Foreword

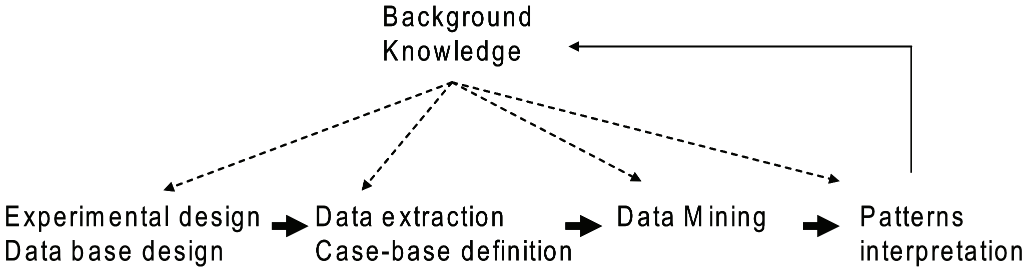
Current research directions are looking at Data Mining (DM) and Knowledge Management (KM) as complementary and interrelated fields, aimed at supporting, with algorithms and tools, the lifecycle of knowledge, including its discovery, formalization, retrieval, reuse, and update. While DM focuses on the extraction of patterns, information, and ultimately knowledge from data (Giudici, 2003; Fayyad et al., 1996; Bellazzi, Zupan, 2008), KM deals with eliciting, representing, and storing explicit knowledge, as well as keeping and externalizing tacit knowledge (Abidi, 2001; Van der Spek, Spijkervet, 1997). Although DM and KM have stemmed from different cultural backgrounds and their methods and tools are different, too, it is now clear that they are dealing with the same fundamental issues, and that they must be combined to effectively support humans in decision making.

The capacity of DM to analyze data and to extract models, which may be meaningfully interpreted and transformed into knowledge, is a key feature for a KM system. Moreover, DM can be a very useful instrument to transform the tacit knowledge contained in transactional data into explicit knowledge, by making experts' behavior and decision-making activities emerge.

On the other hand, DM is greatly empowered by KM. The available, or background knowledge, (BK) is exploited to drive data gathering and experimental planning, and to structure the databases and data warehouses. BK is used to properly select the data, choose the data mining strategies, improve the data mining algorithms, and finally evaluates the data mining results (Bellazzi, Zupan, 2008; Bellazzi, Zupan, 2008). The output of the data analysis process is an update of the domain knowledge itself, which may lead to new experiments and new data gathering (see Figure 1).

If the interaction and integration of DM and KM is important in all application areas, in medical applications it is essential (Cios, Moore, 2002). Data analysis in medicine is typically part of a complex reasoning process which largely depends on BK. Diagnosis, therapy, monitoring, and molecular research are always guided by the existing knowledge of the problem domain, on the population of patients or on the specific patient under consideration. Since medicine is a safety critical context (Fox, Das, 2000),

Figure 1. Role of the background knowledge in the data mining process

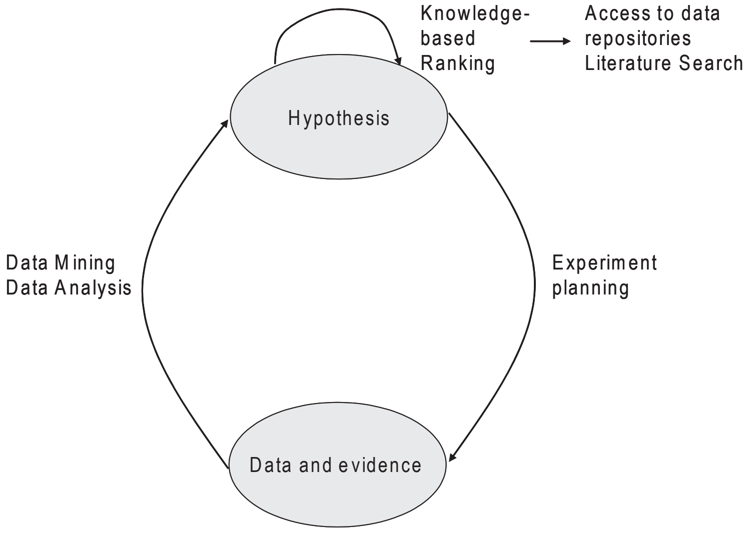


decisions must always be supported by arguments, and the explanation of decisions and predictions should be mandatory for an effective deployment of DM models. DM and KM are thus becoming of great interest and importance for both clinical practice and research.

