AFRL-HE-WP-TR-1999-0241

UNITED STATES AIR FORCE RESEARCH LABORATORY

AIDING THE INTELLIGENCE ANALYST IN SITUATIONS OF DATA OVERLOAD: A SIMULATION STUDY OF COMPUTER-SUPPORTED INFERENTIAL ANALYSIS UNDER DATA OVERLOAD

Emily S. Patterson

COGNITIVE SYSTEMS ENGINEERING LABORATORY
INSTITUTE FOR ERGONOMICS
THE OHIO STATE UNIVERSITY
COLUMBUS OH 43210

Emilie M. Roth

ROTH COGNITIVE ENGINEERING
89 RAWSON ROAD
BROOKLINE MA 02445-4509

David D. Woods

COGNITIVE SYSTEMS ENGINEERING LABORATORY
INSTITUTE FOR ERGONOMICS
THE OHIO STATE UNIVERSITY
COLUMBUS OH 43210

MAY 1999

INTERIM REPORT FOR THE PERIOD MARCH 1998 TO MAY 1999

Human Effectiveness Directorate
Crew System Interface Division
2255 H Street
Wright-Patterson AFB OH 45433-7022

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Air Force Research Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

Federal Government agencies and their contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, Virginia 22060-6218

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TR-1999-0241

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

JOHN M. REISING
Acting Chief, Crew System Interface Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 1999	Interim Report - March 1998 to May 1999

4. TITLE AND SUBTITLE
Aiding the Intelligence Analyst in Situations of Data Overload:
A Simulation Study of Computer-supported Inferential Analysis Under Data Overload

5. FUNDING NUMBERS
C: F41624-94-D-6000
PE: 62202F
PR: 7184
TA: 10
WU: 46

6. AUTHOR(S)
* Emily S. Patterson
** Emilie M. Roth
* David D. Woods

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
* Cognitive Systems Engineering Laboratory, Institute for Ergonomics
The Ohio State University, Columbus OH 43210
** Roth Cognitive Engineering
89 Rawson Road
Brookline MA 02445-4509

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory, Human Effectiveness Directorate
Crew System Interface Division
Air Force Materiel Command
Wright-Patterson AFB OH 45433-7022

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-HE-WP-TR-1999-0241

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
A simulation study of inferential analysis was conducted with ten professional intelligence analysts. Using a process tracing methodology, patterns in vulnerabilities were identified when analysts were asked to analyze something outside their base of expertise, were tasked with a tight deadline, and had a large data set. Study participants were vulnerable to missing critical information. All the participants were observed to use relatively primitive search tactics, quickly narrowing in on a set of documents through the addition of keywords to an initial query. All of the participants missed some of the nine documents that were categorized as high quality. A group of four participants who found and relied upon some of the high quality documents took more time, read more documents, and made fewer inaccurate statements in their verbal briefings than a group of four participants who did not. In addition, three sources of inaccurate statements were identified. First, study participants sometimes relied upon assumptions that would normally be correct, but did not apply in this situation. Second, participants sometimes repeated information that was inaccurate in a document that they had read. Third, participants were observed to rely upon information that was considered accurate at one point in time, but then was later overturned in subsequent updates. The main contribution from this research was a model of potential vulnerabilities in inferential analysis under challenging conditions. These vulnerabilities are informative because they point to a set of challenging design criteria that human-centered solutions to data overload need to meet.

14. SUBJECT TERMS
cognitive engineering, cognitive task analysis, data overload, inferential analysis, information retrieval, information visualization, intelligence analysis, simulation study

15. NUMBER OF PAGES
135

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNL

This page intentionally left blank.

PREFACE

This effort was accomplished under Contract F41624-94-D-6000, Delivery Order 0007 for the Air Force Research Laboratory's Human Effectiveness Directorate, under the direction of the Crew System Interface Division, Information Analysis and Exploitation Branch (AFRL/HECA). It was completed for the prime contractor, Logicon Technical Services, Inc. (LTSI), Dayton Ohio, under Work Unit No. 71841046: "Crew Systems for Information Warfare." Mr. Don Monk was the Contract Monitor.

We thank the study participants for donating their valuable time and expertise, and in particular Mr. Jay Finley for his dedication to this effort. Funding was provided by the Human Effectiveness Directorate of the Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio, with exceptional support from Mr. Gil Kuperman. Additional funding was provided by a National Science Foundation Graduate Fellowship. Any opinions, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

This page intentionally left blank.

1 INTRODUCTION

This research is driven by a formidable challenge in many work domains: generating a coherent, reliable description of a situation by searching, interpreting, and corroborating information sampled from an avalanche of electronic data. Data overload is a fundamental, ubiquitous problem. In almost every complex work domain, practitioners are faced with overwhelming amounts of data. In one sense, the problem seems paradoxical because the participants in these fields of practice almost all agree that access to more data ought to be a benefit. However, the benefit in principle has not been matched by the benefit in practice. The sheer volume of the data creates a situation where it is difficult to determine where to look in the data field, it becomes easy to miss critical information, and determining the significance of data in relation to the ongoing context is challenging.

Intelligence analysis is an outstanding natural laboratory for studying the strategies used to conduct analyses under data overload conditions. The demands of intelligence analysis have always included the need to cope with data overload. In the current modern electronic environment, intelligence analysis clearly suffers from the data overload problem on a daily basis. Analysts are responsible for tracking countries and technologies about which thousands of text documents are generated daily, from secret information generated by field agents to "open source" information such as articles in Aviation Week and Space Technology. Reports on the same world events do not necessarily corroborate each other; rather the information is often discrepant along various dimensions (Schum, 1994). Therefore, a central part of the analysis task involves corroborating critical information and resolving data conflicts.

As with many other domains, the data overload problem in intelligence analysis is expected to become even more difficult in the future as the traditional strategies that have been developed to cope with data overload are challenged by two recurring trends: 1) an explosion in the amount of available electronic data and 2) widespread organizational reductions in staffing and expertise.

The first trend is the continuously growing amount of electronic data that is available to perform an analysis. A typical search on the World Wide Web (WWW) will demonstrate the problem of data overload – so many hits are returned that it is difficult to find the pertinent information (e.g., a recent search on "graduate school ohio state" returned over 6 million hits). Similarly, we have improved our data capture mechanisms to where we can create large "data warehouses" of information that no person has ever reviewed or catalogued. Every time that a credit card is used or a satellite photo is taken, the automatically captured data can be stored in a database. Although we have dramatically improved our ability to capture and distribute

electronic data in these and other ways, we have not correspondingly improved our ability to interpret the data and to determine what is significant.

The second trend which has contributed to the difficulty of dealing with data overload is the nature of changes in the distribution of work in various domains. One coping strategy employed by many organizations confronted with the data overload problem has been to adopt a "watch" style organization. With this strategy, practitioners are assigned to monitor a portion of the overall data field (a subsystem or subfunction), reporting relevant information to supervisors who integrate reports from focused individuals or sub-teams. By distributing the responsibility for different portions of the data field across individuals, and creating organizational mechanisms for coordinating these efforts, an extremely large amount of dynamic data can be effectively managed. However, this strategy relies on a large and continuous deployment of human expertise in order to function well. Economic pressures are forcing many organizations to redistribute human expertise in ways which minimize continuous human involvement during nominal operations and "down times." The tendency is to move towards an "on-call" model of human expertise utilization (Patterson & Woods, 1997). Under this type of operations model, high concentrations of human expertise are deployed only during critical or anomalous situations. While more efficient in terms of resource utilization, the thinning layer of human involvement has partially undermined the most general coping mechanism for data overload: human expertise and experience.

The goal of this research is to predict what vulnerabilities in inferential analysis may arise in the future as traditional strategies for coping with data overload become undermined by technological and organizational changes. With a richer understanding of how analysts may be vulnerable to generating low-quality analytic products in particularly challenging situations, we can proactively design and evaluate training, design, and organizational interventions that reduce these vulnerabilities.

In the intelligence analysis community, there is a growing concern about the potential vulnerabilities in analysis given current and future technological and organizational trends. It is predicted that there will be an increase in situations where analysts will be asked to provide analyses on short deadlines, known as Quick Reaction Tasks (QRTs), in areas in which they will not be as expert as they would like. In addition, there has been an explosion of available data, particularly "open source" data that greatly varies in reliability, and so it is clear that the analysts will be unable to read even a large portion of the accessible data on the area in the amount of time that they will have available.

Given this background and context of a domain in transition, the specific question that this research attempted to address was: *What are potential vulnerabilities in computer-supported inferential analysis under data overload for professional analysts working on a short deadline outside their immediate base of expertise?*

In order to address this question, field observations were conducted of ten professional intelligence analysts simulating the analysis of the causes and impacts of the failure of the maiden flight of the Ariane 501 rocket launcher. Most of the participants had some expertise that was peripherally related to this question, but were not experts on rocket launchers or satellites, or able to adequately answer the question from prior knowledge before gathering information. A customized set of approximately 2000 text reports from open sources such as Aviation Week and Space Technology were available for the study. A baseline set of support aids were provided that is similar in functionality to tools commonly in use by intelligence analysts. Specifically, the participants were trained how to conduct keyword searches, browse by dates and titles of reports returned by a query, and cut and paste information to a text editor in the software environment that was provided.

A set of protocols for each study participant was generated in order to iteratively identify patterns in the data in an exploratory fashion. By tracing and abstracting the processes used by each participant to perform the task, patterns in information sampling and in the sources of inaccurate statements in the verbal briefings across the participants were identified. These patterns provide a rich understanding of the potential sources and forms of vulnerabilities in inferential analysis under data overload conditions. This understanding points to a set of evaluation criteria for solutions designed to address these vulnerabilities.

In the remainder of this report, we will:

- define data overload more precisely and contrast this definition with alternative characterizations,
- describe how the problem of data overload is instantiated in intelligence analysis,
- describe the exploratory simulation study design and the process tracing methodology that was used to analyze the data,
- summarize and discuss the patterns in information sampling and sources of erroneous statements that were observed across study participants,
- discuss the study findings in relation to the conceptual frameworks of search strategies in information retrieval and abductive inference, and
- outline the implications of the study findings, including how the study findings point to a set of challenging evaluation criteria for proposed design, training, and organizational interventions aimed at addressing data overload.

2 DATA OVERLOAD: FINDING THE SIGNIFICANCE OF DATA IN A VAST DATA FIELD

Although everyone agrees that data overload is a commonplace problem that is extremely difficult to address, the precise definition of data overload is far from obvious. Common to most implicit definitions of data overload in supervisory control domains is the notion that somehow an excessive amount of data creates additional cognitive burdens for the human operator. Beyond that, however, the wide variety of design aids touted to "solve data overload" attest to the variability in opinions regarding the core problem of data overload.

There are three basic ways that the data overload problem has been characterized (Woods, Patterson, & Roth, 1998):

1. as a clutter problem where there is *too much data* on the screen: therefore, we can solve data overload by reducing the number of data units that are displayed,
2. as a *workload bottleneck* where there is too much to do in the time available: therefore, we can solve data overload by using automation and other technologies to perform activities for the user or to cooperate with the user during these activities, and
3. as a problem in *finding the significance of data* when it is not known a priori what data from a large data field will be informative: therefore, we can solve data overload by representing the data field in a way such that the significant data naturally emerges from the virtual perceptual field.

2.1 Clutter

> Clutter and confusion are failures of design, not attributes of information.
>
> Tufte, 1990, p. 51

The first way that people have characterized data overload is simply that there is too much stuff. This problem formulation led designers to try to reduce the available data. This approach arose in the early 1980's as a "solution" to the problem of "clutter" in the design of individual displays. The approach led developers to ask: how much is too much for people or what is the maximum rate of data people can process? Developers proposed guidelines for display design that limited the number of pixels that could be lit on the screen (given technological advances this measure of screen density is obsolete, but other ways to define what are too many screen elements can, and have been, proposed).

This has not proven to be a successful or fruitful direction in solving data overload and has faded in large part. This approach has faded because:

- it misrepresents the design problem – see for example Tufte (1990) and Zhang and Norman (1994); one specific thematic example is that reducing data elements on one display increases people's need to navigate across multiple displays (Woods & Watts, 1997),
- it is based on erroneous assumptions about how human perception and cognition work; for example, the questions about maximum human data processing rates are meaningless and misleading because among other things people re-represent problems, re-distribute cognitive work, and develop new strategies and expertise as they confront clutter and complexity,
- it is utterly incapable of dealing with the context sensitivity problem; in some contexts, some of what is removed will be the relevant data.

Systems that reduce or filter available data are brittle in the face of context sensitivity. First, some of usually unimportant data may turn out to be critically informative in a particular situation. For example, one nuclear power plant accident scenario is difficult precisely because the critical piece of data is usually unimportant (Roth, Woods, & Pople, 1992). Second, some data which seems minor now may turn out to be important later, after new events have changed the context. For example, in the recent Zaire civil war, one opposition figure, Kabila, emerged surprisingly as the leader of the rebel forces. Previous data about Kabila would have been considered minor, but it took on new significance after he emerged as a major figure in the events of 1997 (e.g., Kabila's ties to Ugandan and Rwandan leaders forged during their time as rebels is one key to Kabila's rise from obscurity).

It is striking that this "solution" to the clutter problem -- reducing the displayed data unless/until a user asks for more data, or only showing "important" data on a primary screen with less important data removed to secondary screens -- runs counter to the technological trends. If the main benefit of certain technologies is increased access to data, it is ironic that people have to throw away some of that access to cope with the complexity of trying to work with the available data.

A set of navigation techniques could be argued to make solutions based on the "clutter" view more workable. These techniques make it easy to navigate to data that is not in the pre-defined primary focus. For example, techniques such as cone trees (Robertson, Mackinlay, & Card, 1991) and the hyperbolic tree (Lamping, Rao, & Pirolli, 1995) make it easy to access data that is not shown on the screen through direct manipulation browsing. Similarly, "sliders" (Ahlberg, Williamson, & Schneiderman, 1992) have been designed to quickly and easily call up data that is "filtered" by a cutoff point on an ordered dimension.

Although these techniques make navigation much easier, partly addressing the first criticism of this approach, they do not address the other two criticisms: that this approach is based on erroneous assumptions about how human perception and cognition work, and that this approach is incapable of dealing with the context-sensitivity problem.

Fundamentally, approaches designed to solve "clutter" can only "nibble at the edges" of the data overload problem because generic data is still the fundamental unit of analysis, with the assumption that the data cannot be abstracted, organized, or re-represented in other formats. Solutions based on this characterization of data overload do not help the practitioner to recognize the significance of the data or direct the attention of the practitioner *before the practitioner knows where to look or what to look for.*

2.2 Workload Bottleneck

The second characterization of data overload has emerged in settings where access to data has grown quickly and explosively. In these contexts, such as intelligence analysis but also in Web based activities, participants use the words "data overload" in an everyday way that means they are experiencing what Human Factors professionals refer to as a workload bottleneck – there are simply too many individual data units to examine them all manually in the time that is available.

Workload is a potentially useful way to think about data overload as expressed in intelligence analysis-like situations. Previously, analysts were expected to read the vast majority of the reports that were available to them in order to provide a synthesized assessment and recommendations for action on a topic. With the workload characterization of data overload, analysts now express a need for an "agent" to help them with their activities. For example, machine agents could potentially prioritize or summarize reports for the analyst. Notice that with the workload characterization, solutions no longer are focusing on reducing data at the level of individual data units, but now they are focusing on making a person's cognitive activities more tractable.

An important distinction in aiding approaches to solve the workload bottleneck version of data overload is whether or not the approach requires a strong or weak commitment to the automation being "correct." Brittleness of machine processing, particularly in complex, high-consequence domains, is a serious issue in the design of cognitive systems (Smith, McCoy, & Layton, 1997; Roth, Bennett, & Woods, 1987). Approaches such as filters, summarizers, and automated search term selectors (e.g., Maes, 1998; Marx & Schmandt, 1996; Stone, Fishkin, & Bier, 1994; Quintana, 1998; Pratt & Sim, 1995; Chandrasekar & Srinivas, 1998; Cimino & Barnett, 1993; Lee, 1998; Salton, 1986; Srinivasan, 1996; Brann, Thurman, & Mitchell, 1996) are strongly committed to the machine processing being correct. Methods that are more weakly committed to machine pre-processing include using automation to index, cluster, organize, highlight, sort, and prioritize elements in a data field, (e.g., Oakes & Taylor, 1998, Letsche & Berry, 1997) and "cooperative machine agents" that notify, remind, or critique a human partner (e.g., Gruen, Sidner, Boettner, & Rich, 1999; Guerlain et al, 1999; Eckert, 1995, Carroll & Mckendree, 1987; Fischer, Lemke, Mastaglio, & Morch, 1991; Fischer & Reeves, 1992; Rubin, Jones, & Mitchell, 1988).

Although the workload characterization is a potentially useful way to think about data overload, the findings clearly show that automation support is necessary but not sufficient to create useful systems. Introducing autonomous machine agents changes the cooperative structure creating new roles, new knowledge requirements, new judgments, new demands for attention, and new coordinative activities. The automation must be directable and observable in order to avoid patterns of coordination breakdowns such as clumsy automation and automation surprises (Sarter, Woods, & Billings, 1997; Patterson, Woods, Sarter, & Watts-Perotti, 1998; also see Maes & Schneiderman, 1997 for a debate on interface agents vs. direct manipulation techniques that touch on some of these issues).

Similarly, shifting tasks from a human to a machine agent does not eliminate the fundamental difficulties in data overload. Just as we cannot usually eliminate "error" by allocating tasks to machines that humans have been observed to perform erroneously, similarly we cannot expect machine agents to consistently and correctly identify all of the data that is relevant and significant in a particular context in order to bring it to the attention of the human practitioner. All intelligent agent algorithms, from agents programmed by practitioners specifically to flag data items to agents that "learn" rules from observing practitioners, are unable to escape the need to take context into account. The irony here is that sometimes developers believe that shifting the task to a computer somehow makes the cognitive challenges of focusing in on the relevant subset disappear. In fact, all finite cognitive processors face this challenge, whether they are an individual, a machine agent, a human-machine ensemble, or a team of people. It always takes cognitive work to find the significance in data.

For example, attempts in the mid-80's to make machine diagnostic systems handle dynamic processes ran into a data overload problem (these diagnostic systems monitored the actual data stream from multiple sensors). The diagnostic agents deployed their full diagnostic reasoning power in pursuit of every change in the input data streams (see Woods, Pople, and Roth, 1990; Roth et al., 1992; 1992; Woods, 1994b). As a result, they immediately bogged down, dramatically failing to handle the massive amounts of data now available (previously, people mediated for the computer by selecting "significant" findings for the computer to process). To get the diagnostic systems to cope with data overload required creating a front end layer of processing that extracted, out of all of the changes, which events were "significant" findings that required initiating a line of diagnostic reasoning. In this case, determining what were significant events for diagnosis required determining what were unexpected changes (or an unexpected absence of a change) based on a model of what influences were thought to be acting on the underlying process.

2.3 The Significance of Data

> It is of the highest importance in the art of detection
> to be able to recognize, out of a number of facts,
> which are incidental and which are vital.
> Sherlock Holmes

A cognitive systems view of data overload can provide the human-centered component that is the core of the underlying framework for the study design and analysis. The starting point for this approach is recognizing that large amounts of potentially available data stress one kind of cognitive activity – focusing in on the relevant or interesting subset of data for the current problem context. When people are unable to assemble or integrate the relevant data, this cognitive activity has broken down.

People are a competence model for this cognitive activity because people are the *only* cognitive system that we know of that is able to focus in on interesting material in natural perceptual fields, *even though what is interesting depends on context* (Woods & Watts, 1997).

The ability to orient focal attention to "interesting" parts of the natural perceptual field is a fundamental competency of human perceptual systems (Rabbitt, 1984; Wolfe, 1992).

> "The ability to look, listen, smell, taste, or feel requires an animal capable of orienting its body so that its eyes, ears, nose, mouth, or hands can be directed toward objects and relevant stimulation from objects. Lack of orientation to the ground or to the medium surrounding one, or to the earth below and the sky above, means inability to direct perceptual exploration in an adequate way (Reed, 1988, p. 227 on Gibson and perceptual exploration in Gibson, 1966)."

Both visual search studies and reading comprehension studies show that people are highly skilled at directing attention to aspects of the perceptual field that are of high potential relevance given the properties of the data field and the expectations and interests of the observer. Reviewing visual search studies, Woods (1984) commented, "When observers scan a visual scene or display, they tend to look at 'informative' areas . . . informativeness, defined as *some relation between the viewer and scene*, is an important determinant of eye movement patterns" (p. 231, italics in original). Similarly, reviewing reading comprehension studies, Bower and Morrow (1990) wrote, "The principle . . . is that readers direct their attention to places where significant events are likely to occur. The significant events . . . are usually those that facilitate or block the goals and plans of the protagonist."

In the absence of this ability, for example in a newborn, as William James put it over a hundred years ago, "The baby assailed by eye, ear, nose, skin and entrails at once, feels it all as one great blooming, buzzing confusion" (James, 1890, I 488). The explosion in

available data and the limits of computer-based display have left us often in the position of that baby – seeing a "great blooming, buzzing confusion" in the virtual data fields that technology makes it so easy to create.

Designing a virtual environment implies the design of a virtual perceptual field on which human perception then operates. As designers, we need to identify what is significant and therefore needs to stand out in this perceptual field as worthy of the domain practitioner's attention in a particular scenario. The foundation of approaches designed to help practitioners find the significance in data is to use machine intelligence to better organize the data to help people extract meaning despite that fact that what is informative depends on context.

There is an important distinction in solutions to this characterization of data overload between syntactic, statistical clustering that applies across domains and model-based representation aiding for practitioners in a particular domain. Although the first approach is more generalizable, it "finesses" the context sensitivity problem by defining significance based on syntactic or statistical properties of text (e.g., word frequency counts) as cues to semantic content.

Calling this technique a finesse points to a contrast. In one sense, a finesse is a positive pragmatic adaptation to difficulty. By using machine processing on a massive data field to cluster documents, a practitioner is able to work on an ordered data field, able to look at portions of the data one at a time rather than sifting through the unmanageable mass. However, a finesse is a limited adaptation because it represents a workaround rather than directly addressing the factors that make it difficult for people to extract meaning from data.

The syntactic/statistical approach is relied on heavily in keyword search systems, Web search engines, and information visualization algorithms that utilize "similarity" metrics based on statistical properties of the text (e.g., frequency counts of different content words) to place documents in a visual space (e.g., Morse & Lewis, 1997; Wise et al., 1996; Biswas, Weinberg, & Fisher, 1998; Cox, Eick, & Wills, 1997; Eick, 1997; Keim, 1997; Keim & Kriegel, 1994; Keim & Kriegel, 1996; Pirolli, Schank, Hearst, & Diehl, 1996). The primary limitation of this approach is that syntactic and statistical properties of text provide a weak correlate to semantics and domain content. There is rarely a simple one to one relationship between terms and concepts. It is frequently the case that one term can have multiple meanings (e.g., Ariane is both a rocket launcher and a proper name; ESA stands for the European Space Agency, Environmental Services Association, and the Executive Suite Association) and that multiple terms can refer to the same concept. This can either be because of the use of synonyms to refer to the same concept (e.g., the terms 'failed,' 'exploded,' and 'was destroyed' can be used interchangeably) or because descriptions can be slanted by the perspectives of the report writers (e.g., the European Space Agency referred to a launch failure as "did not result in validation").

The problem is compounded by the fact that the 'relevance' metrics employed (e.g., the weighting schemes used by Web search engines) are often opaque to the user. This is the *lack of observability* catch. The user sees the list of documents retrieved based on the query and the relevance weighting generated by the search engine. However, in many cases, how the relevance weighting was generated is unclear, and the resulting document ordering does not accord well with how the user would have prioritized the documents (i.e., documents that come up early with a high weighting can be less relevant than documents that come up later.) This forces the user to resort to attempting to browse through the entire list. Since the generated list is often prohibitively long, it can leave the user unsure about whether important documents might be missed. We have observed that users will often prefer to browse documents ordered by metrics that do not attempt or claim to capture "relevance," such as date or source, rather than by syntactic relevance weighting because the organizing principle is observable and they know how to interpret values along those dimensions.

Attempts to place documents in a visual space based on syntactic properties are also subject to the *over-interpretation* catch. The spatial cues and relationships that are visible to the observer will be interpreted as meaningful even if they are incidental and not intended to be information bearing by the designer. For example, visualizations that attempt to represent multi-dimensional spaces (four or more dimensions) on a two dimensional display can create ambiguities with respect to the position of a document relative to each of the dimensions. Users may assume that two documents that are located close to each other on the display reflect a similar degree of relationship to each of the dimensions represented in the space, when in fact they are not in the same position in the multi-dimensional space – even though it looks that way on the display. Similarly, information visualizations that attempt to reveal thematic relationships between documents through visual patterns are subject to over-interpretation. The visualizations can be dominated by patterns that are unimportant, such as missing data, and the underlying relationships may be distorted in the mapping to the perceptual field.

The model-based representation aiding approach, on the other hand, trades off generalizability of the technique for an increased ability to identify and take advantage of the semantics of the underlying processes or field of activity in order to define the relationships that give data meaning (Vicente & Rasmussen, 1992; Doyle, Charest, Falcone, & Kandt, 1990; Wright, 1995; Vicente, 1996; Potter, Woods, Hill, Boyer, & Morris, 1992). Related techniques would develop expectation-based displays that highlight when events depart from expected or typical behavior and event based displays that capture the flow of events in the world at different levels of abstraction or in comparison to the expected flow of events (Potter & Woods, 1991).

Such techniques become the basis for developing pattern based displays and conceptual spaces that support people's abilities to explore spatially structured environments and recognize patterns across elements. For many kinds of data overload problems there

will be multiple organizing themes each of which defines a perspective on the field of data. Mechanisms to help users coordinate across a set of these perspectives will be needed.

The model-based representation aiding approach also has "catches" associated with it. New representations are subject to the *catch of custom innovation* – each is a unique creation tailored to a specific setting. Model-based methods to depict more than the base data are subject to an *uncertainty catch* – given high uncertainty in the data and significant consequences in possible outcomes, experts tend to revert to raw data, and the *"right" model catch* – how do you know the model that specifies how data is informative is appropriate for the task or situation? Expectation based displays are limited by the fact that it can be difficult to track/compute expectations about a process or about another agent.

One important outcome of this research is additional insight for the design of model-based representation aiding "solutions" (Woods, 1995) to data overload for intelligence analysts. In order to do this, there will need to be ways to use and improve the power of technology to:

- enhance observability,
- take into account context sensitivity, and
- build conceptual spaces.

3 DESCRIPTION OF THE DOMAIN: INTELLIGENCE ANALYSIS

Intelligence analysis in the United States, like many other complex domains, is experiencing the two trends described in Part I that are exacerbating the data overload problem. First, there is an explosion in the amount of data available to intelligence analysts. On an average day, an analyst will receive hundreds of text "messages" through electronic mail that are selected by keyword lists from an overwhelming amount of available information that is generated by such agencies as the National Security Agency. These messages are designed to update analysts on topics related to particular technologies and countries. In addition to these lists, there are massive databases that are accessed when an analysis task arises that is outside the range of their personal databases (that were generated from organizing the incoming messages). For example, an analyst described in an interview that he was asked to "Tell me everything you know about the command and control structures in Somalia in the next 24 hours." Since no analyst had been tracking Somalia, he performed keyword searches in an on-site database that generated 42,000 documents. Theoretically, he could also have searched other databases, such as Lexus Nexus™ and classified and unclassified sites on the World Wide Web. He estimated based on previous experience with a similar task that he could only scan 15,000 messages in a day, making it impossible to "brute force" read them all in the allotted time.

Second, intelligence agencies are undergoing an organizational redistribution of assignments. Resulting from a shift in emphasis from the Cold War paradigm of monitoring a small number of countries for their ability to directly attack the United States to monitoring many more countries for a more diverse set of reasons (e.g., peacekeeping and humanitarian interventions), analysts are now being asked to cover a larger set of countries and technologies. At the same time, there have been large reductions in both staffing and average years of experience. The net result is that intelligence analysts are increasingly asked to analyze situations that are outside their immediate base of expertise on shorter time horizons.

The domain of intelligence analysis could be viewed as a supervisory control process, where the practitioners are always trying to define and focus in on the relevant set of data for a task under data overload conditions. Intelligence analysts monitor situations in other countries in order to identify when and how to intervene. Many of the challenges in more heavily studied supervisory control settings, such as decision making in a dynamic situation under uncertainty, are present in intelligence analysis. However, intelligence analysis in some ways is different than supervisory control in more heavily studied worlds such as aviation flight decks and mission control. There are several factors that complicate the practitioner's cognitive activities in intelligence analysis. These include:

- the kind of processes being monitored,
- the nature of the data available about the state of those processes, and
- the capabilities of the tools available to support analysis.

3.1 Monitoring Human/Organizational Processes

Cognitive engineering studies and designs generally have been targeted at practitioners who monitor and control engineered or sometimes physiological processes. In intelligence analysis, the underlying process that is monitored (what we refer to as the monitored process) is sometimes a technical process (e.g., communications network technology or the technology of specific weapons systems), but often consists of various kinds of human/organizational processes. For example, in analyzing events in one region of the world, an analyst may need to understand current and past ethnic group processes, alternative kinds of political processes, such as those of a theocracy, economic processes, geopolitical processes, and the development and implementation of military doctrine, to name just a few.

Adding such human/organizational processes to the mix leads us to consider the differences between different kinds of monitored processes. Figure 1 indicates that monitored processes can be loosely ordered on a dimension that describes how "definitive" we can be both in understanding and in modeling how the processes work. Figure 1 illustrates this dimension by ordering three classes of monitored processes: engineered, physiological, and human/organizational processes.

Engineered processes are physical systems that are designed and implemented by people, and are exemplified by such systems as the space shuttle, nuclear power plants, and military and commercial aircraft. These processes obey well-understood physical laws. Physiological processes are self-tuning processes that exist naturally in the environment but can be altered by human intervention, as is the case in cardiovascular systems during open heart surgery. Human/organizational processes involve situations or activities in which groups of people interact, such as situations of low-intensity regional conflicts or activities involving supply logistics, economic behavior, or development and application of military doctrine. These processes may be defined or described by sets of rules, but these rules provide only a partial description of the actual behavior of people or organizations (e.g., for various reasons a military unit may deploy in a way inconsistent with standard doctrine).

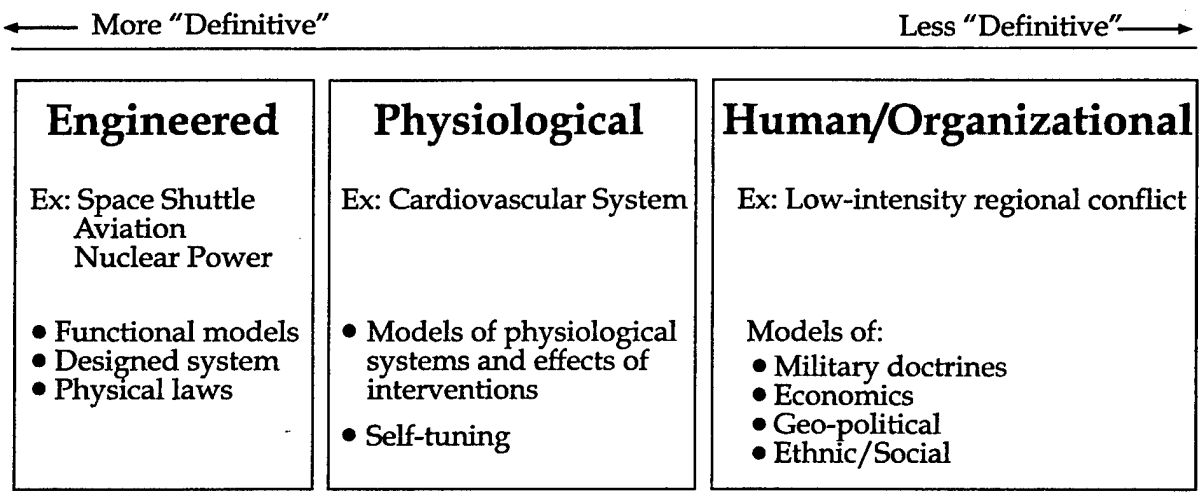


Figure 1. Different kinds of monitored processes can be ordered on a dimension of how "definitive" we can be in understanding, modeling, and predicting how that process works.

Highly "definitive" models, such as models of physical systems that were designed by people to accomplish certain goals, provide comparatively strong analytical frameworks because their component parts obey and are constrained by physical laws (e.g., heat exchangers always work a certain way functionally). Note that for all monitored processes, uncertainty and variability exist, but that the degree of uncertainty and variability changes as we move from less to more "definitive."

Many kinds of monitored processes can be relatively well-modeled at a functional level but are complex enough that many situations arise that are not predicted in advance. For example, regarding physiological systems, we know a great deal about the laws that govern such processes. However, we find that

- the models of physiological systems are not as detailed and accurate as those of the typical engineered process,
- the individual differences in physiological systems are larger between people than they are within analogous components of an engineered process (such as the variations found within examples of a particular model of aircraft),
- physiological processes have built in interactions and self-tuning control loops that are difficult to model completely.

