Data Science Methodology

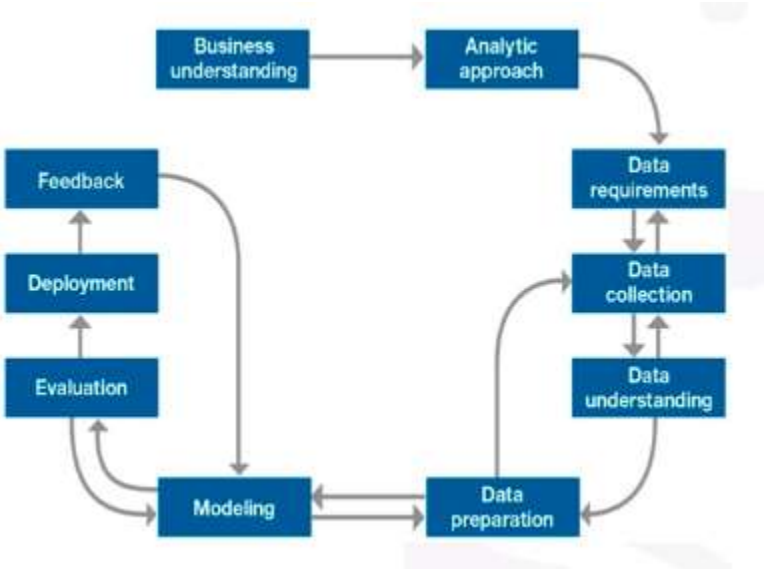
The Data Science Methodology aims to answer 10 following questions:

- From problem to approach:
    1. What is the problem that you're trying to solve?
    2. How can you use data to answer the question?
- Working with the data
    1. What data do you need to answer the question?
    2. Where is the data coming from (identify all sources) and how will you get it?
    3. Is the data that you collected representative of the problem to be solved?
    4. What additional work is required to manipulate and work with the data?
- Deriving the answer:
    1. In what way can the data be visualized to get to the answer that is required?
    2. Does the model used really answer the initial question or does it need to be adjusted?
    3. Can you put the model into practice?
    4. Can you get constructive feedback into answering the question?

Establishing a clearly defined question starts with understanding the GOAL of the person who is asking the question. For example, if a business owner asks: "How can we reduce the costs of performing an activity?" We need to understand, is the goal to improve the efficiency of the activity? Or is it to increase the businesses profitability? Once the goal is clarified, the next piece of the puzzle is to figure out the objectives that are in support of the goal. By breaking down the objectives, structured discussions can take place where priorities can be identified in a way that can lead to organizing and planning on how to tackle the problem. Depending on the problem, different stakeholders will need to be engaged in the discussion to help determine requirements and clarify questions.

Data Science Methodology Flowchart:

- The data science methodology discussed in this course has been outlined by John Rollins, a seasoned and senior data scientist currently practising at IBM



- Properties:

    1. The flowchart is highly iterative.
    2. The flowchart never ends.

- Entities:

    1. **Business Understanding**: Helps clarify the goal of the entity asking the question

    2. **Analytic Approach**: Helps identify what type of patterns will be needed to address the question most effectively.

    3. **Data Requirements**: We learned that the chosen analytic approach determines the data requirements. Specifically, the analytic methods to be used require certain data content, formats and representations, guided by domain knowledge.

    4. **Data Collection**: When collecting data, it is alright to defer decisions about unavailable data, and attempt to acquire it at a later stage

5. **Data Understanding**:
    - Encompasses all activities related to constructing the dataset
    - Answer the question: "*Is the data that you collected representative of the problem to be solved?*"
    - In the case study, working through the Data Understanding stage, it was revealed that the initial definition was not capturing all of the congestive heart failure admissions that were expected, based on clinical experience --> TRUE
6. **Data Preparation**:
    - Answer the question: "*What are the ways in which data is prepared?*"
    - To work effectively with the data, it must be prepared in a way that addresses missing or invalid values and removes duplicates, toward ensuring that everything is properly formatted.
    - *The Data Preparation stage involves correcting invalid values and addressing outliers*
    - `Feature engineering` is also part of data preparation.
        - *It is the process of using domain knowledge of the data to create features that make the machine learning algorithms work*
        - A feature is a characteristic that might help when solving a problem.
        - Features within the data are important to predictive models and will influence the results you want to achieve.
        - *Feature engineering is critical when machine learning tools are being applied to analyse the data*
    - Most-timing consumption along with D-Collection and D-Understanding (about 70% or even 90% of the overall project time)
    - Automating some of the data collection and preparation processes in the database, can reduce this time to as little as 50% -> Increase time for data scientists to focus on creating models
7. **Modelling**:
    - Focus on developing models that are either descriptive or predictive
    - Example of a descriptive model might examine things like: if a person did this, then they're likely to prefer that

