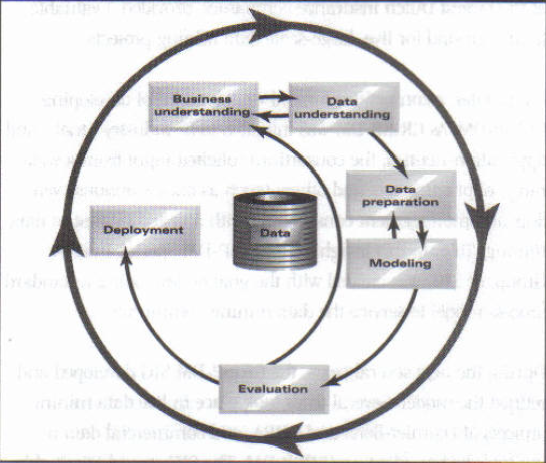
critique of colin shearer’s “the CRISP-DM Model: The New Blueprint for Data Mining”

# introduction

CRISP DM is one of the most widely used methodologies for data mining. CRISP DM stands for Cross Industry Standard Process for Data Mining. This methodology is a cyclic process that comprises of six stages of data mining.

1. Business understanding
2. Data understanding
3. Data preparation
4. Data modeling
5. Evaluation
6. Deployment



1. ***Business understanding:*** This is the first and the most important stage of this model. This phase undergoes the understanding of the objective in a business perspective. It involves four key steps, that are determining business objectives, assessing the situation, determining the goals of data mining and producing a project plan.
2. ***Data Understanding:*** This is the second phase where the raw data is collected, the gross and surface property of the data is analyzed, and the data quality is assessed. The steps in this phase are to collect the initial data, describe the data, explore the data and verify the quality of data.
3. ***Data Preparation:*** In this third phase, the final refined data set which has to be fed to modeling is prepared. The data is formatted, cleansed and integrated in this phase. The steps are selecting data, cleansing data, constructing data, integrating data and formatting data.
4. ***Data Modeling***: In this fourth phase, several modelling techniques are practiced. Going back to data preparation phase is very common in this phase, for modification of the data set pertaining to the respective machine algorithm to be carried out for modeling. The steps involved are selecting the modeling technique, generating test design, building the model and accessing the model.
5. ***Evaluation/Testing:*** It is very important to test the model before the deployment. It is essential as it gives a general idea about how the model will function in the real time environment once it is deployed. The steps are Evaluating results, process reviewing and determine next steps.
6. ***Deployment***: The deployment phase is based on the requirement, if the requirement is a simple task then the deployment will be easy. Normally, the deployment is done by the customer, and not by the analyst. The steps in this phase are plan deployment, plan monitoring and maintenance, produce final report and review project.

Though it is one of the most popular methodologies for data mining, data analysts from novice to experts use CRISP-DM because of its simple approach and the need for a repeatable approach. However, there are some prominent issues and challenges with the implementation of CRISP-DM on bigdata.

# Challenges of Crisp-dm on big data

The challenge arises to the question if CRISP-DM Model is still relevant to Big Data mining in 2018.

***“CRISP-DM remains the most popular methodology for analytics, data mining, and data science projects, with 43% share in latest KDnuggets Poll, but a replacement for unmaintained CRISP-DM is long overdue” - Gregory Piatesky***

CRISP-DM was conceived in late 1996 which was almost 22 years ago. Since then, they haven’t been updated because of the lack of proper maintenance, which in-turn leads to an outdated framework. Because of which the framework faces problems on working with new technologies like Big Data. Any models for that matter must have a constant maintenance with regular updates that evolves itself with the latest technologies in trend to be able to be in the market. Because of this, the model fails to keep up with something as challenging as the big data. Data analysts are switching from CRISP DM to other frame works because of its limitations on handling something as complex as big data. This, however, doesn’t change the fact that CRISP-DM is useless that it need not be practiced anymore. Almost all the phases of this model still plays huge role in most of the data mining methodologies today, irrespective of industry, domain or size of data.

