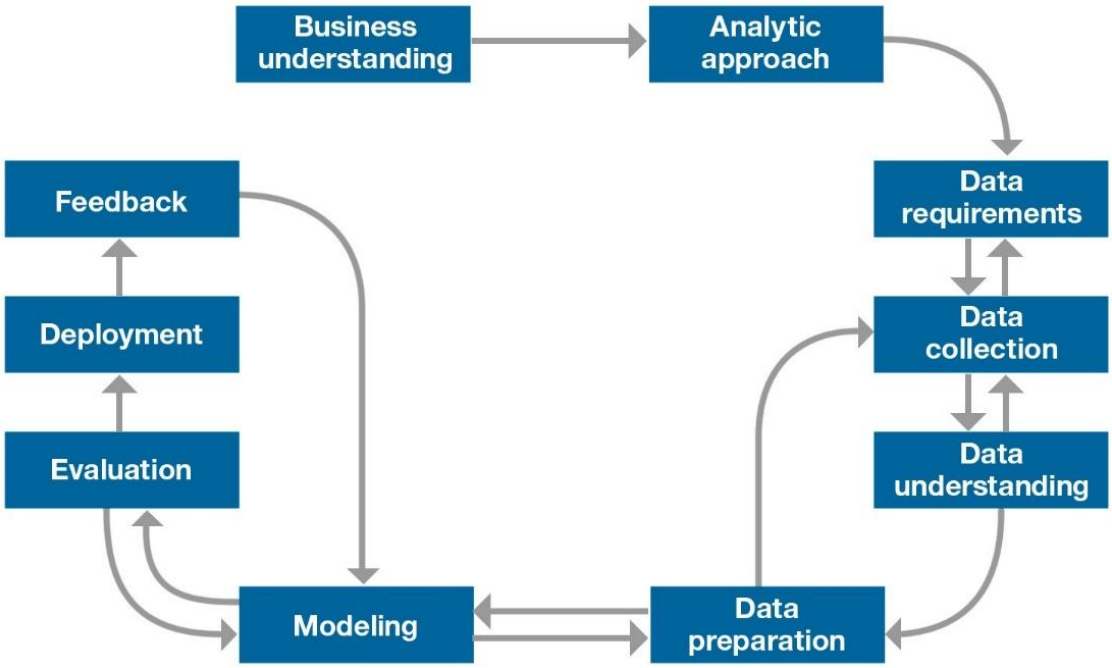
Data Science Methodology - How to design your data science project

Data Science Methodology Series...!!!

Data Science methodology is one the most important subject to know about any data scientist, I have stuck so many times when I was thinking about this problem and always though, like mad man how can data science cycle run and big company's design methodology for data science. my search is completed when I reached out this one of the amazing course of this on Coursera. This is one the best methodology convert your data science, business problem to data science solution. You can learn a whole project cycle here. I write all my learning from this course. After reading this you will know about how to convert business problem to Data Science base Solutions.

Outline of this article

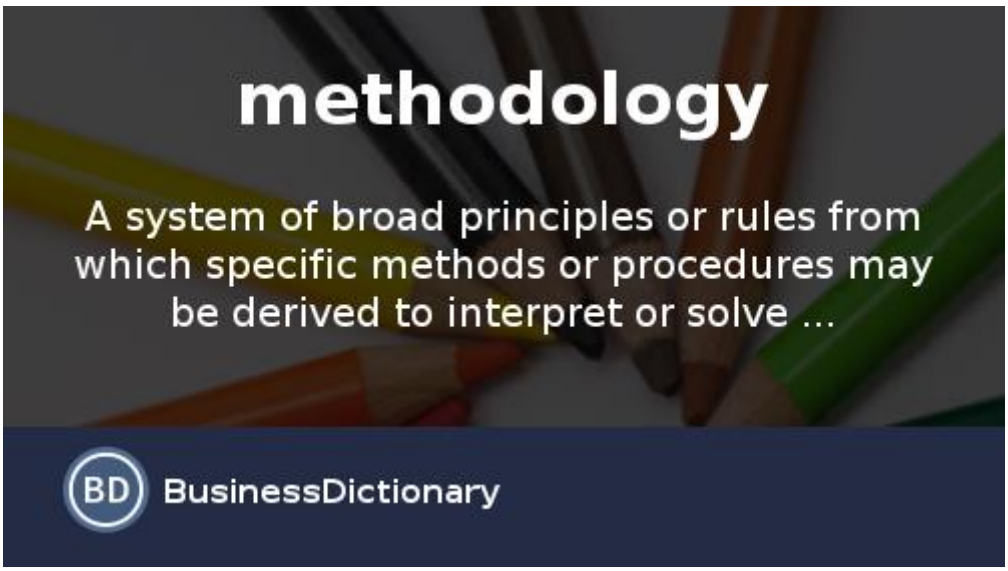


1. **Business understanding**
2. **Analytic approach**
3. **Data requirements**
4. **Data collection**
5. **Data understanding**
6. **Data preparation**
7. **Model Training**
8. **Model Evaluation**
9. **Deployment**
10. **Feedback**

Toward Data Science methodology

Welcome to Data Science Methodology 101! This is the beginning of a story that you will tell others in the years to come. It will not be as you experience it here, but through the stories you share with others as you explain how your understanding of a question led to an answer that changed the way in which something was done. Despite the increased computing power and access to data in recent decades, our ability to use data in the decision-making process is lost or not maximized too often. We do not have a solid understanding of questions that are asked and how the data is correctly applied to the problem in question.

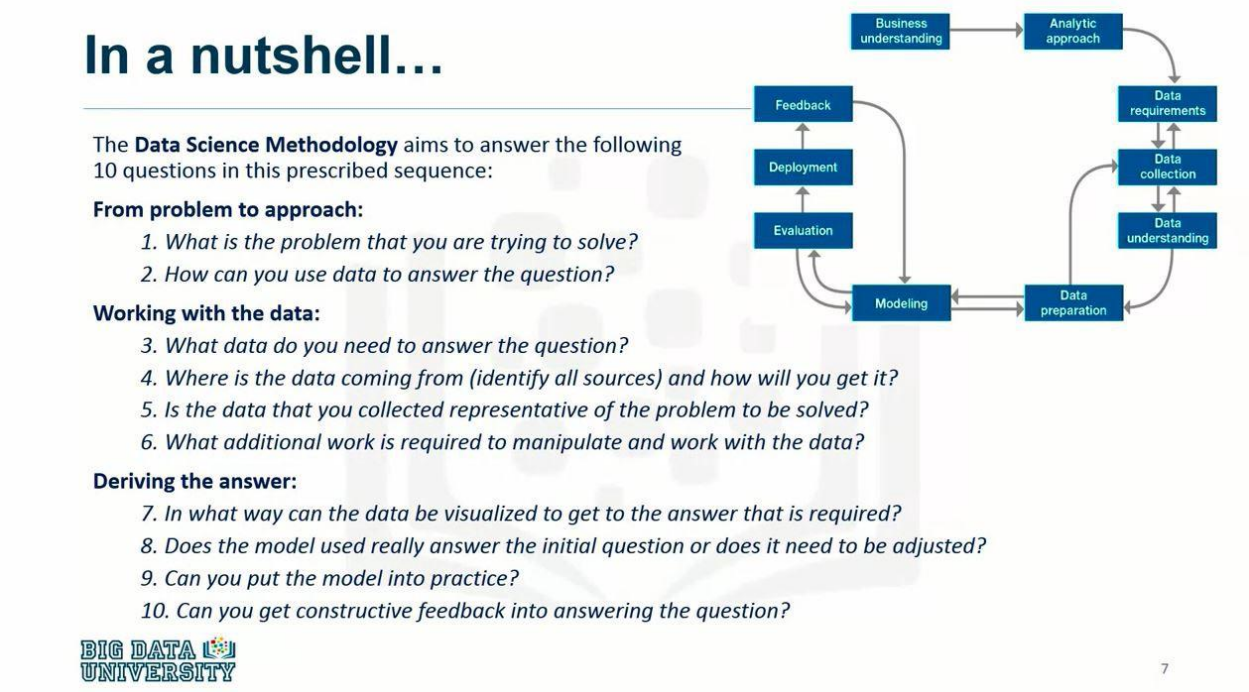
That why methodology come into the picture to design any problem.



Source : Business Dictionary

Here is a definition of the word methodology. It is important to think about it, because the temptation is often great to circumvent the methodology and go directly to the solutions. However, this prevents our best intentions from trying to solve a problem.

Data Science Methodology and Question



The Data science methodology aims to answer 10 basic questions in a given order. As you can see on above image,

1. **Two questions define the problem and determine the approach to use.**
2. **Four questions, you can ask the organization for the data you need.**
3. **Final questions to review the data and how you do it based on four additional questions.**

Take a moment to familiarize yourself with the ten questions that are critical to your success.

This article Series contain 5 modules:

1. **From Problem to Approach**
2. **From Requirement to Collection**
3. **From Understanding to Preparation**
4. **From Modelling to Evaluation**
5. **From Deployment to Feedback**

Now We are Focusing this Article :

#1) Business understanding

What is problem you trying to solve?

Every project, whatever its size, begins with the understanding of the business that forms the basis of an effective solution to the business problem. Business partners who need the analytics solution play a critical role in this phase by defining the problem, the project objectives, and the solution requirements from a business perspective. This is first step for any data science methodology.

#2) Analytic approach

- **How can you use the data to answer the question?**

Once a business problem has been clearly identified, the **Data Scientist** can define the **analytical approach.** To do this, the problem must be expressed in the context of statistical learning and machine learning techniques so that the Data Scientist can identify the techniques to achieve the desired result.

#3) Data requirements

- **What data do you need to answer the question?**

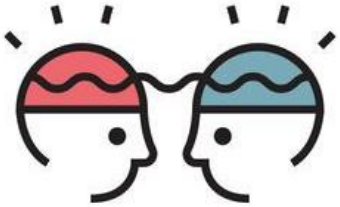
Analytic approach determines the **data requirements** because the **methods of analysis** to be used require specific content, formats, and data representations, based on domain knowledge.

#4) Data collection

- **Where is the data coming from (identify all sources) and how will you get it?**

The **Data Scientist identifies and collects data resources** (***structured, unstructured and semi-structured***) that are relevant to the problem area. If the **data scientist** finds gaps in the data collection, he may need to review the data requirements and collect more data.

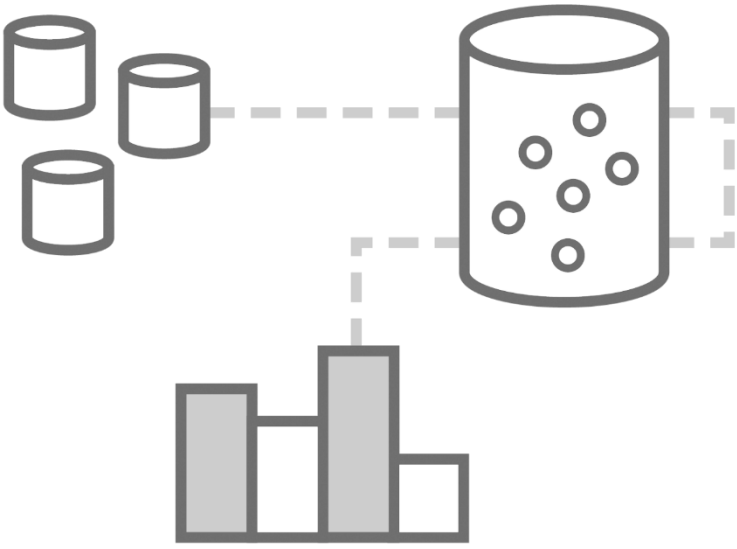
#5) Data understanding



- **Is the data that you collected representative of the problem to be solved?**

Descriptive statistics and visualization techniques can help a data scientist understand the content of the data, assess its quality, and obtain initial information about the data. A recovery from the previous step, data collection, may be necessary to fill the gaps in understanding.

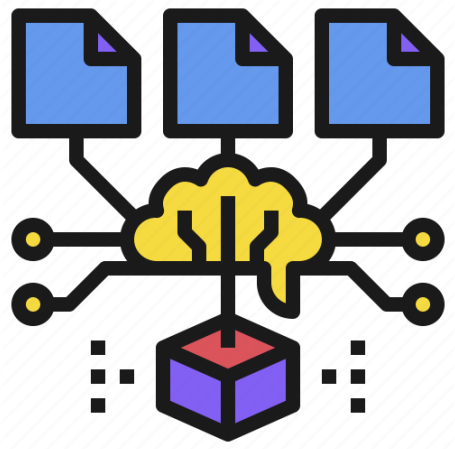
#6) Data preparation



- **What additional work is required to manipulate and work with the data?**

The **Data preparation** step includes all the activities used to **create the data set** used during the **modeling phase**. This includes **cleansing data, combining data from multiple sources, and transforming data into more useful variables**. In addition, **feature engineering and text analysis** can be used to derive **new structured variables** to enrich all predictors and improve model accuracy. **The Data preparation phase is the longest.** Although I have seen that it represents **90% of the total duration of the project**, this figure is usually **70%**. However, **it can go down as much as 50% if the data resources are well managed, well integrated, and analytically clean, not just storage.** Automating some phases of **Data preparation can further reduce the percentage**: *Telecommunications marketing* team members once told me that this team has cut the average time it takes to create and implement promotions from three months to three weeks.

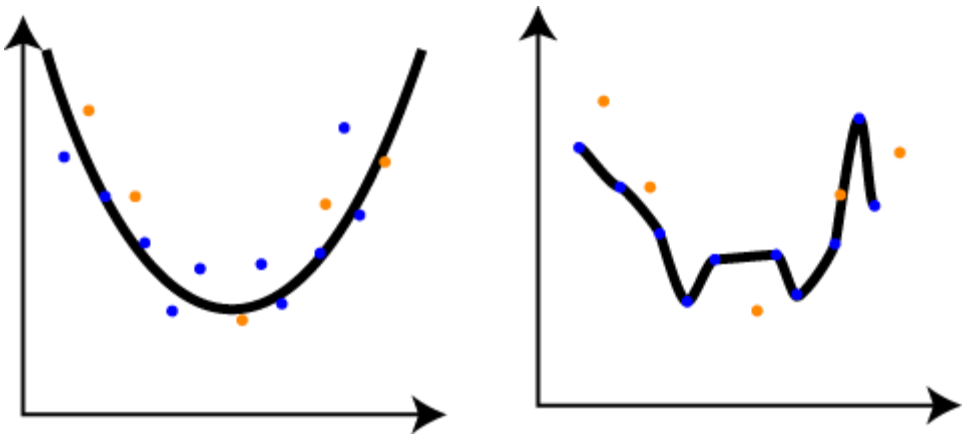
#7) Model Training



- **In What way can the data be visualized to get the answer that is required?**

From the first version of the **prepared data set**, **Data scientists use a Training data set**(***historical data in which the desired result is known***) to develop **predictive or descriptive** models using the described analytical approach previously. The modeling process is very iterative. It may be vary with different situation as per problem.

