Chapter 2
D^3M Methodology

2.1 Introduction

On the basis of the discussions and retrospection on existing data mining methodologies and techniques in Chapter 1, this chapter presents an overall picture of domain driven data mining (D^3M). We focus on the high level of architecture and concepts of the D^3M methodology.

The goals of this chapter consist of the following aspects:

- An overview of the D^3M methodology;
- The main components of the D^3M methodology; and
- The methodological framework of the D^3M methodology.

Correspondingly, this chapter will introduce the following content.

- Section 2.2 presents a concept map of the D^3M methodology, which is composed of the structure, major components, and their relationships.
- In Section 2.3, we outline key methodological components consisting of the D^3M methodology. While some of these can be found in current data mining systems, they will be re-interpreted or revisited under the umbrella of D^3M. We also highlight some elements that are ignored or weakly addressed in classic methodologies and approaches.
- Finally, in Section 2.4, the theoretical underpinnings and process model of the D^3M methodology are introduced. This presents some high-level ideas of how D^3M is built on and what the D^3M process looks like.

2.2 D^3M Methodology Concept Map

Fig. 2.1 illustrates a high-level concept map of D^3M methodology. The concept map consists of the following layers from the outer most layer to the central core.

L. Cao et al., *Domain Driven Data Mining*, DOI 10.1007/978-1-4419-5737-5_2,

- Specific domain problems: In general, this can apply to any domain problems from retail to government to social network, from either a sector or specific business problem perspective. However, since D^3M mainly targets complex knowledge from complex data, we do not concern ourselves with those problems and businesses that have been or can be well-handled by existing data mining and knowledge discovery techniques.
- Fundamental research issues: Driven by specific domain problems and business needs, we here extract the main fundamental research issues emerging from them. We emphasize the following two aspects:
  - Infrastructure capability: We are concerned with key issues including: whether a data mining system can generally handle ubiquitous intelligence surrounding a domain problem or not, how it consolidates the relevant intelligence through what infrastructure, and what will be presented to support decision-making action-taking by end users.
  - Decision-support power: We are concerned with key issues reflecting and enhancing the decision-making power of identified knowledge and deliverables, in terms of key performances such as adaptability, dynamics, actionability, workability, operability, dependability, repeatability, trust, explainability, transferability, and usability.
- D^3M theoretical foundations: D^3M supporting techniques need foundational support from many relevant areas, from the information sciences to social sciences. In particular, we see a strong need to create new scientific fields, such as data sciences, web sciences and service sciences, targeting the establishment of a family of scientific foundations, techniques and tools for dealing with increasingly emergent complexities and challenges in the corresponding areas.
- D^3M supporting techniques: To engage and consolidate the fundamental issues surrounding domain driven actionable knowledge delivery, we need to develop corresponding techniques and tools for involving and utilizing ubiquitous intelligence, supporting knowledge representation and deliverables, catering for project and process management, and implementing decision-making pursuant to the findings.

2.3 D^3M Key Components

The D^3M methodology consists of the following key components.

- Constrained knowledge delivery environment
- Considering ubiquitous intelligence
- Cooperation between human and KDD systems
- Interactive and parallel KDD support
- Mining in-depth patterns
- Enhancing knowledge actionability

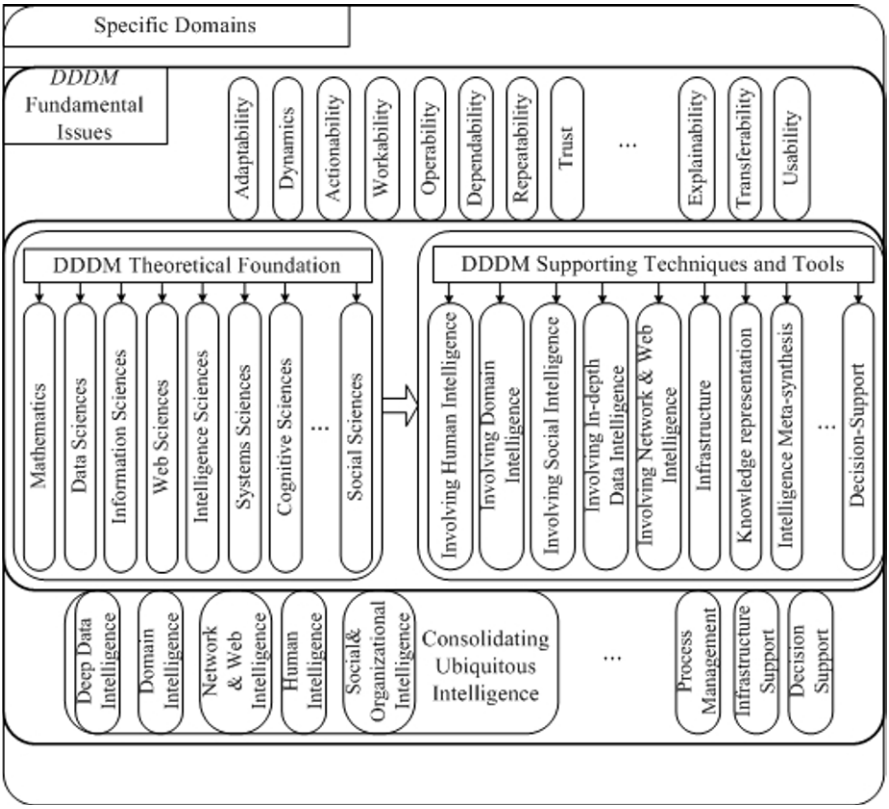

Fig. 2.1 D^3M concept map

- Reference model
- Qualitative research
- Closed-loop and iterative refinement

The nomination of the above key components is based on the support needed to cater for relevant factors in the domain and for actionable knowledge delivery. They have potential for re-shaping KDD processes, modeling and outcomes toward discovering and delivering knowledge that can be seamlessly used for decision-making action-taking, if they are appropriately considered and supported from technical, procedural and business perspectives.

