



European KDD Research and Development

Maarten van Someren

Dunja Mladenic

04 December 2004



INTRODUCTION.....	3
GENERAL AND TECHNOLOGY-ORIENTED PROJECTS	4
INFORMATION GATHERING AND FILTERING, DIGITAL LIBRARIES, SEMANTIC WEB.....	6
LANGUAGE AND MULTIMEDIA LEARNING.....	8
KNOWLEDGE MANAGEMENT, E-COMMERCE.....	9
MARKETING AND PROFILING, BUSINESS MAPPING, COMPETITOR ANALYSIS, RISK MANAGEMENT	10
E-SCIENCE.....	11
E-LEARNING	11
COMPUTER SECURITY AND DATA PRIVACY	13
CULTURAL HERITAGE.....	13
HEALTH CARE	13
VARIOUS	15
DISCUSSION	17
APPENDIX 1: ALPHABETICAL LIST OF WEBSITES OF EUROPEAN PROJECTS.....	19
APPENDIX 2: ALPHABETICAL LIST OF NATIONAL PROJECT WEBSITES	23

Introduction

The purpose of this report is to give an overview of the main European research and development activities in Machine Learning, Data Mining and Knowledge Discovery in Databases (KDD). It was written for the KDNET, the Network of Excellence in Knowledge Discovery. Here we summarize the “what” of projects. More technical details of European research can be found in the KDNET Technological Roadmap document. We included projects under the headings Machine Learning, Data Mining, Knowledge Discovery in Databases and closely related areas, like neural networks, (inductive) adaptive systems.

This report is focused on larger projects in the period of Framework Programmes V and VI. These are mostly European projects but we included some larger national projects, if these projects had at least three different partners. The collection of projects was based in part on automated analysis of the EC database of projects using the system Project Intelligence (see <http://pi.ijs.si/>). Projects were included if evidence was found of past, current or planned activities in the areas of Machine Learning, Data Mining and Knowledge Discovery.

The report is structured as follows. The first section reviews general projects that are not specific to an application area. The following sections review projects classified by application area. The report concludes with a brief discussion and appendices with the lists of projects and their websites.

General and Technology-oriented Projects

Beside *KDNet* itself some other projects have addresses KDD in general. The Cost Action *KNOWLEST* with the primary objective to develop and implement computer systems for extracting previously unknown, non-trivial, and potentially useful knowledge from structurally complex, high-volume, distributed, and fast-changing scientific and R&D databases within the context of current and newly developing global computing and data infrastructures such as the GRID. The development of new discovery methodologies capable of effectively and efficiently extracting knowledge from such databases will be achieved not only by close collaboration of information scientists and database experts, but also by the readiness of both to adequately make themselves familiar with the motivations, goals, methodologies, and languages of the scientific fields involved.

In Framework V, several projects dealt with development of the KDD techniques involving distributed data and resources. *Sol-Eu-Net* with the main technical and scientific goals in research advances in data mining and decision support achieved through partners' involvement in collaborative problem solving of end-users problems. Methods for combining problem solutions and consensus building were developed, failed and successful approaches were analysed contributing to the better understanding of generic methodologies. In addition, a novel methodology was developed, combining data mining and decision support methods for problem solving. *MiningMart* has developed a model for meta-data together with its compiler and implements human-computer interfaces that allow database managers and case designers to fill in their application-specific meta-data. The system supports preprocessing and can be used stand-alone or in combination with a toolbox for the data mining step. A case base of best-practice cases is available on the internet. *AgentAcademy* developed a Data Mining framework for Training Intelligent Agents with a goal to develop an integrated environment for embedding intelligence in newly created agents through the use of Data Mining techniques. This was achieved by developing tools for assembling and maintaining a large repository of data on agent use and behavior, by providing an integrated environment for the systematic study of agent intelligence, and by developing a well-defined, well-specified model for an agent-training facility.

In Framework VI a great emphases have been put on distributed and grid computing (www.cordis.lu/ist/grids/projects.htm) and some of the projects involve KDD. The *SIMDAT* project aims to test and enhance grid technology for product development and production process design as well as to develop federated versions of problem-solving environments by leveraging enhanced Grid services. Further objectives are to exploit data grids as a basis for distributed knowledge discovery as well as to promote de facto standards for these enhanced Grid technologies across a range of disciplines and sectors. The project *DataMiningGrid* is aimed at adapting Data Mining technology to a grid computing environment by developing generic and sector-independent data mining tools and services for the Grid. Several applications from a diverse set of sectors will be used for demonstrating and promoting the developed technology.

Inductive logic programming is a research area lying at the intersection of inductive Machine Learning and logic programming. The general aim is to develop theories, techniques and applications of inductive learning from observations and background knowledge in a first order logical framework. *ILP2* aimed at closing the gap between conceptual work and applications by addressing four key application areas: natural language processing, data mining and discovery, design and configuration, and data-base design.

