



What is CRISP DM?

BY NICK HOTZ | LAST UPDATED APR 28, 2024 | LIFE CYCLE

in

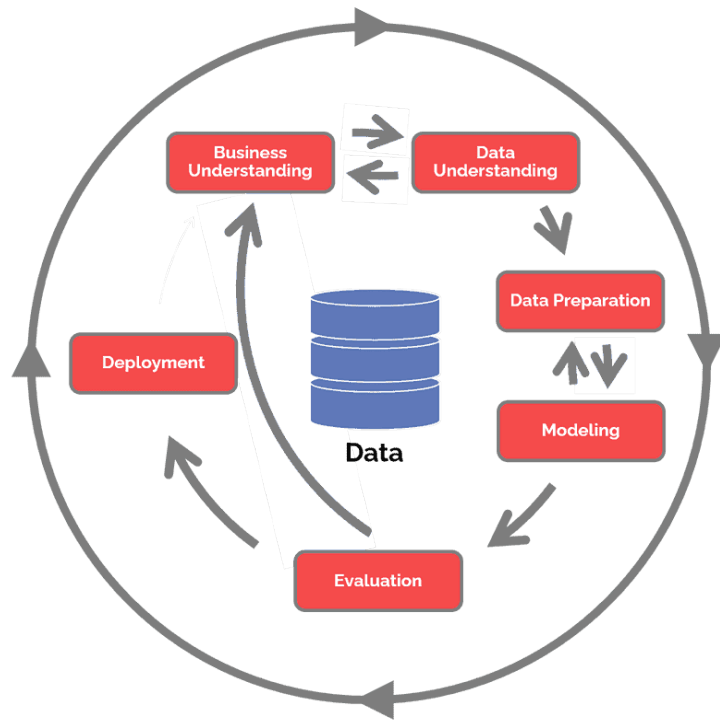
The **C**Ross Industry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) is a process model that serves as the base for a data science process.

It has six sequential 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** – Which model best meets the business objectives?
6. **Deployment** – How do stakeholders access the results?

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok



in

Published in 1999 to standardize data mining processes across industries, it has since become the **most common methodology** for data mining, analytics, and data science projects.

Data science teams that combine a loose implementation of CRISP-DM with overarching team-based **agile** project management approaches will likely see the best results.

What are the 6 CRISP-DM Phases?

I. Business Understanding

Any good project starts with a deep understanding of the customer's needs. Data mining projects are no exception and CRISP-DM recognizes this.

The *Business Understanding* phase focuses on understanding

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok

foundational project management activities that are universal to most projects:

1. **Determine business objectives:** You should first “thoroughly understand, from a business perspective, what the customer really wants to accomplish.” (**CRISP-DM Guide**) and then define business success criteria.
2. **Assess situation:** Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.
3. **Determine data mining 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.

While many teams hurry through this phase, establishing a strong business understanding is like building the foundation of a house – absolutely essential.

II. Data Understanding

Next is the *Data Understanding* phase. Adding to the foundation of *Business Understanding*, it drives the focus to identify, collect, and analyze the data sets that can help you accomplish the project goals. 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.

A common rule of thumb is that 80% of the project is data preparation.

This phase, which is often referred to as “data munging”, prepares the final data set(s) for modeling. It 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.

IV. Modeling

What is widely regarded as data science’s most exciting work is also often the shortest phase of the project. Here you’ll likely build and assess various models based on several different modeling techniques. 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

knowledge, the pre-defined success criteria, and the test design.

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 should continue iterating until they find a “good enough” model, proceed through the CRISP-DM lifecycle, then further improve the model in future iterations.

V. Evaluation

in

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

“Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.”

-CRISP-DM Guide

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.

in

Your organization's work might not end there. As a project framework, CRISP-DM does not outline what to do after the project (also known as "operations"). But if the model is going to production, be sure you maintain the model in production. Constant monitoring and occasional model tuning is often required.

Join 8,000+ readers who get monthly tips to manage **artificial intelligence** projects and products better

Enter your email address

Get email
updates

Is CRISP-DM Agile or Waterfall?

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok

implement it.