2.3.1 Constrained Knowledge Delivery Environment

In human society, everyone is constrained by either social regulations or personal situations. Similarly, actionable knowledge is discovered in a constraint-based con-

text mixing environmental reality, expectations and constraints in the pattern mining process. Specifically, [33] list several types of constraints which play significant roles in a process which effectively discovers knowledge actionable to business. These include *domain constraints*, *data constraints*, *interestingness constraints*, and *deliverable constraints*.

Some major aspects of *domain constraints* include the domain and characteristics of a problem, domain terminology, specific business process, policies and regulations, particular user profiling and favorite deliverables. Potential matters to satisfy or react on domain constraints consist of building domain models, domain metadata, semantics [181] and ontologies [35, 107, 108], supporting human involvement, human-mining interaction, qualitative and quantitative hypotheses and conditions, merging with business processes and enterprise information infrastructure, fitting regulatory measures, conducting user profile analysis and modeling, etc. Relevant hot research areas include interactive mining, guided mining, and knowledge and human involvement.

Constraints on particular data, namely *data constraints*, may be embodied in terms of aspects such as very large volume, ill-structure, multimedia, diversity, high dimensionality, high frequency and density, distribution and privacy, dynamics and changes. Data constraints seriously affect the development of and performance requirements on data mining algorithms and systems, and constitute some grand challenges to real-world data mining. As a result, popular researches on data constraint-oriented issues are emerging such as stream data mining, link mining, multi-relational mining, structure-based mining, privacy mining, multimedia mining, temporal mining, dynamic data mining, and change and difference mining.

What makes a rule, pattern and finding more interesting than the other? This involves *interestingness constraints*. In the real world, simply emphasizing technical interestingness, such as objective statistical measures of *validity* and *surprise*, is not adequate. Social and economic interestingness (we refer to *Business Interestingness*) such as benefit-cost for operating and implementing the identified knowledge, which embodies user preferences and domain knowledge, should be considered in assessing whether a pattern is actionable or not. Business interestingness would be instantiated into specific social and economic measures in terms of the problem domain. For instance, *profit*, *return*, *return on investment*, or *cost-benefit ratio* are usually used by traders to measure the economic performance of a trading rule, and to judge whether a trading rule is in their interest or not.

Furthermore, the delivery of an interesting pattern needs to consider operationalization factors, which consequently involves *deliverable constraints*. Deliverables such as business rules, processes, information flow, presentation, etc. may need to be integrated into the domain environment.In addition, many other realistic issues must be considered. For instance, a software infrastructure may be needed to support the full lifecycle of data mining; the infrastructure needs to integrate with the existing enterprise information systems and workflow; parallel KDD may be involved with parallel support on multiple sources, parallel I/O, parallel algorithms, and memory storage; visualization, privacy and security should receive much-deserved attention; false alarming needs to be minimized.

In practice, constraints on real-world data mining may also be embodied in many other aspects, such as the frequency and density of data, and the scalability and efficiency of algorithms, which have to be facilitated. These plus the above-mentioned constraints form the conditions and environment of data mining from process, operational, functional and nonfunctional aspects. These ubiquitous constraints form a constraint-based context for actionable knowledge discovery and delivery. All the above constraints must, to varying degrees, be considered in the relevant phases of real-world data mining. Thus it is also called constraint-based data mining [17, 115].

In summary, actionable knowledge discovery and delivery will not be a trivial task and should be put into a constraint-based environment. On the other hand, tricks may not only include how to find a right pattern with a right algorithm in a right manner, but may also involve a suitable process-centric and domain-oriented process management methodology and infrastructure support.

2.3.2 Considering Ubiquitous Intelligence

Traditionally, data mining only pays attention to and relies on data to disclose possible stories wrapping a problem. We call such finding *data intelligence* disclosed from data. Driven by this strategic idea, data mining focuses on developing methodologies and methods in terms of data-centered aspects, particularly the following issues:

- Data type such as numeric, categorical, XML, multimedia, composite
- Data timing such as temporal, time-series, sequential and real-time data
- Data spacing such as spatial and temporal-spatial
- Data speed such as data stream, rare and loose occurrences
- Data frequency such as high frequency data
- Data dimension such as multi-dimensional data
- Data relation such as multi-relational data, linkage, casual data
- Data quality such as missing, noisy, uncertain and incomplete data

On the other hand, domain factors consisting of qualitative and quantitative aspects hide intelligence for problem-solving. Both qualitative and quantitative intelligence is instantiated in terms of domain knowledge, constraints, actors/domain experts and environment. They are further instantiated into specific bodies. For instance, constraints may include domain constraints, data constraints, interestingness constraints, deployment constraints and deliverable constraints. To deal with constraints, various strategies and methods may be undertaken; for instance, interestingness constraints are modeled in terms of interestingness measures and factors, such as objective interestingness and subjective interestingness.

In a summary, we list ubiquitous intelligence hidden and explicitly existing in domain problems in terms of the following major aspects.

(1) Domain knowledge aspect

- Including domain knowledge, background and prior information,

(2) Human aspect

- Referring to direct or indirect involvement of humans, imaginary thinking, brainstorming, etc.
- Empirical knowledge
- Belief, request, expectation, etc.