In intelligence analysis, the models that are available to analysts are less "definitive" than the models available in engineered and physiological processes. Rather than a functional model, the frameworks available to analysts tend to be collections of heuristics and knowledge, such as how the military doctrine of a particular country's armed forces would influence behavior in a particular situation. These "models" are

inherently less precise and support weaker predictions about actual behavior in specific situations. Yet these models are still very important, because the skilled use and application of these models is what is responsible for the recognizable differences in performance between more and less experienced analysts.

An additional complication in modeling human/organizational processes is that the division is less clear-cut between the "supervisory controller" and the "monitored process" given that the processes being monitored by intelligence analysts involve people. In engineered processes, for example, people are clearly outside of the processes to be monitored. Even in engineered processes, the roles of different people in the operational system can become quite complex in terms of scope of authority, supervisory control, and field of view. However, in discussing engineered processes, usually the confusion we try to guard against is ambiguity about the different roles different machines can play. The monitored process is technological, but we also now create machines that help us observe, evaluate, diagnose, and act on the monitored process. These support systems and automation are usually better seen as a part of the operational team along with the human monitors and supervisors (Billings, 1996). Similarly, with physiological processes the role of technology can be ambiguous: is it part of the process, (e.g., a programmable pacemaker), or is it part of the treatment team, (e.g., an infusion device)? But another potential complication emerges with physiological processes since the people are both the process being controlled and the controllers, e.g., the patient (the physiological processes in question) can be part of the treatment process (see the case in Obradovich & Woods, 1996).

In the case of human/organizational processes, people, groups of people or human organizations are active in every role. In an attempt to reduce the potential for confusion, Figure 2 provides a very rough schematic of the interacting roles when the monitored process is human/organizational. The figure contains three global roles (represented as the columns):

1. People in other parts of the world in various roles as part of economic, political, religious, ethnic and military processes.
2. People in U.S. organizations in various roles as monitors of those processes (intelligence analysts) and as policy makers who decide about U.S. policies and actions in response to events in those parts of the world.
3. Investigators who try to understand the role of intelligence analysts and help shape new supporting tools to cope with issues like the potential for data overload.

The figure is tremendously oversimplified. There are other groups (e.g., humanitarian) and governments monitoring events in one part of the world that influence or shape the interactions. Governments may be watching and predicting how their people will behave (e.g., polls) or how different subgroups (e.g., constituencies) will react to different events, while outsiders may be monitoring how one group is anticipating how other groups will behave.

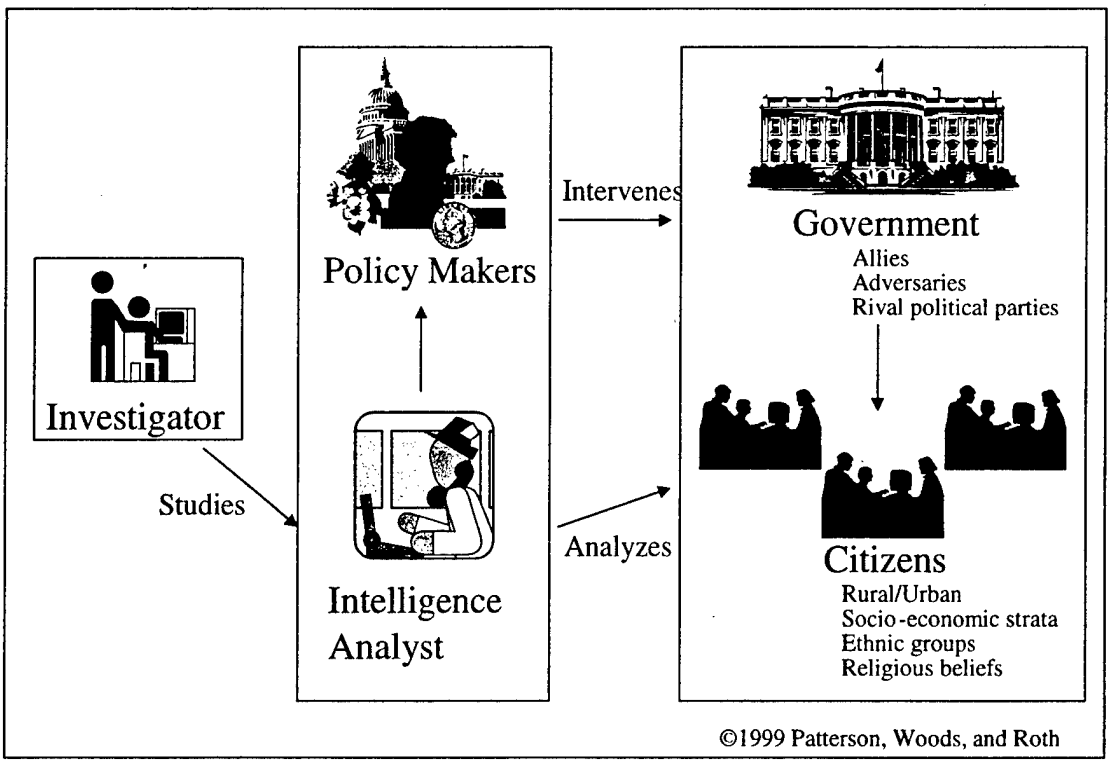


Figure 2. Analysts monitoring a human/organizational process.

3.2 The Nature of Data Available to Intelligence Analysts

In engineered, physiological, and human/organizational processes, data is captured through "sensors." The nature of the data that is available is dependent on how the sensor information is processed, packaged, and displayed. In engineered and physiological processes, there are physical sensors placed at various points that monitor certain variables continuously. In general, the sensors always monitor the same thing in the same way and are displayed as single parameter sensor readings in dedicated locations (although there has been movement away from this one-sensor, one display organization with displays that integrate parameter values based on functional models of how the process works, Vicente and Rasmussen, 1992). In engineered processes, it is possible, though complicated, to define "nominal ranges" and signal an alarm when a parameter goes out of the nominal range. Note, however, that even in engineered processes, it is common to have nuisance alarms that indicate a setpoint crossing that is not abnormal given an ongoing context. In physiological processes, it is also possible to point to possible limit values, but they function much more as landmarks or very general guidance because "significant" values depend so much on the patient and context. For example, what is too much or too little of some parameter for an individual may vary tremendously based on the stage of the surgical procedure, previous disease history, or relative to a baseline established for that particular individual just prior to

surgery. In intelligence analysis the situation is even more difficult. It is not easy to flag abnormal data; indeed that may be part of the analysis process itself. It is often contentious what is an abnormal state, and even when it is not, there are currently no systems that can reliably recognize and flag textual descriptions of abnormal states.

When monitoring less definitive human/organizational processes, the "sensors" are more diffuse, with data about the process gathered remotely, indirectly, or by human observers on the scene. In human/organizational processes, when humans serve as the "sensors," the situation is actually better, in a sense. People can use their intelligence in terms of what variables to sample and what format is best to use to describe their observations. On the other hand, the data becomes more difficult to find and interpret because there is less consistency about what is sampled, how it is sampled, and where the information is displayed. In addition, there is the qualitative difference created by the fact that human/organizational processes are intentional systems. They can realize that they are being monitored and change their behavior or actively attempt to deceive observers. The observational sub-processes may, in fact, be specifically targeted for destruction, disruption, degradation, or denial.

Note that sensor data is not the only form of data available in any of these processes. Direct observation of the process, either by the supervisory controller or other agents in the distributed system, plays a role. In engineered processes, for example, controllers can directly touch a pipe to determine if it is hot. In physiological processes, anesthesiologists can look directly at the surgical field or check the color of the skin (e.g., if one notices the patient turning blue, then it is clear something is preventing adequate oxygenation of tissues). In intelligence analysis, agents can directly perceive information from satellite pictures or receive reports from agents who are dispersed to the area of interest to opportunistically perceive and report information.

In all of these domains, the reliability of the data is a critical concern. Physical sensors in engineered and physiological processes are uncertain indicators because they are placed in only a few locations; they are, in fact, model-based: the parameter of interest is often measured indirectly through other more tractable data, and they can fail. Data that is obtained through direct perception could also be unreliable: the observation relies upon the expertise and perceptual ability of the observer to identify subtle cues. In intelligence analysis, data comes in the form of reports created by humans who serve as the "sensors." The reports integrate a selection of data based on an interpretation and therefore need to be "unpacked" in order to identify the elemental data, which is used to generate an analysis product with a potentially different interpretation frame. People may bring a new set of reporting biases that create new forms of uncertainty. In addition, the difference between a normal state and an abnormal observation is contentious, and there is the added complication that the adversary in human/organizational processes may deliberately attempt to deceive the "sensors." As a result of the potential for unreliable data, similar strategies are observed in all of these different domains where data is cross-checked from independent sources in order to determine if the sensor is providing "invalid" data.

In intelligence analysis, data conflicts can be more subtle than in other domains due to the nature of the data. With engineered and physiological sensor data, there are concerns about effects being masked and sensors failing, but often the practitioners have the ability to check sensors on similar systems that are measuring the same information and see if they agree. Intelligence analysts also employ a variation of this strategy, but it is more difficult to determine if information agrees because it is not known how or when the information was "measured" and the content of the information itself is not identical. Analysts need to break down textual reports to a more elemental data level and then interpret the reports in order to determine the relationship of the data elements. When two or more independent sources give the same description of the same event, the information is more likely to be accurate. Schum (1994) refers to this as corroborative redundancy. When two or more sources provide information that inferentially favor the same hypothesis, this is referred to as convergent evidence which makes a particular hypothesis more likely. If information from two or more sources appear to corroborate or converge but stem from the same information source (e.g., a press release), there is no inferential value.

Conversely, items can be conflicting by saying logically opposing things or favoring different hypotheses. When information is discrepant, judgments of source quality are often important to decide what information to incorporate in the analysis. Factors that are considered in the credibility of a source include competency of the source to understand the issue at hand (e.g., Financial Times is a good source for financial information), predictable biases (e.g., self-reports by individuals or companies tend to be overly optimistic and less judgmental), and even attempts to actively deceive in the past (e.g., reports from countries with controlled media such as China might be publishing inaccurate accounts for political reasons).

Nevertheless, global judgments of source quality, such as "X is a trustworthy source," are under-specified, oversimplifications of how variations across sources play a role in the analysis process. Although certain sources are weighted as more credible than others based on past experience with a source, these judgments need to be tempered by other cues. Reports that are published immediately after the occurrence of an event are missing details that are provided in later updates – in other words, these reports contain "stale" information in relation to later reports. In addition, reports that are "distanced" from the original data are suspect. Having direct access to eyewitnesses, recorded data such as video, and telemetry data improves the quality of the analysis. Similarly, having direct access to people who have interpreted the data in depth, such as the inquiry board after an accident investigation, is important. Reports of other reports suffer from the problems evidenced in the game "Telephone," where the story changes with each telling. This is exacerbated when the reports are translated from foreign languages. Finally, reports that are making predictions about future events are inherently uncertain, regardless of the competency of the person providing the prediction.

3.3 The Nature of Tools Available to Intelligence Analysts

As previously described and as we have observed in other domains, ongoing technological and organizational changes are fundamentally changing the task of intelligence analysis. As a result of data being available more in electronic media, shorter timelines, and a broader range of analytical responsibilities, it is becoming increasingly difficult, if not impossible, for analysts to read all of the potentially pertinent individual messages and potentially relevant reports in a specific problem context. In this new situation, the analyst now needs to search through an electronic data-base/document-base in order to identify relevant information. This is the "new world of data" that has begun to emerge for analysts, and therefore the nature of the tools available to intelligence analysts need to be somewhat different in nature than tools designed for real-time monitoring of sensor data in engineered and physiological processes.

The main complication introduced by this new situation is the relationship between events in the world, database(s) of electronic information about events in the world, and sampled information about events in the world (Figure 3). The intelligence analyst rarely directly observes events in the world. Rather, other humans generate reports about events in the world. These reports make up a set of databases whose characteristics are often opaque to the analyst, particularly since the available information is constantly being updated and the information is generally not indexed. Information is "sampled" from these databases, first by keyword search queries and then by browsing dates and titles through the computer "keyhole," a small CRT screen (Woods & Watts, 1997). The relationship of the sample to the database is generally not available to the analyst (although some ways to characterize the database are being developed that could be used to determine the relationship, e.g., Wise et al., 1996). How does an analyst know if (s)he has read all of the available relevant information or if the information that is retrieved by a keyword search is high quality in comparison to what is available? How does (s)he know what information in the database is contradicted or corroborated by other information in the database? How does s(he) know if the information is significant in the context of other world or related events?

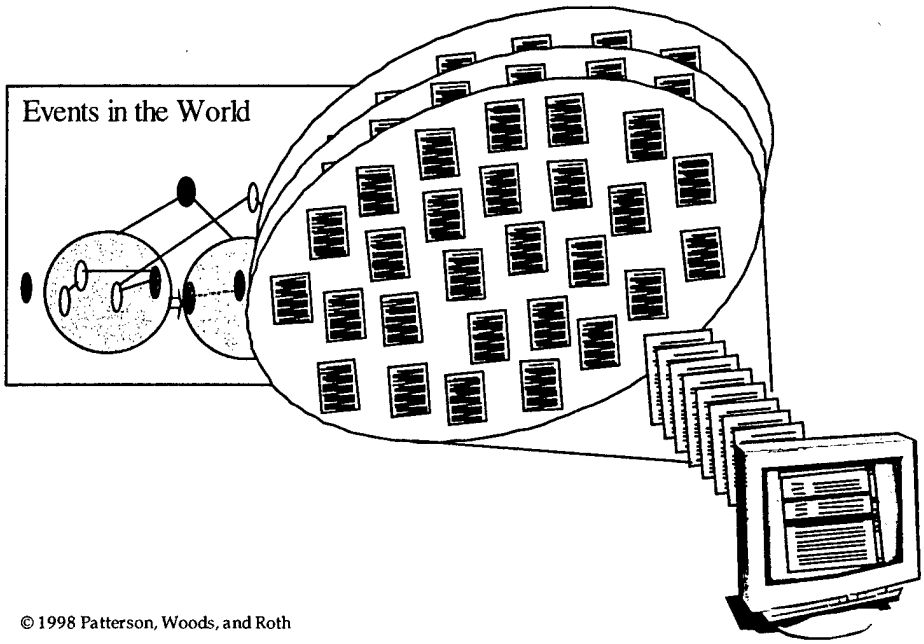


Figure 3. The analyst's new world as information sampling through a computer "keyhole."

Another complicating factor in the search for information is that the report is not an elemental data unit. Intelligence analysts do not make judgments of how information is related at the level of the report. Instead, those judgments occur about selected descriptions taken from reports. The search and retrieval tools available to analysts return "bundles" at the report level, not at the level of selected descriptions within reports. There is no easy way for analysts to search for information that will corroborate a selected description at that level. Analysts would need to look for the selected description in all of the returned reports manually because the date and title information is unlikely to provide clues about the information at the level of a selected description. This process makes it particularly difficult for analysts to know when information about a topic has been updated or changed without reading all of the available documents.

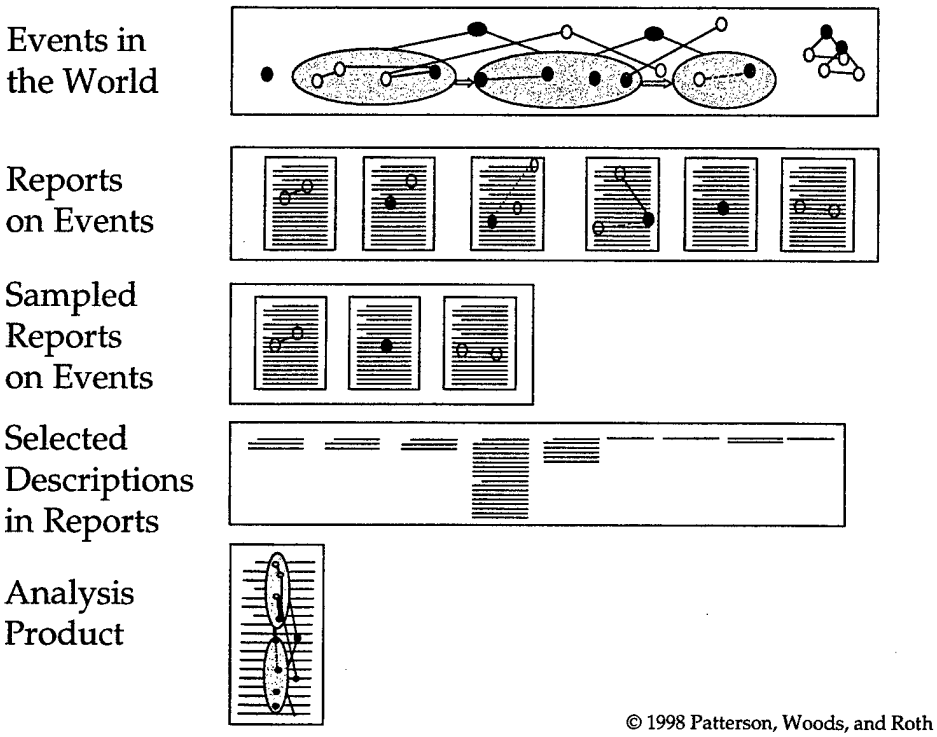


Figure 4. Sequence of information "bundles" in the analytical process.

Figure 4 gives an abstract view of how data is manipulated during the analytical process. Meaningful events occurring in the world are represented as textual descriptions within reports. These reports partially overlap and are distorted by the interpretation of the reporters on what the event was in relation to past, present, and future contexts. An analyst samples a subset of the available reports using keyword search and browsing mechanisms. The analyst then must break down the report into smaller units in order to compare whether related descriptions in different reports are corroborating or discrepant. The corroborated descriptions are then incorporated into a coherent story, the analytic product, which highlights what is significant about the analyzed situation based on an interpretive frame provided by the analyst.

To summarize, intelligence analysis is a domain that has a particularly difficult version of the data overload problem. The exacerbating trends of increased data availability and expanded monitoring responsibilities are transforming the nature of the cognitive task. In addition, there are some features of the intelligence analysis domain that make inferential analysis particularly difficult:

- the humans and organizations are often aware of being monitored and so sometimes compensate or actively engage in deception,
- it is difficult for the analysts to know where to look for informative data in the data field because, unlike in engineered domains, sensor information is not in a dedicated location,
- information must be sampled from data fields, often in ways that leave the relationship of the sample to the available data opaque,
- the task is not direct analysis from first-hand data but rather integration and corroboration of second-hand interpretations of data from multiple sources,
- the models that the analysts use to interpret and find the significance of data are less definitive than in engineered or physiological domains,
- it is controversial if not impossible to define and alarm "normal" operating ranges for variables in the models,
- data comes in the form of text "bundles" in interpretive frames that need to be broken down into more elemental data units, and
- it is difficult to determine if two descriptions of the same event are corroborating or conflicting because of the subtlety with which these differences are represented.

4 SIMULATION STUDY

4.1 Study Methodology Overview

Given that the goal was to identify potential vulnerabilities in computer-supported inferential analysis under data overload and on a short deadline for professional analysts being tasked with an analysis outside their immediate base of expertise, the study was designed to map onto this target as much as possible. The test situation was:

- to analyze a face valid task that had not previously been analyzed and was not in the immediate base of expertise for the participant: the cause and impacts of the June 4, 1996 Ariane 501 rocket launch failure on the Ariane 5 rocket's maiden flight,
- 2000 text documents in a mostly "on topic" database generated by representative searches in Lexus Nexus™ and DIALOG™ by the investigators and a professional search intermediary,
- 3-4 hours to complete the simulated task,
- a "baseline" toolset that supported keyword queries, browsing articles by dates and titles sorted by relevance or date, and cutting and pasting selected portions of documents to a text editor, and
- ten experienced intelligence analysts, one from each major division of the participating intelligence agency.

The data from this study was iteratively analyzed using a process tracing methodology (Woods, 1993). First, a set of detailed protocols of the analysis process were constructed for each participant. These processes were then represented abstractly based on different conceptual emphases in order to identify patterns across the participants.

4.2 Ariane 501 Rocket Launch Failure Scenario

The Ariane 5 rocket is a new European rocket design by Arianespace that is intended to eventually replace the current Ariane 4 rocket series. The Ariane 5 rocket design is larger, can carry two payloads rather than one, and is more powerful in general. The maiden launch on June 4, 1996, of the Ariane 5 vehicle ended in a complete loss of the rocket booster and scientific payload that it was carrying due to an explosion approximately 30 seconds after liftoff.

The Ariane 501 scenario was selected because it was a complex accident that captured many important aspects in the problem of finding the significance of data when it is not known a priori what data from a large data field will be informative. The significance of the Ariane 501 incident lies in how it was a departure from typical launch failures. First, the explosion was due to a design problem in the software rather than the more classic mechanical failure – there was numerical overflow in an unprotected horizontal velocity variable in the embedded software that was re-used from the Ariane 4, which is a slower rocket. Additionally, it was the first launch of a new rocket design, which

raised concern about the viability of the new design. Overall, however, launch failures were relatively common in the industry (the launch failure timeline in Figure 7 lists some of the launch failures that occurred about the time of the Ariane 501 failure) and first launches in particular were prone to fail, so the reputation of the Ariane program was not greatly damaged.

The main sequence of events during liftoff was basically undisputed and were detailed in the official Inquiry Board Report that was released on July 19, 1996 (see the launch event timeline in Figure 5). Approximately 30 seconds after liftoff, there was numerical overflow in an unprotected horizontal velocity variable in the embedded software in both the primary and backup inertial guidance systems. The horizontal velocity number overflowed because the embedded software was re-used from the Ariane 4, which is a slower rocket and could not reach the (three digit) velocities reached by the Ariane 5. Both inertial guidance systems shut down nearly simultaneously as a result of the numerical overflow and went into a diagnostic mode. The assumption during the design process was that only one of the inertial guidance systems would fail at one time, as in typical mechanical failure scenarios, so the other inertial system would theoretically be used to give guidance information while diagnostic data from the faulty system would allow ground personnel to troubleshoot the anomaly. When both systems shut down for the same reason (i.e., a common mode failure occurred), there was no guidance data available to steer the rocket. The diagnostic data was interpreted as guidance data and an abrupt change of course was ordered, initiating the rocket's automatic self-destruct sequence when the rocket began to break up from the strain. Six seconds after the rocket self-destructed, range safety officers on the ground also ordered a destruct command.

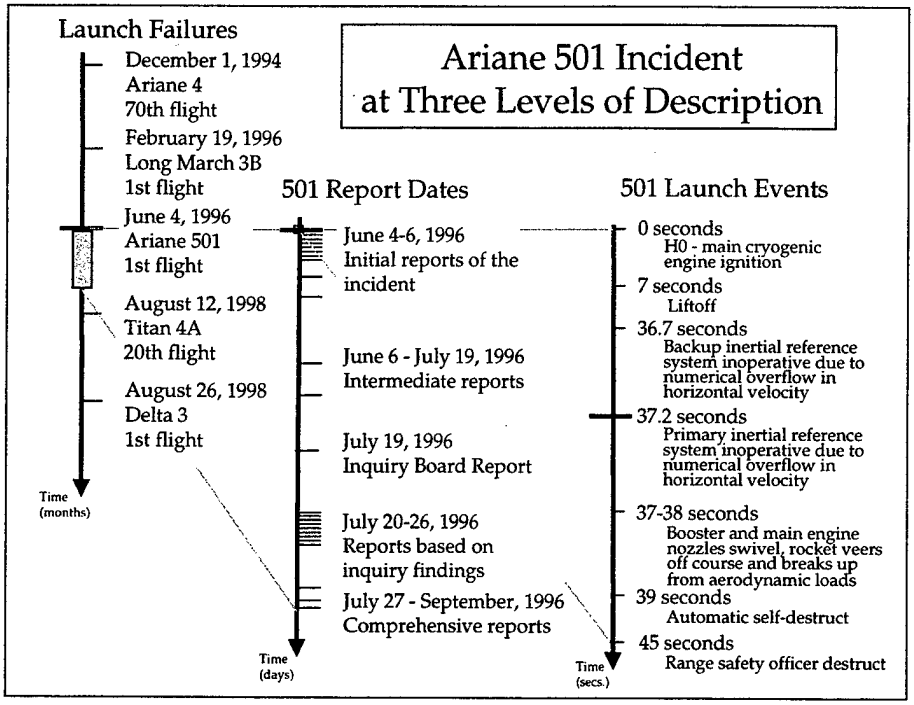


Figure 5. Ariane 501 incident at three timescales.

As is commonly observed following a large-scale failure, a flurry of reports were produced immediately following the accident (see the report date timeline in Figure 5). Generally, reports immediately following the event were detailed, but contained many inaccuracies and were missing information that only became available later. The release of the official Inquiry Board Report released by the European Space Agency six weeks after the accident also generated a flurry of reports. Several of the best reports about the Ariane 501 scenario were written in the months following the release of the Inquiry Board Report that gave a comprehensive picture of why the failure occurred and what the impacts of the failure might be. Even at this time period, however, some documents misreported the technical details of the cause of the failure and appeared overly biased in their assessment of the impacts of the accident.

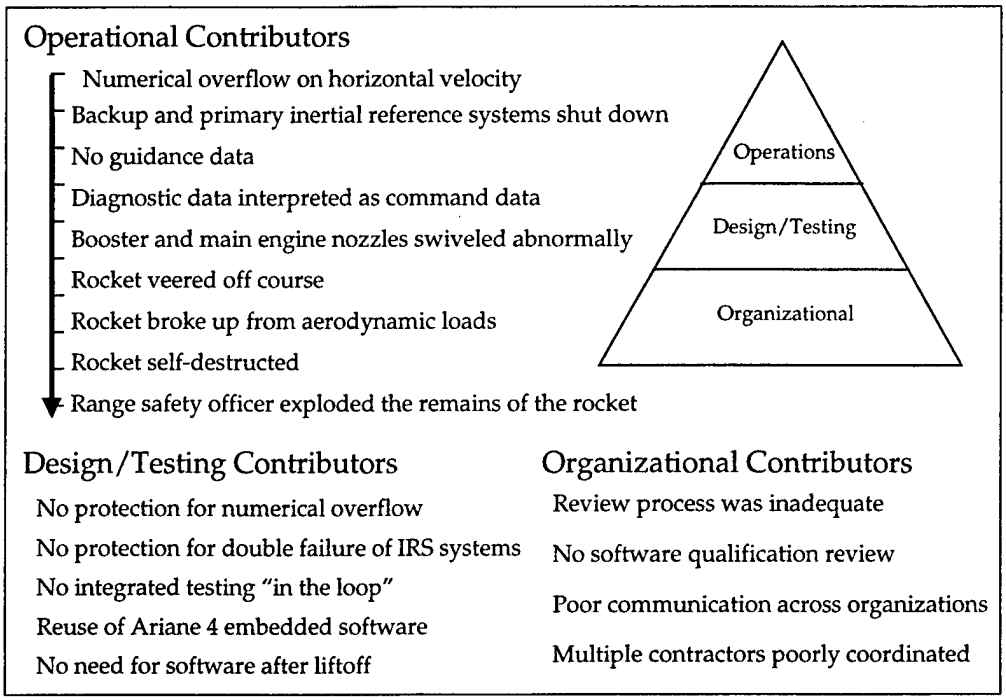


Figure 6. Contributors to the Ariane 501 failure.

One potential description of the cause of the Ariane 501 failure is that the rocket blew itself up. This is a proximal description in that it was very close to the time of the accident. The operational contributors could be backed up along the sequence of events in Figure 6 to the distal operational factor that there was a numerical overflow in the embedded software on a horizontal velocity variable.

As with any accident, there were contributors to the failure that were more removed from the proximal events. Protection against this software problem could have been built into the design of the system, so there are design as well as operational contributors to the accident. First, the software that overflowed was not needed after liftoff. Good design practice suggests that software code should only run when it might be needed to protect against unanticipated interactions between subprograms. Second, the embedded software was reused from the Ariane 4 rocket without carefully checking for assumptions that might be violated with the transfer to the new rocket design – which is a classic failure pattern in the design process of missing side effects associated with a design change. Third, there was no protection against numerical overflow on the horizontal velocity variable even though other variables were protected. Fourth, the designers did not design the rocket to be robust to a failure in both inertial guidance systems. Theoretically, the system could have given calculated guidance information if no guidance data was available. Finally, the inertial systems were never tested "in the loop" before liftoff, which would have caught the software error.

In addition, there are factors that could be viewed as organizational contributors to the launch failure. The review process was inadequate for ensuring that the testing could detect flaws in the design. There was no software qualification review comparable to reviews for mechanical subsystems. There were multiple contractors who worked on the rocket design who were poorly coordinated, and communication across organizations in general was poor.

4.3 Data Set Provided to the Study Participants

The electronic database was constructed by a set of representative queries (e.g., "Ariane") in Lexus Nexus™ by the investigators and in DIALOG™ by a professional search intermediary. The data set that was provided to the participants contained enough information to provide a detailed answer to why the incident occurred and what the short- and long-term impacts might be. There were approximately 2000 text documents from open source literature, such as Aviation Week articles. The majority (~ 60%) of the documents were "on target" in that they contained information that could input to the answer on when it occurred, why, and what the impacts were. Some of the documents (~ 35%) contained information that helped to provide context, such as information about other rocket launch failures, but were not directly relevant to the specific question. Only a small portion contained completely irrelevant information (~ 5%), such as articles about women named Ariane. Nine documents in the database were classified as "high profit" documents by the investigators (Table 1). The high profit categorization is based on both high topicality and utility, which are often combined in relevance definitions in the information retrieval literature (see Mizarro, 1997, for an overview of the factors in relevance definitions; cf. Blair and Maron, 1985, for their distinctions between "vital, relevant, partially relevant, and not relevant" documents in legal analysis). High profit documents were detailed accounts from credible sources that were published some time after the release of the official Inquiry Board Report from the European Space Agency.

Table 1. High Profit Documents

Title	Source (Date)
Ariane 5 Flight 501 Failure: Report by the Inquiry Board	Inquiry Board (July 19, 1996)
Inertial Reference Software Error Blamed for Ariane 5 Failure	Defense Daily (July 24, 1996)
Software Design Flaw Destroyed Ariane 5; next flight in 1997	Aerospace Daily (July 24, 1996)
Ariane 5 Rocket Faces More Delay	The Financial Times Limited (July 24, 1996)
Flying Blind: Inadequate Testing led to the Software Breakdown that Doomed Ariane 5	The Financial Times Limited (July 25, 1996)
Board Faults Ariane 5 Software	Aviation Week and Space Technology (July 29, 1996)
Ariane 5 Explosion Caused by Faulty Software	Satellite News (August 5, 1996)
Ariane 5 Report Details Software Design Errors	Aviation Week and Space Technology (September 9, 1996)
Ariane 5 Loss Avoidable with Complete Testing	Aviation Week and Space Technology (September 16, 1996)

As can be seen in Figure 7, there were naturally occurring discrepancies in the descriptions of the incident in the database (the boxed items had discrepant descriptions). For example, reports on when the rocket blew up ranged from 30 to 66 seconds after lift-off. The discrepancies in the account of the Ariane 501 incident came from several sources that would be expected with any complex, event-driven domain with textual data. First, all reports immediately following the event had some inaccurate or misleading information because data and clarifications were still coming in. For example, it was reported that the rocket was blown up from ground controllers when it really had self-destructed because the first reports were based on seeing the ground controller push the destruct button, although by then the rocket had already self-destructed. Other reports had inaccuracies due to translation from a foreign language, secondary reporting, or a lack of technical expertise. For example, the cause of the swerving of the engine nozzles was described in one report as resulting from a reset of the inertial reference frame and therefore sending inaccurate guidance data midway through the flight as opposed to shutting down and changing to a diagnostic mode where no guidance information at all was provided.

In Figure 8, it is apparent that many of the conflicting descriptions relating to the impacts of the incident are due to the difficulties in predicting future events. Predictions of future events got more consistent the closer the predictions were to the actual events, but most predictions were different from the actual event. For example, the original predictions of the delay to the second flight in the Ariane 5 program (502) were often far too optimistic, and gradually became closer to the actual date near when the event occurred. In some cases, information coming in completely overturned previous predictions. For example, immediately following the 501 incident, it was decided to cancel the Cluster scientific program because of the loss of the $500 million Cluster satellites in the explosion. Later on, one of the four satellites was decided to be replaced, and even later, all four of the satellites were rebuilt.