Waterfall: On one hand, many view CRISP-DM as a rigid waterfall process – in part because of its reporting requirements are excessive for most projects. Moreover, the guide states in the business understanding phase that “the project plan contains detailed plans for each phase” – a hallmark aspect of traditional waterfall approaches that require detailed, upfront planning.

Indeed, if you follow CRISP-DM precisely (defining detailed plans for each phase at the project start and include every report) and choose not to iterate frequently, then you’re operating more of a waterfall process.

Agile: On the other hand, CRISP-DM indirectly advocates agile principles and practices by stating: “The sequence of the phases is not rigid. Moving back and forth between different phases is always required. The outcome of each phase determines which phase, or particular task of a phase, has to be performed next.”

Thus if you follow CRISP-DM in a more flexible way, iterate quickly, and layer in other agile processes, you’ll wind up with an agile approach.

Example: To illustrate how CRISP-DM could be implemented in either an Agile or waterfall manner, imagine a churn project with three deliverables: a voluntary churn model, a non-pay disconnect churn model, and a propensity to accept a retention-focused offer.

CRISP-DM Waterfall: Horizontal Slicing

Learn more about slicing at [Vertical vs Horizontal Slicing Data Science](#)

In a waterfall style implementation, the team’s work would

One “big bang” deliverable is delivered at the end of the project.

in

Deployment			
Evaluation			
Modeling			
Data Preparation			
Data Understanding			
Business Understanding			
	Feature 1: Voluntary Churn	Feature 2: Non-Pay Churn	Feature 3: Offer Acceptance

CRISP-DM Agile: Vertical Slicing

Alternatively, in an agile implementation of CRISP-DM, the team would narrowly focus on quickly delivering one vertical slice up the value chain at a time as shown below.

They would deliver multiple smaller vertical releases and frequently solicit feedback along the way.

Deployment			
Evaluation			
Modeling			
Data Preparation			
Data Understanding			
Business Understanding			

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok

Which is Better?

When possible, take an agile approach and slice vertically so that:

- Stakeholders get value sooner
- Stakeholders can provide meaningful feedback
- The data scientists can assess model performance earlier
- The project team can adjust the plan based on stakeholder feedback

in

How Popular is CRISP-DM?

Definitive research does not exist on how frequently data science teams use different management approaches. So to get an idea on approach popularity, we investigated KDnuggets polls, conducted our own poll, and researched Google search volumes. Each of these views suggests that **CRISP-DM is the most commonly used approach** for data science projects.

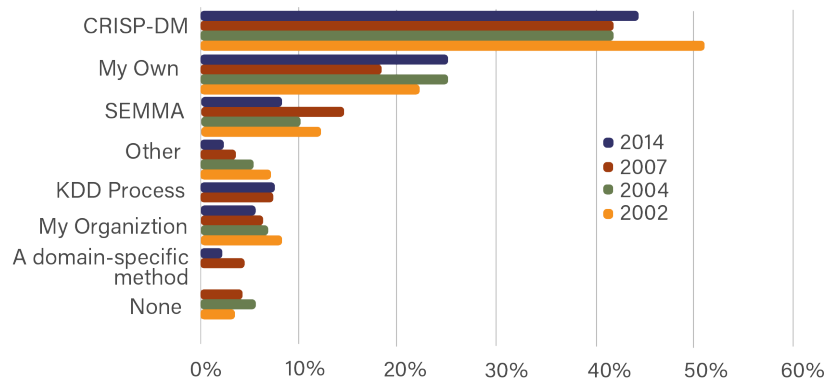
KDnuggets Polls

Bear in mind that the website caters toward data mining, and the data science field has changed a lot since 2014.

KDnuggets is a common source for data mining methodology usage. Each of the polls in 2002, 2004, 2007 posed the question: "What main methodology are you using for data mining?", and the 2014 poll expanded the question to include "...for analytics, data mining, or data science projects." 150-200 respondents answered each poll.

KDnuggets Polls

What main methodology are you using for data mining?



CRISP-DM was the popular methodology in each poll spanning the 12 years.

