, or other user-perceived criteria.

Task 2: With the classified weather and river data from Task 1, participants were asked to investigate the implications and identify correlations among the classified data. In particular, participants were asked to investigate the data correlations focusing on: what and how a weather factor(s) affects river level, and how the river data from different locations (e.g., Waco, Cameron, and College Station) are correlated. They had 10 minutes for this task.

Task 3: With the discovered evidence of relationships among the weather and river data elements from Task 2, participants were asked to explain river level changes/trends based on weather and upstream river flow conditions for 5 minutes. For example, some participants estimated river factors that caused these changes/trends (such as delay time) based on past weather data and other river stream conditions. Other participants interpreted why some river level changes are more or less affected by other river level changes.

5.4 System Conditions

For this study, our main hypotheses concern the effects of the workspace and mixed-initiative recommendations on data exploration and interpretation. To evaluate our hypotheses, we compared four PerCon configurations varying the availability of the visual workspace and recommendations. These are shown in Figure 9 through 12. Without the visual workspace, the two configurations in Figure 9 and 10, the system provided up to two information objects (i.e. application windows) with spatially preset size and location. The two configurations in Figure 11 and 12 show when the mixed-initiative data recommendation system was turned off.

When participants performed the tasks without the visual workspace, they could not use visual attributes and spatial organization to express themselves. In these configurations, they were allowed to use other application(s) such as MS Word or Excel to record notes, intermediate and final task results.

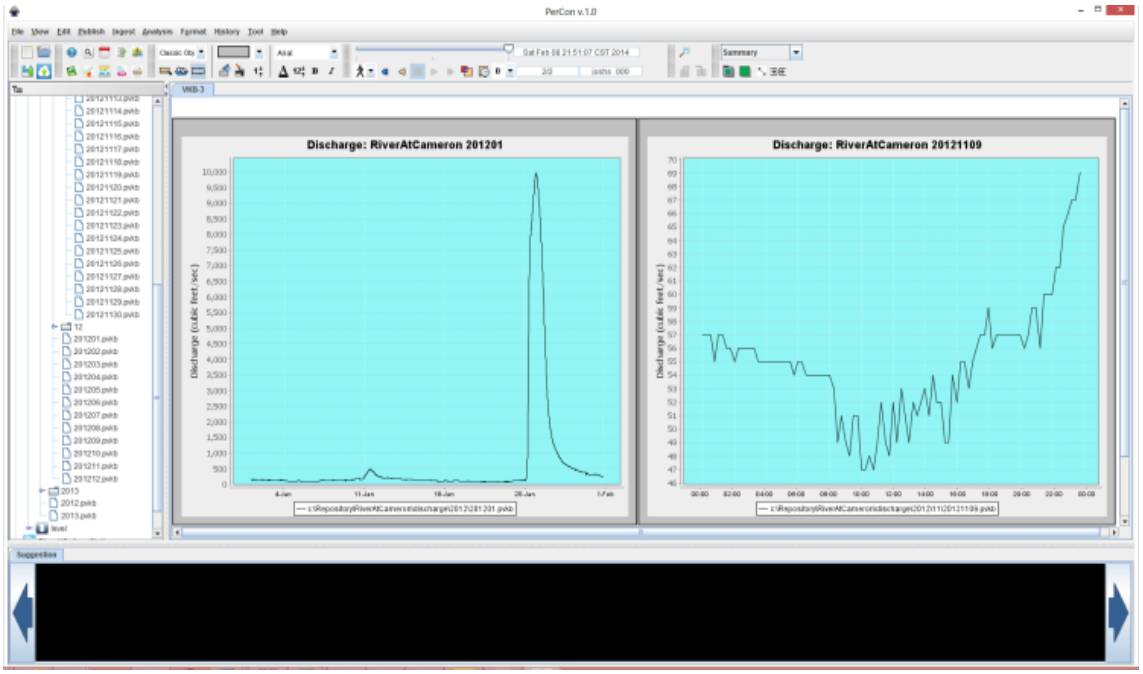


Figure 9. Configuration 1: The visual workspace and the mixed-initiative recommendations are both unavailable.

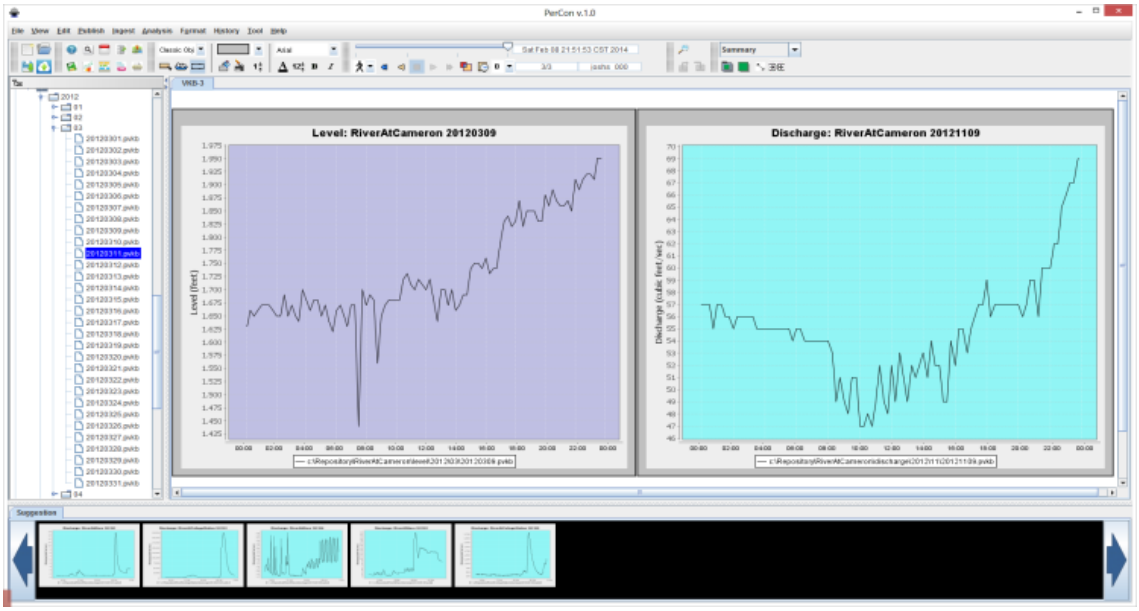


Figure 10. Configuration 2: The visual workspace is unavailable but the mixed-initiative recommendations are available.

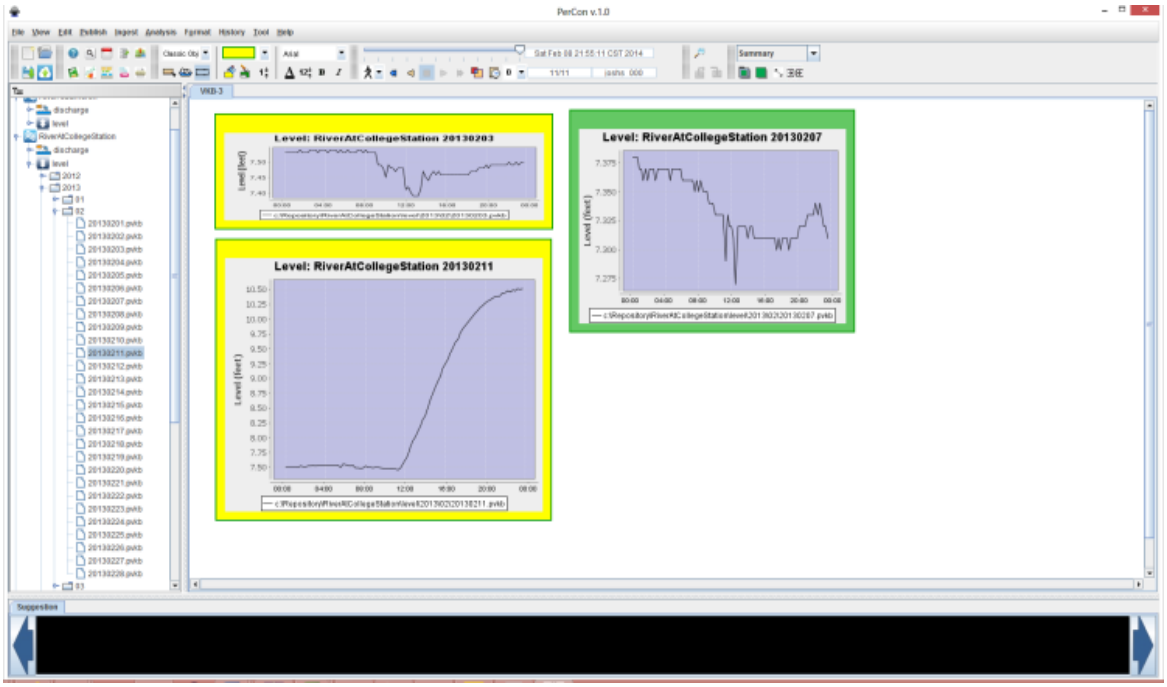


Figure 11. Configuration 3: The visual workspace is available but the mixed-initiative recommendations are not available.

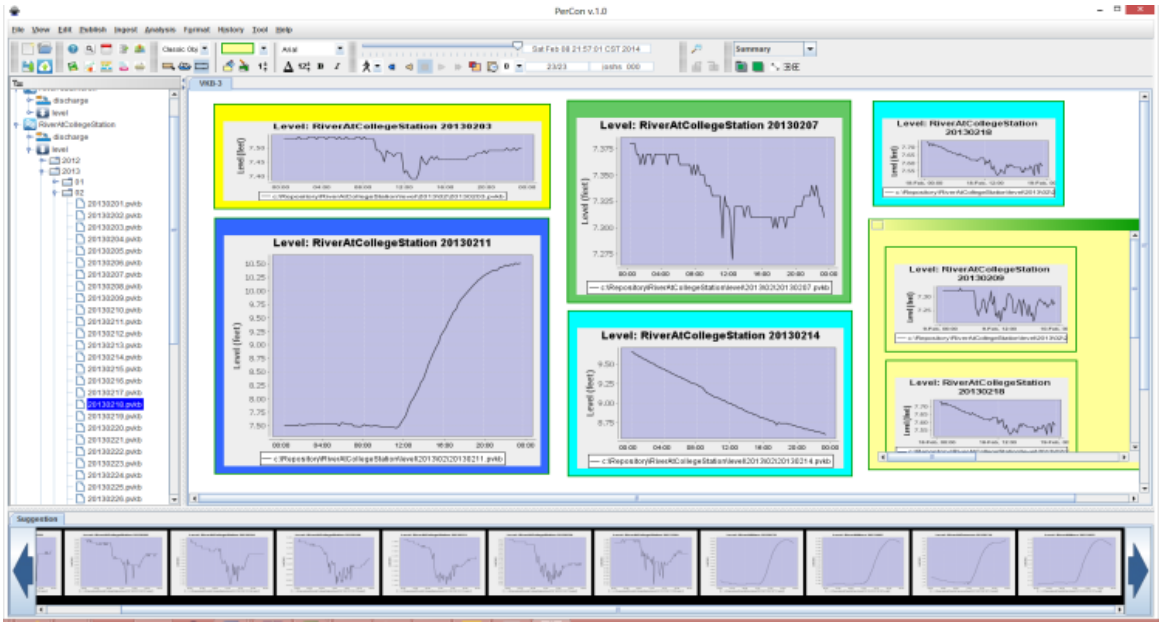


Figure 12. Configuration 4: The visual workspace and the mixed-initiative recommendations are both available.

5.5 Procedure

Before conducting the given tasks, participants were trained in the use of PerCon. This involved watching a 10-minute video tutorial to follow along with a written tutorial/manual to ensure consistent training. After being led through the use of features important to the tasks they would be given, participants had 5-minutes to try out features on their own.

Table 2. Evaluation groups and interface modes.

Group	Subgroup	System interface modes with dataset 1	System interface modes with dataset 2
Group 1	A	Configuration 1	Configuration 2
	B	Configuration 2	Configuration 1
Group 2	A	Configuration 1	Configuration 3
	B	Configuration 3	Configuration 1
Group 3	A	Configuration 1	Configuration 4
	B	Configuration 4	Configuration 1
Group 4	A	Configuration 2	Configuration 3
	B	Configuration 3	Configuration 2
Group 5	A	Configuration 2	Configuration 4
	B	Configuration 4	Configuration 2
Group 6	A	Configuration 3	Configuration 4
	B	Configuration 4	Configuration 3

Each participant was asked to perform the three tasks in Section 5.3 in two of the four system conditions. After each task, the participants were asked to save their task as a file where user events and system logs had been recorded. As shown in Table 2, the order of exposure to system configuration and data set were balanced to account for learning effects, interactions between experiences with configurations, and complexities inherent in the two data sets. Table 2 shows the six evaluation groups covering the combinations of different interface modes. Four participants were in each evaluation group.

At the end of the user study, the participants were given a questionnaire to explore the effects of the visual workspace and recommendations in each condition. The study duration for each participant was 120 minutes: learning how to use the system for 15 minutes, performing the given tasks for 70 minutes, and answering the questionnaire for 20 minutes.

5.6 Result Data Collection

Data about participant activities, system logs, and experiences was collected from four sources: (1) Likert-scale responses about the task and PerCon, (2) open-ended questions, (3) the final and intermediate user-created workspaces, and (4) a record of time-stamped events/interactions (i.e. user activities and system logs) with PerCon throughout the tasks. In particular, the final saved workspaces and the records of the events/interactions include quite a large amount of usage data. We collected 144 workspaces; twenty-four participants performed a total of six tasks in two system configurations. Figure 13 shows an example of the final user-created workspace in

configuration 3 throughout the three tasks. In addition, 1,728 system-recorded files were collected; as shown in Table 3, PerCon is designed to record twelve individual user and system log files along with the workspace. Figure 14 shows a snippet of the recorded event logs in the event_interaction.dat file.

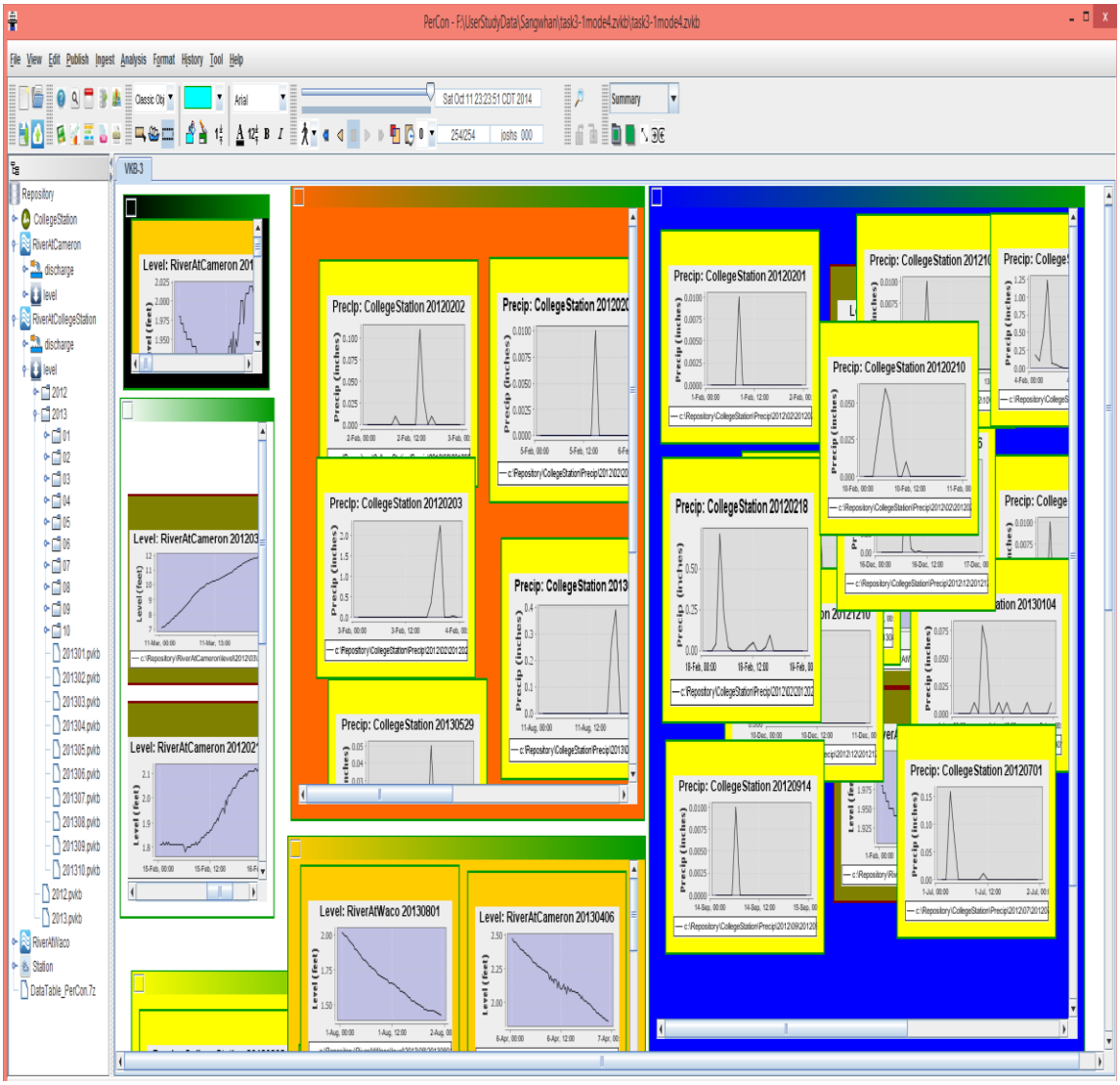


Figure 13. Example of the final user-created workspace in Configuration 3.

