Foresight

Infectious Diseases: preparing for the future

OFFICE OF SCIENCE AND INNOVATION

S4: Data mining and fusion

Colin Upstill, Matthew Addis, Freddy Choi,
Steve Taylor, and Rowland Watkins
University of Southampton IT Innovation Centre

This review has been commissioned as part of the UK Government's Foresight project, Infectious Diseases: preparing for the future. The views expressed do not represent the policy of any Government or organisation.

Contents

1 Introduction 3
2 Data sources 5
3 Data fusion 5
4 Data mining 9
5 Enterprise information integration 12
6 Sequential and stream data 13
7 Unstructured data 13
7.1 Computational linguistics and text mining 14
7.2 Multimedia content 15
8 Distributed data fusion and mining 17
8.1 Web mining 17
8.2 Web content mining 17
8.3 Web usage mining 18
8.4 Web structure mining 18
8.5 Grid-based data fusion and mining 18
9 Privacy 21
10 Provenance 23
11 Realising the vision of DIID 23
11.1 The challenges of DIID 24
11.2 Research communities and technologies 25
11.2.1 Web services 26
11.2.2 Workflow 26
11.2.3 Semantic web 26
11.2.4 Semantic web services 27
11.2.5 Grid services 27
11.2.6 Semantic grid services 27
11.2.7 Peer-to-peer networks 28
11.2.8 Agent-based systems 28
11.2.9 Autonomic Computing 28
11.2.10 Blogs, Wikis and collaborative personal communication 29
12 The future 29
13 Acknowledgements 31
14 References 31

1. Introduction

Data is raw and does not, of itself, have meaning, whereas information is data that has been processed to be useful, given meaning by way of relational connections, semantics, etc. Knowledge results from reasoning over information. The term ‘data’ is often used when, strictly speaking, ‘information’ is the correct word.

Data fusion is a set of techniques for combining data, which may be noisy or conflicting, from multiple, heterogeneous sources. Data mining is the analysis of data to establish relationships and identify patterns.

Data fusion and data mining can be considered as a pipeline that enables data from a wide range of heterogeneous sources to be used in applications such as anomaly detection, hypothesis testing, and epidemiological model calibration for the detection and identification of infectious diseases (DIID) in plants, animals and humans. This pipeline connects sensor networks and data sources with the processes and tools that allow this data to be located, datasets to be selected, and data to be collected, fused, visualised and mined. The information extracted enables predictive models to be constructed, conclusions to be drawn, decisions made and actions taken.

Fusing data from multiple heterogeneous sources and mining a single, homogeneous database are well-established techniques. Distributed data fusion and mining, where the analysis is performed across a network of computers, are emergent technologies. For DIID, the data will be distributed and heterogeneous, due to the wide-ranging (and potentially global) nature of the problem, and the many ways of reporting behaviour and symptoms that may characterise infectious disease. The task is to determine patterns that may occur in a wide variety of data (sometimes linking and cross-referencing data from different sources) in order to detect abnormalities that may indicate infectious disease.

One of the main challenges for DIID is how to analyse data from such distributed heterogeneous sources. Data can be continuous or categorical (e.g. numerical values or discrete tags such as ‘bread’), and it can be structured or unstructured. Structured data sources will be as disparate as parametric data from sensors, medical records, satellite images, CCTV images, audio, and so on. Unstructured data is likely to be important for DIID. Sources such as news feeds and emails are well-established, and new technologies, such as Blogs, Wikis and other methods of personal communication that will supersede them, are spreading rapidly. Analysis of such data poses particular challenges, not least the problem of semantic mapping between domains, but it is likely to yield valuable information for DIID.

We note that some of the challenges facing data analysis in DIID are analogous to those faced by a computer network intrusion detection system (IDS), used to detect unauthorised access and potentially malicious behaviour (Denning 1986). An IDS is characterised by a set of data sources and a process that monitors these sources, with the aim of detecting abnormal

behaviour. This may be performed manually by examining the source data or automatically by, for example, a rule-based system that looks for patterns in the source data.

For DIID, we note the paramount importance of ethics, sovereignty, confidentiality (commercial and otherwise), privacy, data protection, freedom of information, and other rights and obligations when detecting, collecting and processing data. Provenance and trustworthiness of data and information must be factors in any decisions.

Many of the above issues are underpinned by IT security. Security is an attribute of systems and procedures that minimises or manages the risk of undesirable system behaviour, even in the presence of malicious, untrusted parties. The internet presents new challenges for information security. It is no longer possible to think of data as contained within an organisation or administrative domain, accessible only to those with prescribed attributes. The networked enterprise must allow its perimeter to become permeable to support dynamic collaborations, and yet still be capable of controlling access. However, each participating organisation and system has its own security mechanisms, making it very difficult to maintain consistency across different technical, organisational or administrative domains. This presents particular problems for applications involving personal data, where consistent handling is essential in order to maintain the levels of privacy demanded by citizens and required by policies, regulations and laws.

The success of data fusion and data mining depends as much on the adoption of appropriate methodologies and processes as it does on the availability of suitable data and the use of appropriate technology. Without the formulation of a well-defined business or research problem, the assembly of trusted and representative data sources, and a way to validate the results, the output of a data fusion or mining exercise will be untested at best and could be positively misleading at worst. In data mining and data fusion: ‘If you put rubbish in, then you’ll get rubbish out.’ Moreover, the watchword is: ‘Make sure you answer the question.’ If you are not sure what the question is in a data mining or fusion exercise, the outputs cannot be interpreted and acted on with confidence. Uncertainty leads to risk, and risk in turn leads to error and cost, which in the case of DIID could be extremely serious.

An overarching, common methodology and process is important, otherwise *ad hoc* techniques will be used and conclusions will be difficult to justify, as they are unlikely to be repeatable. A methodology is just as important here as in any other branch of science, whereby hypotheses are posed, experiments are conducted to prove or disprove these hypotheses, results are derived and conclusions are drawn. When the results and conclusions are published, others can attempt the same experiment to validate, or contradict, these conclusions.

2. Data sources

Data for analysis for DIID will typically come from a variety of sources. A major issue with such data is that it may be incompatible – from simple mismatches, such as the use of different date representations, to more subtle matters of semantics and interpretation. In general, the semantic gap refers to a mismatch between understandings across domains. Zhao et al. (2002) refer to the semantic gap that results from a user's web search and how the search engines interpret and perform the searches. If the user understands a concept differently from the search engine, a semantic gap arises, as the results will not be what the user expects.

Semantic gaps may arise in DIID due to the heterogeneous nature of the different sources of data. Ensuring that the same term in two different sets of data actually means the same thing, and establishing the appropriate transformation rules, is a major challenge. Even realising that there may be a difference in the meaning of the two terms is generally difficult.

It is also important to distinguish between structured and unstructured data. Structured data (e.g. derived from a form or a database table) is much more easily mined than unstructured data. Traditional data mining tools are designed for structured data. Unstructured data such as. a web page containing a news story is more of a challenge. It either needs to be transformed into structured data (involving disciplines such as natural-language processing and semantic mapping), or specialist data mining tools need to be created.

Often, data from many different sources needs to be combined. The resulting conjunction of more than one dataset may be greater than the sum of its parts, and may produce crucial discoveries. Also, data not directly connected with infectious diseases may be extremely useful. For example, in order to track a disease’s spread, symptom or other primary data will probably need to be cross-correlated with other information from, for example, geographical information systems (GIS), weather reports and travel records.

International awareness, co-operation and joint research efforts have engendered the necessary political platform and international community for global environment monitoring (e.g. UNEP 2005). Cross-border remote sensing and data distribution networks (e.g. UNEP 2004) have been created to make a wide variety of information from different sources and different geographical locations available for analysis.

3. Data fusion

Data fusion is a set of techniques for combining diverse data from multiple sources to create a more structured and coherent view of the data, thus making it possible to conduct further analysis (e.g. data mining, decision support). The technology was created primarily for military applications where there is a regular need to quickly assess a complex situation based on a wide range of observations (e.g. target tracking with multiple cameras) and to

determine the optimal course of action (e.g. launching a missile). The technology is now used in many military and commercial applications for making complex information accessible to decision makers. Example applications include surveillance, situation assessment, robotics, manufacturing, medical diagnosis and remote sensing (Hall and McMullen 2004).

Data fusion is a combination of many disciplines. Communication and data management technologies focus on the organisation, storage, preservation and distribution of data. Mathematics, computer science and artificial intelligence all contribute to the development of automatic and principled methods for combining, restructuring and summarising diverse, incomplete and conflicting information.

Data fusion covers an entire process: data gathering from multiple sources, data format conversion, data combination, conflict resolution, data summarisation and distribution. The process takes input from heterogeneous sources and produces a coherent summary that makes it possible for decision makers to quickly assess a complex situation and determine the best course of action. A decision maker can either be a person (e.g. military officer, business analyst) or an automated system (e.g. for generating alerts).

Rapid advances in sensing, communication and storage technologies have created an explosion in data volume, sources and formats. Data integration makes large volumes of disparate data sources accessible via a common framework, often referred to as the data broker. The process involves establishing a communication channel between a data supplier (e.g. a sensor) and the data broker, to facilitate secured information transfer. The data broker provides the access, selection and transformation capabilities that enable a data consumer (e.g. a user or a data fusion system) to obtain all the relevant data in the required format. This involves gathering all the required data from multiple suppliers in multiple formats and repackaging the data in the required format for delivery to the data consumer.

The emergence and wide adoption of data exchange standards (e.g. Yergeau et al. 2004) has improved data sharing and reuse by encouraging suppliers to provide data in a common format (e.g. XML) and to use metadata to describe the data. This has enabled improvements to be made to automated search and retrieval systems. Advances in distributed mass storage (e.g. RAID 2005) and database management systems (e.g. Oracle 2005) have made it possible to store, transform and distribute large, heterogeneous datasets. Current work in data integration continues to focus on the integration of heterogeneous data from proprietary systems. The main challenges include information extraction from semi-structured data formats (e.g. spreadsheets), business process optimisation, integration and automation (e.g. data validation and delivery), standardisation of secured data exchange channels (e.g. policy-based access control), metadata standards (e.g. data description vocabulary and ontology) and data quality assurance (e.g. accuracy information). Different aspects of a situation may be described by information from multiple sources. Observation accuracy and coverage are improved by combining and comparing overlapping and non-overlapping information. Information can be represented

in disparate and ambiguous forms (e.g. numbers, symbols, rules and natural-language statements). Some can be described in statistical form, whilst others are more difficult to model and process mathematically. Data fusion is about the combination of information for creating a more accurate and coherent assessment of a situation. The process involves combining and consolidating information at different levels of detail using well-established mathematical techniques, including Bayesian inference (Punska 1999), fuzzy logic (Zadeh 1965), neural networks (Aleksander and Morton 1990), decision trees (Quinlan 1993), support vector machines (Cristianini and Shawe-Taylor 2000) and Kalman filters (Kalman 1960). A comprehensive review of the techniques can be found in Klein (2004) and Hall and Llinas (2001).

