LA-UR-00-4850

Approved for public release;
distribution is unlimited.

Title:	Statistical Representations for Information Integration
Author(s):	Alyson Wilson Sallie Keller-McNulty
Submitted to:	Army Conference on Applied Statistics October 16-17, 2000 Rice University Houston, Texas

Los Alamos
NATIONAL LABORATORY

STATISTICAL REPRESENTATIONS FOR INFORMATION INTEGRATION

Alyson Wilson and Sallie Keller-McNulty
LA-UR 00-4850

1. A DECISION-MAKING FRAMEWORK

One of the basic problems in statistics is decision-making under uncertainty. What is the best decision to make given incomplete information? This problem arises in many applications, including deciding whether a new medication is safe and effective, predicting global climate change, and designing a new machine. At Los Alamos, some of the most important decisions that must be made with incomplete information are how to integrate data, information, and knowledge from the experiments, computational models, past tests, sub-system tests, and the expert judgment of subject-matter experts to provide a rigorous, quantitative assessment, with associated uncertainties, of the safety, reliability, and performance of the nuclear stockpile. These decisions are addressed using science-based stockpile stewardship (SBSS).

The complexities of big science problems such as SBSS can quickly become overwhelming, and without careful attention to the whole picture or purpose, the accomplishments of individual scientists can become lost and detached. As some of the key information integrators, we have recently gone back to the "beginning" and reformulated our basic understanding of how decision-making under uncertainty works. This has led to the understanding that is captured in Figure 1. This diagram represents a "time-dependent decision framework." The type of data available and the structure of the analyses in SBSS, as in many other problems, can change dramatically over time; therefore, a useful framework, such as the one depicted in Figure 1, must be flexible enough to capture these changes. This framework represents a dynamic and recursive space where each box has potential to produce new information that can update any other box,

resulting in other updates. The goal is that at any slice in time, the best possible information is available to inform decision-making.

INSERT FIGURE 1 ABOUT HERE

The first piece of decision-making is to define the decision objectives: what is it that must be understood and decided? This step is frequently overlooked by analysts in the rush to collect, model, and analyze data—at least partially because they are well-trained in modeling and data analysis, and not so well-trained in the complexities of understanding decision contexts.

The second piece of decision-making is to try to understand who the "stakeholders" are in the decision process and what their perspectives on the problem are. Groups of stakeholders with similar ways of thinking about the decision are often called "communities of practice." For example, within SBSS, there are many communities of practice. These include physicists, who approach problems by trying to understand the physical processes involved; weapons designers, who are more concerned about harnessing the physical processes; engineers, who think about problems in terms of interacting components; statisticians, who consider issues of uncertainty quantification; computer scientists, who want to understand how complex computer codes work; and politicians, who care about the policy implications of the science. Each of these communities approaches the problem from a different viewpoint, and each represents its data, information, and knowledge in a different way.

The third piece of decision-making is the problem representation and analysis strategy. Before any information is collected, it must be determined how this information will be analyzed and integrated, and how the results will bring better resolution to the decision objectives. These determinations should drive the requirements regarding what to collect.

The fourth piece of decision-making is data, information, and knowledge. Notice that this part of decision-making contains more that just "data" in its traditional narrow sense. All decisions incorporate more than just data: they also include the information and knowledge, i.e., expertise and theory, to do such things as understand the problem, structure the representations, find data sources, and select appropriate models. Even "data" in its narrower sense can include such things as opinions elicited from experts and outputs from computer codes.

The fifth piece of decision-making, as shown in Figure 1, is the "information integration," or the methodologies needed to tie all of the decision objectives, community representations, and information together. If these methodologies are effective, they lead to the sixth piece of decision-making, which is inference (with associated uncertainties) about the decision objectives of interest. This inference must be dynamic, or performed over time, because the data, information, and knowledge (including the problem representation) about the problem change continuously.

The general decision-making structure of Figure 1 leads nicely to a formalization of the statistical problem-solving processes used to approach complex problems. An example used in the development of an automotive system is diagrammed in Figure 2. Here the primary decisions were guided by various performance metrics, e.g., the predicted design reliability of the system and the expected number of failures per thousand units produced. The tasks enumerated in Figure 2 use a statistician's language to describe the general ideas in Figure 1. Each task has a body of research behind it; in particular, statisticians are the most familiar with calculating performance (inference) and performing "what-if" (predictive) analysis. The remainder of this paper will focus on surveying techniques useful for one of the less familiar boxes; specifically on statistical methods for structuring the system and representing data, information, and knowledge.

INSERT FIGURE 2 ABOUT HERE

More information about creating knowledge bases can be found in Meyer and Paton (2000). Additional information about the elicitation and quantification of expert judgment can be found in Meyer and Booker (1991).

2. AN ILLUSTRATIVE EXAMPLE

The example that will be used to demonstrate the statistical methods for problem structuring and data representation is one of current intense interest within the Department of Defense (DoD). Fratricide, the injuring or killing of one's own troops with friendly fire, became a timely issue during the Gulf War, "where most ground units' only combat ID capability was an inverted V taped to their sides" (Carroll 1999). Some estimates put the Gulf War friendly-fire casualty rate at 25%; historical rates are about 15%.

Fratricide is usually associated with a failure of identification-friend-or-foe (IFF), also known as *combat ID*. The purpose of combat ID is to correctly put potential targets into one of three categories: friendly, hostile, or unknown. Once a target is categorized, the rules of engagement determine whether a shot should be taken. For example, in U.S. Army air defense, a weapons control status of "hold" means do not shoot at anyone; a weapons control status of "tight" means shoot only hostiles; a weapons control status of "free" means shoot hostiles and unknowns. Clearly, one would like to be correctly identified as friendly by one's own troops.

