# BIT34503 Data Science

## CHAPTER 2: DATA SCIENCE METHODOLOGY

# Content Title

2.0 The Need for Data Science Methodologies and Frameworks
2.1 Identify problem
2.2 Define question
2.3 Define ideal dataset
2.4 Obtain data
2.5 Analyze data
2.6 Interpret results
2.7 Distribute results

# Data Science Process

What is the goal?

Do you need to classify, estimate, describe?

Do you have the proper data?

What actions are planned?

Are there anomalies or patterns?

How the data look?

Do you have too many or too few variables?

Do you need to impute/transform the data?

Do you need to aggregate/create the data?

What did you learn?

Can you explain the answer with the model?

Can you tell a story?

Can you deploy the model in time?


Which data are relevant?

How many sources are involved?

Do you have access to the data?

Do you have privacy issues?

Will the data be available in production?

Train different models (algorithms and approaches).

Validate all the models.

Test all the models.

Select the best model according to the question/goal.

Score the champion model.

# The data science process

The data science lifecycle is a structured process that data scientists follow to extract insights and knowledge from data. It encompasses various stages and activities from the initial problem definition to the deployment of data-driven solutions. Here's an overview of the data science lifecycle:

1) **The question/Problem Definition:** The process begins with clearly defining the problem or goal. What are you trying to solve or achieve with data? These include understanding the business context, the application domain, the relevant prior knowledge, the goals of the end-user, business objectives, and project parameters.

2) **Data Collection:** This stage involves gathering the necessary data from various sources. It could be structured data from databases, unstructured data from text or images, or external data from APIs and web scraping. Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, from which database, etc.

# The data science process(cont.)

**3) Explore the data:** may involve these processes-**Data Cleaning, Exploratory Data Analysis (EDA), and Feature Engineering.**

- **Data Cleaning -**Raw data often contains errors, missing values, or inconsistencies. Data cleaning involves preprocessing and transforming the data to ensure it's accurate and ready for analysis.
- **Exploratory Data Analysis (EDA) -** to visualize and analyze the data to identify patterns, correlations, and potential insights. EDA helps in understanding the data's characteristics and relationships.
- **Feature Engineering -** creating new variables or features that may be more informative for modelling. It involves selecting, transforming, or combining existing features.

# The data science process (cont.)

**4) Model the data:** may involve **model development** & **model evaluation**.
This is the heart of data science, where predictive or descriptive models are built using machine learning algorithms.

- **model development** -Models are trained on historical data to make predictions or gain insights. Selecting the method to be used for searching for patterns in the data. Deciding which models and parameters may be appropriate
- **model evaluation** - Performance evaluation, study models accuracy, do the results make sense in the context of the problem?

**5) An Answer -** Effective communication of findings and insights is a vital aspect of data science. Data scientists need to explain their results to non-technical stakeholders in a clear and understandable way. This phase also involve **model deployment.** Once a satisfactory model is achieved, it can be deployed in a production environment, making it available for making real-time predictions or decisions.

# Advanced Analytics

- Advanced analytics comprises a set of different techniques used to solve business problems:
  - machine learning
  - statistical analysis
  - forecast
  - text analytics
  - optimization

Machine Learning  Statistical Analysis  Forecast  Text Analytics  Optimization

# The techniques/models can be grouped

1) Functions
    a) Supervised learning
    b) Unsupervised learning

2) Tasks
    a) Classification
    b) Prediction/regression
    c) Association analysis
    d) Clustering

# Models/tools/techniques/approaches:

- **Linear regression**
- **Logistic regression**
- K-Nearest neighbours
- K-means
- **Decision Trees**
- Neural networks
- Fuzzy set theory
- Rule induction
- Evolutionary algorithms
- Swarm Optimization algorithm
- Etc

**Some models are better than others….depending on the dataset used or the problems to be solved**

# Advanced Analytics in Real-World Examples

- Advanced analytics aims to help companies in addressing some business challenges:
  - customer management
  - revenue optimization
  - network analytics
  - data monetization

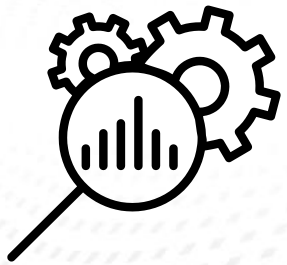
# Customer Management

- **A deeper understanding of the customer experience**
  - Enhanced customer experience
    - Offer more personalized and relevant customer promotions.
  - Churn prediction and prevention
    - Forecast customer issues and prevent them from happening.
  - Next-best-offer
    - Operationalize customer insights by using structured and unstructured data.
  - Target social influencers
    - Identify and track relationships between customers, and target those with the most influence.