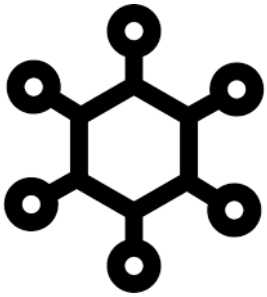
#8) Model Evaluation



- **Does the model used really answer the initial question or does it need to be adjusted?**

The Data Scientist evaluates the quality of the model and verifies that the business problem is handled in a complete and adequate manner. To do this, several diagnostic measures and other results, such as tables and graphs, must be calculated using a set of predictive model tests.

#9) Deployment



- **Can you put the model into practice?**

Once a satisfactory model has been developed and approved by commercial sponsors, it will be implemented in the production environment or in a comparable test environment. Such deployment is often initially limited to allow for performance evaluation. Implementing a model in an operational business process generally involves multiple groups, capabilities, and technologies.

#10) Feedback

- **Can you get constructive feedback into answering the question?**

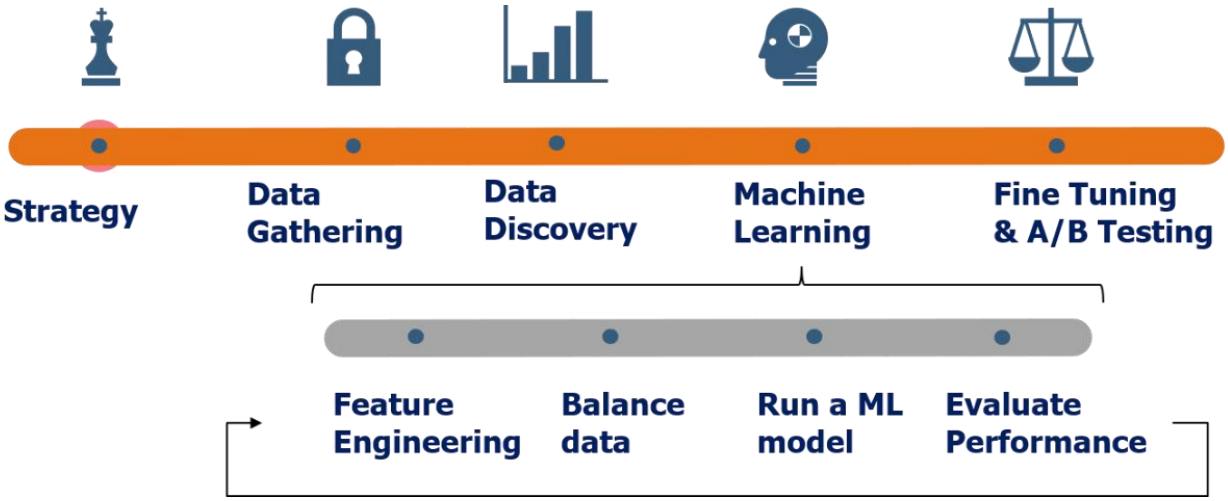
By collecting the ***results of the implemented model, the organization receives feedback on the performance of the model and its impact on the implementation environment.*** By analyzing this information, the data scientist can refine the model, increasing its accuracy and, therefore, its utility.

This phase, often neglected, can have significant additional benefits when carried out as part of the overall process. The flow of this methodology illustrates the iterative nature of the problem-solving process.

I hope you will get the basic understanding of process cycle. How to think on each and every stage that help to direct toward your successful methodology for your Data science project.

Part-1 Data Science Methodology- From Problem to Approach

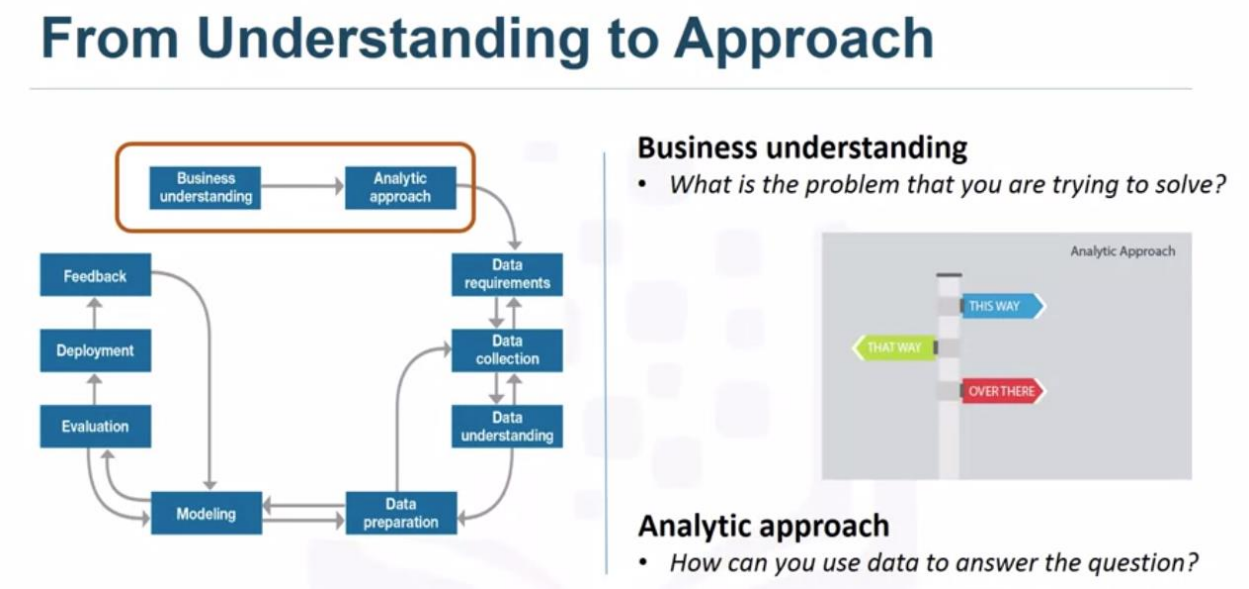
From Problem to Approach...!!!



Data Science methodology I have described basic with the all important question like which question you have to ask on which stage if you haven't read that article and already read here I have explained another module wise process of the same course.

Did that happen to you? Your boss invited you to a meeting in which you were informed about an important task that you should absolutely respect within a very short period of time. They both come and go to make sure all aspects of the task have been taken into account and that the meeting ends with both assurances that things are on the right track. Later in the afternoon, after spending some time investigating the various issues, he realizes that he needs to ask several more questions to truly accomplish his task.

Unfortunately, the boss is not available until tomorrow morning. Now, with the tight deadline in his ears, he begins to feel a sense of excitement. So what are you doing? Do you take the risk or stop to ask for clarification?



Data Science methodology I have described basic with the all important question like which question you have to ask on which stage if you haven't read that article and already read here I have explained another module wise process of the same course.

Article Series:

1. **Overview of Data Science Methodology**
2. **Part-1 Data Science Methodology from Problem to Approach**
3. **Part-2 Data Science Methodology from Requirement to Collection**
4. **Part-3 Data Science Methodology from Understanding to Preparation**
5. **Part-4 Data Science Methodology from Modelling to Evaluation**
6. **Part-5 Data Science Methodology from Deployment to Feedback**

Did that happen to you? Your boss invited you to a meeting in which you were informed about an important task that you should absolutely respect within a very short period of time. They both come and go to make sure all aspects of the task have been taken into account and that the meeting ends with both assurances that things are on the right track. Later in the afternoon, after spending some time investigating the various issues, he realizes that he needs to ask several more questions to truly accomplish his task.

Unfortunately, the boss is not available until tomorrow morning. Now, with the tight deadline in his ears, he begins to feel a sense of excitement. So what are you doing? Do you take the risk or stop to ask for clarification?

Outline of the Article :

- **Module 1: From Problem to Approach**
  - Business Understanding – Concepts & Case Study
  - Analytic Approach – Concepts & Case Study
  - Hands-on Lab & Quiz
- **Module 2: From Requirements to Collection**
  - Data Requirements – Concepts & Case Study
  - Data Collection – Concepts & Case Study
  - Hands-on Lab & Quiz
- **Module 3: From Understanding to Preparation**
  - Data Understanding – Concepts & Case Study
  - Data Preparation – Concepts
  - Data Preparation – Case Study
  - Hands-on Lab & Quiz
- **Module 4: From Modeling to Evaluation**
  - Modeling – Concepts
  - Modeling – Case Study
  - Evaluation – Concepts & Case Study
  - Hands-on Lab & Quiz
- **Module 5: From Deployment to Feedback**
  - Deployment – Concepts & Case Study
  - Feedback – Concepts & Case Study
  - Quiz

#1) Business Understanding

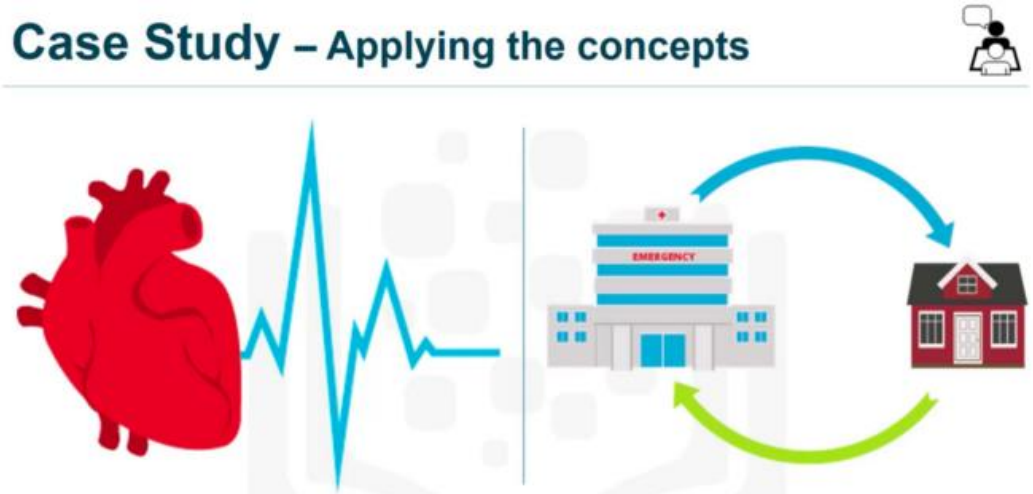
The methodology of data science begins with the search for clarifications in order to achieve what can be called business understanding. This understanding is at the beginning of the methodology because you can determine which data to answer the central question by clarifying the problem to be solved.

- Too often, much effort is spent on answering the question people worry about. Although the methods of solving this question may be useful, they do not solve the problem. Setting a clearly defined question begins with understanding the purpose of the person asking the question.
- **For example**, if a business owner asks, "How can we lower the cost of an activity?" We need to understand if the goal is to improve the efficiency of the activity. Or should the profitability of companies be increased? Once the goal is clear, the next piece of the puzzle determines the goals that support it. The breakdown of objectives can lead to structured discussions that set priorities that can help to organize and plan how to deal with the problem.

Depending on the problem, different stakeholders should participate in the discussion to identify the requirements and clarify the problems.

Case Study:

Let's take a look at the case study on the application of business understanding. The case study asks the following questions:



- ***How to best divide the limited health budget into optimal use to provide quality care? ...*** This issue has become a hot topic for an American insurer.

As public funds for readmission declined, the *insurance company ran the risk of offsetting the difference in costs, which could lead to higher costs for its clients.*

- Knowing that **higher insurance rates would not be popular**, the insurance company contacted local health authorities and **hired Data Science Expert** to learn how data science could be applied. the question. Before ***we could start collecting data, we had to define the objectives***. After spending time setting goals, the team prioritized "**patient readmission**" as an effective area for review.

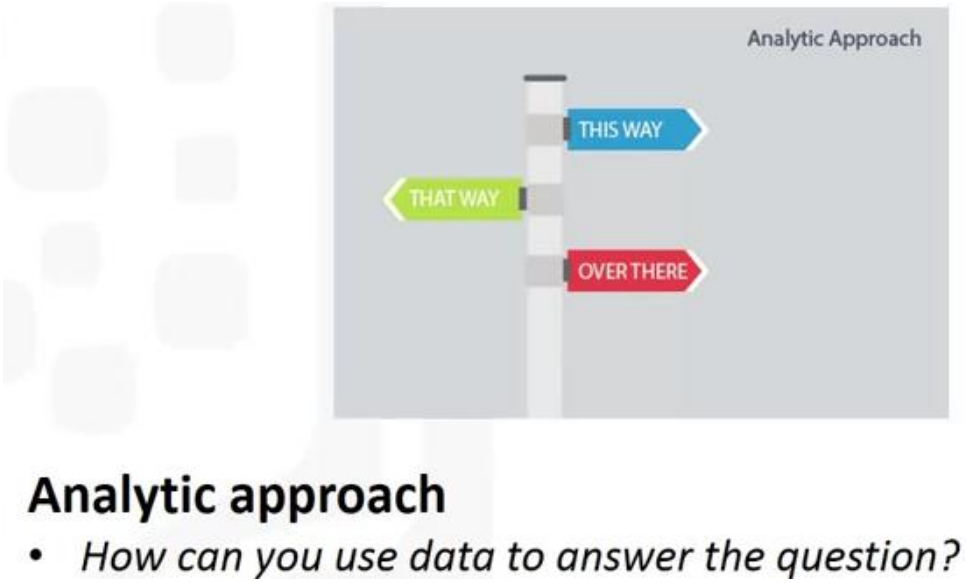
- Taking into account the objectives, it was found that approximately 30% of those who completed the rehabilitation treatment would be reintegrated into a rehabilitation center within one year. and that 50% would be resumed within

five years. After reviewing some records, it was found that patients with heart failure were high on the list of readmission.

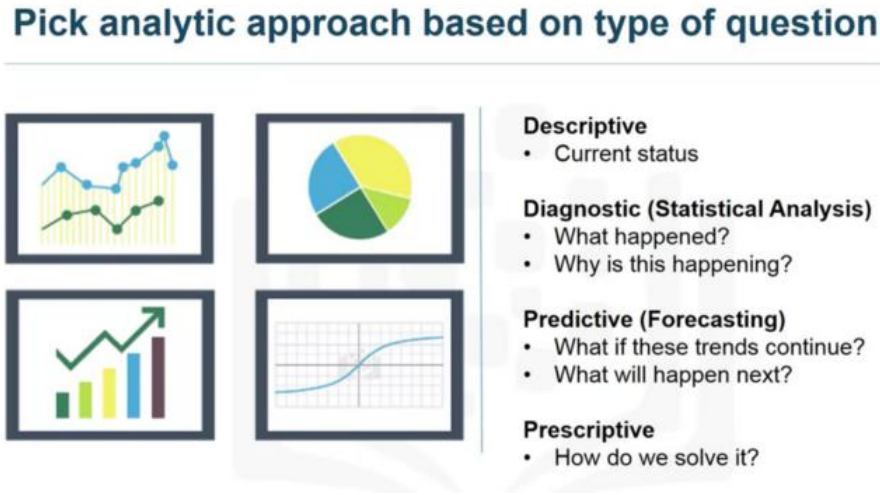
- It has also been found that a ***decision tree model*** can be applied to investigate this scenario to determine the reason for this phenomenon. To gain the business insight that will assist the analysis team in formulating and implementing their first project, ***Data scientists proposed and organized a workshop on-site.***
- The involvement of key commercial sponsors throughout the project has been essential as a sponsor: setting the overall direction. He remained committed and advised. If necessary, he got the necessary support. Finally, four business requirements were identified for each model built.

