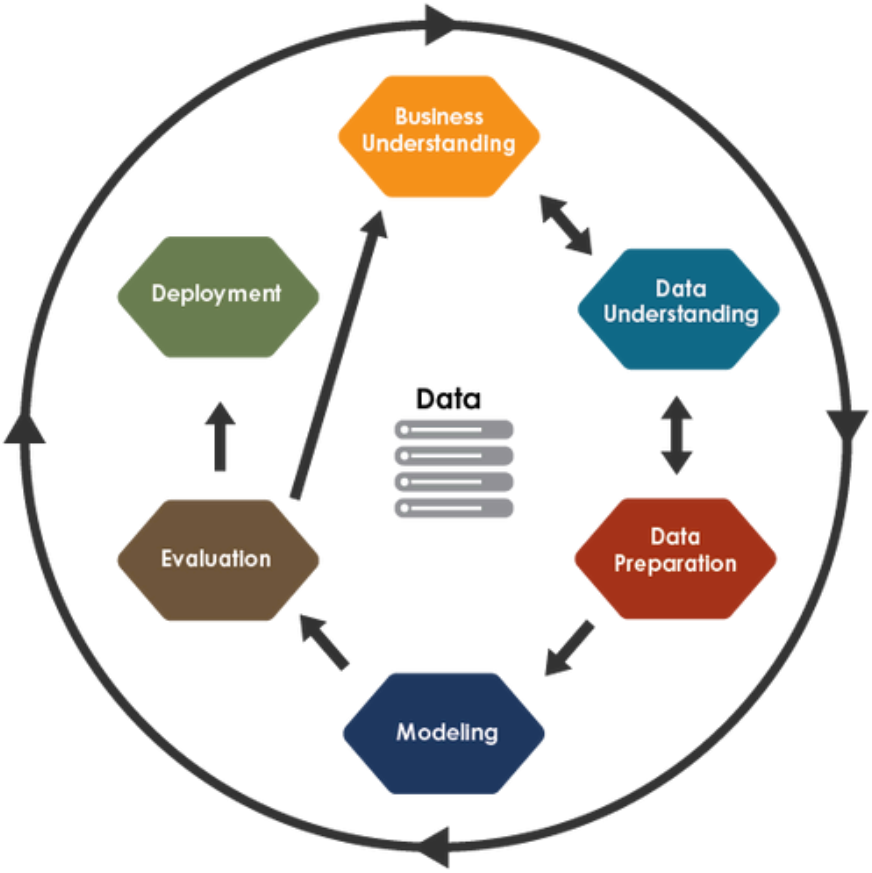
EVALUATING CRISP-DM FOR DATA SCIENCE



How can you use the classic data science life cycle on your next project?

Data Science PM

Integrating data science process effectiveness research with industry leading agile training expertise

Data Science PM

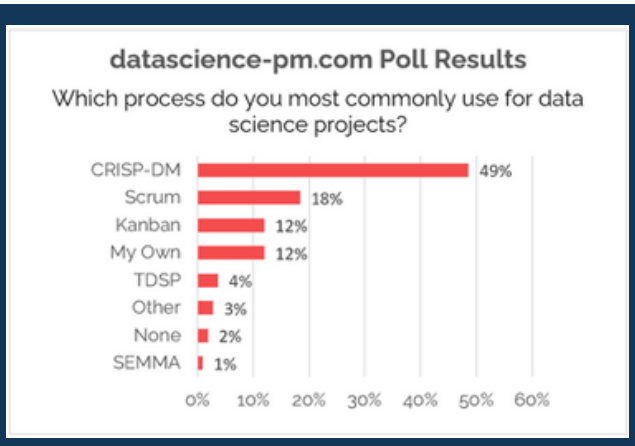
Integrating data science process effectiveness research with industry leading agile training expertise

Executive Summary

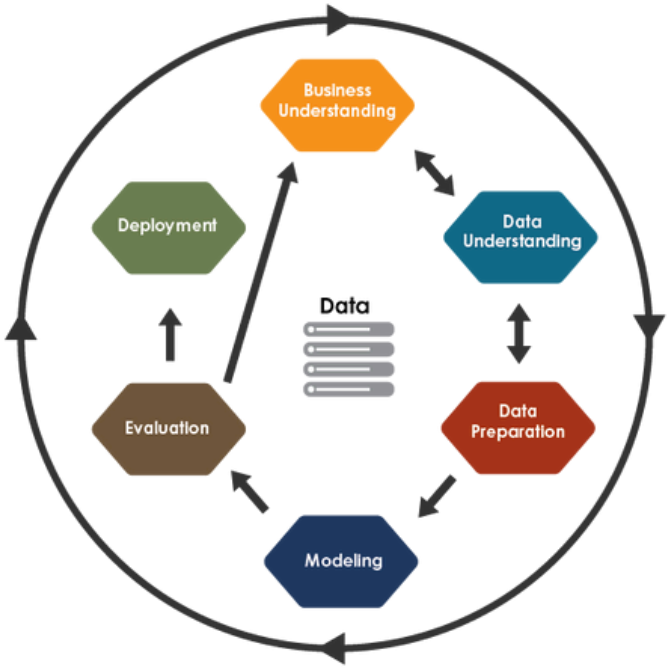
What is CRISP-DM?