(3) Constraint aspect

- Including constraints from system, business process, data, knowledge, deployment, etc.
- Privacy
- Security

(4) Organizational aspect

- Organizational factors
- Business process, workflow, project management
- Business rules, law, trust

(5) Environmental aspect

- Surrounding business processes, workflow
- Linkage systems
- Surrounding situations and scenarios

(6) Evaluation aspect

- Technical interestingness corresponding to a specific approach
- Profit, benefit, return, etc.
- Cost, risk, etc.
- Business expectation and interestingness

(7) Deliverable and deployment aspect

- Delivery manners
- Embedding into business system and process

Correspondingly, a series of issues needs to be studied in order to involve and utilize such ubiquitous intelligence in the actionable knowledge delivery system and process. The involvement of ubiquitous intelligence forms a key element and characteristic of domain driven data mining towards domain driven actionable knowledge delivery. For instance, the following are some such tasks for involving and utilizing ubiquitous intelligence.

- Definition of ubiquitous intelligence
- Representation of domain knowledge
- Ontological and semantic representation of ubiquitous intelligence
- Ubiquitous intelligence transformation between business and data mining
- Human role, modeling and interaction

- Theoretical problems in involving ubiquitous intelligence in KDD
- Metasynthesis of ubiquitous intelligence in knowledge discovery
- Human-cooperated data mining
- Constraint-based data mining
- Privacy, sensitivity and security in data mining
- Open environment in data mining
- In-depth data intelligence
- Knowledge actionability
- Objective and subjective interestingness
- Gap resolution between statistical significance and business expectation
- Domain-oriented knowledge discovery process model
- Profit, benefit/cost, risk, impact of mined patterns

2.3.3 Cooperation between Human and KDD Systems

The real-life requirements for discovering actionable knowledge in a constraint-based environment determine that real-world data mining is more likely to follow man-machine-cooperated mode, namely human-mining-cooperated rather than automated. Human involvement is embodied through the cooperation between humans (including users and business analysts, mainly domain experts) and a data mining system. This is because of the complementation between human qualitative intelligence such as domain knowledge and field supervision, and the quantitative intelligence of KDD systems like computational capabilities. Therefore, real-world complex data mining presents as a human-mining-cooperated interactive knowledge discovery and delivery process.

The role of humans in AKD may be embodied in the full period of data mining from business and data understanding, problem definition, data integration and sampling, feature selection, hypothesis proposal, business modeling and learning to the evaluation, refinement and interpretation of algorithms and resulting outcomes. For instance, the experience, meta-knowledge and imaginary thinking of domain experts can guide or assist with the selection of features and models, add business factors into the modeling, create high quality hypotheses, design interestingness measures by injecting business concerns, and quickly evaluate mining results. This assistance can largely improve the effectiveness and efficiency of identifying actionable knowledge.

In existing data mining applications, humans often take part in feature selection and results evaluation. In fact, a human may play a critical role in a specific stage or during the full stages of data mining on demand. In some cases, for example, mining a complex social community, a human is an essential constituent or the center of a data mining system. The complexity of discovering actionable knowledge in a constraint-based environment determines to what extent humans must be involved. As a result, the human-mining cooperation could be, to various degrees, human-centered mining, human-guided mining, or human-supported or -assisted mining.

To support human involvement, human-mining interaction, otherwise known as interactive mining [6, 9], is absolutely necessary. Interaction often takes explicit forms, for instance, setting up direct interaction interfaces to fine tune parameters. Interaction interfaces may take various forms as well, such as visual interfaces, virtual reality, multi-modal, mobile agents, etc. On the other hand, it could also go through implicit mechanisms, for example accessing a knowledge base or communicating with a user assistant agent. Interaction communication may be message-based, model-based, or event-based. Interaction quality relies on performance such as user-friendliness, flexibility, run-time capability, representability and even understandability.

2.3.4 Interactive and Parallel KDD Support

To support domain driven data mining, it is important to develop interactive mining support for involving domain experts, and human-mining interaction. Interactive facilities are also useful for evaluating data mining findings by involving domain experts in a closed-loop manner. On the other hand, parallel mining support is often necessary for dealing with concurrent applications, distributed and multiple data sources. In cases with intensive computation requests, parallel mining can greatly upgrade the real-world data mining performance.

For interactive mining support, intelligent agents [214] and service-oriented computing [186] are good technologies. They can create flexible, business-friendly and user-oriented human-mining interaction through building facilities for user modeling, user knowledge acquisition, domain knowledge modeling, personalized user services and recommendation, run-time support, and mediation and management of user roles, interaction, security and cooperation.

Parallel KDD [122] is good at parallel computing and management support for dealing with multiple sources, parallel I/O, parallel algorithms and memory storage. For instance, to tackle cross-organization transactions, we can design efficient parallel KDD computing and systems to wrap the data mining algorithms. This can be through developing parallel genetic algorithms and proper processor-cache memory techniques. Multiple master-client process-based genetic algorithms and caching techniques can be tested on different CPU and memory configurations to find good parallel computing strategies.

The facilities supporting interactive and parallel mining can largely improve the performance of real-world data mining in aspects such as human-mining interaction and cooperation, user modeling, domain knowledge acquisition and involvement in KDD, and reducing computation complexity. They are essential parts of the next-generation KDD infrastructure for dealing with complex enterprise data for complex knowledge.

Based on our experience in building agent service-based stock trading and mining system F-Trade [1], agent service and ontological engineering techniques can be used for building interactive and parallel mining facilities. User agents, knowledge management agents, ontology services [35] and run-time interfaces can be built to support interaction with users, take users' requests and to manage information from users in terms of ontologies. Ontologies can represent domain knowledge and user preferences, and further map them to a data mining domain to support a seamless mapping crossing user/business terminology, data mining ontologies, and underlying data source fields. Subsequently, a universal and transparent one-stop portal may be feasible for domain experts and business users to assist with training, supervising and tuning feature selection, modeling and refinement, as well as evaluating the outcomes. This can help avoid the requirements imposed on domain experts and business users for learning technical detail and jargon, and enable them to concentrate on their familiar environment and language, as well as their interests and responsibilities.

