# Cross Industry Standard Process for Data Mining

From Wikipedia, the free encyclopedia

***CRISP-DM*** stands for Cross Industry Standard Process for Data Mining.[1] It is a data mining process model that describes commonly used approaches that expert data miners use to tackle problems. Polls conducted in 2002, 2004, and 2007 show that it isthe leading methodology used by data miners.[2][3][4] The only other data mining standard named in these polls was SEMMA. However, 3-4 times as many people reported using CRISP-DM. A review and critique of data mining process models in 2009 called the CRISP-DM the "de facto standard for developing data mining and knowledge discovery projects."[5] Other reviews of CRISP-DM and data mining process models include Kurgan and Musilek's 2006 review,[6] and Azevedo and Santos' 2008 comparison of CRISP-DM and SEMMA.[7]

## Contents



## Major phases

CRISP-DM breaks the process of data mining into six major phases.[8]

The sequence of the phases is not strict and moving back and forth between different phases is always required. The arrows in the process diagram indicate the most important and frequent dependencies between phases. The outer circle in the diagram symbolizes the cyclic nature of data mining itself. A data mining process continues after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions and subsequent data mining processes will benefit from the experiences of previous ones.

- **Business Understanding**

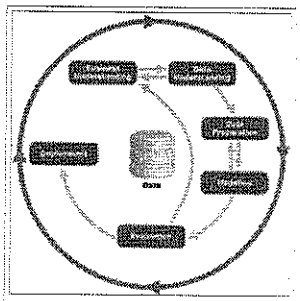
  This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

- **Data Understanding**

  The data understanding phasestarts 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.

- **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 as well as transformation and cleaning of data for modeling tools.



Process diagram showing the relationship between the different phases of CRISP-DM

- **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 have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

- **Evaluation**

  At this stage in the project you have built a model (or 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. Even if the purpose of the model is to increase knowledge of the data, 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 customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front the actions which will need to be carried out in order to actually make use of the created models.

## History

CRISP-DM was conceived in 1996. In 1997 it got underway as a European Union project under the ESPRIT funding initiative. The project was led by five companies: SPSS, Teradata, Daimler AG, NCR Corporation and OHRA, an insurance company.

This core consortium brought different experiences to the project: ISL, later acquired and merged into SPSS Inc. The computer giant NCR Corporation produced the Teradata data warehouse and its own data mining software. Daimler-Benz had a significant data mining team. OHRA was just starting to explore the potential use of data mining.

The first version of the methodology was presented at the 4th CRISP-DM SIG Workshop in Brussels in March 1999,[9] and published as a step-by-step data mining guide later that year.[10]

Between 2006 and 2008 a CRISP-DM 2.0 SIG was formed and there were discussions about updating the CRISP-DM process model.[5][11] The current status of these efforts is not known. However, the original crisp-dm.org website cited in the reviews,[6][7] and the CRISP-DM 2.0 SIG website[5][11] are both no longer active.

While many non-IBM data mining practitioners use CRISP-DM,[2][3][4][5] IBM is the primary corporation that currently embraces the CRISP-DM process model. It makes some of the old CRISP-DM documents available for download[10] and it has incorporated it into its SPSS Modeler product.

## References


1. ^ Shearer C., *The CRISP-DM model: the new blueprint for data mining*, J Data Warehousing (2000); 5:13—22.
2. ^ [a] [b] Gregory Piatetsky-Shapiro (2002); *KDnuggets Methodology Poll* (http://www.kdnuggets.com/polls/2002/methodology.htm)
3. ^ [a] [b] Gregory Piatetsky-Shapiro (2004); *KDnuggets Methodology Poll* (http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm)
4. ^ [a] [b] Gregory Piatetsky-Shapiro (2007); *KDnuggets Methodology Poll* (http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm)
5. ^ [a] [b] [c] [d] Óscar Marbán, Gonzalo Mariscal and Javier Segovia (2009); *A Data Mining & Knowledge Discovery Process Model* (http://cdn.intechopen.com/pdfs/5937/InTech-A_data_mining_amp_knowledge_discovery_process_model.pdf). In Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca, ISBN 978-3-902613-53-0, pp. 438-453, February 2009, I-Tech, Vienna, Austria.
6. ^ [a] [b] Lukasz Kurgan and Petr Musilek (2006); *A survey of Knowledge Discovery and Data Mining process models* (http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=451120). The Knowledge Engineering Review. Volume 21 Issue 1, March 2006, pp 1 - 24, Cambridge University Press, New York, NY, USA doi: 10.1017/S0269888906000737.
7. ^ [a] [b] Azevedo, A. and Santos, M. F. (2008);KDD, SEMMA and CRISP-DM: a parallel overview (http://www.iadis.net/dl/final_uploads/200812P033.pdf). In Proceedings of the IADIS European Conference on Data Mining 2008, pp 182-185.
8. ^ Harper, Gavin; Stephen D. Pickett (August 2006). "Methods for mining HTS data" (http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T64-4KDJSRH-4&_user=793840&_coverDate=08%2F31%2F2006&_rdoc=4&_fmt=full&_orig=browse&_srch=doc-info(%23toc%235020%232006%23999889984%23627946%23FLA%23display%23Volume)&_cdi=5020&_sort=d&_docanchor=&view=c&_ct=17&_acct=C000043460&_version=1&_urlVersion=0&_userid=793840&md5=f7f5b2376172e12b63177a32b03def11). *Drug Discovery Today* 11 (15–16): 694–699. doi:10.1016/j.drudis.2006.06.006 (http://dx.doi.org/10.1016%2Fj.drudis.2006.06.006). PMID 16846796 (//www.ncbi.nlm.nih.gov/pubmed/16846796).
9. ^ Pete Chapman (1999); *The CRISP-DM User Guide* (http://lyle.smu.edu/~mhd/8331f03/crisp.pdf).
10. ^ [a] [b] Pete Chapman, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth (2000); *CRISP-DM 1.0 Step-by-step data mining guide* (ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf).
11. ^ [a] [b] Colin Shearer (2006); *First CRISP-DM 2.0 Workshop Held* (http://www.kdnuggets.com/news/2006/n19/4i.html)


## External links

- Le site des dataminers (http://lesitedesdataminers.free.fr/02_PAGES_WEB/conduite_projet_crisp_dm.html) Article publié par Pascal BIZZARI, Mai 2009

Retrieved from "http://en.wikipedia.org/w/index.php?title=Cross_Industry_Standard_Process_for_Data_Mining&oldid=545188503"
Categories: Applied data mining

---