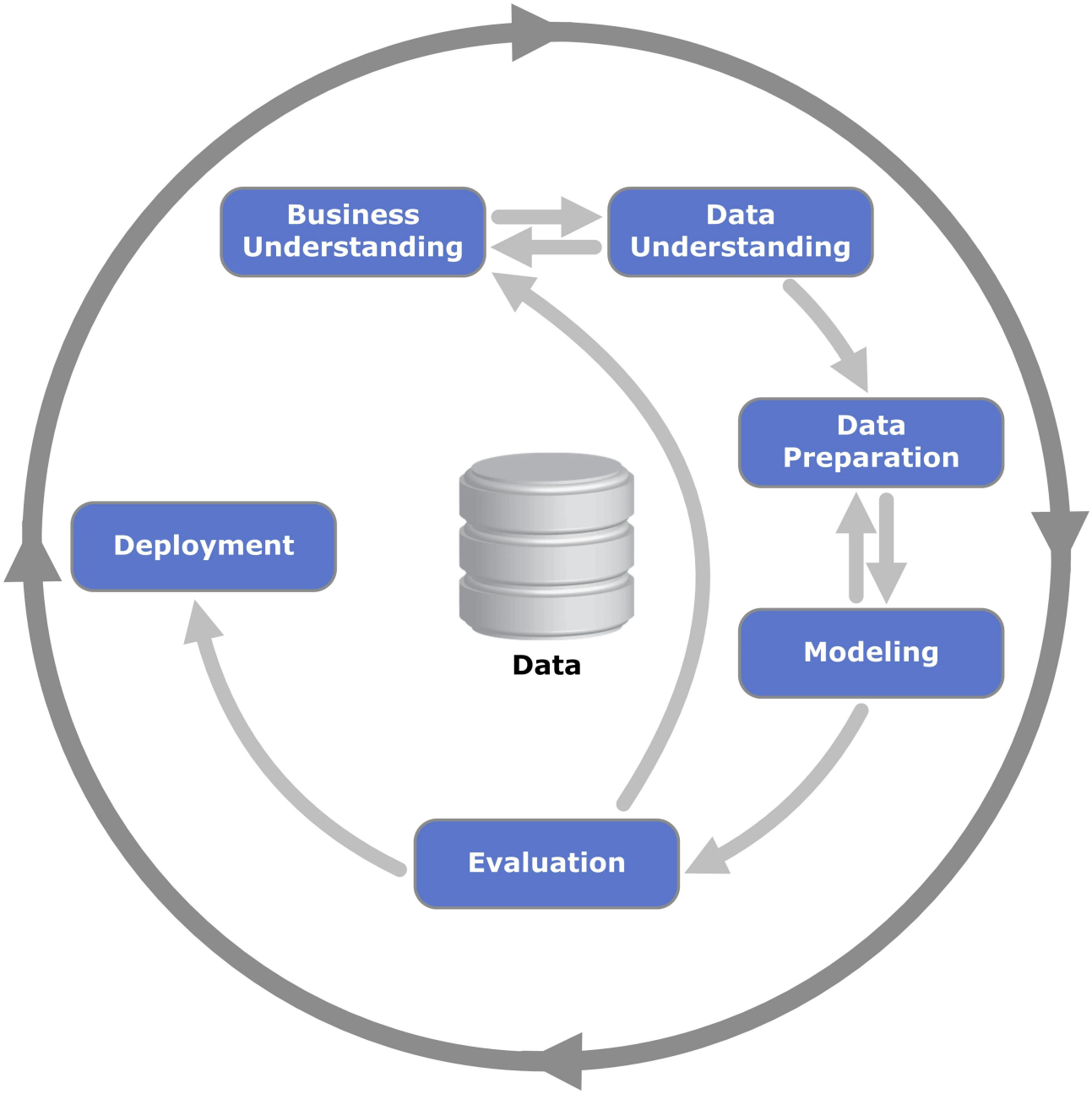
CRISP-DM

Idea In Short

In the early 1990's as data mining was evolving from toddler to adolescent. As a community, we spent a lot of time getting the data ready for the fairly limited tools and computing power. The CRISP-DM that emerged as a result is still valid today in the era of Big Data & Stream Analytics.

As the 90's progressed, the need to standardize the lessons learned into a common methodology became increasingly acute. Two of leading tool providers of the day - SPSS and Teradata - along with three early adopter user corporations, Daimler, NCR, and OHRA convened a Special Interest Group (SIG) in 1996 and over the course of less than a year managed to codify what is still today the CRISP-DM, CRoss Industry Standard Process for Data Mining1. CRISP-DM was not actually the first. SAS Institute had its own version called SEMMA (Sample, Explore, Modify, model, Assess). Nevertheless, within just a year or two many more practitioners were basing their approach on CRISP-DM.



CRISP DM Visualization

CRISP-DM Methodology

The CRISP-DM process or methodology of CRISP-DM is described in these six major steps2:

Business Understanding

Focuses on understanding the project objectives and requirements from a business perspective. The analyst formulates this knowledge as a data mining problem and develops

preliminary plan

Data Understanding

Starting with initial data collection, the analyst proceeds with activities to get familiar with the data, identify data quality problems & discover first insights into the data. In this phase, the analyst might also detect interesting subsets to form hypotheses for hidden information

Data Preparation

The data preparation phase covers all activities to construct the final dataset from the initial raw data

Modeling

The analyst evaluates, selects & applies the appropriate modelling techniques. Since some techniques like neural nets have specific requirements regarding the form of the data. There can be a loop back here to data prep

Evaluation

The analyst builds & chooses models that appear to have high quality based on loss functions that were selected. The analyst them tests them to ensure that they can generalise the models against unseen data. Subsequently, the analyst also validates that the models sufficiently cover all key business issues. The end result is the selection of the champion model(s)

Deployment

Generally this will mean deploying a code representation of the model into an operating system. This also includes mechanisms to score or categorise new unseen data as it arises. The mechanism should use the new information in the solution of the original business problem. Importantly, the code representation must also include all the data prep steps leading up to modelling. This ensures that the model will treat new raw data in the same manner as during model development

Characteristics of CRISP-DM

I believe CRISP-DM's longevity in a rapidly changing area stems from a number of characteristics:

1. It encourages data miners to focus on business goals, so as to ensure that project outputs provide tangible benefits to the organization. Too often, analysts can lose sight of the ultimate business purpose of their analysis – the analysis can become an end in itself rather than a means to an end. The CRISP-DM approach helps ensure that the business goals remain at the centre of the project throughout.
2. CRISP-DM provides an iterative approach, including frequent opportunities to evaluate the progress of the project against its original objectives. This helps minimize risk of getting to the end of the project and finding that the business objectives have not really been addressed. It also means that the project stakeholders can adapt & change the objectives in the light of new findings.
3. The CRISP-DM methodology is both technology and problem-neutral. You can use any software you like for your analysis and apply it to any data mining problem you want to. Whatever the nature of your data mining project, CRISP-DM will still provide you with a framework with enough structure to be useful.

Summary

From today's data science perspective this seems like common sense. Data science has moved beyond predictive modeling into recommenders, text, image, and language processing, deep learning, AI, and other project types that may appear to be more non-linear. If fact, all of these projects start with business understanding. All these projects start with data that must be gathered, explored, and prepped in some way. All these projects apply a set of data science algorithms to the problem. And all these projects need to be evaluated for their ability to generalize in the real world. So yes, CRISP-DM provides strong guidance for even the most advanced of today's data science activities.