The project *cInQ* approached Data Mining from the perspective of databases. The purpose was to develop inductive queries on databases, along with algorithms for answering them, by analogy to database query languages. This result was indeed achieved and resulting in adaptations of existing methods to this new framework and an implemented system based on this principle.

One of the key open questions of artificial intelligence concerns probabilistic logic learning, i.e. the integration of probabilistic reasoning, with first order logic representations and Machine Learning. The overall goal of the *APrIL II* project is therefore to develop a sound theoretical understanding of probabilistic logic learning that enables one to develop effective probabilistic logic learning systems and to apply them on significant real-life applications. The project will develop a number of significant show-case applications of probabilistic logic learning in the area of bio-informatics, more specifically, concerning protein folding, metabolic pathways, and genetics and; develop the needed theory, probabilistic representations, learning algorithms and systems that enables one to learn interesting probabilistic logic models in real-life applications on the basis of data.

Information gathering and filtering, Digital libraries, Semantic Web

An important area of R&D involves using KDD methods for gathering and filtering information from text data also in connection to digital libraries and recently semantic Web. Different techniques have been used in the presented projects including document classification (using Naïve Bayes, Support Vector Machines, Kernel methods, k-Nearest Neighbor), document clustering (using k-Means clustering, hierarchical k-means clustering), user profiling based on content based filtering and collaborative filtering (with k-nearest neighbor as underlying classification algorithm), semi-automatic construction of topic ontologies, word classification for automatic lemmatization for several selected natural languages (using classification rules, inductive logic programming, ripple down rules).

Networks of excellence in Framework V related to information gathering and filtering are ***KDNET*** as a wide and fairly general in topic covering knowledge discovery in different forms and ***NEMIS*** focused on text mining and its applications.

Several R&D projects in Framework V addressed the topic to some extent. ***SOL-EU-NET*** project addressed different methods of Data Mining and Decision Support and a number of their practical applications among others also document filtering, organization, visualization, Web Mining and Link Analysis. More focused on text is the project ***LLAVES*** on unlocking topicality in text via investigates characteristics of clauses in written text with the objective of distinguishing different types of clause. ***KerMIT*** on Kernel Methods for Image and Text classification, clustering, ranking and filtering is the project that concerns the development of algorithms and software for the classification, clustering, ranking and filtering, both in an online and offline setting, of digital documents. In particular, the focus was on investigating the use of kernel methods in processing multilingual and multimedia documents involving text and images. ***CLARITY*** is on cross language information retrieval and organization of text and audio documents.

Different applications of methods for handling text data have been addressed in EU R&D Framework V projects. ***CERENA*** has addressed a problem of personalized information delivery by developing intelligent personal service environments with the objective to make electronic retail banking more competitive by using data mining technology to generate marketing rules for the delivery of services such as personalization to the customers of the bank. Tourism is an interesting application area for decision making and project ***DIETORECS*** has addressed it via intelligent recommendation for tourist destination by providing personalized recommendations based on user profile and contextual information and adapting the dialogue process as it learns more about the user. Tools for innovative publishing in science have been addressed in the project ***TIPS*** aiming to support the activities of document writing, reviewing, publishing, searching,

disseminating and reading, as well as the communication among members of the research community. *ASSAVID* has addressed a problem of automatic segmentation and semantic annotation of sports videos by segmenting the material into shots, and grouping and classify the shots into semantic categories (type of sport). In particular, information is extracted from each shot, based on speech and text recognition, and the highlights from the audio track and from visual audience reactions are identified.

While in Framework V projects we find different applications and development of methods connected to information gathering and filtering from different data sources including text, Framework VI puts more emphasis on the Semantic Web. There are two large Networks of Excellence in the area of semantic Web supported in Framework VI, *MUSCLE* and *KNOWLEDGE WEB*. The first involves multimedia understanding through semantics, computation and learning, while the second aims at extending Semantic Web enabled E-work and E-commerce technology to industry. Techniques of semantic-based systems for handling, acquiring, and processing knowledge embedded in multidimensional digital objects are addressed in Network of Excellence *AIM@SHAPE*. Its mission is to advance research in the direction of semantic-based shape representations and semantic-oriented tools to acquire, build, transmit, and process shapes with their associated knowledge. Another large Networks of Excellence in the area of digital libraries is *DELOS* that aims at developing generic digital libraries technology to be incorporated into industrial-strength Digital Library Management Systems, offering advanced functionality through reliable and extensible services.