As far as clinical practice is concerned, KM can be a key player in the current transformation of healthcare organizations (HCO). HCOs have currently evolved into complex enterprises in which managing knowledge and information is a crucial success factor in order to improve efficiency, (i.e. the capability of optimizing the use of resources, and efficacy, i.e. the capability to reach the clinical treatment outcome) (Stefanelli, 2004). The current emphasis on Evidence-based Medicine (EBM) is one of the main reasons to utilize KM in clinical practice. EBM proposes strategies to apply evidence gained from scientific studies for the care of individual patients (*Sackett, 2004)*. Such strategies are usually provided as clinical practice guidelines or individualized decision making rules and may be considered as an example of explicit knowledge. Of course, HCO must also manage the empirical and experiential (or tacit) knowledge mirrored by the day-by-day actions of healthcare providers. An important research effort is therefore to augment the use of the so-called "process data" in order to improve the quality of care (Montani et al., 2006; Bellazzi et al. 2005). These process data include patients' clinical records, healthcare provider actions (e.g. exams, drug administration, surgeries) and administrative data (admissions, discharge, exams request). DM may be the natural instrument to deal with this problem, providing the tools for highlighting patterns of actions and regularities in the data, including the temporal relationships between the different events occurring during the HCO activities (Bellazzi et al. 2005).

Biomedical research is another driving force that is currently pushing towards the integration of KM and DM. The discovery of the genetic factors underlying the most common diseases, including for example cancer and diabetes, is enabled by the concurrence of two main factors: the availability of data at the genomic and proteomic scale and the construction of biological data repositories and ontologies, which accumulate and organize the considerable quantity of research results (Lang, 2006). If we represent the current research process as a reasoning cycle including inference from data, ranking of the hypothesis and experimental planning, we can easily understand the crucial role of DM and KM (see Figure 2).

Figure 2. Data mining and knowledge management for supporting current biomedical research


Knowledge-based Ranking
Access to data repositories Literature Search
Hypothesis
Data Mining Data Analysis
Experiment planning
Data and evidence

In recent years, new enabling technologies have been made available to facilitate a coherent integration of DM and KM in medicine and biomedical research.

Firstly, the growth of Natural Language Processing (NLP) and text mining techniques is allowing the extraction of information and knowledge from medical notes, discharge summaries, and narrative patients' reports. Rather interestingly, this process is however, always dependent on already formalized knowledge, often represented as medical terminologies (Savova et al., 2008; Cimiano et al., 2005).

Indeed, medical ontologies and terminologies themselves may be learned (or at least improved or complemented) by resorting to Web mining and ontology learning techniques. Thanks to the large amount of information available on the Web in digital format, this ambitious goal is now at hand (Cimiano et al., 2005).

The interaction between KM and DM is also shown by the current efforts on the construction of automated systems for filtering association rules learned from medical transaction databases. The availability of a formal ontology allows the ranking of association rules by clarifying what are the rules confirming available medical knowledge, what are surprising but plausible, and finally, the ones to be filtered out (Raj et al., 2008).

Another area where DM and KM are jointly exploited is Case-Based Reasoning (CBR). CBR is a problem solving paradigm that utilizes the specific knowledge of previously experienced situations, called cases. It basically consists in retrieving past cases that are similar to the current one and in reusing (by, if necessary, adapting) solutions used successfully in the past; the current case can be retained and put into the case library. In medicine, CBR can be seen as a suitable instrument to build decision support tools able to use tacit knowledge (Schmidt et al., 2001). The algorithms for computing the case similarity are typically derived from the DM field. However, case retrieval and situation assessment can be successfully guided by the available formalized background knowledge (Montani, 2008).

