PNNL-15428

Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy

Towards a Unified Approach to Information Integration – A review paper on data/information fusion

P. D. Whitney
C. Posse
X. C. Lei

October 2005

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm

This document was printed on recycled paper.
(9/2003)

PNNL-15428

Towards a Unified Approach to Information Integration – A review paper on data/information fusion

P. D. Whitney
C. Posse
X. C. Lei

October 2005

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Abstract

Data (or information) fusion (or integration) from different sources is found in many applications, from epidemiology, medicine, biology, and business to military applications such as intelligence. Data fusion may involve integration of spectral content with imaging, text, and many other observations or sensor data. In this paper, we review the methodologies and implementations of the data-fusion process used in the literature and illustrate in more detail the methodologies involved by presenting three examples. We propose that the data-fusion process can be viewed as multi-path and multi-stage mappings and that the development of a data-fusion system for each specific application involves integrating four tool boxes: structure tools for data-fusion structure specifications, analytic tools for mapping methodologies, visualization tools for human and machine interface, and evaluation tools for evaluating fusion outcomes.

Contents

Figures

I. Introduction

Data, or information, integration exists in many fields of study, from epidemiology to medicine, biology, business, military applications such as intelligence, and non-destructive testing. For example, in epidemiology, information is often obtained based on many studies conducted by different researchers at different regions with different protocols. In medicine, the diagnosis of a disease is often based on imaging (MRI, x-ray, and CT), clinical examination, patients' description of symptoms, and lab results. In biology, information is obtained based on studies conducted on many different species (DuMouchel and Groer 1983), and many different tools such as electrophoresis and spectrometry. In business, financial as well as political information is gathered and analyzed. In intelligence and military fields, data can be from radar sensors, text messages, chemical/biological sensors, acoustic sensors, optical warnings, and many other sources (Cato and Simmen 1987; Pemberton, Dotterweich et al. 1987; Simpson and Kelley 1987). In non-destructive testing, visual examination, eddy current testing, and other kinds of tests (e.g., ultrasonic test results) are integrated to detect the flaw and its depth and length (Gros 1997).

"Data fusion," "information fusion," "data integration," and "information integration" are all used synonymously. In this paper, we use data fusion and information fusion interchangeably.

There are several versions of generalization of data fusion. Hall and Llinas (2001) pointed out three processing architectures: direct fusion of sensor data, fusion of extracted features data from sensor data, and the fusion of decision data formed from individual sensors. Gros (1997) adopted this integration paradigm to the non-destructive testing situation, as data fusion at the signal level (data), evidence level (features), and dynamics level (decision). This paradigm can be generalized to all types of data, including sensor data fusion, as three types of architectures: raw data fusion, feature data fusion, and decision data fusion. Figure 1.1 depicts this classification.

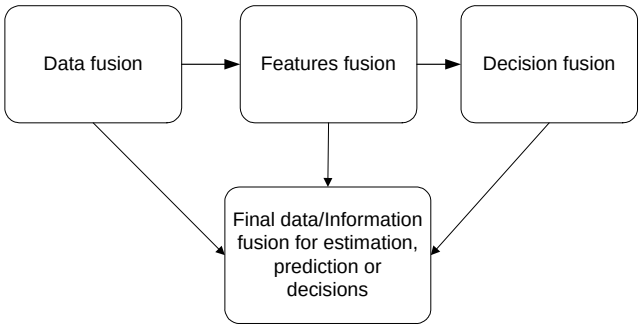

Figure 1.1. Data fusion classified into data fusion, features fusion, and decision fusion (adopted from Hall and Llinas, 2001, and Gros, 1997)

There are situations in which featured data may need to be combined with raw data or decision level data to reveal the correlations among different attributes of variables. In this case, the process is a hybrid of data fusion and feature fusion, or feature fusion and decision fusion.

So, this classification does not capture the interactions among different parts of the data-fusion process.

The Joint Directors of Laboratories (JDL) Fusion Working Group has defined four different levels for the data-fusion process, targeted mainly at military applications (Steinberg and Bowman 2001). This is a four-level functional model for the data-fusion process: level 1 involves object assessment from the raw sensor data; level 2 take the results from level1 to perform situation assessment; level 3 takes results from level 2 to have impact assessment; and level 4 provides resource management and feedback for refinement of the three previous three levels. This definition assumes a sequential flow of information. However, there might be situations for which results from both levels 1 and 2 need to be considered to assess impacts. Most recently, a level 5 process was proposed for this model (Hall and McMullen 2004). This level captures the human interaction with computers in the process of data fusion. JDL's classification is a good overview of the data fusion process but does not indicate how different levels of the process interact. Figure 1.2 illustrates JDL's four-level fusion process.

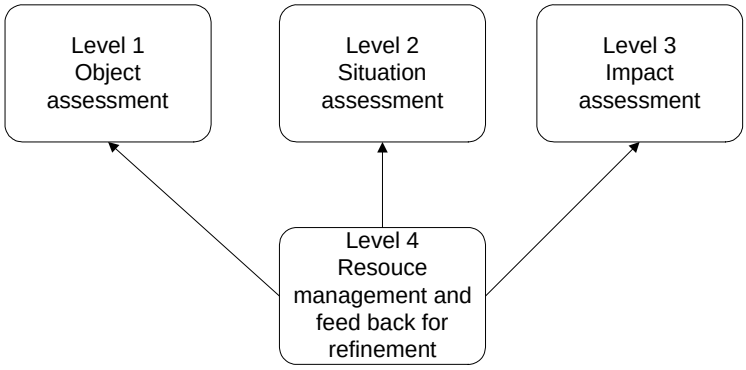

Figure 1.2. Joint Directors of Laboratories (JDL) process model for data fusion targeted at military applications (adopted from Steinberg and Bowman, 2001)

Kokar (Kokar, Tomasik et al. 2004) proposed using category theory from mathematics to classify information fusion systems. This classification, which consists of algebraic specifications and morphisms among the specifications, classifies information fusion into three classes: data fusion, decision fusion, and data association. They claim that decision fusion and data association can be framed as a special case of data fusion. Their classification can potentially provide a framework for computers to automatically specify algorithms and to synthesize and analyze a fusion system.

