**Cinco estruturas de Ciclo de Vida do Projeto Data Science**

Por: [Joji John](https://medium.com/@jojijohn_81482) - <https://medium.com/@jojijohn_81482>.

Feb 18

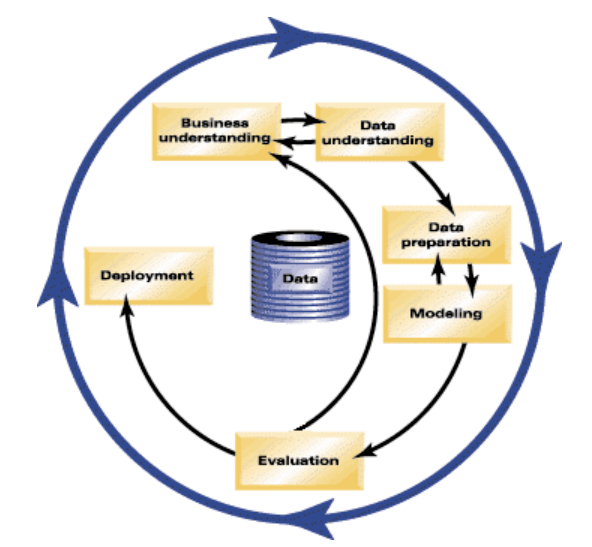
## 1.Cross-Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM é a abreviação de **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining, que pode ser traduzido como Processo Padrão Interindustrial para Mineração de Dados.

Originalmente, como visto na Figura 1, é um modelo de processo de *mineração de dados* que descreve abordagens comumente usadas por especialistas em mineração de dados para atacar problemas. (SHEARER, 2000).

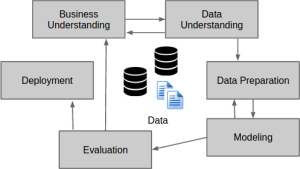
Concebido em 1996, provavelmente é o padrão mais antigo. Foi muito utilizado quando a Ciência de Dados (Data Science) era chamada de Mineração de Dados (Data Mining).

Figura 1. Framework CRISP-DM padrão.



O modelo CRISP-DM tradicionalmente define **seis etapas** no ciclo de vida de mineração de dados. A ciência de dados é semelhante à mineração de dados em vários aspectos, portanto, há alguma semelhança com essas etapas, como pode ser observado na Figura 2. (MANNA, 2014).

Figura 2. Framework CRISP-DM modificado para Ciência de Dados.



Os passos do modelo CRISP-DM, de acordo com a Figura 2, portanto, são:

1. Compreensão Empresarial
2. Entendimento dos Dados
3. Preparação de Dados
4. Modelagem
5. Avaliação e
6. Implantação

De acordo com Manna (2014), dado um certo nível de maturidade em grande volume de dados (big data) e de conhecimento em ciência de dados dentro da organização, é razoável assumir a existência e a disponibilidade de uma biblioteca de ativos relacionados a implementações de ciência de dados. Entre elas, destacam-se:

1. Biblioteca de casos de uso de negócios para aplicativos de big data / data science
2. Requisitos de dados - matriz de mapeamento de casos de uso de negócios
3. Requisitos mínimos de qualidade de dados (casos de teste para garantir um nível mínimo de qualidade de dados para garantir a viabilidade).

Na maioria das organizações, a ciência de dados é uma disciplina incipiente, portanto, os cientistas de dados (exceto aqueles com experiência atuarial) provavelmente terão conhecimento de domínio de negócios limitado - portanto, precisam estar emparelhados com pessoas de negócios e pessoas com experiência em entender os dados. Isso ajuda os cientistas de dados a obter ou trabalhar juntos nas etapas 1 e 2 do modelo CRISM-DM - ou seja, compreensão comercial e compreensão de dados. (MANNA, 2014).

## 2. Ciclo de Vida Padrão de Projetos de Ciência de Dados

Baseado no modelo CRISP-DM, um típico projeto de ciência de dados torna-se um exercício de engenharia em termos de uma estrutura definida de etapas ou fases e critérios de saída, que permitem tomar decisões cientificadas sobre a continuidade dos projetos com base em critérios predefinidos, otimizar a utilização dos recursos e maximizar os benefícios do projeto de ciência de dados. Isso também impede que o projeto seja transformado em um *poço de dinheiro* devido à busca de hipóteses e ideias inviáveis. (MANNA, 2014).

