CRISP-DM: Towards a Standard Process Model for Data Mining

Rüdiger Wirth
DaimlerChrysler Research & Technology FT3/KL
PO BOX 2360 89013 Ulm, Germany
ruediger.wirth@daimlerchrysler.com

Jochen Hipp
Wilhelm-Schickard-Institute, University of Tübingen
Sand 13, 72076 Tübingen, Germany
jochen.hipp@informatik.uni-tuebingen.de

Abstract

The CRISP-DM (CRoss Industry Standard Process for Data Mining) project proposed a comprehensive process model for carrying out data mining projects. The process model is independent of both the industry sector and the technology used. In this paper we argue in favor of a standard process model for data mining and report some experiences with the CRISP-DM process model in practice.

We applied and tested the CRISP-DM methodology in a response modeling application project. The final goal of the project was to specify a process which can be reliably and efficiently repeated by different people and adapted to different situations. The initial projects were performed by experienced data mining people; future projects are to be performed by people with lower technical skills and with very little time to experiment with different approaches. It turned out, that the CRISP-DM methodology with its distinction of generic and specialized process models provides both the structure and the flexibility necessary to suit the needs of both groups.

The generic CRISP-DM process model is useful for planning, communication within and outside the project team, and documentation. The generic check-lists are helpful even for experienced people. The generic process model provides an excellent foundation for developing a specialized process model which prescribes the steps to be taken in detail and which gives practical advice for all these steps.

1 Introduction

Data mining is a creative process which requires a number of different skills and knowledge. Currently there is no standard framework in which to carry out data mining projects. This means that the success or failure of a data mining project is highly dependent on the particular person or team carrying it out and successful practice can not necessarily be repeated across the enterprise. Data mining needs a standard approach which will help translate business problems into data mining tasks, suggest appropriate data transformations and data mining techniques, and provide means for evaluating the effectiveness of the results and documenting the experience.

The CRISP-DM (CRoss Industry Standard Process for Data Mining) project[1] addressed parts of these problems by defining a process model which provides a framework for carrying out data

[1] The CRISP-DM process model is being developed by a consortium of leading data mining users and suppliers: DaimlerChrysler AG, SPSS, NCR, and OHRA. The project was partly sponsored by the European Commission under the ESPRIT program (Project number 24959)

mining projects which is independent of both the industry sector and the technology used. The CRISP-DM process model aims to make large data mining projects, less costly, more reliable, more repeatable, more manageable, and faster.

In this paper, we will argue that a standard process model will be beneficial for the data mining industry and present some practical experiences with the methodology.

2 Why the Data Mining Industry needs a Standard Process Model

The data mining industry is currently at the chasm (Moore, 1991) between early market and main stream market (Agrawal, 1999). Its commercial success is still not guaranteed. If the early adopters fail with their data mining projects, they will not blame their own incompetence in using data mining properly but assert that data mining does not work.

In the market, there is still to some extent the expectation that data mining is a push-button technology. However, this is not true, as most practitioners of data mining know. Data Mining is a complex process requiring various tools and different people. The success of a data mining project depends on the proper mix of good tools and skilled analysts. Furthermore, it requires a sound methodology and effective project management. A process model can help to understand and manage the interactions along this complex process.

For the market, there will be many benefits if a common process model is accepted. The model can serve as a common reference point to discuss data mining and will increase the understanding of crucial data mining issues by all participants, especially at the customers' side. But most importantly, it will create the impression that data mining is an established engineering practice. Customers will feel more comfortable if they are told a similar story by different tool or service providers. On the more practical side, customers can get more reasonable expectations as to how the project will proceed and what to expect at the end. Dealing with tool and service providers, it will be much easier for them to compare different offers to pick the best. A common process model will also support the dissemination of knowledge and experience within the organization.

The vendors will benefit from the increased comfort level of their customers. There is less need to educate customers about general issues of data mining. The focus shifts from whether data mining should be used at all to how data mining can be used to solve the business questions. Vendors can also add values to their products, for instance offering guidance through the process or sophisticated reuse of results and experiences. Service providers can train their personnel to a consistent level of expertise.