Table 3. A list of files which PerCon records.

File name	Description
event_interaction.dat	Events in the repository browser, visual workspace, and suggestion viewer are recorded
history_events.dat	History events in the workspace are recorded.
history_segment.dat	History segments in the workspace are recorded.
object_attributes.dat	Visual and spatial attributes of individual base objects are recorded.
object_values.dat	Data object values (such as annotation) are recorded.
percon_symbols.dat	Events of individual user application objects are recorded.
visual_symbols.dat	Events of individual classic objects or collections are recorded.
symbol_attribute.dat	Visual and spatial attributes of individual collections are recorded.
symbol_values.dat	Values of base objects or collections are recorded.
objects.dat	Events of base objects are recorded.
suggestionList.dat	Recommended data are recorded.
workspaceList.dat	Current data objects in the workspace are recorded and updated.

4	AddSymbol	PLOT	c:\Repository\RiverAtSeymour\level\2013.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
5	AddSymbol	PLOT	c:\Repository\RiverAtSeymour\level\2012.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
6	AddSymbol	PLOT	c:\Repository\RiverAtSouthBend\level\2012.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
7	DeleteSymbol	PLOT	c:\Repository\RiverAtSouthBend\level\2012.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
8	DeleteSymbol	PLOT	c:\Repository\RiverAtSeymour\level\2012.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
9	DeleteSymbol	PLOT	c:\Repository\RiverAtFortGriffin\level\2012.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
10	DeleteSymbol	PLOT	c:\Repository\RiverAtSeymour\level\2013.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
11	AddSymbol	PLOT	c:\Repository\RiverAtFortGriffin\level\2012\01\20120101.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
12	AddSymbol	PLOT	c:\Repository\RiverAtSeymour\discharge\2012\01\20120101.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
13	AddSymbol	PLOT	c:\Repository\RiverAtFortGriffin\level\2012\01\20120102.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
14	AddSymbol	PLOT	c:\Repository\RiverAtSeymour\discharge\2012\01\20120102.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
15	ResizeSymbol	PLOT	c:\Repository\RiverAtFortGriffin\level\2012\01\20120102.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
16	ResizeSymbol	PLOT	c:\Repository\RiverAtFortGriffin\level\2012\01\20120101.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
17	PLOT_EXP	PLOT	c:\Repository\RiverAtSeymour\discharge\2012\01\20120102.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
18	ResizeSymbol	PLOT	c:\Repository\RiverAtSeymour\discharge\2012\01\20120102.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
19	ResizeSymbol	PLOT	c:\Repository\RiverAtFortGriffin\level\2012\01\20120101.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
20	MoveSymbol	PLOT	c:\Repository\RiverAtFortGriffin\level\2012\01\20120102.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
21	ResizeSymbol	PLOT	c:\Repository\RiverAtFortGriffin\level\2012\01\20120102.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
22	MoveSymbol	PLOT	c:\Repository\RiverAtSeymour\discharge\2012\01\20120102.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
23	ResizeSymbol	PLOT	c:\Repository\RiverAtSeymour\discharge\2012\01\20120102.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
24	ResizeSymbol	PLOT	c:\Repository\RiverAtSeymour\discharge\2012\01\20120102.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
25	ResizeSymbol	PLOT	c:\Repository\RiverAtFortGriffin\level\2012\01\20120102.pvkb		joshs_000	1.39164E+12	PerConSymbol	100	200	100
26	AddSymbol	PLOT	c:\Repository\SouthBend\Precip\2012\02\20120229.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
27	AddSymbol	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120101.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
28	AddSymbol	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120102.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
29	ResizeSymbol	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120102.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
30	PLOT_EXP	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120102.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
31	PLOT_EXP	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120102.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
32	PLOT_EXP	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120102.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
33	ResizeSymbol	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120101.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
34	PLOT_EXP	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120101.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
35	PLOT_EXP	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120101.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
36	PLOT_EXP	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120101.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
37	PLOT_EXP	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120101.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
38	PLOT_EXP	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120101.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
39	PLOT_EXP	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120102.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
40	PLOT_EXP	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120101.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100
41	PLOT_EXP	PLOT	c:\Repository\RiverAtCameron\level\2012\01\20120101.pvkb		joshs_000	1.39165E+12	PerConSymbol	100	200	100

Figure 14. Snippet of event logs recorded in event_interaction.dat.

Table 4 lists the eight statements included as 7-point Likert-scale responses (1 means “strongly disagree” and 7 means “strongly agree”) concerning the effect and usefulness of the workspace and mixed-initiative interaction. Participants in Configurations 1 and 2 were told to consider the fixed layout configuration of the workspace combined with their chosen external tools as their effective workspace.

Table 4. Likert-scale questions.

		Statements
Workspace	**Q1**	I had enough support to understand the data content in the workspace
	Q2	I had enough support to express relationships in the way I wanted
	Q3	It was easy to interpret and characterize given/created objects in the workspace
	Q4	I had enough support to effortlessly / quickly browse and select data
Mixed-initiative interaction	**Q5**	I was satisfied with the data suggested
	Q6	I was satisfied with the suggestion request
	Q7	I had enough support to find and interpret data I was interested in
	Q8	I had enough support to find correlations within the dataset

6. RESULTS

Data from the Likert-scale questions will be reported first to give a sense of user perceptions of the different configurations. This will be followed by a more detailed analysis of the activity logs, workspaces, and mixed-initiative recommendations to examine how the different configurations objectively changed data analysis practice. From user experiences with the repository browser and workspace applications, we will find the effects of the other interfaces and draw lessons.

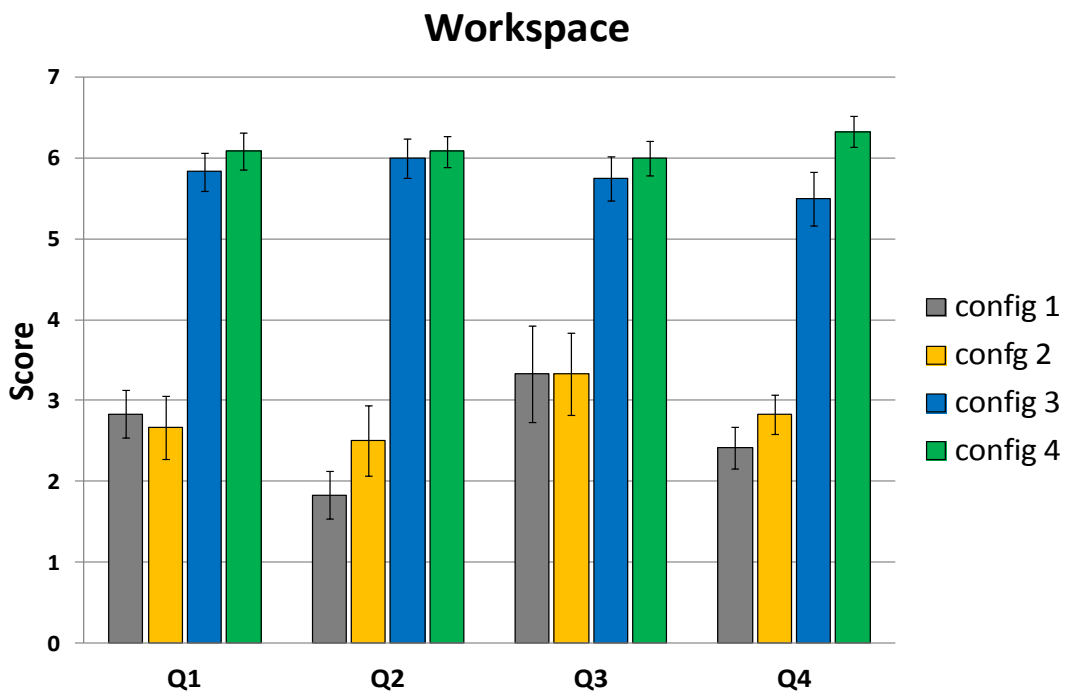


Figure 15. Responses to questions related to workspace.

*Part of this section is reprinted from the following paper: ©2014 IEEE. Reprinted, with permission, from Su Inn Park and Frank Shipman. PerCon: A personal digital library for heterogeneous data. In Proceedings of IEEE/ACM Joint Conference on Digital Libraries (JCDL), pages 97-106, IEEE, Sept. 2014.

6.1 Perceptions of Participants

Figure 15 presents the mean and standard error of participant assessments for the five workspace-related statements after using each interface configuration. The distributions for the five statements are all similar.

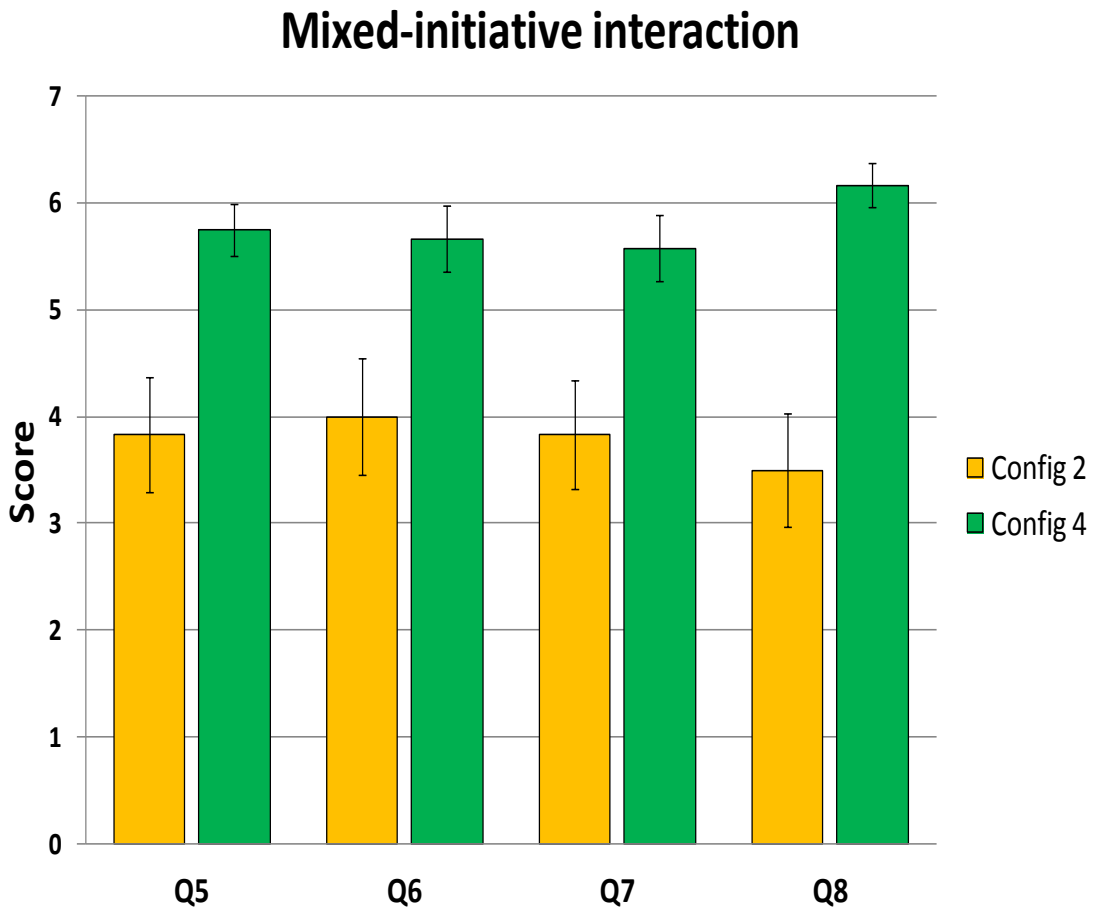


Figure 16. Responses to questions related to recommendations.

The means of the responses for Configurations 3 and 4, which include the workspace, vary between 5.5 and 6.5 out of 7. The means for Configurations 1 and 2,

which did not include the workspace, were between 1.75 and 3.25. In all cases, the difference between "without the workspace" and "with the workspace" is statistically significant ($p < 0.001$ for the closest, Q3). In the case of question 3, a few participants valued the larger fixed-size graphs while analyzing data, explaining that it provided a "*bigger plot*" and was "*easy to check the pattern*" in their open-ended questions (e.g., "*Did the visual workspace help or hinder your tasks regarding data management?*", "*Did the visual workspace help or hinder your interpretation of data*?")

Responses to the statements related to mixed-initiative interaction were requested only after using configurations that included suggestions (configurations 2 and 4). As is clear from the results shown in Figure 16, the impact of the recommendations was different depending on whether users had access to the workspace. Without the visual workspace (configuration 2), participants were more reluctant to explore recommended data as it would take over one of the two plotting areas they had available. This restriction was also commented on in the open-ended questions, (e.g., "*Did mixed-initiative interaction help or hinder your analysis of data?*") as below:

"*My analysis was hindered but I do not think it is because of the mixed initiative feature. Instead, it was the restricted workspace*."

6.2 Participant Work Practices

To investigate the effects in each interface mode, we first examined the work practices of each participant group; we analyzed the sequence and number of events associated with exploring, organizing, and interpreting data/information with timestamps. We also examined the effects of interface configuration on the number of data elements

classified/analyzed and the number of events occurred by participants. In particular, the next two sections examine the effects of the visual workspace and the mixed-initiative recommendations on the number of data elements classified/analyzed throughout the three tasks and the number of events/interactions per analyzed data in the repository browser, respectively. In order to account for the increased possibility of familywise error rate (Type-I error) caused by each individual hypothesis, a Bonferroni-adjusted significance level of 0.025 was calculated. Next we looked at the distribution and pattern of activity in the repository browser and workspace in the different configurations. Finally, we looked at the interleaving of system-initiated and user-requested suggestions.

6.2.1 Number of Data Elements Examined

We first examined and compared the number of data objects classified or analyzed in the two configurations of each group during the three tasks. Table 5 and Figure 17 show the number of the data objects examined and the average number in each group, respectively. Notably, the substantial increases in the number of the data objects among the groups was observed for all the participants in Group 2 (configurations 1 and 3) and 3 (configurations 1 and 4) as shown in Figure 17 and Table 5; the average increased from 13 to 38.8 in Group 2 and from 6.8 to 38.5 in Group 3. Since the common system configuration change in the two groups is the workspace, the potential effect of the workspace is likely to be significant. To quantify the effect of the visual workspace and mixed-initiative interaction, we examined the number of the data objects depending on the system conditions.

