# BROADENING QUANTITATIVE ANALYSIS OF DISTRIBUTED INTERACTIVE SIMULATION WITH DATA MINING FUNCTIONALITIES

**Daniel Dor, Yaniv Minkov**
**Ground Forces Command of the Israeli Defense Forces**
**Tel-Aviv, Israel**
**Daniel.dor.33@gmail.com, yanivminkov@email.com**

## ABSTRACT

A novel approach is proposed in order to get better utilization of the current method for quantitative analysis of military simulated situations during simulation-based combat systems R&D. We argue that this approach increases the usability of the data that is being collected from the simulation, and enables the simulation researchers to find unrevealed data that hardly could be found using trivial methods. For example, the unrevealed information could be recognizing hidden human factors (behavioral patterns) and recognizing irregular events occurring during the simulation. This paper describes the present analysis method, the need for better analysis tools and methods, the proposed broadened method for experiment data analysis, the challenges in using this approach, and the phases that should be added to the current methodology.

## ABOUT THE AUTHORS

**DANIEL DOR** is an Industrial engineer, database designer and military experimenter. Daniel Holds a B.Sc from the faculty of engineering at Tel-Aviv University and is studying for his MBA at Tel-Aviv University. Currently, Daniel serves as a researcher in the IDF-GFC battle-laboratory. His research interests are in application of modeling and simulation of military experiments, Database designing and analysis, and on multi-criteria decision-making processes. daniel.dor.33@gmail.com

**YANIV MINKOV** is a Human-Factors Engineer and military systems expert. Yaniv Holds B.Sc and M.sc from Ben-Gurion University of the Negev. Currently, Yaniv serves as the head of the IDF-GFC battle-laboratory and researches human-machine aspects in military complex systems, specializing on unmanned aerial systems. yanivminkov@gmail.com.

# BROADENING QUANTITATIVE ANALYSIS OF DISTRIBUTED INTERACTIVE SIMULATION WITH DATA MINING FUNCTIONALITIES

**Daniel Dor, Yaniv Minkov**
**Ground Forces Command of the Israeli Defense Forces**
**Tel-Aviv, Israel**
**Daniel.dor.33@gmail.com, yanivminkov@gmail.com**

## 1. INTRODUCTION

Subordinate to Israel's Ground Force Command, the Israeli Army's Battle-Laboratory (Battle-Lab) is a multidisciplinary center serving as a feasibility incubator for Israel's future land arsenal. By applying theoretical and technical realms to existing doctrine, budgetary, and operational constraints, the Battle-Lab can validate or deny new designs before costly projects are initiated. The Battle-Lab engineers and war fighters execute this mission by using land warfare Distributed Interactive Simulation (DIS).

DIS is a government/industry initiative to define an infrastructure for linking simulations of various types at multiple locations to create realistic, complex, virtual worlds for the simulation of highly interactive activities (see SCIS 1995). Although the proof of concept is using DIS, similar methods could be used if the simulation environment used HLA or TENA.

The DIS example dealt with in this paper addresses two main goals: (1) demonstrating real-world military operational conditions and occurrences and (2) analyzing research questions regarding acquisition of new combat systems. For example, demonstrating a real-world Unmanned Air-Vehicle (UAV) via simulation is used to illustrate the UAV's actual user-interface and actual logic.

In general, data is a discrete objective fact about events, information is the message meant to change the receivers' perception, and knowledge is the experience values, i.e. the context applied to the message. Data mining -- the analysis step of Knowledge Discovery in the Database (KDD) -- is the process of discovering new patterns from large data sets by involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract knowledge from a data set in a human-understandable structure, and, along with the raw analysis step, it involves database and data management aspects, data preprocessing, model and interface considerations, interestingness metrics, complexity consideration, post-processing of found structure, visualization and online updating (see Hand et al. 2001). Among the areas being developed, investigated, and provided with identified applications are hypertext and hypermedia data mining, ubiquitous data mining (UDM), phenomenal data mining, distributed and collective data mining, time series, constraint-based, spatial/geographic, and related methods (see Hand et al. 2001).

In order to answer research questions, there is a need to collect the output data of the simulations linked via DIS. This data contains various levels of information regarding the simulation events and the simulation models. For example, this information includes time-series data (the status of the simulated entities during the experiment), sequence data (fire processes modeling) and spatial data (the location of the entities in relation to the other entities and to the battle-field) (see Minkov et al. 2011). This data is then analyzed and transformed into information and knowledge, which are used to answer the research questions. Increasing the amount and quality of information and knowledge that is gathered from the simulation is in the focus of the current paper.