Effective development and assessment of methodologies are vital to integrating the data, whether it is raw data, feature-extracted data, or individual decision data. Depending on the kind of data to be combined, both methodology and evaluation of the effectiveness of the method can vary. For raw data or extracted features, methods include parametric templating, cluster analysis, adaptive neural networks, physical models, knowledge-based methods, and others. For the decision-level data, methods include classical statistical inference, Bayesian inference,

Dempster-Shafer's method, and other heuristic methods (Hall and Llinas 2001; Hall and McMullen 2004).

The framework or classifications of data/information fusion are still evolving. However, in all the classifications mentioned above, the fusion process depends on the types of data (e.g., raw data, extracted-feature data, or decision-type data) and the methodologies used.

It is clear that, no matter how the process is classified or framed, information fusion contains many steps of data processes that entail mapping data from one domain to another with varying methodologies.

In this paper, we divide data fusion into three parallel steps: the data-fusion process, fusion methodologies, and fusion development/implementation. The data-fusion process describes the overall structure of the fusion; fusion methodologies are the methods of linking all structures together; and fusion-system development is the framework for the fusion to be applied. We illustrate these three steps in three examples and propose a multi-path, multi-stage mapping structure as a unified approach to the data-fusion process and system implementation.

II. Information Fusion Models

Every time a decision is made based on analyzing data, a data-fusion process is performed. Three examples are considered below: (1) the discovery of the preventive effect of fluoridated water on tooth decay; (2) the cross-referencing of information among studies on the effect of difference diesels conducted on both animals and humans; and (3) the command and control management of battlefields. In each situation, the data-fusion process, data-fusion methodology, and possible system development are discussed.

II.1. Fluoridated Water and Tooth Decay

In 1908, there was a brown stain (enamel fluorosis) endemic in Colorado among children, ages 12 to 14 (Hilleboe 1956). Subsequently, from 1908 to 1942, several historical studies and observations were conducted to study the cause of the brown stain. Clinical dental and health examinations involving the color of teeth, x-ray estimations of bone maturation, blood counts, eye and ear tests, test for excretion of albumin and for red blood cells, and casts were recorded. It was found that the stain was caused by fluoride in the water. Meanwhile, it was observed that those with brown stain had fewer dental caries. The more brown stains, the fewer dental caries. This led to the hypothesis that fluoridated water prevents the dental caries. In 1945, the Newburgh-Kingston caries-fluorine study, a prospective study, confirmed that the more fluorine in the water given to a study population, the fewer dental caries (Schlesinger, Overton et al. 1950; Dean 1953; Schlesinger, Overton et al. 1953; Ast, Smith et al. 1956; Dean 1956; Schlesinger, Overton et al. 1956). This study led to new public health policies recommending that fluorine be added in drinking water or that toothpastes and fluorine tables or solution be used in the prevention of cavities.

Data-fusion process

In this case, the data fusion process encompasses two components: the hypothesis generated and the hypothesis supported. Historical observations on brown stains and fluorine contents in drinking water and tooth decay were sequentially combined to infer correlations between brown stain and fluorine and brown stain and cavities. From those correlations, a hypothesis was postulated and supported by prospective observations of cavities and fluorine contents. Figure 2.1 depicts the process.