The business understanding phase of CRISP-DM is very exhaustive and the performance will not be 100% as there will be many non-experts involved in it. Besides from understanding what is required for the analysis, it results in an inefficient way of gaining redundant and useless information. Similarly, in the data understanding phase, many issues could arise. The data quality provided by the customer cannot entirely be relied on, verification of the data is required. An added investigation could help in the process of obtaining information about what needs to be done on processing the raw data and to better understand the correlations, associations, anomalies and outliers. The data preparation phase is a very daunting task, as treating such huge dataset costs a lot of time. The data understanding, and data preparation phases consumes most of the time. But when compared to other methodologies CRISP DM is more detailed.

CRISP DM is still widely used by many companies, however, it won’t fit well for the Big data mining because of its vast size as it could be practically impossible for analysts to carry out all the phases on a bigdata. Besides, it does not give clarity on how analysis will help the business problems and does not focus into the details. Analysts could get diverted from the business objective by building interesting models that are irrelevant to the business requirement.

Since it involves multiple stages, confusions could arise between the dedicated teams during the process of hand-off of data/model. The risk of failure or mistakes is potentially high when multiple teams are involved in a single project, as per the saying that too many cooks could spoil a broth. Besides, it would also be practically impossible for a single team to carry out all the phases because of the big data.

There was an attempt however to recreate the magic of CRISP-DM with a new version 2.0 in the 2000s, but for some reason it never happened and because the model has been dormant for so long, some new methodologies emerged over the years, yet they share the same fundamental structure of CRISP-DM or just with mild enhancements of certain phases of the CRISP-DM methodology pertaining to the business needs.

The Microsoft Team Data Science Process (TDSP) is one such methodology that is used by the Microsoft for its data mining. Their say in this is that TDSP is more ‘Big Data-friendly’, as it includes both data acquisition and data understanding in the second phase unlike CRISP-DM that which only involves in data understanding. It is to be also noted that the last phase of TDSP model is Customer Acceptance.

William Vorhies (2016) in his article, “CRISP-DM – a Standard Methodology to Ensure a Good Outcome” has written that some people might argue that data mining has gone beyond predictive modeling into non-linear projects like text mining, recommender systems, image processing, deep learning, Artificial intelligence, etc., but ideally, all these projects should start with fundamental phase of Business Understanding. These projects also have to be started with data that must be acquired and prepared. These models need to apply a set of machine algorithms irrespective of the size of data or perceived linearity of the business problem, and it also needs to be evaluated to test their practicality in the real world. CRISP-DM, old or not still provides strong underlying guidance for even the most advanced data science projects of today.

Despite all this limitation, it still stands tall among the other methodologies as it serves as the fundamentals of mining, which is why it is still practiced in many organizations, it is also because of it adaptability, as it can nullify most of the issues in detailed development phases. It has flaws because it is not maintained properly pertaining to the new technological standards. However, CRISP DM is the most popular data mining techniques when compared to other data mining frame works around the world.

There are very few issues which is stopping the CRISP DM to be called flawless. If CRISP-DM framework is to be updated by addressing its challenges of big data mining, where data which goes inside one module at a time instead of everything at once and if the process is more clear, CRISP DM would be flawless and rule the data mining industry once again.

# References

1. Shearer, Colin, 2000. The CRISP-DM Model: The New Blueprint for Data Mining, Vol. 5, Num. 4, 10.
2. Yenkar, Vitthal, 2014. Review on “Data Mining with Big Data”. Review on “Data Mining with Big Data”, Vol. 3, Issue. 4, 6.
3. Ying Tan, Yuhui Shi, 2016. Data Mining and Big Data. 1st ed. Bali, Indonesia: Springer.
4. Michał Rogalewicz, Robert Sika, METHODOLOGIES OF KNOWLEDGE DISCOVERY FROM DATA AND DATA MINING METHODS IN MECHANICAL ENGINEERING, December 2016, pp. 97–108, DOI: 10.1515/mper-2016-0040
5. KDnuggets, CRISP-DM, Still the top methodology for analytics, data mining, or data science projects, 2014, Retrieved 25.10.2016