Most of the recent discussion of combat ID has been technological, although, as the above discussion points out, there are also command and control issues involved. One standard method for combat ID is using the interrogation/response (or "What's the password?") model: if an interrogation signal is sent out and the target responds correctly (often using a transponder), it is

classified as friendly, if it responds incorrectly, it is classified as hostile, or, if there is no response, it is classified as unknown. It is the "no response" case that causes problems: the target could be friendly, with a broken transponder, or it could be hostile, with its transponder turned off to avoid making an incorrect response.

Another newer method for combat ID is the "situational awareness" (SA) model. The upper echelons of military command have always tried to keep track of where their troops are with respect to the enemy. Modern technology has enhanced this ability through real-time battlefield communications systems like EPLRS, the enhanced position location and reporting system. However, the real future of SA comes from systems like FBCB2 (Force XXI Battle Command Brigade and Below), which promises to give each unit, and perhaps each soldier, the ability to display on a portable device all known friendly and hostile forces on a battlefield. This is accomplished using a combination of EPLRS and portable lightweight GPS receivers (PLGRs) to locate each soldier.

The effectiveness and reliability of combat ID systems is a dynamic and important problem for the DoD, and will be used as the illustrative example in the discussion that follows.

3. THE GENERAL PROBLEM

One example of a formal statistical process for approaching complex problems is diagrammed in Figure 2. Assume that some care has been taken in deciding what decision is to be made and what measures of performance (metrics) need to be estimated to inform that decision. The next step in the process is the creation of several representations of the "system" under study. These representations should:

1. capture all of the factors that affect the measures of performance;

2. outline the components and subsystems of the larger "system" and how they interact;
3. identify the information that feeds into the estimation of the metrics; and
4. specify the methods for "rolling-up" the information and quantifying uncertainty about the metrics.

This process of specifying representations operates very much in terms of a "structural" versus a "predictive" model. Often in statistics, models are created to allow prediction without trying to understand the causes underlying the predictions. Think about the standard plot in elementary regression analysis where height is plotted against weight. Weight is a good predictor of height, but it does not give any causal explanation of height. The representations created in this problem-solving process are geared toward capturing the factors and information that might lead to a causal understanding of the measures of performance.

The general forms of the representations used by the information integration technologies have three parts: icons/pictures/diagrams, rules/statements, and abstract mathematics (Paton 1994). Rules and statements are used when we have observable phenomena that have been characterized by "physical laws" or statistical relationships; abstract mathematics similarly captures "physical laws" about unobservable phenomena. Much of the following discussion concentrates on the diagrams, which are used to gain understanding about systems through metaphor and analogy.

Paton (1994) suggests that "thinking in terms of systems is probably the major reasoning paradigm in the biological sciences." Arguably, the "system" metaphor is applicable to many scientific problems outside the biological sciences. Paton (1994) discusses five systemic metaphors that re-appear throughout science: a circuit (representing flow, transfer, conduit, network, pipeline); a machine (representing input-output, mechanism, purpose, control, balance);

a text (representing context, theme, grammar, hierarchy, interpretation); an organism (representing organized complexity, adaptability, reproduction, openness); an ecosystem (representing niche, competition, environment, distributed processes). Specifically, these metaphorical graphs are often used to represent the relationships between the system and its components: a chain of components (machine), a network of components (circuit), a hierarchy of components (text, organism), or a cluster of components (ecosystem).

On the multi-disciplinary teams (with multiple communities of practice) that are often formed to develop information integration technology processes to address complex problems like fratricide or SBSS, many different metaphors like these will be used to help determine the ways in which the group understands the complex relationships of "the problem." For example, do commanders think of the battlefield as a circuit, ecosystem, organism, text, machine, or something else? For each of these metaphors, there are diagrammatic forms that then can be used as templates to begin creating diagrams of the features and relationships of the problem. Some of these diagrams are used later as a starting point for the knowledge base, and others as planning tools for statistical analyses. For example, where are the data, information, and knowledge sources, and how do their relationships suggest ways to factor them into the analyses? The decision framework outlined in Figure 1 utilizes this multi-perspective information from the multiple communities of practice to produce more robust analyses.

4. REPRESENTING SYSTEMS

There are many types of diagrams that can be used to accomplish the four primary purposes of diagramming and representing the "system" under study. However, they share common features. The basic components of the diagrams are "boxes," which represent where data,

information, and knowledge can be collected, and arcs/arrows/lines, which represent the information flow between boxes.

Figure 3 is notional *scratch net* (Paton 1993, 1996) for the fratricide issue. "Scratch net" is the formal name for the informal picture that is often drawn to start understanding which factors are important for assessing the decision metrics. Scratch nets collect ideas around a central organizing question or issue. If interconnections are added between the peripheral "boxes" in a scratch net, then it becomes a *factor complex*. Scratch nets and factor complexes can be made more complicated by adding hierarchy and directed arrows. The scratch nets, scratch net factors, and more detailed representations of the factors will change over time.

INSERT FIGURE 3 ABOUT HERE

Moving from the identification of the factors influencing the "system" under study to representations that can be populated with data, information, and knowledge for the purpose of analysis draws in many types of representations. There are two types of diagrams familiar to statisticians that are useful for analysis: trees and networks. Trees have a much more limited structure than networks, with a rigid hierarchical structure and a fixed set of relationships along the arcs. Information flows only between parent and child nodes. Networks (also called graphs) can represent more relationships with a richer set of connectors and a more flexible set of allowable connections.