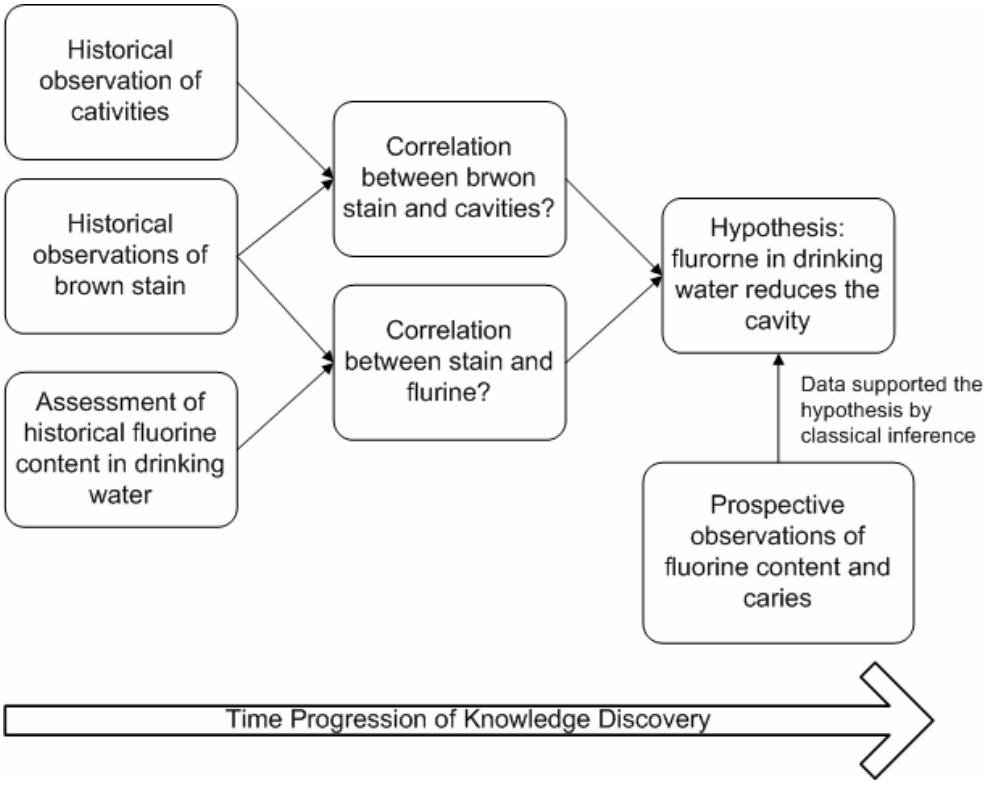

Figure 2.1. Data-fusion process for the fluorine and tooth decay situation

Data-fusion methodology

Exploratory analyses with bar charts, scatter plots, and list of frequency tables were used to model the data. Variables of interest, such as fluorine content in ppm, the number of decayed, missing, or filled (DMF) teeth per 100 erupted permanent teeth, and many other were identified. Regression analysis was used to identify the correlations between brown stain and fluorine, brown stain and dental caries, and dental caries and fluorine. For example, pairs of the index of dental fluorosis and fluoride content in water in ppm were extracted from different cities and states and scatter-plotted (see Figure 2.2) to generate relationships between fluorosis and fluorine content (Dean 1956). Regression analysis of these pairs of data resulted:

$$\text{Fluorosis index} = 0.237 + 2.275 \times \text{Fluorine content}$$

with a 95% confidence interval for the coefficient of fluorine content between 1.589 and 2.961, showing the statistical significance of the fluorine content to the fluorosis content. This same regression analysis and a classical testing of the hypothesis were used to deduce a strong effect of fluorine content on dental caries.

Fluorosis Index	Fluorine Content, ppm
3.4	14.1
3.3	8.0
3.2	7.6
3.3	5.7
2.7	4.4
2.1	4.0
2.3	3.9
1.8	2.9
1.3	2.6
1.4	2.2
0.69	1.9
0.67	1.8
0.46	1.3
0.32	1.2
0.31	0.9
0.17	0.6
0.22	0.5
0.25	0.4
0.09	0.3
0.08	0.2
0.01	0.1
0.01	0.0

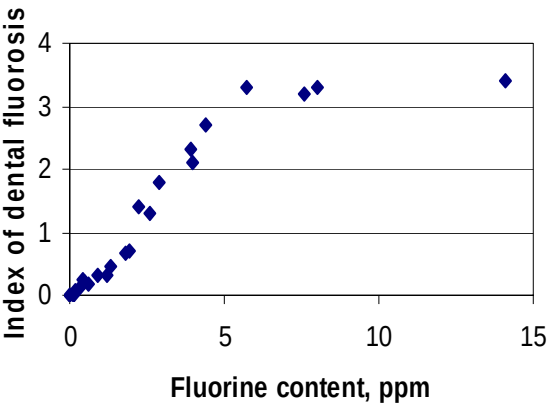

Figure 2.2. Scatter plot of fluorosis versus fluorine content

Possible system development

A software system can be developed to facilitate knowledge discoveries of a similar kind. For example, an outbreak can occur in a small group. To find out the cause of the outbreak, demographic, environmental, and other data potentially related to the outbreak need to be collected. The correlations among different variables should be sought. Analysis tools such as scatter plots, bar charts, contingency tables, regression analysis, classical test of hypothesis, and many others can be put in place to facilitate the process.

II.2. Effect of Chemicals (Diesels) on Humans and Animals

Cancer risk assessment of the effects of chemicals on humans is an important component in government agencies' decisions about chemical regulations. Chemical cancer risk is often studied *in vitro* or in animals. Sometimes, observations of the effect of similar chemicals on particular organs of humans can be observed in occupational studies. In the second example, 37 risk assessments of 10 chemicals were conducted on five cell lines, on animal systems, or on humans. Among those 37 studies, only 4 human lung cancers studies were observed. Combining biological information and similarities among the structures of the chemicals, Bayesian approaches were employed to help extrapolate cancer risk of the chemicals to those species on which no studies were conducted (DuMouchel and Groer 1983). Similar approaches were also applied to assess the effect of plutonium on human bone cancer(DuMouchel and Groer 1989).

Data-fusion process

The data-fusion process in this case encompasses two stages. The first stage is the estimations of the dose response slopes and their coefficient of variations for all 37

experiments/studies separately with their respective dose-response models. The second stage is the estimation/extrapolation of the dose-response slopes and their coefficient of variations for all chemical/species combinations using hierarchical Bayesian models. Figure 2.3 depicts the process.