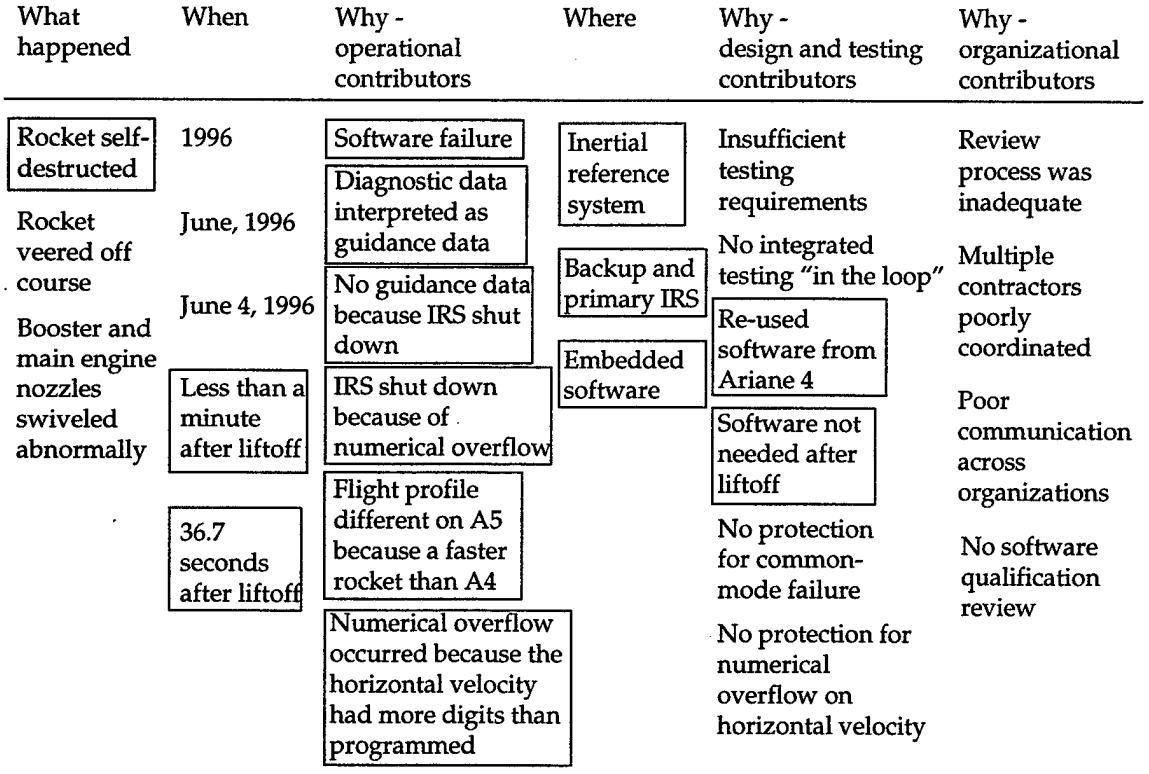


Figure 7. Discrepancies in the causes of the Ariane 501 failure.

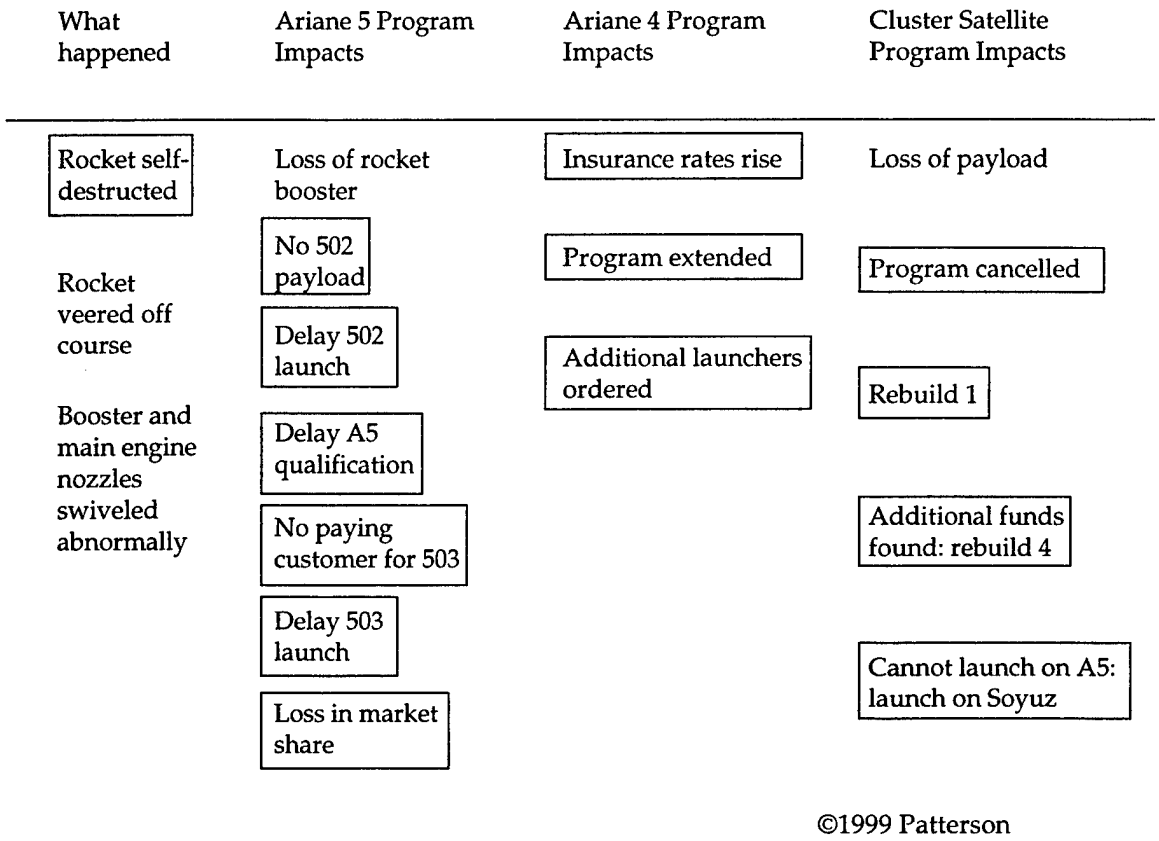


Figure 8. Discrepancies in the impacts of the Ariane 501 failure.

4.4 Overview of Study Session

Each study participant was scheduled for 3-4 hours. The purpose of the study was described to the participants as a means to better understand in detail the processes that experienced intelligence analysts use when performing an analysis by observing them performing a simulated task. The participants were asked for their written consent to audio and video-record the session in order to facilitate data analysis. Following that, the analysts were asked questions intended to better gauge the probability that they would have significant background knowledge related to the scenario and to collect demographic data (Table 2).

Table 2. Questions Asked of Study Participants Prior to Study

Q1	Are you a specialist in satellite or launcher technologies?
Q2	Do you follow closely developments in space technology, particularly satellites/launchers as part of your job responsibilities? As a part of your personal interests?
Q3	Do you read Aviation Week regularly or any other journal that regularly covers the Ariane project?
Q4	What division are you in?
Q5	How would you classify yourself as an analyst (e.g., technology, command and control, global)?
Q6	How many years have you worked as an analyst and in what areas?
Q7	What is your educational background?

The participants were then provided with a short demo of portions of the software environment provided in the study. The software that was used in the study is currently available to intelligence analysts and very similar to other tools presently in use but currently has a relatively small user base at the site where the study was conducted. While the participants observed, a training database was opened, a query was performed while referring to a "cheat sheet" for query formulation that was provided for the participants (Figure 9) and it was explained that the queries are standard Boolean full-text search, the documents were re-sorted by date and several documents were viewed by double-clicking on the view of the date and title. Some other features were demonstrated, including a marking function and a "finder" for keywords. Although there were many more features available in the software, these were the only features that were presented to minimize training time and to better match the tools that are currently used by the participants in their everyday environments.

OPERATOR	DESCRIPTION	EXAMPLES:
&	AND	((large | significant) & (maneuver* | exercis*))
|	OR	
#	EXCLUSIVE OR	(war & game*):%2
!	NOT	
:	ORDERED	((tank!(top secret):%2) & (tank))
;	UNORDERED	
%	PROXIMITY	

Figure 9. Query formulation examples provided to the study participants.

At this point, the participants were provided with a piece of paper that had the written question: "In 1996, the European Space Agency lost a satellite during the first qualification launch of a new rocket design. Give a short briefing about the basic facts of the incident: when it was, why it occurred, and what the immediate impacts were?" The participants were asked to provide a verbal briefing prior to performing any queries and then were asked to provide another verbal briefing when they felt that they had completed the analysis. The simulated task was described as a Quick Reaction Task (QRT) where the person who wrote the question was unavailable to get more information regarding the task. The participants were asked to read the number of each document that they opened to facilitate the note-taking of the investigators and to think aloud during the process. They were sometimes prompted to think aloud when they fell silent. Occasionally, when the investigators noted something of interest, they would ask questions to follow up (e.g., Why do you say that this document is a good one?), so there were several points where opportunistic interviews occurred that were useful in modeling the cues that intelligence analysts use to judge the source quality of a document.

During the session, participants asked about and used other features of the software than were described in the demonstration (Figure 10 is the main browser window with many of these functions directly available on the interface). Many of the participants used the "refine query" function. One participant narrowed a set of documents by date with the help of a facilitator, but no other participants limited the search to anything other than full text. Nearly all of the participants used the marking function for various reasons (note that opened documents were displayed as italicized within the query browser window). One participant displayed the results of one query in a view that showed the words around a keyword hit instead of dates and titles. The same participant also tried to use a "export marked documents" function that crashed the system as well as a Personal Information Manager for taking notes. The other participants used text editors (e.g., WordPerfect™) where information was copied directly from a document open in the software with right and left mouse buttons and blank pieces of paper that were provided for their notes. None of the participants asked for or used the advanced features that were available as buttons on the bottom of the

browser screen (e.g., a visualization technique that displays a set of documents organized by "similarity" based on vectors derived from words in the documents).

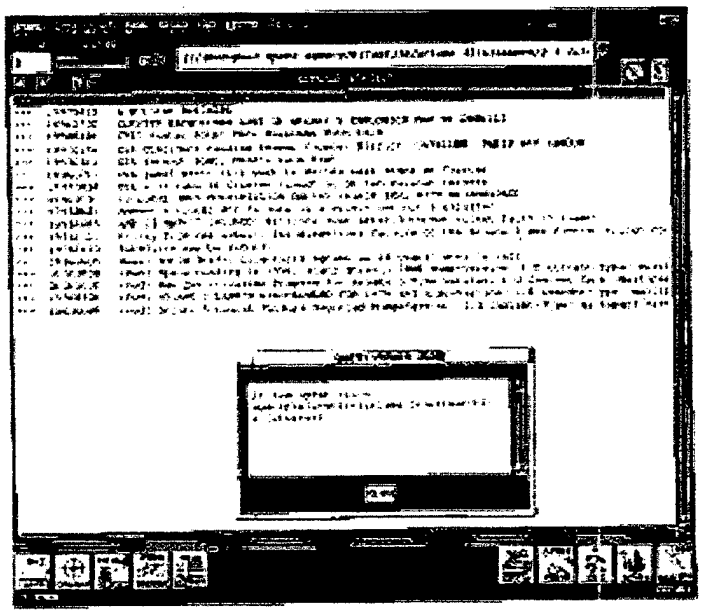


Figure 10. Browser window of software used in the study.

When the study participants felt ready, they provided a verbal response to the written question in a "briefing" style but without visual aids. One participant also provided a written briefing because writing the briefing was an important element in preparing for a verbal briefing. The participants were then asked how confident they were in their analysis and why. Some participants were then asked to perform some other simulated tasks within the Ariane scenario. The follow-up questions were either related to more details about the software failure or estimating the economic impact of the failure on the main competitors in the rocket launcher business. This data was not analyzed because the participants appeared to be less motivated to perform the task because generally they were tired and did not consider the questions to be as face valid.

At the end of the session, the participants were asked a series of follow-up questions mainly related to support tools that tended to generate animated discussion, with all of the participants indicating a great need for better computer support given data overload and workload burdens (Table 3). The session time often ran out before all of these questions were asked.

Table 3. Follow-up Questions Following the Simulated Task

Q8	Do you normally use software tools to support data search or analysis?
Q9	Which ones?
Q10	Infodominator?
Q11	DCARSS?
Q12	How familiar are you with Pathfinder?
Q13	Have you used Pathfinder?
Q14	When do you use it?
Q15	When do you use the TIS intermediary search service to search for you? How often?
Q16	How they felt about the question we posed – Ariane?
Q17	What questions about Ariane would be plausible to get in their own area of expertise?
Q18	How they felt about the tools we provided – how did it compare to the tools they normally use?
Q19	What additional tools or features do they wish were available to help them search/integrate in their actual work?

4.5 Study Participants

Ten experienced intelligence analysts, one from each major division of the affiliated intelligence agency, participated in this study (Table 4). They were selected by their management as being one of the best analysts from their divisions. Another intelligence analyst volunteered to serve as a facilitator for the software that was used in the study, and another volunteered to serve as a pilot participant to help refine the simulated task, training, and interview questions. Because the purpose of the pilot participant session was very different than for the other sessions, much of the data is missing because it was not yet decided what would be collected and how, and the time of the session was shorter, the pilot participant's (i.e., Participant 1's) data was not included in the analysis.

Although all of the study participants were clearly experienced professionals, they varied in their years of experience as intelligence analysts from 7 to 30 years. They also varied in their prior knowledge of the Ariane 501 scenario, the related domains of expertise, and the software that was used in the study. Three of the participants knew more information about the 501 launch failure prior to beginning the analysis than the others, although they did not know enough to feel comfortable giving a briefing without going through an analysis process. None of the participants were current users of the software used in the study, although three had some familiarity with previous versions. Many of the participants had some expertise in an area related to portions of the simulated task, the most relevant being launch vehicles, that clearly made them more representative as a population than surrogates such as college undergraduates would be.

Table 4. Characteristics of Study Participants

Participant	Years Exp.	Prior Knowledge of Scenario	Prior Knowledge of Software	Relevant Expertise	(Completed) Educational Degree	Gender
2	8	Ariane 5	Used to use it, but not now	—	B.S. Psychology	M
3	7	—	Never used it	communication satellites	B.S., M.S. Electrical Engineering	M
4	8	Ariane 5, reused guidance system, software error	Never used it	launch vehicles	B.S. Aerospace Engineering	F
5	17	—	Never used it	ballistic missiles	B.S. Mechanical Engineering, M.S. Strategic Intelligence	M
6	8	Ariane 5	Never used it	space systems	B.S., Electrical Engineering, M.S. Systems Management	F
7	9	Ariane 5, reused software in inertial reference system, data overload, design flaw, inadequate testing	Never used it	launch vehicles	B.S., M.S. Mechanical Engineering	M
8	11	Ariane, assume qualification launch, propulsion-related problem, satellites destroyed, and insurance rates went up	Used previous version a few times	—	B.S. Civil Engineering	M
9	18	—	Never used it	UAVs	B.S. Electrical Engineering	M
10	30	—	Never used it	—	M.S. Physics	M
11	16	—	Used to use it years ago	computer hardware	M.S. Computer Engineering	M

4.6 Base Level Protocols

The participants were asked to think aloud during the process and provide a verbal briefing when they were finished. Two investigators (Emily Patterson, Emilie Roth) directly observed this process for all of the study participants, which was also audio and videotaped. The investigators noted during the session what articles were selected and interesting verbalizations made by the participants. The investigators also electronically saved what queries were used, the documents that were returned with each query, the documents that were marked with the marking function in the software, the workspace configuration of the screen, and periodic snapshots of the electronic notes generated by the participants during the session.

From this set of data, a collection of "base protocols" which emphasized different aspects of the simulated task were generated. These include:

- the search protocol (Figure 11)
- the article protocol (Figure 12)
- the verbal briefing protocol (Figure 13)
- the conflict resolution protocol (Figure 14)

The search protocol (Figure 11) incorporated the queries that were used, the articles that were read, the date and title of the articles that were opened, why they were selected, physical behavior that indicated that an article was particularly important such as by marking it or cutting and pasting from it to a text editor, and verbalizations while they read the article. The other protocols in the set that were generated for each participant included a protocol that focused on the query formulations, a protocol that included verbalizations on the items in the pre-briefing, during the analysis process, and final verbal briefing, and unstructured notes that included the pre- and post- interview data.

No.	Query	Hits	Why	Where got keyword	How sorted
1	ESA | (european & space & agency)	725	first search	written question	by date
1A	(ESA | (european & space & agency)) > (19960601) Infodate (NOTE: Our facilitator helped with this query formulation – described need was to limit to after June 1, 1996)	419	narrow a search by date	got date from info in database (report 1274)	by date

Figure 11. Participant 5's search protocol.

The article trace can be considered to be the main protocol of the analytic process (Figure 12). This protocol was cross-checked by having a single investigator generate the protocol based on the video tape and handwritten notes from multiple investigators and then the other go through the video and adding or revising the protocol. Differing interpretations of the data were identified in this fashion and resolved through additional searches for converging evidence and debate.

Article #	Query	Name and source info	Why selected	Important?	Notes
1380	1	ARIANE 5 EXPLOSION CAUSED BY FAULTY SOFTWARE; SATELLITE NEWS	wants to work backwards so wants a late article		faulty software
1274	1	NEW CLUES TO ARIANE-5 FAILURE; DEFENSE DAILY	title and looking for date of event		June 4, 1996 (limits query results to after June 1 since event is June 4)
253	1A	STRIDE: FIRING TESTS OF NEW H IIA ROCKET ENGINE COMPLETED	time of article close to event		of no interest -- recognizes the HIIA rocket engine is from Japan
1855	1A	<<European>> <<space>> rocket explodes: Work continues with 14 similar models Ottawa citizen; Ottawa Citizen		Cuts and pastes	5 km from launch site 40 seconds 14 rockets on production line – if fault is not generic, the program won't suffer too much (software would classify as not generic according to him)
1223	1A	False computer command blamed in Ariane V failure; Aerospace Daily	6-6-96 date, also title	Cuts and pastes, marks, says good article	computer command Aerospace Daily as a good source says article is "remarkably good" and takes a while reading it June 6 knew false signal and looking closer at it Says what causes were eliminated

Figure 12. An excerpt of participant 5's article protocol.

The verbal briefing protocol (Figure 13) started to group verbalizations in relation to the topics in the written question. This protocol was used to begin to see patterns from having prior knowledge of the scenario, different interpretations of aspects of the Ariane 501 failure, and differences in verbalizations during the process as compared to what they included in an actual briefing.

Topic	Pre-Briefing	During the Analysis	"Official" Briefing
Date and time of event	1996 – from written question	date from 1274 (specifically looked for it and found it in fourth article looked at)	June 4, 1996
Software error	(nothing)	from first article 1380	Failure came as a result of taking both the mechanical equipment and apparently most of the software from Ariane 4 to Ariane 5. The input values from some of the sensors exceeded the input values that the electronic managers expected to see.
Detail of software error: numeric overflow	(nothing)	in 135 has "data overflow" and he says he doesn't understand this – would go to someone to clarify it or might include it without really understanding it if pressed for time in 1301 (mis)interpreted numerical overflow to mean too high not too long	The input values from some of the sensors exceeded the input values that the electronic managers expected to see. In other words, the data was being screened or else if simply overflowed in some respect...the number was too large and so something made the decision that this was a bogus value.
Detail of software error: diagnostic information interpreted as command data	(nothing)	(nothing said – possibly from 1882)	It sent what was essentially diagnostic information downstream to the controls.
Economic impact on Ariane program	(nothing)	in 1855 says if fault is not generic program won't suffer too much in 1882 says long-term results will be small in 1736 commentary says to keep in mind launches are inherently risky	As soon as they...they will get the business back. Still have Ariane 4. Reliable. Economies of scale for them.

Figure 13. An excerpt of participant 5's verbal briefing protocol.

The conflict resolution protocol (Figure 14) evolved over time to include the information shown in the figure. Before the study began, it was anticipated that how the participants resolved conflicting information would be an informative area for analysis. As the discrepant information was identified, this triggered the investigator to look for how all of the participants dealt with that item in their briefings and what information was available to them on that topic in what they read.

Conflicting information	From where	Indication of conflict	Resolution of conflict	Rationale	Method for tracking uncertainty	What was said in debrief
Time the rocket exploded	1725 says 37 seconds 1855 says 40 seconds 1381 says explosion caused by software at 61 seconds and ground crew exploded after that	(unsure)	Selected 37 seconds for final briefing without qualifiers	(not given)	(not done here)	37 seconds
503 as a test flight or a commercial flight	1301 says 503 as test flight another article in notes says no reason to have 503 be a test flight others might say different things	when read 1301, said that it was a confirmation that there will be another test flight which was speculated before but now is known	accepts speculation that 503 will be a test flight	confirmation from different sources	in memory	Was supposed to be a commercial flight but was changed to a test flight about one year later
Cost of the explosion	From "Maiden Ariane 5 Rocket Explodes on Launch" 6/4/96 has "pounds 500 million cargo" article title has 7 billion another article has 7 billion	says 500 million pounds for satellite and 7 billion dollars for the vehicle	keeps both estimates separate and include both	estimates are not incompatible – just focus on different things	cut and pasted info to notes	500 million pounds for the satellite and 7 billion dollar launch vehicle

Figure 14. An excerpt of participant 5's conflict resolution protocol.

4.7 Seeking Patterns Across Participants

Since this simulation study was a discovery-oriented process, the data analysis was iterative. As the base protocols were generated, potential "interesting areas" were noted for more detailed investigation. For example, while putting together the article traces, it was noted that some of the participants clearly spent much more time and relied heavily on a small subset of the documents that were read. This led to the notion of "key" documents and their role in the analysis process. Similarly, several documents were identified prior to the study as being particularly high quality articles. During the process of analyzing the data, several more were discovered. These observations led to the questions of whether or not the documents that were heavily relied upon in the analysis ("key" documents) were high quality documents ("high profit" documents) and how the sampling processes (i.e., queries and browsing) influenced the number of high profit documents that were opened.

At the same time that the protocols were being generated, the "products" of the analysis (i.e., the verbal briefings) were investigated. The verbal briefings were transcribed and items were coded as not mentioned, accurate, vague, and inaccurate for the items in Table 5. These codings were not meant as a performance measure in order to identify the characteristics of "good" and "bad" analysts, but instead were a means of identifying other paths to pursue in the data analysis. The inaccurate statements were grouped and then the process that the individual participant used that arrived at the inaccurate statement was analyzed to identify the cognitive difficulties in analysis that created general sources of inaccurate statements across participants.

Table 5. Items Coded in Verbal Briefings

Incident: Ariane 5 or 501
Date of incident: June 4, 1996
Time rocket exploded: 36, 36.7, 37, or 40 seconds
How exploded: self-destructed (not destroyed by ground controller/range officer commander)
Immediate cause of the accident: software error
Detail of software error: in the inertial reference system, inertial guidance platform, inertial platform
Detail of software error: re-use from Ariane 4
Detail of software error: not needed after liftoff
Detail of software error: numeric overflow
Detail of software error: diagnostic info as command data
Detail of software error: common-mode failure and/or backup system as well as main system failed
Design/organizational cause of the accident: insufficient testing and requirements
Payload: cluster satellites
Detail of payload: not insured
Detail of payload: scientific satellites (to study Sun-Earth relations)
Impact on scientific program: all four Cluster satellites will be rebuilt
Impact on Ariane 5 program: explosion will not stop the program; relatively little impact
Delay to Ariane 502 launch: 502 launch in October 1997 (17 months after 501 launch)
Impact on Ariane 503 launch: was originally a commercial flight but now a qualification flight
Economic cost of the explosion: almost any answer is acceptable if has an acceptable rationale

In general, the analysis process involved bottom-up searching for patterns combined with top-down conceptually driven investigations (Figure 15). The base protocols served as a detailed account of the process from several important frameworks. These protocols were used to identify patterns on particular themes. These patterns were then represented across participants in ways that highlighted similarities and differences along conceptual dimensions.

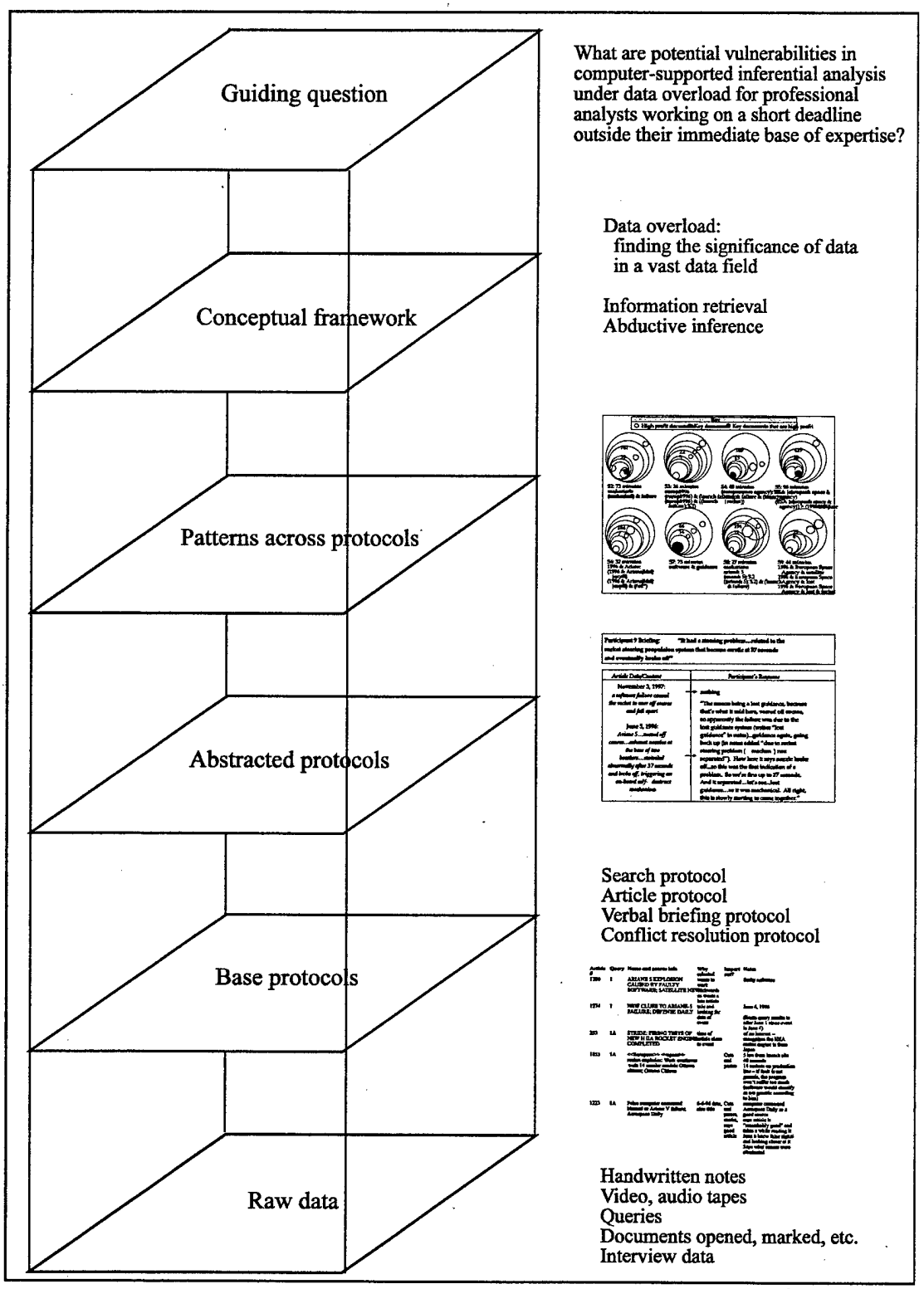


Figure 15. Levels of analysis in the process tracing methodology.

4.8 Walkthrough of the Simulated Task From the Participant's Perspective

Study participant 9 will be used to walk through a simulation session from the participant's perspective because his process was the shortest and least complex process of the participants, and therefore is easiest to describe.

The participant signed the consent form for permission to audio and video-tape the session. He was then asked the initial questions (Table 3), to which he responded that he was not a specialist in satellite or launcher technologies, did not closely follow that area as part of his job responsibilities or personal interests, somewhat regularly read Aviation Week as well as seven other important sources of information. He then described what division he was in, how he would classify himself as an analyst, that he had 18 years of experience as an intelligence analyst from working in six different areas, and his educational background. He was then given a training demonstration of the software to be used in the study and the written question. When asked to provide a verbal briefing before beginning the session, it was clear that he had no specific expertise directly relevant to or prior knowledge of the simulated task because he was unable to do so.

The participant began the analysis process by typing in the query "1996 & European Space Agency & satellite", which returned 250 documents. He stated that he wanted to do a "data reduction" based on looking at the number of hits returned, so refined the query by adding the keyword "lost." This returned 42 documents. He again refined the query by adding the keyword "rocket." This returned 29 documents (Table 6) sorted by a "relevance" metric similar to ones used on standard Web browsers.

Table 6. Information Available to Participant 9 Ranked by "Relevance Score"

Date	Title
00000000	STATUS, CONSTRAINTS, FUTURE THEMES OF SATELLITE COMMUNICATION BUSINESS IN JAPAN
00000000	BMFT REPORT ON GERMAN RESEARCH POLICY, PROGRAMS, FUNDING: MAIN AREAS OF FEDERAL R&D SUPPORT
00000000	FRANCE; ARIANE FAILURE ANALYZED
00000000	ITALY: MINISTER, INDUSTRIALISTS ON NEW SPACE POLICY.
19970723	CONFIDENCE BOOSTERS
19970115	PAYING FOR ROCKET SCIENCE
19970423	A BULLISH BUSINESS
19971103	ARIANE 5 ROCKET LAUNCHED SUCCESSFULLY ON SECOND ATTEMPT Officials Say Several Anomalies Were Relatively Minor
19960605	Ariane V fails on first launch attempt; ESA plans retry
19960702	CLUSTER SATELLITES LOST IN ARIANE-5 EXPLOSION MAY BE REBUILT
19960200	Going up; commercial satellite launch services industry; Industry Overview
19931200	Q&A; interview with Sam Mihara, Staff Director of the Space Transportation Division, McDonnell Douglas Space Systems; Upfront; Interview
19960506	SS/LORAL INKS MULTIMILLION DOLLAR LAUNCH DEAL WITH ARIANESPACE
19960605	Losing a Rocket And a Satellite Edge?; Ariane 5's Failure May Shift Launches to U.S. Firms
19960605	AND IT WASN'T INSURED; Billions lost after European rocket falls in flames
19960605	Front page - first section: European space bid harmed as Ariane 5 explodes: Blow to French company's hopes of dominating satellite launches
19960804	Intuition could have saved Ariane-5 rocket
19961125	Rising from the ashes?; The disastrous failure of the Ariane 5 and Russian Proton rockets dashed the hopes of many scientists. But they may be rescued by a daring plan, codenamed 'Phoenix'.
19960610	Satellite may be rebuilt
19960605	News: World Trade: Scientists aghast as 10 years' work is lost
19960605	INTERNATIONAL BUSINESS; Costly Failure: Space Launch Is Aborted
19960607	Letter to the Editor: Space pays dividends
19970000	THE SHAPE OF THINGS TO COME: 1997: SCIENCE: HEAVEN AND EARTH; Independence day on Mars?
19960420	Europe banks on Ariane 5 to maintain market lead.
19931220	Q&A; interview with Sam Mihara, Staff Director of the Space Transportation Division, McDonnell Douglas Space Systems; Upfront; Interview
19970000	PAYING FOR ROCKET SCIENCE
19941120	What Next for Launchers?
19960506	ARIANE 5 LAUNCH RESCHEDULED FOR LATE MAY
00000000	(Fwd) Space Funding in 1996, Plans Through 2000

At this point, he began browsing the dates and titles. He first opened the document "France: Ariane Failure Analyzed," which was about a launch failure in 1994. He stated that he was surprised that the default was not sorting by date. He then sorted the articles by date (Table 7). During the remainder of the session, he did not change this basic workspace configuration, either by conducting further searches or sorting the articles by other criteria.