Namely: predict readmission results for patients with heart failure, predict the risk of readmission. Understand the combination of events that led to the expected result. Apply a process that is easy for new patients to understand because of their risk of readmission.

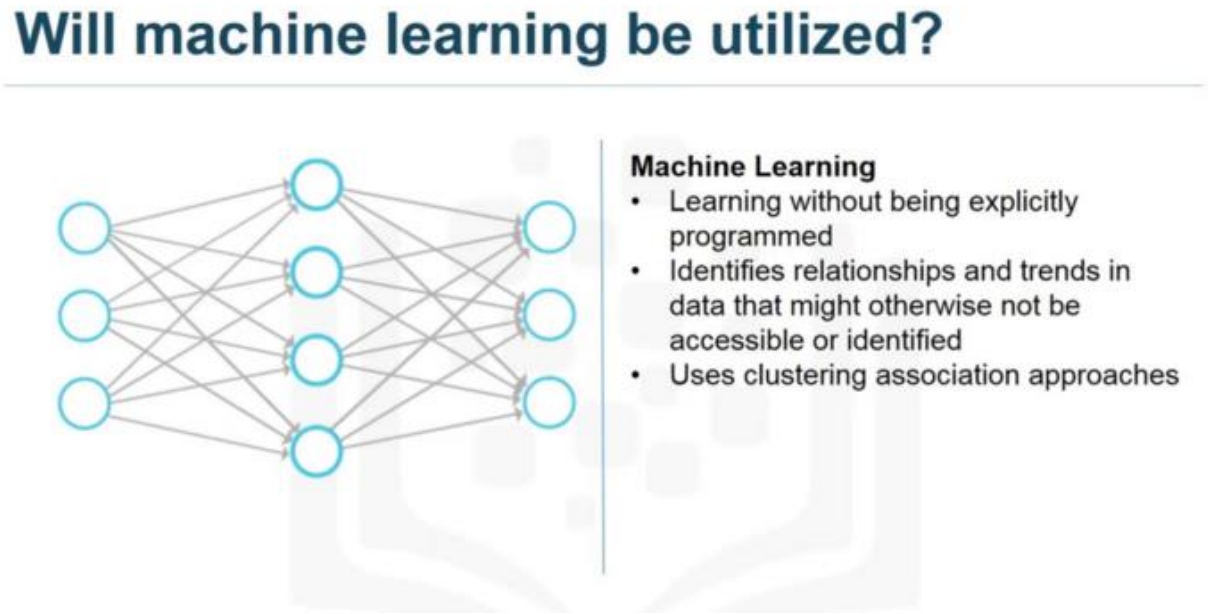
#2) Analytical Approach



- Choosing the **Right analytical approach** depends on the question asked. The approach is to ask the person asking the question to clarify the most appropriate form or approach. here we can understand **second stage of data science methodology.** Once the problem to be addressed is defined, the appropriate analytical approach is selected in the context of the needs of the enterprise. This is the second step in the methodology of data science.



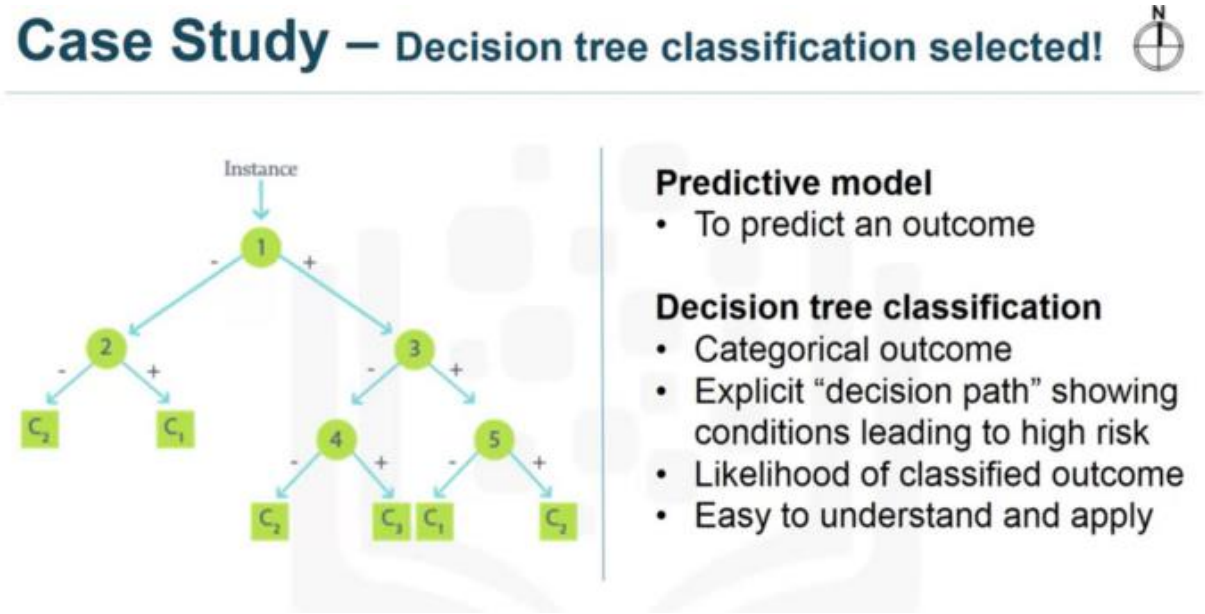
- Once a Deep understanding of the question is established, the analytical approach can be selected. This means identifying what type of pattern is needed to address the problem more effectively.
- When it comes to determining the probabilities of an action, a predictive model can be used.
- When it comes to identifying relationships, a descriptive approach may be necessary. This would be one that analyzes similar activity groups based on events and preferences.



- ***Statistical analysis refers to problems that require accounts. For example, if the question requires a yes / no answer, a classification approach to predicting a response is appropriate. Machine learning is a field of study in which computers can learn without being explicitly programmed. Machine learning can be used to identify relationships and trends in data that would otherwise be inaccessible or identified.***

Case Study :

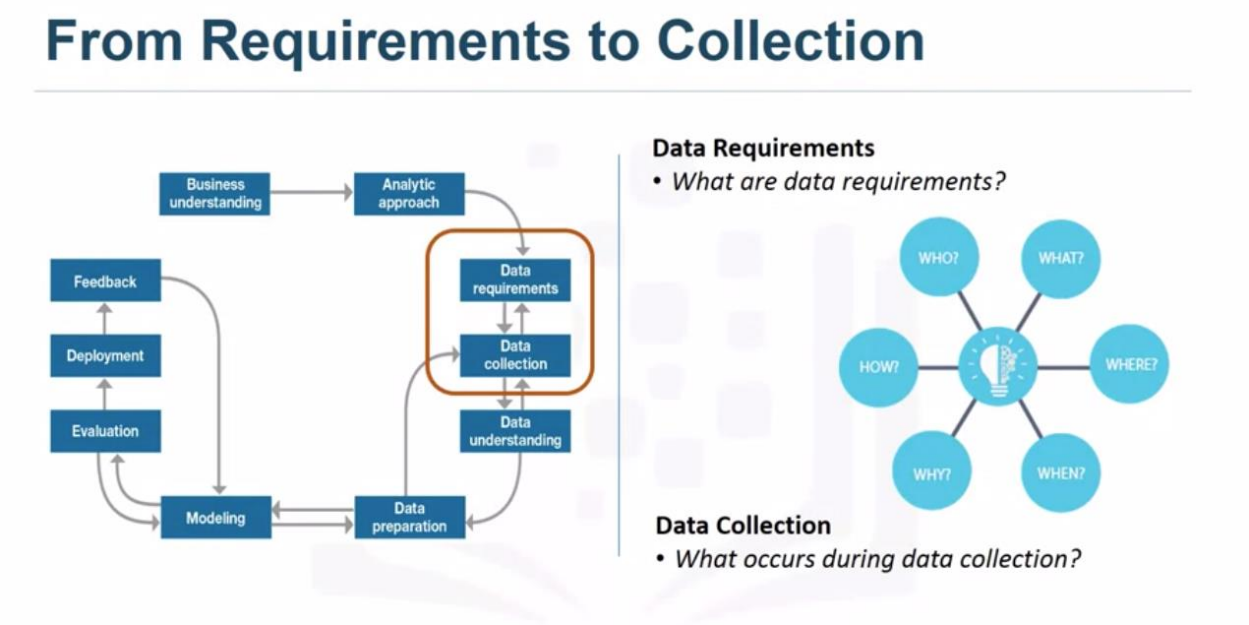
- In the case where the question about human behavior is asked, it would be an appropriate response to use clustering approaches. Let us now examine the case study on the application of the analytical approach. For the case study, a decision tree classification model was used to identify the combination of conditions that resulted in the results of each patient.



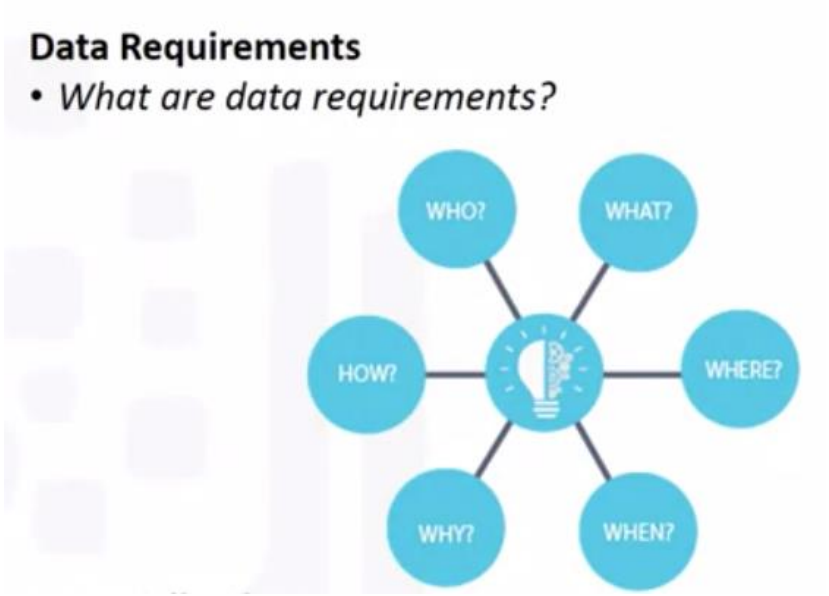
- In this approach, examining the variables in each of the nodes along each path of a leaf resulted in a corresponding threshold. ***This means that the decision tree classifier returns both the expected result and the probability of that result, based on the proportion of the dominant result, yes or no, in each group.*** From this information, analysts can derive the ***risk of readmission*** or ***the probability of a yes for each patient.***
- If the **dominant** result is **yes**, the **risk** is **simply the proportion of patients with yes on the sheet**. Otherwise, **the risk is 1 minus the proportion of a patient on the leaf.** A *decision tree classification model is easy to understand and apply to non-data scientists to assess the risk of readmitting new patients.*
- Doctors can easily identify under which conditions a patient is considered to be at risk, and during hospitalization, multiple models can be designed and used at different times.
- This provides a moving picture of the ***patient's risk and its evolution in the various treatments used***. For these reasons, the **decision tree classification approach was chosen to create the cardiac failure readmission model**.

Part-2 Data Science Methodology - From Requirement to Collection

From Requirement to Collection...!!!



#1) Data Requirement



Imagine that, If your goal is to prepare a spaghetti dinner, but you don't have the right ingredients for this dish, your success will be affected.

Think of this section of data science methodology as cooking with data. Each step is essential for the preparation of the meal.

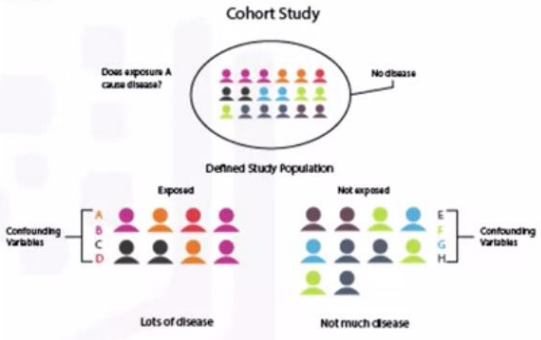
So, if the problem to be solved is, so to speak, the recipe and the data are an ingredient, the data scientist must identify the necessary ingredients, how to obtain or collect them, how to understand or use them, and how to obtain them. data. Ready to achieve the desired result.

- Based on the understanding of the problem and the analytic approach chosen, the data scientist is ready to begin. Let's look at some examples of the data needs of the data science methodology. Before the methodology data collection and processing steps are performed, it is important to define the data requirements for the classification of the decision tree.
- This involves identifying the content, formats, and data sources needed for the initial data collection. Now consider the case study on the application of the "data requirements".

Case Study:

Case Study – Selecting the cohort

- Define and select cohort
  - In-patient within health insurance provider's service area
  - Primary diagnosis of CHF in one year
  - Continuous enrollment for at least 6 months prior to primary CHF admission
  - Disqualifying conditions

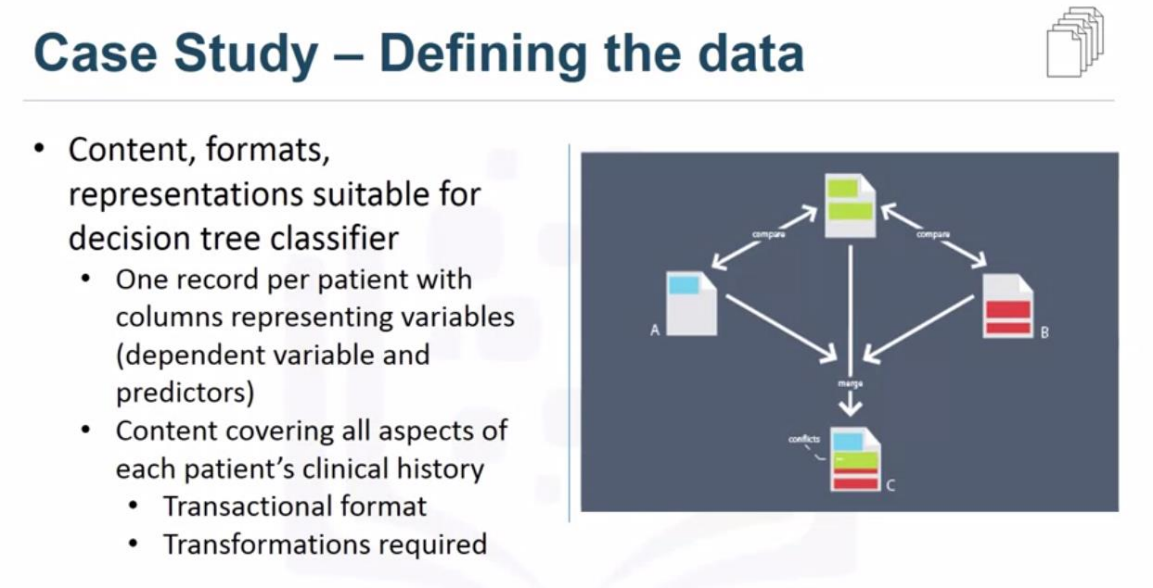


In the case study, the first task was to define the data required for the classification approach of the selected decision tree. This involved selecting a suitable cohort of patients from the members of the health insurance companies.