# Revenue Optimization

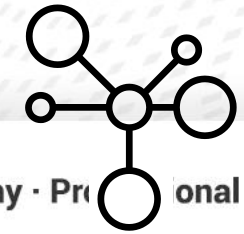
## Better forecast accuracy drives better decisions

- Apply advanced analytics to improve decision-making processes related to product bundles and marketing campaigns.
- Identify situations leading to revenue leakage, whether due to billing and collections, network, or fraud issues.
- Create new and more effective rate plans and bundles.



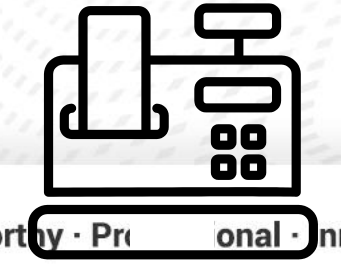
# Network Analytics

- **Better network performance**
  - Network capacity planning
    - Use statistical forecasting and detailed network/supply chain data to more accurately plan capacity.
  - Service assurance and optimization
    - Prevent network/supply chain problems before they happen.
  - Use structured and unstructured data for deeper customer and service performance insights
    - Optimize call center staffing and identify operational changes that could lower the cost to serve while improving service quality.



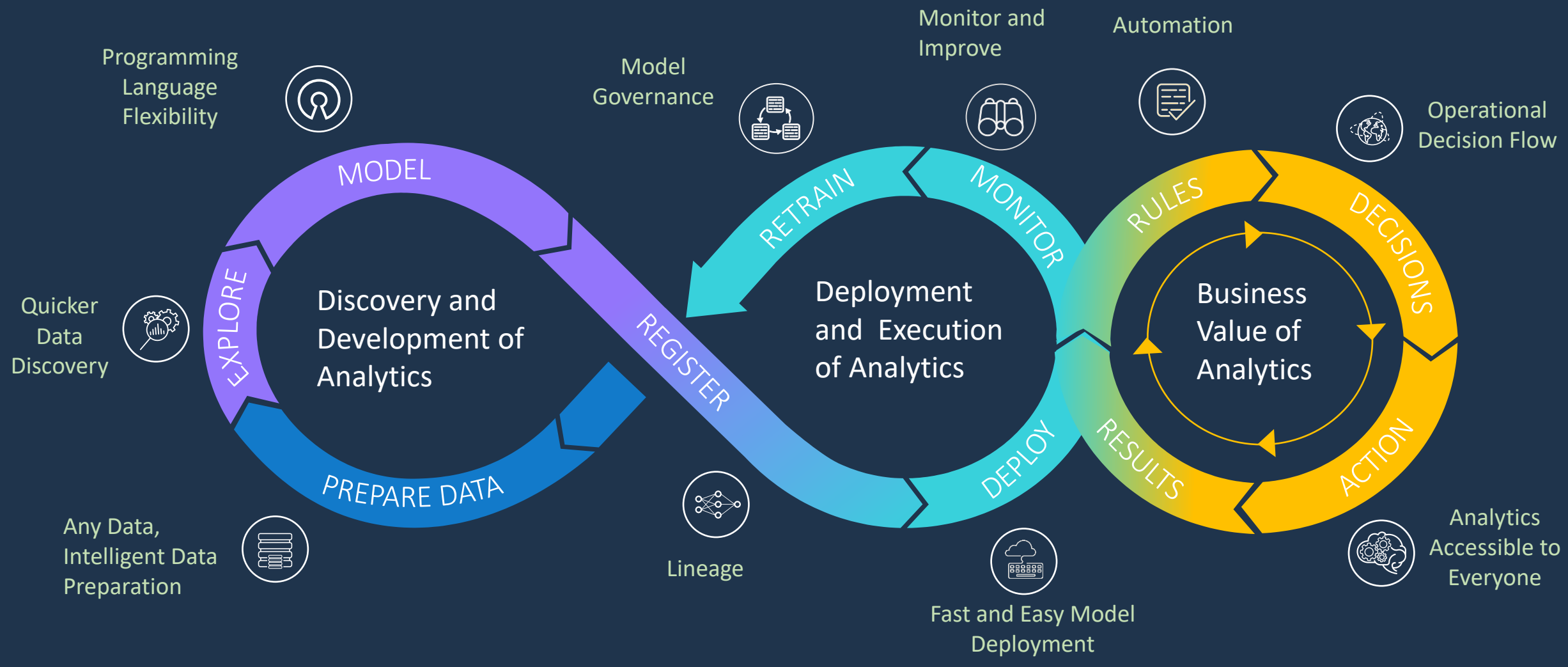
# Data Monetization

- Customer and location-specific campaigns
  - Location-based marketing
    - Develop specialized offers and promotions that are delivered to targeted customers via their mobile devices.
  - Micro segmentation
    - Create highly detailed customer segments that you can use to send highly targeted and timely mobile messages.
  - Real-time data analysis
    - Analyze real-time data streams from wireless subscribers so that you can keep campaigns relevant and effective.