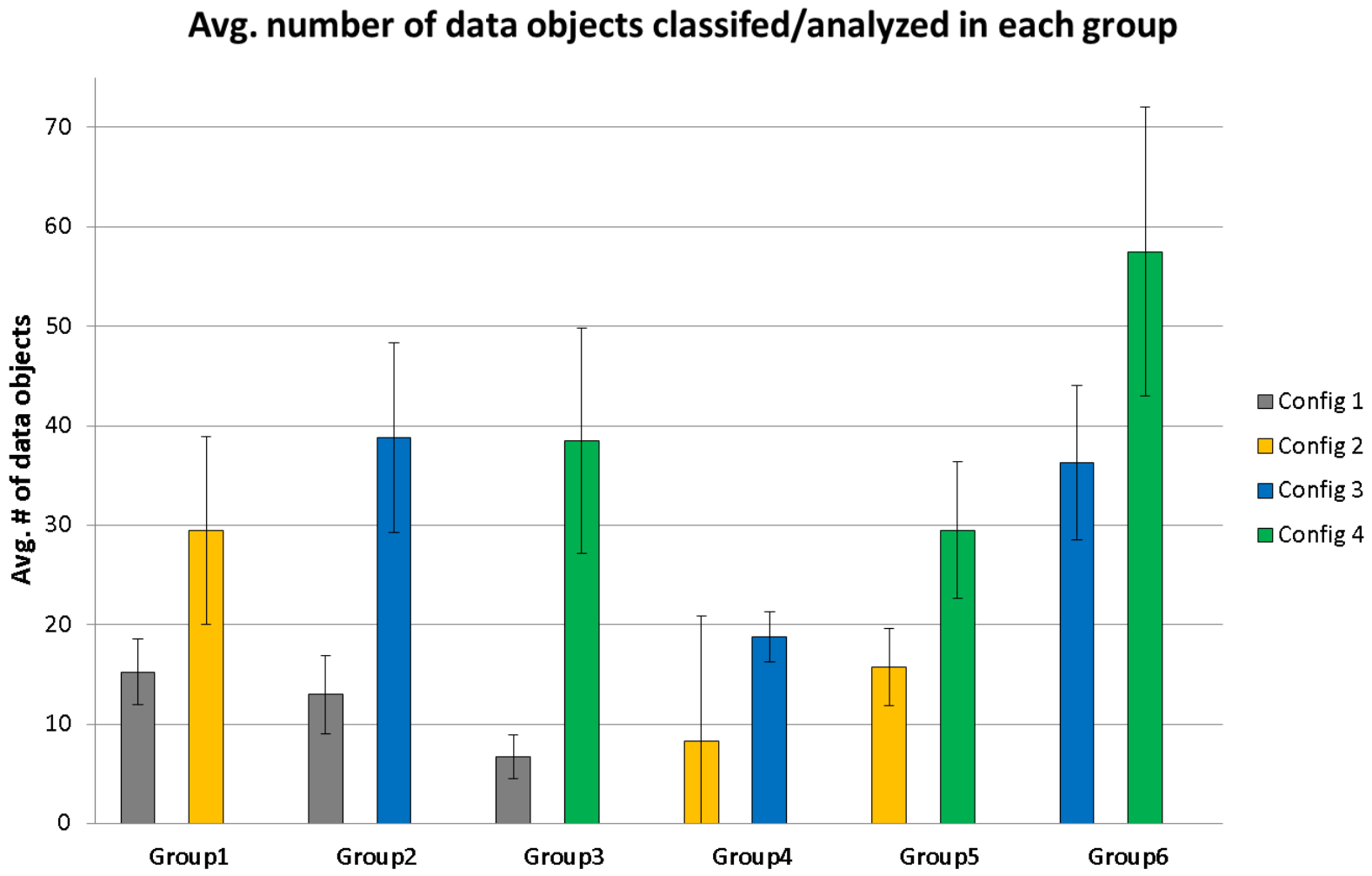


Figure 17. Average number of data objects examined between two interface modes in each group.

Table 5. Number of data classified or analyzed in each group.

Group	Configuration	User 1	User 2	User 3	User 4
Group1	Config. 1	16	7	23	15
	Config. 2	18	9	45	46
Group2	Config. 1	9	24	13	6
	Config. 3	25	31	67	32
Group3	Config. 1	11	2	10	4
	Config. 4	57	17	59	21
Group4	Config. 2	2	27	4	0
	Config. 3	14	22	15	24
Group5	Config. 2	26	15	15	7
	Config. 4	27	48	28	15
Group6	Config. 3	52	37	41	15
	Config. 4	54	95	57	24

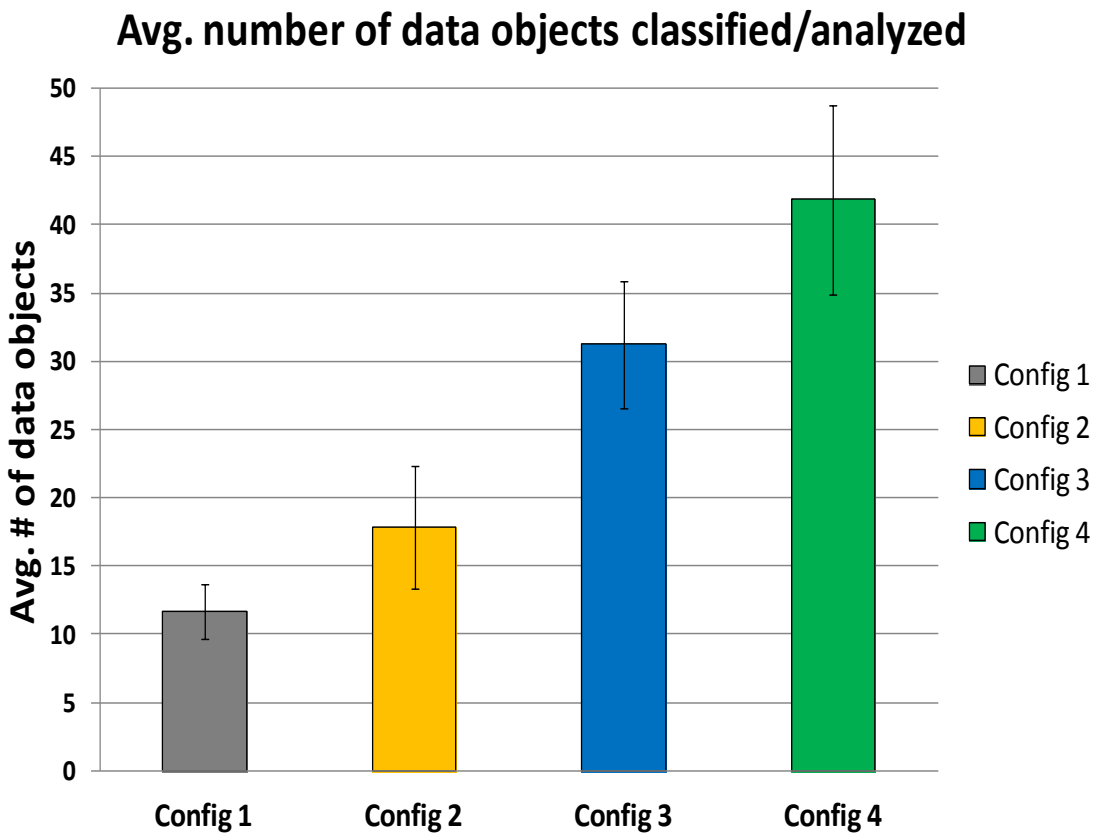


Figure 18. Average number of data objects classified or analyzed during the tasks.

Figure 18 shows how many data elements were classified or analyzed during the 35 minutes spent with a single data set under the four conditions. Without either the workspace or the suggestions, the average was around 11. When only suggestions were added, the mean rose to 17 and when only the workspace was added the mean went to 31. With both the visual workspace and suggestions, participants examined about 41 data elements on average. It is clear that having a drag-and-drop interface that supports the visualization of an open-ended set of data elements facilitates rapid data analysis.

The effect of the workspace on the number of data elements classified or analyzed is strongly significant ($p < 0.001$, t-test, Bonferroni-adjusted significance level = 0.025) when comparing the activity in configurations 1 and 2 to the activity in configurations 3 and 4. Along with the significant difference, the effectiveness of the workspace is large (Cohen's $d = 1.287$).

The larger number of data elements the participants explored, the more evidence they discovered to identify and explain relationships between data elements. Thus, the results of their Tasks 2 and 3 were strengthened. Furthermore, participants receiving recommendations were able to substantiate correlations over broader time periods.

Comments in the open-ended questions confirmed that the workspace supported rapid and persistent analysis:

> "*Obviously, there is still quite a large amount of data and sources to sift through. Being able to collect data objects together, stack them, and reshuffle at will allowed for more opportunity to see potential correlations in data that otherwise might have gone unnoticed.*"

> "*The visual workspace drastically reduced time spent switching between documents and shifting focus away from the data to writing notes. I could leave the most important/relevant data to my task in the workspace and return to it as necessary without having to jot down notes and re-search separate documents.*"

6.2.2 Number of Events/Interactions

Along with the number of data elements classified/analyzed, we also investigated the number of events/interactions in the two configurations of each group throughout the

three tasks. Since the participants could explore and locate data through the repository browser, the number of events in the repository browser indicates how efficiently they located data of interest during the tasks; a smaller number of data previews (i.e. exploration in the repository browser) implies that the users located the data objects efficiently. Table 6 shows the number of the events that occurred in the repository browser with the number of data objects analyzed in each group.

We observed the average number of events in the repository browser per analyzed data object. As shown in Figure 19, the average number of the events substantially decreased for all the participants in Group 3 (configurations 1 and 4) and 6 (configurations 3 and 4). In the two groups, the common system configuration changes are with and without the mixed-initiative recommendation. As the participants in Group 3 and 6 efficiently identified data objects with the recommendation, the mixed-initiative interaction is potentially effective in terms of data location. To evaluate the effect of the mixed-initiative interaction, we examined and compared the average number of the events associated with system configurations. Since the participants in configuration 2 (without the visual workspace and with the mixed-initiative interaction) were reluctant to explore recommended data due to workspace unavailability, we examined the average number of events in configurations 3 and 4.

Table 6. Number of interactions in the repository browser (in bold) and number of data objects analyzed (in parentheses).

Group	Configuration	User 1	User 2	User 3	User 4
Group1	Config. 1	**172** (16)	**85** (7)	**263** (23)	**126** (15)
	Config. 2	**152** (18)	**64** (9)	**58** (45)	**421** (46)
Group2	Config. 1	**441** (9)	**238** (24)	**263** (13)	**103** (6)
	Config. 3	**194** (25)	**284** (31)	**138** (67)	**147** (32)
Group3	Config. 1	**414** (11)	**139** (2)	**234** (10)	**227** (4)
	Config. 4	**188** (57)	**152** (17)	**124** (59)	**174** (21)
Group4	Config. 2	**59** (2)	**43** (27)	**111** (4)	**262** (0)
	Config. 3	**157** (14)	**56** (22)	**103** (15)	**236** (24)
Group5	Config. 2	**398** (26)	**184** (15)	**58** (15)	**57** (7)
	Config. 4	**52** (27)	**136** (48)	**77** (28)	**72** (15)
Group6	Config. 3	**440** (52)	**248** (37)	**354** (41)	**92** (15)
	Config. 4	**94** (54)	**324** (95)	**103** (57)	**67** (24)

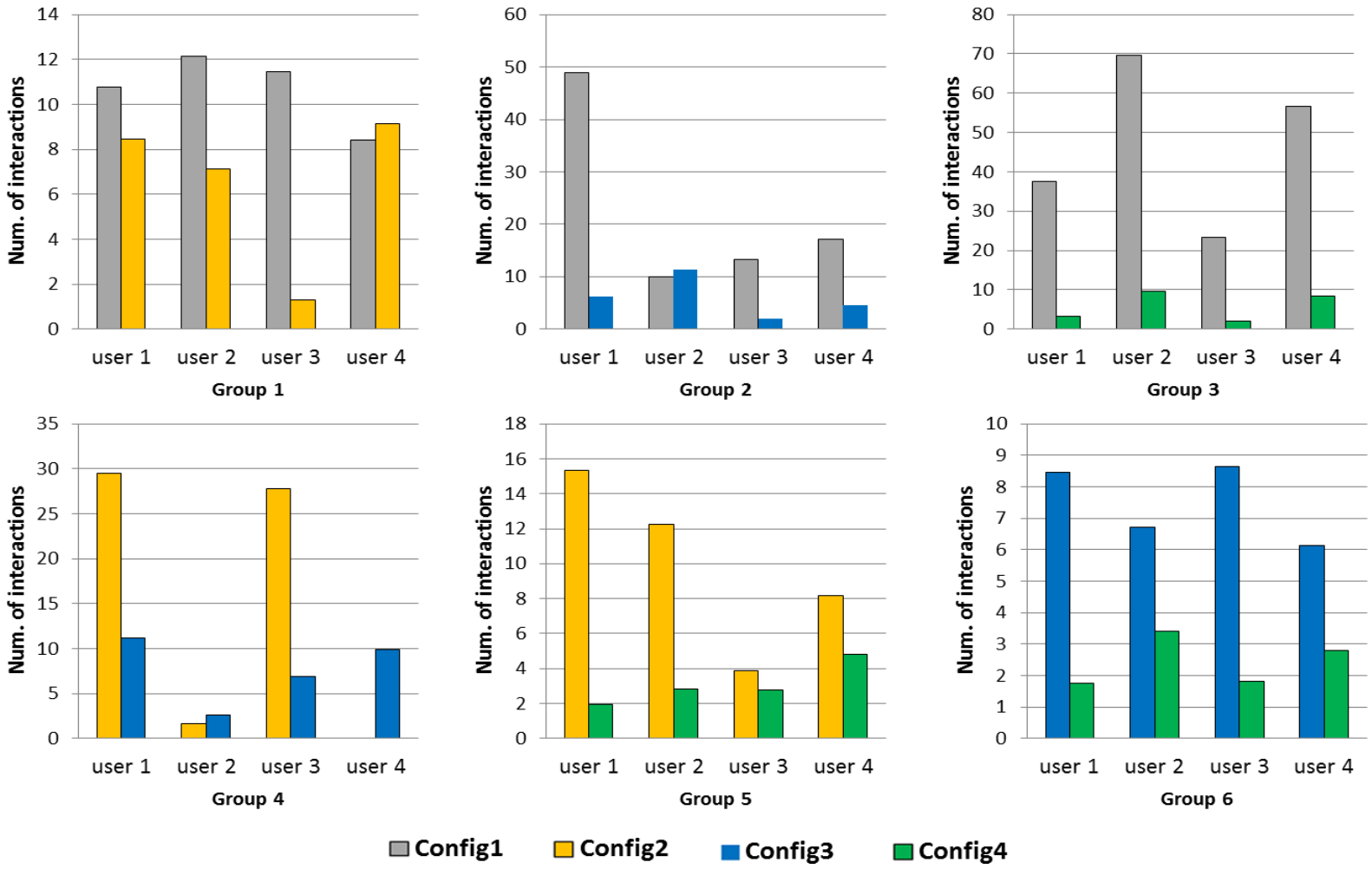


Figure 19. Average number of events in the repository browser for each data object analyzed.

Figure 20 shows how many events in the repository browser occurred with each analyzed data object under the configurations 3 and 4. Without the suggestions, the average was approximately 6.38. When suggestions were provided, the mean decreased to 3.72. This supports the interpretation that the mixed-initiative interaction suggested relevant data and helped the users to locate data of interest while performing the tasks.