Within the different technologies, some methods seem particularly suitable for fostering DM and KM integration. One of those is represented by Bayesian Networks (BN), which have now reached maturity and have been adopted in different biomedical application areas (Hamilton et al., 1995; Galan et al., 2002; Luciani et al., 2003). BNs allow to explicitly represent the knowledge available in terms of a directed acyclic graph structure and a collection of conditional probability tables, and to perform probabilistic inference (Spiegelhalter, Lauritzen, 1990). Moreover, several algorithms are available to learn both the graph structure and the underlying probabilistic model from the data (Cooper, Herskovits, 1992; Ramoni, Sebastiani, 2001). BNs can thus be considered at the conjunction of knowledge representation, automated reasoning, and machine learning. Other approaches, such as association and classification rules, joining the declarative nature of rules, and the availability of learning mechanisms including inductive logic programming, are of great potential for effectively merging DM and KM (Amini et al., 2007).

At present, the widespread adoption of software solutions that may effectively implement KM strategies in the clinical settings is still to be achieved. However, the increasing abundance of data in bioinformatics, in health care insurance and administration, and in the clinics, is forcing the emergence of clinical data warehouses and data banks. The use of such data banks will require an integrated KM-DM approach. A number of important projects are trying to merge clinical and research objectives with a knowledge management perspective, such as the I2B2 project at Harvard (Heinze et al. 2008), or, on a smaller scale, the Hemostat (Bellazzi et al. 2005) and the Rhene systems in Italy (Montani et al., 2006). Moreover, several commercial solutions for the joint management of information, data, and knowledge are available on the market. It is almost inevitable that in the near future, DM and KM technologies will be an essential part of hospital and research information systems.

The book "Data Mining and Medical Knowledge Management: Cases and Applications" is a collection of case studies in which advanced DM and KM solutions are applied to concrete cases in biomedical research. The reader will find all the peculiarities of the medical field, which require specific solutions to complex problems. The tools and methods applied are therefore much more than a simple adaptation of general purpose solutions: often they are brand-new strategies and always integrate data with knowledge. The DM and KM researchers are trying to cope with very interesting challenges, including the integration of background knowledge, the discovery of interesting and non-trivial relationships, the construction and discovery of models that can be easily understood by experts, the marriage of model discovery and decision support. KM and DM are taking shape and even more than today they will be in the future part of the set of basic instruments at the core of medical informatics.

Riccardo Bellazzi
Dipartimento di Informatica e Sistemistica, Università di Pavia

REFERENCES

Abidi, S. S. (2001). Knowledge management in healthcare: towards 'knowledge-driven' decision-support services. *Int J Med Inf,* 63, 5-18.

Amini, A., Muggleton, S. H., Lodhi, H., & Sternberg, M.J. (2007). A novel logic-based approach for quantitative toxicology prediction. *J Chem Inf Model*, 47(3), 998-1006.

Bellazzi, R., Larizza, C., Magni, P., & Bellazzi, R. (2005). Temporal data mining for the quality assessment of hemodialysis services. *Artif Intell Med*, 34(1), 25-39.

Bellazzi, R., & Zupan, B. (2007). Towards knowledge-based gene expression data mining. *J Biomed Inform*, 40(6), 787-802.

Bellazzi, R, & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform*, 77(2), 81-97.

Cimiano, A., Hoto, A., & Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24, 305-339.

Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artif Intell Med,* 26, 1-24.

Cooper, G. F, & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9, 309-347.

Dudley, J., & Butte, A. J. (2008). Enabling integrative genomic analysis of high-impact human diseases through text mining. *Pac Symp Biocomput*, 580-591.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Data mining and knowledge discovery in databases. *Communications of the ACM*, 39, 24-26.