Published in 1999, CRISP-DM (CRoss Industry Standard Process for Data Mining (CRISP-DM) is the most popular framework for executing data science projects. It provides a natural description of a data science life cycle (the workflow in data-focused projects).

However, this task-focused approach for executing projects fails to address team and communication issues. Thus, CRISP-DM should be combined with other team coordination frameworks.



Results of a 2020 datascience-pm.com poll on the use of data science process



Six Phases

1. **Business understanding**
   What does the business need?
2. **Data understanding**
   What data do we have / need? Is it clean?
3. **Data preparation**
   How do we organize the data for modeling?
4. **Modeling**
   What modeling techniques should we apply?
5. **Evaluation**
   What best meets the business objectives?
6. **Deployment**
   How do stakeholders access the results?

How can you use CRISP-DM on your next Project?

Every project, team, and organization is unique. So to evaluate CRISP-DM for your next project, first review its key concepts. Then, assess its strengths and weaknesses. Finally, consider some keys tips for its use.

Evaluating CRISP-DM

1. Review the CRISP-DM framework
2. Explore Strengths & Weaknesses
3. Actions to consider

© Data Science PM 2025

Reviewing CRISP-DM
Diving into the CRISP-DM Phases

I. Business Understanding

The Business Understanding phase focuses on understanding the objectives and requirements of the project. While many teams hurry through this phase, establishing a strong business understanding is like building the foundation of a house – 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.

II. 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.

III. 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.

© Data Science PM 2025

Reviewing CRISP-DM
Diving into the CRISP-DM Phases

IV. 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. regression, 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.

V. 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.

VI. 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.

© Data Science PM 2025

Analyzing CRISP-DM
Strengths and Weaknesses

Strengths & Benefits

- **Common Sense:** Data scientists naturally follow a CRISP-DM-like process. When people are asked to do a data science project without project management direction, they tend toward a CRISP-like methodology and can easily identify with the CRISP-DM phases and doing iterations.
- **Cyclical:** CRISP-DM can support the iterative nature of data science (but how to actually do iterations is not defined)
- **Adopt-able:** CRISP-DM can be implemented without much training, organizational role changes, or controversy.
- **Right Start:** The initial focus on Business Understanding, an often-overlooked step, 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.
- **Flexible**: A loose CRISP-DM implementation can be flexible to provide many of the benefits of agile principles and practices. 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.

Key Strengths:

- Common sense steps
- Easy to understand
- Defines a shared vocabulary for the steps in a project

Weaknesses & Challenges

- **Not a Team Coordination Framework**: 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.
- **Can ignore stakeholders**: CRISP-DM phases and tasks can be done with minimal input from stakeholders.
- **Outdated**: CRISP-DM has not been updated since 1999 and is criticized for not meeting the considerations of modern big data science projects (e.g., operational support).
- **Documentation Heavy**: The full-fledged CRISP-DM approach requires a lot of time-consuming documentation (although most teams seem to skip much of it). In fact, nearly every task has a documentation step. While documenting one's work is key in a mature process, CRISP-DM's documentation requirements might unnecessarily slow the team from actually delivering increments.
- **Slow starts**: The process matches closely with building a waterfall-like approach, which could delay business value delivery by spending too much time on the early phases, without incremental learning.

Key Weaknesses:

- Not clear when to "loop back" to a previous phase
- Missing phases (operational support)
- No structured communication with stakeholders

© Data Science PM 2025

Going Forward
Key Actions to Consider

1. Combine with a team coordination process

- There needs to be a mechanism for the team to communicate and prioritize work.
- The team process should define how the team communicates, prioritizes tasks and "loops back" to previous project phases.
- Teams can leverage the CRISP-DM phases, and then use a framework such as Scrum, Kanban or Data Driven Scrum to prioritize potential tasks.

2. Ensure multiple experiments / iterations

- Iterate quickly and do not fall get pulled into a waterfall of sequential work.
- Rather, try to deliver thin vertical slices of end-to-end value. Your first deliverable might not be too useful. That's okay. Iterate.
- While it's important to do multiple iterations, each team needs to think through how iterations are defined and then evaluated.

3. Define team roles

- CRISP-DM does not include roles (nor a team).
- Data science efforts are increasingly a team sport.
- Roles can include stakeholders / product owners (to ensure the insight is actionable), as well as a process expert.

4. Ensure actionable insight

- CRISP-DM lacks a communication structure with stakeholders.
- How does the team ensure actionable insight?
- Be sure to communicate and set expectations with stakeholders frequently.

5. Add phases (if needed) and define the subitems within each phase

- Add steps or phases for practices like git version control and ML ops.
- Be clear how tasks (within a phase) are defined.
- Some tasks that should be explicitly discussed include: bias checks, accuracy assessments, business validation, and dev dicussions.

6. Document enough…but not too much

- CRISP-DM can be documentation heavy; for example, CRISP-DM calls for 12 reports prior to data collection.
- So, do what's reasonable and appropriate but don't go overboard.

© Data Science PM 2025