2.3.5 Mining In-Depth Patterns

In-Depth Patterns indicate patterns that

- uncover not only appearance dynamics and rules but also inside driving forces; for instance, in stock data mining, not only price movement trends but also the interior driving forces of such movements,
- reflect not only technical concerns but also business expectations, and
- disclose not only generic knowledge but also something that can support straightforward decision-making actions.

Greater effort is essential to uncover in-depth patterns in data. 'In-depth patterns' (or 'deep patterns') are not straightforward such as frequency-based, but can only be discovered through more powerful models following thorough data and business understanding and effectively involving domain intelligence or expert guidance. An example is to mine for insider trading patterns in capital markets. Without deep understanding of the business and data, a naive approach is to analyze the price movement change in data partitions of pre-event, event and post-event. A deeper pattern analysis on such price difference analysis may be considered by involving domain factors such as considering market or limit orders, market impact, and checking the performance of *potential abnormal return*, *liquidity*, *volatility* and *correlation*.

However, in general, the modeling of data mining is only concerned with the technical significance. Technical significance is usually defined in a straightforward manner by reflecting the significance of the findings in terms of the utilized techniques. Consequently, pattern interestingness is measured in terms of such technical metrics. When they are delivered to business people, business analysts either cannot understand them very well or cannot justify their significance from the business

[1] www.F-TRADE.info

end. In many cases, business people may just find them unconvincing, unjustifiable, unacceptable, impractical and inoperable. Such situations have hindered the deployment and adoption of data mining in real applications. Therefore it is essentially critical to develop pattern interestingness catering for business concerns, preferences and expectations. The resulting patterns (P) satisfy both technical and business interestingness ($\forall P, x.tech\_int(P) \wedge x.biz\_int(P) \rightarrow x.act(P)$). As a result, it is more likely that they reflect the genuine needs of business and can support smarter and more effective decision-making.

In-depth pattern mining needs to check both technical ($tech\_int()$) and business ($biz\_int()$) interestingness in a constraint-based environment. Technically, this could be through enhancing or generating more effective interestingness measures [155]. For instance, a series of interestingness measures have been proposed to evaluate associations more properly in association rule mining. It could also be through developing alternative models by involving domain factors and business interestingness for discovering patterns of business interest. Some other solutions include further mining actionable patterns on the initially discovered pattern set. Additionally, techniques can be developed to deeply understand, analyze, select and refine the target data set and feature set in order to find in-depth patterns.

More attention should be paid to business requirements, objectives, domain knowledge and qualitative intelligence of domain experts for their impact on mining deep patterns. Consequently, business interestingness needs to be developed to reflect such business reality, user preferences, needs and expectations. This can be through selecting and adding business features, involving domain knowledge in modeling pattern significance and impact, supporting interaction with users, tuning the parameters and data set by domain experts, optimizing models and parameters, adding organizational factors into technical interestingness measures or building domain-specific business measures, improving the results evaluation mechanism through embedding domain knowledge and expert guidance.

2.3.6 Enhancing Knowledge Actionability

Patterns which are interesting to data miners may not necessarily lead to business benefits if deployed. For instance, a large number of association rules are often found, while most of them are workable in business. These rules are generic patterns satisfying technical interestingness, while they are not measured and evaluated in the business sense.

In traditional data mining, or when data mining methods are used in applications, a common scenario is that many mined patterns are more interesting to data miners than to business people. Business people have difficulties in understanding and taking them over for one of the following reasons:

- they reflect commonsense in business,
- they are uninterpretable, thus incomprehensible by following general business thinking,

- they are too many and are indistinguishable, not indicating which are truly useful and more important to business,
- their significance is not justifiable from a business perspective, and
- their presentation is usually far from that of the legacy or existing business operation systems, and thus they cannot be integrated into the systems directly or there is no advice on how they are to be combined into the business working system.

Any findings falling into one of the above scenarios would have difficulty in supporting business decision-making.

To boost the actionable capability of identified patterns, techniques for further actionability enhancement are necessary for generating actionable patterns useful to business. This may be conducted from several perspectives.

First, the appropriate measurement of pattern actionability needs to be defined. Technical and business interestingness measures should be defined and satisfied from both objective and subjective perspectives. For those generic patterns identified based on technical measures only, business interest and performance need to be further checked so that business requirements and user preferences can be given proper consideration.

Second, actionable patterns in many cases can be created through rule reduction, model refinement or parameter tuning by optimizing and filtering generic patterns. If this is the case, actionable patterns are a revised optimal version of generic patterns, which capture deeper characteristics and understanding of the business, consequently present as in-depth or optimized patterns.

Third, pattern actionability is also reflected in the delivery manner. Some forms of patterns are easily understandable by business people, while others are elusive. To support business operation, it is necessary for the deliverables to be converted into forms that can be easily or even seamlessly fed into business rules, processes and operation systems. For this, one option is to convert patterns into business rules following the business terminology system and the existing business rule specifications.

Finally, more direct efforts are necessary to enhance the KDD modeling and mining process targeting the output of actionable knowledge directly discovered from data set by considering the ubiquitous intelligence surrounding the problem.

2.3.7 Reference Model

Reference models such as those in CRISP-DM [2] are very helpful for guiding and managing the knowledge discovery process. It is recommended that these reference models be respected in domain driven actionable knowledge delivery. However, actions and entities for domain driven data mining, such as considering constraints, and integrating domain knowledge, should be paid more attention in the corresponding modeling and procedures.