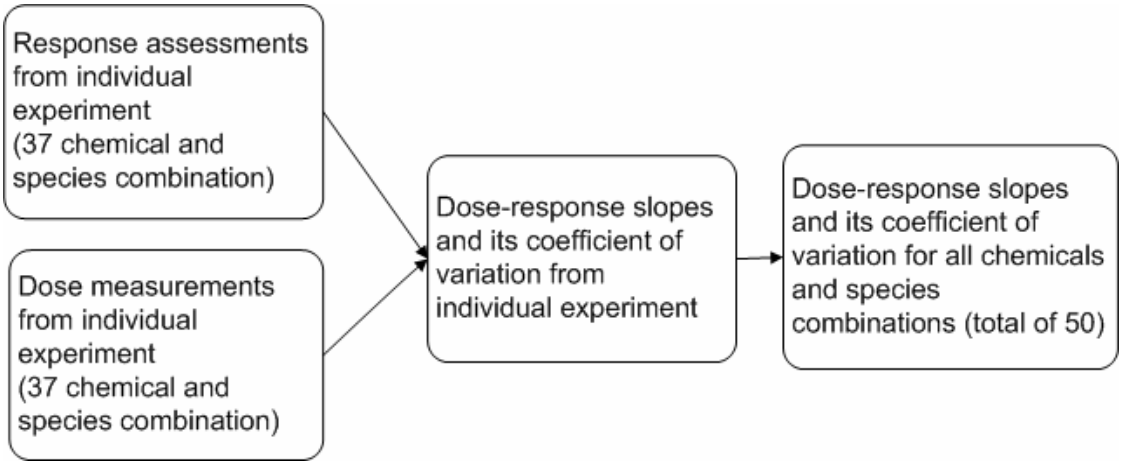

Figure 2.3. Data-fusion process for the effect of chemicals on human and animals

Data-fusion methodology

A hierarchical Bayesian approach, combined with regression analysis, was used to combine the results from the studies conducted on 37 chemical and cell line/animal/human combinations and to extrapolate to all 50 chemical and cell line/animal/human combinations. This modeling was applied to the dose-response summaries obtained from the individual studies. Specifically, suppose the log of the response slope can be decomposed into three components (mean effect, species-specific effect, and agent-specific effect). Then, let Y denote the collection of all observed/estimated log slopes; β the collection of those mean effect, species-specific effect and agent-specific effect; X a design matrix appropriate chosen; δ the random effect from the combination of different species and agents; and ε the overall dose slope estimation error.

$Y = X\beta + \delta + \varepsilon$.

Bayesian hierarchical modeling assumes that, we can specify the following distributions in a hierarchical order:

$\sigma \sim$ (distributed as) $p(\sigma)$,

$(\beta.|\sigma) \sim N(b, V)$, N represents normal distribution,

$(\theta.|\beta, \sigma) \sim N(X\beta, \sigma^2 I)$, *where* $\theta = X\beta + \delta$,

$(Y.|\ \theta, \beta, \sigma) \sim N(\theta, C)$.

The estimate of β was obtained by maximizing the posterior distribution of β given the data Y, $P(\beta|Y)$, which is a mixture of multivariate normal distributions.

Possible system development

In this case, software toolboxes for individual risk assessment can be developed to facilitate data processing for individual experiments. Algorithms for implementing methodologies such as clustering, Bayesian, and others can be put together to facilitate final data fusion for cross-

referencing other experiments. Analysis tools should be flexible enough for prior distribution specifications in Bayesian analysis in different hierarchies.

II.3. Battlefield Management

Development in data fusion, especially in multisensor data fusion, provides a platform for automated battlefield management. Several command and control systems for battlefield management have been developed (Cato and Simmen 1987; Hafner and Thompson 1988). Such a system should be able to manage data from multiple sources, to correlate and evaluate the data, and to provide consistent and coherent tactical support to the commander. Data sources for such systems can include a chemical/biological sensor, radar sensor, acoustic sensor, nuclear detector, free text message, and many others. Developments of methodologies in target detection and tactical situation assessment (Pemberton, Dotterweich et al. 1987; Waltz 1987) are continuing. A fully integrated command and control system for real-time use of battlefield management is not too far from reality.

Data-fusion process

Data from each individual sensor were processed separately to provide state determination (state signature), which includes the identification of objects/targets and their position, velocity, and other physical parameters. Each individual target is further differentiated by a probability assessment and its physical state from all sensors having potential information about it. Target information is further fused/analyzed to provide a tactical assessment (e.g., friend-foe-neutral, attach, or watch) to the command team. Figure 2.4 illustrates the data-fusion process for battlefield management.

Data-fusion methodology

Physical information and templating are used by each sensor to detect targets (Hall and McMullen 2004) and estimate/extract qualities of the physical state (position, velocity, etc.). Weighted, least-squares splines models can be used in tracking targets (Campbell and Samaan 1988). A knowledge-based system with lookup tables is used to correlate results from all sensors for event detection and situation assessment. Both Bayesian probabilistic and Dempster-Shafer's evidential reasoning are used to further assess tactical situations (Carl 2001; Stone 2001).

In the Bayesian approach to combining evidence of multiple sources for target identification, the probability of target identification (e.g., friendly, foe, or neutral) is computed from posterior distributions. The prior specifications of the probabilities of events are required to be mutually exclusive and consistent. However, inconsistent evidence is more realistic. Because of the uncertainties in sensors as well as situations, different sensors may provide different evidence for the same or similar events and conflicting evidence for the same situation or events may occur. In the Dempster-Shafer approach, two beliefs are computed: a) the belief of the event given by the data and b) 1 – belief of the complement of the event. Because the prior specification need not be mutually exclusive, inconsistent evidence for specifying priors is allowed. However, more efficient analytic tools for implementing the Dempster-Shafer method are still needed (Carl 2001; Hall and McMullen 2004).