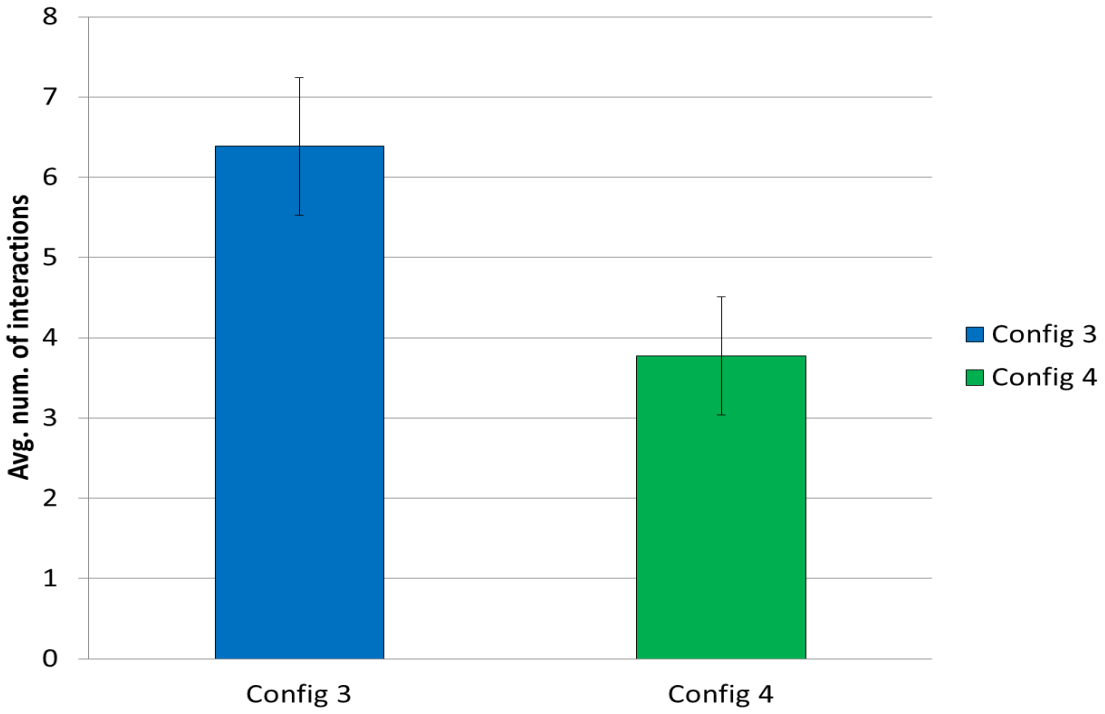


Figure 20. Average number of events in the repository browser per analyzed data object.

When comparing the number of activities in the repository browser for all the participants in configurations 3 and 4, the effect of the mixed-initiative interaction for the efficient data location is significant ($p = 0.025$, t-test, Bonferroni-adjusted

significance level = 0.025). In addition, the effectiveness of data location through the mixed-initiative recommendations is strong (Cohen's d = 0.981).

6.2.3 Distribution of Activity

Given the same tasks but different interface configurations, participants showed different work practices. Without the visual workspace and recommendations (configuration 1), 40% to 95% of the activities/interactions occurred in the repository browser (mean = 0.71, standard deviation (std) = 0.24). On the other hand, with the visual workspace and without recommendations (configuration 3), 11% to 69% of the activities were recorded in the repository browser (mean = 0.41, std= 0.21).

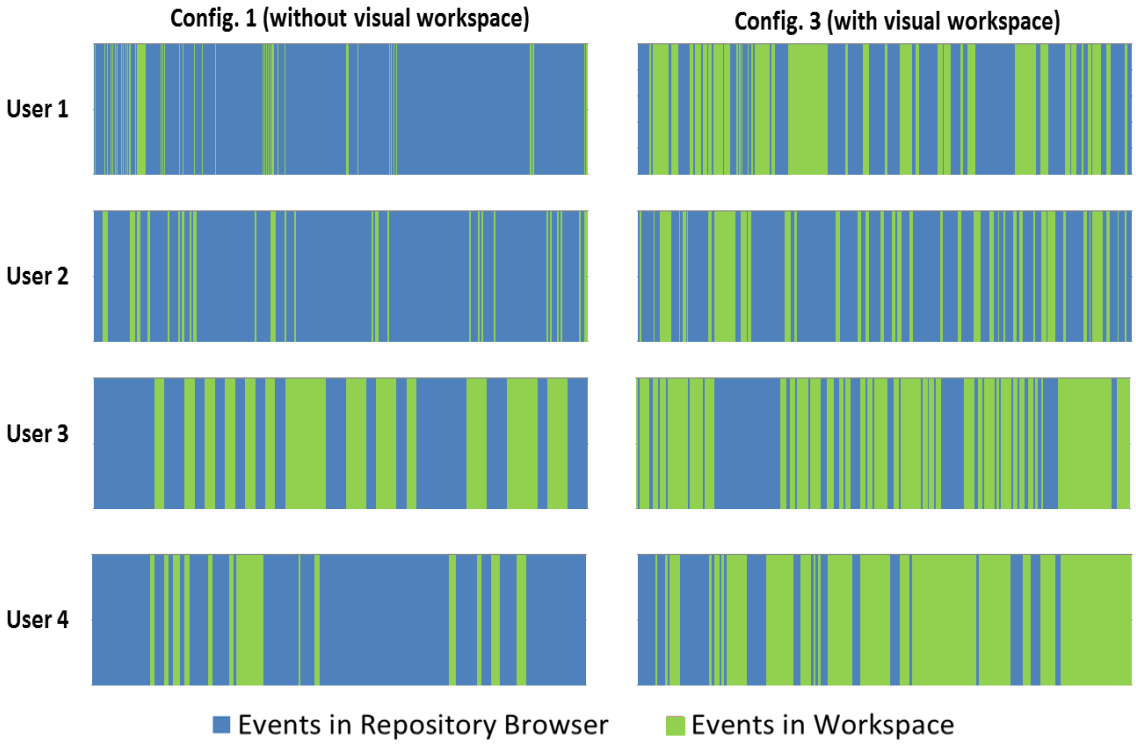


Figure 21. Ordering of user events in the repository browser and workspace shows distinct patterns of work.

Figure 21 shows the event sequences in the repository browser and the workspace for the four participants in Group 2. User1 and user3 were first exposed to configuration 1 and then configuration 3 (subgroup A in Table 2). On the other hand, user2 and user4 were first exposed to configuration 3 and then configuration 1 (subgroup B in Table 2). Regardless of the order of exposure, the availability, or lack thereof, of the workspace had a strong effect on individual work practices.

Without the visual workspace, users focused on and spent more time searching for data using data previews in the repository browser. They then relied on their short-term memory or notes taken in other applications to get back to the data. In particular, the user's short-term memory dependency was commented on by open-ended responses such as:

"*It was not helpful because I needed to find more information and memorize relationships among many variables*"

However, as shown in Figure 21, with the visual workspace, users spent more time exploring potential data relationships in the workspace than searching data of interest in the repository browser. To identify users' data activities/practices with the workspace in detail, we examined the workspace events of all the users in configurations 3 and 4.

In particular, we classified each event in the workspace as being either a data exploration event, one which provided access to or altered the visualization of data, or data interpretation event, where the user is expressing something about the data. Table 7 shows workspace event types and their descriptions.

Table 7. Workspace event types.

Event name	Description	Category of data activity
AddSymbol	Creates a data object (application)	Data exploration
Plot_Exp	Explores data within a data object	
MaximizeSymbol	Maximizes a data object or collection	
RestoreSymbol	Restores a maximized data object or collection	
DeleteSymbol	Deletes a data object or collection	
MoveSymbol	Moves a data object or collection (except moving a data object into a collection)	
ResizeSymbol	Resizes a data object or collection	
MoveToCollection	Moves a data object into a collection	Data interpretation
ChangeContent	Annotates or changes annotation on a data object (including notepad) or collection	
ChangeBackgroundColor	Changes background color of a data object or collection	
ChangeBorderWidth	Changes border line width of a data object or collection	
ChangeBorderColor	Changes border color of a data object or collection	
AddCollection	Creates a collection	
AddNotepad	Creates a notepad object (application) for annotation	

Figure 22. Number and percentage of the users' workspace activities of data exploration and data interpretation in configuration 3.

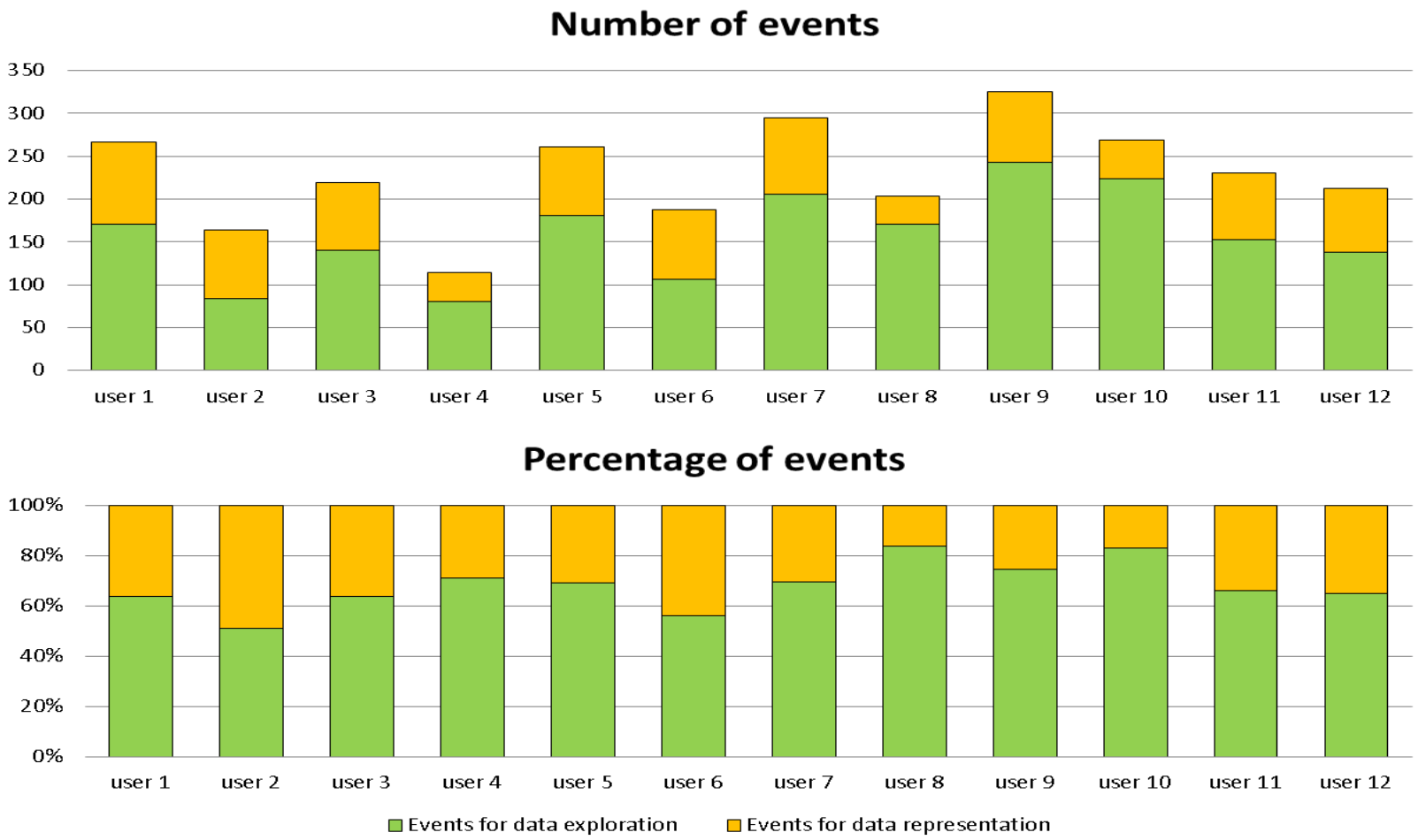


Figure 23. Number and percentage of the users' workspace activities of data exploration and data interpretation in configuration 4.

Along with the categorization, Figure 22 and Figure 23 show the distributions of the number and percentage of the users' workspace activities related to data exploration and data interpretation in configurations 3 and 4, respectively.

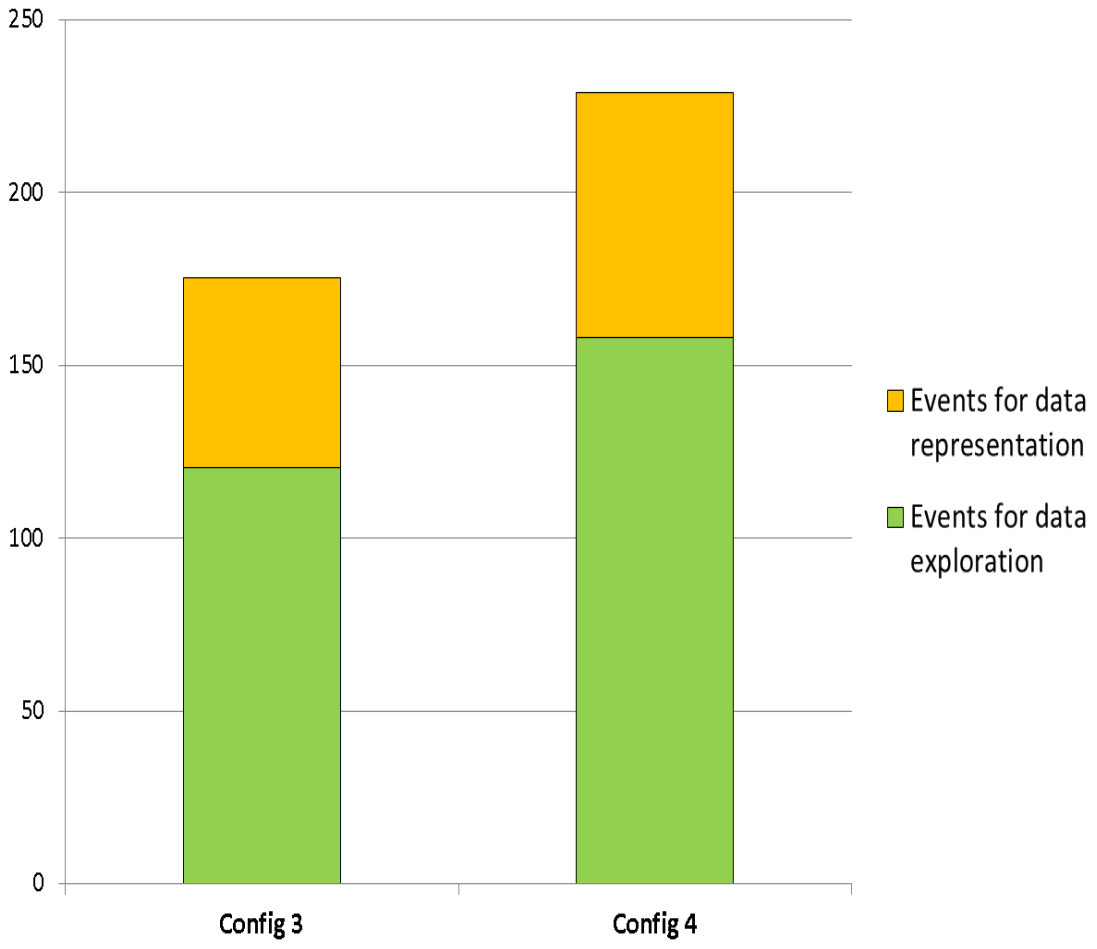


Figure 24. Average number of the occurred workspace events for data exploration and data interpretation in configurations 3 and 4.

As shown in Figure 22 and Figure 23, 11.3% to 54.1% and 16.2 % to 48.8% of the activities for data interpretation occurred in configurations 3 and 4, respectively

(mean = 0.31, std = 0.036 in configuration 3, mean = 0.32, std = 0.028 in configuration 4). After the initial data search in the repository browser, the visual workspace enabled the users to explore and manipulate refined data (e.g. 'Plot_Exp', or 'ResizeSymbol'). While the users explored the data objects, they first attempted to develop an understanding of the data contents and then represented the data relationships they identified; the data objects for correlated data elements were spatially and visually organized (e.g. 'MoveToCollection', 'ChangeBackgroundColor'), annotated (e.g. "ChangeContent') and interacted (e.g. "MoveSymbol') with in ways that made their relationships significant and meaningful.

Figure 24 shows the average number of the events in the workspace for the two configurations. Around 30% of the workspace events involved data interpretation. Notice that, when comparing the number of occurred workspace events in configuration 3 to that in configuration 4, the number with respect to data exploration and data representation increased by 31.3% and 29.5%, respectively. These increases account for the effect of mixed-initiative recommendations, which is covered in section 6.2.4.