# Data alone does not drive organizations. Decisions do.

# Operationalizing Analytics – from Data to Insight to Decision to ROI



# The Need for Data Science Methodologies and Frameworks

- The field of data science has matured greatly in the past decade.
- And yet, teams often struggle to apply an appropriate data science methodology and team-based collaboration framework.
- Consider the following issues:
  - **Production Models**
  - **Team Approach**
  - **Agile Approach**
  - **Ad Hoc Approaches**
  - **Hybrid Approaches**

- **Production Models:**
  - A one-off model often does not provide sustained value.
  - Rather organizations often need a sustainable productized system that delivers model results over a longer period of time.
  - This necessitates a solid data science methodology that expands beyond just model development and into machine learning operations.
- **Team Approach:**
  - Data science is increasingly becoming a team sport.
  - The concept of a back office unicorn data scientist working in isolation is not the norm.
  - But rather, many projects have a diverse team consisting of multiple team roles.
  - Thus, teams need to leverage a modern team-based approach to coordinate their work.
- **Agile Approach:**
  - Data science is a highly iterative process — especially once you extend beyond the classroom and into the real-world with real-time changing market conditions, technological shifts, and ever-evolving business needs.
  - Long and inflexible upfront planning processes won't work.
  - Rather, this gives rise to agile approaches.
  - But how can you effectively apply Agility to data science?

- **Ad Hoc Approaches**
  - Ad hoc processes focus on delivering a specific implementation without concern for broader impact or repeatable processes.
  - In short, we can just “wing it”.
  - This approach may work well for one-off, smaller, and low-impact projects.
  - Think of a toy side project or an academic exercise.
  - Yet, the appropriate use cases for ad hoc in the real world are becoming less frequent.
  - Unfortunately, many people still just result to Ad Hoc.
- **Hybrid Approaches**
  - The reality is that we can mix and match various approaches to design a comprehensive methodology that best suits your team, projects, and organizational needs.
  - This guide highlights two such hybrid approaches — each serving different use cases.

# Data Science Methodologies

- A data science life cycle (also known as a data science methodology) describes the step-by-step approach we take to deliver a project.
- Data scientists (even if they have not explicitly studied various methodologies) intuitively understand these steps.
- Documenting them can help increase repeatably and prevent us from forgetting a step.
- This is increasingly important in the world of distributed teams that extend beyond data science to areas such as legal or business.

- Data Science Methodology indicates the routine for finding solutions to a specific problem.
- This is a cyclic process that undergoes a critic behavior guiding business analysts and data scientists to act accordingly.
- There are dozens of different defined data science methodologies.
- This guide explores the most well-known.

| Approach | Description | Strengths | Challenges | Best For... |
|---|---|---|---|---|
| Waterfall | Plan your work. Work your plan | • Easily understood<br>• Matches traditional corporate culture | • Inflexible<br>• Delays testing<br>• Documentation heavy | • Avoid for data science |
| KDD<br>(Knowledge Discovery in Databases) | 5 Phases from Selection to Evaluation | • Decent explanation of core data mining technical project | • Outdated<br>• Ignores teams<br>• Many same shortcomings as Waterfall<br>• Ignores biz understanding & deployment | • "Toy" projects with a well-defined scope that don't need productized |
| SEMMA<br>(**S**ample, **E**xplore, **M**odify, **M**odel, and **A**ssess) | 5 Phases from Sample to Assess | • Decent explanation of core data mining technical project | • Outdated<br>• Ignores teams<br>• Many same shortcomings as Waterfall<br>• Ignores biz understanding & deployment | • "Toy" projects with a well-defined scope that don't need productized |