Table 7. Information Available to Participant 9 Ranked by Document Date

Date	Title
19971103	ARIANE 5 ROCKET LAUNCHED SUCCESSFULLY ON SECOND ATTEMPT Officials Say Several Anomalies Were Relatively Minor
19970723	CONFIDENCE BOOSTERS
19970423	A BULLISH BUSINESS
19970115	PAYING FOR ROCKET SCIENCE
19970000	THE SHAPE OF THINGS TO COME: 1997: SCIENCE: HEAVEN AND EARTH; Independence day on Mars?
19970000	PAYING FOR ROCKET SCIENCE
19961125	Rising from the ashes?; The disastrous failure of the Ariane 5 and Russian Proton rockets dashed the hopes of many scientists. But they may be rescued by a daring plan, codenamed 'Phoenix'.
19960804	Intuition could have saved Ariane-5 rocket
19960702	CLUSTER SATELLITES LOST IN ARIANE-5 EXPLOSION MAY BE REBUILT
19960610	Satellite may be rebuilt
19960607	Letter to the Editor: Space pays dividends
19960605	Ariane V fails on first launch attempt; ESA plans retry
19960605	Losing a Rocket And a Satellite Edge?; Ariane 5's Failure May Shift Launches to U.S. Firms
19960605	AND IT WASN'T INSURED; Billions lost after European rocket falls in flames
19960605	Front page – first section: European space bid harmed as Ariane 5 explodes: Blow to French company's hopes of dominating satellite launches
19960605	News: World Trade: Scientists aghast as 10 years' work is lost
19960605	INTERNATIONAL BUSINESS; Costly Failure: Space Launch Is Aborted
19960506	SS/LORAL INKS MULTIMILLION DOLLAR LAUNCH DEAL WITH ARIANESPACE
19960506	ARIANE 5 LAUNCH RESCHEDULED FOR LATE MAY
19960420	Europe banks on Ariane 5 to maintain market lead.
19960200	Going up; commercial satellite launch services industry; Industry Overview
19941120	What Next for Launchers?
19931220	Q&A; interview with Sam Mihara, Staff Director of the Space Transportation Division, McDonnell Douglas Space Systems; Upfront; Interview
19931200	Q&A; interview with Sam Mihara, Staff Director of the Space Transportation Division, McDonnell Douglas Space Systems; Upfront; Interview
00000000	STATUS, CONSTRAINTS, FUTURE THEMES OF SATELLITE COMMUNICATION BUSINESS IN JAPAN
00000000	BMFT REPORT ON GERMAN RESEARCH POLICY, PROGRAMS, FUNDING: MAIN AREAS OF FEDERAL R&D SUPPORT
00000000	FRANCE; ARIANE FAILURE ANALYZED
00000000	ITALY: MINISTER, INDUSTRIALISTS ON NEW SPACE POLICY.
00000000	(Fwd) Space Funding in 1996, Plans Through 2000

After sorting by date, the first article in the list "Ariane 5 rocket launched successfully on second attempt" was automatically selected. The participant glanced through it. Then he selected "And it wasn't insured; billions lost after European rocket falls in flames" based on the title. He said "here is the event" and carefully read the document, taking notes and commenting on what he was inferring from the information as he read. He then looked again at the written question and decided that he could answer the question: "As far as I'm concerned, I've got enough information to answer this."

The written question:

> In 1996, the European Space Agency lost a satellite during the first qualification launch of a new rocket design. Give a short briefing about the basic facts of the incident: when it was, why it occurred, and what the immediate impacts were?

The participant's verbal briefing as a response:

> The European Space Agency lost a satellite during the first qualification, so this is a valid statement. According to this, it was their first launch and a new rocket design which was the fifth and not the fourth, so it is a new design. Give a short briefing about the basic facts of the incident. The facts are, it was...date...did I get an absolute date here? Nobody gave me a date...(looks at the last document to retrieve the date) When was it? In the first part of 96. Why did it occur? It had a steering problem that occurred 37 seconds after launch, and the steering problem was related to the rocket steering propulsion system that became erratic at 37 seconds and eventually broke off, starting a self-destruct sequence at the 40-second mark. This destroyed the vehicle. That was what happened, that was when it occurred. OK, what's this? No one asked about payload? That's interesting. The immediate impacts were that this was an additional slip to a program that had already been slipped before. The impact was a sizeable financial venture lost. The bottom line is that there will be another attempt, so it wasn't such a big loss after all.

The participant was then asked how he decided when to stop:

> I could talk for 15 minutes on this. My approach to briefing is that you throw out...and if don't have time to embellish it, you throw bait out and you switch to questions and answers. This would probably be a ten-minute briefing.

Note that this answer is based on a reference to a previous conversation with the investigator. During the analysis, he had indicated that he was treating this as a 15 minute briefing and that he seemed to be basing his decision on when to stop on whether or not he had adequately answered all aspects of the written question:

> OK. Let's see if I've answered them all. Lost a satellite, OK, I have dates and places. First launch. New rocket. I have (names and facts). Again, where it was, where it occurred, and what the main impacts were. I have all of it, just off of this. Do you want me to embellish what I have here? The question is. The data came out pretty quick. Define short briefing. What is a short briefing? Define short. Most briefings are 15 minute briefings. If you're doing command briefings. If you're doing a working group, it could be half an hour to 45 minutes. If you're out on the briefing circuit (Pentagon, NSA), they want a briefing that justifies the cost of bringing you there. I see this as a short briefing. On QRTs sometimes 5-6 a day, sometimes none. It's event-driven. If playing war games, I might get 15 in a week. Other times, I might get none at all. QRTs or requests for information.

The participant was then asked how confident he was in his briefing and he answered:

> 3 on 1-5 because it's a FBIS report. Can it be verified? I'd have to go to the next series of documents to see if I have two independent people saying the same thing – then it would go from 3 to 4. If I can confirm from what I consider to be another independent source.

At this point, the participant was asked to demonstrate how he would go about raising his confidence level. He drew a line in his handwritten notes and stated that once information was verified that he would draw arrows from information above the line to information below the line. He also stated that he would print out the documents and use highlighter pens to indicate that information was on the same topic and from how many sources and whether or not the data was converging or conflicting. Rather than print out the documents, he illustrated this strategy by using colored font on a word processor where he started creating electronic notes in addition to his handwritten notes by cutting and pasting information from documents. Time ran out before he was able to complete this "embellishing," so he declined to provide another verbal briefing but stated that the first one that he gave might be the level at which he might brief his immediate superior on an analysis of relatively little importance.

The participant was then asked some of the follow-up questions. He indicated that he had never previously used the software in the study, that he used a competing package in his work, and that he generally conducts his own searches rather than using the professional search intermediaries at the site, although he uses them for other reasons.

4.9 Using Simulation Capabilities to Conduct Field Research

It has long been recognized that field research techniques are useful early in the process of learning about the cognitive challenges in a domain for discovering unanticipated factors that might influence performance. Through field research methodologies, it is possible to generate a rich understanding of the interplay between the demands of the world, strategies of the practitioners, and the function of artifacts as tools. Traditionally, field methodologies involve very little shaping of the environment, artifacts, participants, or scenarios involved in the investigations, such as critical incident interviews (Flanagan, 1954; Klein, Calderwood, & MacGregor, 1989) and naturalistic observations (e.g., Hutchins, 1995). Without the ability to shape the conditions of observation, it is possible to miss critical cases, it is difficult to evaluate how much the findings derived from a particular case are dependent upon an individual participant or scenario, and it is not possible to study "envisioned worlds" that are predicted for the future but not yet in existence.

These limitations can be overcome through the use of simulation capabilities to provide converging evidence during the exploration of the challenges in a particular domain. Often simulation capabilities are used later in the process of learning about a domain to conduct highly controlled "scaled world" or "microworld" comparison studies, but there is nothing inherent in the simulation methodology that requires it to be used in

this fashion. Simulating complex tasks with real-world participants can be viewed as an opportunity to utilize the field research techniques of naturalistic observation and cued interviewing for situations that cannot easily be naturally observed (e.g., when information on how particular cases were conducted is restricted to protect national security interests).

When simulation capabilities are used in this fashion, investigators can probe areas of interest in ways that are analogous to perturbing a complex engineered system in order to examine how the perturbation affects the process (e.g., breaking an electrical circuit to see how a system will react). For example, in this simulation study, study participants were put under the challenging conditions of a workload bottleneck (i.e., too many documents to read in the time available) and working outside their immediate areas of expertise (i.e., the participants had general analytic skills and technical knowledge, but they could not draw on extensive prior knowledge to perform the task). These conditions were expected to stress the cognitive difficulties in finding the significance of data, which was the main guiding framework. The Ariane 501 scenario used in the simulated task was expected to contain representative challenges in inferential analysis. The Ariane 501 launch failure was an accident with high consequences that was the result of a complex systems failure that was not due to a typical mechanical breakdown. The cognitive engineering research base strongly suggested that this scenario would contain data that was significant in how this accident was a departure from standard cases that would be contained in a vast data field with conflicting descriptions and information coming in over time.

This simulation study serves a critical role in an overall cognitive task analysis process aimed at discovery (Potter, Roth, Woods, & Elm, in press). A base of understanding about data overload generated from previous research in other domains, including space shuttle mission control, nuclear power plants, anesthesiology, and aviation flight decks, was used to frame the data overload problem as one of finding the significance of data against an ongoing context where information needs to be synthesized and integrated from a vast data field. The research base developed during this previous research was synthesized and the simulation study will be used to calibrate against the new domain of intelligence analysis. In addition, the database and environment used in this simulation study will now be able to serve as a "testbed" for future studies aimed more directly at probing the complex system with particular interventions, such as visualizations designed to reduce the vulnerabilities in inferential analysis described in the following section.

5 STUDY FINDINGS

5.1 Cognitive Tasks in Inferential Analysis

The inferential process employed by all the study participants[1] can be abstractly described as following three interrelated stages: information selection, corroborating information and resolving conflicts, and story construction (Figure 16). Information was selected from the database through the refinement of keyword queries and by sequentially browsing the returned reports by dates and titles. Some of the sampled reports were used as the main basis for the analysis, which we refer to as "key" documents. The key documents were used to generate the skeleton of the analysis product. Supporting documents were then used to corroborate the important information and fill in details. Conflicts in the data were flagged and judgments about which data to include were revisited as new information on the topic was identified. When the study participants felt ready, they organized their notes and generated a coherent story to respond to the question.

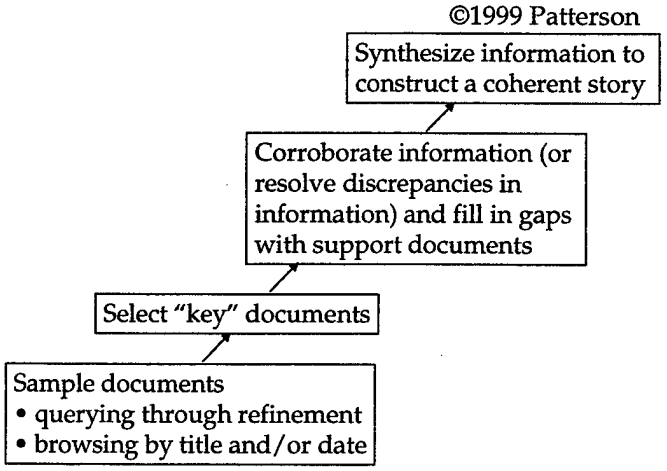


Figure 16. Abstract view of the inferential analysis process.

[1] Two study participants' data were not included in the analysis. Participant 11 attempted to analyze a different satellite failure (SPOT-3) which was not well-supported by the database. Participant 10 did not complete the task because the printer was not working during his session and printing documents before reading them was a key part of his analytic process so that he could see them in parallel and make marks on them.

5.2 Patterns in Information Sampling

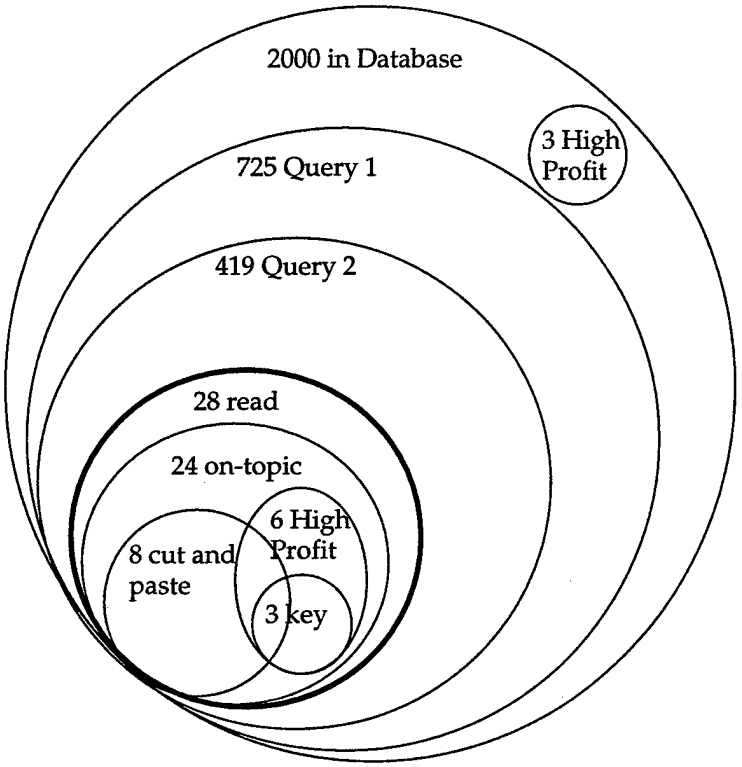
5.2.1 Sampling by Narrowing in

In inferential analysis under data overload in baseline electronic environments with textual databases, information is effectively sampled, generally through querying and browsing. In our study, participants were observed to begin the analysis process by making queries with standard inputs such as keywords and date limits. If a returned set of documents was judged to be too large, the search was narrowed rather than starting with a new set of search terms. Typical narrowing strategies included adding a keyword, limiting to a date range, or enforcing a proximity requirement on a set of keywords. The search was then further narrowed through the process of browsing by summary information about a document, typically dates and titles. Then documents were opened by double-clicking on a report title.

A subset of the opened documents was judged not relevant to the analysis. Of the documents that were judged to be relevant, generally some subset of the information from the document was read or copied to notes in a text editor. Of this set of documents, a small number were used as the basis for the analysis, which we refer to as "key" documents. For this study, the definition of what documents were treated as "keys" was based on converging behavioral and verbal data from the process traces. The key documents were associated with verbalizations such as "Here we go!" or "That's a good one!" In addition, the participants were often observed to spend a longer time reading them than other documents, copy much of the document to their electronic notes, and/or use the marking function in the database software to highlight the title in the browsing window. Convergingly, it could sometimes be determined from the verbal briefings what documents were heavily relied upon in the analysis.

To illustrate this process, consider the information sampling process employed by study participant 5 during the analysis (Figure 17). The participant started with a Boolean keyword search (esa OR (european AND space AND agency)). This search returned 725 hits, so he narrowed the search to documents published after June 1, 1996 after determining that the date of the incident was June 4, 1996 from scanning three articles. After this narrowing criteria, 419 documents remained, which became his "home query" in that he did no more keyword searches. Twenty-eight documents were opened during the analysis (not including two duplicates), 24 of which were on-topic, or relevant to the analysis. Six of the documents that he opened were "high-profit" in that they were judged by the investigators to be highly informative documents. The other three high-profit documents were available in the database but were not returned by either query. The participant cut and pasted portions of eight documents along with references into a word processing file and used a marking function in the software to highlight two documents, one because he stated that it was a remarkably good article and one to mark in case he needed to refer back to it later in the analysis for further information. Three articles were identified as his "key" documents – 1) document 1223

because he remarked that it was "remarkably good" and spent a long time reading it, 2) document 1301 because he spent a long time reading it and made many verbalizations about details of the incident while reading it and said after reading it that now he had a good idea of what had happened, and 3) document 1882 because he said that it was "a definite keeper," that it was like briefings by professional analysts in its quality, spent a long time reading it, cut and pasted the most text from it, and made many verbalizations while reading it. All three of his key documents were high profit documents.



Study Participant: 5
Time: 96 minutes
Experience: 17 years
Query 1: ESA | (european & space & agency)
Query 2: (ESA | (european & space & agency)) > (19960601)Infodate

Figure 17. Searching process employed by study participant 5.

The searching process of study participant 5 was essentially one of continually narrowing in. An initial query was refined to reach a document set that was judged manageable based on the number of hits. A small subset of these documents was then heavily relied upon in generating the analysis product.

Looking at the searching processes for all of the study participants (Figure 18), this process was very representative. All of the participants narrowed their queries to a number that they judged to be manageable (22 – 419 documents) from which they opened documents based on a view of the dates and titles (4 – 29 documents). They then relied heavily on a subset of these documents (1-4 documents) for their verbal briefings.

During the process of searching for information, some study participants verbalized that perhaps they should conduct new searches for specific information, but did not. In addition, comments made by some of the study participants indicated that they did not know what was available in the database and how their queries related to what was available, which made them uncomfortable. In spite of these statements, the study participants appeared reluctant to leave the working area that the home query window represented. The participants developed a familiarity with the titles and dates of the documents returned by the query, the documents had often been sorted by the participant by date, the windows had been resized and placed in a dedicated place on the screen, and some of the documents had been marked for various reasons.

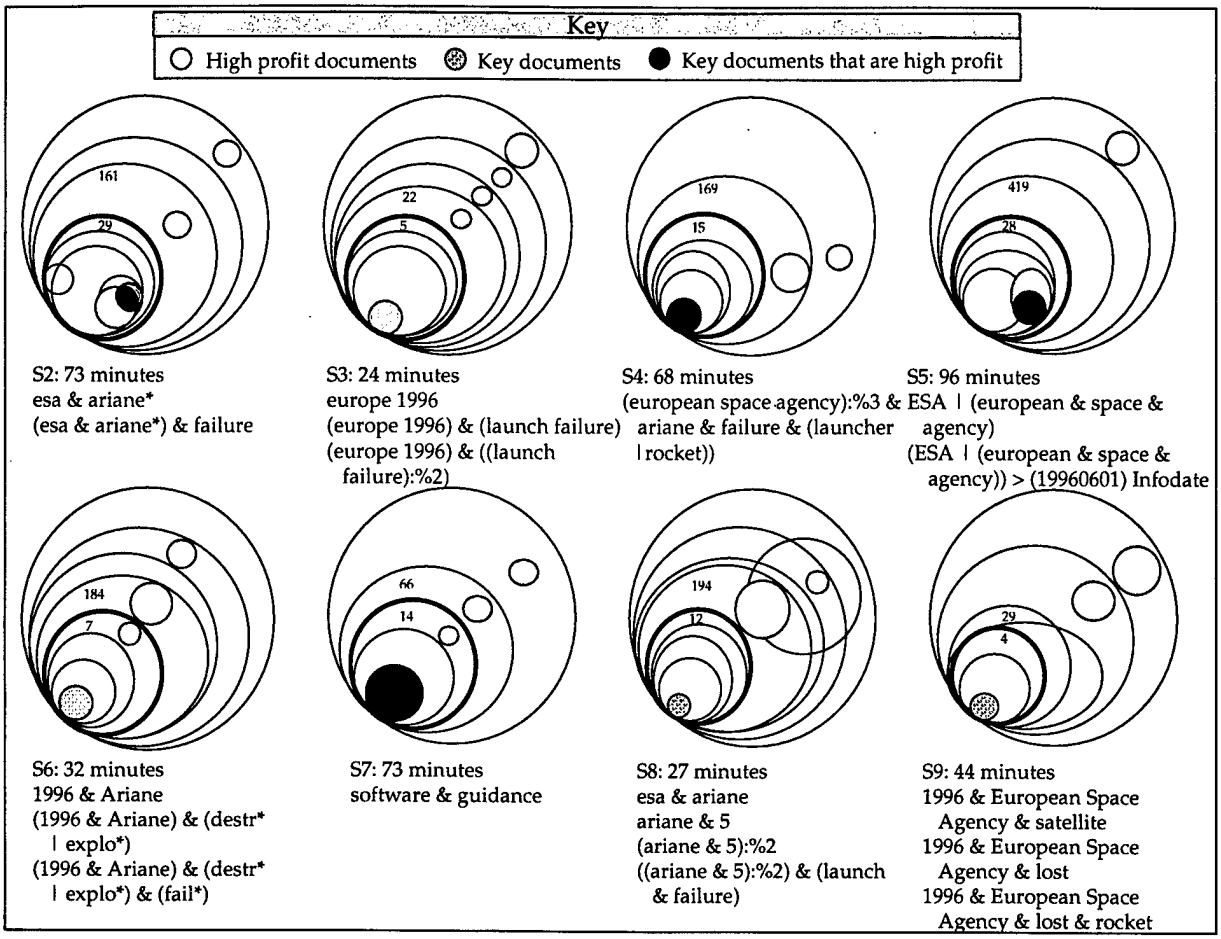


Figure 18. Searching process employed by all study participants.

It is not surprising, given the type of computer support that was provided to the participants, that all of the participants missed high profit documents without being aware of it. Samples that were returned by the keyword searches were essentially opaque in terms of how they related to what was available, such as what high profit documents were left out of the query results. Then documents were sampled based on a view of the dates and titles, which were also relatively weak indicators of whether or not documents were high profit. For example, the first article listed as a low-profit article in Table 8 was a translated, rewritten description of an article originally published in Italy that contained inaccuracies about the details of the cause of the software failure. The second article was a one-paragraph abstract and so contained very little information. The third article contained some inaccuracies because it was published soon after the event and therefore did not have all of the information available when it was published.

Table 8. Dates and Titles of Low and High Profit Articles

"Low-profit" articles	"High-profit" articles
Europe: Causes of Ariane 5 Failure (July 5, 1996)	Software design flaw destroyed Ariane V; next flight in 1997 (July 24, 1996)
Ariane 5 Failure: Inquiry Board Findings (July 25, 1996)	Board Faults Ariane 5 Software (July 29, 1996)
False computer command blamed in Ariane V failure (June 6, 1996)	Ariane 5 loss avoidable with complete testing (September 16, 1996)

5.2.2 Basing Analyses on High Profit Documents

Looking more closely at the process traces in Figure 18, the black circles represent when the key documents were also high profit documents, or in other words, when the documents that were heavily relied upon were the best documents available in the database. Comparing the four participants that used high profit documents as key documents vs. the four that did not, there are some interesting differences between the two groups (Tables 9 and 10). The participants that used high profit documents as key documents spent more time during the analysis, read more documents, and read more of the high profit documents.[2]

[2] Note that non-parametric statistics were calculated to support the interpretation of patterns that were observed across participants. These statistical measures should be interpreted with caution as this study was exploratory and therefore not designed to make strong statements about statistical differences. There were a relatively small number of study participants, variation and interactions were not controlled, and multiple tests (9) were run on the data. The reported patterns and associated statistical data should therefore be viewed as suggestive and converging with the observational data but not confirmatory in their own right.

Table 9. Participants That Used High Profit Documents as Key vs. Not

Participants whose key documents were not high profit documents

Participant	Experience (years)	Time (mins.)	Final query (no. hits)	Documents (no. read)	High profit docs (no. read)
3	7	24	22	5	0
6	8	32	184	7	2
8	11	27	194	12	0
9	18	44	29	4	0
Average:	11	32*	107	7*	0.5*

Participants whose key documents were high profit documents

Participant	Experience (years)	Time (mins.)	Final query (no. hits)	Documents (no. read)	High profit docs (no. read)
2	8	73	161	29	3
4	8	68	169	15	2
5	17	96	419	28	2
7	9	73	66	14	5
Average:	10.5	78*	204	22*	3*

* significant difference using Wilcoxon-Mann-Whitney Non-Parametric test

Table 10. Comparison of Querying and Browsing Breadth

Participants whose key documents were not high profit documents

	Final "Home" Query	No. of Hits in Query	No. of High Profit Hits in Query	Percent of Query Docs that are High Profit	No. of Documents Read	No. of High Profit Documents Opened	Percent of "Key" Docs that are High Profit
3	(europe 1996) & ((launch failure):%2)	22	1	5%	5	0/9	0% (0/1)
6	(1996 & Ariane) & (destr* | explo*) & (fail*)	184	7	4%	7	2/9	0% (0/3)
8	((ariane & 5):%2) & (launch & failure)	194	8	4%	12	0/9	0% (0/1)
9	1996 & European Space Agency & satellite & lost & rocket	29	0	0%	4	0/9	0% (0/1)
	Average:	107	4	3%	7*	0.5/9*	0%

Participants whose key documents were high profit documents

	Final "Home" Query	No. of Hits in Query	No. of High Profit Hits in Query	Percent of Query Docs that are High Profit	No. of Documents Read	No. of High Profit Documents Opened	Percent of "Key" Docs that are High Profit
2	(esa & ariane*) & (failure)	161	6	4%	29	3/9	50% (1/2)
4	(european space agency):%3 & ariane & failure & (laucher | rocket))	169	7	4%	15	2/9	100% (2/2)
5	(ESA | (european & space & agency)) > (19960601) Infodate	419	7	2%	28	2/9	33% (1/3)
7	Software & guidance	66	7	11%	14	5/9	100% (4/4)
	Average:	204	7	5%	22*	3/9*	71%

* significant difference using Wilcoxon-Mann-Whitney Non-Parametric test

Before drawing implications from this data, we will first consider alternative explanations for the differences between these two groups. It is generally recognized in the information retrieval literature that both search and domain expertise is important in information seeking. Therefore, it is possible that the group of analysts that relied on the high profit documents used more effective search strategies to find the documents. Similarly, it is possible that the more experienced professional analysts have developed strategies that help them to perceive high profit documents, or that domain- or scenario-related expertise would make it easier for them to recognize high profit documents.

To that end, we will now examine the following alternative explanations for the differences between the two groups:

A) Differences in searching expertise

S1: Did the group that had some high profit documents as their key documents have a higher percentage of high profit documents in their queries than the other?

S2: Did the group that had some high profit documents as their key documents have higher recall of high profit documents in their queries than the other?

S3: Did the group that had some high profit documents as their key documents use different types of search tactics than the other?

S4: Was the group that had some high profit documents as their key documents more consistent in the search terms that were used than the other?

S5: Did the group that had some high profit documents as their key documents use the search intermediaries less often than the other?

B) Domain, Scenario, System Expertise/Knowledge

D1: Did the group that had some high profit documents as their key documents have more years of experience in intelligence analysis than the other?

D2: Did the group that had some high profit documents as their key documents have more prior knowledge of the question than the other did?

D3: Did the group that had some high profit documents as their key documents have more terms in their queries that came from sources other than the written question?

D4: Did the group that had some high profit documents as their key documents have more prior knowledge of the software used in the study than the other did?

5.2.2.1 Impact of searching expertise.

S1: Did the group that had some high profit documents as their key documents have a higher percentage of high profit documents in their queries than the other?

Looking at Table 10, the group that had some high profit documents in their key documents did not have a significantly higher precision of high profit documents in their queries than the other (p = .34 using the Wilcoxon-Mann-Whitney non-parametric test).

S2: Did the group that had some high profit documents as their key documents have higher recall of high profit documents in their queries than the other?

Looking at Table 10, the group that had some high profit documents in their key documents did not have a significantly higher recall of high profit documents in their queries than the other (p = .34 using the Wilcoxon-Mann-Whitney non-parametric test). Note that it is possible that, as it always is in recall estimations, that there were more high profit documents in the database. Given the nature of the high profit documents, however, it is unlikely to be more than one or two more documents. Also note that since the non-parametric test is a ranking test, the total number of high profit documents has no effect on the significance of the variable.

S3: Did the group that had some high profit documents as their key documents use different types of search tactics than the other?

Bates (1979) introduced the difference between a search strategy, defined as a plan for an entire search task such as citation searching, and a search tactic, a move made to further a search. She described a set of tactics, organized into four groups: monitoring tactics, file structure tactics, search formulation tactics, and term tactics (see Bates, 1979, 1992 for the full set of tactics). She recommends particular tactics to use to expand or narrow the number of hits returned by a query (Table 11).

Table 11. Bates' Tactics to Use to Widen or Narrow a Search

Widening tactics

Tactic	Description
Super	Change term upward to superordinate term
Relate	Change term sideways to a related coordinate term
Reduce	Minimize elements in query
Parallel	Include more synonyms/conceptually parallel terms
Neighbor	Seek terms by looking at neighboring terms (alphabetically, by subject)
Trace	Examine information already retrieved to get additional terms
Vary	Alter/substitute search terms

Narrowing tactics

Tactic	Description
Sub	Change term downward to a more specific term
Exhaust	Include most of all the elements in query
Pinpoint	Minimize or reduce parallel terms
Block	Reject certain terms

Similarly, Wilson (1992) recommends a set of tactics for widening or narrowing a search in free text search tools (Table 12).

Table 12. Wilson's Tactics to Use to Widen or Narrow a Search

Widening tactics

Description
Search additional databases
Use more general terms
Expand search space from titles to abstracts to full text
Relax proximity requirements
Add "OR" terms in a facet
Drop a subject facet
Drop a non-subject facet (time period, language restriction)

Narrowing tactics

Description
Search fewer databases
Use more specific terms
Narrow search space from full text to abstracts to titles
Tighten proximity requirements
Remove "OR" terms in a facet
Add a subject facet
Add a non-subject facet (time period, language restriction)

Based on these lists of narrowing search tactics, the coding categories in Table 13 were developed. The codes associated with each of the two groups are summarized in Table 14. Essentially, all of the participants only used narrowing tactics and no widening tactics. It appears that the group that did not use high profit documents as key documents might have used more narrowing tactics in general than the other group (8 vs. 2), but any inferences drawn on such a small data sample based on counting categorical data should be made with caution. If this is true, then, as might be expected based on previous studies in information retrieval (Blair, 1980), query refinement through narrowing tactics has a negative impact on the ability to find high profit documents in a database. On the other hand, under data overload situations, narrowing is a necessary coping strategy. It is probable that some narrowing tactics might be better than others in locating high profit documents (e.g., using non-content attributes such as dates or proximity restrictions instead of keywords), although there is not really enough data from this study to comment on this issue. Also, this comparison only looks at refinement strategies. It would certainly be possible with this measurement technique that the initial query of a participant could be very restrictive but the participant would still be classified as not using any narrowing strategies.

Table 13. Coding Categories for Narrowing Tactics

Code	Description
CHANGE TERM	Change to more specific terms
NSS	Narrow search space (e.g., from full text to only in title field)
PROX	Add/tighten proximity requirements
DROP OR	Remove "OR" terms within a facet
ADD FACET	Add "AND" terms/facets
ADD ATTRIBUTE	Limit by non-subject attribute such as date
BLOCK	Reject certain terms

Table 14. Narrowing Tactics Used by Two Groups

Participants whose key documents were not high profit documents

Participant	Narrowing tactics
3	ADD FACET, PROX
6	ADD FACET
8	CHANGE TERM, PROX, ADD FACET
9	CHANGE TERM, ADD FACET
Total:	ADD FACET (4), PROX (2), CHANGE TERM (2)

Participants whose key documents were high profit documents

Participant	Narrowing tactics
2	ADD FACET
4	----
5	ADD ATTRIBUTE
7	----
Total:	ADD FACET (1), ADD ATTRIBUTE (1)

S4: Was the group that had some high profit documents as their key documents more consistent in the search terms that were used than the other?

One question that is commonly discussed in the information retrieval literature is whether or not certain groups of searchers are more consistent in their selection of search terms than other groups. It is possible that the group that located the high profit documents did so by using search terms that were particularly good. Similarly, it is possible that the group that did not locate as many of the high profit documents were consistent in using poor terms. Following Saracevic, Kantor, Chamis, and Trivison (1988) and Iivonen (1995), the following asymmetric formula was used to calculate the intersearcher consistency in search terms in order to look into these possibilities:

$$CT_{1,2} = \frac{|T_1 \wedge T_2|}{|T_1|} = \frac{\text{number of search terms in common}}{\text{total number of search terms used by Searcher 1}}$$

and

$$CT_{2,1} = \frac{|T_1 \wedge T_2|}{|T_2|} = \frac{\text{number of search terms in common}}{\text{total number of search terms used by Searcher 2}}$$

The number of search terms in common are:

	S2	S3	S4	S5	S6	S7	S8	S9	Terms	
S2			1	2	1	2	0	2	0	3
S3				2	1	2	0	2	2	4
S4					3	2	0	2	4	7
S5						0	0	0	3	5
S6							0	2	1	5
S7								0	0	2
S8									0	4
S9										6

This translates to percentages as:

	SX2	SX3	SX4	SX5	SX6	SX7	SX8	SX9
S2	--	0.33	0.67	0.33	0.67	0.00	0.67	0.00
S3	0.25	--	0.50	0.25	0.50	0.00	0.50	0.50
S4	0.29	0.29	--	0.43	0.29	0.00	0.29	0.57
S5	0.20	0.20	0.60	--	0.00	0.00	0.00	0.60
S6	0.40	0.40	0.40	0.00	--	0.00	0.40	0.20
S7	0.00	0.00	0.00	0.00	0.00	--	0.00	0.00
S8	0.50	0.50	0.50	0.00	0.50	0.00	--	0.00
S9	0.00	0.33	0.67	0.50	0.17	0.00	0.00	--

In order to perform a Chi-square analysis, the data were categorized for comparison (Table 15).

Table 15. Comparison of Intersearcher Consistency

Percent consistency	k-hp	k-nhp	Combined
0-20	7	4	11
21-40	2	3	5
41-60	3	5	8
61-80	0	0	0
81-100	0	0	0
Total	12	12	24

The Chi-square non-parametric test for independent samples shows that this is not a significant difference (Siegel and Castellan, 1988) with the two-tailed test ($p = .83$). Therefore neither group appeared to be more consistent than the other in the selection of search terms.

S5: Did the group that had some high profit documents as their key documents use the search intermediaries less often than the other?

The responses to whether or not the participants use the search intermediary service are given in Table 16. There do not appear to be any consistent patterns between the two groups. Three out of each of the groups use the search intermediaries at least occasionally for searches. The other participant out of each group does not use the search intermediaries for searches. Nearly all of the participants use the search intermediaries to set up the keyword combinations that select hundreds of daily incoming messages from other intelligence agencies.