Considerado, segundo John (2019), como o framework mais utilizado em projetos de ciência de dados, o ciclo de vida *padrão* de ciência de dados, é portanto, uma versão ligeiramente aprimorada do CRISP-DM.

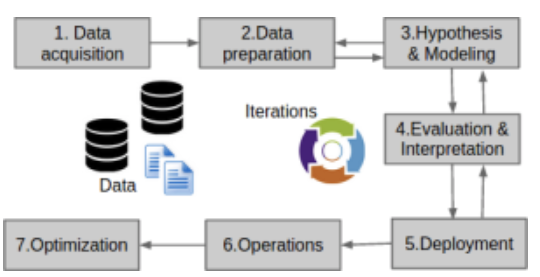
Os passos do modelo *padrão* de ciclo de vida de ciência de dados, de acordo com a Figura 3, são compostos de 7 passos:

1. Aquisição de dados;
2. Preparação de dados;
3. Hipótese e modelagem;
4. Avaliação e Interpretação;
5. Implantação;
6. Operações; e
7. Otimiza.

Em resumo, de acordo com Manna (2014), os passos são:

1. **Aquisição de dados** – envolve a aquisição de dados de fontes internas e externas, incluindo mídias sociais ou *web raspagem* (web scraping). Em uma situação estável, as rotinas de extração e transferência de dados estão em vigor, e novas fontes, uma vez identificadas, são obtidas seguindo os processos estabelecidos.
2. **Preparação de dados** – Esta fase é normalmente também chamada de "*data wrangling*" e envolve a **limpeza** **dos dados** e sua reformulação em uma forma prontamente utilizável para executar os processos de ciência de dados. Esta fase é semelhante, em determinados aspectos, às etapas tradicionais de ETL - um tipo de integração de dados em três etapas extração, transformação, carregamento, usado para combinar dados de diversas fontes – em *data warehousing*, mas envolve uma análise mais exploratória e destina-se principalmente à extração de recursos em formatos utilizáveis.

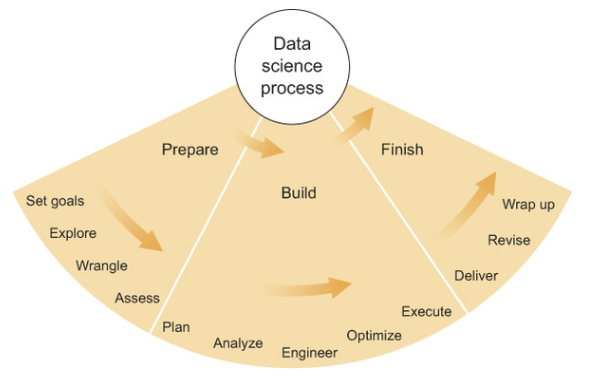
Figura 3. Ciclo de Vida do Projeto de Ciência de Dados.



1. **Hipótese e modelagem** são as etapas tradicionais de mineração de dados - no entanto, em um projeto de ciência de dados, elas não estão limitadas a *amostras estatísticas*. De fato, a ideia é aplicar técnicas de aprendizado de máquina em **todos** **os dados**.
2. **Avaliação e Interpretação -** Uma sub etapa importante realizada aqui é a de **seleção do modelo**. Esta etapa envolve a separação de um conjunto de *dados de treinamento* para **treinar os modelos de aprendizado de máquina** candidatos e também a preparação de um conjunto de *dados de validação e teste***,** com o objetivo de comparar o desempenho dos diversos modelos e selecionar aquele de melhor desempenho, avaliando sua precisão e impedindo o **overfitting** ([sobre-ajuste](https://pt.wikipedia.org/wiki/Sobreajuste) excessivo).
3. **Implementação –** Os *passos 2 a 4 são repetidos várias vezes, conforme necessário*. À medida que a compreensão dos dados e do negócio se tornam mais claros e os resultados dos modelos iniciais e hipóteses são avaliados, outros ajustes são realizados. Às vezes, eles podem incluir a Etapa 5 (implementação) e serem executados em um ambiente de pré-produção ou "piloto" antes da implantação em escala real de "produção" ou podem incluir ajustes rápidos *após* a implantação, com base no modelo de *entrega contínua* (**[continous delivery](https://gaea.com.br/o-que-e-continuous-delivery/)**).
4. **Operação** - Uma vez que o modelo tenha sido entregue para produção, é o momento de realizar manutenção e operações regulares. Essa fase de operações também pode seguir um modelo [DevOps](https://gaea.com.br/o-que-e-devops-conceito/) de destino que combina bem com o modelo de [*entrega contínua*](https://en.wikipedia.org/wiki/Continuous_delivery), dados os requisitos rápidos de tempo de comercialização (time-to-market) nos projetos de *big data*. Idealmente, a implantação inclui testes de desempenho para medir o desempenho do modelo e pode acionar alertas quando o desempenho do modelo se degrada além de um certo limite aceitável.
5. **Otimização -** A fase de otimização é a etapa final no ciclo de vida do projeto de ciência de dados. Esta fase pode ser ocasionada por desempenho deficiente ou devido à necessidade de se adicionar novas fontes de dados e de reciclar o modelo, ou mesmo de implantar versões aprimoradas do modelo com base em melhores algoritmos**.**