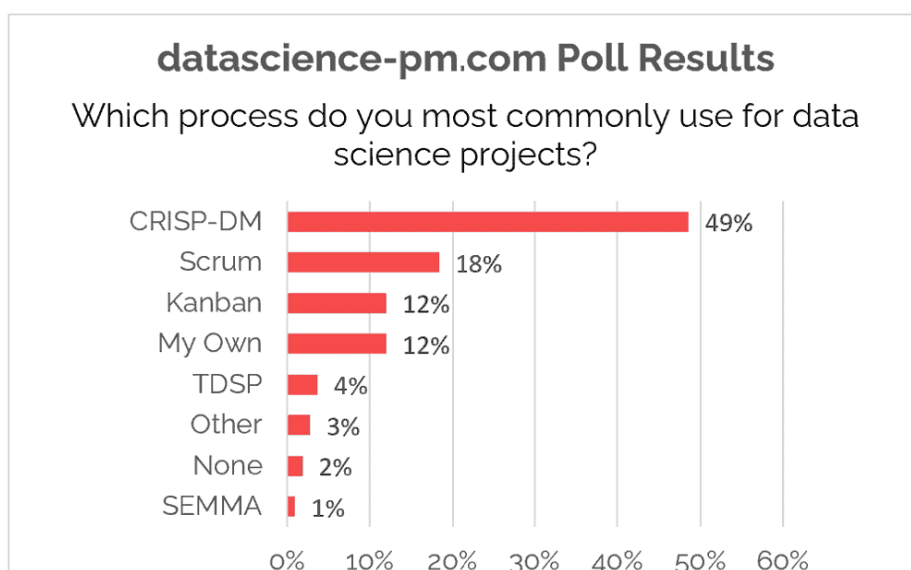
in

Our 2020 Poll

To learn more about the poll, go to [this post](#).

For a more current look into the popularity of various approaches, we conducted our own poll on this site in August and September 2020.

Note the response options for our poll were different from the KDnuggets polls, and our site attracts a different audience.



We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok

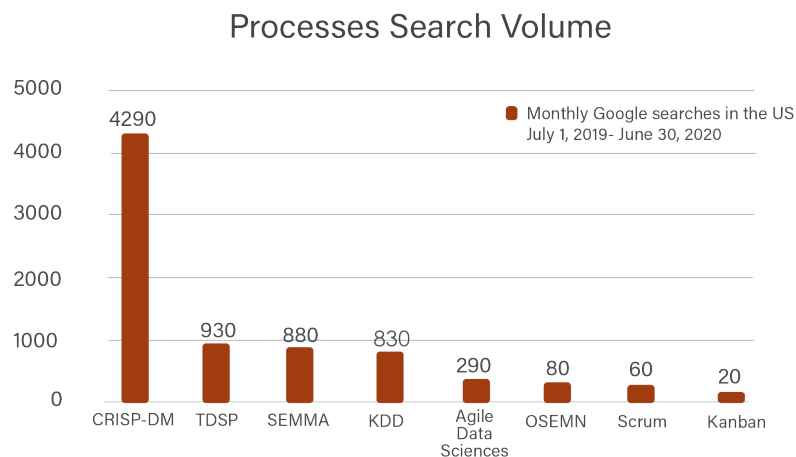
CRISP-DM was the clear winner, garnering nearly half of the 109 votes.

Google Searches

Given the ambiguity of a searcher's intent, some searches like "my own" could not be analyzed and others like "tdsp" and "semma" could be misleading.

For yet third view into CRISP-DM, we turned to Google Keyword Planner tool which provided the average monthly search volumes in the USA for select key search terms and related terms (e.g. "crispdm" or "crisp dm data science"). Clearly irrelevant searches like "tdsp electrical charges" or "semma both aagatha" were then removed.

in



CRISP-DM yet again reigned as king, and this time with a much broader margin.

Should I use CRISP-DM for Data Science?

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok

Like most answers in data science, it's kind of complicated.
But here's a quick overview.

Benefits

From today's data science perspective this seems like common sense. This is exactly the point. The common process is so logical that it has become embedded into all our education, training, and practice.