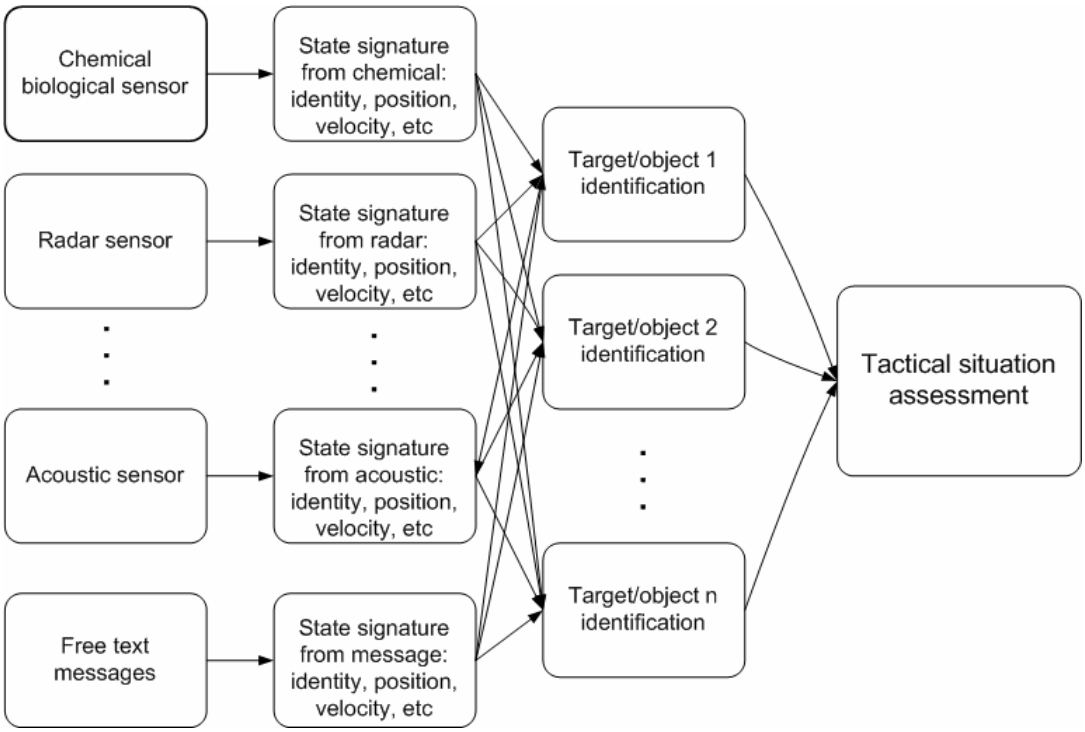

Figure 2.4. Data-fusion process for the battlefield management situation

Possible system development

Caito and Simmen (1987) developed a prototype of a vehicle-integrated defense system. Such a system needs to combine both hardware and software. The hardware includes sensors, such as optical warning sensors, laser detection systems, passive missile detectors, nuclear detectors, and millimeter wave radar detectors. The software includes data-management components as well as collections of algorithms for target detection and for situation and tactical assessments. System development should be done in a way that the addition of methodologies will not trigger a substantial modification of the system.

III. A Unified Approach: Multi-path and Multi-stage Mappings

As described in section II, we divide data fusion into three different steps: the data-fusion process, fusion methodologies, and fusion development/implementation for the three examples illustrated. We propose independent development of the fusion process, the methodology of fusion (mapping), the evaluation criteria, and the visualization tools.

III.1. Data-fusion process

As indicated in section I, the information-fusion process can be categorized in different ways: from the perspectives of architecture, pure data flow, or mathematical category theory. These classifications each try to put a structure to the data-fusion process. However, as can be seen in previous examples, those structures depend on a specific application and cannot be described in one uniform classification.

From an examination of the existing fusion methodologies at all stages of data fusion, it can be seen that the key component is the transformation/mapping of a set of measurements/observations from one domain to another. For example, in the case of a radar sensor used for target tracking (e.g., Stone 2001), observations in measurement space (radar) are mapped to the space of location and time period. Different sensors have different measurement spaces. To combine data from all sensors, data from measurement spaces for target tracking often need to be mapped into the same location and time space. The mapping can take many different forms, depending on the domains and the objectives of the mapping. For example, with radar sensor data, spectral data are mapped to the location and time space, e.g., in Kalman filtering, neural networks, and splines (Campbell and Samaan 1988; Wahba 1990; Kay and Titterington 1999; Hall and McMullen 2004). If the purpose of the radar data is to identify a target, then mapping will be pattern recognition, which can take the form of templating, clustering analysis, an adaptive neural network, or a knowledge-based technique. If we use X to denote the original measurement space of a sensor, Z the (location, time) space, and M_X the mapping (be it Kalman filter, neural network, splines, clustering analysis, or Bayesian analysis, etc.), then we can write: $X \xrightarrow{M_X} Z$ and use (X, Z, M_x: $X \xrightarrow{M_X} Z$) to represent the process of the mapping of space containing X to space containing Y, with method of mapping M_X. The elements in either X or Z can take continuous values or a discrete value, such as 0 or 1.

With the notation above, data fusion can be described as one-time data-mapping process, or a process that maps two or more times. Therefore, it can be a multi-path and multi-stage construction of mappings, depending on the specific data-fusion application.

In the example of the effect of fluoridated water on dental cavities, let X_1 be the collection of observations of the severity of the brown stains and the severity of cavities, Y_1 the correlation coefficient with its uncertainty assessment, and M_1 the regression model. Then, (X_1, Y_1, M_1 : $X_1 \xrightarrow{M_1} Y_1$) represents the mapping between the severity of the brown stains and the severity of the cavities and their correlation assessment. Similarly, (X_2, Y_2, $M_2 : X_2 \xrightarrow{M_2} Y_2$) represents

the mapping between the severity of the brown stains and the fluorine content in the drinking water and their correlation assessment. For Y_1 and Y_2 combined, noted as (Y_1, Y_2), and letting Z denote the decision space as 0 or 1, with 1 = the drinking water reduceing cavities and 0 otherwise, and M_Y be the method of hypothesis, then, ((Y_1, Y_2), Z, MY: $(Y_1, Y_2) \xrightarrow{My} Z$) represents the second stage of the mapping process of the forming of the fluorine-cavity hypothesis. In this example, with observations from the prospective study denoted as W, Z as above and M_W the method of hypothesis testing, the last stage of the data fusion is $W \xrightarrow{M_W} Z$, the mapping of W to Z. There are a total of three stages of mappings to arrive to a final conclusion. Using the multi-stage mapping approach, we can have the schematic shown in Figure 3.1. This schematic is a symbolic version of Figure 2.1.

