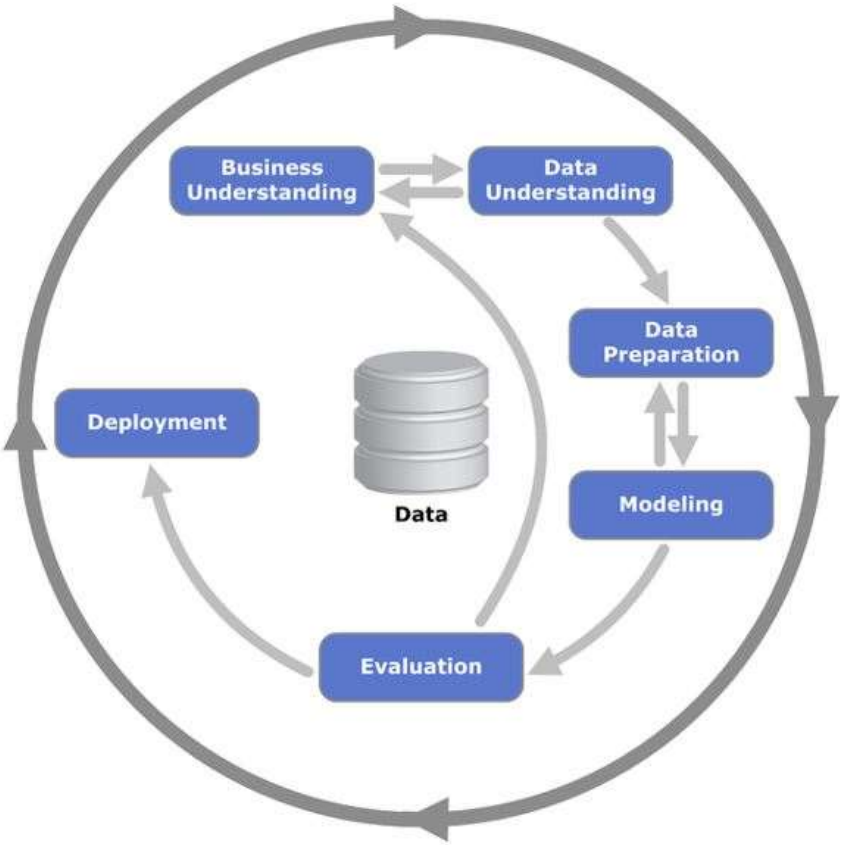
CRISP-DM Overview

Representatives from SPSS, Teradata, Daimler, NCR, and OHRA developed the *Cross-industry standard process for data mining (CRISP-DM)* in 1996 to standardize a data mining process across industries. CRISP-DM describes six major iterative phases, each with their own defined tasks and set of deliverables such as documentation and reports.

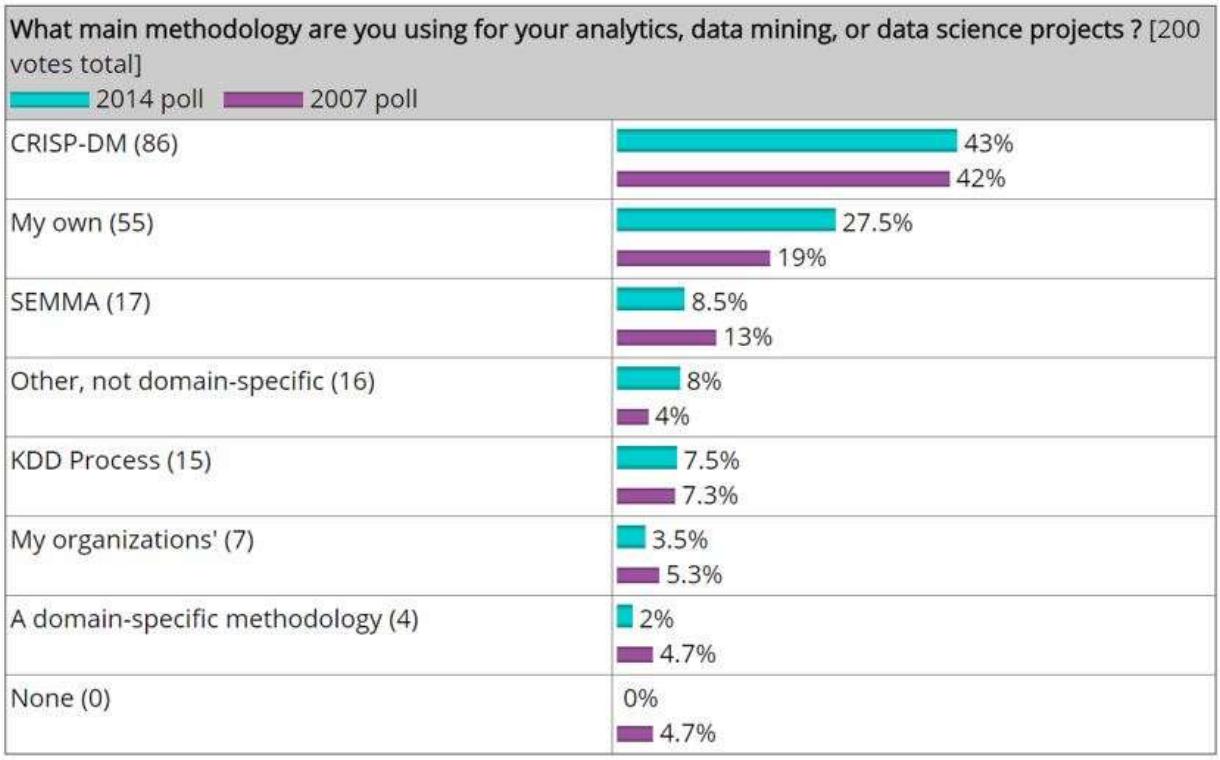
1. **Business Understanding:** determine business objectives; assess situation; determine data mining goals; produce project plan
2. **Data Understanding:** collect initial data; describe data; explore data; verify data quality
3. **Data Preparation** (generally, the most time-consuming phase): select data; clean data; construct data; integrate data; format data
4. **Modeling:** select modeling technique; generate test design; build model; assess model
5. **Evaluation:** evaluate results; review process; determine next steps
6. **Deployment:** plan deployment; plan monitoring and maintenance; produce final report; review project



CRISP-DM process diagram (Wikimedia Commons, 2012).