Table 16. Responses to How Search Intermediaries Are Used

Participants whose key documents were not high profit documents

Participant	When they use search intermediaries (paraphrases)
3	I use them for big, new questions to help clear out the junk, look at other databases, and find stuff in data and use to get converging searches.
6	I use them for profiles. I expect to use them in the future for searches when I have a specific task.
8	If I can do a search by myself in an hour or so I will. If I need to have a search done that incorporates a lot of data or requires access to commercial databases, then I will use them.
9	They helped me with my profile. I do my own searches.

Participants whose key documents were high profit documents

Participant	When they use search intermediaries (paraphrases)
2	Yes, I use it, depending on the expertise of the intermediary.
4	I use them for anything not specific and known on STAIRS and for databases other than STAIRS such as CDs, or when I don't have a good idea of the time frame to get stuff.
5	I use them for profiles and searches; I can't track all the new material alone. If I don't know where it is. For some things in the STAIRS database. I also rely on the library staff to send emails with current contents.
7	I have not used them for searches in 6 years. I only use them to get something specific that I can't get and for specific help on profiles.

5.2.2.2 Impact of domain expertise.

D1: Did the group that had some high profit documents as their key documents have more years of experience in intelligence analysis than the other?

As can be seen in Table 17, there were no significant differences in years of analytic experience between the two groups (p = .44 using the Wilcoxon-Mann-Whitney non-parametric test). Therefore, there does not appear to be a correlation between experience and the ability to locate high profit documents. This means that the difficulty in locating high profit documents is probably not an issue of perception – recognizing high profit documents when they are located – so much as locating likely candidates in the first place.

Table 17. Comparison of Years of Analytic Experience

Participants whose key documents were not high profit documents

Participant	Experience (years)
3	7
6	8
8	11
9	18
Average:	11

Participants whose key documents were high profit documents

Participant	Experience (years)
2	8
4	8
5	17
7	9
Average:	10.5

D2: Did the group that had some high profit documents as their key documents have more prior knowledge of the question than the other did?

The study participants were asked to provide a verbal briefing to answer the written question before conducting any queries for information. There appeared to be three general levels of prior knowledge of the Ariane 501 scenario (Table 18):

1. No prior knowledge of the scenario.
2. Knowledge that the new rocket launcher being developed by the European Space Agency is the Ariane 5 but no details regarding the cause of the 501 failure.
3. Knowledge that the incident was the Ariane 501 rocket launch failure and some details regarding the incident.

Although the group that used high profit documents as keys had a slightly higher number of participants who had prior knowledge of the scenario, the difference is not significant using the Wilcoxon-Mann-Whitney non-parametric test (p = .24).

Table 18. Comparison of Prior Knowledge of Scenario

Participants whose key documents were not high profit documents			Participants whose key documents were high profit documents		
Study Participant	Prior Knowledge of Scenario	Code	Study Participant	Prior Knowledge of Scenario	Code
3	____	1	2	Ariane 5	2
6	Ariane 5	2	4	Ariane 5, reused guidance system, software error	3
8	Ariane, assume qualification launch, propulsion-related problem, satellites destroyed, and insurance rates went up	3	5	___	1
9	___	1	7	Ariane 5, reused software in inertial reference system, data overload, design flaw, inadequate testing	3
	Average:	1.75		Average:	2.25

D3: Did the group that had some high profit documents as their key documents have more terms in their queries that came from sources other than the written question?

The written question given to the study participants was:
In 1996, the European Space Agency lost a satellite during the first qualification launch of a new rocket design. Give a short briefing about the basic facts of the incident: when it was, why it occurred, and what the immediate impacts were?

The query terms in the final "home" query were classified based on where the terms came from in order to see if the participants were able to take advantage of prior knowledge of the scenario to form better queries (Table 19):

- words exactly taken from the written question,
- synonyms of terms in the written question,
- terms that did not come from any of the concepts in the question or documents that were read, and
- terms that came from documents that were read.

These categories are similar to the ones used in Spink and Saracevic (1997), which were the users' written question statements, the users' domain knowledge, terms extracted from retrieved items as relevance feedback, a database thesaurus, and terms derived from intermediaries.

Table 19. Classification of Where the Query Terms Came From

Coded Category	Description
QUES	Exact word from the question
SYN	A synonym to a word in the question
SUBJ	From the study participant's knowledge
READ	From reading documents

The data is summarized in Table 20. There does not appear to be important differences between the two groups in terms of where the query terms came from. On the individual participant level, however, the data is suggestive. As might be expected, the three study participants (3, 5, 9) who had no prior knowledge of the scenario relied heavily on the question for search terms. All of the participants who knew that Ariane was the name of the rocket series that the European Space Agency built took advantage of that information except participant 7 (2, 4, 6, 8). Participant 7 was one of the three participants who had a relatively deep knowledge of the scenario in advance. He took advantage of that knowledge in formulating a query that was based upon knowing the cause of the failure: a software failure related to the guidance platform. In so doing, he was likely to select mostly documents that were aware of what the cause was, and therefore avoiding early documents that were missing much of the data.

Table 20. Comparison of Where Query Terms Came From

Participants whose key documents were not high profit documents				Participants whose key documents were high profit documents			
Subj	Final "Home" Query	Prior Knowledge of Scenario	Coded Queries	Subj	Final "Home" Query	Prior Knowledge of Scenario	Coded Queries
3	(europe 1996) & ((launch failure):%2)	1	QUES, QUES, QUES, SYN	2	(esa & ariane*) & (failure)	2	QUES, SUBJ, SYN
6	(1996 & Ariane) & (destr* | explo*) & (fail*)	2	QUES, SUBJ, SYN, SYN, SYN	4	(european space agency):%3 & ariane & failure & (laucher | rocket))	3	QUES, QUES, QUES, SUBJ, SYN, SUBJ, QUES
8	((ariane & 5):%2) & (launch & failure)	3	SUBJ, SUBJ, QUES, SYN	5	(ESA | (european & space & agency)) > (19960601) Infodate	1	QUES, QUES, QUES, QUES, READ
9	1996 & European Space Agency & satellite & lost & rocket	1	QUES, QUES, QUES, QUES, QUES, QUES, QUES	7	software & guidance	3	SUBJ, SUBJ
		Totals:	QUES(12), SYN(5), SUBJ(3)			Totals:	QUES(9), SYN(2), SUBJ(5), READ(1)

D4: Did the group that had some high profit documents as their key documents have more prior knowledge of the software used in the study than the other did?

As can be seen in Table 21, none of the study participants were familiar with the current version of the software used in the study. Two of the participants (2 and 8) had used a previous version somewhat, but not as their main software tool. It is therefore unlikely that familiarity with the tool explains why some participants located high profit documents in the database and others did not. In addition, only limited portions of the software tool were used to provide a baseline environment of keyword searching and browsing by dates and titles that are common to many applications used by intelligence analysts. The participants appeared to learn the features quickly and asked an intermediary that was provided for all the participants any specific questions on formulating queries that were not answered by the written "cheat sheet" that was provided.

Table 21. Comparison of Prior Knowledge of Software Used in the Study

Participants whose key documents were not high profit documents			Participants whose key documents were high profit documents		
Study Participant	Prior Knowledge of Software in Study	Code	Study Participant	Prior Knowledge of Software in Study	Code
3	Never used it	1	2	Used to use software in study (but not now)	2
6	Never used it	1	4	Never used it	1
8	Used a previous version a few times	2	5	Never used it	1
9	Never used it	1	7	Never used it	1
	Average:	1.25		Average:	1.25

To summarize, the searching process employed by all of the study participants was essentially one of continuously narrowing in on a small set of documents. This narrowing in process left the participants vulnerable to missing the best documents available in the database. The baseline tools of keyword search and browsing by dates and titles were not particularly helpful in identifying the high profit documents or visualizing how the samples related to what was available. Casting a wider net by opening more documents was correlated with finding more of the high profit documents. The participants generally seemed able to recognize the high profit documents once they found them and then relied on them heavily in their analysis products.

If, in fact, as is supported by the lack of noticeable differences in searching strategies and domain expertise, the most likely explanation for the difference between the group of study participants who relied on high profit documents vs. those who relied on lower quality documents is the amount of time that they spent and the number of documents that they opened, then this indicates that one of the ways, given a baseline electronic toolset of keyword querying and browsing by dates and titles, to find the high profit documents in the database might be to cast a wider net by sampling more, either by performing more queries or by opening up more documents. Support tools such as "agents" that remind or critique analysts to be broader in their sampling strategies might be helpful. Given the increasing organizational pressures to do analyses more efficiently, however, these types of support tools might be ineffective because analysts might not have the time to follow the suggestions. These results suggest that analysts may become more vulnerable to missing high profit documents if they are under tight deadlines and are not better supported in locating the best documents in the data set quickly.

5.2.3 Impact of Basing Analyses on High Profit Documents

One of the main findings of this study is that all of the study participants missed some of the high profit documents available in the database and that half of the participants based their analyses on documents that were not classified as high profit documents. Although analysts in past interviews have described that they consider it very important to have high-quality documents to perform their analyses, it is possible that they have developed expert strategies that allow them to use converging information from lower quality sources in such a way as to perform well in spite of having lower quality information. Therefore, an important question is whether or not the group that treated the high profit documents as key documents performed better than the group that did not use the high profit documents as their key documents.

To this end, the study participants' verbal briefings were coded on 20 items relating to the Ariane 501 scenario as accurate, vague, inaccurate, and no information[3] (see Section 4.6 for the items that were coded). Although these items were not at the same level of importance or detail, they were each treated as a single item because of the difficulties in assigning differential weights. Therefore, these codes are not to be viewed as "grades," but as a means for comparing the relative performance of the two groups.

It appears that there are differences in performance between the participants who relied upon the high profit documents and the participants who did not. As would be expected if high profit documents have fewer inaccuracies, the participants whose key

[3] Intercoder reliability by two simultaneous coders (Janet Reynolds and Emily Patterson) was 84% for the eight study participants (agreed on 134 items out of 160). The discrepancies were resolved by discussion and both coders agreed to the final codes.

documents were not high profit documents had more inaccurate statements in their verbal briefings than the participants who had some of their key documents be high profit documents (6 vs. 0, $p = .03$). Note that this difference cannot be explained by one group of participants having more thorough analyses, increasing the likelihood of inaccurate statements, because the participants in the two groups were similar in how many items they included in their briefings.

Table 22. Inaccurate Statements in Verbal Briefings

Participants whose key documents were not high profit documents

Participant	Inaccurate Item
3	the economic loss of the Cluster satellite payload would be recovered by insurance when the payload was not insured
3	after 44 seconds there was a software problem and the rocket blew up when the software problem actually happened at 36.7 seconds
6	the delay to the 502 launch was about 6 months when it was over a year
6	said that it was guidance data at the wrong altitude instead of diagnostic data that was interpreted as guidance data because said the inertial reference system reset instead of shut down
6	the Cluster satellite program was cancelled when it was later fully reinstated
9	the cause of the failure was due to a mechanical problem when it was due to a software problem

Participants whose key documents were high profit documents

Participant	Inaccurate Item
7	the Cluster satellite program was cancelled when it was later fully reinstated

One potential explanation for this difference between these two groups could simply be that the group that made more inaccurate statements just covered more information in the verbal briefings in general. As can be seen in Table 23, this is not the case. The differences between the two groups on the number of items that were covered in the verbal briefings are not significant. Even if they were, they would actually be different in the other direction – the participants in the group that relied on the high profit documents would actually have covered more information and made fewer inaccurate statements in their briefings.

Another potential explanation would be that the participants in the group that made fewer inaccurate statements were more experienced. This is not the case, however, as there are no significant differences in years of experience between the two groups (11 years vs. 10.5 years).

Another potential explanation would be that the participants in the group that made fewer inaccurate statements had more prior knowledge of the incident. Although participants 4 and 7 did have some prior knowledge of the incident, so did participant 8

who was in the other group, and participant 7 made one of the inaccurate statements. Also, none of their verbal briefings before the analysis began were anywhere near as detailed as the final briefings.

Therefore, although this is only one measure of performance and a rather small study, the data suggests, as would be expected, that helping intelligence analysts to efficiently locate high profit documents would allow them to make fewer inaccurate statements in their briefings. Some of the qualities of high profit documents could potentially be recognized by machine processing, such as an article over one thousand words that is not translated or an abstract from a pre-defined source such as Aviation Week, and at least several months after the event of interest. A "recommender" system that uses a combination of these machine-recognizable attributes, which would hopefully be observable and redirectable by the user, to recommend a set of documents to review might be very useful and relatively easy to implement with current technology.

Interestingly, although there were only a small number of participants in this study, there is little evidence to suggest that some participants had better query formulations than others. This raises the question of whether all of the participants were at approximately the same level in terms of performance in information retrieval or if advanced information retrieval techniques would not help to locate high quality documents in a database that is mostly "on topic." Two participants (3, 9) had noticeably fewer high profit documents returned by their query. However, both of these participants had the fewest number of documents overall in their query results, so it is not surprising that they had correspondingly fewer high profit documents. Regarding query term selection, participant 9 could have used "los*" instead of "lost" in his query formulation, which is generally believed to be a better formulation because it is a truncation which does not limit retrieval to a specific verb tense, this would have actually increased the "noise" of the responses in that it would return hits based on words like "loser" which contain that root. Also, several of the participants whose key documents were not high profit documents (3, 6, 9) used "1996" in their query terms, which weighted 1996 documents more heavily[4] but actually did not filter out later documents that contained important updates because they sometimes referred back to the date of the original disrupting event.

[4] Participants 3, 6, and 9 used 1996 in their query formulation and had 14%, 27%, and 24% of the returned documents from their home query published after 1996. In comparison, participants 2, 4, 7, and 8 had 43%, 38%, 33%, and 44% of their documents published after 1996. This difference is significant at the $p = 0.005$ level based on a parametric one-tailed t-test. The dates of participant 5's documents were not analyzed because the large number of documents in his home query renders the analysis extremely time-intensive.

Table 23. Summary of Types of Statements in Verbal Briefings

Participants whose key documents were not high profit documents

Participant	Accurate	Vague	Inaccurate	Nothing
3	5	2	2	11
6	11	1	3	5
8	9	0	0	11
9	5	3	1	11
Average:	7.5	1.5	1.5*	9.5

Participants whose key documents were high profit documents

Participant	Accurate	Vague	Inaccurate	Nothing
2	5	2	0	13
4	11	2	0	7
5	12	3	0	5
7	8	1	0	11
Average:	11	2	0*	6.75

* significant difference using Wilcoxon-Mann-Whitney Non-Parametric test

5.3 Findings in the Context of the Information Retrieval Literature

The findings of the previous section could be viewed as patterns in information seeking under data overload conditions, which is obviously related to findings and concepts from the information retrieval literature. These relationships will be discussed in this section.

First, it is important to highlight that this simulation study differs in several fundamental ways from traditional studies in information retrieval because the main conceptual focus for the study design and analysis was the process that expert practitioners use to find the significance of data in a vast data field in a specific complex, event-driven domain. In contrast, the historical focus of an information retrieval study would be on the performance of professional search intermediaries who employ strategies that apply across multiple domain end users who come to them for help in finding "relevant" information to a question that they would like to answer. In other words, the conceptual framework driving the study design and analysis was looking at how people determine the significance of data, which inherently emphasizes the context-bound and process-oriented nature of the data analysis, instead of traditional information retrieval context-free and product-oriented measures such as precision and recall[5].

[5] Precision is the percent of relevant documents returned in a query. Recall is the percent of the relevant documents returned in a query in relation to the number of available documents in the database.

Stemming from this foundational difference in focus, there are several differences in the details of this study design as compared with a traditional information retrieval study. First, the participants in this study performed the searches themselves (i.e., search intermediaries did not help them with this task), which is the increasingly common situation in their work environment even though several professional search intermediaries are employed at their site. As a result, these participants are often impacted by their ability to retrieve information but in general are much more concerned with their ability to do other tasks (i.e., inferential analysis) than to hone their information retrieval skills.

Second, the study participants did not stop once searches were completed – the searching and analysis was integrated in the simulated task. The historical model of information retrieval implicitly suggests that the retrieval of information and the end-user tasks are sequential (although see Bates, 1989, for a "berrypicking" model of information retrieval that emphasizes a more integrated view of searching and end-user tasks, see Belkin, 1993, for a model of information retrieval that emphasizes interaction of the user with texts and ill-defined search needs, and see Burnett and McKinley, 1998, for a model of information seeking that emphasizes interactive processes). With end-user searching, and particularly end-user searching that is not charged on the basis of "connect time," the two elements are much more interrelated. It was not an uncommon situation that a study participant read an article which triggered a question that was answered by browsing for the information before returning to the ongoing analysis.

Third, the end-user is specifically defined as a professional intelligence analyst rather than an "everyday user" as would be more typical of a library situation. Therefore we can take advantage of an understanding of the demands of the domain in order to have more predictive power about how the users will go about their tasks.

Fourth, in traditional information retrieval studies, the primary measurement standard is precision and recall of the final honed query results, as judged by the end users. In this case, the main findings were patterns in the processes that were used to arrive at analytic products and how these patterns related to inaccurate statements in the verbal briefings. As a means to this end, precision and recall measures were taken as complementary information to provide converging evidence, although with some modifications. Rather than rate all of the documents in the queries as relevant or irrelevant, only high profit documents were investigated (see Mizarro, 1997, for an overview of the factors inherent in relevance definitions including topicality and utility; cf. Blair and Maron, 1985, for their distinctions between vital, relevant, partially relevant, and not relevant documents in legal analysis). High profit documents were used because they were unambiguously relevant, tractable to identify in the analysis process for each query iteration, and because preliminary explorations on precision and recall data on relevant documents did not yield informative patterns. In addition, the judgments of whether documents were high profit or not were made by the

investigators rather than each participant in order to facilitate comparison across the study participants.

Given these differences between the historical model of information retrieval studies and this simulation study, it is tempting to say that concepts and findings from the information retrieval literature are inapplicable. This is not the case, however. Many of the concepts from the historical information retrieval literature provided insight for directions to pursue in the data analysis, such as techniques for widening and narrowing searches that used to be used by professional search intermediaries on indexed (as opposed to unordered full-text) databases. In addition, the accepted model of an information retrieval study is evolving to become closer to the perspective taken in this study: from product-oriented to more process-oriented analyses (Borgman, Hirsh, and Hiller, 1996), from generic end-users interacting with a search intermediary to more professional end-users with domain but not searching expertise, and from sequential tasks of information retrieval and then analysis to integrated searching and analysis (Bates, 1989). Findings from this growing literature provide additional insight on the main findings from this study.

The first main finding from this study related to information seeking is that the study participants used relatively primitive search tactics (Bates, 1979a, 1979b, 1992) as compared to professional search intermediaries. Often, they added words that were either from the question or synonyms of words in the question with a Boolean AND until they reached a number of hits that they felt could be browsed. This is a rather primitive search strategy where the emphasis is on quickly getting to a number of documents that could be browsed as opposed to getting a good, precise, or exhaustive set of information. A somewhat similar strategy discussed in the information retrieval literature is called successive fractionation by Harter (1986, p. 177), which was adapted from Meadow and Cochrane (1981). This strategy involves the successive addition of "facets" to reduce the number of documents returned by a query. With this strategy, a facet that is expected to have the lowest postings or be the most specific is entered first and then additional facets are combined orthogonally with an "AND" combination until a desired number of hits is reached. However, despite the similarity of refining queries by ANDing terms to reduce the number of hits, most of the study participants could not really be described as having used this strategy. Most of the participants did not appear to deeply understand the need for facets in searching. For example, it would be expected that most of the facets would include synonyms within the facets to reduce the chances that important information would not be returned, but in only two cases was a synonym used in an "OR" combination of terms. In addition, sometimes the terms that were ANDed could actually be viewed as being conceptually a part of an existing facet, such as when fail* was ANDed with a query combination that already included (destr* OR explo*).

The finding that the study participants used relatively primitive search strategies is not surprising in the context of the growing information retrieval literature on other domain expert end-users who conduct their own searches but are not search experts (e.g., securities analysts – Kuhlthau, 1999; Baldwin & Rice, 1997; legal analysts – Blair & Maron, 1985; Blair, 1996; Yuan, 1997; health care personnel – Abate, Shumway, Jacknowitz, & Sinclair, 1989; Hersh & Hickam, 1998; Sackett & Straus, 1998; Brown & Agrawala, 1974; Sewell & Bevan, 1976; Leipzig, Kozak, & Schwartz, 1983; energy users – Walton & Dedert, 1983; Case, Borgman, & Meadow, 1986; journalists – Sievert & Glazier, 1990 and academics – Bates, Wilde & Siegfried, 1993; Siegfried, Bates & Wilde, 1993). Across these studies, there is converging evidence that after a short amount of training and/or time on an information retrieval system, users can conduct simple searches. Over time, many of these users do not learn more sophisticated search techniques but instead remain "perpetual search novices."

One caution in interpreting the results about the study participants using relatively primitive search strategies is that this does not necessarily imply that all of the study participants should be using professional search intermediaries to perform all of their searches for them. It is a consistent finding in information retrieval studies that both domain knowledge and search expertise are important in seeking information, and that one is not significantly more important than the other (Saracevic et al., 1988; Wildemuth, de Bliek, Friedman, & File, 1995; Hsieh-Yee, 1993; Fenichel, 1980; Spink & Saracevic, 1997). Also, these two sources of knowledge are only partially decomposable, and may in fact interact in important ways (Shute & Smith, 1992).

The second main finding related to information seeking is that all of the search tactics used by the study participants were narrowing tactics (see Bates, 1979a, 1992; Wilson, 1992 for search tactics that can be used to narrow the number of documents that are returned by a query). This observation suggests that, under data overload conditions, narrowing is a predictable coping strategy. Others have observed this propensity to narrow returned sets based on the number of hits almost indiscriminately when the data sets are large (Blair, 1980 observed this pattern with users of indexed databases and explained the pattern as a result of overestimating the probability of conjunctive sets; Olsen, Sochats, and Williams, 1998 discuss the need to support narrowing tactics that are orthogonal to keyword terms such as document attributes because of the potential of removing interesting documents in data overload conditions through the overuse of adding keyword terms to narrow document sets). Although effective in making the amount of data to be browsed manageable, this coping strategy leaves analysts vulnerable to missing critical information, particularly since the impact of the different narrowing tactics to the relationship of the information that is sampled to what is available is opaque to the end-user.

Another reason why the finding that all of the participants followed the same basic narrowing pattern in their information seeking behavior is significant is that there are few studies in the information retrieval literature where a single variable dominates over the often great variability stemming from other environmental, individual, organizational, domain and search factors (e.g., Bellardo, 1985; Saracevic et al., 1988 discovered great individual variability on many searching metrics; see Fidel and Soergel, 1983 for an extensive lists of the variables that affect on-line searching). This is partly because of the emphasis in traditional information retrieval studies on the "everyday" user as opposed to domain specialists. The results of this study are that there was a consistent pattern across all of the study participants, suggesting that the desire to reduce the tremendous amounts of available data returned from a particular query dominates over the variability stemming from the many other possible factors that influence information seeking.

Blair and Maron (1985, 1990; Blair, 1996) conducted a landmark study with legal analysts using a full-text information retrieval system where the findings were that participants were not well-calibrated to the amount of information that was missed in the search (although see Salton, 1986, for an alternative interpretation of the study findings). The claim from this study is that the participants believed that they stopped searching when they had retrieved about 75% of the information in the database that was relevant to the case that they were investigating, when in fact on average they had retrieved about 20%. The results from this study highlight that sampling from large data sets in ways that leave the relationship of the sample to what is available opaque creates a situation where analysts can easily be miscalibrated as to how much critical information was missed. The significance of these findings in relation to this study is that not only were participants observed to miss critical documents, but it is likely that they did not know how likely they were to miss them, and so would have difficulty providing an accurate assessment of their confidence levels in their analyses, which is critical to the policy makers' decisions of how or whether or not to act on the implications of the analysis.

Discussions and findings related to support tools in the information retrieval literature also offer some insight into the tradeoff dimensions in designing support tools to aid in information seeking. First, there is a tradeoff dimension between precision and recall. Although some domain users might consistently prefer one end of the tradeoff dimension at the cost of the other (e.g., legal analysts might always prefer exhaustive searches over searches where a larger percentage of the returned information is relevant), in general it is difficult to predict whether an individual at a certain time for a certain situation will want a more precise or more exhaustive search, given that with more precise searches important information might be missed but with more exhaustive searches, more of the information is "noise." Therefore, query expansion aids that attempt to reduce the amount of information that is missed might actually work against the desired tradeoff that the user would like to make.

Second, there are findings in the information retrieval literature that help to calibrate the cognitive engineering research base on automation surprises and designing cooperative human-machine architectures to the task of information retrieval. For example, Koenemann and Belkin (1996) demonstrated that study participants using systems which provided interactive relevance feedback performed better than without it, and that the feedback mechanism that provided the most control over the query term selection was more preferred and gave higher performance. Similarly, Salton (1968, 1986) has demonstrated that machine-initiated modifications of query formulations with iterative feedback by the user on output sets is a more powerful method than "the machine does it alone" query expansion techniques. In contrast, Beaulieu and Jones, 1998 found in a comparison study that automated query aids work better than interactive query aids in finding more new relevant documents. Their explanation is that the particular interface design increased the cognitive loading on the human to where the humans declined to use the additional capabilities. These findings make it clear that the following elements of the cognitive engineering research base apply: 1) there are "cooperative" burdens that are introduced by machine agents that need to balanced against the potential benefits of the system, 2) user-initiated vs. machine-initiated strategies need to be balanced against the predicted brittleness of the machine processing, and 3) automated aids need to be directable and observable in order to avoid surprising their human partner by making unexpected actions (Sarter et al., 1997).

5.4 Patterns in the Sources of Inaccurate Statements

There were two main sets of patterns that were investigated during the process tracing analysis of the study data. The first set of patterns related to information sampling, as described above. The second path involved identifying erroneous statements in the verbal briefings and during the analysis process and then tracing the sources of these erroneous statements. By tracing why these inaccurate statements were made with the process tracing methodology, three sources of inaccurate statements were identified that provide insight into the cognitive demands of inferential analysis under data overload: 1) participants relying on assumptions that did not apply, 2) incorporating information that was inaccurate, and 3) relying on outdated information.

The inaccurate statements made by the study participants are displayed against the causes and impacts of the Ariane 501 accident in Figures 19 and 20. The majority of the inaccurate statements relating to the cause of the failure resulted from participants incorporating inaccurate information in the reports that they read about the technical details, either because they did not recognize that information was conflicting in the documents that they read or because they did not open documents that contained the conflicting information. The majority of the inaccurate statements regarding the impacts of the failure, on the other hand, resulted from missing updates that overturned previous predictions of the impacts, either because they were not recognized in documents that were read because the participants' attention was focused elsewhere, or because the participants did not open documents that contained the updates.

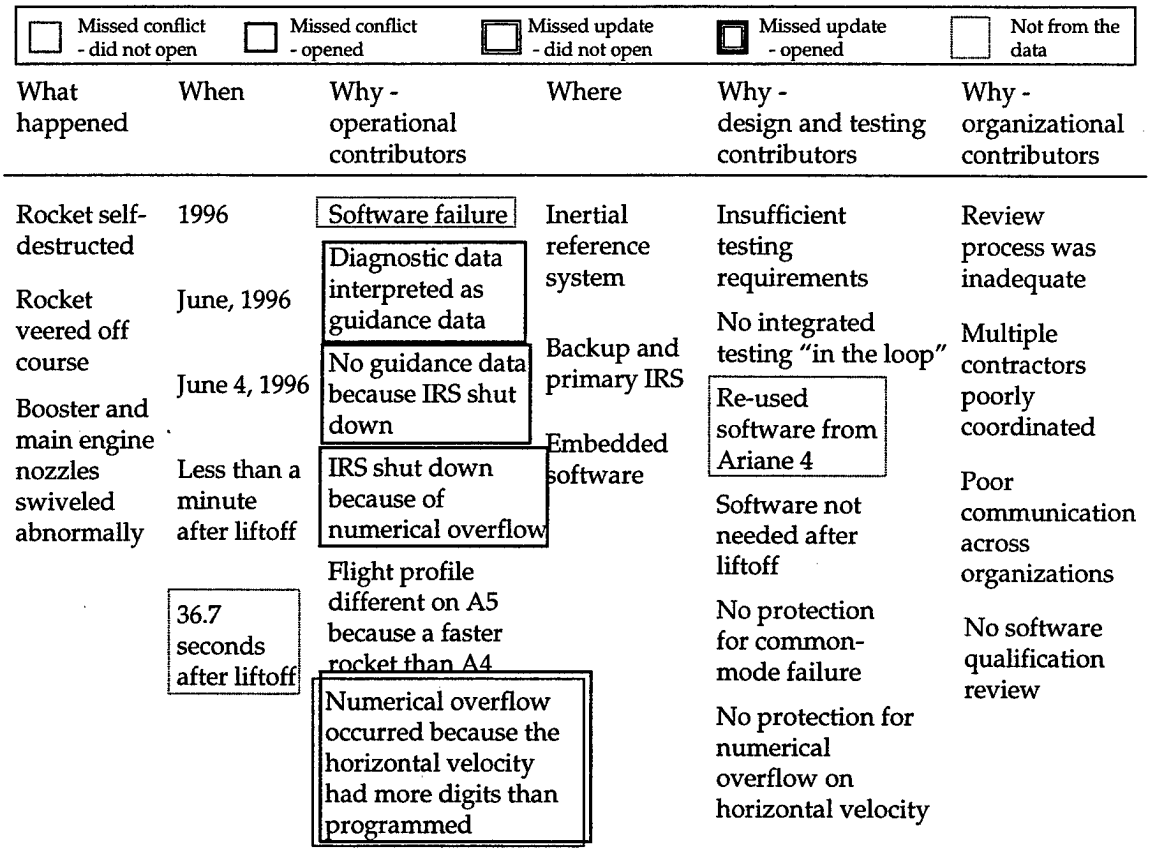


Figure 19. Sources of inaccurate statements for the cause of the failure.

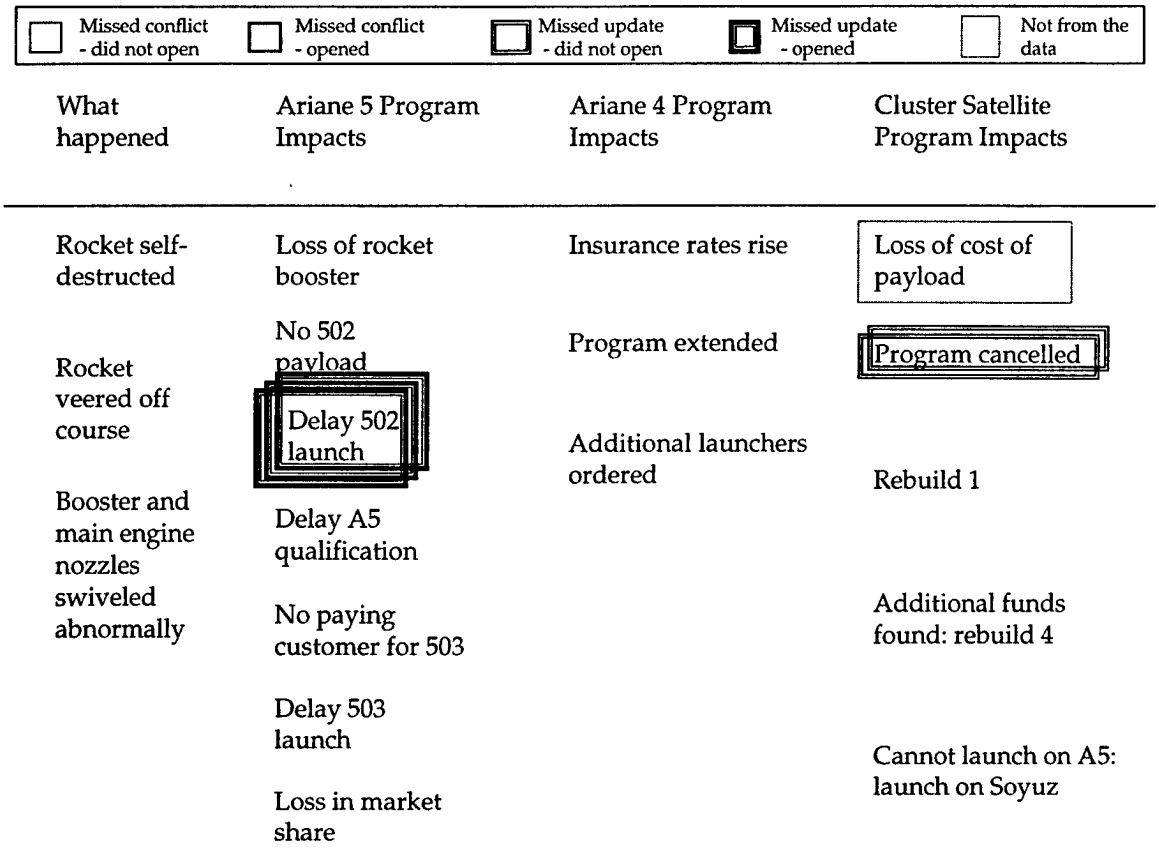


Figure 20. Sources of inaccurate statements for the impacts of 501 incident.