## 3. Prepare — Build — Finish Framework

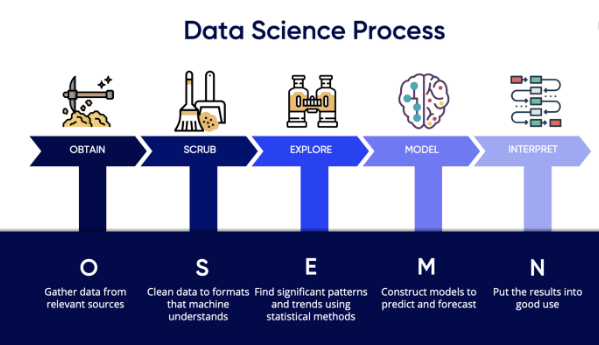
More Agile friendly framework

o

## 4. OSEMN Framework

O método OSEMN é a abreviatura para Obter (Obtain),

Variation of Standard Lifecycle. Added a stage for exploring data. Removed Deployment, operations & Optimization (why?)



I will walk you through this process using OSEMN framework, which covers every step of the data science project lifecycle from end to end.

## 1. Obtain Data

The very first step of a data science project is straightforward. We obtain the data that we need from available data sources.

In this step, you will need to query databases, using technical skills like [MySQL](https://www.mysql.com/) to process the data. You may also receive data in file formats like Microsoft Excel. If you are using [Python](https://www.python.org/) or [R](https://www.r-project.org/), they have specific packages that can read data from these data sources directly into your data science programs.

The different type of databases you may encounter are like [PostgreSQL](https://www.postgresql.org/), [Oracle](https://www.oracle.com/database/), or even non-relational databases (NoSQL) like [MongoDB](https://www.mongodb.com/). Another way to obtain data is to scrape from the websites using web scraping tools such as [Beautiful Soup](https://www.pythonforbeginners.com/python-on-the-web/web-scraping-with-beautifulsoup).

Another popular option to gather data is connecting to Web APIs. Websites such as Facebook and Twitter allows users to connect to their web servers and access their data. All you need to do is to use their Web API to crawl their data.

And of course, the most traditional way of obtaining data is directly from files, such as downloading it from [Kaggle](https://www.kaggle.com/)or existing corporate data which are stored in CSV (Comma Separated Value) or TSV (Tab Separated Values) format. These files are flat text files. You will need to use special Parser format, as a regular programming language like Python does not natively understand it.

## Skills required

To perform the tasks above, you will need certain technical skills. For example, for Database management, you will need to know how to use [MySQL](https://www.mysql.com/), [PostgreSQL](https://www.postgresql.org/) or [MongoDB](https://www.mongodb.com/) (if you are using a non-structured set of data).

If you are looking to work on projects on a much bigger data sets, or big data, then you need to learn how to access using distributed storage like [Apache Hadoop](https://hadoop.apache.org/), [Spark](https://spark.apache.org/) or [Flink](https://flink.apache.org/" \t "_blank).

## 2. Scrub Data

After obtaining data, the next immediate thing to do is scrubbing data. This process is for us to “clean” and to filter the data. Remember the “*garbage in, garbage out*” philosophy, if the data is unfiltered and irrelevant, the results of the analysis will not mean anything.

In this process, you need to convert the data from one format to another and consolidate everything into one standardized format across all data. For example, if your data is stored in multiple CSV files, then you will consolidate these CSV data into a single repository, so that you can process and analyze it.

