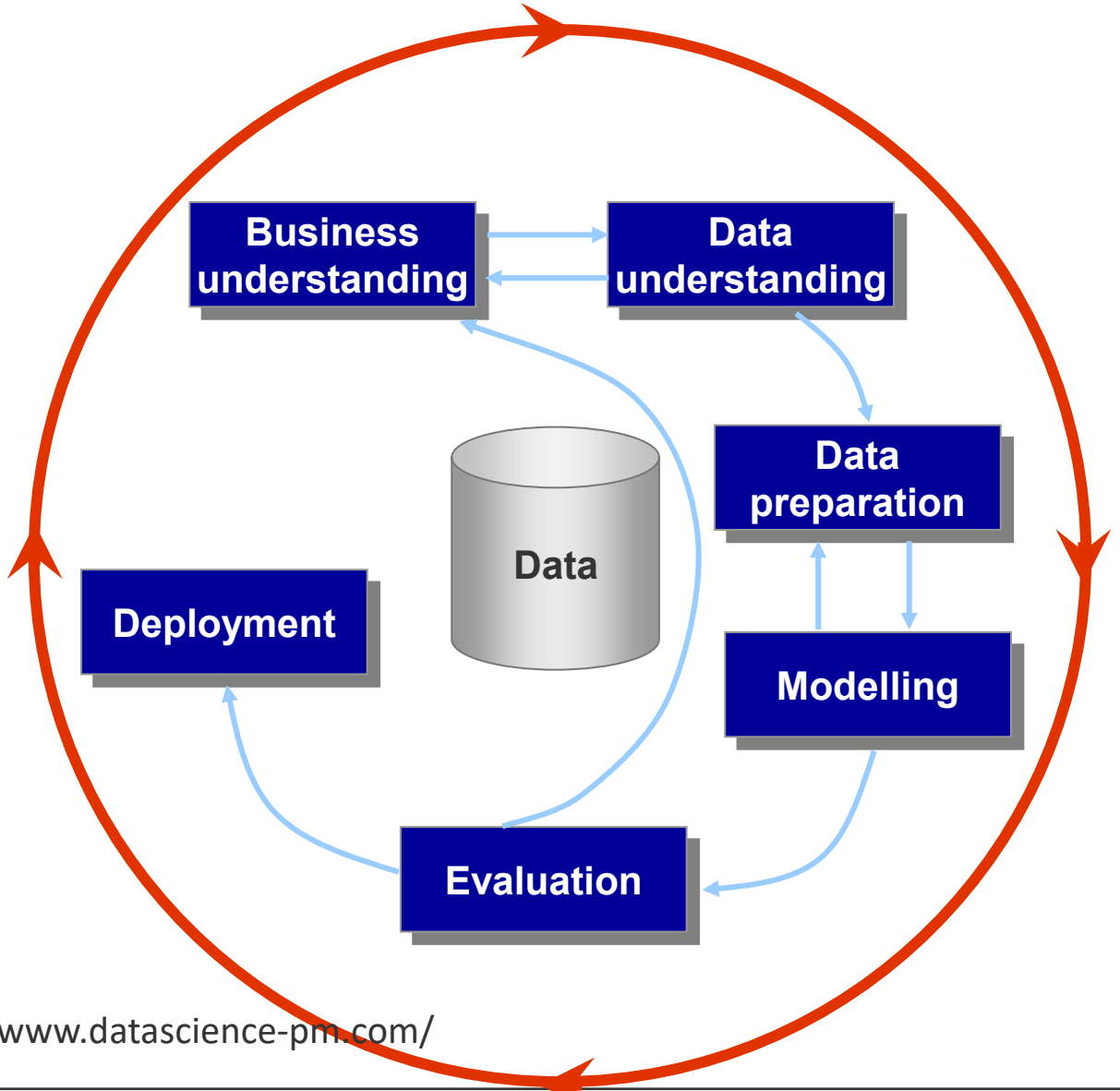
CRISP-DM

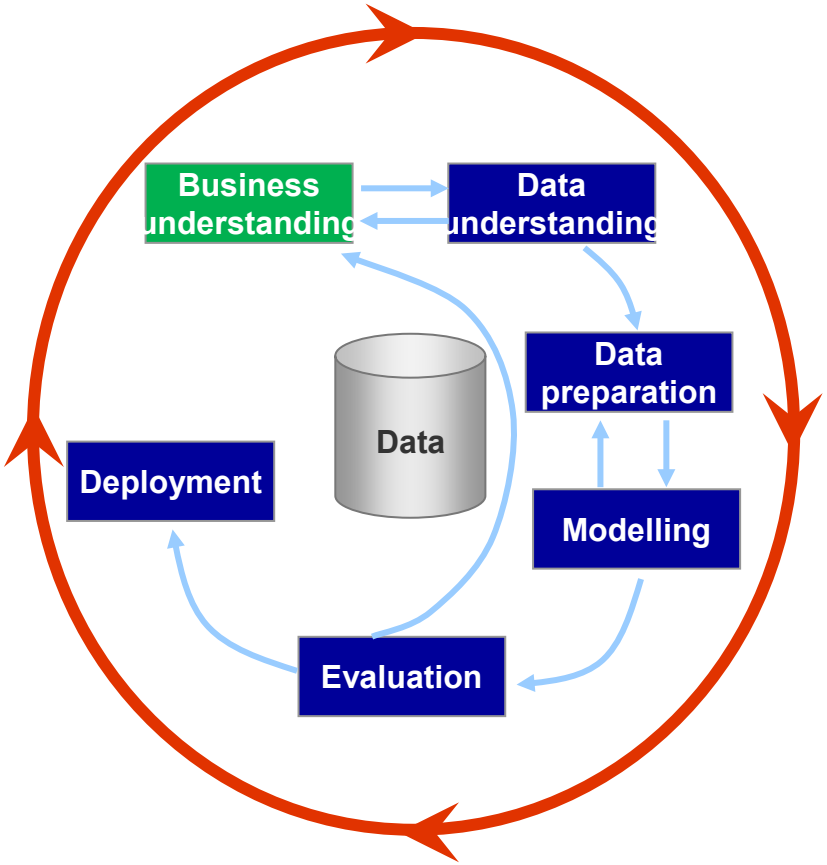


Source: https://www.datascience-pm.com/

Business understanding

The Business Understanding phase focuses on understanding the objectives and requirements of the project. Establishing a strong business understanding is absolutely essential. Aside from the third task, the three other tasks in this phase are foundational project management activities that are universal to most projects:

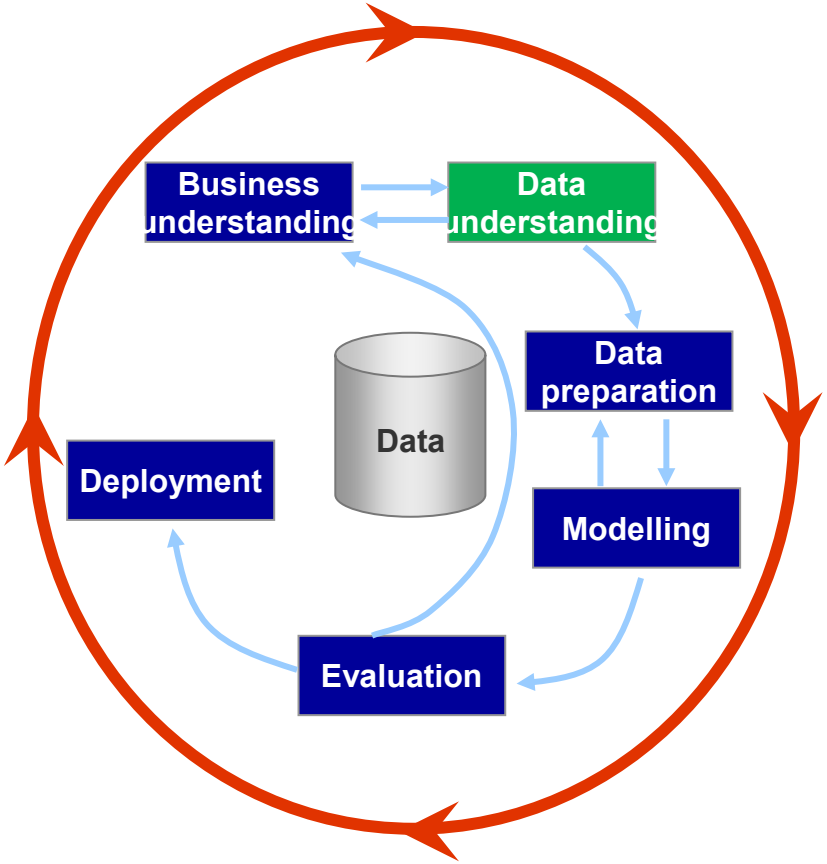
1. Determine business objectives: understand what the customer / client is trying to achieve, including the business success criteria.
2. Assess situation: Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.
3. Determine project goals: In addition to defining the business objectives, you should also define what success looks like from a technical data mining perspective.
4. Produce project plan: Select technologies and tools and define detailed plans for each project phase.