5.4.1 Relying on Assumptions That Did Not Apply

One source of inaccurate statements during the analysis process was, of course, the study participants themselves. There were several inaccurate statements made during the verbal briefings that did not come from any of the documents that were opened. For the majority of these cases, the participants appeared to be relying on assumptions to fill in gaps in the story that did not happen to apply in this case. For example, during his verbal briefing, one participant stated that the monetary loss of the Cluster satellite payload could be recovered by insurance. Although payloads are often insured, in this case the Cluster satellites were not because this was a scientific project under a tight budget.

In one case, although the source of the inaccurate statement is the same (the participant), the cause appeared to be due to forgetting rather than explicitly applying an assumption that did not apply. This case highlights a distinction similar to one that is often discussed in the human error literature between erroneous actions or statements that are intended as opposed to unintended, or slips (Norman, 1981). In this case, the participant stated that the explosion was at 44 seconds, a time that was not in any of the articles that he read. The articles that he read stated that the explosion was at 30 seconds, 37 seconds and 41 seconds after liftoff. It is unlikely in this particular situation that 44 seconds was assumed from background knowledge or prior experience.

5.4.1.1 Example: Software design as the cause of the failure.

As an example that illustrates how relying on assumptions that do not apply can have important impacts on the quality of an analytic product, consider the case where study participant 9 described that the cause of the incident stemmed from a mechanical rather than a software failure. The participant briefed that: "It had a steering problem...related to the rocket steering propulsion system that became erratic at 37 seconds and eventually broke off." This statement implied that the problem was due to a (much more typical) mechanical failure, as opposed to a software failure. Figure 21 shows the information that was available to the participant in the articles that he looked at and what he verbalized while reading the articles relating to this item. The question of what caused the failure is perhaps the most important item to ensure is accurate, particularly since in this case the significance of the event was in the departure from the typical mechanical failure, and yet when we look at the process that was followed, it is easy to understand how this situation can occur when analysts use the normally useful heuristic of applying default assumptions.

Article Date/Content	*Participant's Response*
November 3, 1997: *a software failure caused the rocket to veer off course and fall apart*	→ nothing
June 5, 1996: *Ariane 5...veered off course...exhaust nozzles at the base of two boosters...swiveled abnormally after 37 seconds and broke off, triggering an on-board self-destruct mechanism*	→ "The reason being a lost guidance, because that's what it said here, veered off course, so apparently the failure was due to the lost guidance system (writes "lost guidance" in notes)...guidance again, going back up (in notes added "due to rocket steering problem (mechan) noz separated"). How here it says nozzle broke off...so this was the first indication of a problem. So we're fine up to 37 seconds. And it separated...let's see...lost guidance...so it was mechanical. All right, this is slowly starting to come together."

Figure 21. Process trace of cause of software failure.

Essentially, participant 9's verbal briefing was based upon the description in the June 5, 1996 article, which did not state that the failure was due to a software problem (because the cause was not yet known) and gives the impression that the cause was a more typical mechanical failure (although the reporter does not actually make that statement – all of the information given is accurate). This article was an early article in that it was only one day after the June 4 incident and therefore did not have the information that the cause was a software failure. Although the participant did read an article that described the 501 incident as due to a software failure, this conflict went unnoticed. One possibility as to why the conflict was not detected is because the subject was focusing on the main topic of that article when reading it, which was the description of the 502 launch, rather than the description of the 501 launch.

Interestingly, study participant 9 described his process as being the "quick and dirty" answer. He estimated a confidence level of 3 on a 1-5 scale because the information had not been verified. When asked why he did not verify the answer, he stated that it was a tradeoff of speed and accuracy. "[if] needing to get out fairly quick, I would stop here." If he wanted to raise his confidence in his analysis, then he stated that he would want 3 or 4 documents to verify the same thing. Although clearly this strategy would allow him to ensure that his default assumption was correct, under the condition of considerable time pressure, he did not consider it unreasonable to rely on these assumptions.

The participant was then asked to demonstrate what he would do to raise the confidence in his analysis. He read more articles and elaborated his notes, both handwritten and electronic. He drew a line in his handwritten notes to indicate a

dedicated area below the line where he drew arrows from information above the line from his "first pass" analysis that he considered verified. He described that, if the analysis was particularly important, he would print out all of the documents that he had read and mark in different colors information on a theme that was corroborated by a certain number of sources and information that conflicted. Although he did not finish this process in the remaining few hours because it was time-intensive and mentally tiring, he did catch his earlier inaccurate statement and revised his assessment of the cause of the failure to say that it was due to an integral failure brought on by internal software (Figure 22).

Article Date/Content	*Participant's Response*
November 3, 1997 (second time): *a software failure caused the rocket to veer off course and fall apart*	→ nothing
August 4, 1996: *failure was due to Ariane-5's "brain." It turned out that the computer software in the Ariane-5 was originally designed for the Ariane-4, a much slower rocket. Seconds after take-off, Ariane-5 reached a velocity that exceeded the "brain's" computing capacity. It lost all guidance and attitude information, and the on-board computer tried to supercede the software programme and activated the rocket's solid fuel propellant boosters*	"OK, this is the same one [the Ariane 501 launch]. This is after the fact. Uh oh. Remember I said how data changes? I'm looking...apparently it says a mechanical failure and then I come along. What's this say? Failure was due to the brain. It turned out that computer software which was → designed for 4, which is much slower. So it turns out now my analysis has changed. It now looks like it was an integral failure. Period. Brought on by internal software. So I'll qualify this (draws an arrow from previous note below a line and writes "#1435 wrong software used, software for AR4 used in AR5 launch"). That was the problem. Lost guidance. Launch software."
November 3, 1997 (third time): *a software failure caused the rocket to veer off course and fall apart*	"Software failure. It's a confirmation of the previous message saying it's a software failure... → or is this the same message? Yeah, yeah, that's where the highlighter"

Figure 22. Continuation of the process trace on cause of the failure.

Note that this example also illustrates how the information sampling strategies interact with the potential for inaccurate statements in the verbal briefings. First, study participant 9 missed some critical information, particularly during his initial attempt. None of the documents returned by his "home" query, and therefore also none of the documents that were opened, read, or treated as "key" to his analysis were documents that we identified as high-profit. Therefore the main document relied on during the first pass analysis was published the day after the failure, before the cause had been identified, and none of the documents that were opened had detailed descriptions of the launch sequence or sophisticated analyses of the potential impacts of the launch failure.

Second, although the main cause of this inaccurate statement was relying on a default assumption that did not apply, at least part of the cause of the inaccurate statement was due to breakdowns in the process of corroborating information and checking for conflicts in data. The study participant missed the conflict in the two reports that he opened that described the failure in his first pass analysis, most likely because he was focusing on the main topic of the first report which was about the next flight in the series, 502, rather than the short description of the cause of the 501 failure, when he was reading it. Since he did not verbalize anything while reading the report that stated the cause was due to a software failure, it is possible that he did not read that part of the report. It is clear that when he elaborated his analysis by reading more documents and actively verifying the information, he recognized the conflict: "so it turns out that my analysis has changed." Finally, another breakdown occurred during his analysis process: losing track of whether "converging" information actually stemmed from the same source (note that analysts have referred to this problem in interviews as "creeping validity"). While going through his second pass, the analyst stated that he could not remember if he had previously read the November 7 report or if it was a new article. It was actually the third time that he had read the report.

Overall, the participant could be viewed to have closed the analysis process prematurely, both in terms of sampling and verifying information. He only opened documents from one query during both the first-pass and second-pass analysis, accepted the first hypothesis that he read as the explanation for the main cause of the failure without corroborating it with independent information, did not read very many documents, did not search for any of the "high profit" documents in the database, and did not specifically search for updated information that would render his current analysis incorrect Although the participant stated that he was aware that he was vulnerable to making inaccurate statements because he had not corroborated the information from 3 - 4 documents as he would like to do and so therefore was less confident in his analysis, he stopped and provided a verbal briefing in order to finish the analysis more quickly. Additionally, he stated that if he were under high workload and/or did not judge the analysis to be a high priority, he might stop there in an actual analysis situation.

5.4.1.2 Impact of relying upon assumptions that did not apply.

Partly as a result of relying on assumptions that did not apply:

- study participant 3 stated that the economic loss of the Cluster satellite payload would be recovered by the insurance,
- study participant 9 reported that the cause of the failure was due to a mechanical problem when it was due to a software problem, and a
- study participant 6 stated that they reused the inertial guidance system from the Ariane 4 when it was only the embedded software that was reused, not the whole inertial guidance system.

And apparently as a result of a "slip" where the participant mis-remembered information that he had read:

- study participant 3 stated that after 44 seconds there was a software problem and the rocket blew up when the software problem actually happened at 36.7 seconds.

In summary, one source of inaccurate statements was the application of unverified assumptions to fill in gaps in the verbal briefings. Relying on assumptions is clearly a heuristic that can be applied under time pressure as a coping strategy. Although in the cases discussed above, relying on assumptions led to inaccurate statements, in other cases it did not. For example, in one case, participant 2 used the assumption that the Ariane 5 rocket would eventually replace the Ariane 4 as the standard launch vehicle in his estimation of the impacts of the failure, several participants used the assumption that failures in the rocket launch industry are not uncommon, particularly during maiden flights, in their assessments of the impacts, and several participants assumed that the main competitors to Ariane would pick up some of the market share that Ariane might have had. In addition to filling in gaps in knowledge, default assumptions also proved valuable in knowing what information to seek during the analysis process. For example, participant 4 stated that he assumed that satellites were on the flight and then looked explicitly to see if there were.

5.4.2 Incorporating Information That Was Inaccurate

The second main source of inaccurate statements was from inaccurate descriptions in documents in the database. Intelligence analysts clearly view the elimination of inaccuracies by finding converging evidence across independent sources as a major component of the value of an analytic product. The participants described and employed a variety of strategies for tracking and resolving descriptions that conflicted in order to reduce their vulnerability to incorporating inaccurate information. Partly because this cognitively difficult process of corroborating information and resolving conflicting information was unsupported by the tools that they were provided, nearly every participant experienced some breakdowns in this process. Breakdowns included failing to corroborate information, missing conflicts in documents that were opened, forgetting how many corroborating and conflicting descriptions had been read from independent sources, forgetting the information sources, and treating descriptions that stemmed from the same source as corroborating.

5.4.2.1 Sources of misunderstandings by report writers.

The inaccuracies in the data set appear to have occurred for a variety of reasons. First, a certain level of technical expertise was required to understand the intricacies of why the Ariane 501 rocket exploded, so sometimes there were simply misunderstandings from a lack of technical expertise on the part of the reporter about the cause. Secondly, different reports had different levels of access to the "raw data," such as eyewitnesses, members of the Inquiry Board, engineers who designed the rocket, and machine-produced data such as telemetry, video, audio, and photographs. Reports that were more "distanced" from the data, such as reports that were written from other reports, had inaccuracies that were introduced as a result. Similarly, reports that were translated from a foreign language sometimes introduced inaccuracies during the translation. Finally, note that although in this scenario there was a low probability that reporters were trying to actively deceive analysts, in other scenarios that would be a very important source of inaccuracies.

An important distinction in the types of inaccuracies introduced by report writers should be noted. Some information was clearly either accurate or inaccurate – either the reporters got it "right" or they got it "wrong," such as the date that the incident occurred. In other cases, however, the "accuracy" of the information was inherently contestable, such as the predicted impact of the Ariane 501 failure on the rocket launcher market over the next five years. However, even in these cases, certain interpretations could be judged weaker than others. For example, some of the discrepancies were predictable based on the expected interpretive stance of the source given their goals and associated biases. Specifically, articles that were written to generate public interest in the story (i.e., sensationalistic articles) tended to be overly pessimistic about the impact of the Ariane 501 failure. Similarly, articles from sources that were competitors to Arianespace tended to overemphasize the impacts and articles from the European Space Agency (ESA) and at the other extreme Arianespace tended to downplay the impacts. For example, compare the understated (June 4) description of the Ariane 501 failure in the ESA 19-96 press release:

> The first Ariane-5 flight did not result in validation of Europe's new launcher... A second test already scheduled under the development plan will take place in a few months' time... the skills of all the teams involved in the programme, coupled with the determination and solidarity of all the political, technical and industrial authorities, make us confident of a successful outcome...

with the more sensationalistic (June 5) Washington Post description:

> The flight lasted just a few seconds...blown up in midair...ending with smoke, embarrassment and $ 500 million worth of vaporized satellites... the launch may have shifted the balance in the highly competitive international market for space launches.

5.4.2.2 Inaccurate descriptions in documents about the Ariane 501 Scenario.

In the Ariane 501 scenario, there were several examples of inaccuracies in the available reports about past events due to a lack of technical expertise, distance from the original data, and/or translating from a foreign language that were clearly objectively inaccurate:

- the exact time that the rocket exploded – this was difficult for reporters to determine partly because when the rocket actually lifted off and self-destructed is somewhat ambiguous. Some of the reported times in the data set included 30, 36, 36.7, 37, 40, 41, 45, 61, and 66 seconds after liftoff.
- the cause of the 501 failure – there were several non-technical articles that were presumably written in a way to make the failure easier to understand for readers who were not experts in the technologies. In these cases, descriptions of the failure used analogies such as "a gas leak" or "the brain died" that were inaccurate generalizations or overly vague. In addition, there were articles that attempted to describe the details in technical terms, but the details were clearly misunderstood by the reporter because there were inaccuracies that were not due to information coming in over time. For example, descriptions about why the numeric overflow occurred included the number being too long, too high, too large, higher than a coded limit, higher than the rate of change that could be handled by the processors, and greater than a memory buffer when the accurate description was that the number of digits for the horizontal velocity was greater than the number of digits assigned to the variable. Similarly, there were questions about whether the excessive pitch-over of the rocket was due to altitude information that was not self-generated, a lack of altitude information, a "reset" on altitude information to what would be on the ground, or diagnostic information that was incorrectly interpreted as altitude information.

Interpretations that conflicted with each other but that were more difficult to judge in terms of accuracy because the "right" answer was inherently contestable included:

- the cost of the explosion – judgments about what the actual costs of the explosion were varied widely. Justifications were based not only on what factors were considered relevant but also what amount should be assigned to each factor. The cost of the payload was judged to be $500 million, yet a new payload was constructed much more cheaply than the original one because there were spare parts and they had learned much during the process of constructing the first payload, so the replacement value was less. The launcher was estimated to cost seven billion dollars, yet launchers were not reusable and so in a sense were always lost during a flight. One estimate of the cost to the program to fix the problem was 2-4% of the eight billion dollar investment in Ariane 5, yet they might have discovered other problems during the process that would save money in the future by being discovered then. The delay in the Ariane 5 qualification program had an associated opportunity cost in terms of not being able to provide customers with the ability to launch payloads, although many of the customers could launch on the Ariane 4

rocket instead. Twenty new Ariane 4 launchers were ordered after the 501 failure for US $1.5 billion, but it is not clear how many of them might have been ordered to meet market demand independently of the delay to the Ariane 5 program. Finally, there was the question of whether or not Arianespace would lose market share in relation to what they would have had if 501 had been successful. It is possible that other competitors might have earned some of Arianespace's business, but that depended on the size of the payloads in relation to the available launchers, whether or not the customers changed launch companies as a result of the failure or other considerations, and how the competitors were introducing new features on their own launchers that would compete with the features on the Ariane 5 rocket design.

- the delay to the qualification of the Ariane 5 rocket design – although many of the differences in predictions of delay to the 502 launch were due to updates over time, there were also discrepancies regarding how long 502 would be delayed for time periods where reporters had the same information set. Predictions from ESA tended to be overly optimistic whereas predictions from competitors tended to be more pessimistic. Similarly, predictions from ESA about the possibility of finding a paying customer for the 503 flight were more optimistic than from other sources.

5.4.2.3 Example: Cause of abnormal rocket swiveling.

To illustrate some of the difficulties in the process of eliminating inaccuracies in descriptions stemming from reporters, consider the example of determining the cause for why the rocket swiveled abnormally. Interestingly, participants 6 and 7 both read the same two documents that contained discrepant descriptions but ended up with different outcomes in their verbal briefings (Figures 23 and 24).

Participant 6 based his analysis of why the rocket swiveled mainly on report 858, which described the cause as a reset of the inertial reference frame following a numeric overflow (Figure 23). As he read 858, he was thinking out loud about why the rocket swiveled based on what he was reading. Later, he read 1385, which had a contradictory description of why the rocket swiveled. At that point in time, however, it was the last document that he looked at, and he was focused on a different issue – why testing did not reveal the software error. He gave no evidence that he recognized the conflict. In fact, when asked how he knew when to stop the process, he explained: "It doesn't look like anybody will have any different opinions. From looking at the other titles, it looks like I won't come up with anything new."

Therefore, not only did this participant not explicitly conduct the step on this item of corroborating the information through an independent source; he also did not recognize a conflict in what he read. This indicates that recognizing conflicts is a non-trivial task. Direct attention must be given to interpreting that item of information, remembering what had been read in other articles, and recognizing that the descriptions are incompatible. In the electronic environment, this task is particularly challenging because only one report can be viewed at a time because of space limitations on the

computer screen. Furthermore, the participant was unaware of conflicts in data that he had read, and as well had no way to tell if there were conflicting descriptions in data that he had not looked at, or even in the reports that were not returned from his query but available in the database.

Figure 23. Participant 6's process trace on why the rocket swiveled.

Participant 6 Briefing: "that guidance system, the length of time that it operated, actually interfered with the inertial guidance system which took over after the launch and it confused...they confused each other and decided that they have to reset but by that time the rocket wasn't vertical anymore"

Article Date/Content	*Participant's Response*
July 5, 1996 (Report 858): *Ariane 5 lifts off much faster... information... exhausted the temporary memory (buffer) capacity...both systems simultaneously declared themselves to be in an irredemiable error situation and commenced a reset procedure...when the system was reset, the vehicle's position at that time...was adopted as the reference base* September 16, 1996 (Report 1385): *the active inertial reference system transmitted essentially diagnostic information to the launcher's main computer, where it was interpreted as flight data and used for flight control calculations*	"It's the same system as used on the Ariane 4, but the Ariane 5 takes off faster, much faster, than the Ariane 4. The two inertial guidance systems confused each other. They tried to reset at 37 seconds. It wasn't vertical anymore. It just totally lost its mind...so it couldn't figure out its direction." (talks about a different issue - how it could have been avoided through testing)

In contrast, participant 7 gave an incompatible explanation for the cause of the swiveling rocket as diagnostic information interpreted as command data (Figure 13). This was incompatible because participant 7's description said that there was no command data at all because the guidance platforms had shut down whereas participant 6's description said that there was command data, just that it was incorrect because the guidance platforms had been reset mid-flight.

Participant 7 recognized the conflict in the descriptions in documents 858 and 1440 and resolved it based on a judgment of source quality. He decided to base his analysis on the description in 1440 because it was later and therefore more likely to have all the information, not translated, and from a more authoritative source (analysts had described in interviews that Aviation Week was judged to a more credible source in general than FBIS). Note, however, that even though this was the accurate judgment to

make, he did not notice that another article corroborated the hypothesis that he selected, which would have made the judgment easier. This would have been particularly helpful in this case because, as he pointed out: "[The inaccurate description] sounds good." The description that was inaccurate was written in a way that sounded as if the reporter has sufficient technical expertise to understand the cause in detail. If he had only read article 858 and not found the conflicting descriptions, it is likely that he would have believed the inaccurate description.

Participant 7 Briefing: "numerical values beyond the programmed limits of the flight computer...the platforms initiated a diagnostic "reset" mode that fed incorrect values to the flight computer"

Article Date/Content	*Participant's Response*
September 16, 1996 (Report 1385): *the active inertial reference system transmitted essentially diagnostic information to the launcher's main computer, where it was interpreted as flight data and used for flight control calculations*	nothing
July 29, 1996 (Report 1440): *as a result of the double failure, the active IRS only transmitted diagnostic information to the booster's on-board computer, which was interpreted as flight data and used for flight control calculations*	"We know there was a problem because the guidance platforms shut down. After they shut down, the inertial reference system sent diagnostic information so they're designed to shut down when something goes wrong. Assuming the other system has taken over, it's sending diagnostic information so that the people on the ground can figure out what went wrong with it. Having them both shut down, the guidance computer is interpreting the diagnostic information as where it's at and instead of getting numbers, it's getting other things..."
July 5, 1996 (Report 858): *Ariane 5 lifts off much faster... information... exhausted the temporary memory (buffer) capacity...both systems simultaneously declared themselves to be in an irredemiable error situation and commenced a reset procedure...when the system was reset, the vehicle's position at that time...was adopted as the reference base*	"...In this article, it says when it shut down, it started a reset procedure. In the other article, it says diagnostic information. This article and the other one...are incompatible, inconsistent with each other...Of course messages that can't both be right happen all the time. I'm finding it hard to believe that the vehicle is going to fly without any inertial inputs whatsoever ...let's look at the source...FBIS report. Translated text...the other one was later also...it sounds good. If I had to guess, I would go with the other one.

Figure 24. Participant 7's process trace on why the rocket swiveled.

5.4.2.4 Impact of incorporating inaccurate information from documents.

Partly as a result of incorporating descriptions that were inaccurate because of misunderstandings by reporters:

- study participant 6 gave an inaccurate description of why the rocket unexpectedly swiveled (did not state that the diagnostic information was interpreted as guidance data, said that it was guidance data at the wrong altitude instead of no guidance data at all, and that the inertial reference system reset instead of shut down),
- study participant 7 verbalized at one time during the analysis process that he thought that the cause of the numeric overflow was because the velocity

information was coming at too fast of a rate when actually it was because the actual number had too many digits for the memory that had been assigned to it, and
- study participant 5 stated that the computer rejected a number because it was too large when actually it was that the number had too many digits for the memory that had been assigned to it.

The finding that people sometimes failed to recognize or actively seek out information that would have conflicted with a hypothesis that was believed to be true is not surprising in relation to the results from experimental psychology studies on confirmation bias and fixation effects. The phenomena of "confirmation bias" studied in experimental psychology (Tversky, 1982; see Klayman and Ha, 1987, for an alternative interpretation) is generally described as people having a tendency to check for evidence that confirms their beliefs rather than actively seeking out evidence that would disconfirm them. In some cases, the definition of confirmation bias includes actively discrediting evidence that would counter their beliefs. A similar concept is fixation on a single hypothesis in spite of evidence that indicates that they should revise their hypothesis (De Keyser & Woods, 1990).

During the simulation study, the study participants described and used strategies that were designed to protect against the vulnerability of incorporating "low quality" or "distorted" information in an analysis product. They explained that they did not always use these strategies, both in the simulation study and in actual analysis situations, because they were resource-intensive. Presently, there are no computer-support tools for laying out information in dedicated areas next to each other to enable quick and easy comparisons, so many of the strategies depend on printing out and iteratively highlighting descriptions within documents. Although the methods differed in exactly how they were implemented (e.g., one participant electronically highlighted tentatively accepted information in blue until it was corroborated, another drew an arrow below a line in handwritten notes when information was believed to be true), generally the methods made distinctions between "uncertain," "tentatively accepted," and "verified" information, how many times information had been corroborated or how many times competing evidence descriptions were seen, and from where the information originated[6]. In general, these methods were aimed at making the process less cognitively challenging by reducing the required memory load at the cost of taking much longer to perform the analysis.

---

[6] Although participants were only observed to make these somewhat large-grained distinctions, evidence interactions from a theoretical perspective (Schum, 1987) have been broken down into much finer distinctions. These finer distinctions could potentially be useful from the point of view of training novice analysts and/or forming the basis for tools designed to support this process of corroborating data.

The strategies that the professional analysts described to reduce the vulnerability to failing to identify and adequately resolve conflicts in the data are similar to strategies that have the same purpose in the context of other domains. For example, it is an industry standard that professional journalists are required to obtain the same information from at least two independent sources before publication. In medicine, people often will go to more than one medical practitioner in order to get a "second opinion" on a diagnosis. In the domain of space shuttle mission control at NASA Johnson Space Center, practitioners who diagnose unexpected anomalies on the space shuttle are required to follow an extensive procedure of describing all of the possible explanations for the anomaly, explicitly ruling out as many as possible, and justifying why a particular explanation is believed over competing explanations. Guerlain et al. (1999) have described that expert blood bankers in antibody identification also try to collect independent, converging evidence to both confirm the presence of hypothesized antibodies and to rule out all other potential antibodies.

5.4.3 Relying on OutDated Information

The third source of inaccurate statements was outdated information that once had been considered correct but then later had been overturned when new information became available. This type of "inaccurate" information was much more difficult to detect and resolve than misunderstandings by report writers. There were descriptions from one point in time that could be considered accurate for that point in time but that greatly differed from updated descriptions at later points in time. Because the "findings" or data set on which to base an analysis came in over time, there was always the possibility of missing information that was released after the report that was being read that could overturn or render previous information "stale." This occurred both for descriptions of past events where the information about the event came in over time as well as for predictions about future events that changed as new information became available on which to base the predictions. When these updates occurred on themes that were not central enough to be included in report titles or newsworthy enough to generate a flurry of reports, it was very difficult to know if updates had occurred or where to look for them.

5.4.3.1 Outdated information in the Ariane 501 Scenario.

In the Ariane 501 scenario, there were five examples of updates on a theme that rendered previous information "stale":

1) Immediately after the event, reports indicated that the failure might be due to a mechanical problem. After about a day, it became clear that it was due to a software failure.
2) Reports immediately following the event reported that ground controllers had blown up the rocket. A few days later, it became clear that the rocket had self-destructed before the ground controllers issued the destruct command.

3) It was reported for several weeks after the Ariane 501 failure that the Cluster satellite program would be discontinued as a result of the loss of the Cluster satellite payload. About a month later, it was reported that one of the four Cluster satellites would be replaced. Two months after that, it was reported that either one or all four Cluster satellites would be replaced. On April 3, 1997, it was announced that all four Cluster satellites would be rebuilt.
4) Prior to the 501 launch, the next launch in the Ariane 5 series, 502, was expected to take place about September 1996. Following the 501 failure, it was expected that the launch would be delayed a couple of months. When the Inquiry Board Report was released, the predictions were that the delays would increase in order to eliminate the software problem, so the launch would be between March and June 1997. There were several slips to the predicted 502 launch for a variety of reasons, including the software validation taking longer than expected and some payload-related delays. The 502 launch was finally made on October 30, 1997, over a year after the original predictions prior to the 501 launch.
5) Prior to the 501 launch, PanAmSat was considering being one of two payloads for the first commercial launch of the Ariane 5 vehicle, 503. Partly as a result of the failure of the 501 launch, PanAmSat went with another company. Later, a scientific payload designed to test conditions for re-entry into the atmosphere; the Atmospheric Reentry Demonstrator (ARD) was identified as one of the two payloads for the 503 launch. For many months, there was speculation about several potential commercial customers for the 503 launch, including the Matra Marconi Space Hot Bird 5. After the 502 launch had some somewhat minor anomalies that prevented its payload from reaching the correct altitude, no commercial payload could be found for the 503 launch. In the end, the 503 launch only included the scientific ARD test payload and the flight was used as another qualification launch for the Ariane 5 rocket design.

5.4.3.2 Example: Impacts to the Cluster satellite program.

To illustrate how easy it is to fall prey to relying on stale information, consider the process that study participant 6 employed (Figure 25) to come to the conclusion in his verbal briefing that the Cluster satellite program had been discontinued as a result of the Ariane 501 incident: "The immediate impact were that the solar wind experiment was destroyed. They couldn't afford to build any more satellites so they couldn't pursue that anymore." From a global perspective, this is an inaccurate statement given that later updates overturned this initial assessment of the impacts and the Cluster satellite program was later fully reinstated.

Essentially, participant 6 did not open any documents that contained updates on the impact to the Cluster satellite program. The participant opened seven documents during the analysis. Only two of the documents contained descriptions that predicted what the impact to the Cluster satellite program as a result of the Ariane 501 failure would be. In the first description, a scientist working on the project directly stated that

the project would be discontinued. While reading this report, the participant verbalized that the scientific mission was dead and that the experiment was destroyed. The second description was more vague about the impact and does not directly make any predictions but could be viewed as converging evidence that the Cluster satellite program would be discontinued. It is no surprise given this process that the participant included in the verbal briefing a description similar to the one from the June 5, 1996, article that the experiment was destroyed and that the program would no longer be pursued. In this case, the participant employed the strategy of corroborating information from two independent, authoritative sources that were written in the same time period (which eliminated the first two sources of inaccuracies), incorporated it into the analysis, and yet missed later updates that rendered that information inaccurate.

Article Date/Content	*Participant's Response*
June 5, 1996: *one of the scientists involved in the project said that it was now finished..."There is neither time nor the money to build four more...the mission is dead, dead, dead."...scientific missions tend to be one-offs and therefore irreplaceable..."All our work just gone in seconds."* →	"It wasn't insured...Immediate impact is it was carrying four solar wind experiments and the scientists say that's it, that's all it says, satellites like that are very expensive. The mission is dead, dead, dead...just lost a few satellites. The only immediate impact was that it...and destroyed the experiment."
July 5, 1996: *Why were the cluster satellites, one of the most original, interesting, and costly missions in the space programs, carried on a test flight?...1.8 trillion life for the cluster satellites...down the drain* →	nothing

Figure 25. Participant 6's process trace on the impact to the satellite program.

5.4.3.3 Example: Delay to 502 flight.

Another example illustrates some of the difficulties in dealing with updates over time regarding predictions about a future event, in this case the delay to the next launch, Ariane 502. In the written question, the participants were asked what the short and long-term impacts of the launch failure were, and three of the eight study participants (6, 4, and 5) chose to include the delay to 502 in their description of the impacts. We will now describe how each of these participants approached this portion of the analysis.

Study Participant 6 – 502 delay

The reports that study participant 6 read had descriptions such as "a few months from now" and "mid-semester 1997" that could only be recognized as conflicting if the participant looked at the report date and interpolated the predicted launch date from the report date (Figure 17). While reading the document that predicted the flight would be in the next few months after June 1996, the participant verbalized that the 502 launch would be delayed. The verbal briefing on the delay to 502 was "about six months," which is the same as the information indicated in the June 6, 1996 article. The participant never opened an article that contained an update on this theme that was

later than July 25, 1996, and so underestimated the delay to the 502 launch. He never searched for or opened an article that gave the explicit date of the 502 launch.

In the timeline to the left in Figure 26, the first circle represents the 501 launch failure on June 4, 1996. The second circle represents when the report from the board of inquiry was published on July 19, 1996. The circle on the timeline on the right indicates when 502 actually launched, on October 30, 1997. The lines between the two timelines indicate when 502 was predicted to launch in relation to the date of the report. With the dual timeline representation on the left, it is much easier than with the "text + report date" representation on the right to recognize that the predictions for the 502 launch are conflicting, ranging from September 1996 to June 1997.

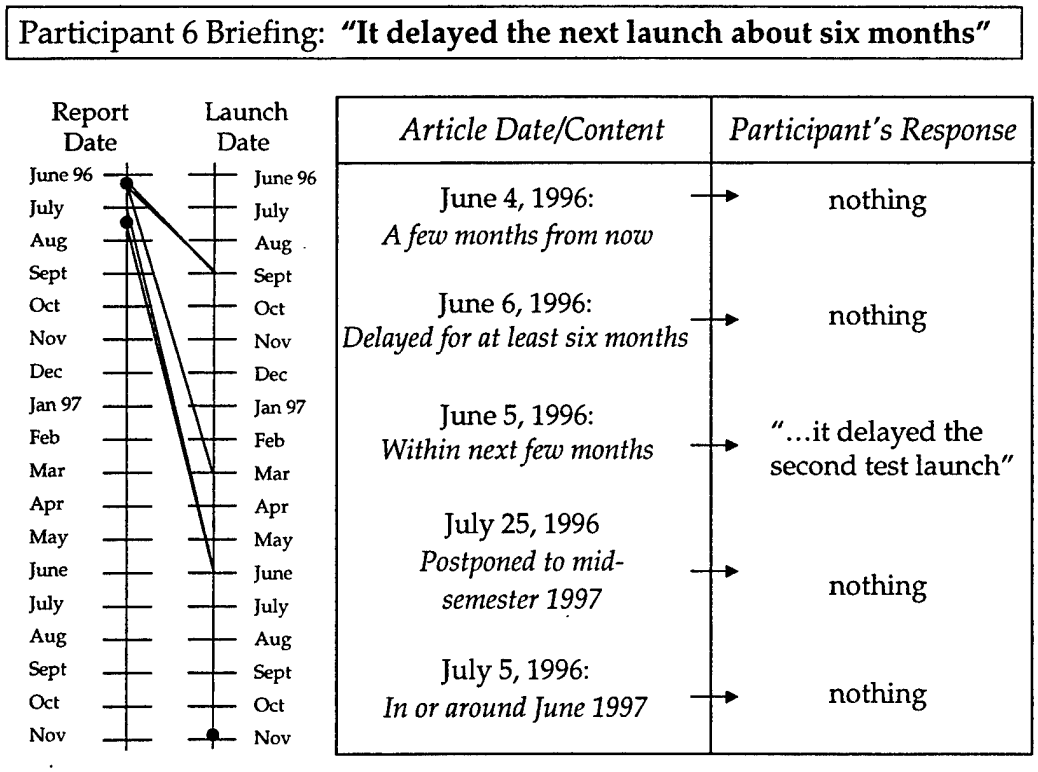
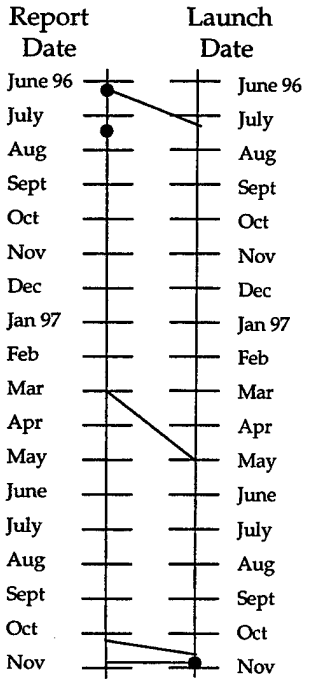


Figure 26. Participant 6's process trace on delay to 502 launch.