-William Vorheis, one of CRISP-DM's authors ([from Data Science Central](#))

in

- **Generalize-able:** 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](#)).
- **Common Sense:** When students were asked to do a data science project without project management direction, they "tended toward a CRISP-like methodology and identified the phases and did several iterations." Moreover, teams which were trained and explicitly told to implement CRISP-DM performed better than teams using other approaches ([Saltz, Shamsurhin, & Crowston, 2017](#)).
- **Adopt-able:** Like [Kanban](#), CRISP-DM can be implemented without much training, organizational role changes, or controversy.
- **Right Start:** 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.

- **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.

Weaknesses and Challenges

in

In a controlled experiment, students who used CRISP-DM “were the last to start coding” and “did not fully understand the coding challenges they were going to face”

–Saltz, Shamshurin, & Crowston, 2017

- **Rigid:** On the other hand, some argue that CRISP-DM suffers from the same weaknesses of **Waterfall** and encumbers rapid iteration.
- **Documentation Heavy:** 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.
- **Not Modern:** 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).
- **Not a Project Management Approach:** 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, &

For a more comprehensive view of recommendations view the [data science process post](#).

CRISP-DM is a great starting point for those who are looking to understand the general data science process. Five tips to overcome these weaknesses are:

- **Iterate quickly:** Don't fall into a waterfall trap by working thoroughly across layers of the project. Rather, think vertically and deliver thin vertical slices of end-to-end value. Your first deliverable might not be too useful. That's okay. Iterate.
- **Document enough...but not too much:** If you follow CRISP-DM precisely, you might spend more time documenting than doing anything else. Do what's reasonable and appropriate but don't go overboard.
- **Don't forgot modern technology:** Add steps to leverage cloud architectures and modern software practices like git version control and CI/CD pipelines to your project plan when appropriate.
- **Set expectations:** CRISP-DM lacks communication strategies with stakeholders. So be sure to set expectations and communicate with them frequently.
- **Combine with a project management approach:** As a more generalized statement from the previous bullet, CRISP-DM is not truly a project management approach. Thus combine it with a [data science coordination framework](#). Popular [agile](#) approaches include:
 - [Kanban](#)
 - [Scrum](#)
 - [Data Driven Scrum](#)

What are other CRISP-DM Alternatives?

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok

in

A few years prior to the publication of CRISP-DM, SAS developed *Sample, Explore, Modify, Model, and Assess* (**SEMMA**). Although designed to help guide users through tools in SAS Enterprise Miner for data mining problems, SEMMA is often considered to be a general data mining methodology. SEMMA's popularity has waned with only 1% of respondents in our 2020 poll stating they use it.

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.

See the main article for **SEMMA**

KDD and KDDS

Dating back to 1989, **Knowledge Discovery in Database** 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. There are different representations of KDD with perhaps the most common having five phases: *Select, Pre-Processing, Transformation, Data Mining*, and *Interpretation/Evaluation*. Like SEMMA, KDD is similar to CRISP but more narrowly focused and excludes the initial *Business Understanding* and *Deployment* phases.

In 2016, Nancy Grady of SAIC, published the **Knowledge Discovery in Data Science (KDDS)** describing it "as an end-to-end process model from mission needs planning to the delivery of value." KDDS specifically expands upon KDD and

build, and improve and five process stages: *plan, collect, curate, analyze, and act*.

KDD tends to be an older term that is less frequently used. KDDS never had significant adoption.

See the main article for *KDD and Data Mining Process*.

Join 8,000+ readers who get monthly tips to manage **artificial intelligence** projects and products better

in

Enter your email address

Get email
updates

Where can I learn more?