| Approach | Description | Strengths | Challenges | Best For... |
|---|---|---|---|---|
| CRISP-DM (CRoss Industry Standard Process for Data Mining) | 6 Phases from Business Understanding to Deployment | • Well-known<br>• More comprehensive than KDD, SEMMA<br>• Defined guide | • Outdated<br>• Ignores teams<br>• Many same shortcomings as Waterfall | • Teams looking for an established practice |
| TDSP (Team Data Science Process) | Combines CRISP-DM and Scrum practices | • Comprehensive open-source documentation | • Includes Agile concepts<br>• Strong team focus | • Teams looking to "modernize" CRISP-DM |
| Domino | Combines CRISP-DM and Agile practices | • Visual roadmap with clear flow and decision points<br>• Includes practical tips | • More of a concept as opposed to a fully vetted approach | • Teams looking to "modernize" CRISP-DM |
| Others | Lesser-known life cycles | • Each includes a novel viewpoint | • Not well-known or vetted | • Good "food for thought" |

# Agile Coordination Frameworks

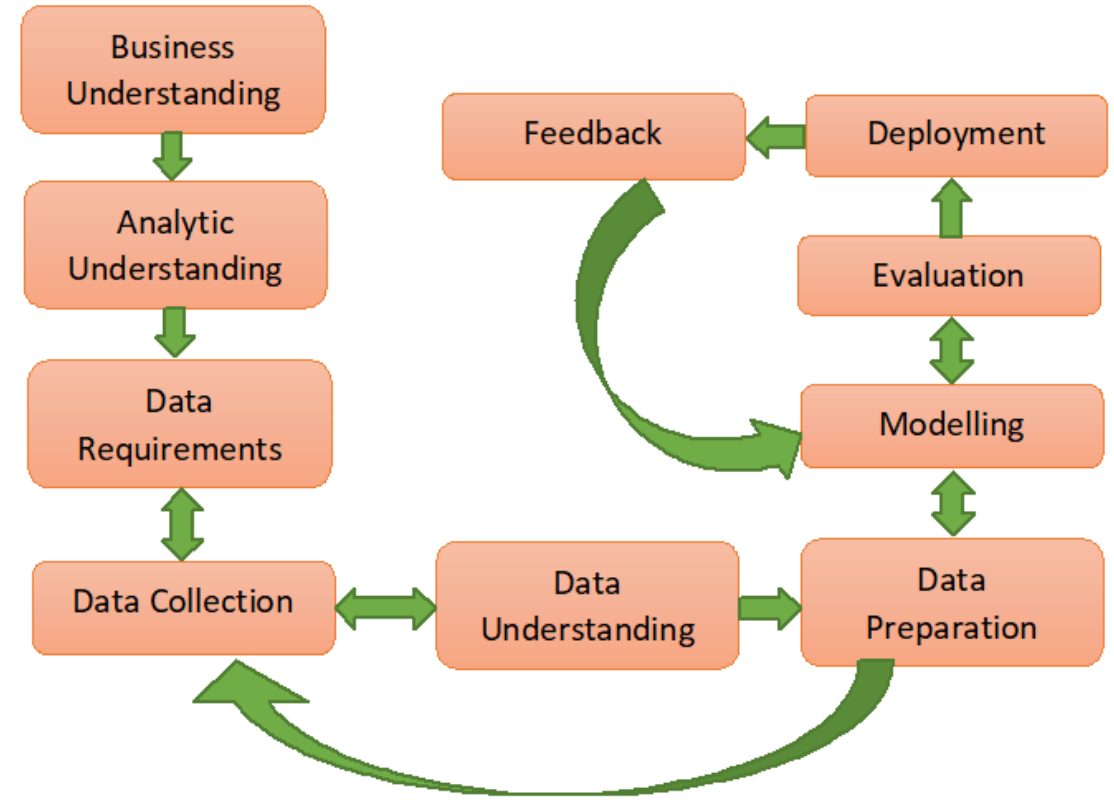
- Agility has taken over the software engineering world.
- Yet, it gets a mixed review for data science.
- However, Agile and data science should go hand-in-hand.
- Don’t focus too much on the specific approach, but rather start with the fundamental principles your team aspires.
- From there, build a framework on top of it that defines how we can sustain team collaboration while also being flexible enough to shift the project’s focus.
- Here are three agile frameworks that can be considered:
  - Kanban is borrowed from manufacturing.
  - Scrum from software.
  - Data Driven Scrum was designed specifically data science.