- A predictive model tries to yield yes/no, or stop/go type outcomes. These models are based on the analytic approach that was taken, either statistically driven or machine learning driven
- *Type I Error (false-positive)*: e.g., when a true, non-readmission is misclassified, and action is taken to reduce that patient's risk, the cost of that error is the wasted intervention
- *Type II Error (false-negative)*: e.g., when a true readmission is misclassified, and no action is taken to reduce that risk, then the cost of that error is the readmission and all its attended costs, plus the trauma to the patient
- *Training set is used for predictive modelling*

8. **Evaluation**:

    - Answers the question: "*Does the model used really answer the initial question or does it need to be adjusted?*"
    - Allows the quality of the model to be assessed but it's also an opportunity to see if it meets the initial request.
    - Model evaluation can have two main phases.
        - `Diagnostic measures phase`: *Ensure the model is working as intended*. If the model is a predictive model, a decision tree can be used to evaluate if the answer the model can output, is aligned to the initial design. It can be used to see where there are areas that require adjustments. If the model is a descriptive model, one in which relationships are being assessed, then a testing set with known outcomes can be applied, and the model can be refined as needed.
        - `Statistical significance testing`: *Ensure that the data is being properly handled and interpreted within the model*. This is designed to avoid unnecessary second guessing when the answer is revealed.
    - Also: *Model Evaluation includes ensuring the model is designed as intended*
    - *ROC (Receiver Operating Characteristic) curve*:
        - *A useful diagnostic tool in determining the optimal classification model*
        - First developed during World War II to detect enemy aircraft on radar

- ROC quantifies how well a binary classification model performs, declassifying the yes and no outcomes when some discrimination criterion is varied
- By plotting the true-positive rate against the false-positive rate for different values of the relative misclassification cost, the ROC curve helped in selecting the optimal model.

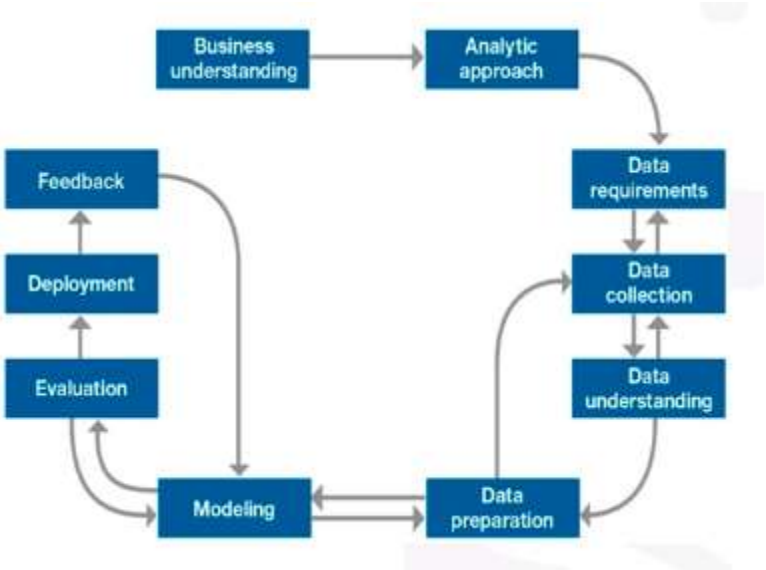
9. **Deployment**

- Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test
- Depending on the purpose of the model, it may be rolled out to a limited group of users or in a test environment, to build up confidence in applying the outcome for use across the board

10. **Feedback**

- Once in play, feedback from the users will help to refine the model and assess it for performance and impact
- The value of the model will be dependent on successfully incorporating feedback and making adjustments for as long as the solution is required
- Throughout the Data Science Methodology, each step sets the stage for the next. *Making the methodology cyclical, ensures refinement at each stage in the game*
- The feedback process is rooted in the notion that, the more you know, the more that you'll want to know (John Rollins)
- We also have to allow for the possibility that other refinements might present themselves during the feedback stage. Also, the intervention actions and processes would be reviewed and very likely refined as well, based on the experience and knowledge gained through initial deployment and feedback
- Finally, the refined model and intervention actions would be redeployed, with the feedback process continued throughout the life of the Intervention program

Look again at the flowchart:



- Properties:

    1. The flowchart is *highly iterative*: All the stages of the Methodology is iterative!
    2. The flowchart *never* ends

- Summary:

    - Its purpose is to explain how to look at a problem, work with data in support of solving the problem, and come up with an answer that addresses the root problem.
    - By answering 10 simple questions methodically, we've taught you that a methodology can help you solve not only your data science problems, but also any other problem.
    - Your success within the data science field depends on your ability to apply the right tools, at the right time, in the right order, to the address the right problem