In some situations, we will also need to filter the lines if you are handling locked files. Locked files refer to web locked files where you get to understand data such as the demographics of the users, time of entrance into your websites etc.

On top of that, scrubbing data also includes the task of extracting and replacing values. If you realise there are missing data sets or they could appear to be non-values, this is the time to replace them accordingly.

Lastly, you will also need to split, merge and extract columns. For example, for the place of origin, you may have both “City” and “State”. Depending on your requirements, you might need to either merge or split these data.

Think of this process as organizing and tidying up the data, removing what is no longer needed, replacing what is missing and standardising the format across all the data collected.

## Skills Required

You will need scripting tools like Python or R to help you to scrub the data. Otherwise, you may use an open-sourced tool like [OpenRefine](http://openrefine.org/" \t "_blank) or purchase enterprise software like [SAS Enterprise Miner](https://www.sas.com/en_my/software/enterprise-miner.html) to help you ease through this process.

For handling bigger data sets require you are required to have skills in Hadoop, [Map Reduce](https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html) or Spark. These tools can help you scrub the data by scripting.

## 3. Explore Data

Once your data is ready to be used, and right before you jump into AI and Machine Learning, you will have to examine the data.

Usually, in a corporate or business environment, your boss will just throw you a set of data and it is up to you to make sense of it. So it will be up to you to help them figure out the business question and transform them into a data science question.

To achieve that, we will need to explore the data. First of all, you will need to inspect the data and its properties. Different data types like numerical data, categorical data, ordinal and nominal data etc. require different treatments.

Then, the next step is to compute descriptive statistics to extract features and test significant variables. Testing significant variables often is done with correlation. For example, exploring the risk of someone getting high blood pressure in relations to their height and weight. Do note that some variables are correlated, but they do not always imply causation.

The term “*Feature*” used in Machine Learning or Modelling, is the data features that help us to identify the characteristics that represent the data. For example, “*Name*”, “*Age*”, “*Gender*” are typical features of members or employees dataset.

Lastly, we will utilise data visualisation to help us to identify significant patterns and trends in our data. We can gain a better picture through simple charts like line charts or bar charts to help us to understand the importance of the data.

## Skills Required

If you are using Python, you will need to know how to use Numpy, Matplotlib, Pandas or Scipy; if you are using R, you will need to use GGplot2 or the data exploration swiss knife Dplyr. On top of that, you need to have knowledge and skills in inferential statistics and data visualization.

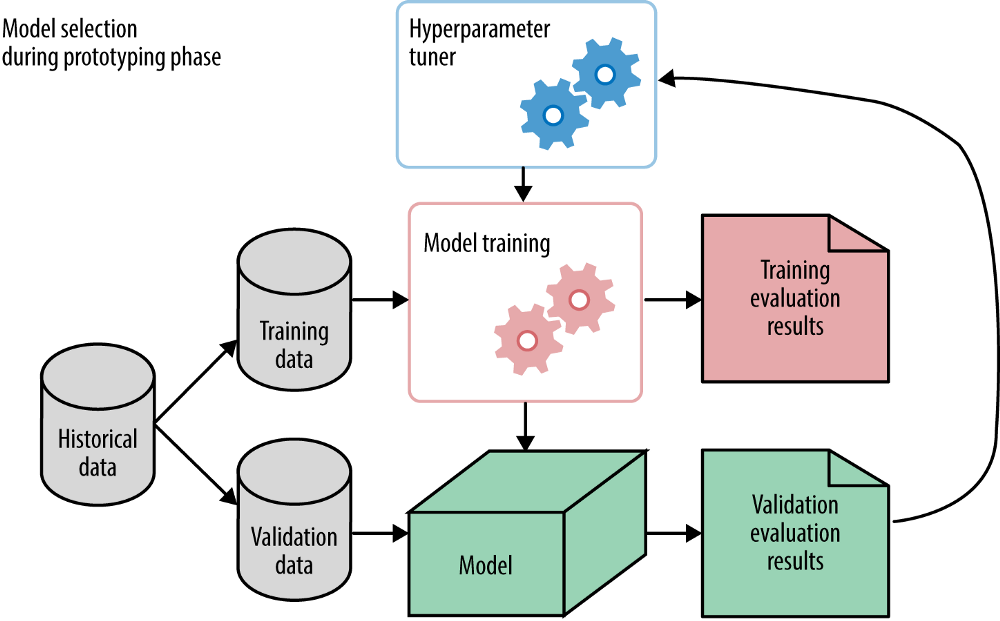
As much as you do not need a Masters or Ph.D. to do data science, these technical skills are crucial in order to conduct an experimental design, so you are able to reproduce the results.