Study Participant 4 – 502 delay

Study participant 4 went farther than participant 6 in trying to track down updates on this theme by explicitly looking for when the actual 502 flight occurred (Figure 27). He looked for and found the actual launch date for 502 as October 30, 1997. His strategy was to determine the delay by looking for when the launch was originally scheduled to occur and comparing the difference between the two dates. He found an article that stated that the launch was originally set for May. He verbalized that "...it was originally scheduled to be launched in May and it was launched in October." With this framing, this would mean that 502 was delayed by 5 months. This report, however, is misleading because it is only the delay from the point of view of the launch date being scheduled as May 1997. There were several delays before the May 1997 launch window. Calculated from the 502 predicted date prior to the 501 launch, the delay was about 15 months. Because of the nature of the reports that he had available to do his analysis, however, he had no way of knowing that the information about the launch "originally" being in May 1997 is actually an update from being "originally" scheduled for August 1996. Note that his verbal briefing is vague on how much the launch was delayed, possibly to avoid making any inaccurate statements on this theme.

Participant 4 Briefing: **"Impacts...delay of second flight of Ariane 5."**



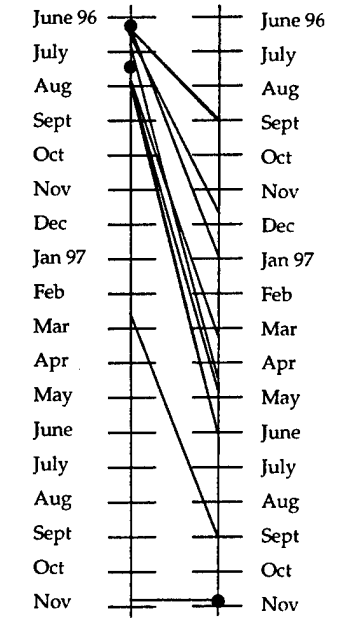
Article Date/Content	*Participant's Response*
June 4, 1996: *Launch preparation will begin June 16*	"Second launch scheduled for June 16, 97" (can't tell if that is a delay from original schedule)
October 13, 1997: *Rescheduled to Oct. 28, 1997. Initially scheduled for June 1997, reset to Sept. 15, then Sept. 30*	"Second launch scheduled for October 28" (would like to find out when second launch was originally scheduled)
March 27, 1997: *Put back to mid Sept...initially set for May, put back to June, then July*	"Originally set for May 1997." (thinks should get the actual date the flight occurred)
October 30, 1997: *Today Ariane 5 rocket was successfully launched*	"So it was originally scheduled to be launched in May and it was launched in October."

Figure 27. Participant 4's process trace on delay to 502 launch.

Study Participant 5 – 502 delay

Participant 5's process was similar to participant 4's in that he also explicitly searched for and found the date that 502 actually launched. His strategy of determining the delay to the 502 launch was somewhat different – rather than searching directly for the original date of the scheduled launch, he sampled many documents in order to get a feel for how much and how often the launch date slipped (Figure 28). He opened 11 documents that contained information relating to the delay to the 502 launch. We have evidence that he read the information on the 502 theme in several of these documents because he verbalized about it while he read it and because of the information contained in his verbal briefing. In his verbal briefing, he indicated that there were conflicting predictions on when the 502 launch was supposed to occur, that the launch in general was getting backed off, and then said when the actual launch occurred.

Participant 5 briefing:
"The next flight of the vehicle was scheduled already. Apparently one article said the first six months of 97. I looked at another article that said March or April and I saw another that said it was going to happen in July, so it was getting backed off and the final launch, I believe, was in October. The second launch, the 502 vehicle, which is Ariane 502."



Article Date/Content	*Participant's Response*
July 25, 1996: *mid-semester 1997*	nothing
March 25, 1997: *until at least September*	(wants to see actual 502 launch date)
June 5, 1996: *within the next few months*	nothing
October 30, 1997: *today...launched*	(notes the launch anomaly on 502)
June 4, 1996: *a few months from now*	nothing
July 28, 1996: *first half of 1997*	"predicts delay for 502"
June 5, 1996: *early 1997*	nothing
June 6, 1996: *later this year*	nothing
June 10, 1996: *early 1997*	"delay qualification of Ariane 5"
July 24, 1996: *March at the earliest*	nothing
July 24, 1996: *mid-1997*	"next flight in 1997"

Figure 28. Participant 5's process trace on delay to 502 launch.

In summary, participant 6 repeated an analysis in a report of how much 502 would be delayed, participant 4 conducted a targeted search for the original date and the actual date by browsing articles in his query, but did not locate the original date prior to the

501 launch, and participant 5 sampled many reports between the 501 date and the 502 date and described that the launch was being backed off and gave the final actual date. It is interesting to note that none of these study participants attempted to pin down exactly how much of the delay to the 502 launch was due to the impacts from the 501 failure versus other reasons, such as payload delays. The difference in defining the impact to 502 as how long after 501 the launch occurred versus how much of the delay was due to the failure of 501 could be viewed as a difference in the estimation of the customer's needs. Perhaps the customer only needs to know when the Ariane 5 rocket launcher will be considered operational, or perhaps the customer wants to look at how much software failures have impacted the satellite business in general. In any case, the definition of the impact as the number of months between 501 and 502 is a simpler analysis task than breaking down the delay by what factors contributed to it. The first task could be viewed as a synthetic task, extracting an uncontroversial relationship across reports, and the second as a contestable task, where different people might determine the delay to be due to a different combination of factors.

5.4.3.4 Impact of relying upon assumptions that did not apply.

As a result of basing an analysis on "stale" information that had been turned over by later updates, study participants made several inaccurate statements at varying levels of importance. The participants would miss the updates either because documents containing the updates were not opened or because his or her attention would be focused on other themes while reading the document containing the update. Partly as a result of this pattern of vulnerability:

- study participants 6 and 7 (written briefing) reported that the Cluster satellite program was cancelled when it was later fully reinstated,
- study participant 8 reported that the 502 launch was originally scheduled for the first half of 97 when the predicted date prior to the 501 launch was September 1996,
- study participant 4 verbalized during the process that the original predicted date for the 502 launch was May 1997 when the predicted date prior to the 501 launch was September 1996, and
- study participant 6 reported that the delay to the 502 launch was about 6 months when it was over a year.

The vulnerability to missing critical information is particularly troubling because it is so difficult for practitioners to gauge whether or not they have missed critical information. It is the *absence* of information, either from not sampling the information or having attention directed on a different theme while reading a document, that creates the vulnerability.

The observation that study participants did not always specifically look for updates during the analysis process or base their levels of confidence on whether or not updates had been located is interesting (see Table 24 in Section 5.6). It is possible that this

observation has implications for training, although many of the study participants indicated an understanding that reports immediately following the event lacked details that came out later. Perhaps there are few strategies that have been developed to deal with this vulnerability because the problem is too difficult given current support tools. Updates could be reported hours, days, weeks, months, or years after an event. Many of the updates on more minor themes do not cause a flurry of reports and are not reflected in the date/title view of the reports. It is possible that "agents" that suggest targeted query formulations and/or "seed" representations with updates on a theme could help address this vulnerability, particularly if the agents had "smart" natural language processing capabilities. This design direction would require artifact-based investigations in order to gain a better understanding of how to make the concept useful despite the fact that the agent's suggestions would likely be incorrect much of the time.

5.5 Findings in the Context of the Abductive Inference Literature

5.5.1 Intelligence Analysis as Abductive Inference

The finding that study participants made inaccurate statements in their verbal briefings is not surprising given that intelligence analysis is an inherently fallible endeavor. Intelligence analysis can be characterized as abductive inference. Abductive inference is a model of reasoning from observations to explanatory hypotheses. Unlike deductive reasoning, where conclusions are guaranteed to be true as long as the premises are true, abductive inference, generally defined as inference to the best explanation (e.g., Josephson & Josephson, 1994), has no such guarantee. The "best" explanation given the available data is not always correct.

Compared with the deductive model of inference, abductive inference is a better descriptive model of how people reason in real-world, complex situations. Diagnosis is an example of a well-known abductive process, where a diagnostic reasoner selects an explanatory hypothesis to explain observed symptoms. The abductive process includes observing deviations from a nominal state, proposing explanatory hypotheses to account for the deviations, and selecting the "best" or most warranted explanation from the set of hypotheses.

More formally, abductive inference follows the pattern (Josephson & Josephson, 1994):

D is a collection of data (facts, observations, givens)
H explains D (would, if true, explain D)
No other hypothesis explains D as well as H does
Therefore H is probably correct.

5.5.2 Second Order Abductive Inference

The Ariane 501 scenario that was used in the simulation study captures many of the complications involved in complex abductive inference (Figure 29). As is often the case, much of the anomalous data could be explained by several hypotheses. For example, the observation that the rocket swiveled abnormally could have been due to poor guidance data, a mechanical failure, or a software failure. The main observation that could be explained by a software failure and not the more typical hypotheses was that both the primary and backup Inertial Reference Systems (IRS) shut down nearly simultaneously. Although this finding made the software failure the most plausible explanation, there was an additional finding that was not covered by this hypothesis -- unexpected roll torque during ascent. The full set of observations was explained by the combination of two hypotheses – a software failure and an unrelated mechanical problem.

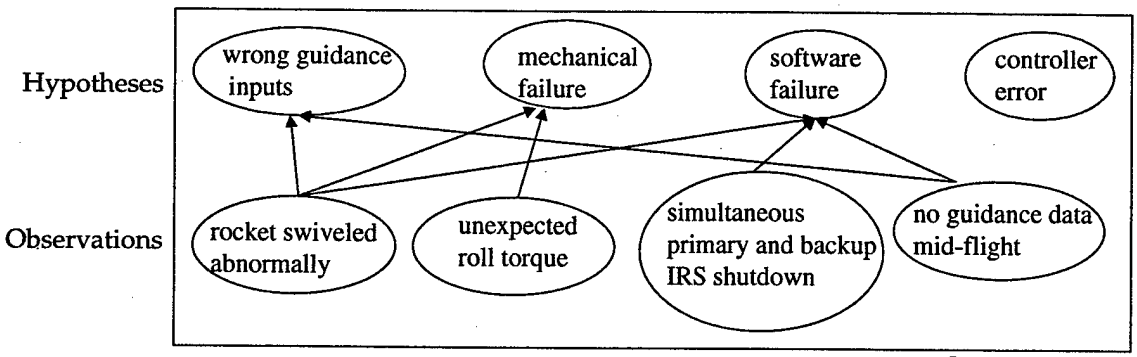


Figure 29. Hypothesis space in Ariane 501 scenario.

Although clearly the abductive inference framework is a very useful one for describing the relationships of the set of observations and hypotheses implicit in the Ariane 501 scenario, the think-aloud protocols and the decisions that the study participants made during the process gave surprisingly little evidence that a standard abductive inference process was being used. Rather than gathering a collection of data, determining what hypotheses would explain the data, and comparing the plausibility for different combinations of hypotheses in order to come up with a best explanation, the study participants appeared to be following a different process. The iterative, interacting steps in the process used in the simulation study could be roughly described as:

- sample documents (generally by keyword searching and browsing dates and titles),
- scan and read portions of the documents,
- break down a document into thematic pieces,
- reorganize the pieces by theme,
- determine if thematically related pieces corroborate or conflict and why,
- come to closure on an interpretation (which could be reopened later in the face of new evidence), and

- integrate and synthesize an interpretation of the thematic elements into a justifiable analytic product such as a written report or verbal briefing.

The main difference between the theoretical pattern of abductive inference and the empirical evidence is that the study participants were not dealing with elemental observations and hypotheses. They were dealing with a "second order" set of data where interpretive frames already existed in which the report writers assumed particular hypotheses and presented data mainly in support of these hypotheses. The main task of the intelligence analyst, therefore, was to improve the veracity of the analytic product by corroborating multiple reports of others who had already performed the task of mapping explanatory hypotheses to a dynamically changing data set.

Given this situation, the main "hypothesis space" that the study participants used is represented in Figure 30. Rather than the "elemental" hypotheses and data given for the Ariane 501 scenario, the think-aloud protocols gave evidence for the study participants dealing at the "second order" level of using cues from the text, document, and source to evaluate whether to incorporate the information. The study participants displayed expertise in recognizing the cues that were used in evaluating the information and in relating those cues to possible hypotheses. Note that this expertise would probably not be available to surrogate participants such as undergraduate students.

Although the main emphasis of the intelligence analysis task was on this "second order" evaluation and integration of others' analyses, occasionally the study participants were observed to revert to the elemental level in order to resolve discrepancies. This strategy relied more heavily upon domain expertise. For example, a study participant stated that one hypothesis for why the inertial reference system shut down was implausible because "I'm finding it hard to believe that the vehicle is going to fly without any inertial inputs whatsoever." By having expert knowledge relating to the area, the participant was able to use that knowledge as additional evidence in determining how to resolve a discrepancy in two explanations.

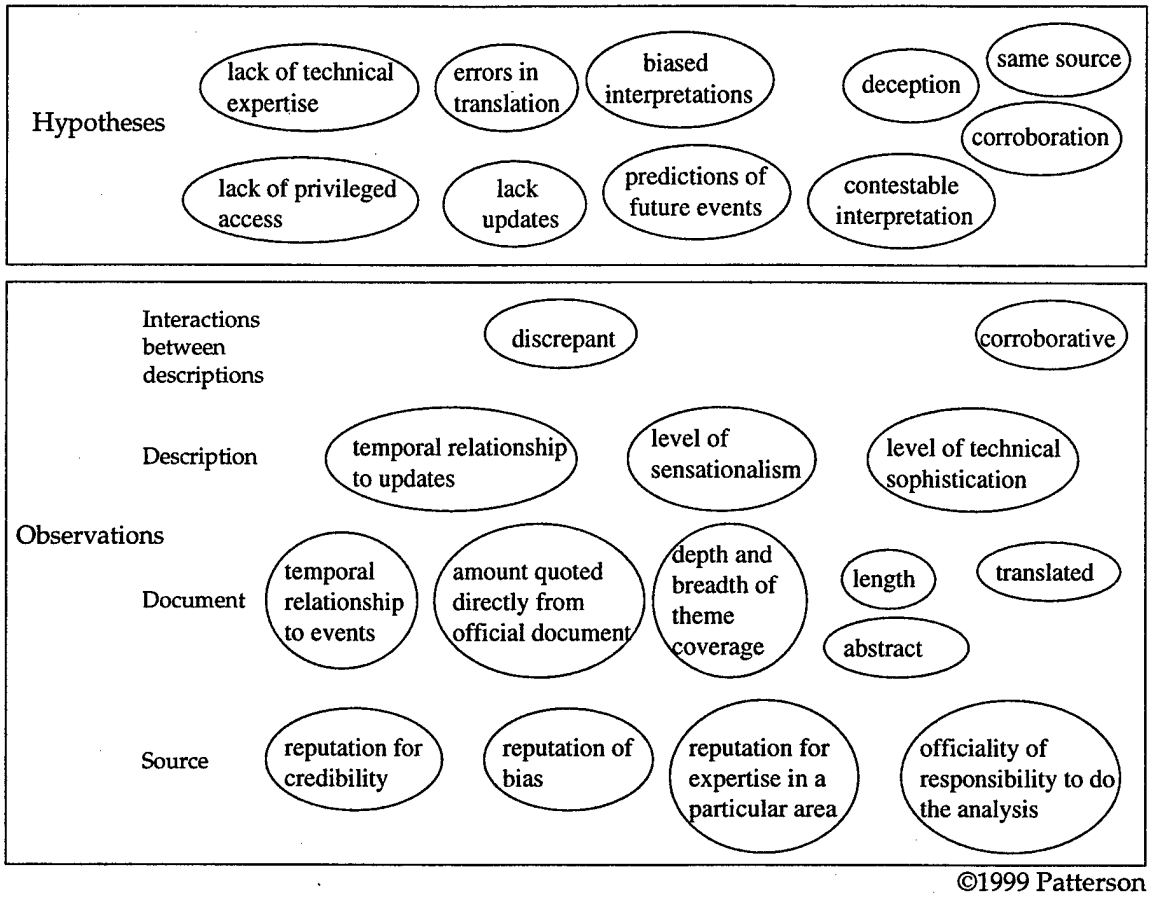


Figure 30. "Second Order" hypothesis space.

5.5.3 Sources of Inaccurate Statements

One of the main findings of this investigation was the sources of inaccurate statements made by study participants in the verbal briefings. These empirical findings will now be discussed in the context of the theoretical literature on abductive inference, including a list of potential errors in abductive inference by Josephson and Josephson (1994), a discussion of the impacts of missing information on inferential analysis (Schum, 1987), and potential errors in abductive inference in the context of disturbance management (Woods, 1994b).

The three sources of inaccurate statements were empirically determined from this investigation to be:

1. relying on assumptions that did not apply,
2. incorporating information that was inaccurate, and
3. relying on outdated information.

Josephson and Josephson (1994) provided a list of theoretical causes for "incorrect" explanations for data in abductive inference processes:

1. There was something wrong with the data such that it really did not need to be explained.
2. There might be causes for the data that were not considered, perhaps because they were unknown or overlooked.
3. Hypotheses were incorrectly judged to be implausible.
4. Hypotheses were incorrectly thought not to explain important findings.
5. The diagnostic conclusion was incorrectly thought to explain the findings.
6. The true answer was underrated, due to faulty knowledge or missing evidence.

Additionally, Schum (1987, p. 3) described how analysis may suffer from missing important information:

> One of the basic suppositions upon which this entire work is based is that what we do not recognize or take into account in our inferences CAN hurt us...A conclusion may be inadequate, not because of the manner in which we evaluated existing evidence, but because of our failure to consider other evidence which might have led us to a more adequate conclusion. The second attribute involves failure to recognize and exploit the wide array of evidential subtleties which are often apparent on close inspection of evidence and the sources from which evidence comes. In short, there is often significant inferential "juice" in evidence which goes unnoticed and, therefore, unincorporated in our inferences. A third attribute concerns possible conclusions we might have entertained but did not, or those we did entertain but dismissed prematurely.

Finally, Woods (1994b) discusses potential biases and errors in abduction in dynamic situations:

1. Fixation errors or cognitive lockup: where the practitioner fails to revise an initial hypothesis despite the presence of cues (new or additional evidence) that should suggest that the earlier assessment is erroneous or that the state of the monitored process has changed.
2. Failure of attentional control:
   A) devoting processing resources to too many irrelevant changes
   B) discarding too many potentially relevant changes as irrelevant
3. Failure to generate plausible alternative hypotheses
4. Selecting a hypothesis or set of hypotheses based on parsimony without considering other factors such as the likelihood of a hypothesis or consequences of acting on that hypothesis.

Relating these theoretical sources of inaccurate statements with the empirical sources, we see that there are informative relationships (Figure 31). First, the empirical source of inaccurate statements of "relying on assumptions that did not apply" points to how study participants were taking "shortcuts" in analysis. The use of time-saving heuristics under high workload and/or time pressure is not often discussed in the abductive inference literature (although see Punch and Josephson, in preparation, for an interesting exception). Relying on assumptions without checking them is a heuristic

that enables analysts to more quickly perform analyses, with the possibility that assumptions could prove wrong in some cases. The heuristic maps onto the theoretical frameworks of Josephson and Josephson (1994) and Schum (1987) in that any or all of the following steps in abductive inference may be shortened or shed:

- determining the findings to be explained
- generating hypotheses to explain the data, and
- evaluating how the hypothesis explains the findings.

Second, the empirically-discovered sources of inaccurate statements of incorporating information that was inaccurate and relying on outdated information points to another time-saving heuristic: believing a description to be reliable without fully corroborating it. The framework from Josephson and Josephson (1994) would consider these sources as part of the category "There was something wrong with the data such that it really did not need to be explained." Although this is not an incompatible statement, this characterization effectively puts problems with inaccurate data out of the scope of the main framework. Schum's (1987) framework incorporates many more subtleties of how evidence may contradict or interact with other evidence, although he does not make the distinction between information that was believed to be true at one point but later was not and information that was always considered to be incorrect by experts. Woods (1994b) directly notes the source of relying on outdated information (in the case where the practitioner accesses updated information) in the failure to revise a hypothesis in the face of new evidence. Although Woods' framework does not contain many of the subtleties in evidence interaction that are described by Schum, implicit in the bias toward selecting single hypothesis explanations for parsimony considerations alone is the problem of incorporating inaccurate information that has a low likelihood of being true. An example is ignoring the high a priori probability of a mechanical sensor failure, such as in the domain of space shuttle mission control. Therefore, one of the common hypotheses to consider is always that some subset of the data is inaccurate due to a sensor problem.

Third, note that the mapping from the empirical sources of inaccurate statements to most of the elements of the Josephson and Josephson (1994) framework that address relationships between findings and hypotheses is missing. This lack of fit between the empirical evidence and the theoretical model of abductive inference highlight an important distinction between standard abductions such as diagnoses and the nature of abductions in intelligence analysis. The intelligence analyst is not performing abductive inference on a base set of findings and mapping them to a hypothesis space. Rather, the intelligence analyst needs to evaluate abductive inferences made by others, compare the relationships between inferences made by others, and then come to an assessment. Because the information that the study participants were evaluating already incorporated hypotheses and findings as well as evaluations of plausibility, the participant could provide an implausible hypothesis as part of a briefing without ever directly evaluating the plausibility of the hypotheses. This could occur either because the participant did not read discrepant descriptions or because (s)he did not notice that

descriptions that were read were discrepant. In other words, it was not commonly observed for the study participants to make judgments about the relationship between the findings and first-order hypotheses (e.g., Figure 29), and therefore failures in that process were not often responsible for inaccurate statements.

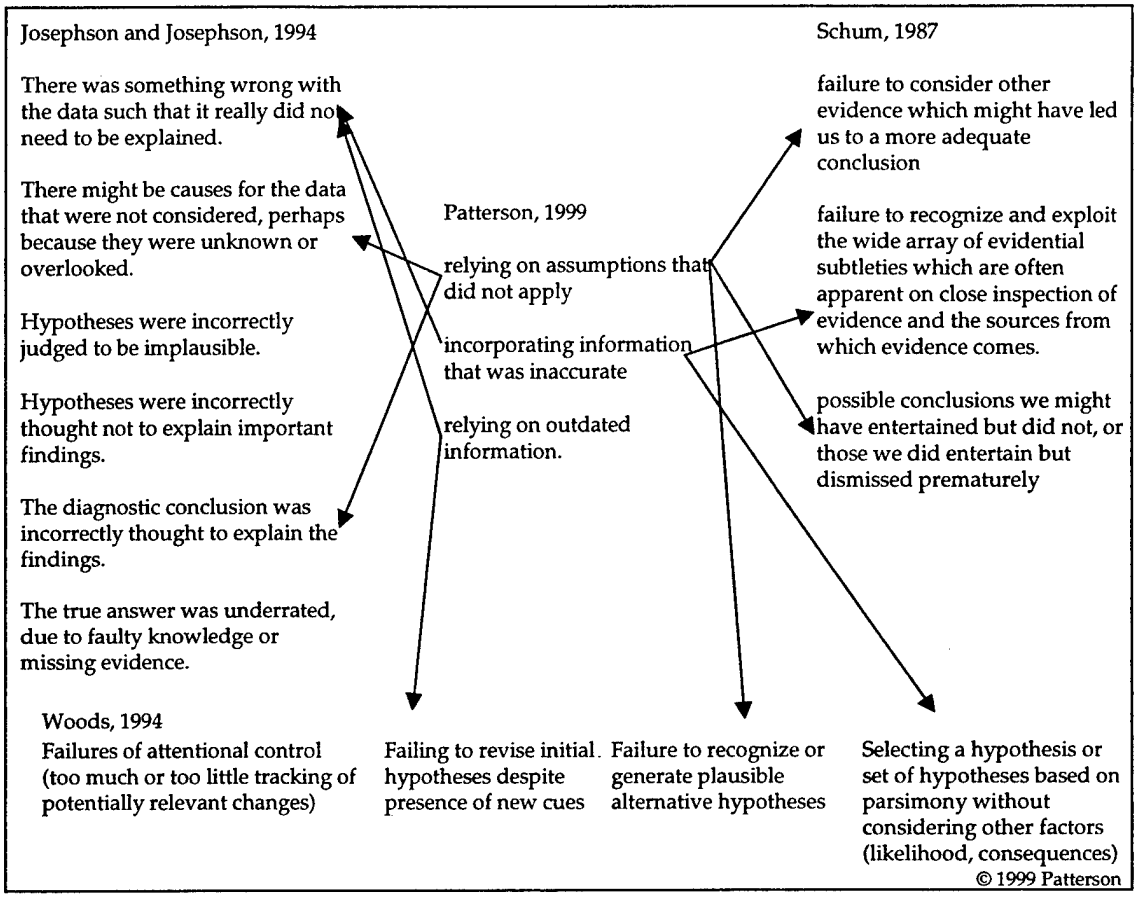


Figure 31. Theoretical and empirical sources of inaccurate statements.

It is interesting to note that the study findings more closely relate to the expectations of Schum (1987) than Josephson and Josephson (1994). An important element of intelligence analysis is that the findings themselves are disputed in different reports. Schum (1987) is one of few abductive inference researchers that takes this complication into account in his theoretical framework, as well as the fact that intelligence analysts are not able to access the full set of findings but rather are vulnerable to missing information. In contrast, the focus of many researchers in artificial intelligence and abductive inference is how hypotheses map onto a full set of undisputed findings that are in a format that eliminates conflicts with each other and makes it easy to compare their relationships to explanatory hypotheses.

Nevertheless, Schum (1987) does not include the difficulties in dealing with findings that come over time in his main theoretical framework (although see Chapter 14 for

some preliminary discussions of the complexities with temporal data). There is scant discussion in the abductive inference literature on the dangers of relying on outdated information in inferential analysis. A major contribution of this research is identifying the need to better support finding updates when performing analysis. In general, the strategies that have been developed to corroborate information (as well as many of the proposed design "solutions" to data overload) largely ignore the vulnerability of relying on outdated information.

Finally, there was a convergence across all three theoretical frameworks and the empirical findings on the centrality of missing data in abductive reasoning. In Josephson and Josephson (1994), a potential failure is underrating the true answer due to faulty knowledge or missing evidence. In Schum (1987), all three sources are described as instantiations of how information that is not known to the practitioner can adversely affect analysis. In Woods (1994b), striking an appropriate balance between considering potentially relevant changes without overwhelming finite resource processors is the motivator behind the failures of attentional control. One of the main findings of the study was that missing information – missing high profit documents, missing conflicting and corroborating information, and missing updates - adversely affected performance and it was difficult to estimate whether or not important information was missed.

5.6 Stopping Rules and Confidence Estimates

Two questions were identified before the simulation study as questions of interest:
1) How did the participants estimate their confidence in their analyses and how did they communicate this to the "customers" of the analytic products?
2) How did the participants determine when to stop?

As can be seen by the verbal responses (Table 24) to the question: "How confident are you in your analysis?" the confidence estimates appeared mainly to be based on judgments of source quality and the analytic process that had been followed. Regarding source quality, the participants judged whether the sources that they had read were "impeachable" sources, whether or not the information from various sources agreed or disagreed, and whether or not the sources were likely to be biased or deceptive. Some of the other answers appeared to be based on the process that was used: whether or not the information had been verified from a number of independent sources, whether or not the participant had worked through the details by writing a briefing[7], and whether or not the participant had gotten a "feel" based on reading reports forward in time from the incident.

[7] Participant 7 said that he did not feel comfortable giving a verbal briefing before generating a written briefing first. At the end of the session, he asked for a copy of the briefing he had written because he wanted to use it in his work if a related question was asked of him.

Table 24. Bases of Confidence in Analyses

Subject	Verbal Response
2	Pretty confident right now – we've got it. I'm not saying its nailed, but its 90%.
3	I am pretty confident. Basing this on the messages. The only thing is that the sources that they've got are not the sources I'm used to. Putting up information, where is it coming from? If it's just open source, then it's a European company that lost Ariane 5. They're not going to give information that sets people off. They're going to say a minor problem and stuff like that. So probably more classified sources would give better information.
4	N/A
5	I am in general less confident unless my source is impeachable. Here we have the statement of the people that built and analyzed it. They told us what, why, what to do about it, etc., like your question. I wasn't content to stop right there with the one final document because sometimes it's better to develop a feel for that. I am as confident as they are. They either believe what they say or they are lying to us. They are slamming themselves and not just one person so they're probably not hiding too much.
6	I am very confident. Everybody agrees. It was the official inquiry board. The reports...they weren't written by the French, they were written by other people and they don't disagree. If they did, they would say so.
7	For what I said, very confident. For what I wrote, pretty confident. There's still that leap in faith in assuming that the guidance update is an automatic thing that happens with Ariane 4 as opposed to having a hold and pushing a button.
8	Everything that I put in there is from open source information. I have couched my assessment in terms of "it was reported that..." such as "The French said..." and "the reported cause of the accident was..." so it is less "it is this" as "it was reported that..." unless I have definitive data that is more distinctive and reliable and trusted as a source.
9	Confidence 3 on 1-5 because it's a FBIS report. Can it be verified? I'd have to say if I have two independent people saying the same thing – then it would go from 3 to 4.

It is interesting to note that some of the bases for confidence related to the first and second sources of inaccuracies, but not the third. None of the study participants indicated that they were basing the confidence estimate on whether or not (s)he felt that they had found all of the updates to the information. For example, participant 6 stated that he was "very confident" in his briefing based on the perception that the evidence was converging: "Everybody agrees." Similarly, participant 9 said that he would rate his confidence in his briefing as a "3 on a 1-5 scale" because he had not found converging evidence from independent sources for the information: "Can it be verified? I'd have to say if I have two independent people saying the same thing – then it would go from 3 to 4." These definitions of converging make no mention of finding information from different periods in time.

Although not all of the participants were asked how they knew when to stop, some of the responses about confidence extended to include information about when to stop and sometimes evidence was provided by the think-aloud protocols that gives some

insight. This information is provided in Table 25. Essentially, the participants based their judgment of when to stop on whether or not all aspects of the written question could be answered and how important the briefing would be. The participants generally described their treatment of this Quick Response Task (QRT) as the least important type of briefing – 10 to 15 minutes to a direct supervisor, for example. The implication was that for more important types of briefings, such as "briefing tours" or presentations to the President of the United States, that the participants would fill in more details and verify the information from more sources. These briefings would obviously take weeks instead of hours to prepare. One participant referred to this additional work as "embellishing" the analysis.

Table 25. How the Participants Decided When to Stop

Subject	Verbal Response
2	I think we really have enough information to answer everything at this point. I'm not saying I would cut it short, but if I had a QRT and they really wanted it done pretty quickly, I would probably do it. Don't think can go any further to give an overall general view.
3	I think with the information I have got so far that I could probably answer, given how vague the question is, with the information that I've got I would go ahead and give a first answer.
4	What do I have to do again? (looks at written question)
5	Here we have the statement of the people that built and analyzed it. They told us what, why, what to do about it, etc., like your question. I wasn't content to stop right there with the one final document because sometimes it's better to develop a feel for that.
6	Everybody agrees. It was the official inquiry board. The reports...they weren't written by the French, they were written by other people and they don't disagree. If they did, they would say so.
7	I think I now know what happened. (when asked if he knew what happened before he did the written briefing): No, I had the general idea. It was just getting down to exactly what really happened, and I would force myself to do this even to do a verbal briefing. I want to work it with this process.
8	OK, got the when and the why pretty well. OK, the French view on the impact... The assumption that I was on, I approached it as this would be a briefing to a senior person who wouldn't have a lot of time to hear the briefing.
9	OK. Let's see if I've answered them all. Lost a satellite, OK, I have dates and places. First launch. New rocket. I have names and facts. Again, where it was, where it occurred, and what the main impacts were. I have all of it, just off of this. Do you want me to embellish what I have here?