4.1 Trees

The basic tree model is the *decision tree*. At each node of a decision tree there is a question or event; arcs coming from each node correspond to the answers to the question or occurrence of an event. A specific kind of decision tree is the *event tree*. These have been commonly used to model accident scenarios. An initiating event is chosen, and then a possible sequence of events is

selected. A tree with 2^N branches is formed by assuming that each event happens/does not happen. Impossible branches are pruned away. Figure 4 shows notional unpruned event tree (4a); the pruned tree (4b) captures some of the factors and their relationships from the scratch net (Figure 3).

INSERT FIGURE 4 ABOUT HERE

A *fault tree* is often used to determine the causes of an undesirable event. It is basically failure oriented. The undesirable event is put at the top of the tree, and intermediate events leading to the "top event" are branched off in their own boxes with lines connecting them. The lines contain the Boolean operators AND and OR. The AND operator requires all connected events to happen before the next event occurs; the OR operator requires one or more of the connected events to happen. Some fault trees use a somewhat richer class of operators, including an INHIBIT gate and an N-of-M gate. Figure 5 shows an example fault tree for the factors in the scratch net (Figure 3) that correspond to the reliability of interrogation/response (using a transponder). The top gate is an AND gate, and the bottom gate is an OR gate.

INSERT FIGURE 5 ABOUT HERE

Process trees are a generalization of fault trees. As used by Eisenhawer and Bott (1999), "For the class of undesired events . . . the tree describes alternative process paths and the phenomenological relationships among the discrete processes." The gates try to capture the ideas of possibility, necessity, and causality. Possibility, necessity, and causality are all defined as metrics on [0,1]. Both fault trees and process trees can be evaluated numerically—the methods for fault trees are commonly known (see Lewis 1987); process trees are evaluated by making Boolean approximations to the possibility and necessity gates.

Classification trees are decision trees that have been developed specifically to determine the group or class for an observation. Each node contains a decision rule that sends the observation down one branch; each "leaf" of the tree is labeled with a classification. *Regression trees* specify a piecewise polynomial fit at each leaf of a classification tree.

Tree structures can be useful for representing system components, information flow, or statistical analyses in particular problems. However, not all communities of practice are comfortable using tree structures. For example, in the SBSS problems, physicists cannot use fault tree structures because the representation is not rich enough to capture how the physics processes relate on a continuous scale and interact with multiple subsystems and components of the nuclear weapon. Network representations, discussed in the next section, are more suitable for their needs.

4.2 Networks and Graphs

Networks and graphs have a more general structure than trees, in that the intermediate nodes can be connected to each other. Two of the most familiar graph representations are *reliability graphs* and *reliability block diagrams.* Reliability graphs and block diagrams, as their names suggest, are used to model system reliability by capturing either the physical interconnection of parts (usually called the reliability graph) or system failure dependencies (usually the reliability block diagram or success graph). Reliability graphs and block diagrams contain much of the same information as fault trees, but they are typically success oriented. The series graphs model the AND gate; the parallel graphs model the OR gate. Non-series and non-parallel connections are also used. Figure 6 shows a sample reliability block diagram for the situational awareness/FBCB2 factor in the scratch net (Figure 3). The first box contains parallel subcomponents; the second contains serial components.

INSERT FIGURE 6 ABOUT HERE

Markov models, or *state-transition graphs*, are used to capture the proportion of time that a system spends in a particular state. In reliability terms, states are defined as combinations of operating and failed components; in software testing, the states are conditions of system use (Poore and Trammell 1998). Sometimes the events are further classified as those that are "successes" and those that are "failures." Directed arcs connect the states, and transition probabilities between the states are assigned to each arc. The usual use for these models is to estimate the proportion of time spent in each state, or perhaps a collection of states. See Figure 7 for an example that depicts a formal connection between the factors in the scratch net (Figure 3).

INSERT FIGURE 7 ABOUT HERE

Bayesian graphical modeling tries to bring together four statistical ideas (Spiegelhalter 1998):

1. extensions of generalized linear models to accommodate more complex dependence structures;
2. Bayesian methods to build probability models and then base inferences on the conditional probability of the quantity of interest given the observed data;
3. Markov chain Monte Carlo computational techniques;
4. pictorial representations of the qualitative conditional independence assumptions underlying a model.

The pictorial representations serve three functions: they are an accessible description of the statistical model, they make it easy to get formal conditional independence statements, and they provide a direct link with computational solutions through MCMC methods.

Formally, Bayesian graphical models are directed acyclic graphs. The nodes are discrete or continuous random variables, and the arrows between nodes represent conditional dependence.

Any node is conditionally independent of all of its non-descendants given its parents. This relationship allows the graph to be used as a "theorem prover" for conditional independence between random variables, and it greatly simplifies the implementation of MCMC techniques. Indeed, these are the representation's primary strengths. Bayesian graphical models are often known as *Bayesian networks* when they are used to estimate the probabilities of unobserved quantities conditional on observed or hypothesized quantities. Figure 7, without the probabilities associated with the arcs, is a sample Bayesian graphical model.

Influence diagrams are commonly used in decision analysis, and share many features with Bayesian graphical models. The arrows between "chance" nodes (random variables, represented by circles) also represent conditional dependence. There are several other types of nodes in influence diagrams: decision nodes (represented by rectangles), deterministic nodes representing mathematical relationships (represented by double circles), and value nodes (represented by diamonds), which represent the quantity to be optimized. Arrows entering decision nodes represent information available to inform the decision at that node. See Figure 8 for an example. In Figure 8, the decision of interest is whether or not to engage a target. There are values assigned to each combination of engaging the target and combat ID (e.g., engage + hostile is "very good," engage + friendly is "very bad," not engage + friendly is "good," not engage + hostile = "bad"). Depending on the values and risk preference, the diagram helps guide what decision should be made about engaging the target.