Analysts performing data mining projects can also benefit in many ways. For novices, the process model provides guidance, helps to structure the project, and gives advice for each task of the process. Even experienced analysts can benefit from check lists for each task to make sure that nothing important has been forgotten. But the most important role of a common process model is for communication and documentation of results. It helps to link the different tools and different people with diverse skills and backgrounds together to form an efficient and effective project.

3 The CRISP-DM Methodology

CRISP-DM builds on previous attempts to define knowledge discovery methodologies (Reinartz & Wirth, 1995; Adriaans & Zantinge, 1996; Brachman & Anand, 1996; Fayyad et al., 1996). This section gives an overview of the CRISP-DM methodology. More detailed information can be found in (CRISP, 1999).

3.1 Overview

The CRISP-DM methodology is described in terms of a hierarchical process model, comprising four levels of abstraction (from general to specific): *phases*, *generic tasks*, *specialized tasks*, and *process instances* (see figure 1).

At the top level, the data mining process is organized into a small number of *phases*. Each phase consists of several second-level *generic tasks*. This second level is called generic, because it is intended to be general enough to cover all possible data mining situations. The generic tasks are designed to be as *complete* and *stable* as possible. Complete means to cover both the whole process of data mining and all possible data mining applications. Stable means that we want the model be valid for yet unforeseen developments like new modeling techniques.

The third level, the *specialized task* level, is the place to describe how actions in the generic tasks should be carried out in specific situations. For example, at the second level there is a generic task called *build model*. At the third level, we might have a task called *build response model* which contains activities specific to the problem and to the data mining tool chosen.

The description of phases and tasks as discrete steps performed in a specific order represents an idealized sequence of events. In practice, many of the tasks can be performed in a different order and it will often be necessary to backtrack to previous tasks and repeat certain actions. The CRISP-DM process model does not attempt to capture all of these possible routes through the data mining process because this would require an overly complex process model and the expected benefits would be very low.

The fourth level, the *process instance level*, is a record of actions, decisions, and results of an actual data mining engagement. A process instance is organized according to the tasks defined at the higher levels, but represents what actually happened in a particular engagement, rather than what happens in general.

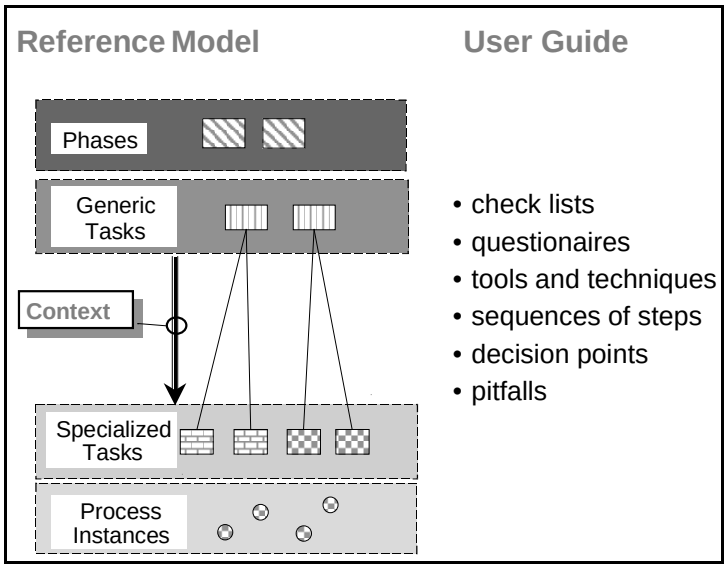

Figure 1: Four Level Breakdown of the CRISP-DM Methodology for Data Mining

The CRISP-DM methodology distinguishes between the *Reference Model* and the *User Guide*. Whereas the Reference Model presents a quick overview of phases, tasks, and their outputs, and describes *what to do* in a data mining project, the User Guide gives more detailed tips and hints for each phase and each task within a phase and depicts *how to do* a data mining project.