CRISP-DM has been consistently the most commonly used methodology for analytics, data mining and data science projects (per KDnuggets polls starting in 2002 up through

the most recent 2014 poll). Despite its popularity, CRISP-DM has not been revised since its creation (Piatesky, 2014).



Does CRISP-DM work for Data Science?

CRISP-DM is good for what it is: a natural description of workflow in data mining projects. As what is possibly "the first step towards defining a data science methodology" (Saltz J. , 2015), CRISP-DM has made a significant impact to bring some order to data science. However, like other KDD approaches, CRISP-DM provides a task-focused approach and fails to address team and communication issues.

Benefits

Although designed for data mining, William Vorhies, one of the creators of CRISP-DM, argues that because all data science projects start with business understanding, have data that must be gathered and cleaned, and apply data science algorithms, "CRISP-DM provides strong guidance for even the most advanced of today's data science activities" (Vorhies, 2016).

CRISP-DM is such a common sense and natural process that, when students were asked to do a data science project without project management directions, they "tended toward a CRISP-like methodology and identified the phases and did several iterations." Moreover, teams that were trained and explicitly told to implement CRISP-DM performed better than teams using other approaches (Saltz, Shamshurin, & Crowston, 2017).

Like Kanban, CRISP-DM can be implemented without much training, organizational role changes, or controversy.

The initial focus on *Business Understanding* is helpful to align technical work with business needs and to steer data scientists away from jumping into a problem without properly understanding business objectives. Its final step *Deployment* likewise addresses important considerations to close out the project and transition to maintenance and operations.

CRISP-DM's flexible, cyclical nature can provide many of the benefits of Agile. By accepting that a project starts with significant unknowns, the user can cycle through steps, each time gaining a deeper understanding of the data and the problem. The empirical knowledge learned from previous cycles can then feed into the following cycles.

Weaknesses and Challenges

On the other hand, some argue that CRISP-DM suffers from the same weaknesses of Waterfall and encumbers rapid iteration. Most notably, the sequential nature relies heavily on documentation; for example, CRISP-DM calls for 12 reports prior to data collection. Such issues were problematic for student teams that used CRISP-DM in a controlled experiment because they "were the last to start coding" and "did not fully understand the coding challenges they were going to face" (Saltz, Shamshurin, & Crowston, 2017).

Counter to Vorheis' argument for the sustaining relevance of CRISP-DM, others argue that CRISP-DM, as a process that pre-dates big data, "might not be suitable for Big Data projects due its four V's" (Saltz & Shamshurin, 2016).

Perhaps most significantly, CRISP-DM is not a true project management methodology because it implicitly assumes that its user is a single person or small, tight-knit team and ignores the teamwork coordination necessary for larger projects (Saltz, Shamshurin, & Connors, 2017). Therefore, structure should be added to help coordinate teamwork.

Recommendations

CRISP-DM is a great starting framework for those who are looking to understand the general data science process. It likewise may serve individual and small teams well, and if augmented with other project management approaches, might suite larger teams. Specifically, emerging approaches that combine agile project management and CRISP-DM are likely more effective. For full details, go to the CRISP-DM Guide.

Other Knowledge Discovery in Database Approaches

SAS and the Science Application International Corporation (SAIC) have also defined and published their own KDD approaches that are variants or expansions of CRISP-DM.

SEMMA

A few years prior to the publication of CRISP-DM, SAS independently developed *Sample, Explore, Modify, Model, and Assess (SEMMA)*. Although SEMMA is designed to help guide users through tools in SAS Enterprise Miner for data mining problems, it is often considered to be a general data mining methodology (Tiwari & Dixit, 2017). SEMMA (8.5%) was the third most popular methodology per the 2014 KDnuggets poll, but its use is down from 13% in 2007.

Compared to CRISP-DM, SEMMA is even more narrowly focused on the technical steps of data mining. It skips over the initial *Business Understanding* phase from CRISP-DM and instead starts with data sampling processes. SEMMA likewise does not cover the final *Deployment* aspects. Otherwise, its phases somewhat mirror the middle four phases of CRISP-DM. Although potentially useful as a process to follow data mining steps, SEMMA should not be viewed as a comprehensive project management approach.

KDD and KDDS

Knowledge Discovery in Database (KDD) is the general process of discovering knowledge in data through *data mining,* or the extraction of patterns and information from large datasets using machine learning, statistics, and database systems.

In 2016, Nancy Grady of SAIC, expanded upon CRISP-DM to publish the *Knowledge Discovery in Data Science (KDDS).* "As an end-to-end process model from mission needs planning to the delivery of value", KDDS specifically expands upon CRISP-DM to address big data problems. It also provides some additional integration with management processes. KDDS defines four distinct phases: a*ssess, architect, build,* and *improve* and five process stages: *plan, collect, curate, analyze,* and *act* (Grady, 2016*).*

KDDS can be a useful expansion of CRISP-DM for big data teams. However, KDDS only addresses some of the shortcomings of CRISP-DM. For example, it is not clear how a team should iterate when using KDDS. In addition, its combination of phases and processes is less straight-forward. Adoption of KDDS outside of SAIC is not known.