INSERT FIGURE 8 ABOUT HERE

As with the tree-based models, network models can be effectively used to represent system components, information flow, or statistical analyses in particular problems. Smith (1990) identifies three purposes of network models:

1. efficient propagation of probabilities;
2. help in eliciting model structure and the relationships between variables from clients; and
3. help in understanding by the analyst/statistician the model's structure.

With respect to the second purpose, he states, "In this way models can be adjusted and elaborated without needing to confront a client with numerical evaluations of uncertainty (e.g., probabilities) early in the analysis—a process about which many clients harbor great suspicion." This is an important observation—the client and the analyst must both have a representation that they are comfortable thinking about.

5. COMBINING DIAGRAMS

As discussed above, icons/pictures/diagrams are only one way, although a very important way, that system structure can be represented. Other tools include if-then relationships, functional relationships, failure modes and effects analysis, and any of the other statistical models (e.g., linear, generalized linear, hierarchical, reliability, competing hypotheses) that are commonly used. The diagrammatic representation of statistical models is powerful in explaining what has to be done to implement the information integration technologies. Specifically, data, information, and knowledge has to be collected, in whatever form, at each box, and then the relationships between the boxes, probabilistic or otherwise, must be accounted for as estimates and uncertainties are propagated.

In attempting to represent a system, often many different diagrams are drawn. This may be because the members of a multi-disciplinary team draw different pictures to represent the different ways they think about the problem, or it may be because different parts of the system are better or more easily represented in different ways. In the integration process, it is constructive to combine these different representations into a single diagram, not forcing it into

the structure of a specific diagram type, but integrating the different representations into a complex diagram that captures the unique features, relationships, and information flows in the system under study. Figure 9 is a notional example of how some of the fratricide diagrams might be integrated into a more complex representation.

6. REPRESENTING DATA

There are many types of data that can be used to populate the boxes/nodes in the system representation: expert judgment, historical test data, data from similar or relevant systems, design specifications, computer simulation model outputs, physical test data. Each type of data and information has natural ways that it can be represented (see Keller-McNulty and McNulty 2000). These include graphs (like histograms, boxplots, and scatterplots), tables, mathematical models, probability distribution functions, and fuzzy distributions and membership functions.

To do mathematically rigorous information integration, all of the system representations and data representations must be collected and then turned into probability distributions, probabilistic dependency relationships, and statistical models. These models and distributions may not be parametric, but they must somehow be in the language of distributions, dependencies, correlations, conditional independences, and statistical models of various forms. The formal methods for converting the diverse representations of systems and data into statistical models are the subjects of other papers.

7. REFERENCES

S. Carroll (1999). *The Mark of Cain: Avoiding Fratricide*, Journal of Electronic Defense, 22:1, http://jed-prod.weblabs.com/index.html.

D. Edwards (2000). Introduction to Graphical Modeling 2nd Edition. Springer-Verlag, New York.

S. Eisenhawer and T. Bott (1999). Application of approximate reasoning to safety analysis. LA-UR-99-1932.

R. Howard (1990). From Influence to Relevance to Knowledge, in *Influence Diagrams, Belief Nets, and Decision Analysis*, eds. R. M. Oliver and J. Q. Smith, John Wiley and Sons, New York.

S. Keller-McNulty and M. McNulty (2000). Show me the data: statistical representations. To appear in *Theoria et Historia Scientarium*.

E. Lewis (1987). Introduction to reliability engineering. John Wiley & Sons, New York.

M. Meyer and J. Booker (1991). *Eliciting and Analyzing Expert Judgment: A Practical Guide*. Academic Press, San Diego.

M. Meyer and R. Paton (2000). Interpreting, Representing and Integrating Scientific Knowledge from Interdisciplinary Projects. To appear in *Theoria et Historia Scientarium*.

R. C. Paton (1993). Understanding Biosystem Organisation I: Verbal Relations. *International Journal of Science Education*, 15: 4, 395-410.

R. C. Paton (1996). On an Apparently Simple Modelling Problem in Biology. *International Journal of Science Education*, 18: 1, 55-64.

R. C. Paton, D. M. Jones, and M. J. R. Shave (1994). The Role of Verbal and Visual Metaphors in Scientific Theories, Conference on Information-Oriented Approaches to Language, Logic, and Computation, St. Mary's College, Moraga, California, June 13-15.

J. Poore and C. Trammell (1998). Application of statistical science to testing and evaluating software intensive systems. In *Statistics, Testing, and Defense Acquisition*, National Academy Press, Washington, DC.

J.Q. Smith (1990). Statistical Principles on Graphs, in *Influence Diagrams, Belief Nets, and Decision Analysis*, eds. R. M. Oliver and J. Q. Smith, John Wiley and Sons, New York.

D. Spiegelhalter (1998). Bayesian graphical modeling: a case-study in monitoring health outcomes. Applied Statistics 47, Part 1, 115-133.

