**Data Science Methodology**

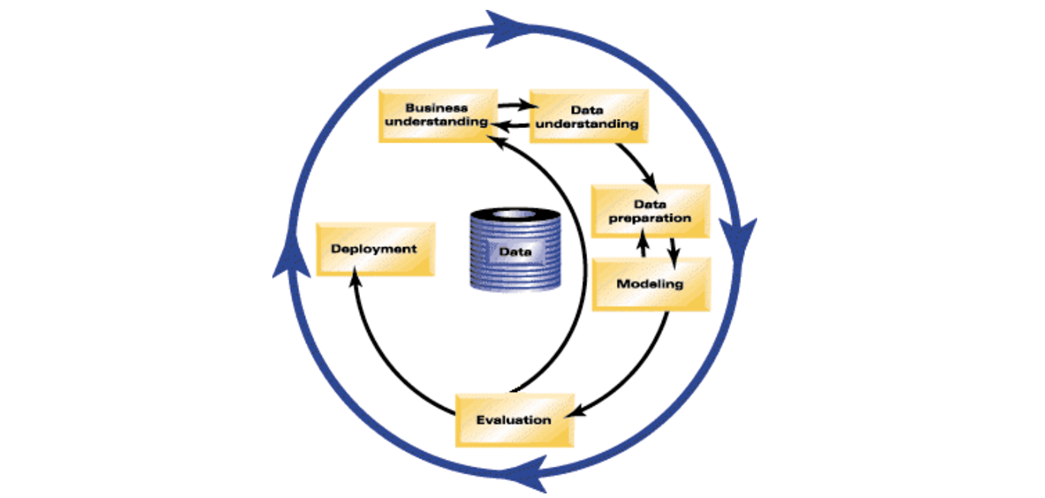
1. Methodology: a system of methods used in a particular area of study or activity.
2. 10 basic questions:
   1. What is he problem that you are trying to solve?
   2. How can you use data to answer the questions?
   3. What data do you need to answer the question?
   4. Where is the data coming from (identify all sources) and how will you get it?
   5. Is the data that you collected representative of the problem to be solved?
   6. What additional work is required to manipulate and work with the data?
   7. In what way can he data be visualized to get to the answer that is required
   8. Does the model used really answer the initial question or does it need to be adjusted?
   9. Can you put the model into practice?
   10. Can you get constructive feedback into answering the question?

**What is CRISP-DM?**

1. Process aimed at increasing the use of data mining over a wide variety of business applications and industries.
   1. intent is to take case specifi scenarios and general behaviors to make them domain neutral
2. Comprised of six steps with an entity that has to be implemented in order to have a reasonable chance of success

**Six steps**

1. Business understanding
   1. Most important – where the intention of the project is outlined
   2. Foundational Methodology and CRIPS-DM are aligned here.
   3. Stakeholders have different objectives and biases, without clear, concise and complete perspective of wht the project goal are, resources will be needlessly expended
2. Data Understanding
   1. Relies on business understanding.
   2. Data is collected at this stage and understanding of what the business wants and needs will determine what data is collected, from what sources, and by what methods
   3. CRISP-DM combines the stages of Data Requirements, Data Collection and Data Understanding
3. Data Preparation
   1. Once collected, it must be transformed into a useable subset
   2. Once a dataset is chosen it must be checked for missing or ambiguous cases.
4. Modeling
   1. Once prepared for use, data must be expressed through whatever appropriate models, give meaningful insights and new knowledge
   2. Purpose of data mining: to create knowledge information that has meaning and utility. Models reveal patterns and structures within the data that provide insight into the features of interest.
   3. Models selected on a portion of the data and adjustments made if necessary
5. Evaluation
   1. Selected model must be tested
   2. Usually done by having pre-selected test, set to run the trained model on
   3. This will allow you to see the effectiveness of the model on a set
   4. Results determine efficacy of the model and foreshadows its role in the next and final stage
6. Deployment
   1. Model is used on new data outside the scope of the dataset and by new stakeholders.
   2. New interactions may reveal the new variables and needs for the dataset and model
   3. New challenges could initiate revision of either business needs and actions, or the model and data, or both



**Business Understanding**

1. Seek clarification – clarity around the problem to be solved will allow you to choose which data will be used to answer the core question
2. Have a clearly defined question – what is the goal?
3. Figure out the objectives of the goal
   1. Break down objectives
   2. Stakeholders involved to determine requirements and clarify questions