In order to compile the complete medical records, three criteria were identified that should be included in the cohort.

- **First**, *a patient had to be hospitalized in the service area of the provider to gain access to the required information.*
- **Second**, *for one year, they focused on patients with a primary diagnosis of heart failure.*
- **Third**, *a patient must have had a continuous record of at least six months prior to initial heart failure for a complete medical history.*

Patients with congestive heart failure who have been diagnosed with other serious conditions have been excluded from the cohort, as this may result in above-average rates of re-entry and may therefore distort results.

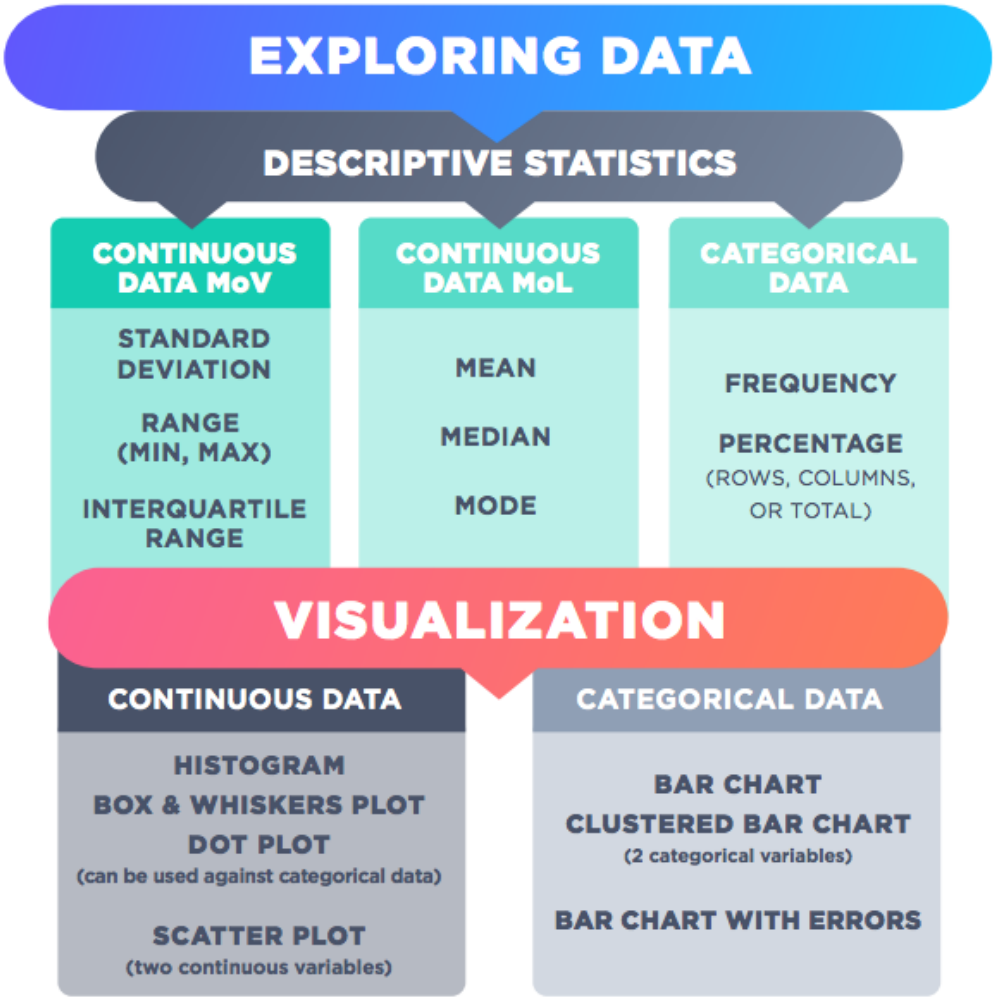


- Then he defined the content, format and representations of the data needed to classify the decision tree.
- This modeling technique requires one registration per patient, with columns representing the variables of the model. To model readmission results, data covering all aspects of the patient's medical history should be available.
- This content includes authorizations, primary, secondary and tertiary diagnoses, procedures, prescriptions and other services provided during hospitalization or visits by patients / doctors.

In this way, a given patient can have thousands of records that represent all their attributes. To obtain a record by patient format, the data analysis specialists collected the transaction records from patient records and created a set of new variables to represent that information. It was a task for the data preparation phase, so it is important to anticipate the next phases.

#2) Data Collection

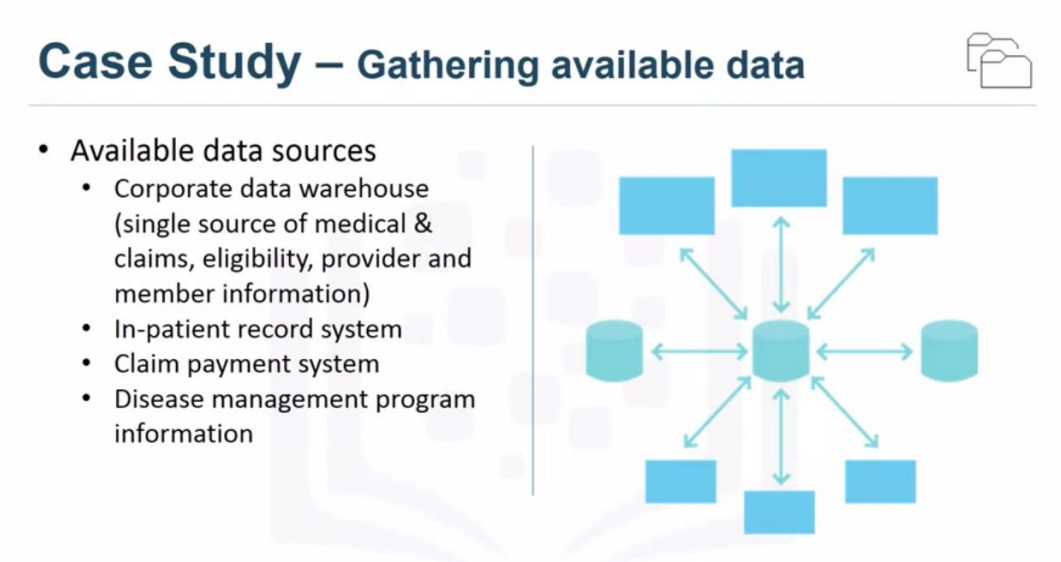
- Once The data collection is completed, the ***Data Scientist performs a score to determine if he has the required resources***. As with the purchase of ingredients for making a meal, some ingredients may be out of season and more difficult to obtain or cost more than originally planned.
- At this stage, the data requirements are reviewed and a decision is made as to whether more or less data is required for the collection.



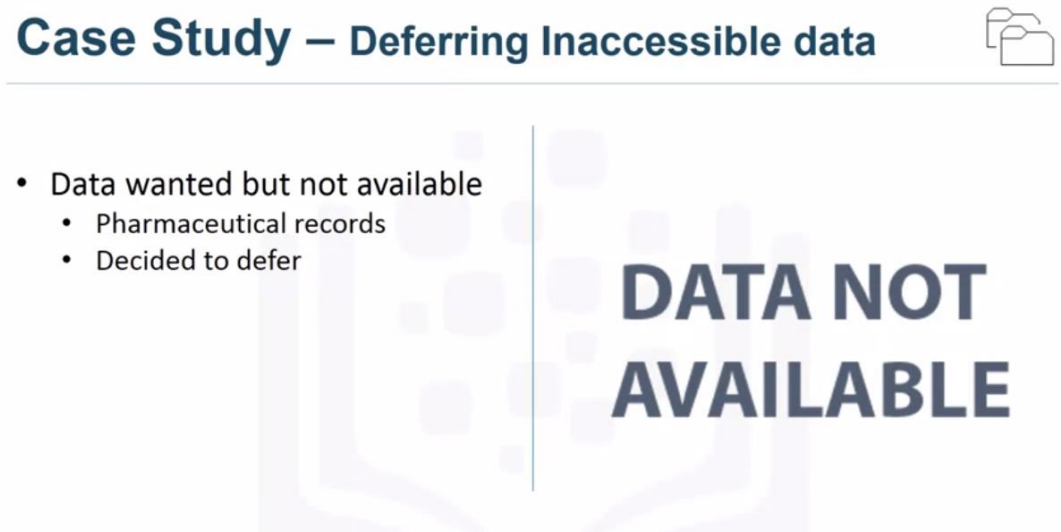
- Once the data components have been collected, the data scientist will have a good understanding of what data he will be working on during the data collection phase.
- Techniques such as ***descriptive statistics*** and ***visualization*** can be applied to the dataset to evaluate the content, quality and information of the original data. The gaps in the data are identified and plans for filling or replacement must be made.

Essentially, the ingredients are now sitting on the cutting board. Now let's look at some examples of the data collection phase in the data science methodology. This step is performed as a result of the data request step. Let us now consider the case study on the application of "data collection". To capture data, you must know the source or know where the required data items are located.

Case Study :

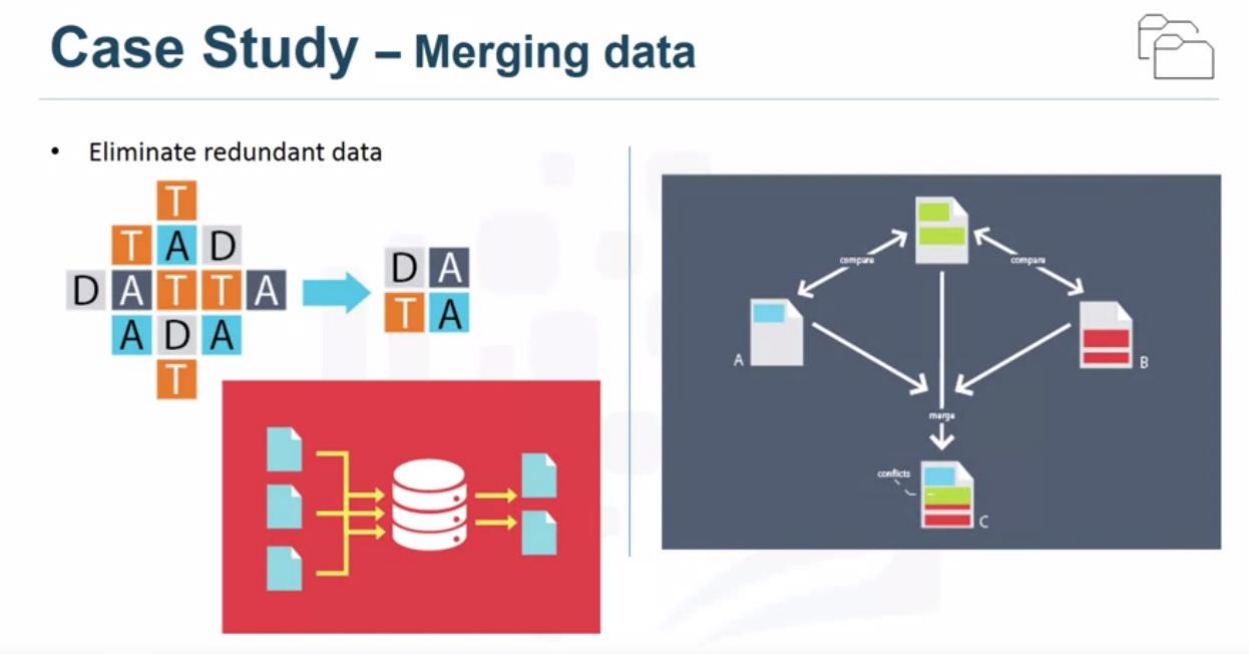


- In our case study, this information may include ***demographic, clinical and patient care information, provider information, claims records, as well as pharmaceutical and other information related to all heart failure diagnoses.***



- This case study also required specific information on ***drugs, but this data source*** was not yet integrated with the rest of the data sources.
- This brings us to an important point: ***it is correct to postpone decisions about unavailable data and to try to capture them later.***
- For example, this can happen even after obtaining intermediate results from predictive modeling. If these results indicate that drug information may be important for a good model, you will spend time trying to get it.

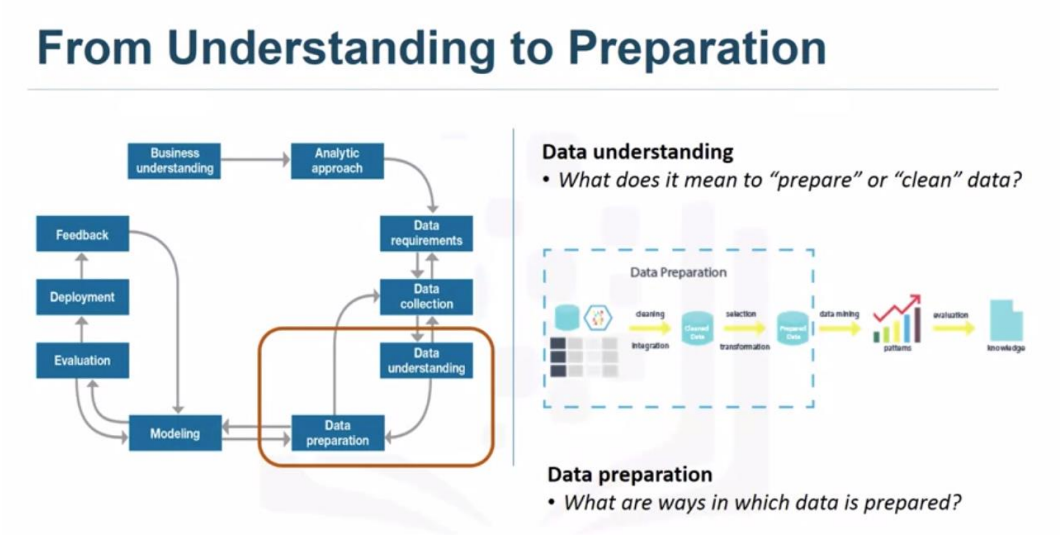
However, it turned out that they could build a reasonably good model without this information about drugs.



- Database administrators and programmers often work together to extract data from different sources and then combine them.
- In this way, the redundant data can be deleted and made available to the next level of methodology, namely the understanding of the data.
- At this stage, scientists and analysis team members can discuss ways to better manage their data by automating certain database processes to facilitate data collection.