This work is based on the assumption that the collection and analysis of the DIS output data is one of the researchers' main goals. For example, the researchers may use quantitative analysis in order to examine the human user performance, which makes it possible to answer the research questions. Therefore, the simulation development process should be done in the context of the research questions. To enhance this goal, the simulation output database design should be data mining oriented, in order the fully exploit of the data gathered during the experiment.

Since the current framework of simulation development doesn't consider data mining needs, it, therefore, hardly ever reveals the high value of the simulation output data. This paper demonstrates the superiority of the new approach that contains data mining capabilities that are superior to the current conventional approach (mentioned in Section 3 of this paper).

This paper is organized as follows: Section 2 sets the background by showing the current research abilities and shows the need for broadening of the quantitative analysis. Section 3 reviews the principles of the approach and the challenges in broadening the current process by adding data mining functionalities. Section 4 describes the Military DIS Data Mining Process. Section 5 describes a proposed case study that would examine spatial, sequence, and time-series data mining analysis in the Battle-Lab. Section 6 concludes the paper with the proposed approach and discusses points for future work.

## 2. PRELIMINARIES

### 2.1 Center of Operational Research Branch

The Israeli Defense Forces (IDF) Ground Forces Command (GFC) Center of Operational Research (COR) Branch is a multidisciplinary research center supporting GFC needs for the development of fighting concepts, weapon systems and doctrine (Minkov et al. 2010). The COR research helps decision making in the context of GFC strengthening. The COR hosts two research domains: the operations research domain, which focuses on analytical and simulated quantitative analysis, and the Battle-Laboratory (Battle-Lab) domain, which performs simulated human centered experimentations. The Battle-Lab mainly focuses on human decision making during combat, on Human-Machine Interfaces (HMI) and on human factors engineering. The Battle-Lab performs studies in the Research, Development and Acquisition (RDA) and the Advanced Concepts and Requirements (ACR) domains. Being a demonstrative medium, The Battle-Lab uses Real time Virtual reality environments (Wilcox, Burger and Hoare 2000; Mitchell 1997), e.g. 3D computer games, which enable executives to look at the operational potential of a proposed technology, as well as to develop efficient operational techniques and tactics. The Battle-Lab has few main roles: (a) to determine and evaluate operational requirements and concepts for future combat systems, (b) to evaluate operational effectiveness and contribution of future combat systems (c) to evaluate HMIs for new combat systems and (d) to evaluate fighting techniques, doctrine and combat organization for current and future combat systems. The Battle-Lab's “arsenal” contains a wide spectrum of experimental methods including: usability tests, workload tests, quantitative and qualitative analysis and performing campaigns of experimentation that enable test and evaluation of operational technology-based concepts in various operational scenarios. The Battle-Lab uses a wide variety of high-end simulation tools, quantitative analysis tools, and information tools.

### 2.2 Center of Operational Research Branch

#### 2.2.1 Background

The Battle-Lab experimentation method contains two stages: the development of the simulation and the research of the data it reveals. The method divides those two stages into 5 phases that define Battle-Lab DIS research and development process (see Figure 1 on page 4).

#### 2.2.2 First stage - Simulation development

The Battle-Lab Simulation development process is composed of mixture of techniques that had been taken from several worlds: software engineering, software development, experimentation design, statistics, process engineering, human factors, and information technologies (IT).

The simulation development stage is defined by two phases:

1. Phase 1 - Specification of research questions and the appropriate analysis method.
2. Phase 2 - Software development which includes two parts: (1) Simulation software development which is based upon conventional and agile development techniques (see Martin 2003) and (2) Quantitative analysis software development which is based upon a self-developed methodology (see Minkov et al. 2011).

#### 2.2.3 Second stage - Data analysis

The Battle-Lab simulation data analysis contains two levels: qualitative analysis and quantitative analysis which is the main focus of this paper. The quantitative analysis follows data collection and is defined by three main phases:

1. Phase 3 - Experiment execution and data recording.
2. Phase 4 - Experiment data improvement using Extracting, Transforming and Loading (ETL) methods.
3. Phase 5 - Quantitative analysis and concluding.