**Analytic Approach**

1. Selecting the right analytic approach depends on the question being asked
2. Second stage of data science methodology
3. Identify what type of patterns will be needed to address the question most effectively
   1. Question is to determine probabilities: Predictive model
   2. Show relationships: Descriptive model
   3. Requires yes/no: Classification model
4. Machine learning
   1. Learning without being explitly programmed
   2. Identifies relationships and trends in data that might otherwise not be accessible or indentified
   3. Human behaviour? Uses clustering association approaches

**Data Requirements**

1. Vital to define the data requirements for decision tree classification
   1. Content
   2. Format
   3. Sources

**From Understanding to Preparation**

**Data Understanding**

1. All activities relating to constructing the data set
2. Is the data you collected representative of the problem needing to be solved?
3. Descriptive statistics
   1. Univariate statistics: Hirst, univariants, statistics on each variable (mean, median, minimum, maximum, standard deviation
   2. Pairwise correlations: how closely certain variables are related
   3. Histograms: understand distributions. How values or variables are distributed and which sorts of data preparation more useful in a model
      1. Used to assess data quality
      2. Missing values, invalid or misleading

**Data Preparation**

* + - 1. Most time consuming (70-90%)
      2. Automation can reduce this time to 50%
      3. Process of getting the data into a state where it may be easier to work with
         1. Invalid values
         2. Missing data
         3. Remove duplicates
         4. Formatting
      4. Feature engineering: Process of using domain knowledge of the data to create features that make the machine learning algorithms work
      5. Text: when working with text text anaylsis steps are reuired to be able to manipulate the data
         1. Must know what you are looking for to address the question
         2. Proper groupings are set
         3. Not overlooking what is hidden within

**Modeling to Evaluation**

1. Geared towards answering two key questions:
   1. What is the purpose of data modeling
   2. What are the characteristics of this process
2. Data modeling focuses on creating models that are either descriptive or predictive
   1. Descriptive: if a person did this, they are likely to prefer that
   2. Predictive: Yes/no stop/go outcomes
3. Data scientist uses training sets for predictive modeling
   1. Set of historical data where the outcomes are already known
   2. Does the model need to be calibrated?
4. The framework is geared to do three things
   1. Understand the question
   2. Select an analytic approach or method to solve the problem
   3. Obtain, understand, prepare and model the data
5. End goal is to build a model to answer the question
6. Model evaluation, deployment and feedback loops ensure the answer is near and relevant

**Evaluation**

1. Modeling and evaluation stages are done iteratively
2. Model evaluation is performed during model development and before deployment
3. Modeling: In what way can the data visualized get to the answer that is required?
4. Evaluation: Does the model used really answer the initial question or does it need to be adjusted?
5. Model Evaluation has two main phases
   1. Diagnostic measures phase: Ensure the model is working as intended
      1. Predictive model: Decision tree can be used
      2. Descriptive (relationships are being assessed): Testing set with known outcomes
   2. Statistical significance testing: Properly handled and interpreted
6. ROC: Reciever Operative Characteristic Curve
   1. Optimal classification model

**From Deployment to Feedback**

1. Stakeholders familiar with the tool produced
2. Solution owner, marketing, developers, administration
3. Assimilate knowledge and gain trust

**Feedback**

1. Problem solved? Question answered?
2. Actual real time use in the field
   1. Once it is deployed and tested
3. Measure effectiveness of intervention
4. Refine model: After all data is compiled

**Summary**

1. Forming a concrete business or research model
2. Collecting and analyzing data
3. Building a model
4. Understanding the feedback after model deployment
5. Understanding the question, business goals and objectives
6. Picking the most effective analytic approach
7. Learning to work with data
   1. Determining the data requirements
   2. Collecting the appropriate data
   3. Understanding the data
   4. Preparing the data for modeling
8. Once approach is selected
   1. Evaluating an deploying the model
   2. Getting feedback constructively
   3. Using that feedback constructively so as to improve the model
9. THE STAGES OF THIS METHODOLOGY ARE ITERATIVE
   1. It can always be improved

**Data Collection**

1. Data revised and decisions made as to whther the requirements require less or more data
2. Descriptive techniques and visualization can be applied to assess
   1. Content
   2. Quality
   3. Initial incites
3. Gaps identified
   1. Plans to either fill or substitutes
4. Gathering data
   1. Requires you know the source
   2. Or find data elements that are needed
5. You can defer data and acquire it at a later stage
6. Eliminate