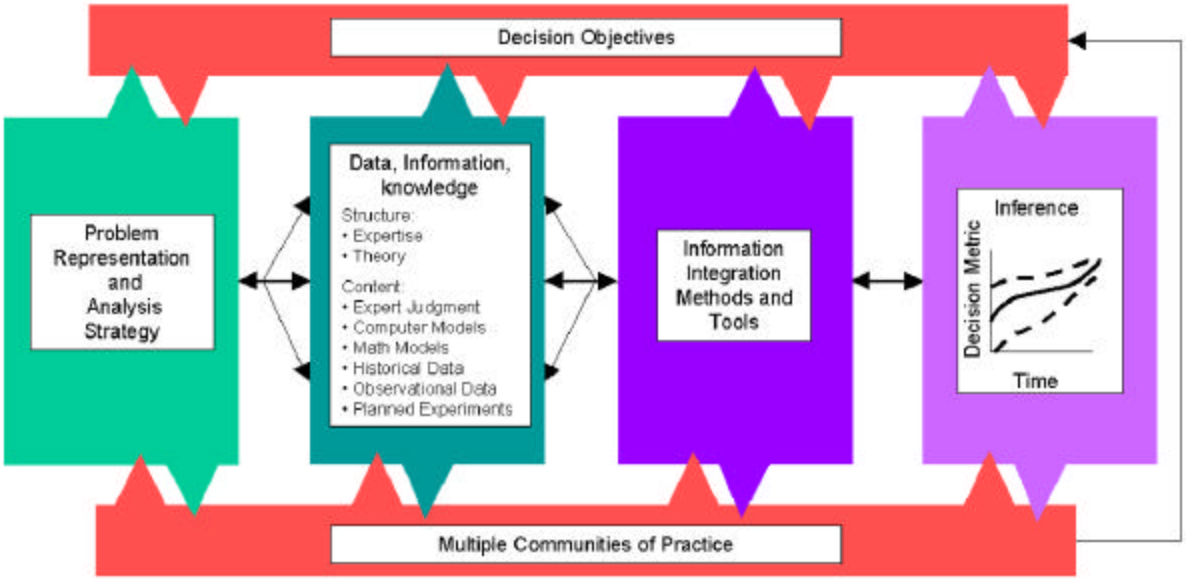


Figure 1. Information Integration Technology Framework

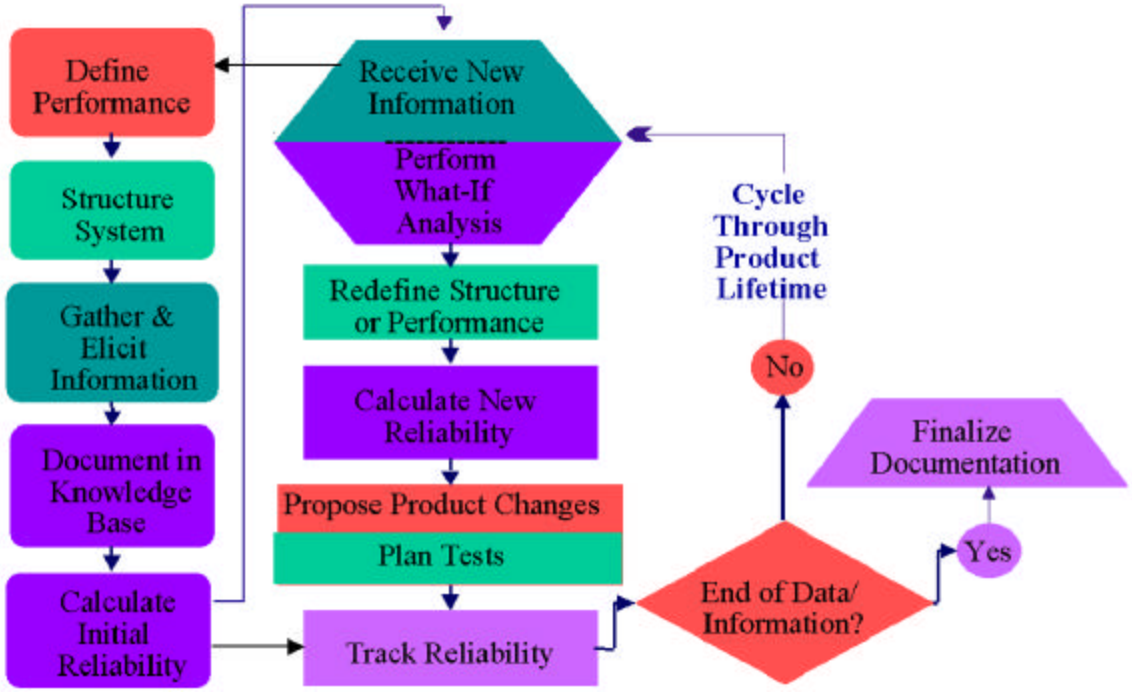


Figure 2. Sample Problem-Solving Process