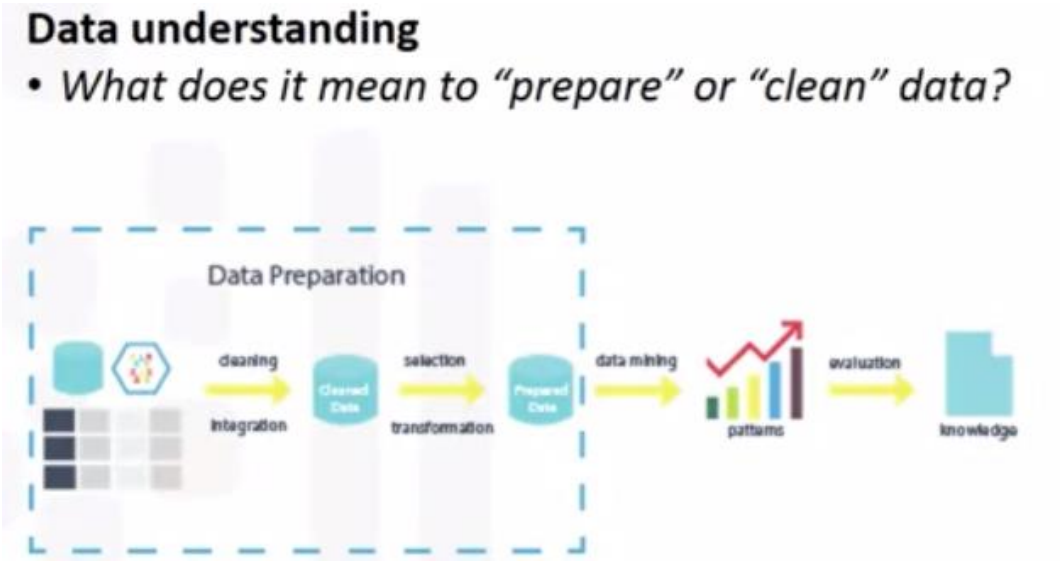
Part-3 Data Science Methodology From Understanding to Preparation

From Understanding to preparation

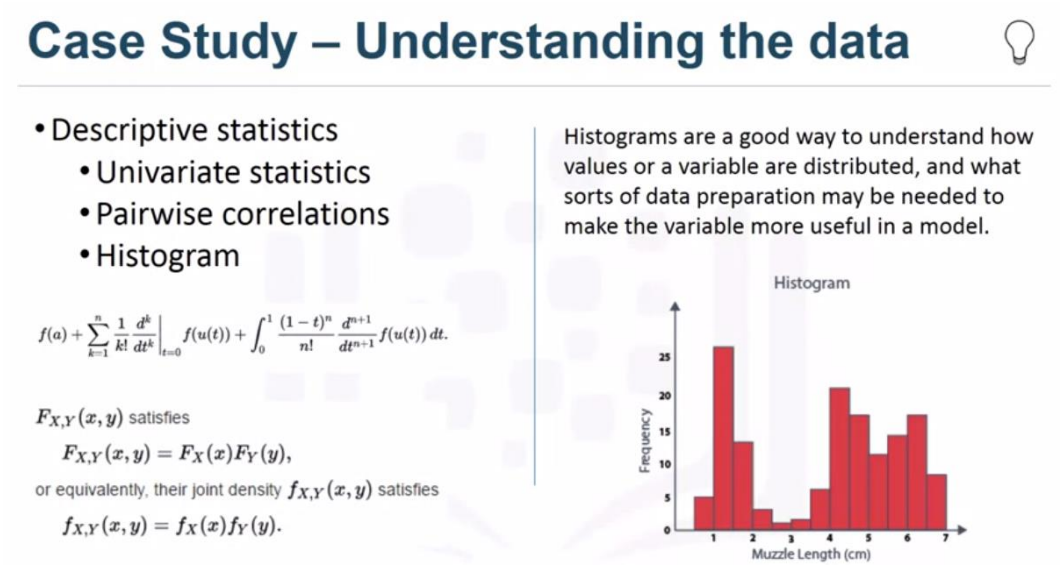


#1) Data Understanding

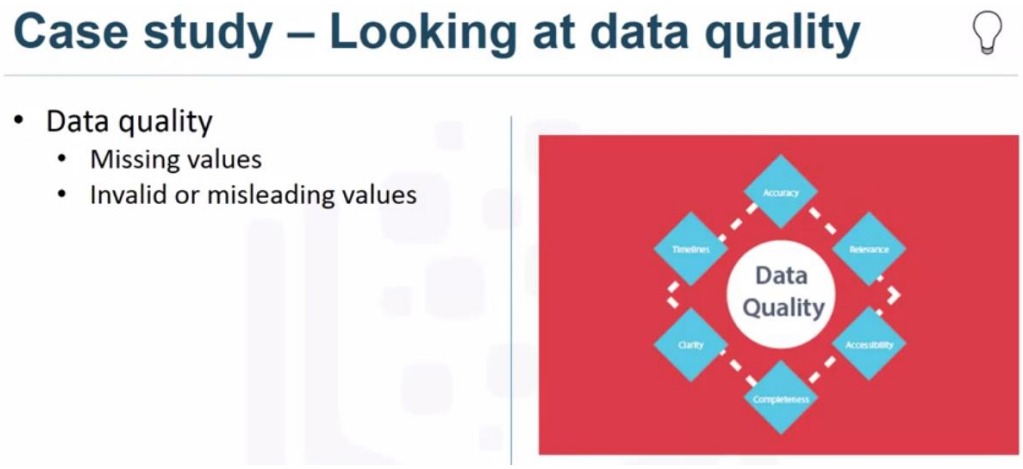
- Understanding the data includes all activities related to creating the data-set. The "Understanding Data" section of the Data Science methodology answers the question:
- ***Is the data you collect representative of the problem you are trying to solve?***



Apply the ***understanding of our methodological data*** to the case study we are studying. To understand data on the onset of heart failure, ***descriptive statistics had to be established in the data columns that would become variables in the model.***



- **First**, these statistics included ***Hearst, Uni-variate, and statistics for each variable, such as mean, median, minimum, maximum, and standard deviation.***
- **Second**, ***pairwise correlations have been used to determine the degree of correlation between the linked variables*** and those that, ***if any, are highly correlated, meaning that they are essentially redundant, making it only relevant for the modeling***.
- **Third**, the ***histograms of the variables were examined to understand their distributions***. ***Histograms are a good way to understand how values or variables are distributed and what kind of data preparation may be needed to make the variable more useful in a model.***For example, **if a categorical variable contains too many different values to be meaningful in a model**, the histogram can help decide how to consolidate those values.

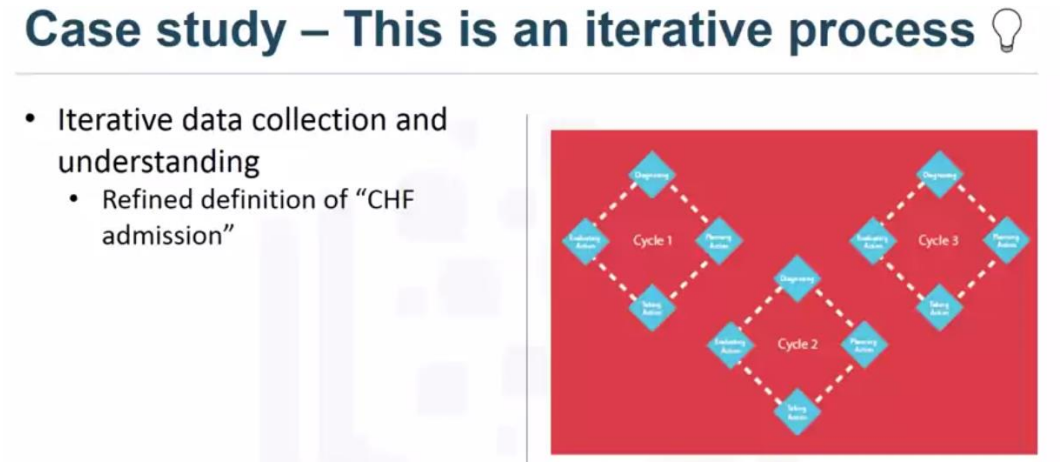


- Univariate, statistics and histograms are also used to assess the quality of the data.

On the basis of the data provided, some values can be recorded or deleted if necessary, e.g. **For example**, if a particular variable has a lot of missing values.

- *The question then arises as to whether "missing" means something. Sometimes a missing value means "no" or "0" (zero), or sometimes simply "we do not know".*

- Or *if a variable contains invalid or misleading values; For example, a numeric variable called "age" containing 0 to 100 and 999, where "triple-9" actually means "missing", will be treated as a valid value unless we have corrected it.*



- First, the importance of heart failure was determined based on a primary diagnosis of heart failure. However, the data comprehension study revealed that the initial definition did not cover all expected cases of heart failure due to clinical experience.
- This involved returning to the data collection phase, adding secondary and tertiary diagnoses, and creating a more complete definition of heart failure approval.
- This is just an example of the interactive processes in the methodology. The more you work with the problem and the data, the more you learn and the more the model can be adjusted, **which ultimately leads to a better resolution of the problem.**

#2) Data Preparation:

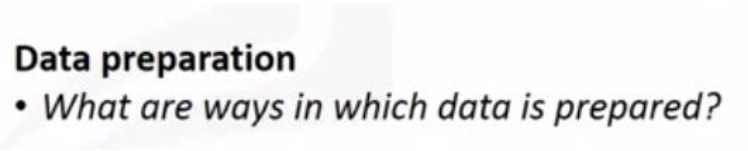
In a way, data preparation is like washing freshly cut vegetables, as long as they remove unwanted elements such as dirt or insects.

- With data collection and understanding, data preparation is the slowest phase of a data science project. As a rule, ***it takes up 70% or 90% of the total***

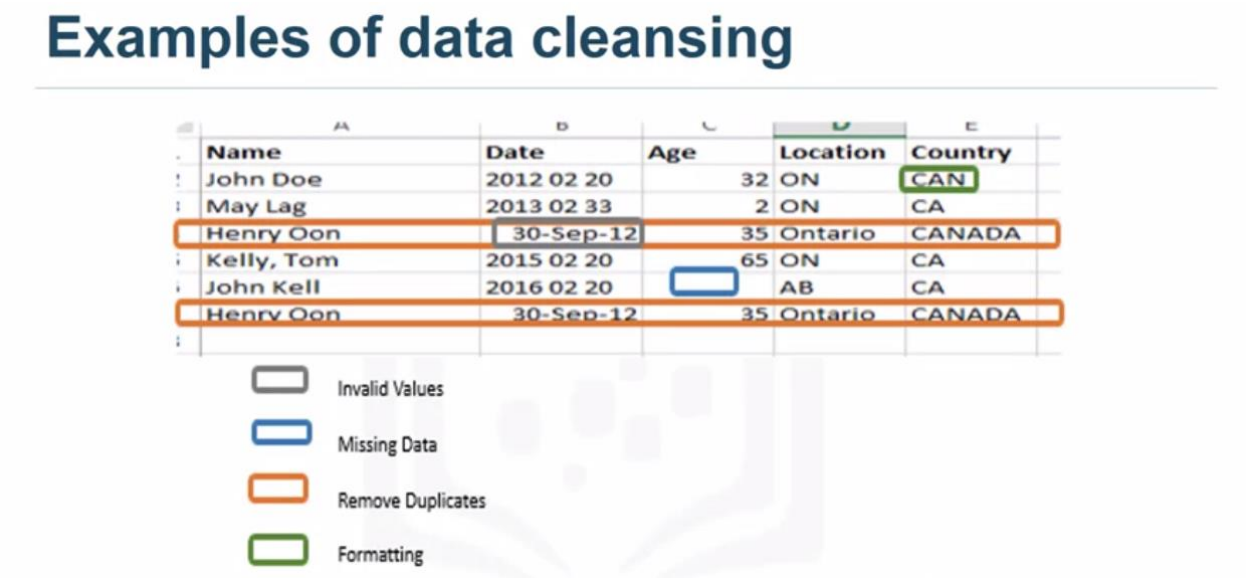
project time. By automating certain data collection and preparation processes in the database, this time can be reduced to only 50%.

- This ***saving of time means that data scientists should focus more on creating models***.

To continue with our metaphor of cooking, we know that the process of cutting onions in a finer state allows the flavors to spread more easily in a sauce than if we dropped the whole onion in the pot.

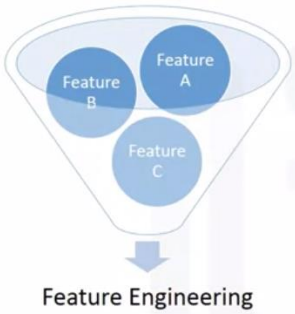


- Similarly, data transformation in the data preparation phase involves putting the data in a state where it may be easier to work.



- More specifically, the data preparation phase of the methodology answers the question: ***how are the data processed?*** To work efficiently with data, missing or invalid values must be changed and duplicates removed to ensure proper formatting of all data.

Using domain knowledge

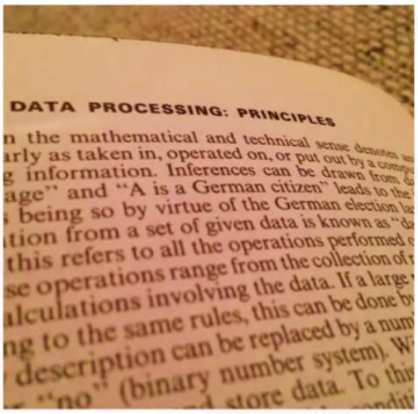


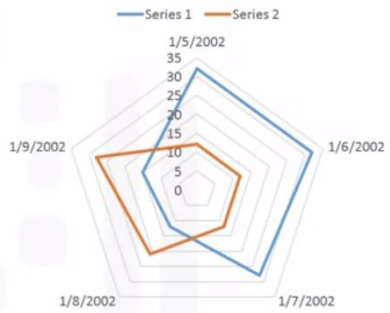
Feature engineering is the process of using domain knowledge of the data to create features that make the machine learning algorithms work.

Feature engineering is critical when machine learning tools are being applied to analyze the data.

- **Feature engineering** is also part of the ***data preparation***. Use domain knowledge on data to create features that work with machine learning algorithms. ***A feature is a property that can be useful for solving a problem. The functions in the data are important for the predictive models and influence the desired results.***

Working with text analysis

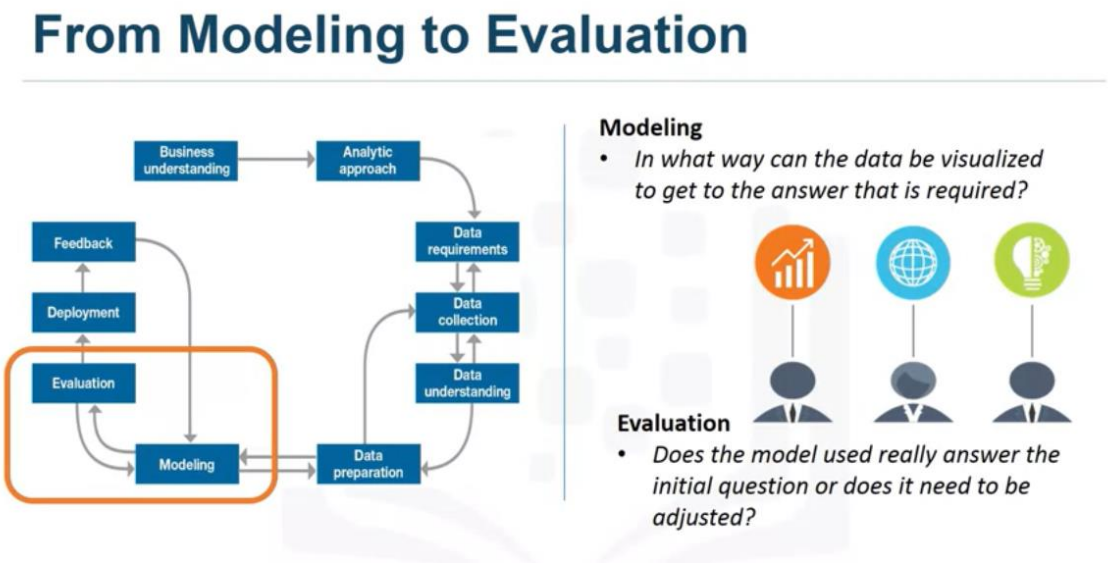




- ***Feature engineering*** *is essential when machine learning tools are used to analyze data. When working with text, text analysis steps are required to code the data for manipulation.*
- The data scientist must know what he is looking for in his file to answer the question. Text analysis is essential to ensure that the correct groups are defined and that programming does not overlook what is hidden inside.
- The data preparation phase is the basis for the next steps to answer the question. Although this phase may take some time, the results will support the project if done correctly. If this is omitted, the result is not at the same level and can sit on the drawing board.
- Make sure you spend time in this area and use the tools available to automate common steps to speed up data preparation. Pay attention to details in this area. After all, all you need is a bad ingredient to ruin a good meal.