The effects of the availability of the workspace on users can be inferred in comments for participants. When the configuration changed from "without the workspace" to "with the workspace", many of the corresponding participants (Groups 2, 3, 4, and 5 in Table 2) explicitly expressed their relief with data exploration and interpretation. But, for those participants who went from "with the workspace" to "without the workspace" conditions, comments indicated that they found it stressful to explore and interpret data in the second task.

6.2.4 Mixed-Initiative Recommendations

As already indicated, participants were more willing to make use of recommendations when their configuration had a workspace. Thus, to evaluate the effectiveness of recommendations, we examined how the twelve users of configuration 4 located data objects, which included both the workspace and recommendations. Figure 25 shows the distributions of the number of data examined associated with how the users located data objects: (1) the user explored/examined in the repository browser, (2) the user accepted user-requested recommendations, and (3) the user accepted system-provided recommendations. All twelve of the participants in this configuration accepted and explored some of the suggested data. Indeed, 37.5% to 55.9% (mean = 45.8, std = 9.6) of the data objects in their workspaces were from recommendations. All the users in the configuration used system-provided recommendations for their analyses and 7.1% to 51.6% (mean = 28.0, std = 13.1) of the data objects were from the system-provided recommendations. In addition, ten of the twelve participants actively used the ability to request suggestions. 5.3% to 37% (mean = 21.4, std = 10.7) of the data objects were from the user-requested recommendations. Both implicit (system-provided) and explicit (user-requested) inference-based recommendations were observed.

Along with how the users located the data objects, we also investigated workspace events related with suggestions. Figure 26 shows the temporal sequence of suggestion events triggered by the system and requested by users during the task. This shows that at least half of the participants requested suggestions fairly frequently over the 35 minute period. In addition, Figure 27 demonstrates another temporal sequence of

data creation events in the workspace when the users located the data objects. All the users started to explore the data objects in the repository browser, but, depending on individual work practices, events of data creation through user-requested or system-provided recommendations were interleaving with events when they explored the data in the repository browser. This implies the users were often encouraged to explore and identify the data context in the repository browser. This continued use implies that the recommendations were seen as being valuable. This was corroborated by open-ended comments such as:

"*When I wanted to find data with a similar pattern/trend, the recommendations reasonably provided me with data objects that resembled what I was looking for. There were other times when I wanted to have [data with new patterns] recommended, e.g. only considerable amounts of rain in College Station, and this is where I tended to have to look more for myself or hunt more in the requested recommendations to find what I was looking for.*"

Thus, through the mixed-initiative recommendations, PerCon helped users significantly reduce unnecessary additional effort in information seeking or exploration.

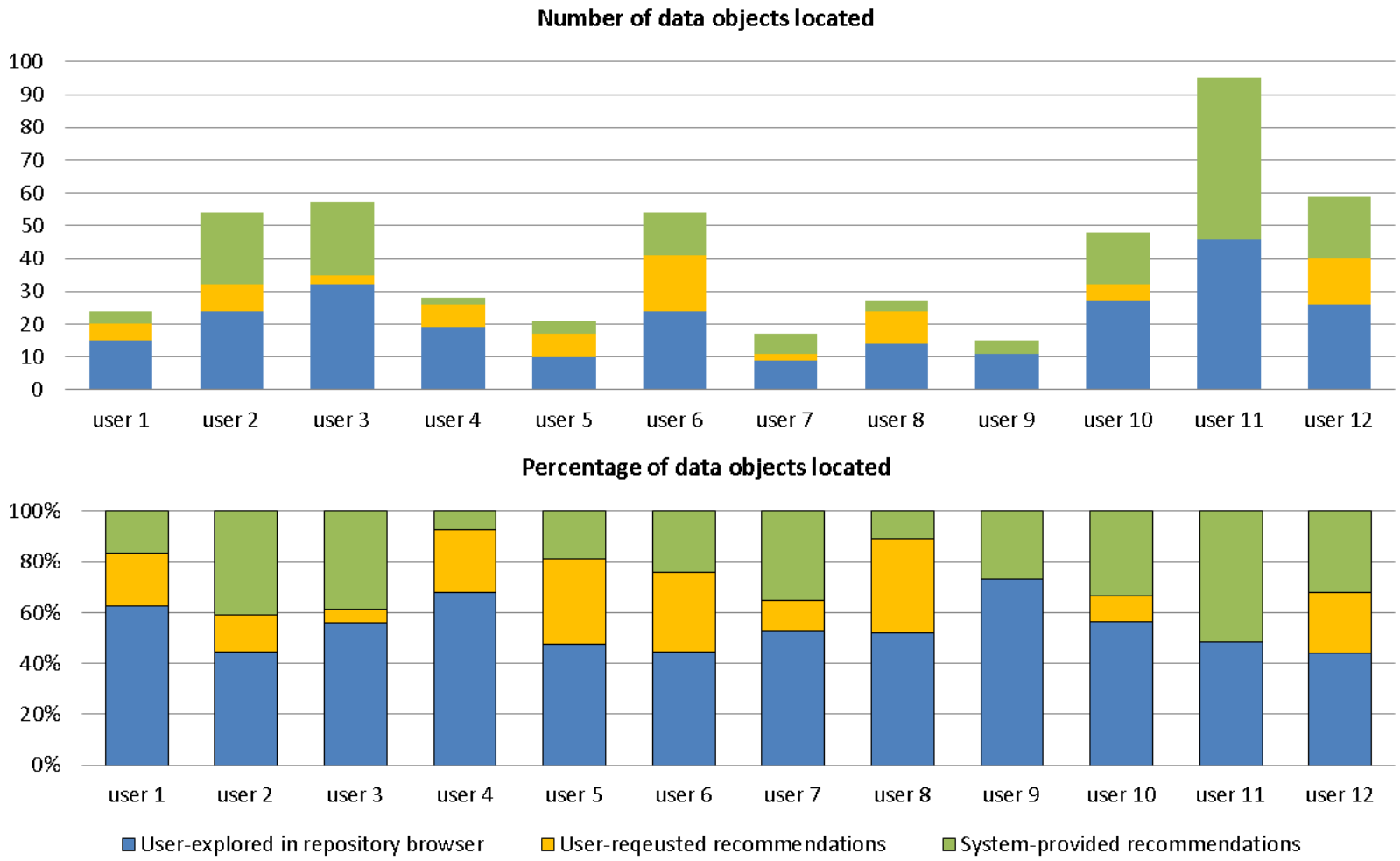


Figure 25. Number and percentage of data objects located.

Figure 26. A sequence of suggestion events.

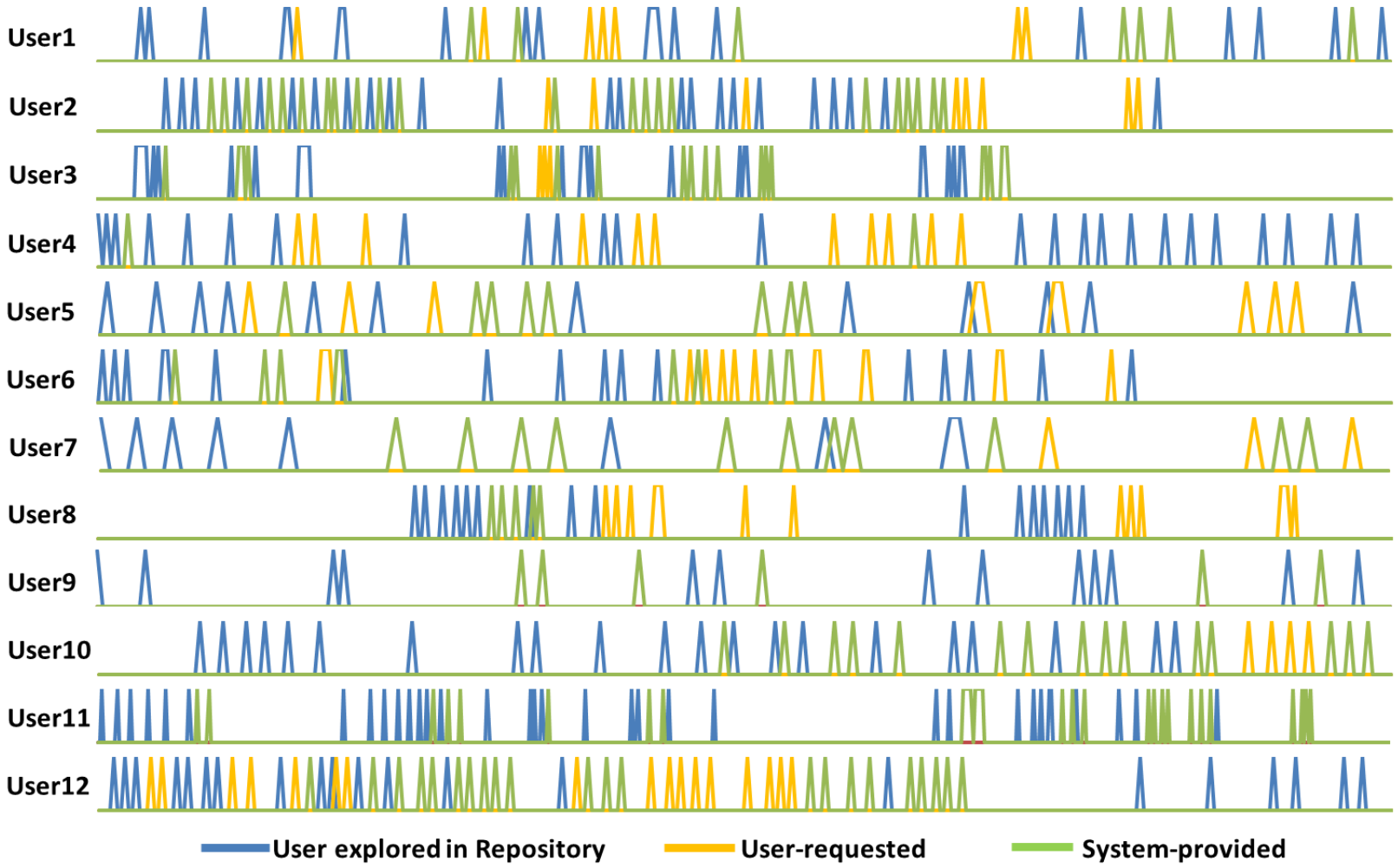


Figure 27. A sequence of data creation events in the workspace.

6.3 Effects of Other Interfaces

Besides the effects of the workspace and recommendations, through the user study we confirmed the design rationale and importance of the system interface and model. In particular, data visualization in the repository browser was valued for initial data exploration, filtering, and identification of the context. In addition, the integrated workspace model for unanticipated data types was justified in contrast to using individually separate applications.

6.3.1 Effect of Visualization in Repository Browser

In our questionnaire, a statement about the repository browser interface was included and the participants responded to the question of *"It was useful to explore data using a thumbnail preview (previsualization) in the repository view before instantiating it in the workspace."*

Figure 28 shows the average responses to the question. In the four configurations, the average varies between 5.9 and 6.5. Regardless of the configurations associated with the workspace or recommendation, the value of the data preview in the repository browser was assessed as important. Also, 35% to 97% of the activities/interactions occurred in the repository browser for all configurations and users. Functionally, the repository browser served as an interface of initial exploration, which led to a reduction in the workspace overload by previsualizing and filtering the objects of interest. Furthermore, beyond the intended design, it played the role of an interface between the workspace and recommendation. Rather than instantiating thumbnailed data (i.e. recommended data) from the suggestion browser to the workspace directly, the

participants showed common work practices of identifying the data context through the repository browser; they checked the information about the data (e.g., date, location, etc.), confirmed the data object with the preview, or explored neighbor data objects around the recommended data. As a result, the usefulness of data hierarchy and previsualization in the repository browser was demonstrated by the assessment and activities of the users.

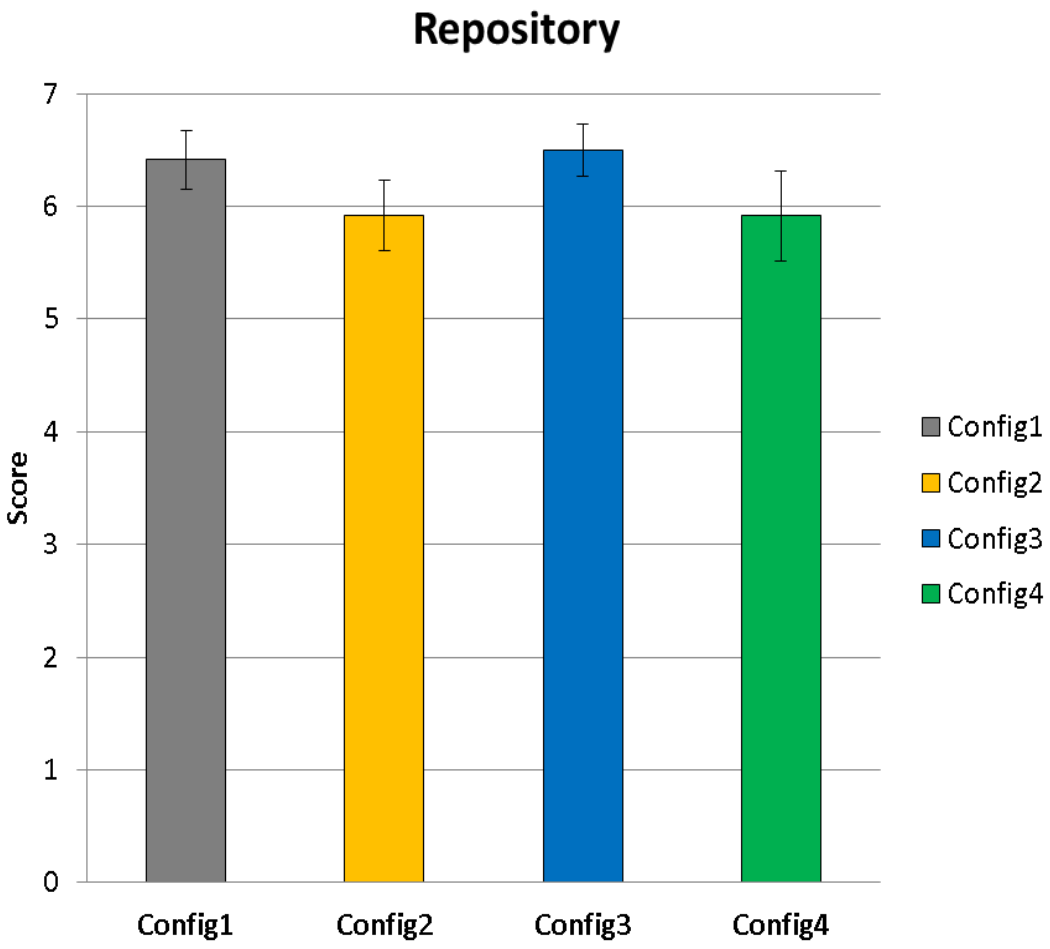


Figure 28. Reponses to a question to related to repository browser.

6.3.2 Resistance to Individual Applications/Representations

When we designed the workspace model in 2-D space, i.e. the separation of system base panel and application panel, our design rationale of the model was initiated to support various multiple representations or applications without constraints. In particular, we compared an integrated data environment that supports multi-applications to another conventional environment where users just use existing individual applications (when necessary). Along with the effect of workspace regarding the data exploration and interpretation, our user study strengthens the value and design rationale of the multi-application supported workspace.

Though the participants in configuration 1 or 2 (without the workspace) were allowed to freely use other applications such as MS Word or Notepad to annotate or write notes during the tasks, no one used the separate applications as a means of data analysis or to write intermediate results. The participants only used them as a final answer sheet relying on their short-term memory or confirming the same data objects by exploring in the repository browser several times. Some participants explicitly expressed that they wanted to annotate their intermediate findings using a PerCon notepad application. Namely, they tried to switch their focus minimally between PerCon and other applications. On the other hand, the participants in configuration 3 or 4 (with the workspace) did not show any resistance to annotating or writing their intermediate and final results on the PerCon notepad object.

Obviously, the individually available but separate applications can lead to cognitive overhead of or resistance to their usage. This justifies not only our workspace

model, but also strengthens the workspace effectiveness in the perspective of a data analysis platform.