Framework VI projects involving knowledge systems are mostly connected to semantics and semantic Web. *SEKT* a large, integrated project connecting three core technologies Ontology-based Metadata, Human Language Technology and Knowledge Discovery and developing several applications. The project aims at delivering software to: semi-automatically learn ontologies and extract metadata, and to maintain and evolve the ontologies and metadata over time; to provide knowledge access; besides middleware to effect integration of all the developed components. The project also aims at developing a methodology for using semantically-based knowledge management. The software components and the methodology are evaluated and refined through three case studies, in the legal, media and telecoms industries. *DIP* as an integrated project on Semantic Web Services is based around the claim that providing actual support in information processing and information exchange requires machine-processable semantics of data and information. The idea is that by using ontologies, the computer will be enabled as a device for querying and managing semi-structured information. Software programs can be accessed and executed via the web based on the idea of Web Services. The major mission of *DIP* is to further develop Semantic Web and Web Services and especially to enable their combination. Specific targeted project *ALVIS* is connecting semantics with search engines in developing superpeer semantic search engine. The project involves research in the design, use and interoperability of topic-specific search engines with the goal of developing an open source prototype of a distributed, semantic-based search engine.

In Slovenia, there are two national R&D projects on information gathering, digital libraries and text mining. One is on “*Construction of archive for Slovenian Web publications*” developing methodology and software for collection and analysis of Slovenian electronic publications on the Web. This includes development of methodology for constructing electronic publications archive based on the main characteristic of the publications, defining standards for bibliographic annotation of Web publications, preparation of legal material addressing copyright of the Web publications, design of software and the related infrastructure, definition of protocols for long-term maintenance of the archive. The other project builds on some of that results going beyond Web publications and aims at “*Development and analysis of Slovenian digitalized electronic publications of national importance*”, where the goal is to develop a prototype system for national archive of electronic publications with the corresponding analytical functions of electronic documents and taking into account bibliographic standards.

In Germany, there is a national research project *SemIPort* on Semantic Methods and Tools for Information Portals that address several problems related to data mining in general and especially to semantic content, usage and structure mining. Tools for dealing with complex, semantically enriched information have been developed, as for instance, the focused crawler METIS that uses an ontology for navigating the Web, the portal infrastructure SEAL for building a semantic portal on top of existing information sources, the semantic usage logging and domain ontology evolution tool OntoManager, the query expansion tool LibraryAgent, and the semantic data mining workbench Artemis. One of the main applications of the developed tools is the semantic integration and enrichment of several large, heterogeneous bibliographic metadata repositories (DBLP, CompScience, etc.) with over 1.5 million entries.

Language and Multimedia Learning

One of the major R&D areas is the use of learning methods for the construction of systems that perform tasks involving language. The overall methodology is to construct a corpus of language items with the associated meaning and use learning methods to construct a system that can assign meanings to new language items. Language can be text but it can also be speech or text combined with images (and speech). The main research problems in this area are the complexity of data (many dimensions and much information).

In Framework VI, beside *KDNET*, two Networks of excellence have addressed this area. The network *PASCAL* on Pattern Analysis, Statistical Modelling and Computational Learning Network of excellence addresses data mining, knowledge discovery and machine learning with an emphasis on statistical approaches. The network *MUSCLE* on Multimedia Understanding through Semantics, Computation and Learning aims at collaboration between research groups in multimedia data mining on the one hand and machine learning on the other in order to make breakthrough progress towards the following objectives: i) Harnessing the full potential of machine learning and cross-modal interaction for the (semi-) automatic generation of metadata with high semantic

content for multimedia documents, ii) Applying machine learning for the creation of expressive, context-aware, self-learning, and human centered interfaces that will be able to effectively assist users in the exploration of complex and rich multimedia content.

In Framework V and VI there are several European projects that involve multimedia technology but few of these involve substantial work on machine learning, data mining or knowledge discovery in databases. Examples from Framework V are *IMAGINE* on Interfacing Mobile Applications with Voice Natural Language Interface, *MAGELAN* on Multimedia And Game Enhanced Learning And Networking, *BINDEX* on Bilingual Automatic Parallel Indexing and Classification, *MUSA* on MULTilingual Subtitling of multimedia content, *MATCHPAD* on MACHine Translation systems for the use of Hungarian and Polish Administrations. *SIMAC* on developing methods for indexing, annotating and personalising music for consumers via learning to categorize, characterize and recommend popular music handled as audio signal. *CROSSMARC* applies state-of-the-art language engineering and machine learning tools and techniques to achieve commercial strength technology for information extraction from web pages. The project developed a platform for cross-lingual information management. Project *aceMedia* is an integrated project in Framework VI aimed at optimising multimedia presentations. The main technological objectives are to discover and exploit knowledge inherent to the content in order to make content more relevant to the user; to automate annotation at all levels; and to add functionality to ease content creation, transmission, search, access, consumption and re-use. Collecting information about users and automated indexing of materials is likely to require KDD methods but information about this is not available.

The aim of a large national project in Slovenia, *Language resources for the Slovene language*, is to develop text corpora and software tools for researching Slovene texts and the Slovene language in general including a system for lemmatization based on machine learning techniques. It is designed as the qualitative (written and spoken corpus, internet texts) and quantitative (200.000.000 words) upgrading of the existing Slovene reference corpus FIDA developed by the same partners. The project represents a major step forward in developing research and language-policy infrastructure in linguistics, social studies and information technology for Slovenian language.