The theoretical foundation for data fusion (Goodman and Nguyen 1985; Goodman et al. 1997; Daley and Vere-Jones 1988; Hall and Llinas 2001) is a rigorous mathematical formulation that is hard to translate into an engineering solution (Mahler 2004). Data manipulation, modelling, integration and forecasting algorithms are increasingly mature. Theoretical research continues to focus on the formulation of a unifying and purely probabilistic framework (Mahler 1994; Goodman et al. 1997) for combining diverse and ambiguous information (Mahler 2000). New approximation techniques are being sought to facilitate the integration of more information sources. Current research focuses on the computational complexity of generalised system-level solutions and the application of principled approximation techniques to make a solution computationally tractable.

A standard model for data fusion was proposed by the US Department of Defense to facilitate discussions, component reuse and system integration. The Joint Directors of Laboratories (JDL) data fusion model (White 1988) offers a multi-level functional model that describes how processing is organised in a military data fusion system. The model can be generalised, and it has been widely adopted in commercial applications and academic research. It is continuously being revised, refined and expanded to accommodate new requirements (Bowman 1994; Steinberg et al. 1999; Llinas et al. 2004; Steinberg and Bowman 2004). The JDL data fusion model is recognised as the *de facto* standard in data fusion and is likely to remain so for the foreseeable future.

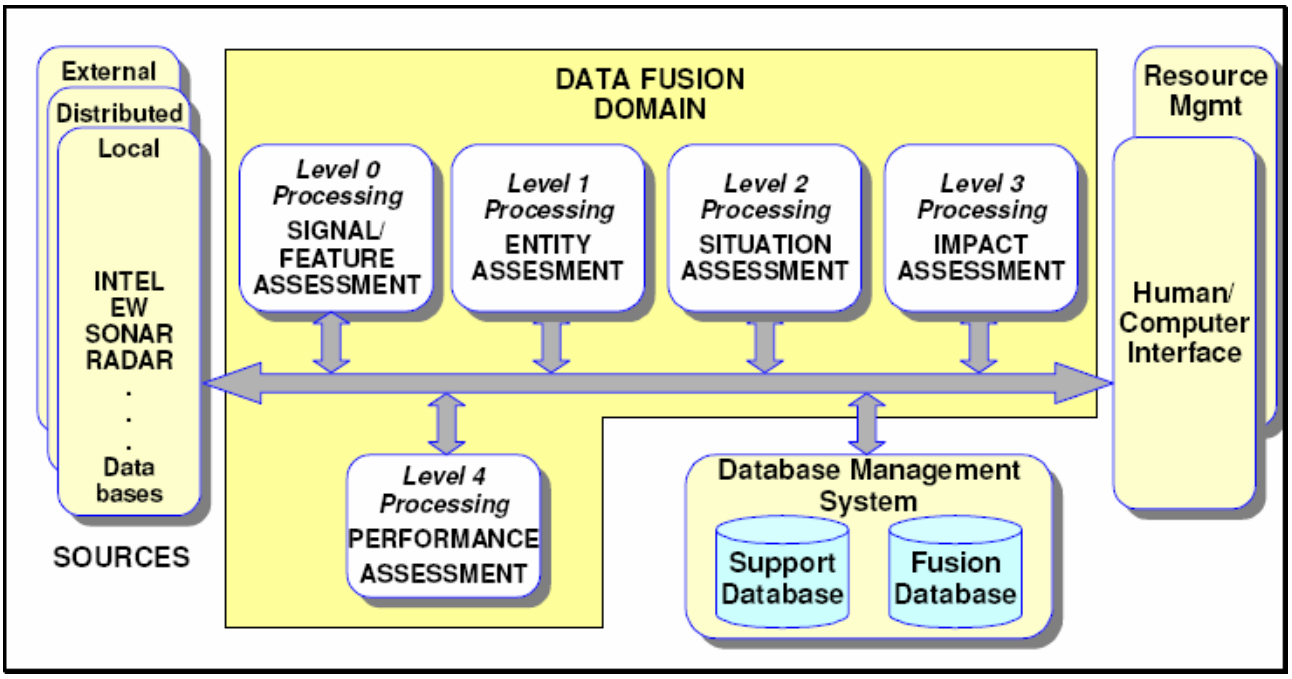

Figure 1:. The JDL data fusion model (Steinberg and Bowman 2004)

The latest incarnation of the JDL model (Llinas et al. 2004; Steinberg and Bowman 2004) defines five functional levels and a complementary set of resource management levels. The purpose of each functional level can be summarised as follows:

- Level 0 extracts interesting features from raw data, e.g. enhancing a satellite image to highlight colour variations in a field of crops.
- Level 1 gathers information about individual entities, e.g. the colour, location and health condition of each plant or cluster of plants.
- Level 2 focuses on the relationships between entities and contextual implications, e.g. estimating and predicting the spread of disease among plants based on wind direction and geographical factors.
- Level 3 assesses the consequence of applying known plans on the current situation, e.g. predicting the impact of the disease on crop production if remedial action is or is not carried out within the next 24 hours.
- Level 4 measures the performance and effectiveness of the system to facilitate refinement, e.g. comparing the predicted and observed spread of a disease to determine and adjust the contribution of each feature to the overall prediction.

Signal processing (Level 0) and multi-sensor integration (Level 1) are increasingly mature fields of research, whereas the higher levels of information fusion (Level 2–4) are in their infancy due to the absence of unifying, theoretical foundations (Mahler 2004) and computationally tractable, algorithmic solutions.

Advances in materials science, computer science, electronics, nanotechnology, communication technology, engineering and manufacturing

have all contributed to the development of robust, precise and affordable sensor technologies and the associated data management framework. Stationary and mobile wireless sensor networks for measuring a wide range of parameters in extreme environments have been developed (see e.g. Karl and Willig 2005; Zhao and Guibas 2004).

Advances in sampling techniques, data modelling, classification algorithms, mathematically rigorous inference methods and computation hardware will have a significant impact on data fusion. Principled sampling techniques and faster computational hardware enable systems to handle more data. Data modelling algorithms make it possible to process noisy data and identify abnormalities in it. Classification algorithms identify salient features in a dataset. Current work focuses on the development of methods that can solve high-dimensional, non-linear classification problems, thus enabling a system to assess more complex situations. Mathematically rigorous inference methods facilitate principled and complex reasoning with disparate data. The combination of all these advances will enable future data fusion systems to use high-resolution observation data from diverse information sources to accurately assess more complex situations, thus enabling decision makers to make informed decisions about large, real-world events.

4. Data mining

Data mining is a relatively mature technique aiming to achieve business benefit by providing tools that assist in the discovery of patterns and relationships in large amounts of data, and in the prediction of the values of unseen data based on information gained from seen data.

Current technology is based on well-established mathematical techniques for identifying patterns in data. These are techniques for understanding and modelling data. Summaries (aggregations of data), visualisation, clustering of data records, and the discovery of associations and correlations between datasets aid the understanding of data, and enable the acquisition of insights for more detailed analysis. Mathematical models enable predictions of the classification and value of unseen data.

The techniques are mature for data that has been structured into records with clearly defined data types and collected in one place. Such a data warehouse is a single, authoritative source of data, integrated from distributed (and possibly differently structured) databases to facilitate a global overview and comprehensive analysis of data at or up to a specific time. This is in contrast with operational data systems e.g. transaction processing systems, where the data is changing with every transaction (order, payment, or whatever). See e.g. Data Warehousing Wikipedia.

The maturity of these data mining techniques is indicated by their deployment in a number of commercial products and their migration into undergraduate syllabuses. Goharian (2004) describes a process where undergraduates not only use data mining tools, but also build them. Where mature techniques exist, data mining is moving from the lab of the computer scientist to the office

of the domain expert. These mature techniques are well described elsewhere: a brief introduction to data mining is given by Moss (2003), and data mining tools and techniques are introduced and discussed by Witten and Frank (2005).

The use of XML (eXtensible Markup Language) as a representation format can result in any amount of structure in data. XML is similar to the HTML used for web pages but is more flexible. XML documents can also import namespaces and refer to ontologies which provide meaning to the document, and can help to address semantic gaps. The mathematics of Graph Theory offers a means of mining semi-structured data such as HTML and XML. Graph-based data mining aims to find structures embedded in semi-structured data. See Washio and Motoda (2003) for an overview.

When faced with massive amounts of data, sampling may be used to reduce this to manageable proportions. How the sampling is performed is critical to ensuring that the sample accurately represents the dataset as a whole. Examples of algorithms that attempt to automate the sampling problem are FAST (Chen et al. 2002) and EASE (Brönnimann et al. 2003).

How do we determine when we have found a significant pattern? It is one thing to find correlations, but the key to exploitable results is to uncover causal relationships. For example, we discover in a shopping basket analysis that most customers who bought butter also bought bread, and we make a rule that says: ‘Butter implies bread.’ This seems like a good discovery, until we realise that almost everyone bought bread, whether they bought butter or not. The bare correlation of butter implying bread does not constitute information gain, since almost everyone bought bread anyway. Only if we know that butter is not universally purchased, and that the presence of bread significantly increases the chances of butter being found in the same basket, do we have a knowledge gain. This is known as ‘lift’.

Current work in this field concerns the theory of expected information (see e.g. Robson 2003). This involves the computation of degrees of mutual dependency between variables, and the expected information gain from a pattern. Based on Robson's work, IBM has demonstrated a tool (see CliniMiner) that provides heuristics and other tools to detect rules that do not contain significant information gain and prune them from a larger set of discovered rules.