7. CONCLUSION AND FUTURE WORK

We started this project with the goal of building an environment that would integrate much of the data ingestion, management, and analysis activities for an on-going data-intensive research project; the initial use of PerCon was in the domain of personal physiological, psychological, and contextual data using wearable sensors and devices. As such, the original system architecture and development focused on the system-side perspective of storing, processing, and visualizing data that were appropriate for the particular types of data. As development continued, our focus shifted to the human activities of locating, annotating, and interpreting data. Also, as new data collection, types, and formats were ingested to PerCon, the system has the potential to be valuable for many different research goals as a generic data platform.

In terms of data location, PerCon's main interface provides metadata-based access to the collection via the repository browser. A weakness of our initial implementation was that it just showed the limited metadata of the data elements (e.g. data type, date/time, location or name). This completely obscured the data contents, making it difficult to know what the data was really like until it was brought into the workspace by the user, resulting in workspace management issues (e.g. continually rearranging and deleting content). This observation led to the addition of thumbnails for previewing data in the repository browser.

---

*Part of this section is reprinted from the following paper: ©2014 IEEE. Reprinted, with permission, from Su Inn Park and Frank Shipman. PerCon: A personal digital library for heterogeneous data. In Proceedings of IEEE/ACM Joint Conference on Digital Libraries (JCDL), pages 97-106, IEEE, Sept. 2014.

In addition to browsing the data collection, PerCon supports queries that filter the content presented in the browser. Because many domains require understanding temporal relations among data elements, query results can also be examined in a calendar view. The calendar visualization uses a combination of labels and colorized thumbnails to present the types and contents of data elements.

PerCon is built around the main workspace for visualizing and interpreting data. The workspace enables the visualization of heterogeneous data types and the expression of interpretation through annotation and visual structure. The user study showed that the workspace has a large effect on data analysis practice in terms of the number of data elements classified, the time spent locating vs. time spent interpreting data, and the users' perceptions of system support. The ability to develop persistent task-oriented workspaces from data collections seems crucial to efficient and effective data analysis.

Data location is traditionally a user-driven activity. Our efforts to recommend data elements based on user activity in the workspace aimed to overcome the difficulty of users not knowing what data is available in a collection. The development of a multi-faceted approach to similarity assessment and probabilistic reasoning about user interests combine to generate recommendations. The user study showed that recommendations increased the number of data elements classified with and without the workspace and that many of the users valued the recommendations enough to take the initiative to ask for recommendations and to accept the recommended data.

There are a number of directions for future work: exploring the various workflows (e.g. [20]) surrounding data collection, ingestion, and analysis, adopting

available metadata standards and repository communication, improving workspace interactions, extending the recommendation subsystem, and exploring the user of PerCon in new domains and with new user communities.

With the ability to add most any Java application as an element in the workspace, PerCon has the potential for broad use considering interoperability with other scientific data. In practice, most data analysis proceeds via a simple graphing object type. We want to explore data visualization objects that are more dynamic and tailorable. For example, we envision workspace objects that synchronize the presentations of time-aligned data elements. Another example is extending the data graph object to allow users to merge data elements into a single presentation. If the user drags one data graph object into another data graph object, the two are presented in a single graph. This new data object type would allow the user to pull a data line back out of the graph to separate the two again. This mode of interaction is common in tabbed web browsers where users can put windows together and pull tabs out to create new windows.

The recommendation subsystem shows the potential for data analysis environments to be more proactive in supporting users. The current approach to recommendation generation assumes the user wants more data of the same data types with similar characteristics to those already included in the workspace. A natural extension is to look for similarities across data types. Because the phenomena being measured in the data are often dissimilar, an alternative similarity metric is likely to be needed.

PerCon was originally designed with the expectation that users are either professionals engaged in this type of work or are otherwise knowledgeable about data types and data manipulations. The user study showed the system was usable by such a population. We would like to explore the use of a PerCon-like interface for more casual use in the domain of personal health. Enabling people to explore the data collected on their smart watches, health monitoring devices, cell phones, etc. could lead to improved mental models about how their lifestyle and health are intertwined.

REFERENCES

[1] Ahmed Abbasi and Hsinchun Chen. Categorization and analysis of text in computer mediated communication archives using visualization. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, pages 11-18, 2007.

[2] Neal Audenaert and Richard Furuta. What humanists want: How scholars use source materials. In Proceedings of the 10th Annual Joint Conference on Digital Libraries, pages 283-292, 2010.

[3] Neal Audenaert, George Lucchese, and Richard Furuta. Critspace: A workspace for critical engagement within cultural heritage digital libraries. In Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries, pages 307-314, 2010.

[4] Soonil Bae, DoHyoung Kim, Konstantinos Meintanis, Michael Moore, Anna Zacchi, Frank Shipman, Haowei Hsieh, and Catherine Marshall. Supporting document triage via annotation-based multi-application visualizations. In Proceedings of the 10th Annual Joint Conference on Digital Libraries, pages 177-186, 2010.

[5] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. Nucleic Acids Research, 41(D1):D36-D42, 2013.

[6] Jürgen Bernard, Tobias Ruppert, Maximilian Scherer, Jörn Kohlhammer, and Tobias Schreck. Content-based layouts for exploratory metadata search in scientific research data. In Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, pages 139-148, 2012.

[7] Ann Blandford, Nick Bryan-Kinns, and Harold Thimbleby. Interaction modelling for digital libraries. In Workshop on Evaluation of Information Management Systems, pages 1-10, 2000.

[8] John Booker, Timothy Buennemeyer, Andrew Sabri, and Chris North. High resolution displays enhancing geo-temporal data visualizations. In Proceedings of the 45th Annual Southeast Regional Conference, pages 443-448, 2007.

[9] Christine L Borgman, Jillian C Wallis, Matthew S Mayernik, and Alberto Pepe. Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, pages 269-277, 2007.

[10] Leonardo Candela, Donatella Castelli, Nicola Ferro, Georgia Koutrika, Carlo Meghini, Pasquale Pagano, Seamus Ross, Dagobert Soergel, Maristella Agosti, and Milena Dobreva. The DELOS digital library reference model. Foundations for Digital Libraries (version 0.98). 2008.

[11] Rick Cattell. Scalable sql and nosql data stores. ACM SIGMOD Record, 39(4):12-27, 2011.

[12] Mark Claypool, Phong Le, Makoto Wased, and David Brown. Implicit interest indicators. In Proceedings of the 6th International Conference on Intelligent User Interfaces, pages 33-40, 2001.

[13] Constanze Curdt, Dirk Homeister, Christian Jekel, Sebastian Brocks, GuidoWaldhoff, and Georg Bareth. TR32DB-Management and visualization of heterogeneous scientific data. In Proceedings of the 19th International Conference on Geoinformatics, pages 1-6, 2011.

[14] Edward W Davis. Application of the massively parallel processor to database management systems. In Proceedings of National Computer Conference, pages 299-307, 1983.

[15] Ruxandra Domenig and Klaus R Dittrich. A query based approach for integrating heterogeneous data sources. In Proceedings of the 9th International Conference on Information and Knowledge Management, pages 453-460, 2000.

[16] Edward A Fox, Robert Hall, and Neill Kipp. NDLTD: Preparing the next generation of scholars for the information age. New Review of Information Networking, 3(1):59-76, 1997.

[17] Michael Franklin, Alon Halevy, and David Maier. From databases to dataspaces: A new abstraction for information management. ACM Sigmod Record, 34(4):27-33, 2005.

[18] Marcos Andre Goncalves, Edward A Fox, Layne T Watson, and Neill A Kipp. Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. ACM Transactions on Information Systems, 22(2):270-312, 2004.

[19] Christopher G Healey, Sarat Kocherlakota, Vivek Rao, Reshma Mehta, and Robert St Amant. Visual perception and mixed-initiative interaction for assisted visualization design. IEEE Transactions on Visualization and Computer Graphics, 14(2):396-411, 2008.

[20] Jeffrey Heer and Sean Kandel. Interactive analysis of big data. XRDS: Crossroads, The ACM Magazine for Students, 19(1):50-54, 2012.

[21] EA Henneken, A Accomazzi, CS Grant, MJ Kurtz, D Thompson, E Bohlen, and SS Murray. The SAO/NASA astrophysics data system: A gateway to the planetary sciences literature. In Proceedings of Lunar and Planetary Science Conference, page 1873, 2009.

[22] Eric Horvitz. Principles of mixed-initiative user interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 159-166, 1999.

[23] Intel. Intel Healthcare and Life Sciences. Accessed Feb. 2015. http://www.intel.com/content/www/us/en/healthcare-it/healthcare-overview.html.

[24] Dongyeop Kang and Ho-Jin Choi. Robot task planning for mixed-initiative human robot interaction in home service robot. In Workshops on IEEE Computer and Information Technology, pages 49-54, 2008.

[25] Carola Kanz, Philippe Aldebert, Nicola Althorpe, Wendy Baker, Alastair Baldwin, Kirsty Bates, Paul Browne, Alexandra van den Broek, Matias Castro, Guy Cochrane, *et al*. The EMBL nucleotide sequence database. Nucleic Acids Research, 33(suppl. 1):D29-D33, 2005.

[26] Thorsten Karrer, Malte Weiss, Eric Lee, and Jan Borchers. Dragon: A direct manipulation interface for frame-accurate in-scene video navigation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 247-250, 2008.

[27] Daniel A Keim. Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics, 8(1):1-8, 2002.

[28] Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: A bibliography. ACM SIGIR Forum, 37(2):18-28, 2003.

[29] Andruid Kerne, Eunyee Koh, Steven M Smith, Andrew Webb, and Blake Dworaczyk. Combinformation: Mixed-initiative composition of image and text surrogates promotes information discovery. ACM Transactions on Information Systems, 27(1):5, 2008.

[30] Steve Lawrence, C Lee Giles, and Kurt Bollacker. Digital libraries and autonomous citation indexing. Computer, 32(6):67-71, 1999.

[31] Jae W Lee, Jianting Zhang, Ann S Zimmerman, and Angelo Lucia. DataNet: An emerging cyberinfrastructure for sharing, reusing and preserving digital data for scientific discovery and learning. AIChE Journal, 55(11):2757-2764, 2009.

[32] Huajing Li, Isaac G Councill, Levent Bolelli, Ding Zhou, Yang Song, Wang-Chien Lee, Anand Sivasubramaniam, and C Lee Giles. Citeseer : A scalable autonomous scientic digital library. In Proceedings of the 1st International Conference on Scalable Information Systems, page 18, 2006.

[33] Wei Luo, Alan M MacEachren, Peifeng Yin, and Frank Hardisty. Spatialsocial network visualization for exploratory data analysis. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-based Social Networks, pages 65-68, 2011.

[34] Alexa T McCray and Marie E Gallagher. Principles for digital library development. Communications of the ACM, 44(5):48-54, 2001.

[35] Robert E McGrath, Joe Futrelle, Ray Plante, and Damien Guillaume. Digital library technology for locating and accessing scientific data. In Proceedings of the 4th ACM Conference on Digital Libraries, pages 188-194, 1999.

[36] Microsoft. Microsoft HealthVault. Accessed Feb. 2015. http://www.healthvault.com/us/en.

[37] Thomas Moser, Richard Mordinyi, and Stefan Bi. An ontology-based methodology for supporting knowledge-intensive multi-discipline engineering processes. In Ontology-Driven Software Engineering, page 2, 2010.

[38] Edward T O'Neill, Brian F Lavoie, Rick Bennett, Thornton Staples, Ross Wayland, Sandra Payette, Makx Dekkers, Stuart Weibel, Sam Searle, Dave Thompson, *et al*. Trends in the evolution of the public web, 1998-2002; The fedora project: An open-source digital object repository management system; State of the dublin core metadata initiative, April 2003; Preservation metadata; How many people search the eric database each day?. D-lib Magazine, 9(4):n4, 2003.

[39] Patricia Ordonez, Tim Oates, ME Lombardi, Genaro Hernandez, Kathryn W Holmes, Jim Fackler, and Christoph U Lehmann. Visualization of multivariate time-series data in a neonatal ICU. IBM Journal of Research and Development, 56(5):7-1, 2012.

[40] Adam Perer and Ben Shneiderman. Integrating statistics and visualization: Case studies of gaining clarity during exploratory data analysis. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 265-274, 2008.

[41] Saverio Perugini and Naren Ramakrishnan. Personalizing web sites with mixed-initiative interaction. IT Professional, 5(2):9-15, 2003.

[42] Hamid Pirahesh, C Mohan, Josephine Cheng, TS Liu, and Pat Selinger. Parallelism in relational data base systems: Architectural issues and design approaches. In Proceedings of the 2nd International Symposium on Databases in Parallel and Distributed Systems, pages 4-29, 1990.

[43] Unni Ravindranathan, Rao Shen, Marcos Andre Goncalves, Weiguo Fan, Edward A Fox, and James W Flanagan. Etana-dl: A digital library for integrated handling of heterogeneous archaeological data. In Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, pages 76-77, 2004.

[44] Jeremy Rowe, Anshuman Razdan, and Arleyn Simon. Acquisition, representation, query and analysis of spatial data: A demonstration 3d digital library. In Proceedings of Joint Conference on Digital Libraries, pages 147-158, 2003.

[45] Frank M Shipman III, Haowei Hsieh, Preetam Maloor, and J Michael Moore. The visual knowledge builder: A second generation spatial hypertext. In Proceedings of the 12th ACM Conference on Hypertext and Hypermedia, pages 113-122, 2001.

[46] Frank M Shipman III and Catherine C Marshall. Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems. Computer Supported Cooperative Work, 8(4):333-352, 1999.

[47] Frank M Shipman III and Raymond J McCall. Incremental formalization with the hyper-object substrate. ACM Transactions on Information Systems, 17(2):199-227, 1999.

[48] Ben Shneiderman, David Feldman, Anne Rose, and Xavier Ferre Grau. Visualizing digital library search results with categorical and hierarchical axes. In Proceedings of the 5th ACM Conference on Digital Libraries, pages 57-66, 2000.

[49] Ben Shneiderman and Pattie Maes. Direct manipulation vs. interface agents. Interactions, 4(6):42-61, 1997.

[50] Terence R Smith and James Frew. Alexandria digital library. Communications of the ACM, 38(4):61-62, 1995.

[51] Robert St Amant and Paul R Cohen. Interaction with a mixed-initiative system for exploratory data analysis. In Proceedings of the 2nd International Conference on Intelligent User Interfaces, pages 15-22, 1997.

[52] State of Texas. Brazos River Authority. Accessed Feb. 2015. http://www.brazos.org/USGSGaugingSystem.asp/.

[53] Lucy A Suchman. Office procedure as practical action: Models of work and system design. ACM Transactions on Information Systems, 1(4):320-328, 1983.

[54] Robert Tansley, Mick Bass, David Stuve, Margret Branschofsky, Daniel Chudnov, Greg McClellan, and MacKenzie Smith. The dspace institutional digital repository system: current functionality. In Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, pages 87-97, 2003.

[55] Loren G Terveen. Intelligent systems as cooperative systems. Journal of Intelligent Systems, 3(2-4):217-250, 1993.

[56] Loren G Terveen. Overview of human-computer collaboration. Knowledge-Based Systems, 8(2):67-81, 1995.

[57] Giannis Tsakonas, Sarantos Kapidakis, and Christos Papatheodorou. Evaluation of user interaction in digital libraries. In Notes of the DELOS WP7 Workshop on the Evaluation of digital libraries, 2004.

[58] United States Department of Commerce. National Oceanic and Atmospheric Administration. Accessed Aug. 2015. http://www.weather.gov/climate/.

[59] Howard D. Wactlar, Michael G. Christel, Yihong Gong, and Alexander G. Hauptmann. Lessons learned from building a terabyte digital video library. Computer, 32(2):66-73, 1999.

[60] Jillian C Wallis, Matthew S Mayernik, Christine L Borgman, and Alberto Pepe. Digital libraries for scientific data discovery and reuse: From vision to practical reality. In Proceedings of the 10th Annual Joint Conference on Digital Libraries, pages 333-340, 2010.