Part-4 Data Science Methodology from Modelling to Evaluation

From Modelling to Evaluation



Welcome to the data science methodology. Till now we have seen all ***3 stages of data science methodology*** from ***Problem to approach, Requirement to collections, Understanding to preparation.*** We have discussed amazing example with case study approach if you haven't read this article series read from below links. and already read that go directly with this article. In this article, you can learn about how to select the model and how to evaluate that model or this model is ready for deployment or not.

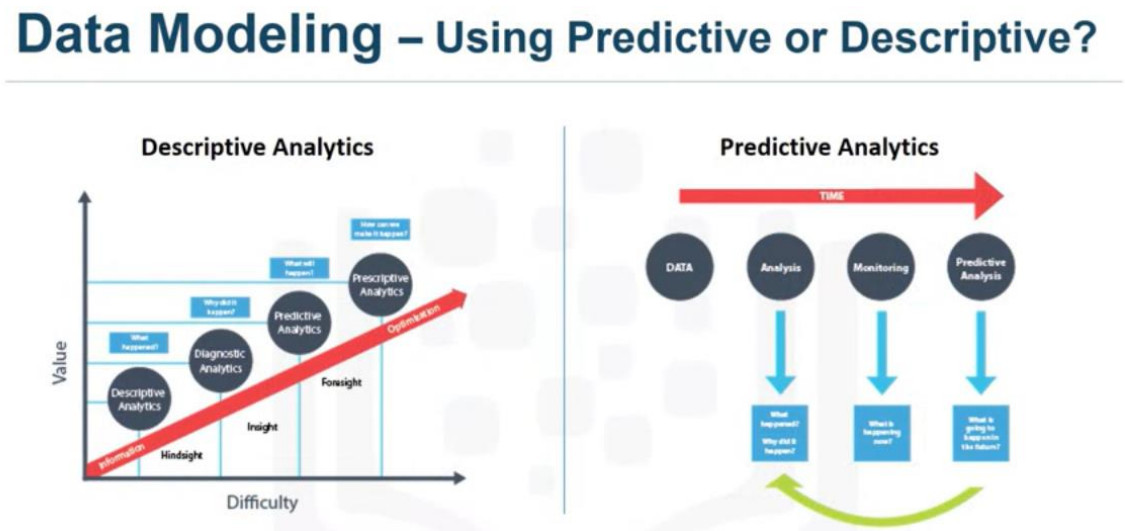
#1) Modeling

Modeling is the phase of the methodology of data science in which the data scientist has the opportunity to taste the sauce and determine if it needs more seasoning or if it needs more seasoning!

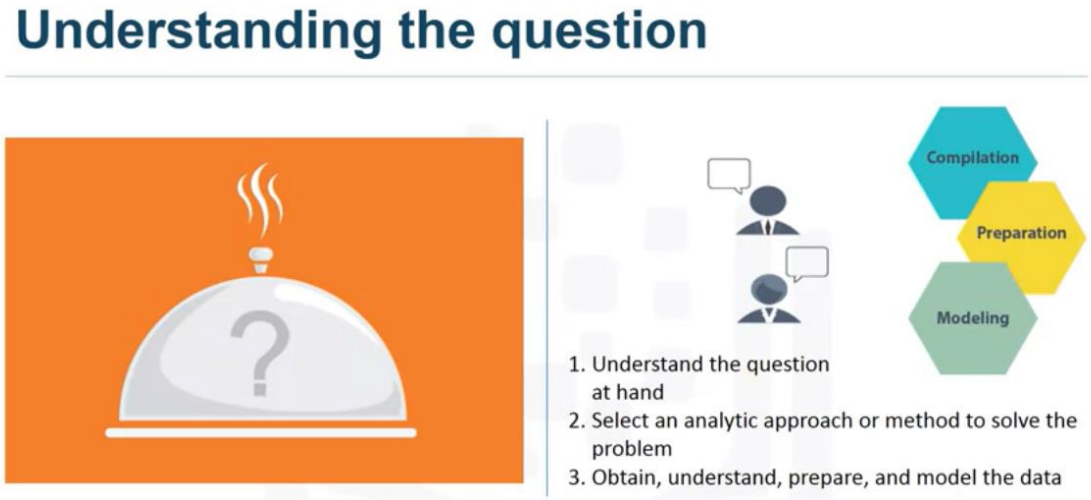
This part of the course is designed to answer two key questions:

- ***First, what is the purpose of data modeling, and***
- ***Second, what are the characteristics of this process?***

Data modeling focuses on the development of ***descriptive or predictive models.***



- An example of a descriptive model might be the following: if someone did it, they probably prefer it.
- A **predictive model** attempts to provide **yes / no** results or to **stop / continue.** These models are based on an analytic approach learned either statistically or Machine learning. The Data Scientist will use a training set for predictive modeling.
- ***A training set is a set of historical data in which the results are already known***. The training set serves as an ***indicator to determine if the model needs to be calibrated***.
- At this point, the data scientist will use several algorithms to ensure that the variables involved are really needed.

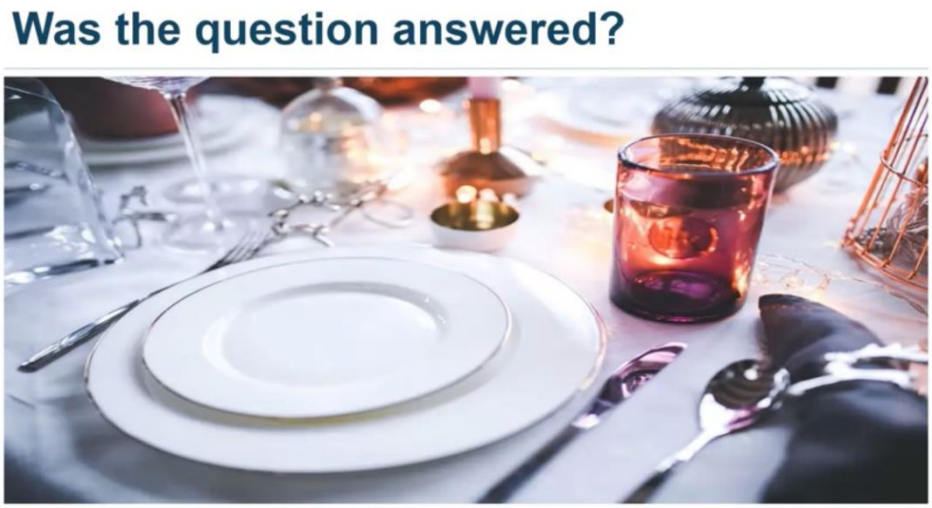


- The **success** of ***data collection, preparation and modeling depends on an understanding of the problem in question and the appropriate analytical approach.***
- The data support the answer to the question and the quality of the ingredients in the kitchen is the basis of the result.
- Each step requires constant improvements, adjustments and tweaking to ensure the strength of the result.

In the descriptive data science methodology of John Rollins, the framework is designed for three things:

- ***First, understand the question that concerns you.***
- ***Secondly, choose an analytical approach or method to solve the problem.***
- ***Thirdly, obtaining, understanding, preparing and modeling data.***

The ultimate goal is to bring the data scientist to a point where it is possible to create a data model to answer the question.



- While dinner is being served and a hungry guest sits at the table, the key question is: have I prepared enough to eat? *We hope that at this stage of the methodology, model evaluation , deployment and feedback cycles of the models will ensure that the response is relevant and near to the result.*
- This relevance is essential for the whole field of data science, as it is a relatively new field and we are interested in the possibilities it offers.
- The more people benefit from the results of this practice, the more the field develops.

Case study:

- The modeling is the phase of the methodology of data science during which the data scientist has the opportunity to taste the sauce and determine if it breaks or if it needs additional seasoning! Now apply the case study to the modeling phase as part of the data science methodology.

Here we will discuss one of the many aspects of model construction, in this case optimizing the parameters to improve the model.

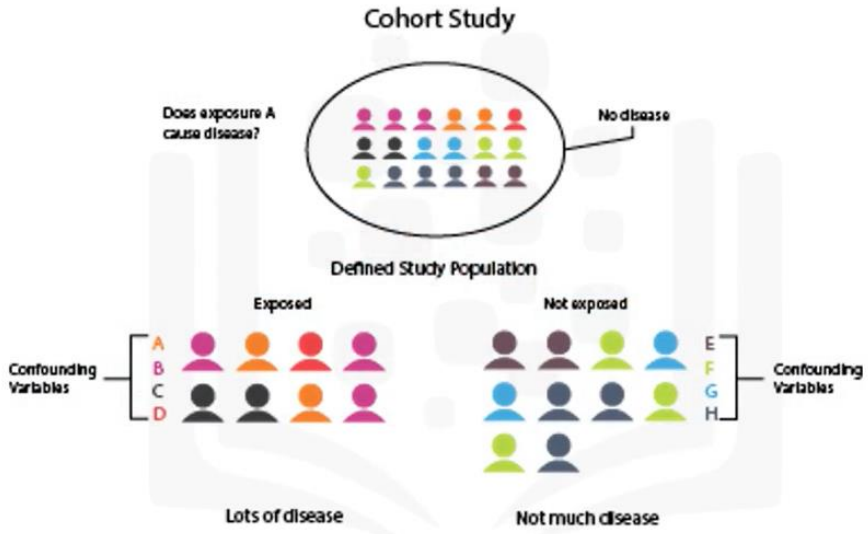
Case Study – Analyzing the 1st model

Initial decision tree classification model
- Low accuracy on "Yes" outcome

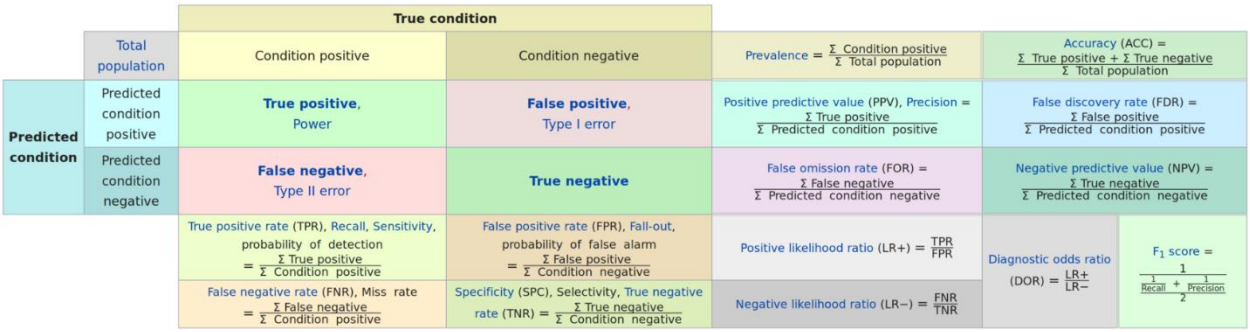
Model	Relative Cost Y:N	Overall Accuracy (% correct Y & N)	Sensitivity (Y accuracy)	Specificity (N accuracy)
1	1:1	85%	45%	97%
2	9:1	49%	97%	35%
3	4:1	81%	68%	85%

- With a set of prepared training data, it is possible to construct the first classification model of the decision tree for congestive readmission for heart failure. We are looking for patients with high risk readmission. The result that will interest us will be a congestive readmission for heart failure equivalent to "yes". In this first model, the overall accuracy of the classification of the results was 85% and not 85%. It sounds good, but represents only 45% of the "yes". Actual readmission are ranked correctly, which means that the model is not very accurate.
- The question is : ***how to improve the accuracy of the model to predict the outcome itself ?***. For the classification of the decision tree, the best parameter to adjust is the relative cost of the results yes and not classified incorrectly.

Case Study – How to improve the model?



- ***Think of it this way:*** When a true ***non-readmission is misclassified*** and **actions are taken to reduce the risk of this patient**, the *cost of this error is a wasted intervention.*



- A statistician calls this a **Type I error or a false positive.** But ***when a real readmission is misclassified and no action is taken to reduce this risk***, the cost of such an error is ***readmission*** and all associated costs, as well as trauma to the patient.
- It's a **Type II error or a false negative**. Then ***we can see that the costs of the two different types of incorrect classification errors*** can be very different. For this reason, it is reasonable to adjust the relative weights of the incorrect classification of the results yes and no.
- The default is between 1 and 1, but the decision tree algorithm allows you to set a higher value for yourself.

Case Study – Analyzing the 2nd model

Second model

- High accuracy on "Yes" but poor on "No"

Model	Relative Cost Y:N	Overall Accuracy (% correct Y & N)	Sensitivity (Y accuracy)	Specificity (N accuracy)
1	1:1	85%	45%	97%
2	9:1	49%	97%	35%
3	4:1	81%	68%	85%

- For the **second model**, the relative cost was set at **9/1**. This report is very high, but provides more information about the behavior of the model. This time, the ***97% model worked well***, but at a ***very low cost***, ***with a general accuracy of only 49%.*** Obviously, ***this is not a good model.***
- The **problem** with this result is the ***large number of false positives***, **suggesting unnecessary** and **costly interventions** for patients that have never been re-admitted.
- Therefore, the data scientist must try again to get a better balance between the ***yes and no data***.