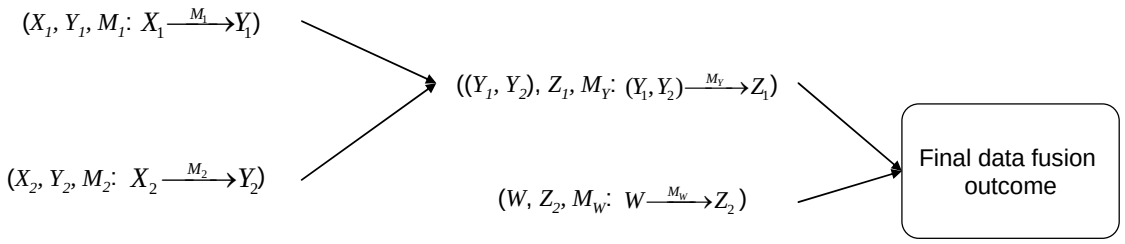

Figure 3.1. Schematic of multi-stage mapping of fluorine-cavity situation

In the case of the effect of diesel on humans and animals, each individual experiment (i^{th} experiment) has mappings of the dose-response slope and coefficient of variation (Y_{i1}) to the dose (X_{i1}) through the dose-response model (M_{i1}). Letting *(Y_{11}, Y_{21}, …, $Y_{37,1}$)* be the collection of all the (Y_{i1}), Z the collection of all the dose-response slopes and their coefficients of variation for all the species and chemicals (thus, there are 50 components in Z), and M the Bayesian methodology, then (*(Y_{11}, Y_{21}, …, $Y_{37,1}$), Z, M*) represents the final mapping of the data-fusion process. So, in this case, there are two constructions of mappings. See Figure 3.2 for a schematic.

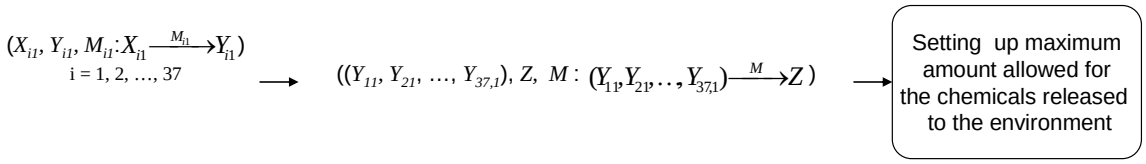

Figure 3.2. Schematic of multi-stage mapping of the effect of diesels on human and animals

The mappings get more complicated in the case of battlefield management. Letting X_C be the collection of measurements from a chemical/biological sensor for target identification and Y_C the collection of values resulted from transforming/estimating X_C with mapping method M_C, then, (X_C, Y_C, M_C : $X_C \xrightarrow{M_C} Y_C$) represents the process. Similarly, we can have (X_R, Y_R, M_R : $X_R \xrightarrow{M_R} Y_R$) for the radar sensor, (X_A, Y_A, M_A : $X_A \xrightarrow{M_A} Y_A$) for the acoustic sensor, (X_T, Y_T,

$M_T: X_T \xrightarrow{M_T} Y_T$) for the text message, etc. The next natural stage is to combine sources of information obtained to identify potential targets, 1, 2, …, n. This process can be *n* separate processes for those *n* target assessments, or it can be one process. Figure 3.3 depicts *n* separated processes/mappings. The last stage of the data fusion is the tactical situation assessment for commanders to take necessary actions.

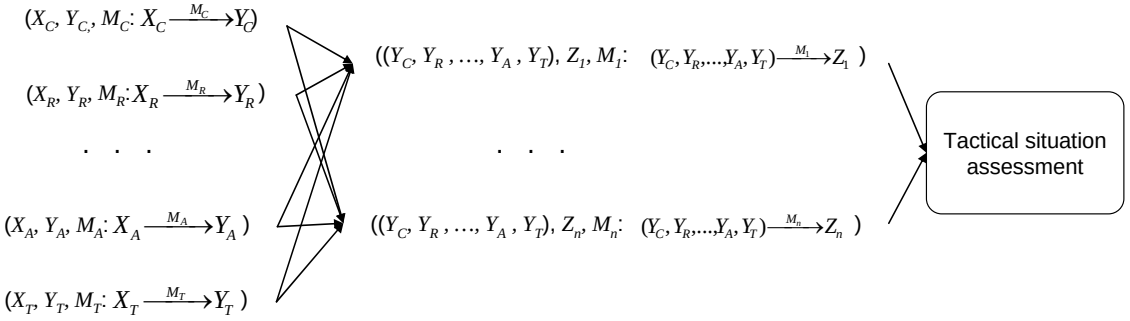

Figure 3.3. Schematic of multi-stage mapping of battlefield management

III.2. Data-fusion methodologies

Methodologies or mappings for data fusion include pixel-level fusion, Bayesian theory, the Dempster-Shafer theory of evidence, neural networks, the Newman-Pearson criteria, fuzzy logic, knowledge based systems, or Markov random fields (Gros 1997; Hall and McMullen 2004).

Data can be generated from radar or biological sensors, microarrays, or other spectral data, can be sets of new data, often referred to as kinematic and attribute estimation (extracted features) in target tracking or as biological signature generation in biological biomarker discovery, and can involve mapping methods such as a Kalman-filter, Alpha-Beta filter, least squares estimation, and principle component analysis (Hilario, Kalousis et al. 2003; Hall and McMullen 2004). From either a set of raw or extracted features to a set of identities, the methods used include clustering, physical templating, pattern recognition, and knowledge-based database matching (Hilario, Kalousis et al. 2003; Hall and McMullen 2004; Gasparini and Hayes 2005; Johnson, Davis et al. 2005). For mapping a set of identities to another set of identities, methods used include Bayesian, Dempster-Shafer, and fuzzy logic (Carl 2001; Mahler 2001; Stone 2001; Hall and McMullen 2004).

Data fusion is not a simple combination of data from all sources; it includes consolidation, re-organization, and abstraction of data (Antony 2001). The purpose of fusion is to optimize the total information content from multiple sources. Antony (2001) pointed out that total information content can be enhanced in at least four approaches for the case of multiple sensors fusion:

1. New sensors can be used to provide more data, and old sensors can be improved.
2. Similar sensors can be added to provide more coverage or more confidence for observed data.