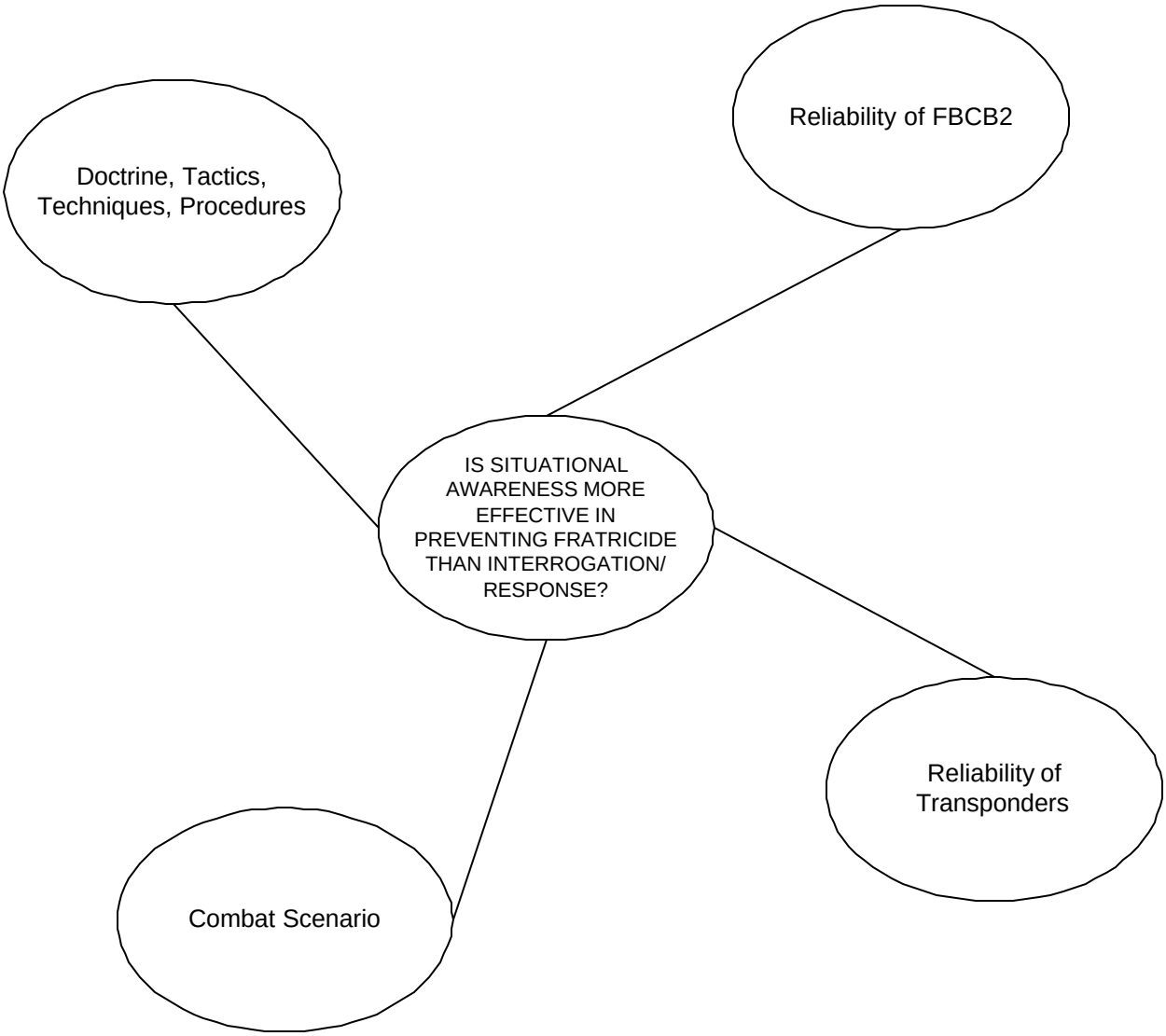
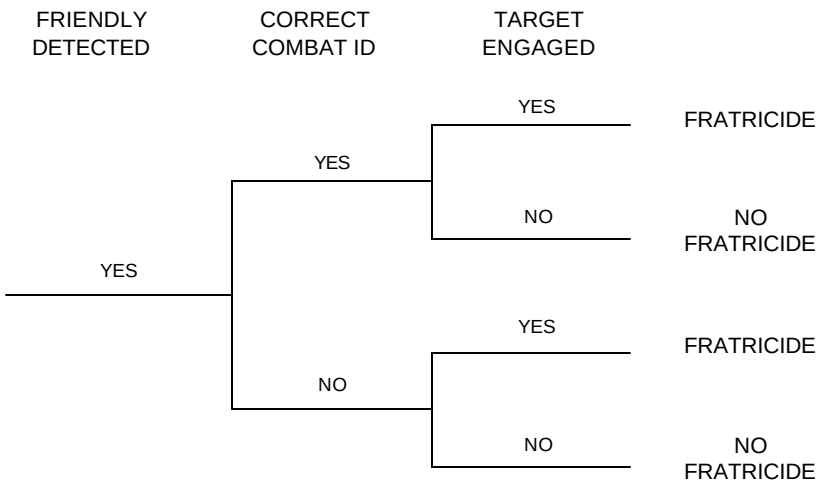
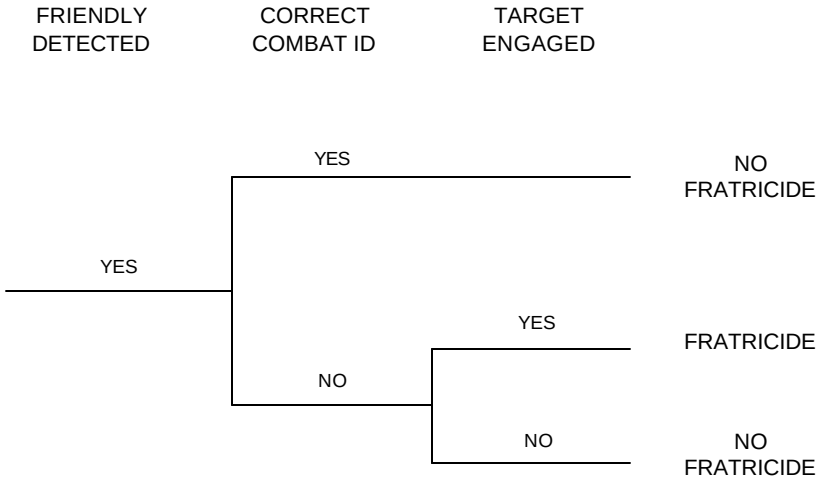


Figure 3. Fratracide Scratch Net



(a)



(b)