Knowledge Management, E-commerce

Application of knowledge discovery techniques in knowledge management and e-commerce brings an important contribution to these applications until with rather strong emphases on development.

In Framework V several R&D projects covered different application problems related to knowledge management and e-commerce. *ONTOWEB* as a Network of Excellence was centered on ontology-based information exchange for knowledge management and electronic commerce. *SMART2EAM* aimed at knowledge management for smart technology transfer while sharing of corporate knowledge was addressed in *COMMA* through corporate memory management using intelligent agents and in *CORMA* via development of practical tools and methods for corporate knowledge management,

sharing and capitalising engineering know-how in the concurrent enterprise. Ontology driven temporal text mining was addressed in *PARMENIDES* focused on organisational knowledge management via developing an ontology driven systematic approach for integrating the entire process of information gathering, processing and analysis. In addition a large Network of excellence *KMFORUM* was established with the objective to bring together the available critical mass of knowledge management experts in Europe in order to share and exchange the latest developments in the domain and to develop visions for the future.

Framework VI has put additional emphases on knowledge management and semantic technologies. One of the biggest project in the area are already mentioned *SEKT* on semantically-enable knowledge technologies and *DIP* on data, information, and process integration with semantic web services (see Section on Information gathering). *KB20* as a kind of successor of KM Forum aims at developing the European knowledge space. Multimedia in knowledge management is addressed in *MUSCLE*, a project on multimedia understanding through semantics, computation and learning (see Section on Language and Multimedia Learning). *PALLIANET* addresses decision support and knowledge driven collaborative practices in palliative care, where Text Mining methods are used to extract knowledge from text. The *Dot.kom* project uses Machine Learning for Information Extraction and automated annotating text documents. Machine Learning is a key technology for this because manually constructing annotation rules is more difficult than manually tagging examples and inducing rules.

Marketing and profiling, Business mapping, Competitor Analysis, Risk Management

One use of Machine Learning and Data Mining is the construction of predictive models of processes from observations. Projects in Framework V that take this approach in a business environment address problems like predicting risks and expected benefits of commercial projects (*PRIMA* on Project risk management) and various social-economic developments as in *TERRA2000*. Both projects use tools from statistics and machine learning to discover and extrapolate trends but do not aim at extending the existing technology. Another example is *EICSTES* on European Indicators, Cyberspace and the Science-Technology-Economy System. This project is aimed at discovering social network structures in the area of science, technology and their clients in industry and government. This project uses sophisticated analysis methods in novel ways. The project *SHOWRISK* aims to develop procedures to predict the risk of business loans using balance sheet data of enterprises. It uses an ensemble of many plausible models for prediction. This allows to derive the uncertainty of predictions and classifications as well as to estimate the value of new information.

In Framework VI an example of this type of project is *COCOMBINE* on Competition Contents and Broadband for the Internet in Europe. This project is aimed at market analysis/modeling. It uses methods from Data Mining and Machine Learning with

emphasis on visualization and discovery of relations between actors in a market. Personalization is addressed in *VIP ADVISOR* on virtual, independent advisor for personal insurance and finance risk management. The main idea is to develop a virtual personal insurance and finance assistant with new means of interaction. This personal assistant will be specialized in risk management counseling for small and medium enterprises but could be extended towards general insurance counseling for private persons.

E-science

E-science applications involve the integration of Data Mining and KDD into scientific research. Learning methods are used to construct models of (large amounts of) data. An important aspect of E-Science is collaborative work by scientists in different disciplines and at different locations. Currently European projects of this type address genetics and chemical engineering.

In Framework V the *SOL-EU-NET* project has developed a number of prototype solutions for customers in different areas including industrial data mining to explore the possibility of distributed collaborative work on problems needing data mining and decision support. *OpenMolGRID* (Open computing GRID for MOLEcular science and engineering) uses data mining for relating properties of chemical structures to functions/effects to support prediction of effects and design. One of the goals of this project is to develop a grid-based infrastructure that enables the use of distributed data and reduces turn around time of complex computations by distributed computing. The emphasis is not so much on developing new (KDD) technology but on developing an integration of existing tools and functionalities. ADC's Data Mining Resources for Space Science Data mining resource guide for space sciences -- investigating data mining of large scientific databases, with specific astronomy applications to the new Virtual Observatory initiatives. The goal of the *BioMinT* project is to develop a generic text mining tool that (1) interprets diverse types of query, (2) retrieves relevant documents from the biological literature, (3) extracts the required information, and (4) outputs the result as a database-slot filler or as a structured report. The consortium consists of biologists and data/text mining groups.

E-learning

Several projects pursue the use of machine learning methods for personalization in the context of e-learning. Specifically, information about a user (documents read and studied, education, CV information) is used to infer a user profile that is then used to select or actively search for suitable educational material. A more ambitious task is to adapt teaching material to the user on the basis of an induced profile.