Fox, J., & Das, S. K. (2000). *Safe and sound: artificial intelligence in hazardous applications*. Cambridge, MA: MIT Press.

Galan, S. F., Aguado, F., Diez, F. J., & Mira, J. (2002). NasoNet, modeling the spread of nasopharyngeal cancer with networks of probabilistic events in discrete time. *Artif Intell Med*, 25(3), 247-264.

Giudici, P. (2003). *Applied Data Mining, Statistical Methods for Business and Industry*. Wiley & Sons.

Hamilton, P. W., Montironi, R., Abmayr, W., et al. (1995). Clinical applications of Bayesian belief networks in pathology. *Pathologica*, 87(3), 237-245.

Heinze, D. T., Morsch, M. L., Potter, B. C., & Sheffer, R.E Jr. (2008). Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology. *J Am Med Inform Assoc*, 15(1), 40-3.

Lang, E. (2006). Bioinformatics and its impact on clinical research methods. Findings from the Section on Bioinformatics. *Yearb Med Inform*, 104-6.

Luciani, D., Marchesi, M., & Bertolini, G. (2003). The role of Bayesian Networks in the diagnosis of pulmonary embolism. *J Thromb Haemost*, 1(4), 698-707.

Montani, S. (2008). Exploring new roles for case-based reasoning in heterogeneous AI systems for medical decision support. *Applied Intelligence*, 28(3), 275-285.

Montani, S., Portinale, L., Leonardi, G., & Bellazzi, R. (2006). Case-based retrieval to support the treatment of end stage renal failure patients. *Artif Intell Med*, 37(1), 31-42.

Raj, R., O'Connor, M. J., & Das, A. K. (2008). An Ontology-Driven Method for Hierarchical Mining of Temporal Patterns: Application to HIV Drug Resistance Research. *AMIA Symp*.

Ramoni, M., & Sebastiani, P. (2001). Robust learning with Missing Data. *Machine Learning*, 45, 147-170.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R B., & Richardson, W. S. (2004). Evidence based medicine: what it is and what it isn't. BMJ, 312 (7023), 71-2.

Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D., & Chute, C. G. (2008). Mayo clinic NLP system for patient smoking status identification. *J Am Med Inform Assoc,* 15(1), 25-8.

Schmidt, R., Montani, S., Bellazzi, R., Portinale, L., & Gierl, L. (2001). Case-based reasoning for medical knowledge-based systems. *Int J Med Inform*, 64(2-3), 355-367.

Spiegelhalter, D. J., & Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20, 579-605.

Stefanelli, M. (2004). Knowledge and process management in health care organizations. *Methods Inf Med*, 43(5), 525-35.

Van der Spek, R, & Spijkervet, A. (1997). Knowledge management: dealing intelligently with knowledge. In J. Liebowitz & L.C. Wilcox (Eds.), *Knowledge Management and its Integrative Elements*. CRC Press, Boca Raton, FL, 1997.

***Ricardo Bellazzi** is associate professor of medical informatics at the Dipartimento di Informatica e Sistemistica, University of Pavia, Italy. He teaches medical informatics and machine learning at the Faculty of Biomedical Engineering and bioinformatics at the Faculty of Biotechnology of the University of Pavia. He is a member of the board of the PhD in bioengineering and bioinformatics of the University of Pavia. Dr. Bellazzi is past-chairman of the IMIA working group of intelligent data analysis and data mining, program chair of the AIME 2007 conference and member of the program committee of several international conferences in medical informatics and artificial intelligence. He is member of the editorial board of Methods of Information in Medicine and of the Journal of Diabetes Science and Technology. He is affiliated with the American Medical Informatics Association and with the Italian Bioinformatics Society. His research interests are related to biomedical informatics, comprising data mining, IT-based management of chronic patients, mathematical modeling of biological systems, bioinformatics. Riccardo Bellazzi is author of more than 200 publications on peer-reviewed journals and international conferences.*