## Additional Tips:

* Be curious. This can help you develop your spidey senses to spot weird patterns and trends.
* Focus on your audience, and understand their background and lingo. So that you are able to present the data in a way that makes sense to them.

## 4. Model Data

This is the stage where most people consider interesting. As many people call it “*where the magic happens*”.



Once again, before reaching this stage, bear in mind that the scrubbing and exploring stage are equally crucial to building useful models. So take your time on those stages instead of jumping right to this process.

One of the first things you need to do in modelling data is to reduce the dimensionality of your data set. Not all your features or values are essential to predicting your model. What you need to do is to select the relevant ones that contribute to the prediction of results.

There are a few tasks we can perform in modelling. We can also train models to perform classification to differentiating the emails you received as “*Inbox*” and “*Spam*” using logistic regressions. We can also forecast values using linear regressions. We can also use modelling to group data to understand the logic behind those clusters. For example, we group our e-commerce customers to understand their behaviour on your website. This requires us to identify groups of data points with clustering algorithms like k-means or hierarchical clustering.

In short, we use regression and predictions for forecasting future values, and classification to identify, and clustering to group values.

## Skills Required

In Machine Learning, the skills you will need is both supervised and unsupervised algorithms. For libraries, if you are using Python, you will need to know how to use [Sci-kit Learn](https://scikit-learn.org/); and if you are using R, you will need to use [CARET](https://topepo.github.io/caret/).

After the modelling process, you will need to be able to calculate evaluation scores such as precision, recall and F1 score for classification. For regressions, you need to be familiar with R² to measure goodness-of-fit, and using error scores like MAE (Mean Average Error), or RMSE (Root Mean Square Error) to measure the distance between the predicted and observed data points.

## 5. Interpreting Data

We are at the final and most crucial step of a data science project, interpreting models and data. The predictive power of a model lies in its ability to generalise. How do we explain a model depends on its ability to generalise unseen future data.

Interpreting data refers to the presentation of your data to a non-technical layman. We deliver the results in to answer the business questions we asked when we first started the project, together with the actionable insights that we found through the data science process.

Actionable insight is a key outcome that we show how data science can bring about predictive analytics and later on prescriptive analytics. In which, we learn how to repeat a positive result, or prevent a negative outcome.

On top of that, you will need to visualise your findings accordingly, keeping it driven by your business questions. It is essential to present your findings in such a way that is useful to the organisation, or else it would be pointless to your stakeholders.

In this process, technical skills only are not sufficient. One essential skill you need is to be able to tell a clear and actionable story. If your presentation does not trigger actions in your audience, it means that your communication was not efficient. Remember that you will be presenting to an audience with no technical background, so the way you communicate the message is key.

## Skills Required

In this process, the key skills to have is beyond technical skills. You will need strong business domain knowledge to present your findings in a way that can answer the business questions you set out to answer, and translate them into actionable steps.

Apart from tools needed for data visualization like Matplotlib, ggplot, [Seaborn](https://seaborn.pydata.org/), [Tableau](https://www.tableau.com/), [d3js](https://d3js.org/)etc., you will need soft skills like presentation and communication skills, paired with a flair for reporting and writing skills will definitely help you in this stage of the project lifecycle.