Several methodologies have been developed for data mining, of which the most widely known and commonly adopted (kdnuggets 2004) is CRISP-DM (Cross-Industry Standard Process for Data Mining; see CRISP-DM project), created by a European consortium in the mid-1990s.

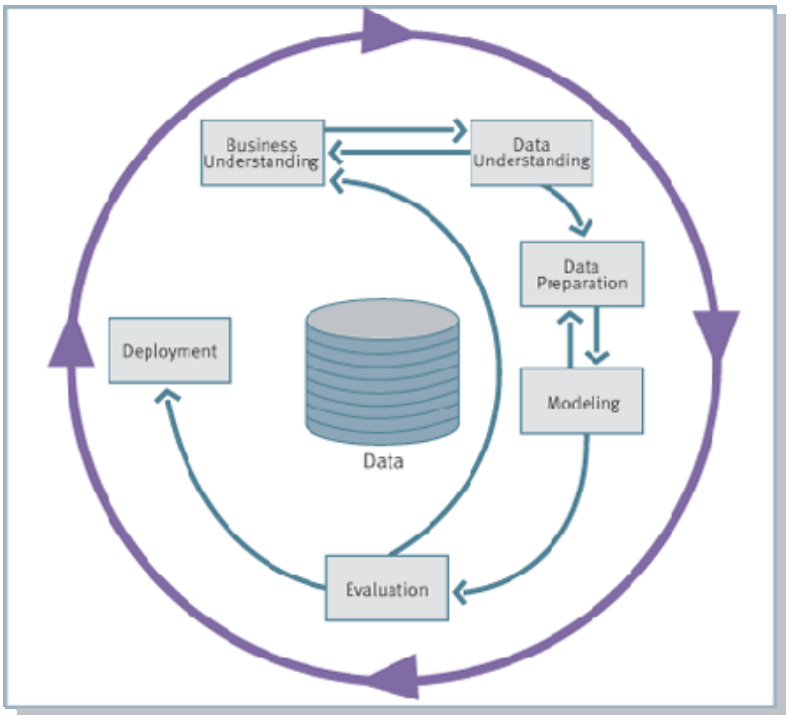

Figure 2: The CRISP-DM model (reproduced from the CRISP-DM project)

Figure 2 shows the CRISP-DM model. In it, the lifecycle of a data mining project consists of six phases. The sequence of the phases is not strict – moving back and forth between different phases is always required, depending on the outcomes of each phase, and which phase or which particular task of a phase has to be performed next. The overall process is cyclic, iterative and continues after initial deployment of the results. The six main phases are as follows.

1. Business Understanding, which focuses on understanding the project objectives and requirements from a business perspective, and converting this into a data mining problem definition and plan.

2. Data Understanding, which starts with data collection and includes activities to discover first insights into the data, to detect interesting subsets and form hypotheses about hidden information, and to identify data quality problems. There is a strong element of 'playing' with the data in this phase – the analyst can cut the data, visualise it, find out ranges and perform analyses on it in many different ways. The result is a deeper understanding of the data and clues for the next phases of the process: for example, which features may be the most promising for further analysis.

3. Data Preparation, which covers all activities to construct the final dataset (data that will be fed into the modelling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order.

4. Modelling, where various techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

5. Evaluation follows construction of a model (or models) that, from a data analysis perspective, appears to have high quality. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and to review the steps executed to construct it, in order to be certain that it properly achieves the business objectives.

6. Deployment is where the knowledge gained will need to be organised and presented in a way that can be used, i.e. action can be taken.

Other data mining processes include the SAS Institute's SEMMA, an acronym for Sample, Explore, Modify, Model, Assess (see SEMMA). SEMMA is not a data mining methodology in itself, and can be used within an overarching framework such as CRISP-DM, where steps such as formulating a well-defined business or research problem and assembling quality representative data sources are critical to the overall success of a data mining project.

CRISP-DM, SEMMA and other methodologies and processes have typically arisen to deal with data mining in relatively controlled and centralised environments, for example, within an enterprise. As distributed computing and grid technologies have developed, data mining is starting to be applied in much more distributed and heterogeneous environments, where there is no longer centralised control or enforcement of standards, practices and processes. This is exactly the scenario in which DIID is likely to take place.

5. Enterprise information integration

Enterprise information integration is now regarded as fundamental to corporate life, and is appearing in commercial and off-the-shelf products. Its emphasis is on making the large amount of disparate information owned by a typical enterprise indexable, searchable and behave like a single repository, even though it may be physically distributed and highly heterogeneous.

An example of this type of tool is the IBM Information Integrator (see DB2II), which enables searching of many different types of data, from databases to text, and on different platforms, internal and the web. Information Integrator also provides a single point of access with an SQL-like interface directly connectable to data mining suites. The data mining is, however, still centralised. Distributed data mining, performed across a network of computers, and the mining of data within transaction processing systems, where the data is changing with every transaction (e.g. stream data mining), are emergent technologies in this area.

6. Sequential and stream data

Data fusion algorithms can make use of data that is changing over time by adding a time parameter to the calculations. There are several well-established techniques for processing data which is sequential in time, for instance, ARIMA (AutoRegressive Integrated moving Average; see e.g. Pankratz 1983; Brockwell and Davis 1996) and GARCH (Generalised AutoRegressive Conditional Heteroskedasticity, due to Bollerslev 1986; see also Hamilton 1994), which can keep track of a changing mean and variance.

If a database or data source is changing over time, the addition of time series in data mining allows the discovery of rules that concern not only a set of antecedents, but also their order in time (their sequence). An example quoted by Masseglia et al. (2003) is that if someone rents the film *Star Wars*, followed by *The Empire Strikes Back*, there is a high chance that they will rent *Return of the Jedi*. Here, we can see the sequence of operations in the rental of the films in the antecedent. In this example, the rule will only become relevant after rental of *The Empire Strikes Back*, which is a subsequent event to rental of *Star Wars*. The body of discovered knowledge will also change over time as more events are added to the database. Masseglia et al. (2003) propose an algorithm named ISE (Incremental Sequence Extraction) that computes the frequent sequences in the updated database when new transactions and new customers are added to the original database.

The sequential aspect is also represented in stream-based data mining. This is where the data arrives as a stream, and a goal can be the detection of abnormalities in this stream. An example is in stock market share prices (Zhu and Shasha 2003), which are streamed to their consumers. Zhu and Shasha (2003) present an algorithm that uses a window of varying width, which is applied over the data stream to determine trends in the data. For example, how is a particular stock changing over time? We have its value over history, and when we get a new value, the trend will be adjusted given this new information. We may have a trigger that uses the gradient of the price change to determine if we buy or sell that stock.

The temporal aspect of data fusion and mining is highly relevant to DIID, as the detection of disease may occur in the change pattern of the source data rather than patterns found in static datasets.

7. Unstructured data

Well-established data mining techniques require the data to be structured into records with clearly defined data types and to be accessible as a single, authoritative source integrated from distributed (and possibly differently structured) databases. However, in addition to the structured data sources (e.g. record-based, forms, fields-with-values) that have fed traditional data mining systems and algorithms, a wealth of information is stored in unstructured data, and these sources are likely to contribute enormously to DIID.

There is much interest in applying search technology developed for the web to the files on a computer (e.g. Google Desktop). Tools such as these specialise in unstructured searches of many different types of data on a user's computer, including files of many formats, email and also web searches. Thus, the user can run queries on data without caring what format the data is in.

7.1 Computational linguistics and text mining

Computational linguistics and text mining are essential techniques for working with unstructured text and natural speech. This is a valuable capability for DIID.

Computational linguistics has been an active field of research for some 40 years. The aim is to create machines that can understand and generate natural language. Early work focused on surface-level and syntactic analysis, distinguishing abbreviations and sentence boundaries, identifying structured information such as time and date, and determining how words are combined to form a grammatical sentence. More recent work has been concerned with semantic analysis, determining and representing the meaning of words, sentences or texts, identifying people, locations, concepts and their relationships. The techniques have been widely applied to intelligence gathering, information retrieval and translation, for example.

The ability to extract information from different languages and present the result in a single language has been identified by the US Government as one of the grand challenges in national security (see GALE project). The Topic Detection and Tracking (TDT) and Message Understanding Competitions (MUC) have contributed to the rapid development of information extraction and retrieval technologies in recent years (see NIST for further information). Current systems can identify structured information (e.g. time, date, address), named entities (e.g. organisations, people, places), concepts (e.g. actions, objects) and their relationships (e.g. ‘he’, ‘she’, ‘it’ or ‘they’ references) in unstructured text (see ACL for further information). Existing solutions are typically domain-dependent. The research challenge is open domain, multi-lingual information extraction: a system that can understand everything in many languages.

Text mining is an increasingly mature field of research (Hearst 1999). State-of-the-art systems use information extraction technologies to gather structured information from free text (e.g. personal names), and information retrieval technologies to relate documents in a text collection, thus discovering new knowledge (e.g. friends of a friend). Text mining has been applied to bioscience literature to discover causes of rare diseases (Swanson and Smalheiser 1997) and biomedical literature to discover medically interesting genes (see BioText). Chung and& McLeod (2003) describe a topic mining framework that supports the identification of meaningful topics (themes) from news stream data. It aims to utilise the mapping from news feeds to content descriptions (ontologies) in order to determine the higher-level meanings of the stories. This involves clustering and hierarchical document searching in order to provide classifications that can be mapped onto the ontologies.

Advances in computational linguistics and text mining will enable an automated system to gather up-to-date information in a variety of languages from unstructured information sources and generate a translated summary of the information to aid decision support.

7.2 Multimedia content

Much of the data used in DIID will be non-textual, for example, satellite imagery, video surveillance footage, or photographs of the symptoms of a disease in the various stages of its lifecycle. As a result, image and video processing using content-based analysis techniques are likely to become significant features of data mining in DIID. Some examples are: automatically analysing satellite images to detect regions of crop failure based on colour; detecting body temperature anomalies using infrared video as people move through airports or other public places; and automatic diagnosis of a condition by analysing photographs of skin lesions and comparing them with known cases in a database. Analysis of video surveillance footage has already been applied to situations relevant to DIID, including surveillance in public transport (Sun 2004), and image analysis has been used as the basis for automated condition classification (Lewis 2004).

