# Cross-industry standard process for data mining

**Cross-industry standard process for data mining**, known as **CRISP-DM**,[[1]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-Shearer00-1) is an [open standard](https://en.wikipedia.org/wiki/Open_standard) process model that describes common approaches used by [data mining](https://en.wikipedia.org/wiki/Data_mining) experts. It is the most widely-used [analytics](https://en.wikipedia.org/wiki/Analytics) model.[[2]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-2)

In 2015, [IBM](https://en.wikipedia.org/wiki/IBM) released a new methodology called [*Analytics Solutions Unified Method*](https://en.wikipedia.org/w/index.php?title=Analytics_Solutions_Unified_Method&action=edit&redlink=1) *for Data Mining/Predictive Analytics*[[3]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-3)[[4]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-4) (also known as ASUM-DM) which refines and extends CRISP-DM.

## Contents

1. [History](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#History)
2. [Major phases](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#Major_phases)
3. [Polls](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#Polls)
4. [References](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#References)

## History

CRISP-DM was conceived in 1996 and became a European Union project under the [ESPRIT](https://en.wikipedia.org/wiki/European_Strategic_Program_on_Research_in_Information_Technology) funding initiative in 1997. The project was led by five companies: [Integral Solutions Ltd (ISL)](https://en.wikipedia.org/w/index.php?title=Integral_Solutions_Ltd_(ISL)&action=edit&redlink=1), [Teradata](https://en.wikipedia.org/wiki/Teradata), [Daimler AG](https://en.wikipedia.org/wiki/Daimler_AG), [NCR Corporation](https://en.wikipedia.org/wiki/NCR_Corporation) and [OHRA](https://en.wikipedia.org/w/index.php?title=OHRA&action=edit&redlink=1), an insurance company.

This core consortium brought different experiences to the project: ISL, later acquired and merged into SPSS. The computer giant NCR Corporation produced the Teradata [data warehouse](https://en.wikipedia.org/wiki/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,[[5]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-crispDMbrussels-5) and published as a step-by-step data mining guide later that year.[[6]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-crispDMguide-6)

Between 2006 and 2008 a CRISP-DM 2.0 SIG was formed and there were discussions about updating the CRISP-DM process model.[[7]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-Shearer-7) The current status of these efforts is not known. However, the original crisp-dm.org website cited in the reviews,[[8]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-kurgan-8)[[9]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-AzevedoSantos-9) and the CRISP-DM 2.0 SIG website[[7]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-Shearer-7) are both no longer active.

While many non-IBM data mining practitioners use CRISP-DM,[[10]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-KDnug2002-10)[[11]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-KDnug2004-11)[[12]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-KDnug2007-12) IBM is the primary corporation that currently uses the CRISP-DM process model. It makes some of the old CRISP-DM documents available for download[[6]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-crispDMguide-6) and it has incorporated it into its [SPSS Modeler](https://en.wikipedia.org/wiki/SPSS_Modeler) product.

Based on current research CRISP-DM is the most widely used form of data-mining model because of its various advantages which solved the existing problems in the data mining industries. Some of the drawbacks of this model is that it does not perform project management activities. The fact behind the success of CRISP-DM is that it is industry, tool, and application neutral.[[13]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-13)

## Major phases

Diagrama

Descrição gerada automaticamente

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

CRISP-DM breaks the process of [data mining](https://en.wikipedia.org/wiki/Data_mining) into six major phases:[[14]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-Harper06-14)

* Business Understanding
* Data Understanding
* Data Preparation
* Modeling
* Evaluation
* Deployment

The sequence of the phases is not strict and moving back and forth between different phases as it 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.