[61] Chris Weaver. Cross-filtered views for multidimensional visual analysis. IEEE Transactions on Visualization and Computer Graphics, 16(2):192-204, 2010.

[62] Ian H Witten, Stefan J Boddie, David Bainbridge, and Rodger J McNab. Greenstone: A comprehensive open-source digital library software system. In Proceedings of the 5th ACM Conference on Digital Libraries, pages 113-121, 2000.

[63] Pak Chung Wong, Han-Wei Shen, Christopher R Johnson, Chaomei Chen, and Robert B Ross. The top 10 challenges in extreme-scale visual analytics. IEEE computer graphics and applications, 32(4):63, 2012.

[64] Hong Iris Xie. Shifts of interactive intentions and information-seeking strategies in interactive information retrieval. Journal of the American Society for Information Science, 51(9):841-857, 2000.

[65] Yan Zhang, Ruisheng Zhang, Qiuqiang Chen, Xiaopan Gao, Rongjing Hu, Ying Zhang, and Guangcai Liu. A hadoop-based massive molecular data storage solution for virtual screening. In Proceedings of the 7th Annual Conference on ChinaGrid, pages 142-147, 2012.

APPENDIX A

IRB Protocol Summary (IRB2014-0115D)

PerCon: Support for Weather and River Data Management and Analysis

Background and Rationale for Study:

Weather data include various variables that exhibit a great deal of spatial and temporal correlations with one another. Each variable in weather data, such as precipitation, wind, humidity, and temperature, has an impact on environmental variables. River stream level is a representative environmental variable significantly affected by weather conditions. Depending on geographic relationships (i.e. relative location), river stream levels exhibit a strong relation (engagement) between upstream and downstream.

Public repositories such as the National Oceanic and Atmospheric Administration (NOAA) provide weather data to the public via web services. State water organizations such as the Brazos River Authority in Texas provide river level data for comprehensive monitoring and water quality management. Thus, weather and river data collections are easily accessible. However, even though a large amount of the data collection is available, its management and analysis pose many difficulties. Since data measurements are in numeric or text format, exploring and examining the meaning of each data measurement is difficult. Even though each data can be viewed with proper visualization tools (e.g. graphs and color maps on web browsers), sharing and intercommunicating data between tool instances may not be possible or may demand expensive effort.

Research Design:

To manage and analyze these heterogeneous and interrelated weather data and other related data, we have developed a digital library system called PerCon (**Per**sonalized and **Con**textual Data Environment). PerCon pre-processes raw weather and river data. It provides users with a visual workspace in which to visualize both weather and environmental data and to express interpretations of the data. Thus, the system enables users to build their own knowledge space to interpret visualized data. Interactively, the user-created information and knowledge in the workspace are shared with the system as evidence to infer a user's goal or interest. In order to enhance data exploration capability among a large amount of data, PerCon enables a user to view data before generating data objects in the workspace. To support the analysis, users can examine the data in a variety of resolutions or scales. PerCon adopts a collaborative interaction with the user; it is designed to support a mixed-initiative interaction for data analysis. PerCon monitors user activity with a goal of understanding user interest. It also analyzes the data in order to identify similarity, correlation, and clusters of data variables. The user activity and

analysis of data together are used to automatically recommend additional data to the user. Users can also request suggestions if desired.

Data Being Analyzed:

For this user study, PerCon has processed 2 years of recorded weather and river data (from 2011 Oct. to 2013 Oct.). The weather data includes temperature, precipitation, relative humidity, wind speed, and wet bulb temperature, measured at six different weather observation stations in Texas (Dataset1 collected from College Station, Waco, and Temple, Dataset2 collected from South Bend, Seymour, and Fort Griffin). The river data includes the river level and discharge recorded from the Brazos River near the aforementioned weather stations. To facilitate data management and analysis, various granular levels of data are provided. A data directory of each data element is provided with annual, monthly, and daily (hourly) data available within a data hierarchy.

The user study involves three specific tasks with respect to using weather and river data. The tasks are: (1) classifying and organizing the data in the workspace, (2) investigating and identifying data correlations, and (3) interpreting and estimating data events. To perform the tasks, a system user manual (tutorial) and a user task sheet will be given to participants.

Task Procedure:

Before conducting given tasks, participants will be given a user manual which includes a 10-minute video clip to understand PerCon. Then, they will have an additional 5-minute trial and learning time to practice how to use it.

For the study, there are four different system interface modes depending on the availability of visual workspace and mixed-initiative interaction: the visual workspace and mixed-initiative interaction will and will not be provided, respectively. Table 1 shows evaluation groups which are all permutations of two different interface modes (considering the order of interface modes). The participants will be randomly assigned to one of the groups. In each group, two system interfaces will be evaluated and the same three tasks will be asked to the participants in each interface mode. The entire assignments to each group will have equal numbers of the participants to be balanced. The given weather and river datasets in the first and second tasks will be equivalent: the only difference is the geographic location of data recorded. In brief, after learning about the system, the participants will be asked to perform the three tasks in each interface mode according to the group they belong to.

Table 1. User study groups and interface modes

	Tasks with dataset1	Tasks with dataset2
Group 1	Mode 1	Mode 2
Group 2	Mode 1	Mode 3
Group 3	Mode 1	Mode 4
Group 4	Mode 2	Mode 1

Group 5	Mode 2	Mode 3
Group 6	Mode 2	Mode 4
Group 7	Mode 3	Mode 1
Group 8	Mode 3	Mode 2
Group 9	Mode 3	Mode 4
Group 10	Mode 4	Mode 1
Group 11	Mode 4	Mode 2
Group 12	Mode 4	Mode 3

* Mode 1: The **visual workspace** and the **mixed-initiative interaction** are both **unavailable**
Mode 2: The **visual workspace** is **unavailable** but the **mixed-initiative interaction** is **available**
Mode 3: The **visual workspace** is available but the **mixed-initiative interaction** is unavailable
Mode 4: The **visual workspace** and the **mixed-initiative interaction** are both **available**.

The following four figures (Figure 1, 2, 3, and 4) describe the characteristics of the workspace and suggestion panel according to the system interface modes.

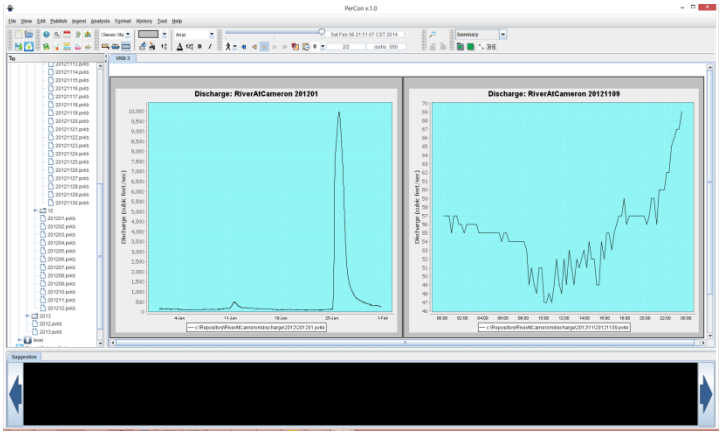


Figure 1. Configuration 1: The visual workspace and the mixed-initiative interaction are both unavailable. PerCon allows users to view up to two information objects in the workspace to explore weather and river data.

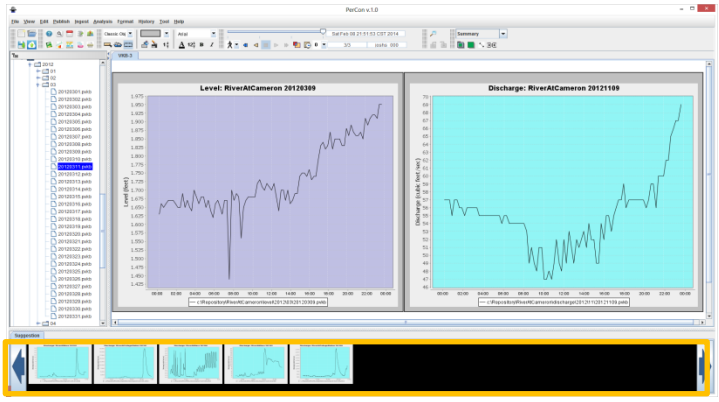


Figure 2. Configuration 2: The visual workspace is unavailable but the mixed-initiative interaction is available. Compared with Configuration1, data suggestion is available.

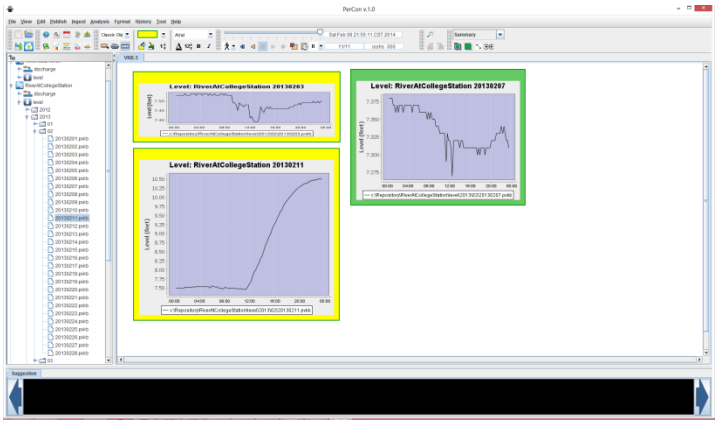


Figure 3. Configuration 3: The visual workspace is available but the mixed-initiative interaction is unavailable. PerCon allows users to use all features of the visual workspace. PerCon does not suggest any data to users.

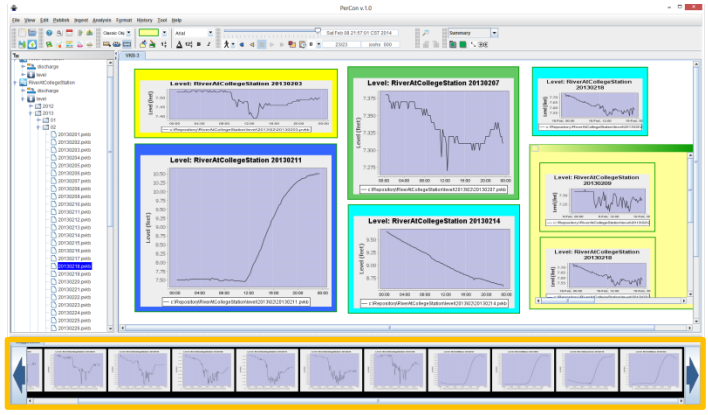


Figure 4. Configuration 4: The visual workspace and the mixed-initiative interaction are both available. PerCon allows users to use all features in PerCon.

Once participants are done with the given tasks, they will be asked to answer a questionnaire for about 20 minutes.

*When they perform the tasks **without the visual workspace**, they can use a text editor such as MS word or notepad to write down task results including data object name(s) for identification. After completing each task, participants will need to **save their workspace** (File – Save in Menu or ctrl + 's') and include each task number in the saved file name. Total estimated time to complete the tasks is 70 minutes.

APPENDIX B

TASK SHEET

Task 1. Classifying and organizing data (40 minutes)

To observe and discover how weather and river data are correlated, one of the possible initial approaches is to classify and organize data objects based on individual changes, trends, and patterns in the data.

With Dataset 1

1-1. Classify and organize **daily** river level and precipitation data according to its changes, trends, quantity, duration or user-perceived criteria. You can use "Collection" to classify similar datasets or to represent classification hierarchy. (20 minutes)

With Dataset 2

1-2. Classify and organize **daily** river level and precipitation data according to its changes, trends, quantity, duration or user-perceived criteria. You can use "Collection" to classify similar datasets or to represent classification hierarchy. (20 minutes)

Task 2. Investigating and identifying data correlation (20 minutes)

With the classified weather and river data from Task 1, we would like you to

- Investigate the implications
- Identify correlations among the classified data
  (e.g. how river level is affected by various weather components and other river data)

With Dataset 1

2-1. Based on your workspace data, investigate what and how weather factor(s) affects river level. Also, investigate how rivers at different places (Waco, Cameron, and College Station) are correlated. Briefly explain the rationale with evidence and explain how they are related. (10 minutes)

With Dataset 2

2-2. Based on your workspace data, investigate what and how weather factor(s) affects river level. Also, investigate how rivers at different places (South Bend, Cameron, and College Station) are correlated. Briefly explain the rationale with evidence and explain how they are related. (10 minutes)

Task 3. Interpreting and estimating river data events/causes (10 minutes)

If you found evidence of correlation and identified various possible impacts between the weather and river data, we would like you to interpret river level changes/trends under the given weather and upstream river flow conditions. Also, we expect you to estimate river factors that causes these changes/trends (such as delay time) based on past weather data and other river stream conditions. In addition, estimate why some factor(s) are more or less affected by river level changes.

With Dataset 1

3-1. Based on your workspace data, interpret the river level changes. Also, estimate the (average) time delay regarding the flow if you find any. Briefly explain the changes considering weather factors and other river stream flows. (5 minutes)

With Dataset 2

3-2. Based on your workspace data, interpret the river level changes. Also, estimate the (average) time delay regarding the flow if you find any. Briefly explain the changes considering weather factors and other river stream flows. (5 minutes)

APPENDIX C

USER MANUAL

PerCon: PERSONALIZED & CONTEXTUAL DATA ENVIRONMENT

THE USER'S MAUNAL

Version 1.0

CENTER FOR THE STUDY OF DIGITAL LIBRARIES

TEXAS A&M UNIVESRITY

1

Chapter 1: About PerCon

Per*sonal and* ***Con****textual Data Environment (****PerCon****) is designed to support the management and analysis of heterogeneous weather data with other potentially related data (e.g. environmental data). PerCon provides an integrated environment for managing and analyzing those data.*

A vast amount of research has been performed on managing and analyzing weather data. It involves studying weather data trends and impacts on the environment. These types of studies pose many difficulties such as analyzing massively recorded data and gross differences between datasets. Thus, the design and development of an environment to support weather data management and analysis is indispensable.

PerCon is a combination of a digital library and a data analysis platform for weather data and other potentially related data (e.g. river). As an information infrastructure, PerCon manages, accesses, preserves, and shares various types of weather data and information objects created by individual research participants. In addition, to enhance user's data analysis ability, PerCon suggests a relevant dataset according to user's interests or tasks.

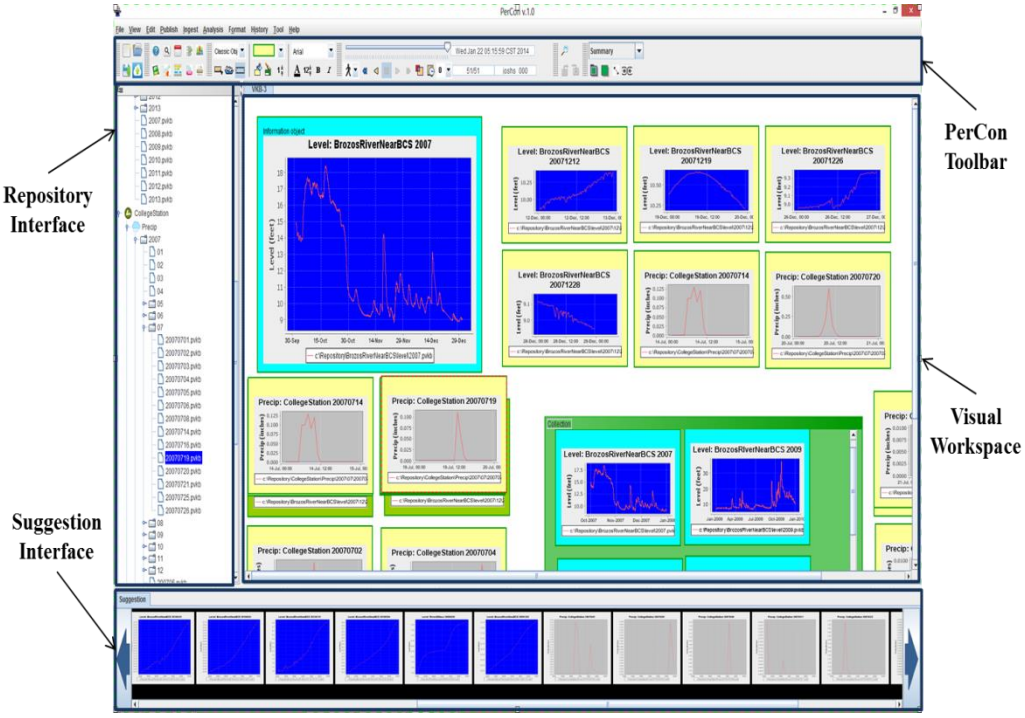
The following two chapters describe the main user interfaces and components in PerCon and provide instructions of how to use PerCon with a video demo.