Image processing is a subset of the wider field of signal processing, and also forms the functional Level 0 of data fusion. Data fusion techniques would rarely be applied directly to raw image data because of the complexity involved in making inferences using such a large amount of data. Typically, image data is pre-processed using standard segmentation techniques before data fusion is applied.

Image and video processing is a large and active research field, and we do not review it in detail here. The state of the art can be assessed by examining the topics of discussion and paper presentations at international conferences in this area, including the Conference of Image and Video Retrieval (see CIVR), an international forum for discussing research challenges and exchanging ideas among researchers and practitioners in image/video retrieval technologies. It addresses innovative research in the broad field of image and video retrieval. A unique feature of the conference is the high level of participation from practitioners.

There is also significant activity at the European level through several projects supported by the European Commission, including: SCHEMA (see SCHEMA project) which aims to bring together a critical mass of industrial partners, end users, universities and research centres in order to improve the systematic exchange of information on content-based semantic scene analysis and& information retrieval; AceMedia (see AceMedia project) which is developing and implementing a system based on an innovative concept of knowledge-assisted, adaptive multimedia content management; and PrestoSpace (PrestoSpace project), which aims to apply image processing to particular problems such as indexing and the retrieval of audiovisual material.

A good survey of the underlying techniques used for content analysis can be found in the PrestoSpace report on the State of the Art of Content Analysis

Tools for Video, Audio and Speech (Bailer et al 2005), a survey of the tools and algorithms for analysis of audiovisual content for the purpose of metadata extraction.

Applications of image processing include video processing (for example, shot boundary detection and motion tracking), as well as analysis of single images (for example, Content Based Image Retrieval in large image libraries). Image processing techniques are often used to extract features that represent a particular aspect of an image in a compact and machine-processable form. These features can be used, for example, to measure the similarity between images as part of content-based retrieval. Also referred to as content descriptors, they can be generated at several levels. At the lowest level, image analysis is used to produce basic descriptors of colour (in various colour spaces), texture (repeating patterns in the image), and shape (for example, boundary and contour detection).

Lower-level descriptors are then used to extract higher-level features, for example, salient region identification and image segmentation. Image segmentation uses lower-level features such as edge detection and colour to segment an image into regions that are relatively uniform. Segmentation is often used as the basis for differentiating between foreground and background objects within images. Segmentation can be applied in a hierarchical (tree-like) scheme to segment at increasingly finer granularity. Image analysis can also be tailored to specific application areas such as human face location. In the case of face location within an image, a range of descriptors are used, for example, colour as the basis of segmenting an area with human skin tones and then looking for specific features within this region, e.g. changes in intensity to locate the eyes.

It should be noted that image descriptors extract information such as colour, shape, texture, segments; they do not identify the domain semantics of the subject matter of an image. For example, image analysis can detect a disc of a white or yellow colour within an image, but cannot determine whether this is the sun, the moon, a tennis ball, or the top view of a round cheese.

Bridging the semantic gap between the image analysis domain and the user's application domain is one of the major problems of image classification and content-based search and retrieval. Techniques do exist to help bridge this gap, but these are relatively immature compared with the larger body of work on content descriptors. In general, supervised learning techniques (e.g. training of neural networks) are used with an example set of images with known application-domain semantics to build classifiers that can label images based on the value of one or more content descriptors.

Instead of developing ever more sophisticated content descriptors, one approach to the problem of 'bridging the semantic gap' is to propagate existing human annotations, e.g. semantic annotations extracted from existing textual descriptions, across a collection of content items. We rely on people to describe the semantics of a subset of images or video, either by providing new annotations, or by using existing metadata. These human-authored annotations are then propagated to similar items in a database. Propagation is

done via content-based analysis to identify similar items. In this way, content analysis is not used to attach semantics to content *per se*, but instead to propagate high-quality, manual annotations from items with known semantics to items that need further semantic annotation.

The semantic gap in the field of image and video content-based analysis is clearly a research challenge that needs to be addressed if these techniques are to be successfully applied in DIID. There is a considerable gap to bridge, for example, between the low-level information that can be extracted from colour-based segmentation of satellite images and the application-domain semantics of failing crops or animal migration, which might be the purpose for analysing the images in a DIID scenario.

8. Distributed data fusion and mining

As distributed computing and grid technologies have developed, data fusion and mining is starting to be applied in far more heterogeneous environments. As this is exactly the scenario in which DIID is likely to take place, developments in this area are particularly relevant.

8.1 Web mining

The web is a massive source of information. To attempt to mine this huge resource is an obvious target, and this is a current field of research. There are three different types of web mining (Liu 2004):

- web content mining, where the actual content of web pages is analysed
- web usage mining, where common patterns of the web's usage are found from access logs
- web structure mining, where the focus is on deriving patterns from the structure of the hyperlinks in the web pages.

8.2 Web content mining

Web content mining offers considerable opportunities and presents considerable challenges due to its characteristics. These are discussed below, paraphrased and adapted from Liu (2004).

- There is a huge amount of information on the web. This is probably the main driver for web content mining, coupled with the fact that it is easily accessible.
- On the web, there is information about almost anything. Again, this is a primary driver for web content mining.
- There exist many different data types, often on the same page (for example, plain text, multimedia, tables, etc.).
- A great deal of information on the web is redundant, i.e. it is repeated in many pages.

- The web is very noisy. A page may contain irrelevant information (for example, adverts, or links to unrelated subjects). Sorting the useful information from the useless is a considerable challenge (especially as the concept of ‘useful’ depends on one’s point of view).
- The web is dynamic; it is always changing. What is present and relevant today may be gone or out of date tomorrow.
- Above all, the web is a community. It represents society more than it represents information, data and computers. Many lessons regarding society can be learned from the web. However, there are downsides due to its societal nature. Anyone can author and host a page containing erroneous information, and if this is relied on in mining, there will be problems. A useful counter measure to this problem is the redundancy feature of the web highlighted above. If a fact can be cross-checked, a consensus can be formed. Wikipedia (see Wikipedia project) is an interesting reflection of this and promotes self-regulation. Anyone can alter an article, but everyone is free to dispute what someone else has written, thus promoting a consensus on a particular topic.

8.3 Web usage mining

In general, web mining involves the automatic discovery of user access patterns from one or more web servers. An overview and taxonomy of the applications of web usage mining is presented by Srivastava (2000). Briefly, they are: personalisation (target marketing for e-commerce sites); system improvement (determine usage patterns with the aim of improving performance and other service quality attributes); site modification (the improvement of sites’ design based on feedback from user patterns); and business intelligence (how the customers are using a site).

Note that, when dealing with usage data that may be tracked back to individuals, privacy issues are important, and data protection guidelines should be adhered to.

8.4 Web structure mining

Much less work has been done regarding web structure mining, but a large part of what has been done involves considering hyperlinks in terms of graphs. Desikan (2004) considers this with the added dimension of time, and how the link structure can change over time.

8.5 Grid-based data fusion and mining

The term ‘grid computing’ originated in the early 1990s as a metaphor for making computer power as easy to access as an electricity distribution grid. The original view of grid computing in e-Science was to use the resources of many separate computers connected by a network (usually the internet) to solve large-scale computation problems. Many people still identify grid computing solely with 'number-crunching', but for some years the term has embraced all kinds of networked resources, notably data. Today, there are many definitions of grid computing (see e.g. Grid Computing Wikipedia), and

unfortunately the term has been adopted for marketing purposes by vendors, e.g. to describe software for cluster computing. But the sharing of resources across administrative domains sets grid computing apart from traditional computer clusters or traditional distributed computing.

An oft-cited definition (Foster et al., 2001) is ‘flexible, secure, coordinated resource-sharing among dynamic collections of individuals, institutions, and resources’. We say that: ‘Grid computing involves sharing heterogeneous resources which are under different ownership or control, over a network using open standards.’ In short, it involves virtualising computing resources.

Grid computing is clearly of enormous importance to DIID. It is the subject of much current research and development (R&D), but, in itself *per se*, it is outside the scope of this review, and we confine our attention to some especially relevant work.

There is as yet little published research on grid-based data fusion, though grid computing clearly has great potential as an enabling technology for decentralised data fusion: see e.g. the ESA SpaceGRID project (Marchetti et al. 2002), the Foresight DARP ARGUS II project (ARGUS II), the ESRC INWA project (INWA) and the NASA GENESIS project (Wilson 2005).

Gathering all relevant data in one data warehouse is unlikely to be fruitful for the vast amount of disparate time-varying data needed for DIID. Middleware to facilitate access and integration of data from separate sources is therefore a key requirement which is being addressed by e.g. the Open Grid Services Architecture Data Access and Integration project (see OGSA-DAI project). This is supported by several current grid middleware projects (see Globus, GRIA and OMII).

Grid-based data mining, and the workflow necessary to orchestrate it, is at the leading edge of current work.. (Au et al. 2004) describe work done in the Discovery Net e-Science project (see Discovery Net project), and highlights the following major benefits:

- Data from disparate sources may be mined and patterns found in the data as a whole rather than in its source components.

- Workflows can dynamically integrate many different (non-co-located) data and analysis services.

- The large computational resources available to a grid user permit different (more computationally intensive) analyses.

Discovery Net has already been applied in areas relevant to DIID, including SARS (Curcin et al. 2004), crop monitoring (Hassard et al. 2004), urban air pollution (Ghanem et al. 2004) and processing of satellite imagery (Liu et al. 2003). In each case, an e-Science data analysis process has been developed for a scientific knowledge discovery process conducted in an open environment and making use of distributed data and resources.

In the study of urban air pollution (Au et al. 2004), a sensor array generates huge amounts of data, and the large number of computational resources available on a grid is ideal for processing this data and also permits the correlation with other data sources of different types (for example, weather on the day of collection, traffic concentration, and a wide variety of data about the population's health).

In the analysis of the evolution of the SARS epidemic (Curcin et al. 2004), multiple sources of data were used and integrated into a workflow that correlated aspects of the data from the different sources. This workflow combines elements of data mining and grid computing:

- The data is gathered from disparate sources, filtered and cleaned.
- Visualisation tools are used when needed to explore the data.
- Computation is outsourced where required to remote compute clusters on a grid, thus providing considerable reductions in execution time.