3.2 The Generic CRISP-DM Reference Model

The CRISP-DM reference model for data mining provides an overview of the life cycle of a data mining project. It contains the phases of a project, their respective tasks, and their outputs.

The life cycle of a data mining project is broken down in six phases which are shown in Figure 2. The sequence of the phases is not strict. The arrows indicate only the most important and frequent dependencies between phases, but in a particular project, it depends on the outcome of each phase which phase, or which particular task of a phase, has to be performed next.

The outer circle in Figure 2 symbolizes the cyclic nature of data mining itself. Data mining is not finished once a solution is deployed. The lessons learned during the process and from the deployed solution can trigger new, often more focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones (cf. the virtuous cycle of (Berry and Linoff, 1997)).

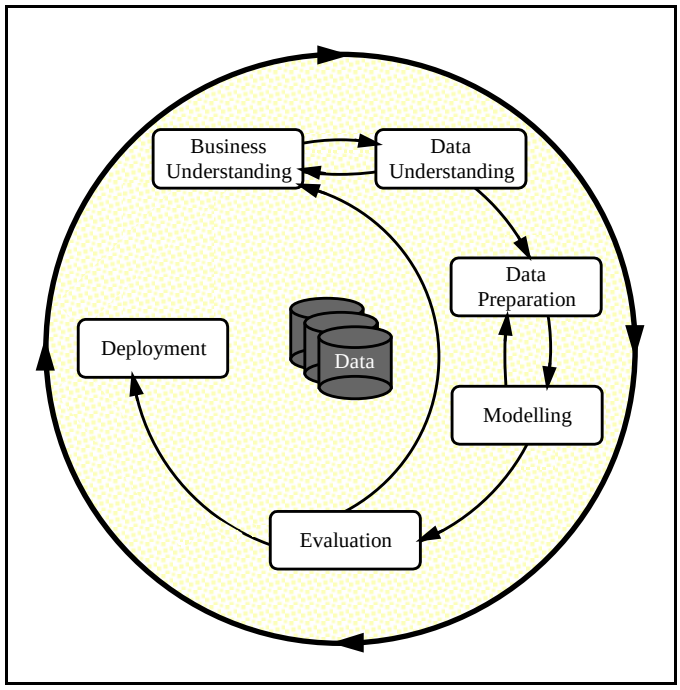

Figure 2: Phases of the Current CRISP-DM Process Model for Data Mining

In the following, we outline each phase briefly:

- *Business Understanding*

  This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives.

- *Data Understanding*

  The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

  There is a close link between Business Understanding and Data Understanding. The formulation of the data mining problem and the project plan require at least some understanding of the available data.

- *Data Preparation*

  The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and

attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools.

- *Modeling*

  In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques require specific data formats.

  There is a close link between Data Preparation and Modeling. Often, one realizes data problems while modeling or one gets ideas for constructing new data..