Case Study – Analyzing the 3rd model

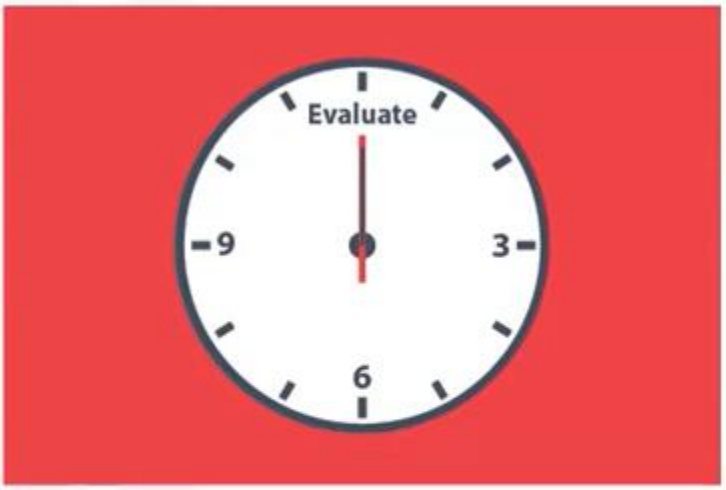
Third model

- Better balance on "Yes" and "No" accuracy

Model	Relative Cost Y:N	Overall Accuracy (% correct Y & N)	Sensitivity (Y accuracy)	Specificity (N accuracy)
1	1:1	85%	45%	97%
2	9:1	49%	97%	35%
3	4:1	81%	68%	85%

- For the **third model**, the relative cost was set to a more reasonable ***4: 1 ratio***. This time, **68% was obtained yes**, but statistician called it **sensitivity**, and **85% accuracy for the no**, called **specificity**. , with an overall **accuracy of 81%.**
- This is the best balance that can be achieved with a relatively limited training set of workouts by adjusting the relative cost of the misclassified yes and no result parameters. Of course, modeling requires much more work, including an iteration in the data preparation phase, to redefine some of the other variables to better represent the underlying information and thus improve the model.

#2) Model Evaluation



Evaluation

- *Does the model used really answer the initial question or does it need to be adjusted?*

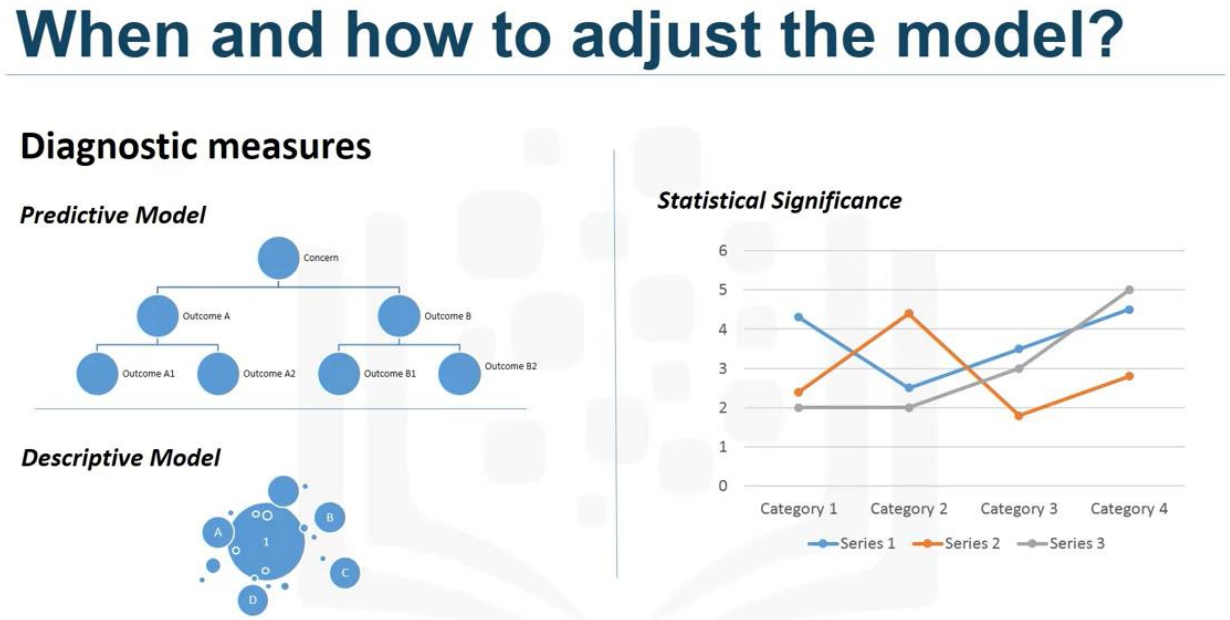
A model evaluation goes hand in hand with the creation of models. The modeling and evaluation steps are performed iteratively. The evaluation of the model is carried out during the ***development of the model and before deployment.***

- ***The evaluation evaluates the quality of the model, but also provides the opportunity to determine if it meets the initial requirements.***

The evaluation answers the question:

- ***Does the model used really answer the original question or should it be adapted?***

The evaluation of the model can have two main phases.



- **The first phase** is the diagnostic measurement phase, ***which ensures that the model works as intended.*** If the **model is predictive**, a ***decision tree can be used to assess whether the response provided by the model matches the original design***. This allows areas to be displayed where adjustments are required. **If the model is a descriptive model** that **evaluates the relationships,** a ***set of tests with known results can be applied and the model refined as necessary.***
- **The second evaluation phase** that can be used is the **statistical significance test.** This type of ***evaluation*** can be applied to the ***model to ensure that the model data is processed and interpreted correctly.*** This is to ***avoid a second unnecessary assumption when the answer is revealed.***

Case study :

- Let's go back to our case study to apply the ***Evaluation component*** in the data science methodology.

Case Study – Misclassification costs

Misclassification cost tuning

- Tune the relative misclassification costs
- Balance true-positive rate and false-positive rate for best model

Model	Relative Cost Y:N	True Positive Rate (Sensitivity)	Specificity (accuracy on N)	False Positive Rate (1 – Specificity)
1	1:1	0.45	0.97	0.03
2	1.5:1	0.60	0.92	0.08
3	4:1	0.68	0.85	0.15
4	9:1	0.97	0.35	0.65

- Let's look for a way to find the *optimal model* through a diagnostic measurement based on the configuration of one of the model's construction parameters. *We will examine more closely how the relative costs of misclassifying positive and negative results can be adjusted. As shown in this table, four models were constructed with four different relative misclassification costs.*

Case Study – Relative costs

Misclassification cost tuning

- Tune the relative misclassification costs
- Balance true-positive rate and false-positive rate for best model

Model	Relative Cost Y:N	True Positive Rate (Sensitivity)	Specificity (accuracy on N)	False Positive Rate (1 – Specificity)
1	1:1	0.45	0.97	0.03
2	1.5:1	0.60	0.92	0.08
3	4:1	0.68	0.85	0.15
4	9:1	0.97	0.35	0.65

- As we see, each value of this ***model construction parameter increases the true positive rate***, or the ***sensitivity***, of the ***accuracy in the prediction yes***, to the detriment of a ***lower accuracy in the prediction no***. that is, an increasing ***rate of false positives.***
- The question is, ***which model is best based on setting this parameter?*** For budgetary reasons, the risk reduction intervention could not be applied to most

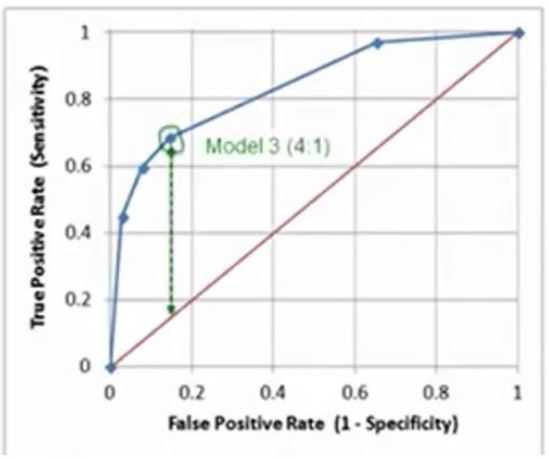
patients with heart failure, many of whom would not have been readmitted anyway.

- On the other hand, the intervention would not be as effective as it should be to improve patient care, since the number of patients with high-risk heart failure was not enough.

Cost Study – Using the ROC curve

Diagnostic tool for classification model evaluation

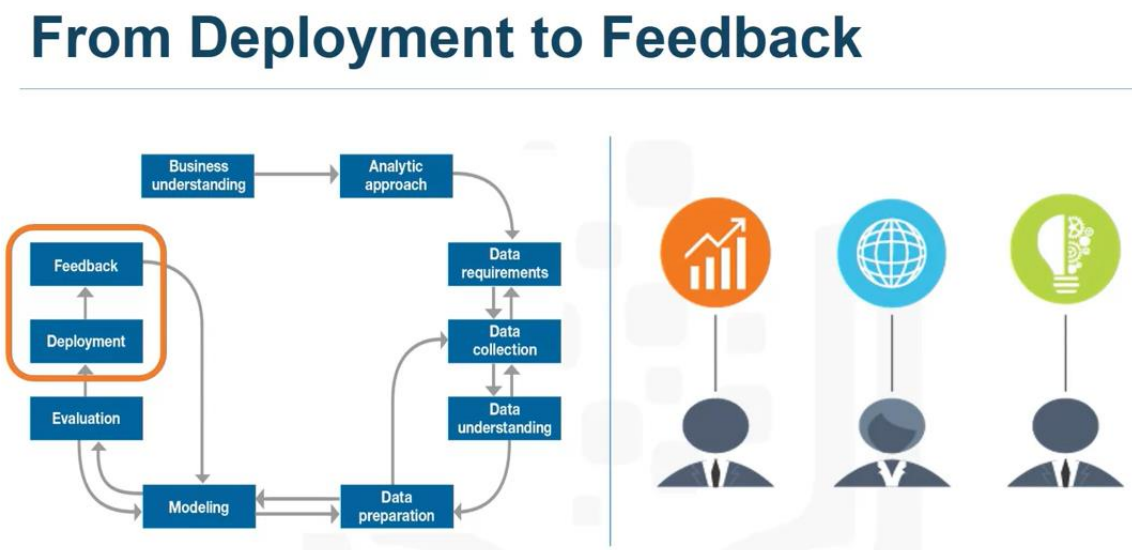
- Classification model performance
- True-Positive Rate vs False-Positive Rate
- Optimal model at maximum separation



- ***So how do we determine which model was optimal?*** As you can see on this image above , the ***optimal model is the one that provides the maximum separation between the blue ROC curve and the red baseline.***
- We can see that **model 3**, with a **relative cost of misclassification** of **4 to 1**, ***is the best of the 4 models***. And if asked, ***ROC represents the characteristic operating curve of the receiver***, which was **first developed during World War II to detect enemy aircraft on a radar.**
- Since then, it has also been used in many other areas. Today, it is commonly used in machine learning and data mining. ***The ROC curve is a useful diagnostic tool to determine the optimal classification model.***
- This curve quantifies the performance of a **binary classification model,** declassifying the results yes and no when a discrimination criterion is changed.
- In this case, the criterion is a relative cost of misclassification. By plotting the true positive rate against the false positive rate for different values of the relative cost of misclassification, the ROC curve facilitated the selection of the optimal model.

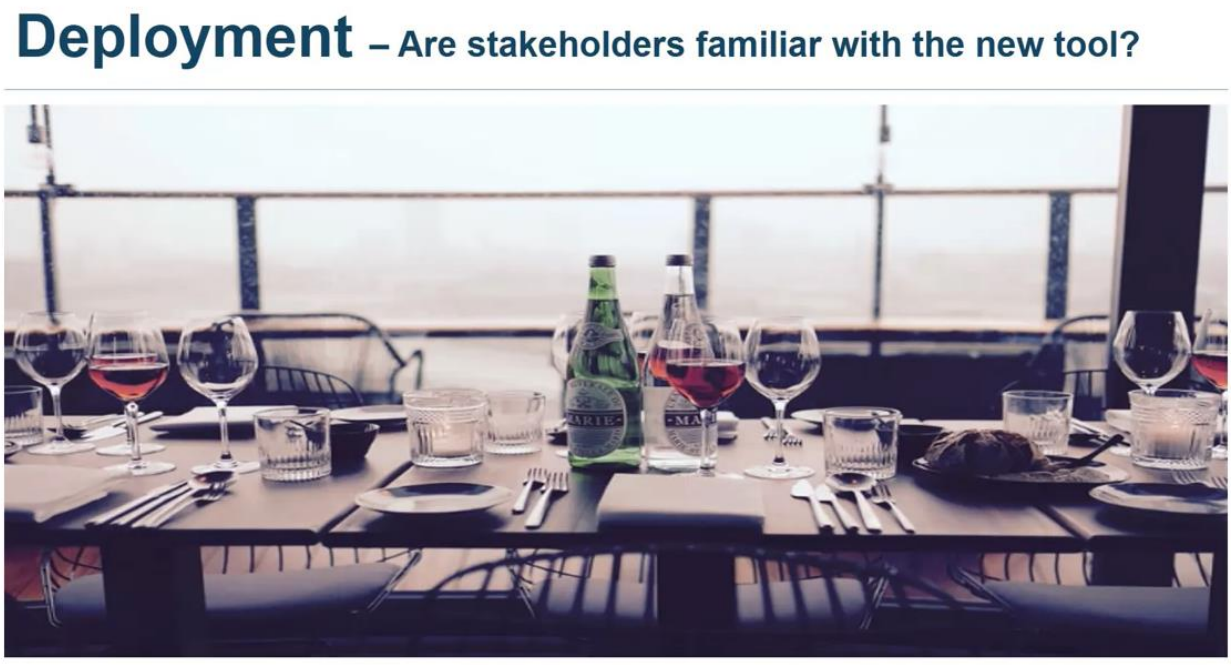
Part-5 Data Science Methodology From Deployment to Feedback

From Deployment to Feedback



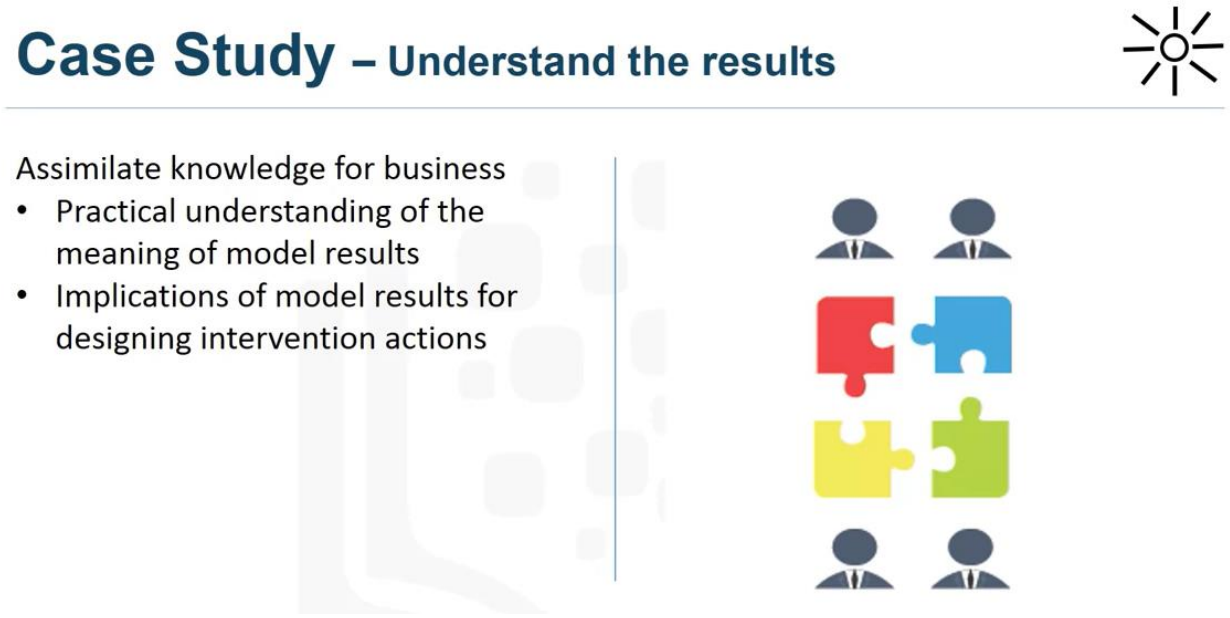
Welcome to the data science methodology. Till now we have seen all **4 stages of data science methodology from Problem to approach, Requirement to collections, Understanding to preparation, Modeling to Evaluation**. We have discuss amazing example with case study approach if you haven't read this article series read from below links. and already read that go directly with this articles. In this article, You can learn about how to Model deploy and how to take feed back a model so model become more mature by the time.