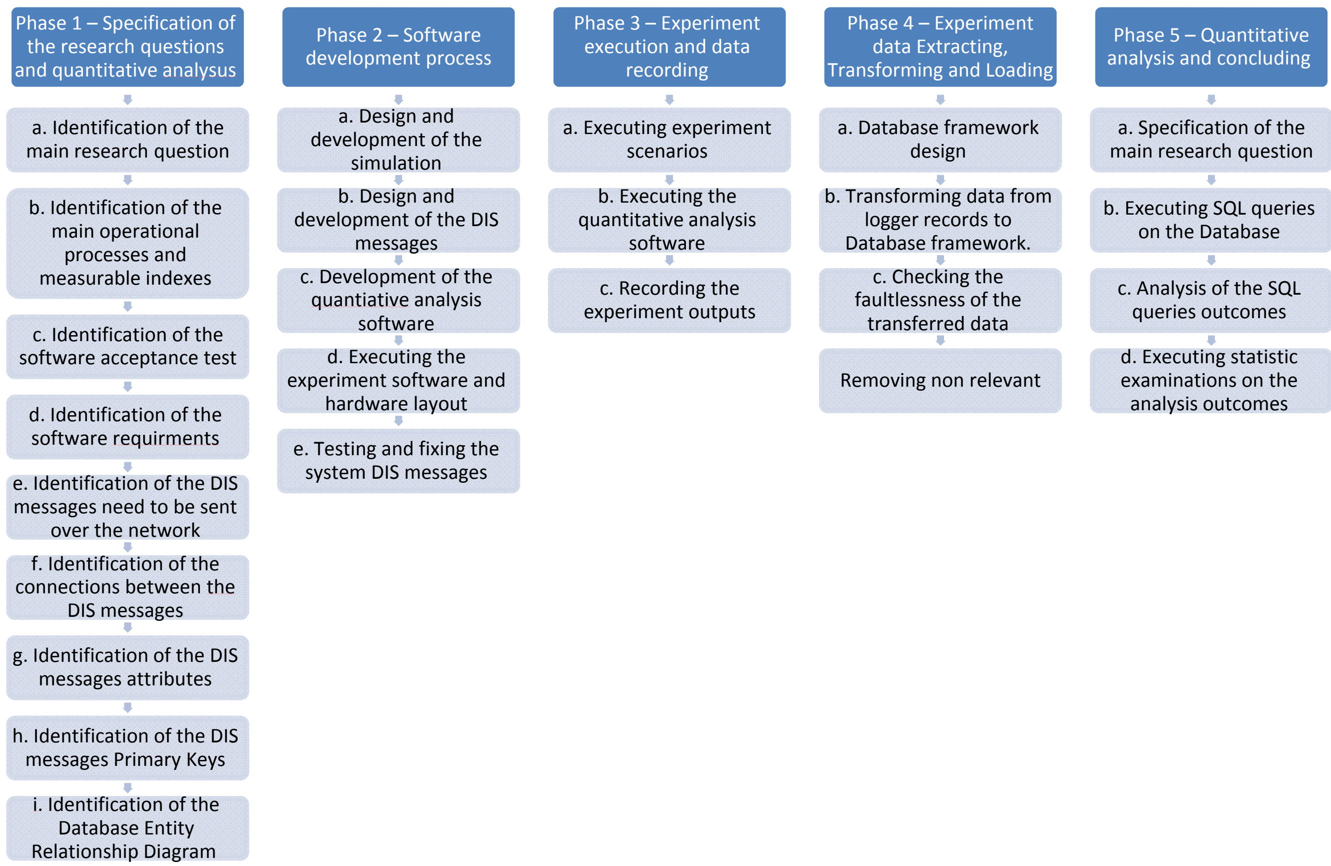


**Figure 1. Battle-Lab DIS research and development process**

### 2.2.4 Battle-Lab studies

Literature on simulation (see Ferscha and Johanson 1997; see Salitan and Eingel 2010) distinguishes between three main branches: human-in-the-loop (HITL) simulation, technical simulation and constructive simulation. The former branch of simulation involves human operating the simulated systems, the second branch includes analysis of the forecasted technical performance of a system, while the latter includes simulated autonomous entities with decision making mechanisms (see Fujimuto 2001; see Hofer and Loper 1995).

Different types of studies demand different types of simulation usage. These types of simulations stand as a rich source of data for researchers, and sometimes provide feedback to one another in an iterative loop. Anyway, the data being collected during the simulation depends upon the research questions that were defined. Examples for the information collected from the simulation are: fire, damage status of entities, movement of forces, time stamps, systems' state, and special events.