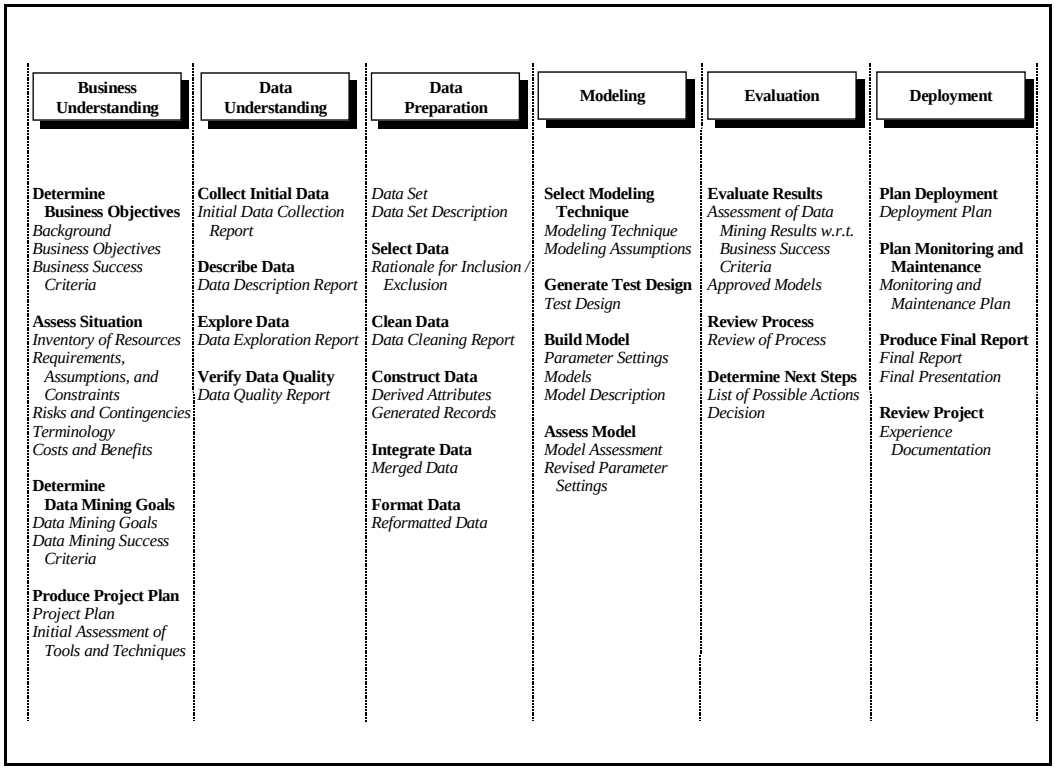

Figure 3: Overview of the CRISP-DM tasks and their outputs.

- *Evaluation*

  At this stage in the project you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

- *Deployment*

Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.

4 CRISP-DM in Action

Response modeling is an approach to improve the efficiency and effectiveness of mailing actions in marketing. It allows to increase the response rate while decreasing the costs of a campaign. By generalizing the knowledge we already have on our prospects with the help of data mining methods, we are able to predict the likelihood of potential customers to reply to our mailings.

In the context of a mailing action, the main objective of response modeling is to restrict the mailings only to people belonging to the target group. Apart from this, response modeling helps us to learn about and to understand our prospects. Therefore, we always aim at getting an understandable, actionable profile of the target group as a further requirement.

In order to be able to train models according to our goals we rely on data from the past that contain information on the behavior we want to predict, e.g. contain an attribute that indicates whether a customer changed his car brand and switched to a Mercedes. We obtain such data from our own customer database at DaimlerChrysler, from general surveys, or buy them from address providers.

The goal of our project is to establish a standardized process which can be reliably performed by marketing people with only little data mining skills and little time to experiment with different approaches. There is a core project team which does the initial case studies, develops and maintains the process, trains the marketing people, and which will later support the marketing people with more challenging applications scenarios outside the standard procedure. The project team consists of both experienced data miners and marketing specialists.

The initial case studies focused on acquisition campaigns, i.e., the challenge is to select prospects from an address list which are likely to buy a Mercedes for the first time. Although this sounds like a textbook data mining application, there are many problematic implications.

First of all, the process is not as stable and predictable as one might expect. There are many factors, like the available data or the market situation, that make every response modeling project unique in a certain sense. This gets even worse if we want the process to be applicable in different European countries with differences in language, culture, law, and competitive situation.

Consequently, we get into the following dilemma: On the one hand it is obviously impossible and after all even not desirable to give a detailed and exhaustive process description. This is due to the complexity resulting from the many unknown factors and unforeseeable situations that will always remain. We experienced that even from campaign to campaign in the same country the circumstances may differ fundamentally. On the other hand, the opposite extreme, a very high

level description like the generic CRISP-DM, is not a solution either. While it covers the whole process and is useful for experienced people, it is not suitable for the kind of users one is confronted with when moving into normal business processes. The resulting process description should guide the user as much as possible but, at the same time, enable him to handle difficult unexpected situations.