Figure 4. Event Trees

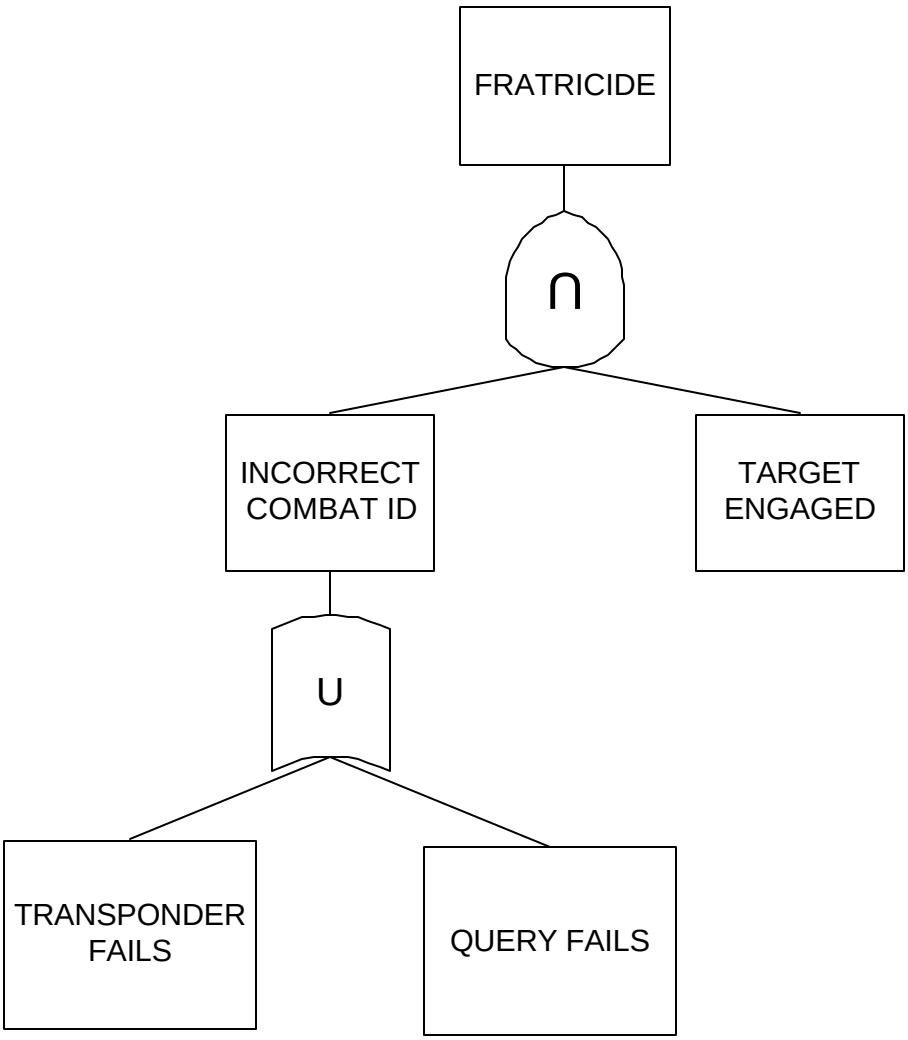


Figure 5. Fault Tree

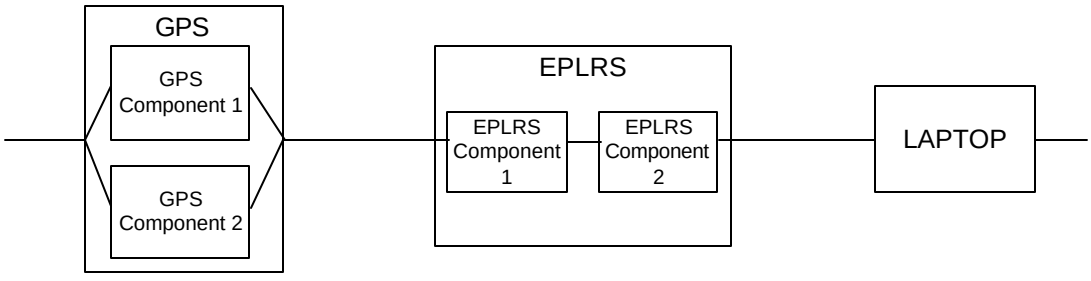


Figure 6. Reliability Block Diagram

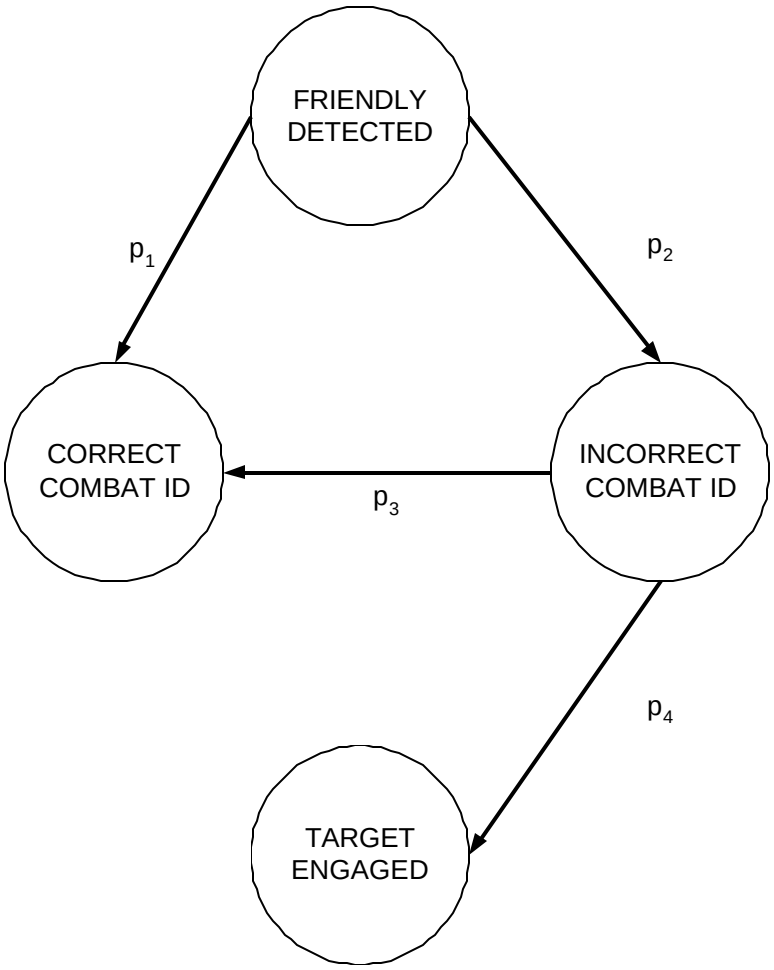