[2] http://www.crisp-dm.org/

On the other hand, new reference models are essential for supporting components such as in-depth modeling and actionability enhancement.

For instance, Fig. 2.2 illustrates the reference model for actionability enhancement in domain driven data mining.

2.3.8 Qualitative Research

In developing real-world data mining applications, qualitative research methodology [88] is very helpful for capturing business requirements, constraints, requests from organization and management, risk and contingency plans, expected representation of the deliverables, etc. For instance, a questionnaire can assist with uncovering human concerns and business specific requests.

During the process of conducting real-world data mining, a questionnaire can be used to collect feedback on business requirements, constraints, interest, expectations, and requests from domain experts and business users. The information collected may involve aspects, factors, elements, measures and critical values about the organization, operation, management, business processes, workflow, business logics, existing business rules, risk and contingency plans, expected representation of the deliverables, and so on.

It is recommended that questionnaires be designed for every procedure in the domain driven actionable knowledge delivery process. Analytical and contingency reports are then developed for every procedure. Follow-up interviews and discussions may be necessary. Data and records must be collected, analyzed, clarified and finally documented in a knowledge management system as the evidence and guidance for business and data understanding, feature selection, parameter tuning, result evaluation, and deliverable presentation. The findings will also guide the design of user interfaces and user modeling, as well as working mechanisms for involving domain knowledge and experts' roles in the process.

In addition, reference models are helpful for guiding the implementation and for managing the actionable knowledge discovery and delivery process and project. For instance, Fig. 2.2 illustrates the reference model for actionability enhancement.

2.3.9 Closed-Loop and Iterative Refinement

Actionable knowledge discovery in a constraint-based context is more likely to be a closed-loop rather than open process. A closed-loop process indicates that the outputs of data mining are fed back to change relevant parameters or factors in particular stages. The feedback and change effect may be embodied through analyzing and adjusting the relationships between outputs and particular parameters and factors, and eventually tuning the parameters and factors accordingly.

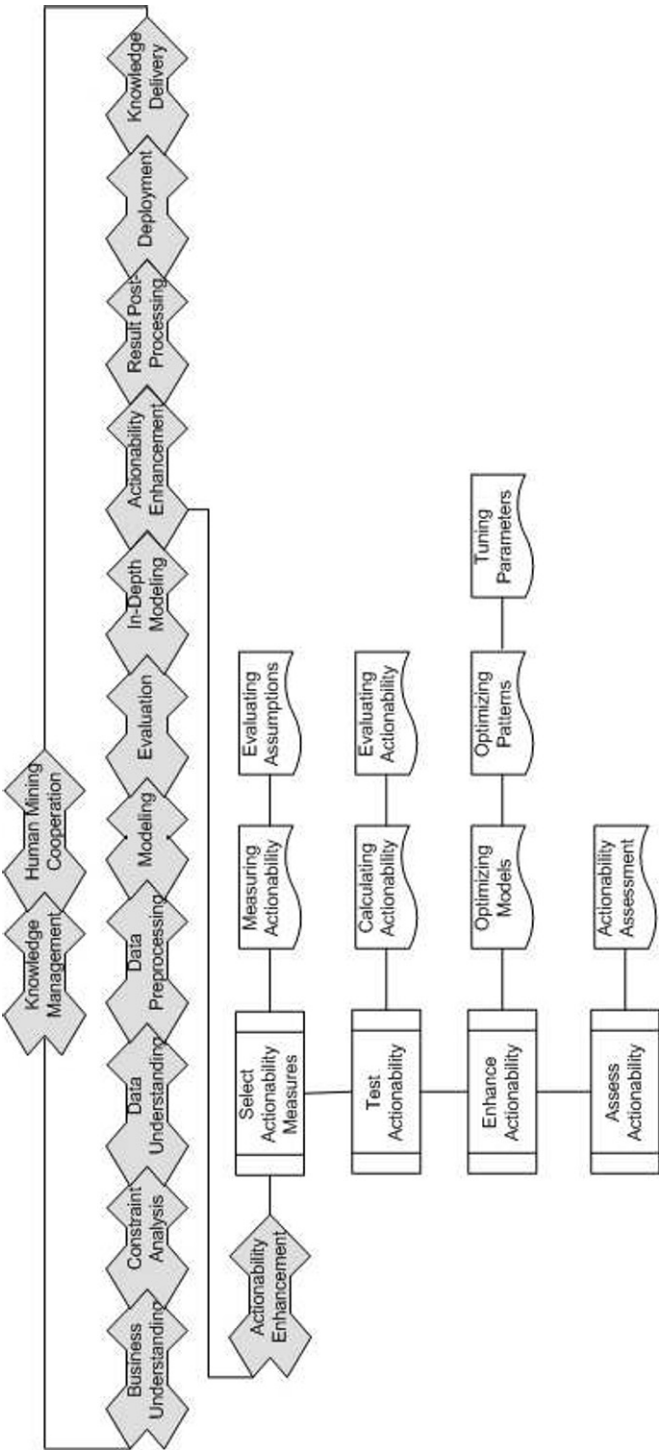

Fig. 2.2 Knowledge actionability enhancement in D^3M

The real-world data mining process is likely iterative because the evaluation and refinement of features, models and outcomes cannot be completed in a one-off way. Iterative interaction may be conducted to varying stages such as sampling, hypothesis, feature selection, modeling, evaluation and/or interpretation before reaching the final stage of knowledge and decision-support report delivery.