6 DISCUSSION

6.1 Summary of Findings

When analysts are asked to analyze something outside their immediate base of expertise, are tasked with a tight deadline, and are under data overload conditions, they are vulnerable to missing critical information that could potentially overturn their analyses. In order to cope with data overload, all of the participants in this study were observed to refine their initial queries by using narrowing tactics such as adding keywords until they reached a manageable set for browsing and then treat that manageable set as a home base rather than conduct additional queries or expand the set in other ways. Partly as a result of this coping strategy, all of the study participants missed some of the nine high-quality or "high profit" documents available in the database. The four participants who spent the most time and read the most documents did find and heavily rely on some of the high profit documents, whereas the four who spent less time and read fewer documents found fewer of them, did not rely upon them heavily in their verbal briefings, and made more inaccurate statements in their verbal briefings. These findings suggest that the participants who relied on the high profit documents did not employ expert strategies to recognize them, but rather that "persistence" was the main tactic in finding them, probably because the "number of hits" and "date and title" views of document sets did not help the participants in estimating the document quality.

One of the main challenges in inferential analysis is that there is always inaccurate and stale information in the data set that is used to perform an analysis. By tracing why inaccurate statements were made in the verbal briefings, three sources of inaccurate statements were identified: 1) participants relying on assumptions that did not apply, 2) incorporating information that was inaccurate, and 3) relying on outdated information. The first two sources could be eliminated by checking for converging information from independent (non-deceptive) sources. The last source of inaccurate statements was very difficult to eliminate as high-quality, independent sources published during one time period could provide converging evidence that was later overturned as new information was discovered. Predicting and locating updates on a theme was extremely difficult to do, leaving analysts particularly vulnerable to making inaccurate statements from relying on outdated information.

Although, in general, analysts were aware of these sources of inaccurate statements, the calibration on how vulnerable an analysis is to having inaccuracies in the analytic product is difficult given that it is based on the *absence* of information. The confidence judgment is particularly difficult with relation to updates because of the potential for basing confidence estimates on finding converging evidence from high-quality independent sources, which would eliminate the first two sources of inaccuracies but

not the third. This situation leaves analysts open to being over- or under-confident in the veracity of their analytic products.

Strategies were observed that were aimed at reducing the inaccurate statements in an analytic product. In general, these strategies were difficult, resource-intensive, and time-consuming. For example, one strategy involved printing out the entire set of documents and using highlighters as memory aids for how many times conflicting and corroborating information from independent sources had been read on a theme. The baseline electronic environment that was used in the study provided only crude support for this process, with a general lack of memory aids and targeted notation tools to help with the process of identifying, tracking, and resolving conflicts in the data.

At the broadest level, some of the participants could be viewed as having prematurely closed the analysis process. During prior interviews, we discovered that it is generally recognized among professional analysts that there is a vulnerability to premature closure during the analysis process. Note that there are also concerns about premature closure in inferential analysis tasks in other domains such as medical diagnosis (McSherry, 1997; Baldwin & Rice, 1997; Fraser et al., 1989) due to the costs of obtaining further information and in general about the effects of time pressure on human judgment and decision making (Svenson & Maule, 1993). The potential impacts of premature closure include a degraded quality of the analytic product and poorly calibrated confidence in the veracity of the analytic product. In addition, analysts might be less able to effectively respond to questions. More specifically, as a result of premature closure, analysts might:

- make inaccurate statements in an analytic product,
- not rule out as many competing hypotheses as might be desirable,
- cover fewer items in the analytic process (information that is searched as well as what topic items are included in the briefing)
- cover topic items less thoroughly,
- not be as broad in the coverage of the data,
- or be diffuse in terms of the specificity of predictions.

6.2 Implications of the Study Findings

6.2.1 Evaluation criteria for Proposed Solutions to Data Overload

The main contribution from this study is a model of vulnerabilities in inferential analysis under data overload conditions. These vulnerabilities are useful because they point to a set of challenging design criteria that human-centered solutions to data

overload must meet in order to be useful. These criteria can serve, not only to guide the next cycle in design, but are also useful in generating scenarios to test the effectiveness of proposed designs:

1. Solutions should bring analysts' attention to highly informative or definitive data and relationships between data, even when the practitioners do not know to look for that data explicitly. Informative data includes data that deviates from expectations, data that eliminates potential hypotheses, and data that contradicts or corroborates other data. A particularly difficult criterion to meet that should be designed into evaluation scenarios is to help analysts recognize updates that overturn previous information.
2. Solutions should aid analysts to manage data uncertainty. In particular, solutions should help analysts identify, track, and revise judgments about data conflicts.
3. Solutions designed to deal with data overload should help analysts to avoid prematurely closing the analysis process. They should broaden the search for recognition of pertinent information, break fixations on single hypotheses, and/or widen the hypothesis set that is considered to explain the available data.

These evaluation criteria are interesting, in part, because they are so difficult to address. They are not amenable to simple, straightforward adjustments or feature additions to current tools. Meeting these design criteria will require innovative design concepts.

6.2.2 Towards Context-Sensitive Design Aids

The characterization of data overload as finding the significance of data in a vast data field leads to the conclusion that, because of the context-sensitivity problem, we must direct our efforts towards techniques which do not rely on knowing in advance the relevant subset of data. Additionally, the research base in cognitive engineering has shown that methods that rely centrally on machine processing are vulnerable to brittleness, particularly in complex real-world domains such as intelligence analysis. These observations point to a useful region of the design space for solutions to data overload: helping people to recognize or explore the portions of the data field that might be relevant for the current context so that they can focus attention on those areas.

The design approach of Woods, Patterson, and Tinapple (in preparation) therefore involves two parallel strategies. The first is to use models of the domain semantics as the foundation for visualizations that provide a structured view of the data field for observers. The intent is to take advantage of the context-sensitive properties of human cognition by giving observers the perceptual leverage needed to focus in on relevant sub-portions of the data space. The second strategy is to use active machine intelligence in supplemental, cooperative roles to aid human observers in organizing, selecting, managing, and interpreting data.

6.2.2.1 Model-based visualizations.

This tactic is similar in philosophy to other methods in the literature which use models of domain semantics as a way to structure displays of data (e.g., Vicente & Rasmussen, 1992). Taking advantage of the context-sensitive nature of human cognition presumes a structured data field on which our attentional processes operate. The idea therefore is to build a conceptual space for organizing the data based on a model of the fundamental relationships, objects, and events in the domain. For example, a high profit document for the Ariane 501 scenario used in the simulation study could be modeled as:

- a relatively long document that was released several months after the original event (and certainly after the Inquiry Board Report was officially released from the European Space Agency),
- from a credible source on rocket launcher and satellite technologies such as Aviation Week and Space Technology,
- not an abstract,
- not reporting information from another news agency (i.e., not "secondhand"),
- not translated from another language, and
- a report that had been opened several times by others.

In order to support skillful shifting of attention, these visualizations would have to include mechanisms allowing observers to perceive changes or potentially interesting conditions which are not necessarily in direct view, and to re-orient their attention to the new data. They must also emphasize anomalies and contrasts by showing how data departs from or conforms to expectations. For example, the following views are being developed as a coordinated workspace (Woods, Patterson, and Tinapple, in preparation):

- longshot and detailed views of the "report space," with disrupting events emergent in the visualization as clusters of reports in time (given a set of documents retrieved by a query mechanism),
- longshot and detailed views of the "event space," with the focus of attention on particular event themes emergent in the visualization, and
- overlay perspectives of the judgments of how evidence conflicts and corroborates at various scales (detailed text descriptions of several words, selected descriptions of several paragraphs, and at the document level).

6.2.2.2 Cooperative roles for machine intelligence.

One prevalent view in the design of human-machine systems is to "allocate functions" between machine and human agents. With this approach, designers must choose which functions are better performed by machines or humans and tasks are assigned accordingly. This framing often leads either to over-commitment to an immature

machine intelligence or over-reliance on unaided human expertise. The philosophical underpinnings of Cognitive Systems Engineering have redefined this debate of whether to trust the machine or the human as an issue of coordination between team players, where neither "does it all," both are limited resource processors, and both are subject to brittleness in processing. This is not a rejection of technology, but rather a redefinition of how technology and people should interact as team players to make a system that is more robust than the individual elements.

There are a variety of cooperative roles that machine intelligence could play in the approach of Woods, Patterson, and Tinapple (in preparation) that relax the need for machine intelligence to always be correct. For example, a support system that leveraged the model of a high profit document could be implemented different ways depending on the cooperative architecture that was desired:
1) the user could mark a document as high profit and the computer could then display and categorize that information in various ways,
2) a computer algorithm could determine similarities in documents that were marked as high profit and suggest a combination of attributes as representing a model of high profit documents that the user could observe and redirect,
3) the computer system could present potentially "similar" documents to the set that were marked as high profit by the user,
4) the computer system could "seed" potentially high profit documents for the user to browse based on a designer-defined model of a high profit document,
5) the user could give feedback to computer-generated sets of high profit documents to "sharpen" the definition and/or "train" the computer system,
6) the computer system could remind the user to search for high profit documents during the analysis process, and
7) the computer system could critique the user's selection of high profit documents or the reliance upon documents that are not considered high profit by the definition in the computer software.

6.2.3 Methodology

The results of this simulation study increase our understanding of the inferential analysis process under data overload. In particular, we have a richer model of the cognitive subtasks in inferential analysis: identifying what is meaningful in a vast field of data, identifying, tracking and resolving conflicts in the data set, and constructing an explanatory story. We have a much better understanding of the cues associated with the source, the document, and the text descriptions, that professional analysts use to judge the quality of the data. In addition, we have a richer view of the document sets and "bundles" of information that analysts need to manipulate in order to break down the information that they receive and build it up in a new interpretive frame. The understanding that was gained from this reasonably high fidelity simulated task and in-context interviewing is much richer and more detailed than was gained from prior interviews. In the prior interviews, analysts tended to generalize their descriptions of

the strategies they use and the demands that they face in inferential analysis. For example, the description of how they judged information quality was that they based their judgment on the source, glossing over many subtleties that could provide valuable design ideas that were observed to be used when the analysts actually had to perform the task. For example, high profit articles tended to be relatively long documents from sources that are considered credible that were published more than a week after the release of the official Inquiry Board Report. Similarly, some articles were considered low profit because they were abstracts, translated, secondhand, were from sources that were likely to be biased, or were judged to be sensationalistic.

The goal of the simulation study was to discover vulnerabilities in inferential analysis under challenging conditions. The process tracing analysis methodology was useful for discovering patterns across multiple participants and then linking those patterns to impacts on performance. If a more traditional study had been employed where the purpose would have been to verify a priori hypotheses, it would have been easy to miss the importance of "key" documents during analysis, the correlation between participants that used the high profit documents in the database as their key documents and the amount of time that they spent and the number of documents that they opened, the breakdowns in the process of corroborating information, and the impact of missing updates to an analysis that was considered accurate at one point in time. These variables provide valuable insight for what might be useful support tools, training interventions, or organizational restructuring, as well as variables to pursue in a more targeted way in follow-up studies.

This project serves as an illustration for a cognitive task analysis process aimed at discovery (Potter et al., in press). A base of understanding about data overload generated from previous research in other domains, including space shuttle mission control, nuclear power plants, anesthesiology, and aviation flight decks, was used to jump-start the project. The research base developed during this previous research was synthesized and calibrated against the new domain of inferential analysis through interviews and a simulation study. The results of these relatively few investigations allowed us to generate a rich understanding of intelligence analysis under data overload. As a result of this process, we now have valuable insight into the demands that make inferential analysis under data overload conditions fundamentally hard.

REFERENCES

Abate, M.A., Shumway, J.M., Jacknowitz, A.I., & Sinclair, G. (1989). Recording and evaluating end-user searches on a personal computer. *Bulletin of the Medical Library Association, 77*(4), 381-383.

Ahlberg, C., Williamson, C., & Schneiderman, B. (1992). Dynamic queries for information exploration: an implementation and evaluation. *Proceedings of CHI '92, ACM Conference on Human Factors in Computing Systems,* New York, 619-626.

Baldwin, N.S., & Rice, R.E. (1997). Information-seeking behavior of securities analysts: individual and institutional influences, information sources and channels, and outcomes. *Journal of the American Society for Information Science, 48*(9), 674-693.

Bates, M.J. (1979a). Information Search Tactics. *Journal of the American Society for Information Science, 30,* 205-214.

Bates, M.J. (1979b). Idea Tactics. *Journal of the American Society for Information Science, 30,* 280-289.

Bates, M.J. (1992). Search and Idea Tactics. In *For Information Specialists: Interpretations of Reference and Bibliographic Work,* ed. Howard D. White, Marcia J. Bates, and Patrick Wilson, 183-200. Norwood, NJ: Ablex.

Bates, M.J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*(5), 407-424.

Bates, M.J., Wilde, D.N., & Siegfried, S. (1993). An analysis of search terminology used by humanities scholars – the GETTY online searching project report –1. *Library Quarterly* 63(1), 1-39.

Beaulieu, M. & Jones, S. (1998). Interactive searching and interface issues in the Okapi best match probabilistic retrieval system. *Interacting with Computers, 10,* 237-248.

Belkin, N.J. (1993). Interaction with texts: Information retrieval as information seeking behavior. In *Information Retrieval 93: von der Modellierung zur Anwendung.* Proceedings of the First Conference of the Gesselschaft fur Informatik Fachgruppe Information Retrieval (pp. 55-66). Konstanz: Universitatsverlag Konstanz.

Bellardo, T. (1985). An investigation of online searcher traits and their relationship to search outcome. *Journal of the American Society for Information Science, 36(4),* 241-250.
Billings, C. E. (1996). *Aviation Automation: The search for a human-centered approach.* Hillsdale, NJ: Erlbaum.

Biswas, G., Weinberg, J.B., & Fisher, D.H. (1998). ITERATE: a conceptual clustering algorithm for data mining. *IEEE Transactions on Systems, Man, and Cybernetics, 28C*(2).

Blair, D.C. (1996). STAIRS Redux: Thoughts on the STAIRS Evaluation, Ten Years after. *Journal of the American Society for Information Science, 47*(1), 1-29.

Blair, D.C. (1980). Searching biases in large interactive document retrieval systems. *Journal of the American Society for Information Science, 31*, 271-277.

Blair, D. C. & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system, *Commun. ACM 28*, 3, 289-299.

Blair, D. C. & M. E. Maron (1990). Full-text information retrieval: Further analysis and clarification. *Information Processing & Management, 26*(3), 437--47.

Borgman, C.L., Hirsh, S.G., & Hiller, J. (1996). Rethinking online monitoring methods for information retrieval systems: from search product to search process. *Journal of the American Society for Information Science, 47*(7), 568-583.

Bower G. & Morrow, D. G. (1990). Mental models in narrative comprehension. *Science, 24*, 44-48.

Brann, D.B., Thurman, D.A., & Mitchell, C.M. (1996). Human interaction with lights-out automation: a field study. In *Human Interaction with Complex Systems '96*, Dayton, OH.

Brown, R.N. & Agrawala, A. (1974). On the behavior of users of the MEDLINE system. In C. Fenichel (Ed.) *Changing patterns in information retrieval: Tenth Annual National Information Retrieval Colloquium*, Washington, D.C.: American Society for Information Science, pp. 36-38.

Burnett, K. & McKinley, E.G. (1998). Modelling information seeking. *Interacting With Computers, 10*(3), 285-302.

Carroll, J. M., & Mckendree, J. (1987). Interface design issues for advice-giving expert systems. *Communications of the ACM (30)*: 14-31.

Case, D., Borgman, C.L., & Meadow, C.T. (1986). End-user information seeking in the energy field: Implications for end-user access to DOE/RECON databases. *Information Processing & Management, 22*, 299-308.

Chandrasekar, R., & Srinivas, B. (1998). Glean: Using syntactic information in document filtering. *Information Processing and Management, 34*(5), 623-640.

Cimino, J.J. & Barnett, G.O. (1993). Automatic Knowledge Acquisition from MEDLINE. *Methods of Information in Medicine, 32*(2), 120-130.

Cox, K.C., Eick, S.G., & Wills, G.J. (1997). Visual Data Mining: Recognizing Telephone Calling Fraud. *Journal of Data Mining and Knowledge Discovery, 1*(2), 225-231.

De Keyser, V., & Woods, D. D. (1990). Fixation errors: Failures to revise situation assessment in dynamic and risky systems. In A. G. Colombo & A. Saiz de Bustamante (Eds.), *Systems Reliability Assessment* (pp. 231-251).

Doyle, R.J., Charest, L.K., Falcone, L.P., & Kandt, K. (1990). Addressing information overload in the monitoring of complex physical systems. In *4th International Qualitative Physics Workshop*. Lugano, Switzerland, July 9-12.

Eckert, C. (1995). Intervention Strategies for Critiquing Professional Designers. *Open University Centre for the Design of Intelligent Systems Research Report 9504.*

Eick, S.G. (1997). Graphically Displaying Text. *Journal of Computational Graphics and Statistics, 3*(2).

Fenichel, C. (1980). The process of searching online bibliographic databases: a review of research. *Library Research, 2*, 107-127.

Fidel, R. & Soergel, D. (1983). Factors affecting online bibliographic retrieval: a conceptual framework for research. *Journal of the American Society for Information Science, 34*(3), 163-180.

Fischer, G. Lemke, A. C., Mastaglio, T. & Morch, A. I. (1991). The role of critiquing in cooperative problem solving. *ACM Transactions on Information Systems, 9*(3), April 1991, 123-151.

Fischer, G. & Reeves, B. (1992). Beyond intelligent interfaces: exploring, analyzing, and creating success models of cooperative problem solving. *Journal of Applied Intelligence, 1*, 311-332.

Flanagan, J. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-359.

Fraser, J. M., Strohm, P., Smith, J. W. J., Galdes, D., Svirbely, J. R., Rudmann, S., Miller, T. E., Blazina, J., Kennedy, M., & Smith, P. J. (1989). Errors in abductive reasoning. *Proceedings of the 1989 IEEE International Conference on Systems, Man, and Cybernetics*, 1136-1141.

Gibson, J.J., (1996). The senses considered as perceptual systems. Boston: Houghton Mifflin.

Gruen, D., Sidner, C., Boettner, C., & Rich, C. (1999). A collaborative assistant for email. In *CHI 99 ACM Conference on Human Factors in Computing Systems*, New York: ACM Press. Pittsburgh, PA. 196-197.

Guerlain, S., Smith, P.J., Obradovich, J. H., Rudmann, S., Strohm, P., Smith, J.W., Svirbely, J., and Sachs, L. (1999). Interactive Critiquing as a Form of Decision Support: An Empirical Evaluation. *Human Factors, 41*(1), 72-89.

Harter, S.P. (1986). *Online information retrieval. Concepts, principles, and techniques.* New York: Academic Press.

Hersh, W.R., & Hickam, D.H. (1998). How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. *JAMA, 280*, 1347-1352.

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science, 44*(3), 161-174.

Hutchins, E. (1995). *Cognition in the Wild.* Cambridge, MA: MIT Press.

Iivonen, M. (1995). Consistency in the selection of search concepts and search terms. *Information Processing and Management, 31*(2), 173-190.

Josephson, J., & Josephson, S. (1994). *Abductive Inference.* New York, NY, Cambridge University Press.

Keim, D.A. (1997). Pixel-oriented visualization techniques for exploring very large data bases. *Journal of Computational Graphics and Statistics 5*(1).

Keim, D.A. & Kriegel, H. (1994). VisDB: Database Exploration Using Multidimensional Visualization. *IEEE CG&A, 14*(5), 40-49.

Keim, D.A., & Kriegel, H.-P. (1996). Visualization Techniques for Mining Large Databases: A Comparison. *Transaction on Knowledge and Data Engineering, 8*(6), 923-938.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211-228.

Klein, G.A., Calderwood, R., & MacGregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(3), 462-472.

Koenemann, J., & Belkin, N.J. (1996). A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *CHI '96. Proceedings of the ACM SIG CHI Conference on Human Factors in Computing Systems.* New York: ACM. 205-212.

Kuhlthau, C.C. (1999). The role of experience in the information search process of an early career information worker: perceptions of uncertainty, complexity, construction, and sources. *Journal of the American Society for Information Science, 50*(5), 399-412.

Lamping, J., Rao, R., & Pirolli, P. (1995). A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems.* ACM, May.

Lee, J.H. (1998). Combining the evidence of different relevance feedback methods for information retrieval. *Information Processing and Management, 34*(6), 681-691.

Leipzig, N., Kozak, M.G., & Schwartz, R. (1983). Experiences with end-user searching at a pharmaceutical company. In M.E. Williams & T.H. Hogan (Eds.) *Proceedings of the 4th National Online Meeting,* Medford, NJ: Learned Information, pp. 325-332.

Letsche, T.A. & Berry, M.W. (1997). Large-scale information retrieval with latent semantic indexing. *Information Sciences,* 100, 105-137.

Maes, P. (1998). Agents that reduce work and information overload. In M.T. Maybury, W. Wahlster *Readings in Intelligent User Interfaces.* Morgan Kaufmann Publishers ISBN 1-55860-444-8.

Maes, P. & Schneiderman, B. (1997). Direction manipulation vs. interface agents: a debate. *Interactions, 4*(6), ACM Press.

Marx, M., & Schmandt, C. (1996). CLUES: Dynamic Personalized Message Filtering. In *CSCW '96 Proceedings,* pp.113-121.. Boston, MA.

McSherry, D. (1997). Avoiding premature closure in sequential diagnosis. *Artificial Intelligence in Medicine, 10,* 269-283.

Meadow, C.T. & Cochrane, P. (1981). *Basics of online searching.* New York: John Wiley & Sons, pp. 136-141.

Mizzaro, S. (1997). Relevance: the whole history. *Journal of the American Society for Information Science, 48*(9), 810-832.

Morse, E & Lewis, M. (1997). Why information retrieval visualizations sometimes fail. In *Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics,* Oct. 12-15, 1997, 1680 – 1685.

Norman, D. A. (1981). Categorization of action slips. *Psychological Review, 88*(1), 1-15.

Norman, D. A., & Draper, S. W. (1986). *User Centered System Design: New Perspectives on Human-Computer Interaction.* Hillsdale, New Jersey: Lawrence Erlbaum.

Oakes, M.P., & Taylor, M.J. (1998). Automated assistance in the formulation of search statements for bibliographic databases. *Information Processing and Management, 34*(6), 645-668.

Obradovich, J. H. & Woods, D. D. (1996). Users as designers: How people cope with poor HCI design in computer-based medical devices. *Human Factors, 38*(4), 574-592.

Olsen K.A., Sochats K.M., & Williams, J.G. (1998). Full text searching and information overload. *International Information and Library Review, 30*(2), 105-122.

Patterson, E.S., & Woods, D.D. (1997). Shift Changes, updates, and the on-call model in space shuttle mission control. *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting.* Albuquerque, NM., 243-247.

Patterson, E. S., Woods, D. D., Sarter, N. B., & Watts-Perotti, J. (1998). Patterns in cooperative cognition. *Proceedings of COOP '98, Third International Conference on the Design of Cooperative Systems.* Cannes, France, 26-29 May.

Pirolli, P., Schank, P., Hearst, M., & Diehl, C. (1996). Scatter/Gather browsing communicates the topic structure of a very large text collection. *Proceedings of CHI '96, ACM Conference on Human Factors in Computing Systems,* New York, 213-220.

Potter, S. S., Roth, E. M., Woods, D. D. & Elm, W. (in press). Bootstrapping Multiple Converging Cognitive Task Analysis Techniques for System Design. In Schraagen, J.M.C., Chipman, S.F., & Shalin, V.L. (Eds.), *Cognitive Task Analysis.* Mahwah, NJ: Lawrence Erlbaum Associates.

Potter, S. S. & Woods, D. D. (1991). Event-driven timeline displays: Beyond message lists in human-intelligent system interaction. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, IEEE.*

Potter, S. S., Woods, D. D., Hill, T., Boyer, R., & Morris, W. (1992). Visualization of dynamic processes: Function-based displays for human-intelligent system interaction. In *Proceedings of IEEE International Conference on Systems, Man, and cybernetics.*

Pratt, W. and Sim, I. (1995). Physician's information customizer (PIC): using a shareable user model to filter the medical literature. *Medinfo 8*(2), 1447-1451.

Punch, B., & Josephson, J. (unpublished document). A real-time algorithm for diagnosis using abductive assembly.

Quintana, Y. (1998). Intelligent medical information filtering. *International Journal of Medical Informatics, 51,* 197-204.

Rabbitt, P. (1984). The Control of Attention in Visual Search, in R. Parasuraman and D. R. Davies (eds.)., *Varieties of Attention,* New York: Academic Press.

Reed, E.S. (1988). *James J. Gibson and the Psychology of Perception.* New Haven CT: Yale University Press.

Robertson, G.G., Mackinlay, J.D., & Card, S.K. (1991). Cone trees: Animated 3D Visualizations of Hierarchical Information. *Proceedings of CHI '91, ACM Conference on Human Factors in Computing Systems,* New York, 189-194.

Roth, E.M., Bennett, K., & Woods, D.D. (1987). Human interaction with an "intelligent" machine. *International Journal of Man-Machine Studies, 27,* 479-525.

Roth, E.M., Woods D.D., & Pople, H.E. Jr. (1992). Cognitive simulation as a tool for cognitive task analysis. *Ergonomics, 35,* 1163–1198.

Rubin, K.S., Jones, P.M., & Mitchell, C.M. (1988). OFMspert: inference of operator intentions in supervisory control using a blackboard architecture. *IEEE Transactions on Systems, Man, and Cybernetics, 18*(4), 618-637.

Sackett D.L, & Straus S.E. (1998). Finding and applying evidence during clinical rounds: the "evidence cart." *JAMA, 280,* 1336-1338.

Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM, 29*(7), 648-656.

Saracevic, T., Kantor, P., Chamis, A.Y., & Trivison, D. (1988). A study of information seeking and retrieving (3 parts). *Journal of the American Society for Information Science, 39,* 161-216.

Sarter, N., Woods, D. D., & Billings, C. E. (1997). Automation Surprises. In G. Salvendy (ed.). *Human Factors/Ergonomics* (2nd edition). Wiley, NY.

Sarter, N. B., & Woods, D. D. (1997). "Teamplay with a Powerful and Independent Agent": A Corpus of Operational Experiences and Automation Surprises on the Airbus A-320. *Human Factors, 39*(4), 553-569.

Sarter, N. B., & Woods, D. D. (1994). Pilot Interaction with Cockpit Automation II: An Experimental Study of Pilot's Model and Awareness of the Flight Management System. *International Journal of Aviation Psychology, 4,* 1-28.

Sarter, N. B., & Woods, D. D. (1992). Pilot Interaction with Cockpit Automation I: Operational Experiences with the Flight Management System. *International Journal of Aviation Psychology, 2,* 303–321.

Schum, D.A. (1994). *The Evidential Foundations of Probabilistic Reasoning.* New York: John Wiley and Sons.

Sewell, W., & Bevan, A. (1976). Nonmediated use of MEDLINE and TOXLINE by pathologists and pharmacists. *Bulletin of the Medical Library Association, 64*(4), 382-391.
Shute, S. J., & Smith, P. J. (1992). Knowledge-based search tactics. *Information Processing & Management 29*(1), 29-45.

Siegel, S. & Castellan, Jr., J.N. (1988). *Nonparametric statistics for the behavioral sciences,* 2nd ed., New York: Mc Graw-Hill.

Siegfried, S., Bates, M.J., & Wilde, D.N. (1993). A profile of end-user searching behavior by humanities scholars: The Getty online searching project report no. 2. *Journal of the American Society for Information Science, 44,* 273-291.

Sievert, M. & Glazier, R. (1990). Journalists' searching of a menu-driven multi-database retrieval system: A pilot study. In D. Henderson (Ed.), *ASIS '90: Proceedings of the 53rd ASIS Annual Meeting* Medford, NJ: Learned Information, pp. 39-46.

Smith, P.J., McCoy, E., & Layton, C. (1997). Brittleness in the design of cooperative problem-solving systems: the effects on user performance. *IEEE Transactions on Systems, Man, and Cybernetics, 27*(3), 360-370.

Spink, A., & Saracevic, T. (1997). Interaction in information retrieval: selection and effectiveness of search terms. *Journal of the American Society for Information Science, 48*(8), 741-761.

Srinivasan, P. (1996). Query expansion and MEDLINE. *Information Processing and Management, 32*(4), 431-443.

Stone, M.C., Fishkin, K., & Bier, E.A. (1994). The Movable Filter as a User Interface Tool. In *CHI 94 ACM Conference on Human Factors in Computing Systems,* New York: ACM Press. Boston, MA, 306-312.

Svenson, O., & Maule, J. (1993). *Time pressure and stress in human judgment and decision making.* New York: Plenum Press.

Tufte, E. R. (1990). *Envisioning Information.* Chesire, CT: Graphics Press.

Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Vicente, K. J. & Rasmussen, J. (1992). Ecological interface design: Theoretical Foundations. *IEEE Transactions on Systems, Man, and Cybernetics, 22*(4), 589-606.
Vicente, K.J. (1996). *Improving dynamic decision making in complex systems through ecological interface design: A research overview.* System Dynamics Review, *12*(4), 251-279.

Walton, K.R., & Dedert, P.L. (1983). Experiences at Exxon in training end-users to search technical databases online. *Online, 7*(5), 42-52.

Wildemuth, B.M., de Bliek, R., Friedman, C.P., & File, D.D. (1995). Medical students' personal knowledge, searching proficiency, and database use in problem solving. *Journal of the American Society for Information Science, 46*(8), 590-607.

Wilson, P. (1992). "Searching: Strategies and Evaluation." In *For Information Specialists: Interpretations of Reference and Bibliographic Work,* ed. Howard D. White, Marcia J. Bates, & Patrick Wilson, 153-181. Norwood, NJ: Ablex.

Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1996) Visualizing the Non-Visual: Spatial analysis and interaction with information from text documents. *Proceedings of Info Viz 96.*

Wolfe, J. M. (1992). The parallel guidance of visual attention. *Current Directions in Psychological Science, 1,* 124-128.

Woods, D.D. (1995). Toward a theoretical base for representation design in the computer medium: Ecological perception and aiding human cognition. In Flach, J.M., Hancock, P.A., Caird, J.K., and Vicente, K.J. (Eds.) *An Ecological Approach to Human Machine Systems, Vol. I: A Global Perspective.* Hillsdale NJ: Erlbaum, pp. 157-188.

Woods, D.D. (1994a). Visual Momentum: A concept to improve the cognitive coupling of person and computer. *International Journal of Man-Machine Studies, 21,* 229-244.

Woods, D.D. (1994b). Cognitive demands and activities in dynamic fault management: abductive reasoning and disturbance management. In N. Stanton (Eds.), *Human factors in alarm design* Bristol, PA: Taylor and Francis.

Woods, D. D. (1993). Process tracing methods for the study of cognition outside of the experimental psychology laboratory. In G. Klein, J. Orasanu, and R. Calderwood (Eds.), *Decision Making in Action: Models and Methods.* Norwood, NJ: Ablex Publishing Corporation.

Woods, D. D. (1984). Visual momentum: A concept to improve the cognitive coupling of person and computer. *International Journal of Man-Machine Studies, 21,* 229-244.

Woods, D. D., Patterson, E. S., Roth, E. M., & Redenbarger, W. J. (1998). Aiding the Intelligence Analyst in Situations of Data Overload: The Zairean Civil War Scenario. Institute for Ergonomics/Cognitive Systems Engineering Laboratory Report, *ERGO-CSEL 98-TR-01,* The Ohio State University, Columbus OH, January 1998. Prepared for Armstrong Laboratory Crew Systems Integration Branch (AL/CFHI), WPAFB.

Woods, D.D., Patterson, E.S., & Roth, E.M. (1998). Aiding the Intelligence Analyst in Situations of Data Overload: A Diagnosis of Data Overload. Institute for Ergonomics/Cognitive Systems Engineering Laboratory Report, *ERGO-CSEL 98-TR-03,* The Ohio State University, Columbus OH. Prepared for Armstrong Laboratory Crew Systems Integration Branch (AL/CFHI), WPAFB.

Woods, D.D., Patterson, E.S., & Tinapple, D. (in preparation). Context-sensitive, model-based representation aiding for data overload in inferential analysis.

Woods, D.D., Pople, H.E., Jr. & Roth, E.M (1992). Cognitive Environment Simulation: A tool for modeling intention formation in human reliability analysis. *Nuclear Engineering and Design, 134,* 371–380.

Woods, D. D. & Watts, J. C. (1997). How not to have to navigate through too many displays. In M. Helander (Ed.), *Handbook of Human-Computer Interaction, 2nd edition.* Elsevier Science Publishers, B.V., North Holland.

Wright, W. (1995). Information animation applications in the capital markets. *Proceedings of InfoVis '95,* IEEE Symposium on Information Visualization, New York, 19-25.

Yuan, W. (1997). End-user searching behavior in information retrieval: a longitudinal study. *Journal of the American Society for Information Science, 48*(3), 218-234.

Zhang, J. & Norman, D.A. (1994). Representations in distributed cognitive tasks. *Cognitive Science, 18,* 87-122.