3. Dissimilar sensors can be used to complement the other sensors.
4. Domain knowledge can be used to constrain the decision process.

These four approaches can be extended to integrating information other than from sensors. Specifically, these approaches can be used to create composite knowledge signatures from multiple sources. Suppose that multiple signatures have been created from each individual source (e.g., remote imaging, text documents, and spectral analysis, etc.). The composite knowledge signatures can be formed by 1) creating new signatures and improving the existing ones from raw data; 2) adding additional signatures to the existing ones to increase coverage; 3) studying the dissimilarity among signatures, and creating signatures that complement each other; or 4) using expert knowledge to facilitate the above three ways to fusion.

Selection of mapping methodology depends on the effectiveness measures. There is no uniform standard for choosing one mapping over the other. In the case of the Bayesian approach, the evaluation is done through minimizing cost functions or maximizing the posterior distribution functions (Hall and McMullen 2004). The development of methodologies is coupled with the development of effectiveness assessments of the methods and data-fusion outcomes. Research into the effectiveness of assessments/criteria and of methodologies is needed to broaden data fusion to the next level of advancement.

III.3. Data-fusion system development

To facilitate data-fusion activities, a software or hardware system needs to be developed. This system should be user-friendly and easily modified. Current data-fusion systems are methodology-specific and need to be reworked when new methodologies are implemented. With the framework proposed here, the fusion-process structures and mapping methodologies can be developed independently. A collection of methodologies can be managed independently in an analytical tools database and can be retrieved when needed. To further facilitate the fusion process, collections of visualization and evaluation tools for comparing different fusion outcomes should be developed as well. With those four components independently managed, specific data-fusion activities can be performed by retrieving elements from those four components, as illustrated in Figure 3.4.

Many data-fusion systems are implemented or prototyped for military applications (Dannenberg and Smetek 1984; Cato and Simmen 1987; Hafner and Thompson 1988; Antony 1995; Antony 2001; Bowman and Steinberg 2001). The need for more standard effectiveness measurement/criteria to evaluate a fusion system, as well as to evaluate an individual algorithm, is well recognized (Bowman and Steinberg 2001; Hall and McMullen 2004; Kokar, Tomasik et al. 2004). Either methodologies or effectiveness criteria should be generalized from examination of specific fields of application fields, such as military, medical diagnosis, and genomics and proteomics, which would benefit most from such fusion-system development. In fact, software such as BLAST, MASCOT, SEQUEST, In-SPIRE, OmniViz, and ACQUIRE are examples of fusion systems currently applied to genomics and proteomics.

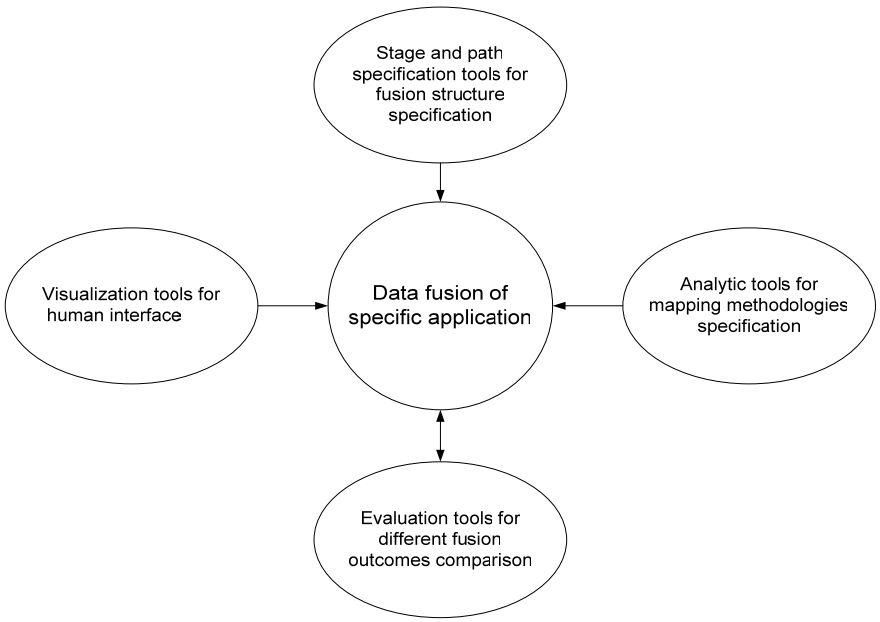

Figure 3.4. Schematic of a unified approach to data fusion

IV. Future Work of Data/Information Fusion

Data fusion (information integration) takes many forms, from simple exploratory data summarizing to sophisticated expert system with evidential reasoning. Developments in both methodology and system implementation are still evolving. This paper presents three specific examples in which data fusion is present and points out that a software system will benefit the knowledge-discovery process and facilitate further understanding of the problem at hand. Because methodologies are often specific to a scientific discipline and dependent on the evaluation criteria selected, new problems and new criteria will trigger the development of new methodologies. The unified approaches proposed here allow the collection of methodologies and evaluation criteria to expand and evolve independently of the fusion process. System development facilitates the selection of methodologies based on chosen evaluation criteria, combined with the selection of the mapping stages and paths. Extending the system developments shown in the three examples in section II into similar disciplines will demonstrate the advantages of the proposed unified approach. Future work should, therefore, be in selecting a collection of mappings, methodologies, and evaluation tools, and putting together sample systems for selected applications. Once a system for a specific application is ready, new methodologies with various evaluation criteria can be evaluated and compared. In turn, this will facilitate new knowledge discovery in the application fields.

V. References

Antony, R. (2001). "Data fusion automation: a top-down perspective." Handbook of multisensor data fusion. D. L. Hall and J. Llinas. New York, CRC Press**:** 6-1 to 6-25.