Consequently, real-world data mining cannot be undertaken just using an algorithm. Rather, It is necessary to build a proper data mining infrastructure to discover actionable knowledge from constraint-based scenarios in a closed-loop iterative manner. To this end, an agent-based data mining infrastructure [23, 49] can provide good facilities for interaction and message passing amongst different modules, and support user modeling and user agent interaction toward autonomous, semi-autonomous or human-mining-cooperated problem-solving.

2.4 D^3M Methodological Framework

2.4.1 Theoretical Underpinnings

Research and development, and the effective use of D^3M involves multiple disciplines. Its theoretical underpinnings involve analytical, computational and social sciences. We interpret the theoretical infrastructure for D^3M from the perspectives of methodological support, fundamental technologies, and supporting techniques and tools.

From the methodological support perspective, D^3M needs the support of multiple fields, including the information sciences, intelligence sciences, system sciences, cognitive sciences, organizational sciences, and social sciences. *Information and intelligence sciences* provide support for intelligent information processing and systems. *System sciences* furnish methodologies and techniques for domain factor modeling and simulation, closed-loop system design and analysis, and feedback mechanism design. *Cognitive sciences* incorporate principles and methods for understanding human qualitative intelligence such as imaginary thinking, empirical knowledge, belief and intention which is important for understanding and analyzing complex domain problems. *Social sciences* supply foundations for conceiving organizational and social factors and business processes surrounding problem domain.

In particular, we highlight the need to involve a few new scientific fields: data sciences, knowledge sciences, web and network sciences, service sciences, and complexity sciences. We need them because they are critical for handling the increasingly emergent issues and complexities, as well as increasingly for the breadth and depth of their involvement in our business, data and environment.

- Data sciences, on top of the current efforts on data engineering, offer a systematic and fundamental understanding and exploration of the ever-increasing data, which certainly forms one of the essential foundations for deep data understanding, exploration and analysis in D^3M;

- Knowledge sciences, on top of the current efforts on knowledge engineering, entails the systematic and fundamental understanding and exploration of the ever-increasing atock of knowledge, from both prior, empirical and human knowledge, to emerging knowledge from discovery, interaction and computing; it certainly forms one of the essential foundations for knowledge representation, transformation, reasoning, emergence, transferring and use in D^3M;
- Web sciences and network sciences, as an attempt to understand and explore the ever-growing phenomenon of the World Wide Web and increasingly emerging networks, contribute to the understanding, identification, facilitation and involvement of networks and networking in D^3M.
- Service sciences, which are a melding of technologies for an understanding of business processes and organizations, and services systems, contribute to the infrastructure establishment and knowledge delivery from data mining systems to business operations.
- Complexity sciences is a discipline studying complex systems, which can provide methodologies and techniques for involving and managing surrounding factors in understanding and analyzing complex data.

In addition, areas and knowledge bodies such as optimization theory, risk analysis, economics and finance are also important for understanding and measuring business impact and the interestingness of identified patterns.

Besides the mainstream KDD techniques, fundamental technologies needed also involve user modeling, formal methods, logics, representation, knowledge engineering, ontological engineering, semantic web, and cognitive engineering. The modeling of pattern impact and business interestingness may refer to the relevant technologies such as statistical significance, impact analysis, benefit-cost analysis, risk management and analysis, and performance measurements in economics and finance. To understand domain-specific ubiquitous intelligence, and the evolution of user modeling and group thinking, we refer to techniques and tools in fields like systems simulation, communication, artificial social system, open complex systems, swarm intelligence, social network analysis, reasoning and learning. The deliverable presentation may involve means in knowledge representation, business rule presentation, visualization and graph theory.

2.4.2 Process Model

The existing data mining methodology, for instance CRISP-DM, generally supports autonomous pattern discovery from data. By contrast, the idea of domain driven knowledge discovery is to involve ubiquitous intelligence into data mining. The D^3M highlights a process that discovers in-depth patterns from a constraint-based environment with the involvement of domain experts and their knowledge. Its objective is to maximally accommodate both naive users as well as experienced analysts, and to satisfy business goals. The patterns discovered are expected to be integrated into business systems and to be aligned with existing business rules. To make do-

main driven data mining effective, user guides and intelligent human-machine interaction interfaces are essential through incorporating both human qualitative intelligence and machine quantitative intelligence. In addition, appropriate mechanisms are required for dealing with multiform constraints and domain knowledge.

The main functional components of the D^3M process model are shown in Fig. 2.3, in which we highlight those processes specific to D^3M in thick boxes. The lifecycle of the D^3M process is as follows, but be aware that the sequence is not rigid, some phases may be bypassed or moved back and forth in dealing with a real problem. Every step of the D^3M process may involve ubiquitous intelligence and interaction with business users and/or domain experts.

P_1 . Problem understanding (to identify and define the problems, including its scope and challenges etc.);
P_2 . Constraints analysis (to identify constraints surround the above problems, from data, domain, interestingness and delivery perspectives);
P_3 . Definition of analytical objectives, and feature construction (to define the goal of data mining, and accordingly features selected or constructed to achieve the objectives);
P_4 . Data preprocessing (data extraction, transformation and loading, in particular, the data preparation such as processing missing and privacy data);
P_5 . Method selection and modeling (to select the appropriate models and methods for achieving the above objectives); or
P'_5 . In-depth modeling (to apply deep modeling either by using more effective models disclosing the very core of the problem, or by using multi-step mining or combined mining);
P_6 . Initial generic results analysis and evaluation (to analyze/assess the initial findings);
P_7 . It is quite possible that each phase from P_1 may be iteratively reviewed through analyzing constraints and interaction with domain experts in a back-and-forth manner; or
P'_7 . In-depth mining on the initial generic results where applicable;
P_8 . Actionability measurement and enhancement (to check the interestingness from both technical and business perspectives, and to enhance the performance by applying more effective methods etc.);
P_9 . Back and forth between P_7 and P_8;
P_{10}. Results post-processing (to post-analyse or post-mine the initial resulting data);
P_{11}. Reviewing phases from P_1 may be required;
P_{12}. Deployment (to deploy the results into business lines);
P_{13}. Knowledge delivery and report synthesis for smart decision making (to synthesize the eventual findings into decision-making report to be delivered to business people).