The DataMiningGrid project (see DataMiningGrid project) recognises that, currently, there is no coherent framework for developing and deploying data mining applications on a grid. It is addressing this gap by developing generic and sector-independent data mining tools and services for grid computing. To demonstrate the technology developed, the project is implementing a range of demonstrator applications in e-Science and e-Business.

DataMiningGrid has identified a set of requirements (Stankovski and Trnkoczy 2004) that need to be addressed. Requirements areas include: identifying (locating) resources by using metadata; accessing and selecting subsets of data; data transfer; data (pre-) processing; and data mining tasks. These can immediately be seen to align with the data mining processes and methodologies outlined earlier. However, DataMiningGrid also identifies further requirements that are concerned with the wider issues of finding data sources, moving data, and the need for security. These include: text mining and ontology learning; workflow editing and submission; data privacy, security and governance; integration of domain knowledge, grid infrastructure and middleware functionality; usability, response times and user-friendliness.

DataMiningGrid has also identified a set of challenges and issues for grid-based data mining (Stankovski and Trnkoczy 2004), including the problems of distributed data, distributed operations, and data privacy, security, and governance.

Grid-based data mining operations must be orchestrated in a workflow, with data passing between them in such a way that the operations can use the data. This means that there must be published interfaces, so that the data may be transformed in such a way that the operations may run.

Given that data is under the control of its owner, another party including it in a workflow is relying on the data owner to keep the data available and accurate. If the workflow is saved, and the data is required again, how can users

executing the workflow be sure that they will be retrieving the same data as the first time the workflow was enacted?

We note that there is, as yet, no established best practice for building secure systems or detecting security breaches in distributed, multi-owner systems like grids. This is a key research challenge that has only just begun to be addressed (see e.g. Surridge 2002; Herveg et al. 2004).

Privacy is an issue of sufficient relevance and magnitude to merit discussion at length.

9. Privacy

There is a considerable ongoing debate about breaching privacy with data fusion and mining. Central to this is what is known as the inference problem, described by Farkas (2002). This is the situation whereby a user may deduce sensitive information from raw data that is essentially public. For example, a set of employees' names and a set of numbers representing salaries may be published separately. However, if it were possible to infer which employee earned which salary, the result would be private. Such problems are highly complex and involve technology, sociology and law (Thuraisingham 2002). The debate has intensified post-9/11 as counterterrorism and civil liberties advocates argue from both sides. Whatever conclusions are reached from time to time will crucially influence the progress and adoption of data fusion and mining in circumstances where privacy is an issue.

An architecture has been proposed for privacy considerations to be integral to database design in the so-called 'Hippocratic Database' (Agrawal 2002). This takes its name from the Hippocratic Oath, whereby medical doctors swear they will keep confidential anything discovered as a result of their professional relationship with a patient, thus protecting the patient's privacy regarding their health. The Hippocratic Database takes its basic principles from the OECD data protection guidelines (OECD 1980). Countries around the world have used these as the basis for data protection laws. The principles, adapted for the Hippocratic Database, are listed below and are quoted verbatim from Agrawal (2002).

1. Purpose Specification. For personal information stored in the database, the purposes for which the information has been collected shall be associated with that information.

2. Consent. The purposes associated with personal information shall have consent of the donor of the personal information.

3. Limited Collection. The personal information collected shall be limited to the minimum necessary for accomplishing the specified purposes.

4. Limited Use. The database shall run only those queries that are consistent with the purposes for which the information has been collected.

5. Limited Disclosure. The personal information stored in the database shall not be communicated outside the database for purposes other than those for which there is consent from the donor of the information.

6. Limited Retention. Personal information shall be retained only as long as necessary for the fulfilment of the purposes for which it has been collected.

7. Accuracy. Personal information stored in the database shall be accurate and up-to-date.

8. Safety. Personal information shall be protected by security safeguards against theft and other misappropriations.

9. Openness. A donor shall be able to access all information about the donor stored in the database.

10. Compliance. A donor shall be able to verify compliance with the above principles. Similarly, the database shall be able to address a challenge concerning compliance.

Central to the architecture is the concept of purpose – the purpose for which the data is accessed. In compliance with the OECD data protection guidelines, this must be stated and available to the person the data represents. An example of a purpose could be the purchase of a book from an online bookseller. To complete the transaction, the bookseller needs to know certain information (namely, the customer's name, address, credit card number and the book they want). The OECD guidelines also state that the data must not be kept for longer than the stated purpose. If the only purpose is the purchase of a book, the bookseller is obliged to delete the data as soon as the transaction has been completed. (In reality, this may be impractical for customers, since they have to re-enter their address every time they want to buy a book. So, in this instance, there may be another purpose, namely registration. This has a different lifetime than the purchase, namely the lifetime of a customer's relationship with the bookseller).

Work has continued using the Hippocratic Database concept, and has resulted in IBM's Hippocratic Database Technology (HDB), described by Agrawal (2005). This is a commercial product based on the Hippocratic Database Architecture.

In terms of security in grid computing, we note that there is, as yet, no established best practice for building secure systems or for detecting security breaches in distributed, multi-owner systems like grids. This is a key research challenge which has only just begun to be addressed (see e.g. Surridge 2002; Herveg et al. 2004).

Privacy is most acute in distributed data fusion and mining due to the very distribution of data. Once data has been released to a third party it is no longer completely under the control of its owner, and thus the potential for violations is magnified. The inference problem is also magnified. The data

owner may not necessarily know with which other data theirs will be cross-referenced and thus has no way of knowing which patterns will emerge. The only simple solution is to ensure that any data used for distributed data fusion and mining is not personal data. Otherwise, data fusion and mining activities will have to be highly specific and have informed consent from the individuals concerned.

10. Provenance

The provenance of data is of paramount importance for DIID, as it is a measure of the credibility and reliability of the data. In essence, it is an audit trail of where the data originated and any transformation it has undergone.

Mark-up languages, such as XML and its derivatives, and their use for metadata, are key to self-describing datasets. Provenance information may mark up the data it applies to, and thus will be carried with the data, but validating and using provenance metadata is a significant challenge.

For example, the internet is a largely unregulated community. Anyone can author and host a web page that may contain erroneous information. Establishing the provenance of the page enables one to rely on the information contained within it. However, this is seldom straightforward.

Provenance is also extremely important to the e-Science community and is the focus of a great deal of current work. The research agenda is to provide mechanisms that allow information to be proven and trusted. This means that the history of the information, including the processes that created and modified it, are documented in a way that can be inspected, validated and reasoned about by authorised users who need to ensure that information controls have not been altered, abused or tampered with (see e.g. the PROVENANCE project).

In the myGrid project (see myGrid project), data is generated by experiments and its provenance is measured in two ways: its derivation path (i.e. how the data was created); and annotation (a means of adding notes relevant to the data, for example, when it was updated, what was changed, who changed it, etc). (Greenwood et al. (2003;) Zhao et al. 2004). As the data progresses through a workflow, annotations can be added describing the transformations, or, if new data is generated, derivation paths must be added to the new data.

11. Realising the vision of DIID

Effective DIID will only become a reality when the underlying challenges outlined below are addressed in a cohesive way by the research community. Many of these challenges are shared by other domains where intensive research is already underway, for example, in e-Science, autonomic systems and pervasive computing.

The advancement of DIID needs to build on, combine and stimulate further work in these research communities. In this section, we review the major challenges of DIID, identify the research communities that are addressing some of these challenges, and enumerate some of the key enabling technologies that are emerging from these communities.

11.1 The challenges of DIID

- Data generation. The results of fusion and mining can be expected to feed back into sensors and collection. For example, a sensor network configuration and the data it collects might be altered in response to the data needed for a particular mining scenario or the development and calibration of an epidemiological model. Sensors will need to be adjusted or added to a DIID scenario as it unfolds, which creates new challenges of how to do this in real time and in a way that is integrated into the process of analysis or hypothesis testing.

- Discovery, integration and semantic mapping. Inputs to data fusion and mining for DIID are wide-ranging and include biosensing, remote sensing, global positioning systems (GPS) and tracking, web content, and hospital records, to name but a few. Each new DIID scenario will raise questions of what data to collect, how to combine more than one source of data, and how to analyse that data. Integrative models will emerge to combine data at various scales, from earth observation down to cellular function and genetics. Such a wide range of heterogeneous and distributed data sources creates real challenges for locating and combining relevant data sources. Achieving interoperability between data sources through well-defined semantics is an essential yet challenging part of this process.

- Trust and provenance. Support for provenance is a fundamental requirement to enable a scientific approach to be taken to DIID. Provenance covers: data sources (what the data is, where it came from, and its quality and accuracy); analysis processes (who did what using the data, and what tools and methods they adopted); and derived results (what conclusions were drawn, what interpretations were made, and confidence levels). In addition to a DIID infrastructure that records provenance, there is a need for policies and processes that ensure provenance recording and provenance assessment in DIID.

- Privacy and security. Distributed and unstructured data in multi-owner and distributed systems creates major challenges for privacy and security. Flexible but highly managed control over who can access what data, and how they can use it, is needed. Distributed, dynamic and semantic-based security is a major challenge.

- Distribution, optimisation and robustness. The data sources for fusion and mining are manifold, highly distributed, heterogeneous, with variable reliability, and do not have guaranteed availability. The volumes of data are potentially huge, and there will be many computationally intensive processing steps that need to be applied. Centralising the data and the computation is not a scalable or robust option. Distributed approaches will

be required that are resilient to failures or disruption to networks, data sources, analysis services (overload, deliberate attack, censorship of information, quarantine of sources etc.). This, in turn, requires processes and strategies distributing the steps in DIID and optimising the use of resources.

- Collaborative and team-based approaches. Collaborative approaches to DIID will be necessary to combine the expertise of various geographically distributed experts and analysts and to bring this to bear on a particular DIID scenario. This requires the assembly, operation and management of distributed collaborations supported by tools and infrastructure for collaborative working.

- Agile systems and methods. The data that needs to be captured for DIID and how it needs to be analysed will change over time, for example, as a disease spreads. Changes will include geographical areas, data volumes, and the type of data required. Interpretation of data sources will change as the provenance and reliability of data becomes clearer, or when further scientific research is done on the cause and origin of a disease, which changes the initial interpretation of the data. The dynamics of DIID (e.g. the transition from a monitoring phase to the outbreak of an epidemic) require agile use of data fusion and mining – very different from most current applications, which have relatively static geographies, sensor configurations, data types and data volumes.