Antony, R. (2001). "Data management support to tactical data fusion." Handbook of multisensor data fusion. D. L. Hall and J. Llinas. New York, CRC Press**:** 18-1 to 18-25.

Antony, R. T. (1995). Principles of Data Fusion Automation. Boston, Artech House.

Ast, D. B., D. J. Smith, et al. (1956). "Newburgh-Kingston caries-fluorine study. XIV. Combined clinical and roentgenographic dental findings after ten years of fluoride experience." J Am Dent Assoc **52**(3): 314-25.

Bowman, C. L. and A. N. Steinberg. (2001). "A system engineering approach for implementing data fusion systems." Handbook of multisensor data fusion. D. L. Hall and J. Llinas. New York, CRC Press**:** 16-1 to 16-38.

Campbell, J. and J. Samaan. (1988). "Algorithm for decisions on merging and linking target tracks." Proceedings of the 1988 Tri-Service Data Fusion Symposium (DFS-88) **1**: 414-424.

Carl, J. W. (2001). "Contrasting approaches to combine evidence." Handbook of multisensor data fusion. D. L. Hall and J. Llinas. New York, CRC Press**:** 7-1 to 7-31.

Cato, A. A. and R. L. Simmen. (1987). "Development of the vehicle integrated defense system feasibility demonstration." Proceedings of the 1987 Tri-Service Data Fusion Symposium (DFS-87) **June** 402-414.

Dannenberg, K. and R. T. Smetek (1984). "The challenge of technology for military applications." J. Electron. Defense **January**: 29-50.

Dean, H. T. (1953). "Some reflections on the epidemiology of fluorine and dental health." Am J Public Health **43**(6:1): 704-9.

Dean, H. T. (1956). "Fluorine in the control of dental caries." J Am Dent Assoc **52**(1): 1-8.

DuMouchel, W. and P. G. Groer. (1989). "A Bayesian methodology for scaling radiation studies from animals to man." Health Phys **57 Suppl 1**: 411-8.

DuMouchel, W. H. and P. G. Groer. (1983). "Bayes methods for combining the results of cancer studies in humans and other pecies (with discussion)." J Am Dent Assoc **78**(382): 293-315.

Gasparini, G. and D. Hayes. (2005). Biomarkers in breast cancer : molecular diagnostics for predicting and monitoring therapeutic effect. Totowa, N.J., Humana Press.

Gros, X. (1997). NDT Data Fusion, Elsevier.

Hafner, A. and B. Thompson. (1988). "Naval command and control system - AFLOAT NCCS(A), the composite warfare commander's tactical decision support system." Proceedings of the 1988 Tri-Service Data Fusion Symposium (DFS-88): 80-99.

Hall, D. L. and J. Llinas. (2001). "Multisensor Data Fusion." Handbook of multisensor data fusion. D. L. Hall and J. Llinas. New York, CRC Press**:** 1-1 to 1-10.

Hall, D. L. and S. A. H. McMullen (2004). Mathematical techniques in multi-sensor data fusion. Boston, Artech House.

Hilario, M., A. Kalousis, et al. (2003). "Machine learning approaches to lung cancer prediction from mass spectra." Proteomics **3**(9): 1716-9.

Hilleboe, H. E. (1956). "History of the Newburgh-Kingston caries-fluorine study." J Am Dent Assoc **52**(3): 291-5.

Johnson, R. S., M. T. Davis, et al. (2005). "Informatics for protein identification by mass spectrometry." Methods **35**(3): 223-36.
Kay, J. and D. Titterington, Eds. (1999). Statistica and neural netwokrs: advances at the interface, Oxford University Press.
Kokar, M., J. Tomasik, et al. (2004). "Formalizing classes of information fusion systems." Information Fusion **5**: 189-202.
Mahler, R. (2001). "Random set theory for target tracking and identification." Handbook of multisensor data fusion. D. L. Hall and J. Llinas. New York, CRC Press**:** 10-1 to 10-30.
Pemberton, W. G., M. S. Dotterweich, et al. (1987). "An overview of ATR (automatic target recognition) fusion techniques." Proceedings of the 1987 Tri-Service Data Fusion Symposium (DFS-87) **June** 115-123.
Schlesinger, E. R., D. E. Overton, et al. (1950). "Newburgh-Kingston Caries-Fluorine Study; pediatric aspects--preliminary report." Am J Public Health **40**(6): 725-7.
Schlesinger, E. R., D. E. Overton, et al. (1953). "Newburgh-Kingston caries fluorine study. V. Pediatric aspects; continuation report." Am J Public Health **43**(8): 1011-15.
Schlesinger, E. R., D. E. Overton, et al. (1956). "Newburgh-Kingston caries-fluorine study. XIII. Pediatric findings after ten years." J Am Dent Assoc **52**(3): 296-306.
Simpson, W. R. and B. A. Kelley. (1987). "Multidimensional context representation of knowledge-base information." Proceedings of the 1987 Tri-Service Data Fusion Symposium (DFS-87) **June** 239-246.
Steinberg, A. N. and C. L. Bowman. (2001). Revision to the JDL data fusion model. Handbook of multisensor data fusion. D. L. Hall and J. Llinas. New York, CRC Press**:** 2-1 to 2-18.
Stone, L. D. (2001). "A Bayesian approach to multiple-target tracking." Handbook of multisensor data fusion. D. L. Hall and J. Llinas. New York, CRC Press**:** 10-1 to 10-30.
Wahba, G. (1990). Spline models for observational data. Philadelphia, US: Society for industrial and applied mathematics.
Waltz, E. (1987). "Data fusion functions required for tactical situation assessment." Proceedings of the 1987 Tri-Service Data Fusion Symposium (DFS-87) **June** 176-186.

Distribution

No. of Copies

ONSITE

3 Pacific Northwest National Laboratory

X.C. Lei	K6-08	P.D. Whitney	K6-08
C. Posse	K6-08		