The D^3M process highlights the following aspects that are critical for the success of a data mining in the real world. They are

- context and environment (including the factors from data, domain, organizational social aspects),

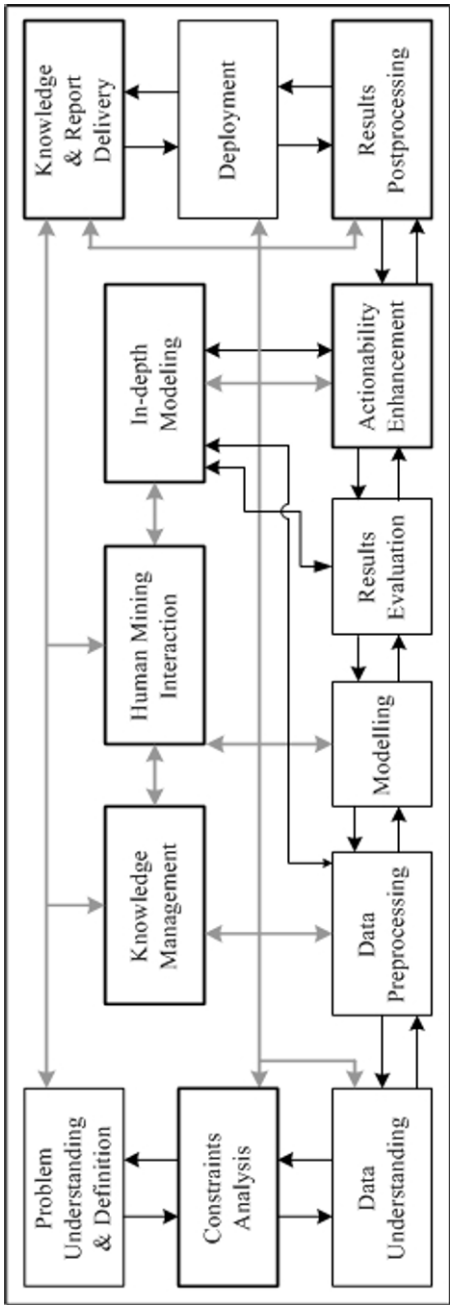

Fig. 2.3 D^3M process model

- constraints (including constraints from data, domain, interestingness and deliverable perspectives),
- domain knowledge (including domain expert knowledge and knowledge from business systems),
- organizational and social factors (including factors from aspects such as organizational rules, relationships, social networks),
- human qualitative intelligence and roles (including human empirical knowledge, imaginary thinking etc.),
- user preferences (including user expectations and needs),
- interaction and interfaces (including the tools and interfaces for interaction between a user and data mining systems),
- cooperation between humans and the data mining system (reflecting the allocation of tasks between human and a data mining system),
- in-depth pattern mining (to discovery deep patterns reflecting geniune interest of business owners and decision-makers),
- parallel support (including the support for mining patterns in parallel),
- business impact and interestingness (reflecting business concerns from objective and subjective perspectives),
- knowledge actionability (including concerns from technical and business aspects),
- feedback (from business modelers, as well as business owners and decision-making), and
- iterative refinement (as needed, back to corresponding steps).

These aspects are consistent with the key components in D^3M, as we discussed in Section 2.3.

2.4.3 D^3M Evaluation System

The D^3M evaluation system caters for significance and interestingness ($Int(p)$) of a pattern (p) from both technical and business perspectives. $Int(p)$ is measured in terms of *technical interestingness* ($t_i(p)$) and *business interestingness* ($b_i(p)$) [41].

$$Int(p) = I(t_i(p), b_i(p)) \tag{2.1}$$

where $I(.)$ is the function for aggregating the contributions of all particular aspects of interestingness.

Further, $Int(p)$ is described in terms of *objective* (o) and *subjective* (s) factors from both *technical* (t) and *business* (b) perspectives.

$$Int(p) = I(t_o(), t_s(), b_o(), b_s()) \tag{2.2}$$

where $t_o()$ is objective technical interestingness, $t_s()$ is subjective technical interestingness, $b_o()$ is objective business interestingness, and $b_s()$ is subjective business interestingness.

We say p is truly *actionable* (i.e., $\widetilde{p}$) to both academia and business if it satisfies the following condition:

$$Int(p) = t_o(\mathbf{x}, \widetilde{p}) \wedge t_s(\mathbf{x}, \widetilde{p}) \wedge b_o(\mathbf{x}, \widetilde{p}) \wedge b_s(\mathbf{x}, \widetilde{p}) \tag{2.3}$$

where '$\wedge$' indicates the interestingness 'aggregation'.