- Methodologies and best practice. There is normally a delay between the development and establishment of new tools and infrastructure and the development of methodologies and processes for best use of those tools and infrastructure. Whilst methodologies and processes exist for both data fusion and data mining, a unified and integrated approach has yet to be developed to cover both. The need for appropriate best practice in distributed data mining is becoming evident, just as it did for enterprise data mining in the late 1990s.

11.2 Research communities and technologies

Many of the challenges faced in DIID are currently research topics in the e-Science, autonomic computing, agent, semantic grid, and pervasive networks and computing communities. Relevant research is being carried out in universities, defence and other government research agencies, and in medical and health protection, aerospace, automotive and robotics communities worldwide.

e-Science is an exemplar research community that is tackling many of the same challenges faced by DIID. e-Science refers to the large-scale science that is increasingly carried out through distributed global collaborations enabled by the internet. . Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large-scale computing resources and high-performance visualisation. e-Science addresses many of the issues relevant to data fusion and mining in DIID. For example, myGrid, Integrative Biology (Gavaghan et al. 2004) and

Discovery Net (DiscoveryNet project) are UK e-Science projects undertaking computer science research into how data-intensive scientific analysis activities can be supported by grid and knowledge technologies. In each case, an e-Science data analysis process has been developed for scientific knowledge discovery conducted in an open environment making use of distributed data and resources.

The key challenges in data analysis for DIID will be addressed by combining a number of emergent software technologies. Data fusion and mining systems have hitherto been static client/server-orientated architectures. Distributed, dynamic, autonomous, intelligent data fusion and mining will be possible through the integration of current research activities. This is likely to be in a service-oriented architecture (SOA), in which resources are made available as independent services that are accessed in a standardised way. Web services are the leading current technology, but one can implement SOA using any service-based technology.

11.2.1 Web services

Web services provide access to modern and legacy systems through XML protocols, such as SOAP over the web. Web services are now at a stage where data can be transmitted between heterogeneous systems while maintaining data integrity and confidentiality (see WS-Policy, WS-Security Policy). Workflow standards such as BPEL4WS (Business Process Execution Language for Web Services) describe how to tie together different web services, for example, in automatic business processing systems. UDDI (Universal Description, Discovery and Integration protocol) adds the ability to discover services through centralised repositories.

11.2.2 Workflow

There is an active research community concerned with how to support scientific workflows involving data processing, in particular using data from multiple, distributed and heterogeneous sources. Examples include: the Taverna workflow workbench (Oinn et al. 2004), which includes support for semantic service discovery and provenance; the Kepler workflow system (Altintas et al. 2004), developed for scientists with a range of interests and built on Ptolemy II (Ptolemy project); the Triana data analysis problem-solving environment (Shields and Taylor 2004); Geodise (Jiao et al 2004), which has a focus on engineering design search and optimisation; and Pegasus (Gil et al. 2004) which abstracts the user from the detail of workflow design. This collective body of research forms the building blocks for developing processes, systems and best practice for the agile analysis systems needed for data mining and data fusion in DIID.

11.2.3 Semantic web

‘The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation’ (Berners-Lee 2001). The aim of the semantic web is to enhance the web with metadata that is machine-readable. This metadata can then be searched for and aggregated with existing knowledge.

By far the most promising aspect of the semantic web is the use of ontologies for describing domains of knowledge. Ontologies are used as the basis for semantic inferencing, where new, implicit knowledge can be generated from existing, explicit knowledge using rule-based systems. These techniques are beginning to be researched in areas relevant to DIID (see e.g. Crubézy et al. 2005).

Currently, two competing standards exist for representing data on the semantic web, namely the Resource Description Framework (RDF) and Topic Maps. The Web Ontology Language (OWL) is the only description logic standard.

The semantic web has the potential to help solve the interoperability problems that exist in DIID. Semantic mark-up provides a semi-structured and standardised format for data interchange; ontologies provide formal semantics for concepts and relationships in datasets as well as semantic interoperability; and semantic inferencing yields new knowledge.

11.2.4 Semantic web services

Semantic web services extend basic web services. Such services are accompanied by a semantic description, written in an OWL-based Web Service Ontology (OWL-S), which helps find services of a certain type and can also help dynamically create new services. Whilst many semantic web services exist, these tend to be rather simple, reflecting the learning curve required to utilise semantic frameworks effectively. Currently, the Semantic Web Services Framework (SWSF) appears to be the future path for semantic web service development.

11.2.5 Grid services

Unlike web services, which tend to exist in isolated environments within organisations, grid services are intended to be deployed within virtual organisations. Grid services can be viewed as web services that hold state that can be referenced by other grid Services within the virtual organisation. The Web Service Resource Framework (WSRF) is the proposed definition of an open framework for modelling and accessing stateful resources using web services.

11.2.6 Semantic grid services

The semantic grid is the anticipated result of research by both grid and semantic web communities. Described by De Roure et al. (2003, 2005), the semantic grid is envisioned as an extension of the current grid, in which information and services are given well-defined meaning, better enabling computers and people to work in co-operation.

Advances in semantic web technology will allow grid services to include semantic inferencing. This will be especially useful for data fusion and mining, as computational overhead could be reduced through intelligent resource management and task decomposition.

11.2.7 Peer-to-peer networks

Peer-to-peer (P2P) networks go beyond the classic client/server paradigm, treating a computer on a network as both a consumer and producer of services or resources. Since each peer on the network can serve and consume, the emergent behaviour of the system normally leads to higher availability of services and resources – a highly desirable feature for DIID.

Availability is an important issue, especially when we consider the increased occurrences of Denial-of-Service (DoS) attacks against corporate and public web-hosted services. Effective DIID demands access to data and to services that perform the required data fusion or mining when needed. P2P systems like BitTorrent (Cohen 2003) alleviate the effects of DoS for file sharing, which could also be applicable for service-based architectures for DIID.

P2P networks are also highly dynamic, allowing peers to join and leave the network quickly and painlessly. File-sharing networks are among the most successful P2P systems. Popular files maintain high availability due to the number of peers holding portions or full copies of that file.

It should be perfectly feasible for future P2P networks to provide many of the features that exist in SOAs of all flavours (web, grid, semantic web, semantic grid, etc.).

11.2.8 Agent-based systems

Agent-based computing is also a service-oriented model, in which software agents can be the producers, consumers and brokers of services. An agent is defined as ‘an encapsulated computer system that is situated in some environment, and that is capable of flexible, autonomous action in that environment in order to meet its design objectives’ (Wooldridge 1997).

De Roure et al. (2005) argue that the autonomous behaviour and other qualities of multi-agent systems are vital to realise virtual organisations (VO) in grid computing. Foster el al. (2004) note a convergence of interests in the grid and agent communities. Grid computing requires the autonomous nature of agents, while agents require a robust infrastructure.

Agents or the agent philosophy would be applicable for deployment in a DIID environment, especially considering the dynamic and autonomous decision making that components in such a system would be required to make.

11.2.9 Autonomic Computing

Autonomic computing is a systemic view of computing modelled on self-regulating biological systems (IBM 2001). Its aim is to overcome the rapidly growing complexity of modern computer systems and to help enable further growth. This growing complexity is seen as the limiting factor to future development.

At the core of autonomic computing is the concept of closed-control loops, a well-known term taken from process control theory. Control loops are used to manage selected resources that automatically self-regulate to maintain

parameters within a predefined range. Furthermore, autonomic computing systems are characterised as being self-configuring. They are able to adapt automatically in dynamic environments. They are self-healing: able to discover, diagnose and react to disruptions. They are self-optimising: able to monitor and tune resources automatically. And they are self-protecting: able to anticipate, detect, identify and protect themselves from attack, independent of location (Ganek and Corbi 2003). Work in this field has led IBM to release an autonomic computing toolkit and to integrate a number of autonomic features into their enterprise product range.

11.2.10 Blogs, Wikis and collaborative personal communication

There is an implicit expectation that the discovery of the spread of infectious diseases will be through analysing clinical and biosensor data. However, the rapid spread of disruptive technologies such as Blogs and Wikis and new collaborative means for personal communication will also be important sources – perhaps the most important. Using these new technologies, people share a wide range of personal information including health matters and medical concerns. Such publicly -accessible information is becoming universally available. The dynamic social networks by which it spreads can be analysed and may reveal the emergence of new diseases much more rapidly than any other method. Targeted analysis of clinical and biosensor data may then be used to discover if there is any substance behind the concerns being expressed in the personal communications. This is clearly a worthwhile area for further research.

12. The future

The detection and identification of infectious disease through distributed data fusion and data mining creates many challenges. Intelligent and agile semantic data integration is needed in order to deal with the complexity and heterogeneity of DIID data. Distributed processing, analysis and knowledge management techniques are the only way to solve the problem of such large and dynamic datasets. New methodologies, processes and best practice have to be developed to ensure confidence in the results. Last, but certainly not least, privacy, trust, provenance and security are bedrocks for addressing the wider issues of ethics, sovereignty, data protection, freedom of information, and other rights and obligations when detecting, collecting and processing the data involved in DIID.

We believe that the underpinning processing power, data storage and network bandwidth will continue to increase for many years to come. When current technologies reach physical limits, new technologies such as quantum computing and holographic storage will come into play. These will not only enable ever-increasing volumes of data to be collected and stored, but will also underpin the use of new software technologies to analyse it.

While available processing power will continue to increase, the explosion in data volumes will nevertheless require advances in computational efficiency. Many data fusion and mining algorithms are computationally expensive. If

massive amounts of data are being analysed, large amounts of computing will be required on demand. Grid computing is an emergent solution, but much remains to be done before we can readily share heterogeneous resources that are under different ownership or control.

Intelligent semantic web services and grid services will find patterns in massive datasets from remote sensors, news feeds and publicly accessible personal communications. Massive distributed systems in dynamic, virtual organisations will deliver the required analytical capability. Grid systems will allow for automated load balancing of tasks, and workflow engines will help choreograph task decomposition. Agents will be important in dynamic and autonomous environments, especially if coupled with sensor networks. Agents will be used in conjunction with grid and semantic web technologies to produce intelligent and robust services that can reason about data being fused and mined.