Chapter 2: Getting Started

In this chapter, we describe PerCon user interfaces. The user interface involves three main components for repository view (data source view), visual workspace, and data suggestion view. Detailed features of each user interface component are described.

PerCon User Interface

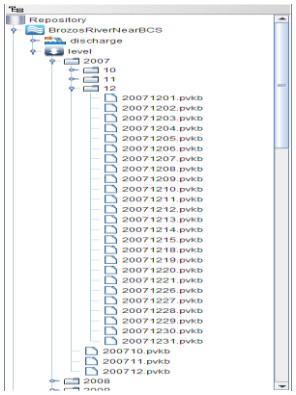
The PerCon user interface involves three major components: the repository view, the visual workspace, and the suggestion view. Each component is depicted in the below figure.



PerCon Desktop Application (User Interface)

Repository View

A repository stores and preserves (1) the original weather data objects (e.g. data streams), and (2) computed and filtered datasets. The interface for the repository displays a hierarchical structure of the data objects. The hierarchy levels relate to data observation station (location), data type/element, and recorded date and its period.



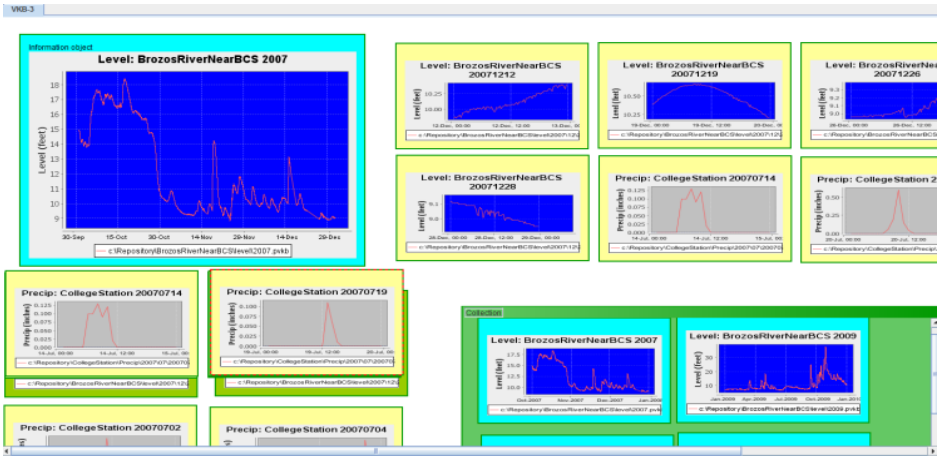
Visual Workspace

The PerCon workspace embeds the designs of Visual Knowledge Builder (VKB) workspace and its features. VKB's user manual or web tutorial can be found here:

http://www.csdl.tamu.edu/vkb/Download/VKBManual.PDF (user manual v0.7)

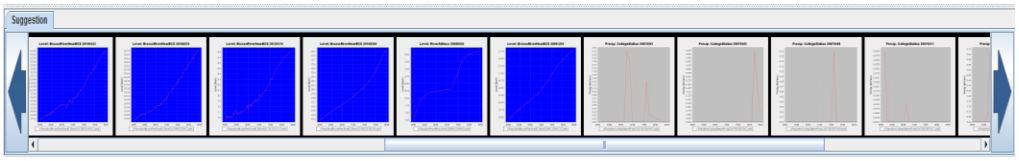
http://www.csdl.tamu.edu/vkb/tutorial/UsersTutorialFinal.html (web tutorial)

The visual workspace allows users to visualize data and organize information. Once a user queries data through a mouse operation or a menu selection, the proper data is visualized.

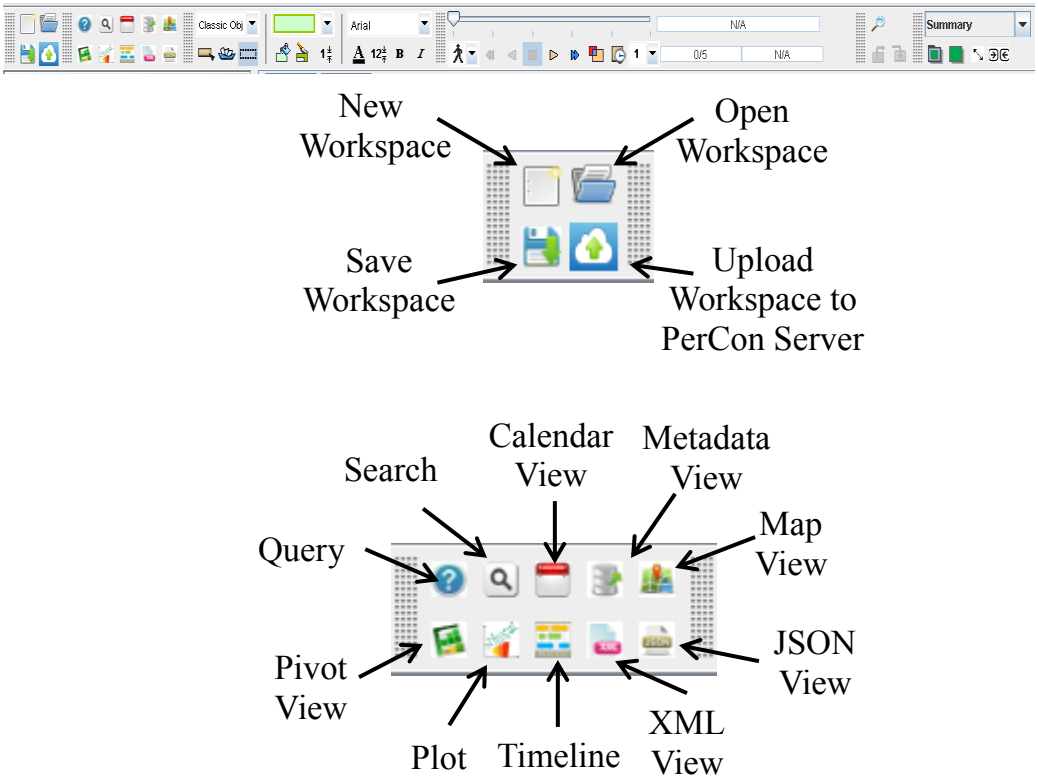


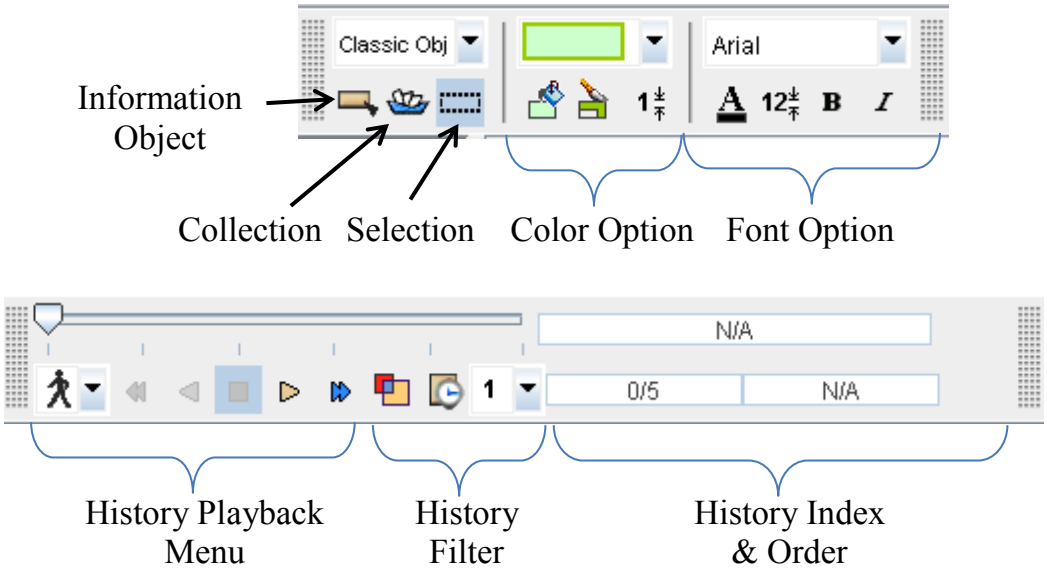
Suggestion View

Based on user events and activities within the visual workspace, PerCon infers the user's interests. Depending on the inferred interests, PerCon suggests and visualizes relevant datasets in the suggestion view.



Tool Bar



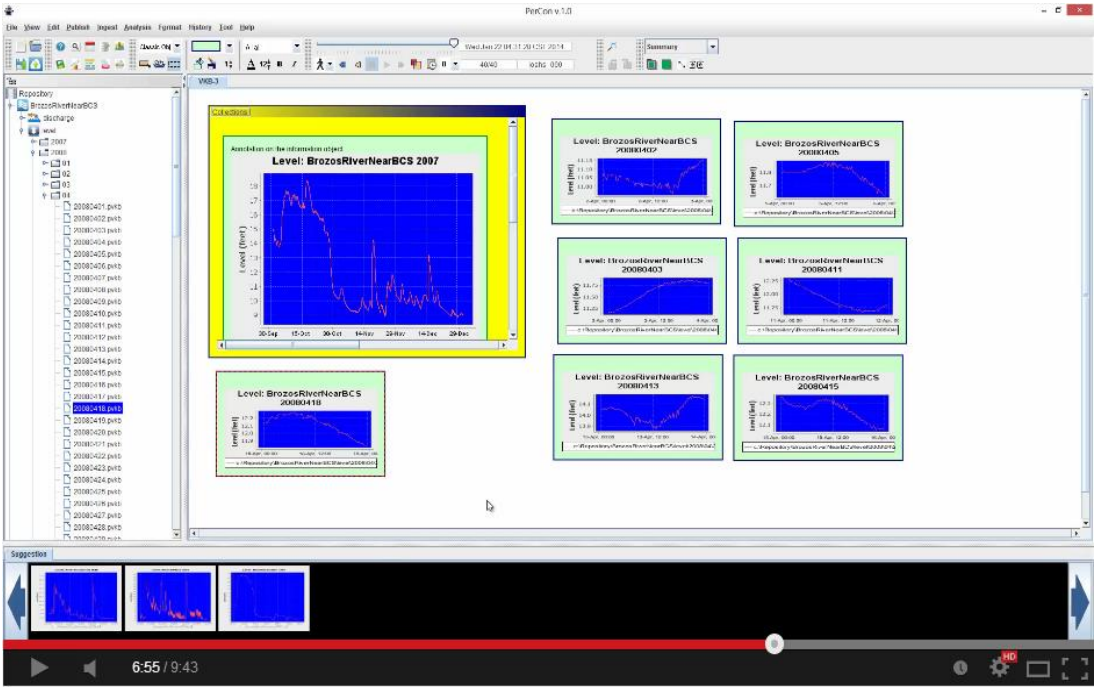

Classic Obj
Arial
Information Object
Collection
Selection
Color Option
Font Option
N/A
1
0/5
N/A
History Playback Menu
History Filter
History Index & Order

3

Chapter 3: Short Video Demo

The following video clip gives instructions on how to use system.

To see the video clip: (Click here)



URL:

https://www.dropbox.com/s/111laysf649ovzw/UserManual_VideoClip.mp4

http://www.youtube.com/watch?v=drPMD-2F9cQ

APPENDIX D

QEUSTIONNAIRE

Question Set 1.

1. Gender
   - Male
   - Female

2. Highest Degree (finished or currently pursuing)
   - Bachelors
   - Masters
   - Doctorate
   - Other (please specify):

3. Field of Study (Major)

4. Age
   ☐ 18 ~ 20
   ☐ 20 ~ 25
   ☐ 26 ~ 30
   ☐ 31 ~ 35
   ☐ 36 ~ 40
   ☐ Over 40

Question Set 2. Workspace

5. It was useful to explore data using thumbnail preview (previsualization) in the repository view before instantiating it in the workspace

1 2 3 4 5 6 7

Strongly disagree Neutral Strongly agree

6. I had enough support to understand data content in the workspace

1 2 3 4 5 6 7

Strongly disagree Neutral Strongly agree

7. I had enough support to express data relationships by organizing the dataset in the way I wanted:

1 2 3 4 5 6 7

Strongly disagree Neutral Strongly agree

8. It will be easy for someone else to understand the way I organized the dataset in the workspace

1 2 3 4 5 6 7

Strongly disagree Neutral Strongly agree

9. It was easy to visualize and represent targeted objects in the workspace

1 2 3 4 5 6 7

Strongly disagree Neutral Strongly agree

10. I had enough support to effortlessly / quickly browse and select co-occurrent data:

1	2	3	4	5	6	7
Strongly disagree			Neutral			Strongly agree

11. The visualization methods provided for different data were sufficient:

1	2	3	4	5	6	7
Strongly disagree			Neutral			Strongly agree

12. It was easy to interpret and characterize given/created objects in the workspace:

1	2	3	4	5	6	7
Strongly disagree			Neutral			Strongly agree

13. What criteria did you apply to organize the data of interest in visual workspace? In case of multiple selections, rank your selections.

- ☐ Data granularity
- ☐ Particular events / occurrences / changes in data
- ☐ Data component (source)
- ☐ Visualization types (plot, web browser, etc.)
- ☐ Content-based data similarity
- ☐ Other (please be specific)

_____

14. During the tasks, what activities/attributes in the workspace do you think were the main focus of your attention? In case of multiple selections, rank your selections.

    ☐ Textual information (annotation) in data object

    ☐ Temporal attributes in data

    ☐ Spatial structures in workspace

    ☐ Visual patterns in data

    ☐ Other (please be specific)

    _____

15. What was useful when you explored data/information objects and performed the several tasks in the workspace? Ex) Mouse Drag and Drop, hierarchical view (tree structure) of repository, data-source view synchronization with suggested data, pre-visualization, etc.

    _____

    _____

    _____

    _____

16. Did the visual workspace help or hinder your tasks regarding data management? Explain.

    _____

    _____

    _____

    _____

    _____

17. Did the visual workspace help or hinder your interpretation of data? Explain.

_____

_____

_____

_____

Question Set 3. Mixed-Initiative Interaction (MI)

18. I was satisfied with the data suggested by the MI agent.
(I had enough support to find and select other related data through mixed-initiative interaction)

1 2 3 4 5 6 7

Strongly disagree Neutral Strongly agree

19. I was satisfied with the suggestion request functionality to MI agent.

1 2 3 4 5 6 7

Strongly disagree Neutral Strongly agree

20. I was satisfied with the suggested data I requested from the MI agent.

1 2 3 4 5 6 7

Strongly disagree Neutral Strongly agree

21. I was satisfied with the suggestion frequency of the MI agent.

1 2 3 4 5 6 7

Strongly disagree Neutral Strongly agree

22. I had enough support to find and interpret data I was interested in with MI:

1 2 3 4 5 6 7

Strongly disagree Neutral Strongly agree

23. I had enough support to find correlations within the data set:

1 2 3 4 5 6 7

Strongly disagree Neutral Strongly agree

24. Overall, I think PerCon is useful to analyze the given weather and river data collections:

1 2 3 4 5 6 7

Strongly disagree Neutral Strongly agree

25. What attribute(s) was useful for data analysis? In case of multiple selections, rank your selections.

☐ Trend in data

☐ Particular incidence in data

☐ Pattern in data

☐ Measurement date (co-occourrent data)

☐ Other (please be specific)

_____

26. Did mixed-initiative interaction help or hinder your analysis of data? Explain.

_____

_____

_____

_____

27. Did the recommendation request to the mixed-initiative interaction agent help you? Explain.

_____

_____

_____

_____