In general, $t_o()$, $t_s()$, $b_o()$ and $b_s()$ of practical applications can be regarded as independent of each other. With their normalization (expressed by $\hat{}$), we can get:

$$\begin{aligned} Int(p) &\rightarrow \hat{I}(\hat{t_o}(), \hat{t_s}(), \hat{b_o}(), \hat{b_s}()) \\ &= \alpha\hat{t_o}() + \beta\hat{t_s}() + \gamma\hat{b_o}() + \delta\hat{b_s}() \end{aligned} \tag{2.4}$$

The AKD optimization problem in D^3M can be expressed as follows:

$$\begin{aligned} AKD^{e,\tau,m\in M} &\longrightarrow O_{p\in P}(Int(p)) \\ &\rightarrow O(\alpha\hat{t_o}()) + O(\beta\hat{t_s}()) + \\ &\quad O(\gamma\hat{b_o}()) + O(\delta\hat{b_s}()) \end{aligned} \tag{2.5}$$

The *actionability* of a pattern p is measured by $act(p)$:

$$\begin{aligned} act(p) &= O_{p\in P}(Int(p)) \\ \rightarrow O(\alpha\hat{t_o}(p)) &+ O(\beta\hat{t_s}(p)) + \\ O(\gamma\hat{b_o}(p)) &+ O(\delta\hat{b_s}(p)) \\ \rightarrow t_o^{act} + t_s^{act} &+ b_o^{act} + b_s^{act} \\ \rightarrow t_i^{act} &+ b_i^{act} \end{aligned} \tag{2.6}$$

where t_o^{act}, t_s^{act}, b_o^{act} and b_s^{act} measure the respective actionable performance in terms of each aspect.

Due to the inconsistency often existing at different aspects, we often find that the identified patterns only fit in one of the following sub-sets:

$$\begin{aligned} Int(p) \rightarrow \{&\{t_i^{act}, b_i^{act}\}, \{\neg t_i^{act}, b_i^{act}\}, \\ &\{t_i^{act}, \neg b_i^{act}\}, \{\neg t_i^{act}, \neg b_i^{act}\}\} \end{aligned} \tag{2.7}$$

where '$\neg$' indicates the corresponding element is not satisfactory. Ideally, we look for actionable patterns p that can satisfy the following condition:

IF

$$\begin{aligned} \forall p \in \widetilde{P}, \exists \mathbf{x} : t_o(\mathbf{x}, p) \wedge t_s(\mathbf{x}, p) \wedge b_o(\mathbf{x}, p) \\ \wedge b_s(\mathbf{x}, p) \rightarrow act(p) \end{aligned} \tag{2.8}$$

THEN:

$$p \rightarrow \widetilde{p}. \tag{2.9}$$

In the real-world data mining, it is often very challenging to find the most actionable patterns that are associated with both 'optimal' t_i^{act} and 'optimal' b_i^{act}. Clearly, D^3M favors patterns confirming the relationship $\{t_i^{act}, b_i^{act}\}$. There is a need to deal with possible conflict and uncertainty amongst respective interestingness elements. Technically, there is an opportunity to develop techniques to balance and combine all types of interestingness metrics to generate uniform, balanced and interpretable mechanisms for measuring knowledge deliverability. Under sophisticated situations, domain experts from both computation and business areas need to interact with each other, ideally through an m-space with intelligence meta-synthesis facilities such as letting one run models with quantitative outcomes to support discussions with other experts. If $t_i()$ and $b_i()$ are inconsistent, experts argue and compromise with each other through m-interactions in the m-space, like what happens in a board meeting, but with substantial online resources, models and services.

2.4.4 D^3M Delivery System

Well experienced data mining professionals attribute the weak executable capability of existing data mining findings to the lack of proper tools and mechanisms for implementing the deployment of the resulting models and algorithms ideally by business users rather than analysts. In fact, the barrier and gap comes from the weak, if not none, capability of existing data mining deployment systems, existing in presentation, deliverable and execution aspects. They form the D^3M delivery system, which is much beyond the identified patterns and models themselves.

- Presentation: studies how to present data mining findings that can be easily recognized, interpreted and taken over as they need;
- Deliverable: studies how to deliver data mining findings and systems to business users so that the findings are handy to be re-formated, transformed, or cut and pasted into their own business systems and presentation on demand, and the systems can be understood and taken over by end users; and
- Execution: studies how to integrate data mining findings and systems into production systems, and how the findings to be executed easily and seamlessly in an operational environment.

Supporting techniques need to be developed for AKD presentation, deliverable and execution. For instance, the following lists some of techniques.

- Presentation: typical tools such as visualization techniques are essentially helpful, visual mining could support the whole data mining process in a visual manner;
- Deliverable: business rules are widely used in business organizations, one of the methods for delivering patterns is to convert them into business rules; for this

we can develop a tool with underlying ontologies and semantics to support the transfer from pattern to business rules;
- Execution: tools to make deliverables executable in an organization's environment need to be developed, one of the efforts is to generate PMML to convert models to executables so that the models can be integrated into production systems, and run on a regular basis to provide cases for business management.

2.5 Summary

This chapter has presented the basic concept and overall picture of domain driven data mining methodology. We have discussed the concept map of D^3M, key methodological components consisting of D^3M, and its theoretical underpinnings.

Conclusions from this chapter consist of:

- Driven by challenges and complexities from specific domain problems, domain driven data mining provides a systematic solution and guideline from identifying fundamental research issues, to developing corresponding techniques and tools;
- Major methodological components of D^3M reflect the corresponding problem-solving solutions to tackle key challenges and issues existing in traditional data mining;
- The identified key components within D^3M exhibit tremendous new opportunities for us to explore in the data mining area, by engaging knowledge and lessons from many other disciplines, including traditional ones such as information sciences, as well as new fields such as data sciences, web sciences, service sciences, knowledge sciences and complexity sciences;
- It is far from mature as a new research area, and there are many great opportunities and prospects for us to further investigate in domain driven data mining.

In Chapter 3, we specifically explore ubiquitous intelligence surrounding and contributing to domain driven data mining.

http://www.springer.com/978-1-4419-5736-8

Domain Driven Data Mining
Cao, L.; Yu, P.S.; Zhang, C.; Zhao, Y.
2010, XVI, 248 p., Hardcover
ISBN: 978-1-4419-5736-8