Distributed data fusion and data mining using pervasive and grid techniques will be truly enabled only when we have addressed one of the key challenges in distributed computing: homogeneous access to, and use of, heterogeneous data. Much remains to be achieved in enabling the analysis of unstructured data, particularly multimedia data. Bridging semantic gaps, semantic mapping and semantic interoperability are all key research areas.

Privacy and security will always be major factors in DIID. Increasing volumes of distributed, unstructured and multi-owner data only serve to increase the scale of the problem. What matters is who can access what and what they can do with it in some defined context, not their logon credentials or what their role is, or where they are or where the data is. Dynamic semantic security and (business) process-based access control are among the research issues. Web services and semantic web technologies present an opportunity to introduce security as a fundamental aspect of shared information. Web services have well-defined interfaces with message format and transport specifications, including security headers for, among others, authentication and encryption. Semantic web technologies provide a means to annotate documents with machine-readable descriptions of their content and meaning, including their relationships to other documents and resources. If these mechanisms were applied to service security policies, they would become semantically aware (based on the meaning or the contents of documents) and dynamic (sensitive to the process or context in which access is requested).

No single technology or research focus taken in isolation will deliver clear benefits. For example, the grid may bring efficient and scalable data fusion and mining, but this will fall short of the truly autonomous computing that DIID requires. However, once other technologies are added to the mix, there is the long-term potential for powerful, intelligent systems that can analyse vast datasets in remote locations, categorise information according to known, agreed taxonomies, discover patterns using semantic inferencing, and reason over the results to provide augmented cognition for decision support – all achieved securely and reliably. It is a combination of technologies and the interdisciplinary collaboration of a wide range of research communities that is essential to the success of DIID.

13. Acknowledgements

Special thanks go to Graham Bent of IBM and Veronica Rapley of DSTL, who read and commented on earlier drafts of this review, and made valuable suggestions and contributions.

14. References

AceMedia project: http://www.acemedia.org/aceMedia

ACL, The Association of Computational Linguistics: http://www.aclweb.org

Agrawal, R., Kiernan, J., Srikant, R. and Xu, Y. (2002) *Hippocratic databases*. Proceedings of the 28th International Conference on Very Large Databases,. Hong Kong, China, August.

Agrawal, R., Asonov, D., Bayardo, R., Grandison, T, Johnson, C. and Kiernan, J. (2005) *Managing disclosure of private data with Hippocratic databases*. IBM White Paper. http://www.almaden.ibm.com/software/quest/Publications/papers/nc_hdb_white_paper_health.pdf

Aleksander, L. and Morton, H. (1990) *An introduction to neural computing*. New York: Van Nostrand Reinhold.

Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludscher, B. and Mock, S. (2004) Kepler: *Towards a grid-enabled system for scientific workflows*. In Workflow in Grid Systems Workshop in GGF10, Berlin, March.

ARGUS II project: http://www.robots.ox.ac.uk/~argus/

Au, A.K.T.P., Curcin, V., Ghanem, M., Giannadakis, N., Guo, Y., Jafri, M.A., Osmond, M., Oleynikov, A., Rowe, A.S., Syed, J., Wendel, P. and Zhang, Y. (2004). *Why grid-based data mining matters? Fighting natural disasters on the grid: from SARS to land slides.* In Proceedings of the 3rd UK e-science All-Hands Conference, Nottingham, UK., 121–-126. ISBN 1-904425-21-6.

Bailer, W., Höller, F., Messina, A., Airola, D., Schallauer, P. and Hausenblas, M. (2005) *State of the art of content analysis tools for video, audio and speech.* PS_WP15_JRS_D15.3_SOA_Content_Analysis_v1.0, 10/3/2005. Available from: http://www.prestospace.org

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) *The semantic web*,. Scientific American, May.

BioText: http://biotext.berkeley.edu

Bowman, C. (1994) *The data fusion tree paradigm and its dual*. In Proceedings of the 7th National Symposium on Sensor Fusion.

Bollerslev, T. (1986) *Generalized autoregressive conditional heteroskedasticity*., Journal of Econometrics 31, 307–327.

Brockwell, P.J. and Davis, R.A. (1996) *Introduction to Time Series and Forecasting.* New York: Springer. Sections 3.3 and 8.3.

Brönnimann, H., Chen, B., Dash, M., Haas, P. and Scheuermann, P. (2003), *Efficient data reduction with EASE*, Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington D.C., 59-68.

Chen, B, Haas, P. and Scheuermann, P. (2002) *FAST: a new sampling-based algorithm for discovering association rules*. 18th International Conference on Data Engineering. 263.

Chung, S. and McLeod, D. (2003) *Dynamic topic mining from news stream data*., In: R. Meersman, R. et al. (eds), Eds., On the move to meaningful internet systems 2003: CoopIS, DOA, and ODBASE. Springer LNCS 2888, 653–670.

CIVR: http://www.comp.nus.edu.sg/~civr/

CliniMiner: http://www.alphaworks.ibm.com/tech/cliniminer

Cohen, B. (2003) Incentives build robustness in BitTorrent:, http://www.bittorrent.com/bittorrentecon.pdf, 2003

CRISP-DM project: http://www.crisp-dm.org

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines (and other kernel-based learning methods)*. Cambridge: Cambridge University Press.

Crubézy, M., O'Connor, M., Pincus, Z., Musen, M.A. and Buckeridge, D.L. (2005) *Ontology-centered syndromic surveillance for bioterrorism*,. IEEE Intelligent Systems., 20, 26–-35.

Curcin, V., Ghanem, M. and Guo, Y. (2004) *SARS Analysis on the Grid*. In Proceedings of the 3rd UK e-Science All-Hands Conference, Nottingham, UK, 114–120. ISBN 1-904425-21-6.

Daley, D. and Vere-Jones, D. (1988) *An introduction to the theory of point processes*. New York: Springer-Verlag.

DataMiningGrid project: http://www.datamininggrid.org/

Data Warehousing Wikipedia: http://en.wikipedia.org/wiki/Data_warehousing

DB2II: http://www-306.ibm.com/software/data/integration/db2ii/

Denning, D.E. (1986) *An Intrusion-Detection Model*. IEEE Symposium on Security and Privacy., 118–-133.

De Roure, D., Jennings, N.R. and Shadbolt, N. (2003) *The semantic grid: a future e-Science infrastructure*. I, in F. Berman, F., G. Fox, G. and A.J.G. HeyEds (eds). Grid computing: - making the global infrastructure a reality., John Wiley and Sons Ltd. 437–470.

De Roure, D. Jennings, N.R. and Shadbolt, N.R. (2005) *The semantic grid: past, present, and future.*, Proceedings of the IEEE, 93, 669–681.

Desikan, P. and Srivastava, J. (2004*) Mining Temporally Evolving Graphs*. In B. Mobasher, B. Liu, B. Masand and O. Nasraoui (eds)Eds,. Proceedings of the 6th WEBKDD workshop: Webmining and Web Usage Analysis in conjunction with the 10th ACM SIGKDD conference, Seattle, Washington, USA., 22 August.

DiscoveryNet project: http://www.discovery-on-the.net/

Farkas, C. and Jajodia, S. (2002) *The inference problem: a survey*. ACM SIGKDD Explorations Newsletter 4, 2. 6–11.

Foster, I., Kesselman, C. and Tuecke, S. (2001) *The anatomy of the grid: enabling scalable virtual organizations.* International Journal of. Supercomputer Applications 15(3).

Foster, I., Jennings, N.R. and Kesselman, C., (2004) *Brain meets brawn: why grid and agents need each other*..

GALE project: http://www.darpa.mil/ipto/solicitations/open/05-28_PIP.htm

Ganek, A.G., Corbi, T.I. (2003) *The dawning of the autonomic computing era*. IBM Systems Journal 42, 1.

Gavaghan, D.J., Lloyd, S., Boyd, D.R.S., Jeffreys, P.W. and Simpson, A.C. (2004) *Integrative biology: exploiting e-Science to combat fatal diseases*. Proceedings of the 3rd UK e-Science All-Hands Conference, Nottingham, UK. http://www.integrativebiology.ox.ac.uk/publications/IB%20Overview%20Paper%20AHM2004.pdf

Ghanem, M., Guo, Y., Hassard, J., Osmond, M. and Richards, M. (2004) *Grid-based Data Analysis of Air Pollution Data*. In 4th International Workshop on Environmental Applications of Machine Learning.

Gil, Y., Deelman, E., Blythe, J., Kessleman, C. and Tangmunarunkit, H. (2004). *Artificial Intelligence and Grids: Workflow Planning and Beyond*. IEEE Intelligent Systems (special issue on e-Science), 19(1), 26–-33.

Globus: http://www.globus.org/toolkit/

Goharian, N. , Grossman, D., Raju, N. (2004) *Extending the Undergraduate Computer Science Curriculum to Include Data Mining*. Proceedings of the International Conference on Information Technology: Coding and Computing.

Goodman, I., Mahler, R. and Nguyen, H. (1997) *Mathematics of data fusion.* Dordrecht: Kluwer Academic Publishers.

Goodman, I. and Nguyen, H. (1985) *Uncertainty models for knowledge based systems*. Amsterdam: North-Holland.

Google Desktop: http://desktop.google.com/Unstructured data mining

Greenwood, M., Goble, C., Stevens, R., Zhao, J., Addis, M.J., Marvin, D.J., Moreau, L. and Oinn, T. (2003) *Provenance of e-Science experiments - experience from bioinformatics*. In Proceedings of tithe UK OST e-Science, second All-Hands Meeting, Nottingham, UK. 223–-226.

GRIA: http://www.gria.org

Grid Computing Wikipedia: http://en.wikipedia.org/wiki/Grid_computing

Hall, D. and Llinas, J. (2001) *Handbook of multisensor data fusion*. Boca Raton FL: CRC Press.

Hall, D. and McMullen, A. (2004) *Mathematical techniques in multisensor data fusion*. Boston: Artech House.

Hamilton, J.D. (1994) *Time series analysis*,. Princeton NJ: Princeton University Press, Chapter 21, Section 2.

Hassard, S., Osmond, M., Pereira, F., Howard, M., Klier, S., Martin, R. and Hassard, J. (2004) *Distributed biosensors in genetically modified crop trial monitoring*. In Proceedings of the 3rd UK e-Science All-Hands Conference, Nottingham, UK. 156-–161. ISBN 1-904425-21-6.