One of our basic decisions was to follow the CRISP-DM methodology as much as possible. We used the generic reference model for planning the case studies, for communication within the project team, for communication outside the project, and for documenting results and experiences. While it was very useful for these purposes, it is too abstract to describe repeatable processes for our end users. According to CRISP-DM, the proper level would be a specialized process model, whose context is response modeling for acquisition campaigns using Clementine[2].

Basically, we used two resources for the development of the specialized process model, the CRISP-DM Generic User Guide and the documentation of three case studies, which correspond to CRISP-DM process instances.

We worked simultaneously top-down and bottom-up. We took the Generic User Guide as a structure. We deleted tasks and activities which were not relevant for our application, renamed tasks to make them more concrete, and added a few tasks at various places. The additions, however, were not due to omissions in the generic model. It was more like splitting one abstract task into several more concrete tasks. The main work was to generate check lists of specific activities for each task. There again, the generic check lists were an excellent guideline.

In addition we enriched the specialized process model with examples gained from case studies. Furthermore we derived templates from the most important Clementine streams[3] that we developed during these initial projects. These templates proved to be quite helpful especially for people who are not proficient with our particular data mining tool.

5 Lessons learned

In this section, we try to summarize some of the experiences we made in both applying the generic CRISP-DM process model and in developing a specialized process model.

We expected the generic process model to be useful for planning and documentation, and this turned out to be the case. However, the use of the model for communication both within and outside the project was much more advantageous than we originally anticipated. Presenting the project plan and status reports in terms of the process model and, of course, the fact that we followed a process, inspired a lot of confidence in users and sponsors. It also facilitated status meetings because the process model provided a clear reference and a common terminology.

Although we relied on the reference model, we did not always follow the advice of the user guide. Sometimes it was faster to just go ahead (and this is ok with the CRISP-DM

[2] Clementine is a trademark of SPSS, Inc.

[3] Clementine relies on a visual programming interface. In the sense of this interface a stream is basically the visualization of a sequence of actions to be executed on some data.

methodology). However, sometimes, we encountered problems because we did not follow the model. Occasionally, we skipped planning and documentation tasks because they are time-consuming and we thought, we (the experts !) could do without. But in the end, we spent probably more time than we would have spent if we had done the proper level of explicit planning. Sometimes, we overlooked something which led us to a blind alley or which caused us to waste our effort on the wrong problem. In other cases, we failed to communicate some information to other team members who then did their part of the job based on false assumptions or re-did something that was done before. Also, documentation at the end is difficult, if you try to reconstruct what and why you did something. Preparing the documents suggested by the CRISP-DM model is worth the effort.

A major challenge is to put a project management framework around this highly iterative, creative process with many parallel activities. In our case, we employed external service providers for some tasks. Therefore, there must be firm deadlines for a task to be completed to ensure timely completion and proper usage of resources. But when is a task complete in a data mining project? This is a question that cannot be answered by a generic process model. This must be addressed by the specialized process model. But even there, it is not trivial to come up with useful criteria. So far, we have not yet a complete, satisfactory and operational set of criteria in our application.

In our first case studies, we encountered an unexpected difficulty with the generic process model. Although in the CRISP-DM documents it is stated at several places, that the phases and tasks are not supposed to be strictly sequential, the clear and obvious presentation of the process model inevitably created this impression in our decision makers. Despite our arguments, we found ourselves forced to very tight deadlines, which in the end let to sub-optimal solutions. In our case, this was not a serious problem because we still learned enough about the problem and about our specific process. But in general, this could jeopardize your project.

On the other hand, the process model gave us a structured documentation which allowed us to justify how much effort was spent in which task. This made it fairly easy to argue for more realistic resources in later applications.

Based on this experience, we strongly advise to plan for iterations explicitly. As a rough guide we suggest to plan for three iterations where the second and third iterations are allowed half the time and a quarter of the time of the first iteration. Also take into account that it is never the case that a phase is completely done before the subsequent phase starts. The relation between the phases is such that a phase cannot start before the previous one has started.