ELEARN (on E-learning) and **PROLEARN** (Professional Learning) are networks of excellence that address the use of information technology in learning and teaching. Both address the use of automated construction of learner profiles and their use in educational activities. Key issues in these efforts are the development of standardized representations and concepts for user profiles and corresponding indices for annotating learning materials. Other issues that are mentioned are the problem of privacy and the evaluation of personalization. Examples of such projects are **CELEBRATE** on Context ELearning with BROADband TEchnologies, **LEACTIVEMATH** on Language-Enhanced, User Adaptive, Interactive eLearning for Mathematics.

OpenMolGRID uses UNICORE to integrate the various applications needed, from databases to molecular engineering and prediction modules. Building on the basic Grid middleware, substantial functionality is added to the client, and in the abstraction layers used to hide the system complexity from the user.

STARDEX on Statistical and Regional dynamical Downscaling of Extremes for European regions is part of a co-operative cluster of projects exploring future changes in extreme events in response to global warming. The other members of the cluster are **MICE** and **PRUDENCE**. Statistical methods are applied to a variety of data to identify changes in the climate. Machine learning methods are a complement to traditional statistical analysis techniques.

We already mentioned the Network of Excellence **MUSCLE** (Multimedia Understanding through Semantics, Computation and Learning) aimed at creating and supporting a pan-European to foster close collaboration between research groups in multimedia Data Mining on the one hand and machine learning on the other in order to make breakthrough progress towards the following objectives: (1) Harnessing the full potential of machine learning and cross-modal interaction for the (semi-) automatic generation of metadata with high semantic content for multimedia documents. (2) Applying machine learning for the creation of expressive, context-aware, self-learning, and human-centered interfaces that will be able to effectively assist users in the exploration of complex and rich multimedia content.

The German project **DaMiT** aimed at developing an intelligent tutoring system where the users can learn about the foundations, techniques and application of data mining. The content is provided in different media, e.g. text, animations, video, and last not least by the integration of real data mining environments. The system supports learning by doing; for example, typical complex data mining tasks are given to the students in form of competitive exercises. The tutoring system is adaptable and adaptive, e.g. based on the user model it gives recommendations about the suitability of the chosen learning content. Another example is that the user can select its preferred presentation style. The system also supports different user roles with different access rights and is equipped with an electronic payment system.

Computer Security and Data Privacy

Protection of privacy and in general security is an important issue for KDD. *PISA* on Privacy Incorporated Software contributes at building a model of a software agent within a network environment, to demonstrate that it is possible to perform complicated actions on behalf of a person, without the personal data of that person being compromised. *PAMPAS* on Pioneering Advanced Mobile Privacy And Security and *PRIME* on Privacy and Identity Management for Europe develop technology for protecting privacy issues in the context of mobile systems such as mobile phones and systems collecting data about cars (PAMPAS) and more in general in networked systems (PRIME).

Cultural Heritage

In this application area there are hardly any projects that involve work on Machine Learning. Some examples involve the use of learning methods for building user profiles for personalisation. Examples are the Framework V project *ARCHEOGUIDE* on Augmented Reality-based Cultural Heritage On-site Guide that intends to add personalisation to a system with presentations based on virtual reality. *M-PIRO* on Multilingual Personalised Information Objects developed the concept of a personalized information object capable of responding to a request for information. The project addressed the domain of heritage environment working closely with museums, galleries, and other "memory institutions" to develop technology which is responsive to their particular needs. The developed technology allows written and spoken descriptions of exhibits to be generated automatically from an underlying language-neutral database and existing free-text descriptions. The resulting descriptions, which are generated in English, Greek or Italian, are tailored according to the user's interests, background knowledge, and language skills.

Health care

Health care is one of the first applications areas where Machine Learning and KDD methods were used and is a constant source of ideas and research problems for KDD.

In Framework V several R&D project addressed different problems connected to health care. *EPI-MEDICS* on Enhanced Personal, Intelligent and Mobile system for Early Detection and Interpretation of Cardiological Syndromes, involved monitoring the ECG of a patient using a simple portable and intelligent personal ECG monitor. Based on build-in intelligent self-adaptive serial ECG processing and decision making techniques, a number of different levels of alarm could be generated. Using wireless communication protocols like Bluetooth and GSM/GPRS alarm messages can be automatically transmitted to an emergency call center. Another project on health monitoring is *Lifebelt* on intelligent wearable device for health monitoring during pregnancy. The key objective of the project is research, development and validation of a wearable device equipped with

advanced biosensors for the monitoring of fetal and maternal health during pregnancy. The system is designed to monitor and to evaluate human vital signs such as fetal and maternal ECG, heart rate, uterine contractions, blood pressure, temperature, weight, stress and abdomen growth. The wearable device transmits the collected vital signs to a centralized hospital system for monitoring. The project also aimed at providing a decision support tool for the obstetrician, who is enabled to remotely monitor patients, evaluate automated diagnosis results, access patients' medical data anytime, anyplace and be alerted when potential pregnancy complications occur. The main goals of the project included research on knowledge-based methods and intelligent classification algorithms for the signal processing and automated diagnosis and developing statistical medical data generation and prediction of trends.