## Polls

Polls conducted at the same website (KDNuggets) in 2002, 2004, 2007 and 2014 show that it was the leading methodology used by industry data miners who decided to respond to the survey.[[10]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-KDnug2002-10)[[11]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-KDnug2004-11)[[12]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-KDnug2007-12)[[15]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-KDnug2014-15) The only other data mining approach named in these polls was [SEMMA](https://en.wikipedia.org/wiki/SEMMA). However, SAS Institute clearly states that SEMMA is not a data mining methodology, but rather a "logical organization of the functional toolset of SAS Enterprise Miner." 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."[[16]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-16) Other reviews of CRISP-DM and data mining process models include Kurgan and Musilek's 2006 review,[[8]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-kurgan-8) and Azevedo and Santos' 2008 comparison of CRISP-DM and SEMMA.[[9]](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#cite_note-AzevedoSantos-9) Efforts to update the methodology started in 2006, but have, as of June 2015, not led to a new version, and the "Special Interest Group" (SIG) responsible along with the website has long disappeared (see [History of CRISP-DM](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#History)).

## References

Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22.

[What IT Needs To Know About The Data Mining Process](https://www.forbes.com/sites/metabrown/2015/07/29/what-it-needs-to-know-about-the-data-mining-process/#2065f3a3515f). Published by Forbes, 29 July 2015, retrieved June 24, 2018

[Have you seen ASUM-DM?](https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/), By Jason Haffar, 16 October 2015, SPSS Predictive Analytics, IBM [Archived](https://web.archive.org/web/20160308065035/https:/developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/) 8 March 2016 at the [Wayback Machine](https://en.wikipedia.org/wiki/Wayback_Machine)

[Analytics Solutions Unified Method - Implementations with Agile principles](ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf) Published by IBM, 1 March 2016, retrieved October 5, 2018

Pete Chapman (1999); [The CRISP-DM User Guide](http://lyle.smu.edu/~mhd/8331f03/crisp.pdf).

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth (2000); The CRISP-DM User Guide ([entry on semantic scholar, including links to PDFs](https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72)), ([PDF version with high-resolution graphics](https://www.the-modeling-agency.com/crisp-dm.pdf)).

Colin Shearer (2006); [First CRISP-DM 2.0 Workshop Held](http://www.kdnuggets.com/news/2006/n19/4i.html)

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.

Azevedo, A. and Santos, M. F. (2008); [KDD, SEMMA and CRISP-DM: a parallel overview](http://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf). In Proceedings of the IADIS European Conference on Data Mining 2008, pp 182–185.

Gregory Piatetsky-Shapiro (2002); [KDnuggets Methodology Poll](http://www.kdnuggets.com/polls/2002/methodology.htm)

Gregory Piatetsky-Shapiro (2004); [KDnuggets Methodology Poll](http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm)

Gregory Piatetsky-Shapiro (2007); [KDnuggets Methodology Poll](http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm)

Gregory Piatetsky-Shapiro (2014); [KDnuggets Methodology Poll](http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html)

Mariscal,G.,Marban,O.,Fernandez,C. "A Survey of Data Mining and knowledge discovery process Models and methodologies". The Knowledge Engineering Review. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1017/S0269888910000032](https://doi.org/10.1017%2FS0269888910000032).

Harper, Gavin; Stephen D. Pickett (August 2006). ["Methods for mining HTS data"](https://caridokumen.com/download/methods-for-mining-hts-data-_5a462410b7d7bc7b7af27f4a_pdf). [Drug Discovery Today](https://en.wikipedia.org/wiki/Drug_Discovery_Today). 11 (15–16): 694–699. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1016/j.drudis.2006.06.006](https://doi.org/10.1016%2Fj.drudis.2006.06.006). [PMID](https://en.wikipedia.org/wiki/PMID_(identifier)) [16846796](https://pubmed.ncbi.nlm.nih.gov/16846796).

Martínez-Plumed, Fernando; Contreras-Ochando, Lidia; Ferri, Cèsar; Flach, Peter; Hernández-Orallo, José; Kull, Meelis; Lachiche, Nicolas; Ramírez-Quintana, María José (19 September 2017). "CASP-DM: Context Aware Standard Process for Data Mining". [arXiv](https://en.wikipedia.org/wiki/ArXiv_(identifier)):[1709.09003](https://arxiv.org/abs/1709.09003) [[cs.DB](https://arxiv.org/archive/cs.DB)].