## Recap

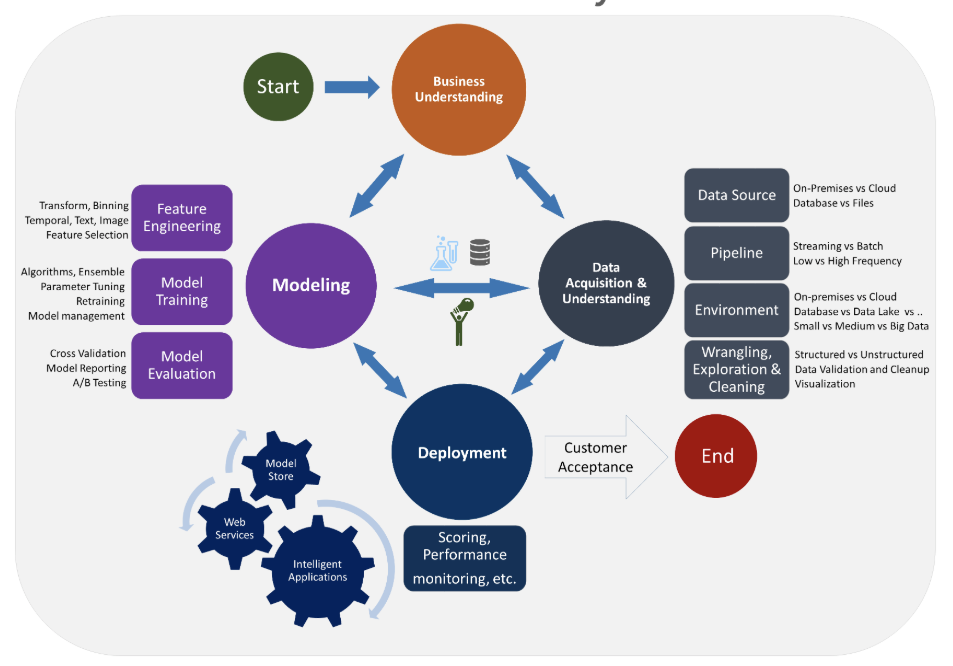
What I have presented here are the steps that data scientists follow chronologically in a typical data science project. If it is a brand new project, we usually spend about 60–70% of our time just on gathering and cleaning the data. Since it is a framework, you may use it as a guideline with your favorite tools.

The true north is always that business questions we defined, before even started the data science project. Always remember that solid business questions, clean and well-distributed data always beat fancy models.

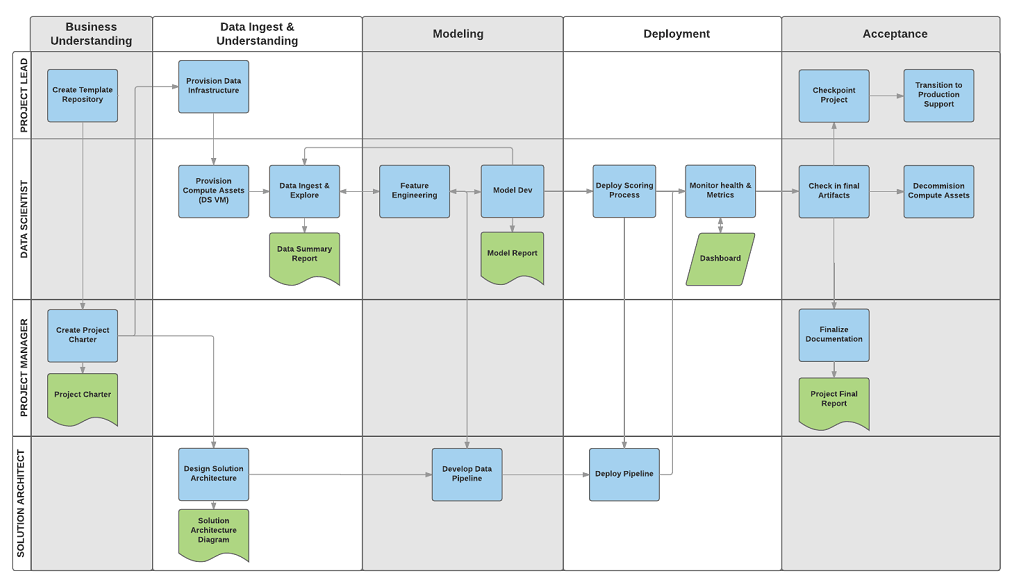
I hope you learned something today. Feel free to leave a message if you have any feedback, and share with anyone that might find this useful.

## 5. The Team Data Science Process (TDSP) lifecycle

Published by Microsoft in 2017. At the core it is very similar to CRISP-DM, but perceived Data Science as a collaborate team excersice and gave the details on the tasks and deliverables for each stages and roles.



The following diagram provides a grid view of the tasks (in blue) and artifacts (in green) associated with each stage of the lifecycle & roles



Bibliografia

Shearer. C. The CRISP-DM model: the new blueprint for data mining. J Data Warehousing 2000; 5:13—22.

Manna, Maloy. The data science project lifecycle, Data Science Central, December 18, 2014. Disponível em: <https://www.datasciencecentral.com/profiles/blogs/the-data-science-project-lifecycle>. Acesso em: 09` de abril de 2019.