In Framework VI, there is a Network of Excellence *SemanticMining* on semantic interoperability and Data Mining in Biomedicine that aims at development of generic methods and tools supporting the critical tasks of the field; data mining, knowledge discovery, knowledge representation, abstraction and indexing of information, semantic-based information retrieval in a complex and high-dimensional information space, and knowledge-based adaptive systems for provision of decision support for dissemination of evidence based medicine. *CARE-PATHS* on intelligent support environment to improve the quality of decision processes in health communities, studies methods and systems for monitoring and analytical tools for clinical pathways variance. Managing clinical pathways requires first to gather historical data regarding the application of pathways to all patients, in an original or modified way. The evaluation of a patient's variance can first be used by the decision support tools to propose changes in the patient's pathway.

There are several projects in where KDD methods fit naturally and can be used but it is not completely clear from the project description if that is the case. For instance, in Framework V projects on health mobile systems, such as *HealthMate*, *MobiHealth* or intelligent information systems such as *IHE-E*. In Framework VI projects involving collecting data and improving decision making process such as *PALLIANET* and *AMICA* probably involve KDD methods. *HealthMate* on personal intelligent health mobile systems for Telecare and Tele-consultation aimed at developing portable personal systems for health Tele-care and Tele-consultation based on innovative wireless technologies to configure a secure information exchange media between the personal systems and the health service providers and assure service continuity at any time and place. *MobiHealth* on mobile health care aims at developing and trialing new mobile value-added services in the area of healthcare, thus bringing healthcare to the patient with a system customized to their individual health needs. Physical measurements such as blood pressure or ECG transmitted wirelessly to their doctor, the hospital or their health call centre. *IHE-E* on integrating the health care enterprise in Europe aims at developing and providing application scenarios as well as technical requirements for the integration of the health care information systems, imaging systems and other modalities. *PALLIANET* on Decision Support and knowledge driven collaborative practices in Palliative care, aims to collect documents (emails, chat logs, articles, interaction with information systems) and extract knowledge that is made accessible to physicians in palliative care. A particular goal is to identify and publish best practice. *AMICA* on

assembling data and knowledge at the point of care to improve medical decision making and prevent errors, focuses more on collection information about patients in combination with management of documents. **PROTCURE** on integrating formal methods in the development process of medical guidelines and protocols aims at developing medical protocols. In all these projects systematic structured representation of medical information is a key issue and ontologies are an approach for this. Machine Learning plays a minor role in analysis of patient data and in analyzing collections of documents for indexing or extracting information.

The Dutch project **DUMPERS** on Distributed User Modeling for Personalised Exploring Recommender Systems, develops methods for user modeling and recommending for an information system for elderly people. Machine Learning methods are used to construct models of users based on click stream data. The resulting model is the basis of a recommender systems or improvements to the navigation structure.

Various

Some projects involving KDD related to other areas not mentioned in the previous Sections are described here.

One of the projects (DEMETRA) uses neuro-fuzzy approach to learning, in particular integrating the learned and manually built models using statistical integration, fuzzy aggregation and connectionist integration.

In Framework V, georeference data was addressed in project **SPIN!** with the main objective to offer new possibilities for the analysis of georeferenced data. To this end a Spatial Data Mining system was developed that integrates state of the art Geographic Information System and Data Mining functionality in an open, highly extensible, internet-enabled plug-in architecture. **HYPERGEO** aimed at providing technical tools to enable value-added information services related to Geographic information. The technical objectives were to develop and integrate, in single system, innovative software components enabling the users to formulate advance requests, to access information in both pull and push modes, and to display the information in efficient multi-layered form. Research and development activities included: distributed information access and management (data mining), natural language processing, user profiling, information presentation. **DEMETRA** on Development of Environmental Modules for Evaluation of Toxicity of pesticide Residues in Agriculture aims to develop software for a quantitative prediction of the toxicity of a molecule, in particular molecules of pesticides, candidate pesticides, and their derivatives. The project involves development of new approaches to knowledge representation for predictive toxicology, knowledge representation for hybrid intelligent systems and soft computing techniques, machine learning-based models for QSARs, integration of developed models. **ReBuilder** is an intelligent software design tool that helps the software designer to create new system designs easily. It takes advantage of software reuse and artificial intelligence techniques to: suggest new design approaches,

search helpful past designs, promote the designer creativity and problem solving, verify automatically new created designs.

OASIS (Open Advanced System for Improved Crisis Management) develops integrated systems for crisis management in control rooms. The R&D activities include Knowledge management and knowledge-based decision support (data analysis support tool, event recognition and prediction tools, “intelligent checklist” for finding and building emergency action plans, etc.) and advanced technologies for handling decision complexity including, data mining, data aggregation, multi-criteria optimisation, visual data analysis etc.