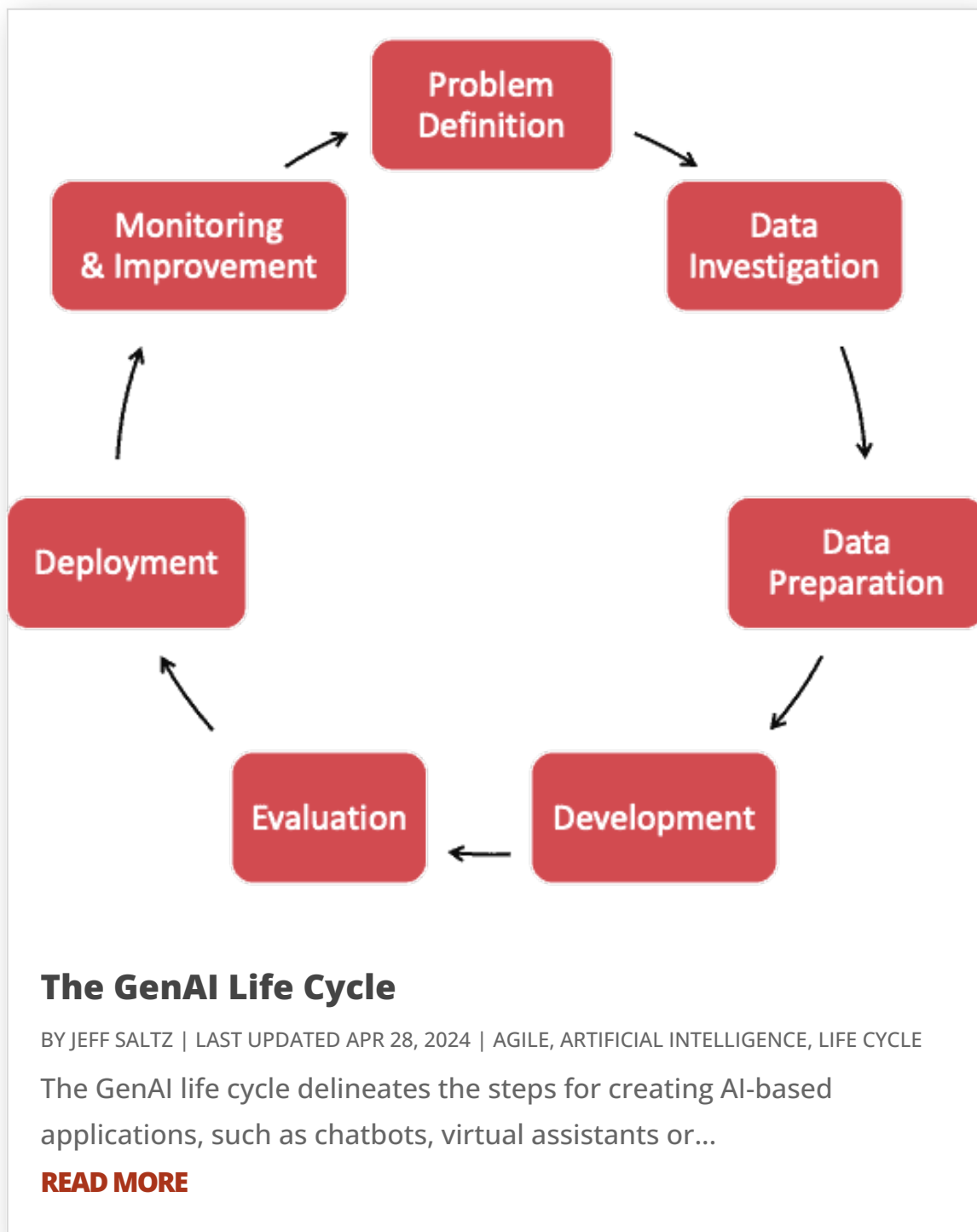
- Blog Post: [What is a Data Science Life Cycle?](#)
- Blog Post: [What is a Data Science Workflow?](#)
- Blog Post: [What is the Data Science Process?](#)
- Blog Post: [Steps to Define an Effective Data Science Process](#)
- Blog Post: [CRISP-DM for Data Science – 5 Actions to Consider](#)
- Blog Post: [CRISP-DM is still the most Popular Framework](#)
- Blog Post: [Data Science vs Software Engineering](#)
- Explore the [Consulting services](#) to learn CRISP and other processes
- (external): [Official CRISP-DM Guide](#)