#1) Deployment



- While a ***data science model provides an answer***, the *key to making the answer relevant to answering the initial question is to familiarize people with the product tool.* In a business scenario, stakeholders have different characteristics, such as: ***The solution owner, marketing, application developers and IT administration.***
- Once the model has been evaluated and the Data scientist is convinced that it will work, it will be used and subjected to the final test.
- Depending on the purpose of the model, it can be extended to a limited group of users or in a test environment to increase confidence in the application of the result to global use.

Case Study:

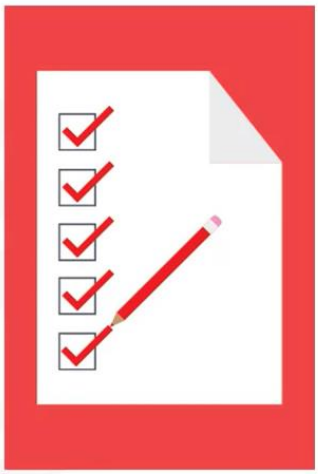


- Let's take a look at the case study of the implementation application "**In preparation, To** provide the solution, the next step was to ***gather the knowledge*** of the *stakeholder group responsible for the design and management of the intervention program to reduce the risk of readmission.*
- In this scenario, **entrepreneurs have translated the results of the model** so ***that clinical staff can understand how to identify high-risk patients and design appropriate interventions.***
- Of course, the ***objective was to reduce the risk of readmission of these patients within 30 days of discharge***. During the **operational requirements phase**, the ***intervention program director and her team looked for an application that could assess the risk of heart failure almost automatically in real time.***

Case Study – Gathering application requirements

Application requirements

- Automated, near-real-time risk assessments of CHF inpatients
- Easy to use
- Automated data preparation and scoring
- Up-to-date risk assessment to help clinicians target high-risk patients



- It should also be *easy for clinical staff to use,* preferably through a ***browser*** and ***tablet-based application*** that *any employee could carry with them*. These ***patient data were generated throughout the hospital stay***. It will be ***generated automatically in a format required by the model*** and ***each patient will be noticed shortly before discharge.***
- Then, ***doctors would have the most up-to-date risk assessment for each patient to help them choose which patients to treat after discharge.***
- As part of providing the **solution,** the intervention team would develop and offer training to clinical staff.

Case Study – Additional requirements?

Additional requirements

- Training for clinical staff
- Tracking / monitoring processes

- In addition, in collaboration with ***IT developers and database administrators, monitoring and monitoring processes should be developed for patients receiving the intervention***, so that the ***results can go through the feedback phase*** and the **model can be mature over time.**

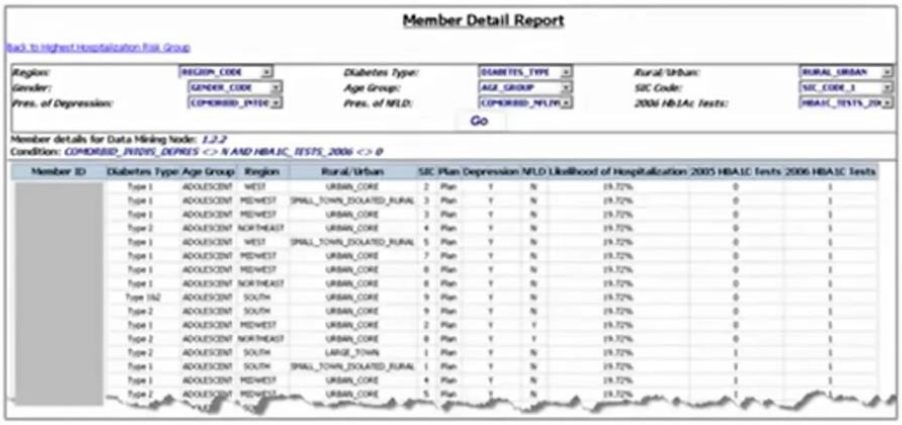
Example 1 – Solution deployment

Hospitalization risk for juvenile diabetes patients

- This Map is an example of a solution implemented through a **Cognos application(IBM Cognos Business Intelligence is a web-based integrated business intelligence suite by IBM).** In this case, the case study focused on the ***risk of hospitalization of patients with juvenile diabetes***. Similar to *congestive heart failure, he used the classification of the decision tree to create a risk model that would form the basis of this application.*

Example 2 – Solution deployment

Risk summary report by decision tree model note



- The map provides an overview of **hospital risk nationwide**, with a planned interactive **risk assessment for different patient conditions and other characteristics.** ***This above image provides an interactive summary report on the risk per patient population in a given node of the model so that doctors can understand the combination of conditions for that subset of patients***.

Example 3 – Solution deployment

Individual patient risk report

Member Management Report

Member Summary

Member ID: [redacted] Risk: 0.197 Confidence: 0.197 Previous Hospitalization: N

Demographic Information

Diabetes Type:	Type 1&2
Rural / Urban:	URBAN_CORE
Gender:	M
Age:	17
Region:	SOUTH

Clinical Comorbidities

Depression:	Y	Nephropathy:	N
MLD:	N	Dyslipid:	Y
Anxiety:	Y	Obesity:	Y
Neuropathy:	N	Gestation Diabetes:	N
Hypertension:	Y	Celiac:	N

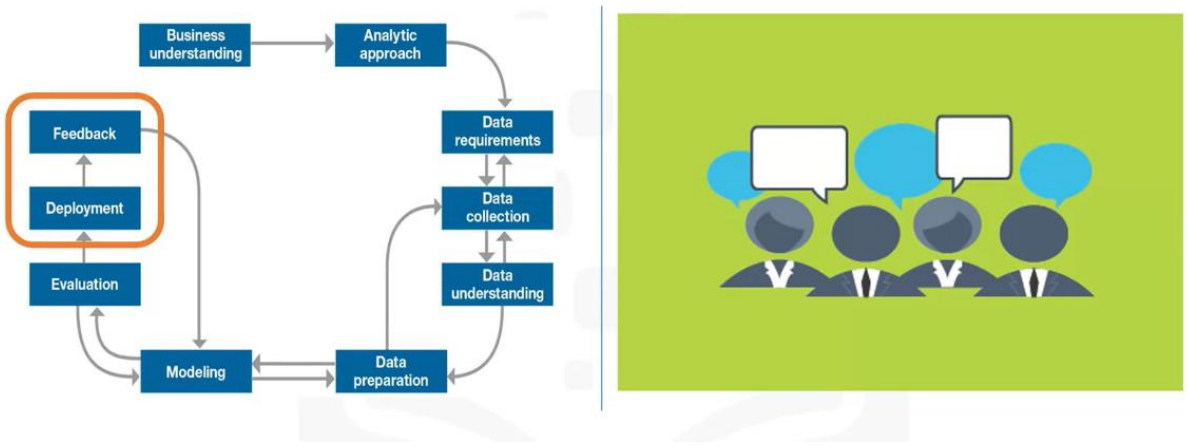
Utilization Metrics

HbA1c Tests 2006:	1	LDLC Tests 2006:	0
Eye Exams 2006:	Y	Diabetic Education 2006:	N
Dietitic Consultations 2006:	N	Micro Albumin Tests 2006:	N
Mental Health Consultations 2006:	Y	Distinct Physicians 2006:	1-5
Flu Shots 2006:	N	Inpatient Admissions 2006:	N

- This report provides a detailed summary of a single patient, including details of the ***patient's history and expected risk, and provides the doctor with a brief summary.***

#2) Feedback

Feedback – Problem solved? Question answered?



- Feedback ! After playing, ***user feedback help refine the model and evaluate its performance and impact***. The **value of the model depends on the successful integration of feedback and customization whenever the solution is needed.**
- Throughout the methodology of data science, *each step paves the way for the next.* By making the methodology cyclical, you ensure a refinement at each stage of the game.
- ***The feedback process is based on the idea that the more you know, the more you want to know.***

From Deployment to Feedback

Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test

- Actual real-time use in the field

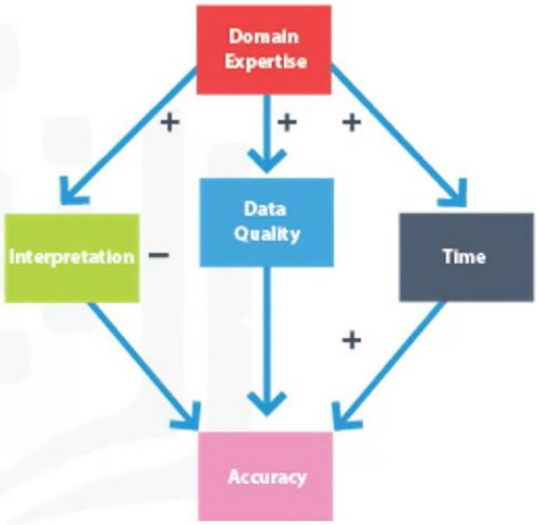
- Once the *model has been evaluated* and the ***data scientist trusts that it will work***, it will be implemented and will undergo the final test: ***its real use in real time in the field***.

Case Study:

Case Study – Assessing model performance

Define review process

- To measure results of applying the risk model to the CHF patient population
- Track patients who received intervention
  - Actual readmission outcomes
- Measure effectiveness of intervention
  - Compare readmission rates before & after model implementation



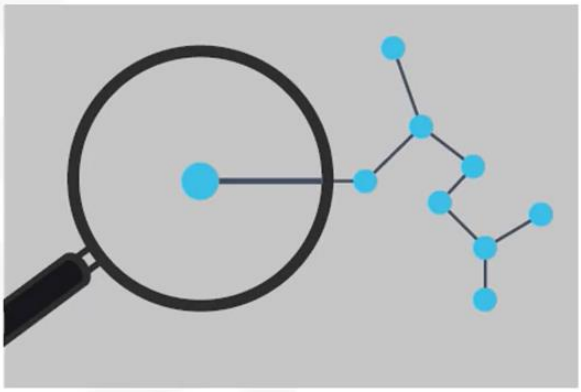
- Now let's review our case study to see how the part of the feedback methodology is applied.
- The feedback phase plan included the following steps: ***First, the review process would be defined and established, with the overall responsibility of measuring the results of a flight risk model of the heart failure risk population.***Clinical management has overall responsibility for the review process.

- **Second**, ***patients with heart failure who receive an intervention would be monitored and their readmission results recorded.***
- **Third**, the ***intervention would be measured to determine its effectiveness in reducing the number of readmissions.***
- For ethical reasons, patients with heart failure would not be divided into controlled groups and treatment groups. Readmission rates are compared before and after the implementation of the model to measure the impact.

Case Study – Refinement

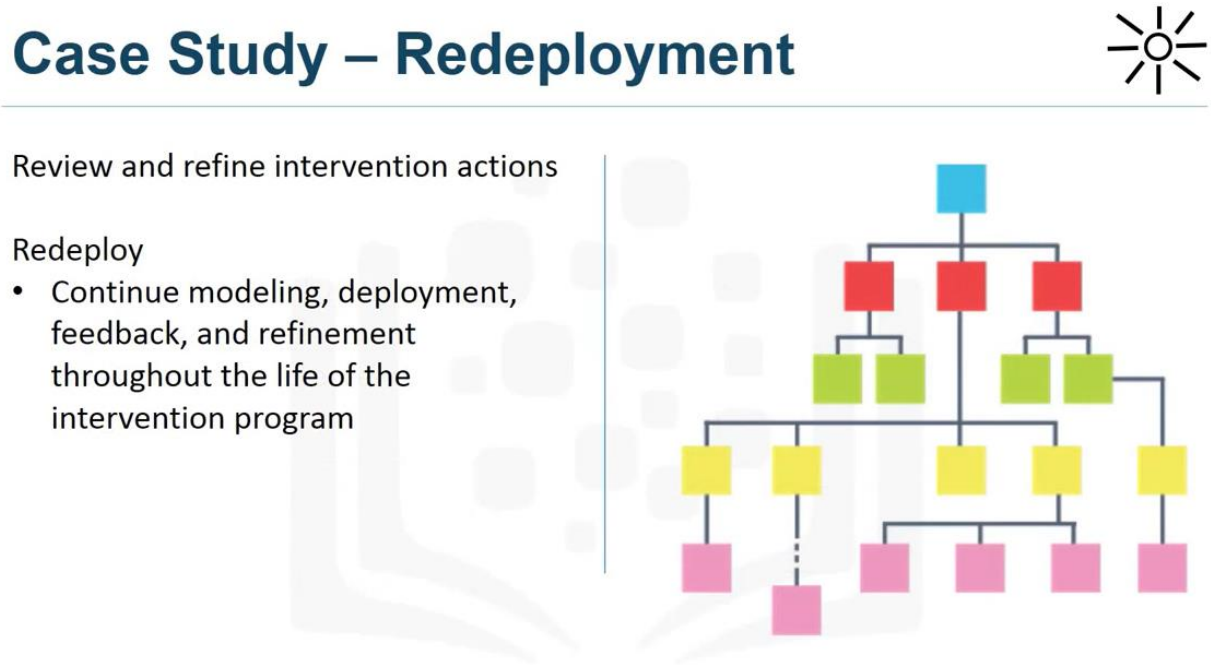
Refine model

- Initial review after the first year of implementation
- Based on feedback data and knowledge gained
- Participation in intervention program
- Possibly incorporate detailed pharmaceutical data originally deferred
- Other possible refinements as yet unknown



- **After deployment and feedback**, the impact of the ***intervention program on readmission rates*** will be reviewed after the ***first year of implementation***.
- Then, ***the model would be refined based on all data compiled*** after the implementation of the **model** and the **knowledge** acquired in these steps. **Other improvements** include the **inclusion of information on participation in the intervention program and possibly the refinement of the detailed pharmaceutical data model.**
- If you remember, data collection was initially delayed because drug data was not available at that time. However, after feedback and practical experience of the model, it can be said that adding this data can be worth the investment of time and money.We must also consider the possibility of new adjustments during the feedback phase.

Case Study – Redeployment



- In addition, response actions and processes are reviewed and probably refined according to the experience and knowledge acquired during the initial implementation and feedback.
- Finally, the refined model and intervention would be redeployed, and the feedback process would continue throughout the intervention program.

Conclusion :

I hope you are enjoying this article series Thanks for reading…!!! Happy Learning…!!!

References :

1. https://www.coursera.org/learn/data-science-methodology