Analysis of data is the focus of the Battle-Lab studies. This analysis enables identification of the following information: (1) Processes that take place in the virtual combat environment (for example, the process that starts with some entity firing a missile and ends with the damage result of the missiles' hit), (2) high value events and behavior patterns (for example, frequent repetitions of tank commander behavior after firing on an enemy target) and (3) Time dependent questions and variables (for example, the time it takes to fire on an enemy that just became exposed). Therefore, data analysis is crucial for the success of the research and for later decision making.

### 2.2.5 Need to broaden the current analysis method

Analysis of experiments is done using two different approaches -- quantitative and qualitative analysis – explained here: (1) In quantitative analysis, samples of human actions are being taken and analyzed in order to find evidence of expected behaviors or patterns. After recording samples of interesting events, one has to evaluate and examine the data in order to answer the research questions, usually using statistical and mathematical tools. For example, examining the data would include counting the number of events in which an armored platoon was destroyed in a specific area. (2) In Qualitative analysis it is more difficult to implement numerical indicators. Thus, an experimenter that follows the course of actions in the simulation evaluates the performance of the participant in a subjective way, for example, by evaluating the participant's stress during the experiment.

Data mining technology can generate new research opportunities by: (1) automated prediction of trends and behaviors: Data mining automates the process of finding predictive information in a large database. Questions that traditionally required extensive hands-on analysis can now be directly answered from the data. A typical example of a predictive problem is "targeted marketing". Data mining uses data on past experimenter actions to identify what future action is most likely to happen. Other predictive problems include forecasting some forms of defaults, and identifying segments of a population likely to respond similarly to given events. (2) Automated discovery of previously unknown patterns: Data mining tools sweep through databases and identify previously hidden patterns. An example of pattern discovery is the analysis of experimenter actions data to identify seemingly unrelated actions that often occurred together. Other pattern discovery problems include identifying anomalous data that could represent data errors (see Hand et al. 2001).

The current Battle-Lab quantitative analysis method doesn't consider data mining needs during the determination of the research questions and the development of the simulation. The lack of data mining capabilities narrows the researcher's ability to produce information and knowledge from the experiments' data. For example, there is slim chance that the Battle-Lab researchers can identify seemingly unrelated actions that often occur together during the experiment. Therefore, using data mining methods may help to achieve deeper and wider knowledge through use of the experiment data.

The proposed approach covers this gap by enabling execution of data mining actions, in order to get better utilization of the simulation output data. This better utilization would be expressed by getting more information and knowledge from the simulation outcome than in the current situation.

## 3. PRINICPLES AND CHALLENGES IN BROADENING QUANITATIVE ANALYSIS OF DIS WITH DATA MINING FUNCTIONALITIES

### 3.1 Conventional analysis and Process Oriented Development

The conventional analysis method is highly based upon the communication between applications participating in the simulation (see Hasan and Wybarnietz 1990). The applications in the Lab-Battle communicate via the DIS protocol: their messages are broadcast throughout the network and are then accessible to any application listening to the network. This is indeed the approach taken by COR: specifically COR researchers use of a logger application that records all transmitted messages and enables exportation of these messages into a relational database.

Then, quantitative analysis of the relational database is performed and SQL queries are executed in order to analyze the database. The queries' results then serve the researchers in preparing the final experiment report.

To this conventional approach, COR researchers have added new methodology called Process Oriented Development (POD) which enables those focused data investigations (see Minkov et al. 2011). Investigation using POD is based upon identifying logical processes in the system, and designing the relational database in consideration of those processes. The POD takes the disconnected tables from the basic DIS output, joins them through the logical processes, and creates a full database design that fits exactly to the experiment's purposes.

### 3.2 The addition to the current process

In order to expand the researcher's ability to produce information and knowledge, the following method suggests adding data mining processes to the POD methodology described in paragraph 2.2. In order to do so, there is a need to broaden the current DIS research and development (R&D) process in two places: (1) Phase 2 – Software development should deliver solutions for current requirements that define the data mining oriented database. These requirements include identification of the new research questions, design and development of new DIS messages, design and development of the quantitative analysis software and validation of simulation outcomes with data mining analysis. Therefore, all DIS output would be in the context of data mining oriented database. (2) Phase 5 – Quantitative analysis and conclusions should deliver solutions for current data mining processes that enable the execution of data mining queries. In fact, some of the quantitative analysis must be done with a data mining process model that helps the researchers tackle problems regarding the data mining process. For example, the Cross Standard Process for Data Mining (CRISP).