Discussion

This report gives a compact and high-level overview of recent and current projects in the areas of Machine Learning, Data Mining and Knowledge Discovery. The overview does not include smaller national research projects at universities, research centers and industry nor does it take the success of projects into account. Therefore it is biased towards the types of projects that are favoured by the European Commission. Keeping this in mind, we can make a few observations.

Table 1 gives the number of projects by area. Numbers include networks and regular projects. Projects that appear under several headings are counted under each heading.

Area	Number
General and technology-oriented	10
Information gathering and filtering, Digital libraries, Semantic Web	20
Language and Multimedia Learning	12
Marketing and profiling, Business mapping, Competitor Analysis, Risk Management	6
E-Science	3
E-Learning	9
Computer Security and Data Privacy	3
Cultural Heritage	2
Health care	13
Various	3

Table 1: Number of projects by area

The overview gives rise to several observations.

Most projects are application-oriented and not technology-oriented

The number of projects in the 6th framework programme of the European Commission in which the technology of Machine Learning, Data Mining or Knowledge Discovery is a central focus is quite small. The methodology-oriented projects are Almost all projects are organised around application objectives and not about technical objectives such as new methods, paradigms or theoretical understanding. Exceptions are APRIL (on Relational Learning) and CinQ (on inductive databases). This is probably a consequence of the fact that EC policy is aimed at economic development and not on scientific research (although this may change in the near future with the advent of the European Research Area). We do not know to which extent the results of the application-oriented projects contribute to scientific knowledge.

The most popular application areas are related to Semantic Web, Language, Multimedia and Health Care.

Although it must be interpreted with care, Table 1 shows clear difference in the volume of projects. The most popular areas are those that involve language, images and sound (multimedia), application in information and knowledge management and in health care.

APPENDIX 1: Alphabetical list of websites of European projects

<i>PROJECT</i>	<i>WEBSITE</i>
Acemedia	www.acemedia.org
ADC - NASA's Astronomical Data Center	adc.gsfc.nasa.gov/adc/
AgentAcademy: A Data Mining framework for Training Intelligent Agents	agentacademy.iti.gr
AIM@SHAPE	www.aimatshape.net
ALVIS - Superpeer Semantice Search Engine	www.alvis.info
APRIL - Application of Probabilistic Inductive Logic Programming II	choyu.informatik.uni-freiburg.de
ARCHEOGUIDE - Augmented Reality-based Cultural Heritage On-site Guide	archeoguide.intranet.gr
ASSAVID	viplab.dsi.unifi.it/ASSAVID/
ASSO	www.info.fundp.ac.be/asso/
BINDEX - Bilingual Automatic Parallel Indexing and Classification	ews.e-technik.uni-ulm.de/deutsch/bindex.htm
BioMinT (Biological Text Mining)	www.biomint.org
CARE-PATHS	www.carepaths.eupm.net/my_spip/index.php
CERENA	www.cerena.org
cInQ	www.cinq-project.org
CLARITY	www.dcs.shef.ac.uk/research/groups/nlp/clarity/
COCOMBINE - Competition Contents and Broadband for the Internet in Europe	www.cocombine.org
Computational Methods in Medical Genetics and Expression Data Analysis	www.cs.helsinki.fi/research/fdk/
COMMA - Corporate Memory Management through Agents	www.ii.atos-group.com/sphia/comma/
CORMA - Practical Tools and Methods for Corporate Knowledge Management - Sharing and Capitalising Engineering Know-How in the Concurrent Enterprise	www.corma.net
CROSSMARC - Cross-lingual multi-agent retail comparison	www.iit.demokritos.gr/skel/crossmarc/
DaMiT - Data Mining Tutor	damit.dfki.de
DataMiningGrid	www.datamininggrid.org