Data understanding

Adding to the foundation of Business Understanding, the Data Understanding phase focuses on identifying, collecting, and analyzing data sets that can help the project. This phase also has four tasks:

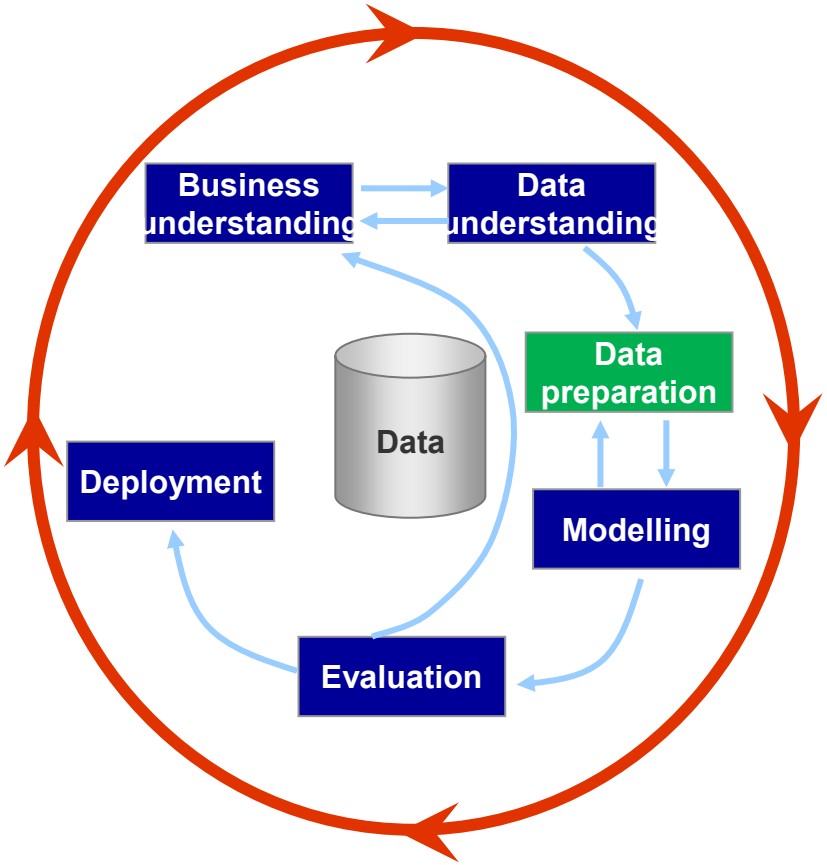
1. Collect initial data: Acquire the necessary data and (if necessary) load it into your analysis tool.
2. Describe data: Examine the data and document its surface properties like data format, number of records, or field identities.
3. Explore data: Dig deeper into the data. Query it, visualize it, and identify relationships among the data.
4. Verify data quality: How clean/dirty is the data? Document any quality issues.



Data preparation

This phase, which is often referred to as “data munging”, prepares the final data set(s) for modeling. A common rule of thumb is that 50% to 80% of the project effort is in the data preparation phase. This phase has five tasks:

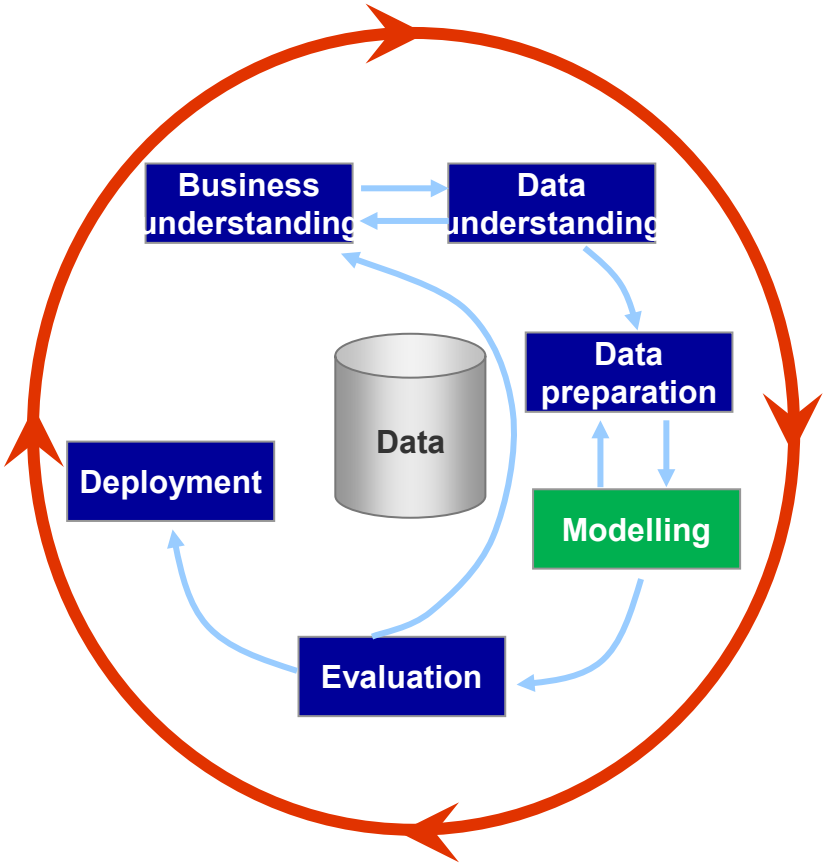
1. Select data: Determine which data sets will be used and document reasons for inclusion/exclusion.
2. Clean data: Often this is the lengthiest task. Without it, you’ll likely fall victim to garbage-in, garbage-out. A common practice during this task is to correct, impute, or remove erroneous values.
3. Construct data: Derive new attributes that will be helpful. For example, derive someone’s body mass index from height and weight fields.
4. Integrate data: Create new data sets by combining data from multiple sources.
5. Format data: Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations.