Hearst, M. (1999) *Untangling text data mining*. In Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics, University of Maryland, 20–-26 June.

Herveg, J., Crazzolara, F., Middleton, S.E., Marvin, D.J. and Poullet, Y. (2004) GEMSS: *Privacy and security for a medical grid*. In Proceedings of HealthGRID 2004, Clermont-Ferrand, France.

IBM (2001) *Autonomic computing*: IBM's perspective on the state of information technology:. http://www-1.ibm.com/industries/government/doc/content/resource/thought/278606109.html

Intel (2005) Intel Research Sensor Network Operation. Technical Report, Intel Corporation.

INWA project: http://www.epcc.ed.ac.uk/~inwa/

Jiao, Z., Wason, J.L., Song, W., Xu, F., Eres, H., Keane, A.J. and Cox, S.J. (2004). *Databases, workflows and the grid in a service- oriented environment*. In Euro-Par 2004, Parallel Processing, 972–979, Pisa.

Kalman, R. (1960) *A new approach to linear filtering and prediction problems. transactions of the ASME*. Journal of Basic Engineering 82, 35–-45.

Karl, H. and Willig, A. (2005) *Protocols and architectures for wireless sensor networks*,. John Wiley and Sons LtdWiley.

kdnuggets (2004). *Poll on the use of data mining methodologies*: http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm

Klein, L. (2004) *Sensor and data fusion: a tool for information assessment and decision making*. SPIE Press Monograph PM138.

Lewis, P.H., Martinez, K., Abas, F.S., Ahmad Fauzi, M.F., Addis, M.J., Lahanier, C., Stevenson, J., Chan, S.C.Y., Mike J.B. and Paul, G. (2004) *An integrated content and metadata based retrieval system for art*. IEEE Transactions on Image Processing 13(3) 302–313.

Liu, J.G., Mason, P.J, Clerici, N., Chen, S. and Davis, A. (2003) *Landslide hazard assessment in the three gorges area of the Yangtze river using ASTER imagery.* International Geoscience and Remote Sensing Symposium, Learning from Earth's Shapes and Colors, IEEE IGARSS2003, Toulouse, France., July, 21–-25 July2003. pp. 1302–-1304.

Liu, B., and Chang, K.C.-C. (2004) *Editorial: Special issue on web content mining.*, ACM SIGKDD Explorations Newsletter 6, 2, 1–4.

Llinas, J., Bowman, C., Rogova, G., Steinberg, A., Waltz, E. and White F. (2004) *Revisiting the JDL data fusion model II*. In Proceedings of the 7th International Conference on Information Fusion, Stockholm, Sweden., 1218–-1230.

Mahler, R. (1994) *A unified foundation for data fusion*. In Proceedings of the 7th Joint Service Data Fusion Symposium, Laurel, Maryland., 1, 153-173.

Mahler, R. (2000) *An introduction to multisource-multitarget statistics and its applications*. Technical Monograph. Lockheed Martin, Eagan, MN.

Mahler, R. (2004) *The levels 2, 3, 4 fusion challenge: fundamental statistics*. In Proceedings of the 7th International Conference on Information Fusion, Stockholm, Sweden. , pp. 535-536.

Marchetti, P.G., Beco S. and Cantalupo, B. (2002) *SpaceGRID: how to foster an Earth Observation GRID, Euroweb 2002 Conference*. The Web and the GRID: from e-Science to e-Business, St Anne's College Oxford, UK, 17–18 December. http://www.w3c.rl.ac.uk/Euroweb/poster/108/SpaceGRID.htm

Masseglia, F., Poncelet, P. and Teisseire, M. (2003*) Incremental mining of sequential patterns in large databases,.* Data and& Knowledge Engineering 46, 97–121.

Moss, L.T. (2003) *Defining data mining.* http://www.businessintelligence.com/ex/asp/code.64/xe/article.htm, excerpted

from Larissa T. Moss and Shaku Atre, the book Business intelligence roadmap: the complete lifecycle for decision support applications (Addison-Wesley): http://www.businessintelligence.com/ex/asp/code.64/xe/article.htm,

myGrid project: http://www.mygrid.org.uk

NIST,: http://www.nist.gov

OECD guidelines on the protection of privacy and transborder flows of personal data. (1980). OECD Document C(80)58(Final), 1 October. Available online: http://www.oecd.org/document/18/0,2340,en_2649_34255_1815186_1_1_1_1,00.html

OGSA-DAI project: http://www.ogsadai.org.uk/

Oinn, T., Addis, M.J., Ferris, J., Marvin, D.J., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A. and Li, P. (2004) Taverna: *A tool for the composition and enactment of bioinformatics workflows*. Bioinformatics Journal 20(17), 3045–-3054.

OMII: http://www.omii.ac.uk

Oracle (2005). Oracle Database 10g Release 2 Enterprise Edition Data Sheet. Oracle Corporation.

Pankratz , A. (1983) *Forecasting with univariate Box-Jenkins models: concepts and cases*., John Wiley and Sons Limited.

PrestoSpace project: http://www.prestospace.org

Provenance project: http://www.gridprovenance.org/

Ptolemy project: http://ptolemy.eecs.berkley.edu

Punska, O. (1999) *Bayesian Approaches to multi-sensor data fusion*. M.Phil. Thesis., Signal Processing and Communications Laboratory, Department of Engineering, University of Cambridge.

Quinlan, J. (1993) *C4.5: programs for machine learning*,. San Francisco: Morgan Kaufmann Publishers.

RAID (2005) *Phoenix SATA storage solutions product guide*. RAID Incorporated.

Robson, B. (2003) *Clinical and pharmacogenomic data mining: 1. generalized theory of expected information and application to the development of tools.,* Journal of Proteome Research 2, 283–302.

SCHEMA project: http://www.iti.gr/SCHEMA/

Schroeder, F. and Petersen, W. (2000) *The ferry-box as a monitoring tool for marine waters*. In Proceedings of TechnoOcean 2000, Kobe, Japan.

SEMMA. Data mining process used in SAS Enterprise Miner: http://www.sas.com/technologies/analytics/datamining/miner/semma.html

Shields, M. and Taylor, I. (2004) *Programming scientific and distributed workflow with Triana Services. In Workflow in Grid Systems.* Workshop in GGF10, Berlin, March 2004.

Stankovski, V. and Trnkoczy, J. (2004) Datamining Grid Deliverable D11(1): *Common requirements analysis, specification and evaluation of DataMiningGrid interfaces and services:* http://www.datamininggrid.org/deliverables/DataMiningGrid-d-D11(1)-s-v14.pdf

Steinberg, A., Bowman, C. and White, F. (1999) *Revisions to the JDL data fusion mode*l. In Sensor Fusion: Architectures, Algorithms and Applications, Proceedings of the SPIE, Vol. 3719.

Steinberg, A. and Bowman, C. (2004) *Rethinking the JDL data fusion model*. In Proceedings of NSSDF, JHAPL.

Srivastava, J., Cooley, R., Deshpande, M. and Tan, P-N. (2000) *Web Usage Mining: Discovery and Applications of Usage Patterns* from Web Data, SIGKDD Explorations 1, 12–23.

Sun, J., Velastin, S.A., Lo, B., Vicencio-Silva, M.A. and Khoudour, L. (2004) *A distributed surveillance system to improve personal security in public transport.* Proceedings of European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, London, UK., November. 25–-26,.

Surridge, M. (2002) *A rough guide to grid security*. UK e-Science Technical Report UKeS-2002–-05:. http://www.nesc.ac.uk/technical_papers/RoughGuidetoGridSecurityV1_1a.pdf

Swanson, D. and Smalheiser, N. (1997) *An interactive system for finding complementary literatures: a stimulus to scientific discovery*. Artificial Intelligence 91, 183–203.

Thuraisingham, B. (2002) *Data mining, national security, privacy and civil liberties*, ACM SIGKDD. Explorations 4, 1–5.

Washio, T. and Motoda, H. (2003) *State of the art of graph-based data mining*, ACM SIGKDD. Explorations 5, 59–68.

White, F. (1988) *A model for data fusion*. In Proceedings of the 1st National Symposium on Sensor Fusion, Vol. 2.

Wikipedia project: http://www.wikipedia.org/

Wilson, B. (2005) *Grid Workflow Execution using GENESIS SciFlo*: http://isandtcolloq.gsfc.nasa.gov/fall2005/speaker/wilson.html

Witten, I.H. and Frank, E. (2005*) Data mining: practical machine learning tools and techniques.* Morgan Kaufmann Series in data management systems, 2nd edition. ISBN 0120884070.

Wooldridge, M. (1997) *Agent-based software engineering*. IEEE Proceedings. Software Engineering 144, 26–37.

WS-Policy: http://www-128.ibm.com/developerworks/library/specification/ws-polfram/

WS-Security Policy: http://www-128.ibm.com/developerworks/library/specification/ws-secpol/

Yergeau, F., Bray, T., Paoli, J., Sperberg-McQueen, C. and Maler, E. (2004) *Extensible Markup Language (XML) 1.0* (3rd Edition). W3C Recommendation.

Zadeh, L.A. (1965) *Fuzzy sets*. Information and Control 8, 338–353.

Zhao, F. and Guibas, L. (2004) *Wireless sensor networks : an information processing approach.*, CA: Morgan Kaufmann Publishers.

Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D. and Greenwood, M. (2004) *Using semantic web technologies for representing e-Science provenance*. In Proceedings of 3rd International Semantic Web Conference (ISWC2004), Hiroshima, Japan., November. 2004, Springer-Verlag LNCS 3298, 92–-106.

Zhao, R. and Grosky, W.I. (2002) *Narrowing the semantic gap – improved text-based web document retrieval using visual features*. IEEE Transactions on Multimedia 4. 189–200.

Zhu, Y. and Shasha, D. (2003) *Efficient elastic burst detection in data streams.* Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington D.C.., 336–-345

All the reports and papers produced within the Foresight project 'Infectious Diseases: preparing for the future,' may be downloaded from the Foresight website (www.foresight.gov.uk). Requests for hard copies may also be made through this website.

First published April 2006. Department of Trade and Industry. www.dti.gov.uk