### 3.3 Challenges of adding Data Mining functionalities

Implementation of the data mining tools leads to some challenges. The following list exemplifying these challenges:

1. Database complexity. Data mining oriented database design adds complexity in relation to process oriented database design - In the current status, the database design is mainly process oriented, and enables the researchers to get conclusions regarding processes taking place in the experiments. Adding Data Mining requirements to the database design might make it very complex to handle. In this case, the solution depends on prioritizing the most important information needed to the experiment report and designing the Database in the context of this prioritization.
2. Database size. Large Database size can conflict with Data Mining needs. The data that needs to be mined is actually the messages that were sent throughout the simulation session at a very high rate. This situation creates very large Databases. In order to get high usability of the data gathered in the experiment, there is a need to collect some subset of this large amount of data. Due to this conflict there is a need to balance the Data Mining needs with the will to decrease Databases size. Therefore, the solution to this problem is using a smaller rate of saving messages to the Database which would be enough to satisfy Data Mining needs.
3. Database quality. Handling noise, uncertainty, and incompleteness of data is always an issue. The DIS protocol is by definition unreliable (Ferscha and Johanson 1997), and loss of messages is common. Furthermore, there are other errors that are caused due to the distributed nature of the simulation; for instance, these errors may include a mismatch of inconsistent time-stamps of the internal clocks of the applications that sent them or a bug in a single application leading to partial or incorrect information within all messages sent by this application. Due to the POD methodology, it is very hard to trace the bugs, because these bugs do not affect the way the combat simulation works or displays, and are thus difficult to identify in advance. Those examples, among others, lead to uncertainty, and incompleteness of data. Adding this characteristic to the fact that the database is full of noise, we get very hard to handle data. A solution to this problem is to execute more efficient Extracting, Transforming and Loading (ETL) process and to divide the database into 3 distinct groups of data.

### 3.4 Examples for new questions the Battle-Lab researchers might ask

Despite the aforementioned challenges, implementing data mining functionalities in the DIS environment may contribute valuable improvement to the quantitative analysis process. As mentioned earlier, this approach intends to increase the usability of the data collected from the simulation, and to enable the simulation researchers to find unrevealed and non-trivial information, which they couldn't find without using this methodology. In the following paragraph there is a list of examples of questions that the methodology can help the researchers to answer:

1. Is there an association between two frequent events that occurred during the experiment?
2. Is there any connection between two or more seemingly unrelated events that occurred during the experiment?
3. What kind of classification of simulation entities exists?
4. What are the reasons for some special events that took place?

## 4. DESCRIPTION OF THE BROADENED APPROACH

The broadened approach is divided it into 3 phases: (1) Phase 1 is DIS POD and quantitative analysis (2) Phase 2 is segmentation of the experiment data and (3) Phase 3 is delivery of the experiment's outputs for data mining. The input for the entire broadened approach is the set of experiment research questions, which are later transferred to the simulation software that enables COR to execute experiments. The experiment's outputs are data that is analyzed and processed using conventional quantitative analysis. Afterwards, this data is segmented and data mined. The outputs of entire broadened approach process is information and knowledge which serves the COR researchers to write their reports and thus, to later help in decision making in the context military systems R&D.

### 4.1 Phase 1 - DIS POD and quantitative analysis

**Phase** 1 (Minkov et al. 2010) is a standalone process detailed in the preliminaries of this paper and Figure 1.

### 4.2 Phase 2 - Segmentation of the experiment's data

Phase 2 of the broadened approach divides the experiment output database into three distinct groups of the information in the Database: sequence, time-series and spatial data. The objective of spatial data mining is to find patterns in geographical behaviors of objects (or simulated entities in our case). The objective of sequence data mining is to find patterns and frequencies in sequences of events. The objective of time-series data mining is to find patterns related to the events occurring over simulation time. That kind of data mining offers potential benefits for Battle-Lab researchers. This division enables the COR researchers to take each of the three divided databases and examine it through execution of data mining methods (see Hand, Mannila and Smyth 2001). The three distinct groups of databases are:

1. Sequence Database, According to the POD concept, every simulation event (and thus, communication message) does not stand by itself, but rather is associated with a process (or sequence) of events or messages. Therefore, in this relational database each table represents group of events and the connection between the tables represent a process. These "logical" processes can be mined through use of one joined table that contains all the process' information.
2. Time-series Database. Most of the DIS database is temporal. It contains time dependent events, and the correlations between those events also depend on time. Consequently, the analysis typically focuses on sequences of events. For instance, the movement of an entity from one location to another is modeled as a sequence of location update messages. These messages are being collected in different tables which construct the time-series database.
3. Spatial Database. One of the main characteristics of Battle-Lab simulation is being Geographic Information System (GIS) based. Therefore, the produced data also has geo-location characteristics (Shekhar and Chaula 2003) meaning that the database is geographic. It contains events that are dependent on location, and the correlations between those events are also dependent on location. For example, information regarding the distance and angles between two simulation entities.