Figure 7. State Transition Graph

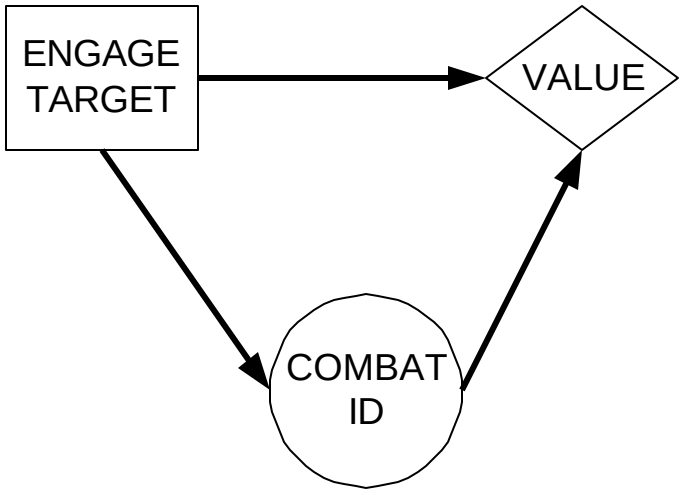


Figure 8. Influence Diagram

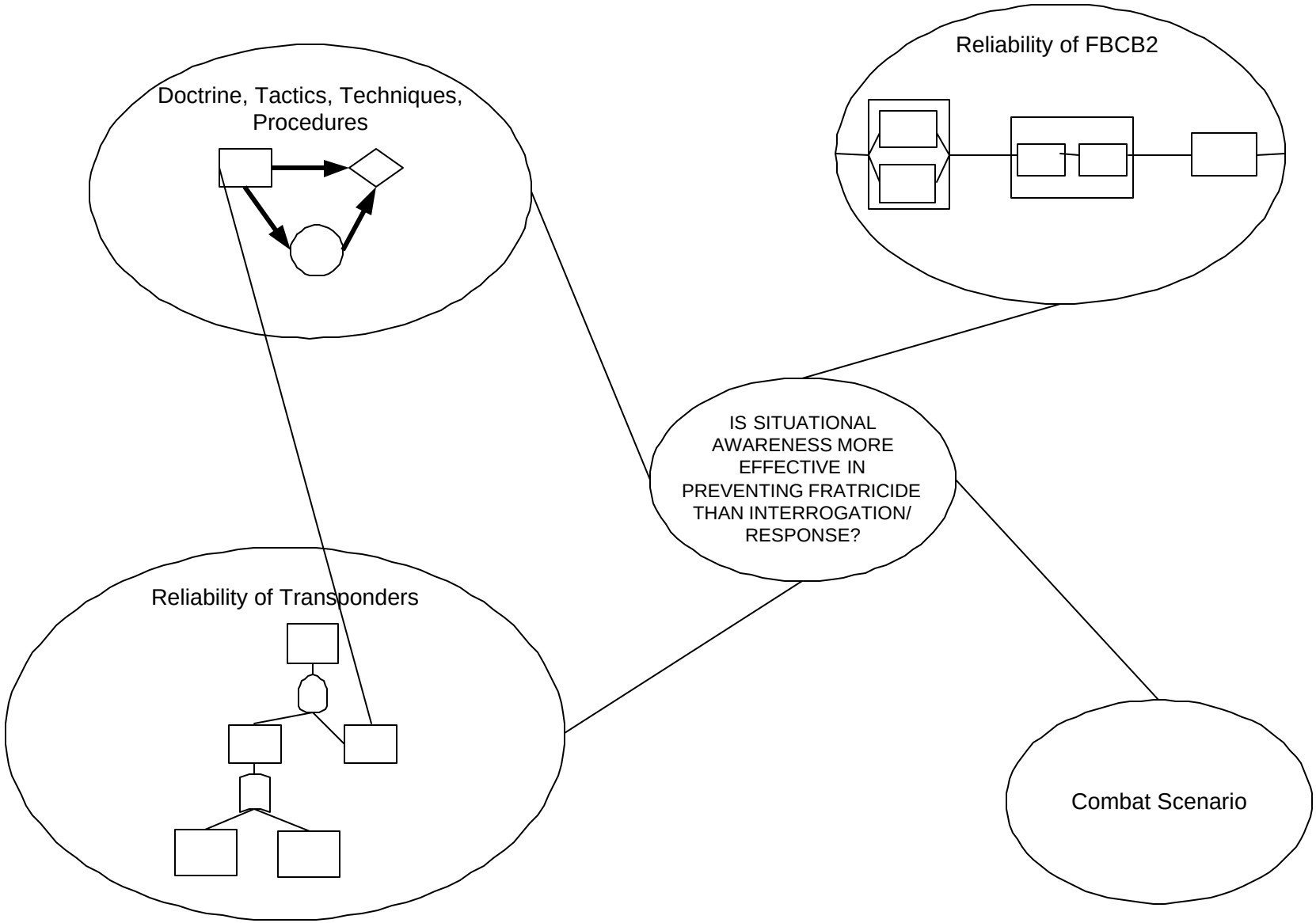


Figure 9. Complex Representation