Modeling

Modeling is often regarded as data science's most exciting work. In this phase, the team builds and assesses various models based, often using several different modeling techniques. Although the CRISP-DM guide suggests to "iterate model building and assessment until you strongly believe that you have found the best model(s)", in practice teams might iterating until they have a "good enough" model. This phase has four tasks:

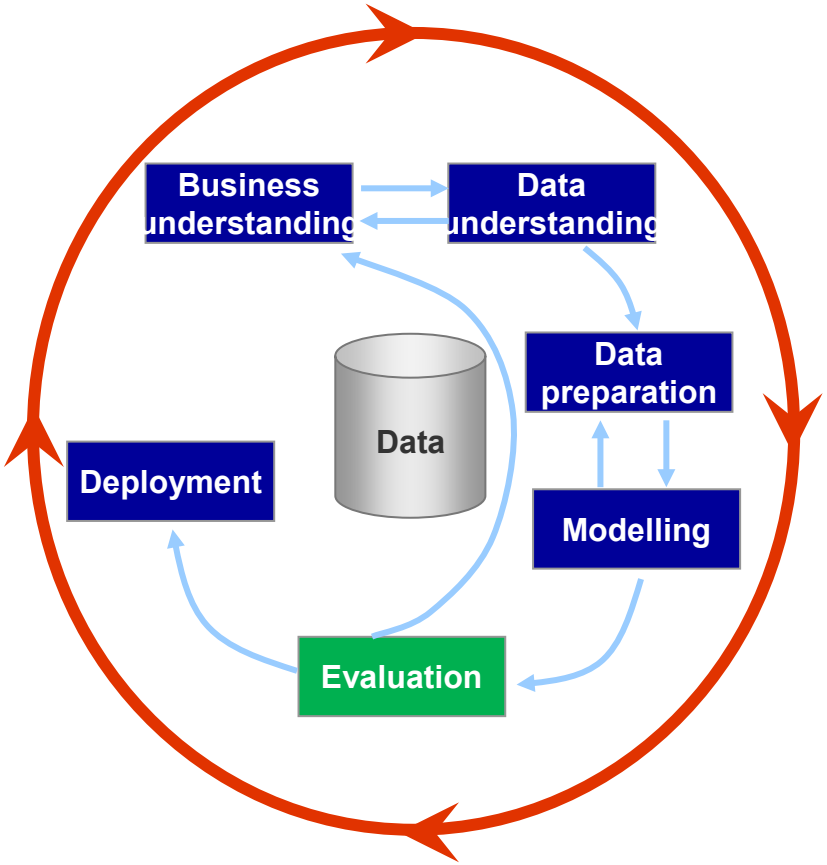
1. Select modeling techniques: Determine which algorithms to try (e.g. decision tree, kNN, neural net).
2. Generate test design: Pending your modeling approach, you might need to split the data into training, test, and validation sets.
3. Build model: As glamorous as this might sound, this might just be executing a few lines of code like "reg = LinearRegression().fit(X, y)".
4. Assess model: Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design.



Evaluation

Whereas the Assess Model task of the Modeling phase focuses on technical model assessment, the Evaluation phase looks more broadly at which model best meets the business and what to do next. This phase has three tasks:

1. Evaluate results: Do the models meet the business success criteria? Which one(s) should we approve for the business?
2. Review process: Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed.
3. Determine next steps: Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects



Deployment

A model is not particularly useful unless the customer can access its results. So, deployment should be thought of in terms of what does it take to actually use the results of the project. Depending on the project, this can be as simple as sharing a report or as complex as implementing a live real-time predictive model. This final phase has four tasks:

1. Plan deployment: Develop and document a plan for deploying the model.
2. Plan monitoring and maintenance: Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase (or post-project phase) of a model.
3. Produce final report: The project team documents a summary of the project which might include a final presentation of data mining results.
4. Review project: Conduct a project retrospective about what went well, what could have been better, and how to improve in the future