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok

Explore Related Content

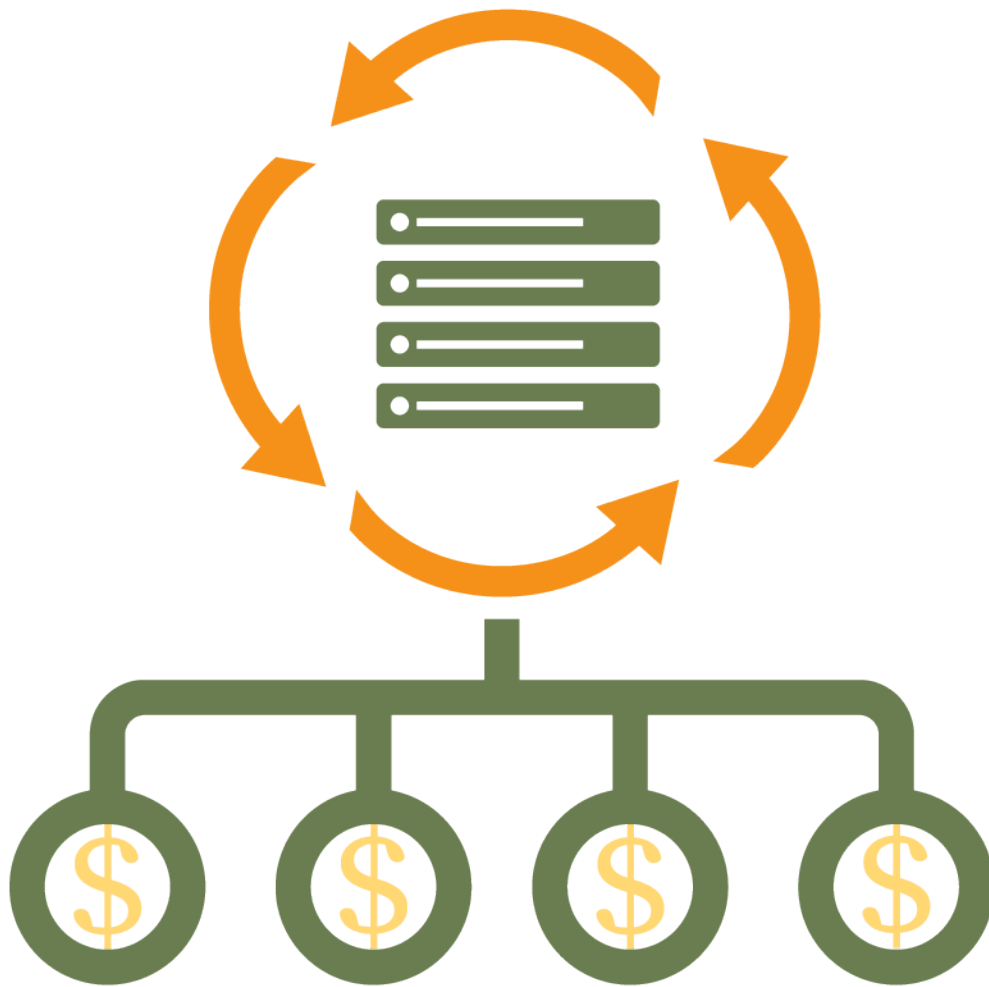
in



We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok

in



Managing Generative AI Projects

BY JEFF SALTZ | LAST UPDATED APR 10, 2024 | AGILE, ARTIFICIAL INTELLIGENCE, LIFE CYCLE, PROJECT MANAGEMENT

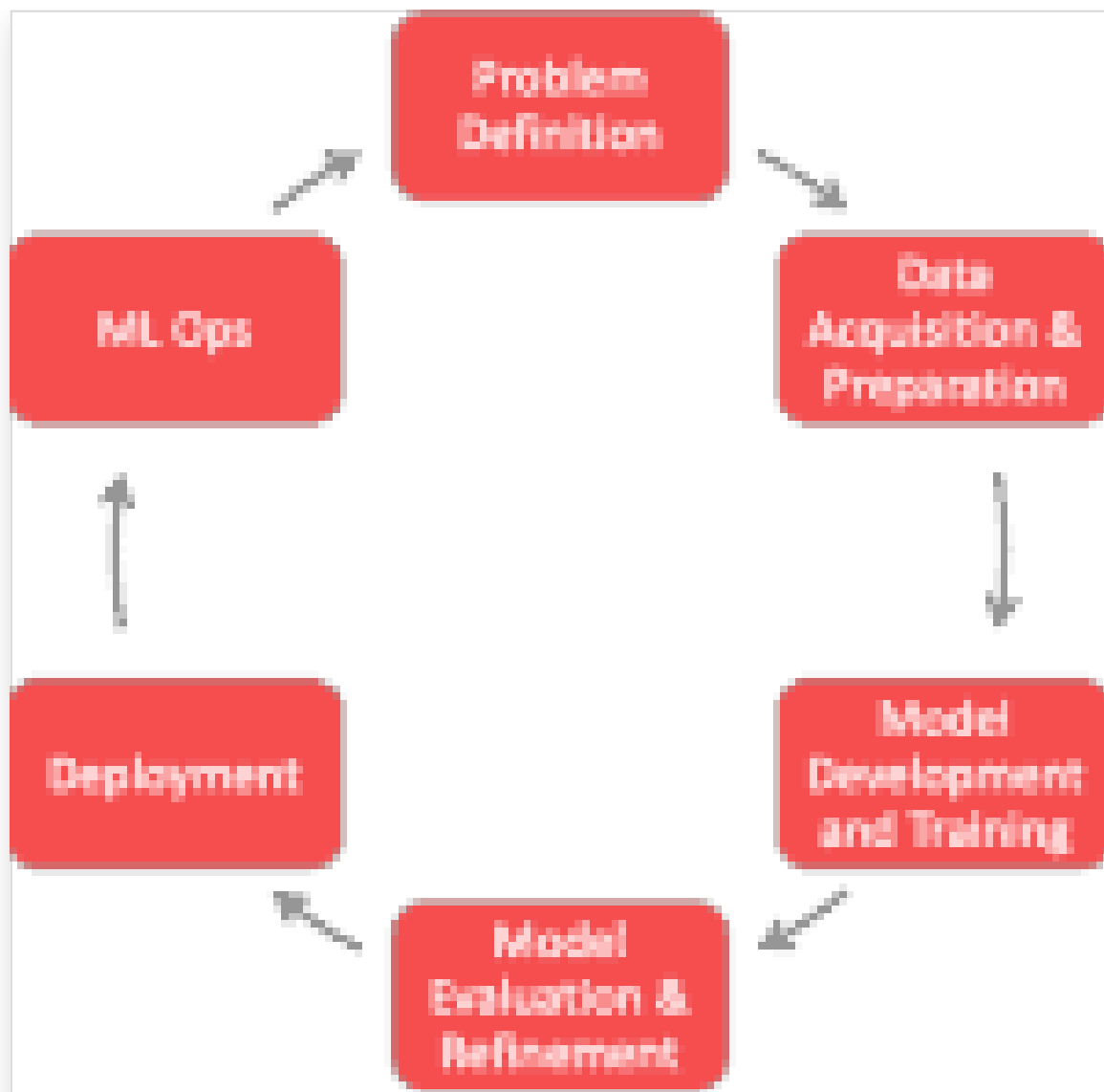
Not stopping at merely utilizing apps like ChatGPT, many companies are building, or exploring the possibility of...

[READ MORE](#)

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok

in



What is the AI Life Cycle?

BY JEFF SALTZ | LAST UPDATED MAR 31, 2024 | ARTIFICIAL INTELLIGENCE, LIFE CYCLE

In the rapidly evolving world of artificial intelligence (AI), project management can be as complex as the...

[READ MORE](#)

[« Older Entries](#)

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok

Frameworks

Life Cycles

CRISP DM

Microsoft TDSP

SEMMA

Agile Data Science

Kanban

Scrum

Data Driven Scrum

in

Popular Posts

What is a DS project manager?

Agile AI

What is an AI product manager?

DS documentation best practices

How to lead data science teams

The data science project checklist

Data science ethics courses

Managing data analytics projects

About DSPA

The Data Science Process Alliance helps individuals and teams apply effective project management techniques and frameworks to improve data science project outcomes

About Us

Terms of Service

Privacy Policy

Contact Us

Subscribe

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok

Email

Subscribe

Designed by Nue Hue LLC. Copyright 2024 @Data Science Process Alliance. All rights reserved.

in

We use cookies to ensure that we give you the best experience on our website. If you continue to use this site we will assume that you are happy with it.

Ok