### 4.3 Phase 3 - Experiment outputs Data Mining

Phase 3 of the broadened approach enables COR researchers to execute data mining operations on the databases mentioned above. There is a variety of data mining operations that can be undertaken, and using them may help COR researchers to discover the following information and knowledge (see Nisbet et al. 2009):

1. Frequent patterns. In the current quantitative analysis, frequent patterns can be identified by manual examination of the information. Executing data mining queries on the experiment output data may enable COR researchers to automatically identify frequent patterns of events. For example, in the context of tank platoon behavior, identification of a repeated action that the platoon commander tank is doing after shooting may be useful in the future.
2. Association, correlation vs. causality. In the current quantitative analysis, association and correlation between groups of events can be identified only when there is special statistical examination of these events. Therefore, executing data mining queries on these groups of events may often show correlation. For example, in the context of performance for a tank platoon, identification of those tanks that have similar performance may be useful.
3. Classification and label prediction. In the current quantitative analysis, it is very hard to classify groups of data in a way to enable prediction of the behavior of this group. This feature should enable us to describe and distinguish classes or concepts for future prediction (Lin 2006). For example, we would try to predict how much ammo the platoon intends to use as a function of the time duration of an experiment.
4. Cluster Analysis. Maximizing intra-class similarity and minimizing interclass similarity would enable COR researchers to identify characteristics of similar events that occurred in the simulation and to group these similar events into clusters. This feature would help the researchers get insights into the data distribution. This type of useful data could include, for example, identifying groups of destroyed enemy forces according to their type and geographical location.
5. Outlier analysis. Identifying data objects that do not comply with the general behavior of the data would enable COR researchers to focus on the most cardinal data that has been

gathered. Furthermore, this feature may enable the recognition of special phenomenon that occurred during the simulation. For example, identification of distinctive portion of the event that had a special characteristic that caused the outlier may be useful.

## 5. SUGGESTED CASE STUDY

In order to examine the broadened approach, a suggested case study would examine the assumption that this approach increases the usability of the data that is being collected from the simulation, and enables the simulation researchers to find otherwise unrevealed data. This proposed case study would examine spatial, sequence, and time-series data mining analysis in the Battle-Lab by executing data mining actions on our POD database. The case study data mining would follow the same functions mentioned above in phase 2.

The suggested case study would look at two alternative applications of the DIS data principles. In order to compare between the cases, we would show an example of an environmental company that is trying to assess the impact of varying land-use patterns on climate change. Afterwards, the Battle-Lab capability would be compared in the context of this example. The main similarity between the DIS world and the environmental company case world, is the fact that both of them deal with time-based processes that need to be analyzed. The main difference is that in previous discussions of the Battle-Lab case, we looked at discrete human interactions, and the environmental company case will look at continuous simulation events.

On the environmental company case, a fully distributed, physically-based hydrologic modeling system, MIKE SHE, was used to investigate whole-watershed hydrologic response to land use changes within the Gyeongancheon watershed in Korea (Sangjun et al. 2008). A grid of 200 × 200 meters was established to represent spatial variations in geology, soil, and land use. Initial model performance was evaluated by comparing observed and simulated stream flow from 1988 to 1991. Results indicated that the calibrated MIKE SHE model was able to predict stream flow well during the calibration and validation periods. Proportional changes in five classes of land use within the watershed were derived from multi-temporal Landsat TM imageries taken in 1980, 1990, and 2000. The imagery revealed that the watershed experienced conversion of approximately 10% non-urban area to urban area between 1980 and 2000. The calibrated MIKE SHE model was then programmed to repeatedly analyze (using data mining tools) an artificial dataset under the various land use proportions identified in the Landsat TM imagery. The analysis was made to quantitatively assess the impact of land use changes (predominantly urbanization) on watershed hydrology. There were increases in total runoff (5.5%) and overland flow (24.8%) as a response to the land use changes. In fact, the analysis shows coherent connection between the human actions (land use proportions) and the geo-spatial conditions (watershed hydrology).