From this, it follows that the distinction between phases is not clear cut. But this is a rather academic question. In practice, we only need to come up with a sensible grouping of tasks.

Quality assurance is very important and cannot be stressed too much. Within the CRISP-DM process model quality assurance is not really stressed because it is considered to be a management issue common to all kinds of projects. But it turned out that there are many process quality issues which need to be addressed by a specialized process model.

After we did and documented the first case studies, it became very obvious, that many problems are much worse than expected. E.g., data preparation turned out to be a much more time-

consuming and error-prone than we thought (and we already expected it to be bad). Now, we have much more realistic estimates for tasks, and a handle for the project planning. But even now, we expect our estimates to be updated with every future case study.

And this is probably the most important lesson from this exercise. A data mining process is a living process which will be changed by future experiences. Therefore, all documentation and all process models must be flexible and living as well.

6 Conclusions and Future Work

We can conclude that CRISP-DM works. The generic process model is useful for planning, documentation and communication. It is fairly easy to write specialized process models based on generic check lists. Finding the right level of detail is still difficult. But the process is living and therefore all the documents must be living documents, too.

The claims of the CRISP-DM projects are not easy to evaluate, especially in terms of speed and costs. For a single, small scale projects, the improvements are probably less than expected. But CRISP-DM really pays off for repeatable processes and for large projects with several people involved.

In the future, we will constantly adapt the specialized process model as new experience is gathered. On the technical side, one of our immediate goals is the definition of key performance indicators to make it easier to judge and to control the progress of a particular project.

Acknowledgements:

We gratefully acknowledge the contributions and discussions with our colleagues from the CRISP-DM consortium; Thomas Reinartz from DaimlerChrysler; Julian Clinton, Tom Khabaza and Colin Shearer from SPSS, and Pete Chapman, Jens Hejlesen and Randy Kerber from NCR.

References

Adriaans, P.; Zantinge, D. (1996). *Data Mining*. Addison-Wesley. Harlow, England.

Agrawal, R. (1999). Data Mining: Crossing the Chasm. Presentation at International Conference on Knowledge Discovery and Data Mining, San Diego.

Brachman, R.J.; Anand, T. (1996). The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., & Uthurasamy, R. (eds.). *Advances in Knowledge Discovery and Data Mining*. AAAI /MIT Press

Berry, M. J. A.; Linoff, G. (1997*)*: *Data Mining Techniques. For Marketing, Sales and Customer Support*. Wiley Computer Publishing

Brodley, C. E.; Smyth, P. (1995): The Process of Applying Machine Learning Algorithms. Presented at the Workshop on Applying Machine Learning in Practice, 12th International Machine Learning Conference (IMLC 95), Tahoe City, CA.

Cabena, P.; Hadjinian, P.; Stadler, R; Verhees, J.; Zanasi, A. (1998): *Discovering Data Mining. From Concept To Implementation*. Prentice Hall, Upper Saddle River, New Jersey.

The CRISP-DM process model (1999), http://www.crisp-dm.org/

Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. (1995): From Data Mining to Knowledge Discovery: An Overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.). *Advances in Knowledge Discovery and Data Mining*. MIT Press, Cambridge, MA.

Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. (1996): The KDD Process for Extracting Useful Knowledge from Volumes of Data. In *Communications of the ACM*, November 1996, Vol. 39, No. 11, pp. 27-34.

Hand, D. J. (1997): *Construction and Assessment of Classification Rules*. John Wiley & Sons Ltd., Sussex, England.

Moore, G. (1991): *Crossing the Chasm*. Harper Business.

Reinartz, T., & Wirth, R. (1995). The Need for a Task Model for Knowledge Discovery in Databases. In: Kodratoff, Y., Nakhaeizadeh, G., & Taylor, C. (eds.). Workshop Notes Statistics, Machine Learning, and Knowledge Discovery in Databases. MLNet Familiarization Workshop, April, 28-29, Heraklion, Crete, pp. 19-24.