DELOS	www.delos.info
DEMETRA - Development of Environmental Modules for Evaluation of Toxicity of pesticide Residues in Agriculture	www.demetra-tox.net
DIETORECS	dietorecs.itc.it
DIP - Data, Information, and Process Integration with Semantic Web Services	dip.semanticweb.org
Dot.KOM	nlp.shef.ac.uk/dot.kom/
EICSTES - European Indicators, Cyberspace And The Science-Technology-Economy System	www.eicstes.org
ELEARNTN	www.elearntn.org
EPI-MEDICS - Enhanced Personal, Intelligent and Mobile system for Early Detection and Interpretation of Cardiological Syndromes	epi-medics.univ-lyon1.fr
HealthMate	www.healthmate-project.org
HYPERGEO	www.soi.city.ac.uk/~dmm/hypergeo/
IHE-E - Integrating the Health Care Enterprise in Europe Concerted Action	www.ihe-europe.org
ILP2	www.cs.kuleuven.ac.be/~ml/esprit/esprit.ilp2.20237.html
IMAGINE - Interfacing Mobile Applications with Voice Natural Language Interface	www.hltcentral.org/projects/detail.php?acronym=imagine
KB20 - The European Knowledge Space	www.knowledgeboard.com
KDNET	www.kdnet.org
KERMIT- Kernel Methods for Images and Text	www.eurokermit.org
KM FORUM - European Knowledge Management Forum	www.knowledgeboard.com
KNOWLEDGE WEB	knowledgeweb.semanticweb.org
KNOWLEST (COST Action 282)	www.mpa-garching.mpg.de/~opmolsrv/COST282/index.html
LEACTIVEMATH - Language-Enhanced, User Adaptive, Interactive eLearning for Mathematics	www.leactivemath.org
Lifebelt - An intelligent wearable device for health monitoring during pregnancy	www.lifebelt.eu.com
LLAVES	www.sics.se/humle/projects/llaves/
MATCHPAD - MACHine Translation systems for the use of Hungarian and Polish Administrations	www.hltcentral.org/projects/detail.php?acronym=matchpad
METAL - A Meta-Learning Assistant in	www.metal-kdd.org

Machine Learning and Data Mining	www.liacc.up.pt/ML/METAL/
MICE – Modeling the impact climate extremes	www.cru.uea.ac.uk/cru/projects/mice/
MiningMart	www-ai.cs.uni-dortmund.de/MMWEB/index.html
MobiHealth - Innovative GPRS/UMTS mobile services for applications in healthcare	www.mobihealth.org
M-PIRO - Multilingual Personalised Information Objects	www.ltg.ed.ac.uk/mpiro/
MUSA - MULTilingual Subtitling of multimedia content	sifnos.ilsp.gr/musa/
MUSCLE - Multimedia Understanding through Semantics, Computation and Learning	www.cwi.nl/projects/muscle/
NEMIS Network of Excellence in text Mining and Its applications in Statistics	nemis.cti.gr
OASIS - Open Advanced System for Improved Crisis Management	http://www.oasis-fp6.org/
ONTOWEB	www.ontoweb.org/misc.htm
OPENMOLGRID	www.openmolgrid.org/project/index.html
PALLIANET - Decision Support and Knowledge driven Collaborative practices in Palliative Care	www.pallianet.eupm.net/my_spip/index.php
PAMPAS - Pioneering Advanced Mobile Privacy And Security	www.pampas.eu.org
PARMENIDES	www.crim.co.umist.ac.uk/parmenides/
PASCAL	www.pascal-network.org
PISA - Privacy Incorporated Software Agent	pet-pisa.openspace.nl/pisa_org/pisa/
PRIMA - Project risk management	www.esi2.us.es/prima/
PRIME - Privacy and Identity Management for Europe	www.prime-project.eu.org
PROLEARN	www.prolearn-project.org
PROTOCURE - Integrating Formal Methods in the Development Process of Medical Guidelines and Protocols	www.protocure.org
PRUDENCE - Prediction of Regional scenarios and Uncertainties for Defining European Climate change risks and Effects	prudence.dmi.dk
ReBuilder - Intelligent Reuse of Software Objects	eden.dei.uc.pt/~pgomes/ReBuilder/
STARDEX - Statistical and Regional dynamical Downscaling of Extremes for European regions	www.cru.uea.ac.uk/cru/projects/stardex/
SEKT - Semantically-Enable Knowledge Technologies	sekt.semanticweb.org

SEMANTICMINING - Semantic Interoperability and Data Mining in Biomedicine	www.semanticmining.org
SHOWRISK	ais.gmd.de/KD/showrisk.html
SIMAC - Semantic Interaction with Music Audio Contents	www.semanticaudio.org
SIMDAT	www.scai.fraunhofer.de/index.php?id=587&L=1
SMART2EAM	www.smartcities.co.uk
Sol-Eu-Net	soleunet.ijs.si/
SPIN!	www.ais.fraunhofer.de/KD/spin.html
STARDEX - Statistical and Regional dynamical Downscaling of Extremes	www.cru.uea.ac.uk/cru/projects/stardex/
TERRA2000	www.terra-2000.org/
TIPS	tips.sissa.it
VIP ADVISOR	Vip-advisor.fi.upm.es/

APPENDIX 2: Alphabetical list of national project websites

<i>PROJECT</i>	<i>WEBSITE</i>
ADC - NASA's Astronomical Data Center	adc.gsfc.nasa.gov/adc/
DaMiT - Data Mining Tutor	damit.dfki.de/
Development and analysis of Slovenian digitalized electronic publications of national importance	sicris.izum.si/search/prj.aspx?lang=eng&id=3452&opt=1
DUMPERS	www.swi.psy.uva.nl/dumpers/
Language resources for the Slovene language	sicris.izum.si/search/prj.aspx?lang=eng&id=3515&opt=1
SemIPort - Semantic Methods and Tools for Information Portals	km.aifb.uni-karlsruhe.de/semiport