The Battle-Lab case analysis also shows coherent connection between the human actions and the geo-spatial conditions. In the Battle-Lab suggested case study, tanks platoon and enemy entities performances would be evaluated by analyzing observed and simulated movements and actions throughout the simulation scenarios. The data mining results will identify the influence of the tanks platoon and their enemy's geographical location and actions over-time, on their performance and on the events related to them. For example, we would be able to identify how the tanks platoon height affect the type of ammunition being shot at them. This analysis would be done by data mining tools that would try to find connection between the platoon height and the type of missiles launched towards them.

To illustrate the advantages of the broadened Data Mining methodology, a data mining analysis of a large-scale military experiment was proposed. The methodology was tested, by executing some Data Mining operations on the database gained from the experiment. The database was gathered using Process Oriented Development and quantitative analysis methodologies; during the case study, the observation indicated that the new approach allowed the Battle-Lab researchers to gain much more information from the current database, exhibiting surprising and unthinkable results. The following paragraph shows a portion of the case study results by exemplifying 4 different data mining research questions.

1. 1$^{st}$ research question: Spatial data query. Which enemy was frequently shooting from a certain altitude to targets on a certain altitude? This Frequent Patterns Data Mining query was executed on a table that described the process of enemy's attack on a blue force in the experiment (enemy's fire->enemy's detonation result->blue force's damage result). The table also included the geospatial characteristics of each event. This query managed to identify, using a decision tree, that substantial amount of the shooting (~40%) came from the altitude range of 965.1m > UTM_Z > 772.6m. The query was also pointed that when the target was above the altitude of 798.02m, most of the attackers were from a specific type, and when the target was below that altitude, it was

another attacker type that carried out most of the attacks.

2. 2nd research question: Sequence data query. Which blue tanks were attacked frequently by the same enemy? This Associate Data Mining query was executed on a table that described the process of enemy's attack on a blue force in the experiment (enemy's fire->enemy's hit->blue force's damage result). The table also included the various entities characteristics. This association algorithm examined a dataset in order to find blue tanks that appear to be attacked by the same enemy. This set of tanks constituted a case. The algorithm generated a rule that tried to predict the connection between the attacked units and the attacker. The COR researchers noticed that if the 3rd tank and the 1st tank in the 1st platoon were attacked by an enemy entity, then the likelihood that the 2nd tank from the 1st platoon will be attacked by the same enemy is 71.4%. This rule was labeled "important"[1].
3. 3rd research question: Sequence data query. Which ammunition was used by the blue battalion as a function of the number of shells used in the scenario? This classification Data Mining query was executed on a table that described the process of the blue tank attack on a red force in the experiment (blue tank's fire->blue tank's detonation result ->red force's damage result). This algorithm extracts patterns that predict the individual values of one column (count of ammunition) based on the values in other columns (type of ammunition and type of scenario). In this case, the COR researchers noticed that in the scenarios where the blue forces were shooting above 23 shells, the main ammunition type is A, and in the scenarios where the blue forces were shooting 23 shells and below, the main ammunition type is B.
4. 4th research question: Time-series data query. Are there any patterns regarding the time it takes the blue tank platoon to close "Fire Cycle"[2] on an enemy entity? This clustering Data Mining query was executed on a table that described the process of closing "Fire Cycle". This process is based on the following actions: enemy attacks blue forces in the experiment, the fire event causes a detection of the enemy and then the blue force attack and destroy the enemy (enemy's fire->enemy's hit->blue force's damage result->enemy's detection by blue force->blue force's fire->blue force's hit->enemy's damage result). The clustering algorithm detects groups of rows in the database that share similar characteristics. This query managed to identify that its takes a huge amount of time to close "Fire Cycle" on enemy entity type A, but its takes much less time to close "Fire Cycle" on enemy entity type B.

Our new approach should ensure that the COR researchers can answer the above questions easily by using data mining as described. Furthermore, the new approach would help the researchers to recognize patterns they would not otherwise have anticipated or noticed.

This suggested case study shows the high potential in broadening the current process by adding data mining functionalities. Without these functionalities, the COR researchers would be unlikely to notice these types of patterns, so the data mining function will help them explain patterns in events that took place in the simulation.

## 6. CONCLUSIONS

We presented an approach that enables an expanded quantitative analysis of complex operational processes and events that take place in the simulated world. The approach is based upon novel software development methodology called Process Oriented Development. We explained the main principles of our broadened approach and highlighted the added functions to the former methodology. We've also showed a suggested case study that explains four of the many potential uses of this approach. Our experience shows that the new method may offer for the following advantages:

1. This approach would increase usability of the data collected from the simulation.
2. This approach may enable the simulation researchers to find unrevealed data that they couldn't find using current methods.
3. This approach is easy to use by leveraging currently available tools.
4. Similar methods could be used if the simulation used HLA or TENA (although the proof of concept is using DIS)

[1] Importance of a rule – Importance is designed to measure the usefulness of a rule. The importance shows you which rules are more likely to be significant.

[2] Closing "Fire Cycle" – **Fire Cycle** is a collaborative process that involves multiple entities of the simulative environment. Closing "Fire Cycle" means, that the blue tanks finished a process, which begins with fire initiated by some enemy entity ***A*** at some entity ***B***. The fire event then causes a detection of the enemy entity ***A*** by some entity ***C.*** Entity ***C*** then reports to some entity ***D*** about this detection; Entities **B, C** and **D** then combine to attack entity ***A***.

In spite of that, some issues of the DIS R&D process still require improvements, and there is a need to further develop the statistical understanding before using this method. Therefore, further work is needed in the following disciplines: Data Mining oriented Database design, ETL processes, Database dividing, Data Mining statistics, and Data Mining analysis execution. Above all, there is a need to validate this new approach.

This approach expands a world of opportunities and improved data utilization for the COR researchers. Until now, we were drowning in data but starving for knowledge. Today, we start eating.

## ACKNOWLEDGEMENTS

We would like to acknowledge and extend our heartfelt gratitude to the following persons who have made the completion of this paper possible: To Einav Kiperman for reviewing the paper and for Alon Peleg for his vital support.

## REFERENCES

Dasu, T. & Johnson, T., (2003). *Exploratory Data Mining and data cleaning*, New Jersey: John Wiley & Sons

Ferscha, A. & Johanson, J., (1999). Distributed Interaction on Virtual Spaces, *Distributed Interactive Simulation and Real-Time Applications Proceedings on 3rd IEEE International Workshop*, 5-13.

Fujimuto, R. M., (1999), Parallel and distributed Simulation, *Simulation Conference Proceedings 1999 winter*, 122 - 131 vol.1.

Haban, D. & Wybarnietz, D., (1990). A Hybrid Montior for Behavioral Performance Analysis of Distributed Systems, *in IEEE Transactions on Software Engineering*, 197-211 vol. 16.

Hand, D.J., & Mannila, H., & Smyth, P., (2001). *Principles of Data Mining*, Cambridge: MIT Press.

Hofer, R.C., & Loper, M.L., (1995). DIS Today, *in Proceedings of the IEEE*, 1124-1137 vol. 83.

Kamalpur, S., & Deshpande, N., (2009). *Distributed Systems*, San Francisco: Technical Publications.

Minkov, Y., & Deutch, D., & Novoselski, M., & Dor, D., & Kiperman, E., (2011). Process Oriented Development for effective data analysis of military distributed simulation, *Spring Simulation Interoperability Workshop 2011 Proceedings*, 48-56.

Martin, R., (2003). *Agile Software Development, Principles, Patterns, and Practices.* New Jersey: Prentice-Hall.

Mitchell, T.M., (1997). *Machine Learning*, New York: Mcgraw Hill

Nisbet, R., & Elder, J., & Miner, G., (2009). *HB of statistical analysis and Data Mining Applications*, Atlanta, GA: Elsevier.

Sangjun, I., & Hyeonjun, K., & Chulgyum, K., & Cheolhee, J., (2008), Assessing the impacts of land use changes on watershed hydrology using MIKE SHE, *Environmental Geology, Volume 57, Number 1, 231-239*

Shekhar, S., & Chaula, S., (2003). *Spatial Databases: A Tour*, New Jersey: Prentice-Hall.

Standard Committee on Interactive Simulation, (1995). *IEEE Standard for DIS - Application Protocols*. Retrieved May 22, 2009, from http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=729408&url=http%3A%2F%2Fieeexplore.ieee.org%2Fstamp%2Fstamp.jsp%3Ftp%3D%26arnumber%3D729408

Tan, P.N., & Steinbach, M., & Kumar, V., (2005). *Introduction to Data Mining*, New Jersey: John Wiley & Sons

Wilcox, P.A, & Burger, A.G, & Hoare, P., (2000). *Advance Distributed Simulation,* New Jersey: Prentice-Hall