| Approach | Description | Strengths | Challenges | Best For... |
|---|---|---|---|---|
| Kanban | Visualize flow. Minimize work-in-progress. | • Simple<br>• Combines well with other frameworks<br>• Maximizes throughput<br>• Minimizes waste | • Least definitive<br>• Lots of ambiguity | • Starting with a solid core set of principles and building a framework on top of it |
| Scrum | Well-known Agile approach focused on fixed-length iterations | • Quick, incremental value focus<br>• Well-defined feedback loop<br>• Strong team focus | • Time-boxing can be restrictive<br>• Often poorly implemented<br>• Management might get in the way | • Agile teams who need discipline provided by fixed time cycles<br>• Radical innovation cultures |
| Data Driven Scrum | Agile framework specifically designed for data science teams | • Most of same benefits of Scrum and Kanban<br>•Caters to experimentation<br>• Relaxes Scrum pain points | • Not as vetted as Scrum<br>• Adds challenges of managing concurrent iterations | • Teams with strong experimental culture<br>• Data science teams that struggled with Scrum |

# Hybrid Approaches

| Approach | Description | Strengths | Challenges | Best For... |
|---|---|---|---|---|
| Waterfall-Agile | Attempts to combine best of Agile and waterfall | • Allows for some flexibility while catering to broader constraints | • "Best of both worlds" can water down the advantages from either | • Highly-regulated projects that require rigid administrative processes |
| Research & Development | Combines open-research phases followed by structured development | • Gives flexibility for open-ended research<br>• Adds structure when needed to coordinate deliverables | • Difficult to monitor<br>• Can suffer from ad hoc chaos<br>• High trust and discipline required | • Mature teams who don't need heavy oversight<br>• Research-focused teams needing freedom |

1. Identify problem
2. Define Question
3. Define Ideal Dataset
4. Obtain data




Data Science Methodology

Ref: https://www.ibm.com/blogs/journey-to-ai/

# Business Understanding

- Before solving any problem in the Business domain it needs to be understood properly.
- Business understanding forms a concrete base, which further leads to easy resolution of queries.
- We should have the clarity of what is the exact problem we are going to solve.

# From Problem to Approach

- Every customer's request starts with a problem, and Data Scientists' job is first to understand it and approach this problem with statistical and machine learning techniques.
- The **Business Understanding** stage is crucial because it helps to clarify the goal of the customer.
- In this stage, we have to ask a lot of questions to the customer about every single aspect of the problem; in this manner, we are sure that we will study data related, and at the end of this stage, we will have a list of **business requirements**.

# Analytic Understanding/Approach

- Based on the above business understanding one should decide the analytical approach to follow.
- The approaches can be of 4 types:
  - Descriptive approach (current status and information provided)
  - Diagnostic approach(a.k.a statistical analysis, what is happening and why it is happening)
  - Predictive approach(it forecasts on the trends or future events probability)
  - Prescriptive approach( how the problem should be solved actually).

- The next step is the **Analytic Approach**, where, once the business problem has been clearly stated, the data scientist can define the analytic approach to solve the problem.
- This step entails expressing the problem in the context of statistical and machine-learning techniques, and it is essential because it helps identify what type of patterns will be needed to address the question most effectively.
- If the issue is to determine the probabilities of something, then a predictive model might be used; if the question is to show relationships, a descriptive approach may be required, and if our problem requires counts, then statistical analysis is the best way to solve it.
- For each type of approach, we can use different algorithms.

# Data Requirements

- The above chosen analytical method indicates the necessary data content, formats and sources to be gathered.
- During the process of data requirements, one should find the answers for questions like ‘what’, ‘where’, ‘when’, ‘why’, ‘how’ & ‘who’.

# From Requirements to Collection

- Once we have found a way to solve our problem, we will need to discover the correct data for our model.
- **Data Requirements** is the stage where we identify the necessary data content, formats, and sources for initial data collection, and we use this data inside the algorithm of the approach we chose.

# Data Collection

- Data collected can be obtained in any random format.
- As a result, according to the approach chosen and the output to be obtained, the data collected should be validated.
- Thus, if required one can gather more data or discard the irrelevant data.

- In the **Data Collection** Stage, data scientists identify the available data resources relevant to the problem domain.
- To retrieve data, we can do web scraping on a related website, or we can use repository with premade datasets ready to use.
- Usually, premade datasets are CSV files or Excel; anyway, if we want to collect data from any website or repository, we should use Pandas, a useful tool to download, convert, and modify datasets.

- Here is an example of the data collection stage with pandas.

```
import pandas as pd # download library to read data into dataframe

pd.set_option('display.max_column', None)
dataframe = pd.read_csv("csv_file_url")
print("Data read into dataframe!")

dataframe.head() # show the first few rows
dataframe.shape # get the dimensions of the dataframe
```

# Data Preparation

- Let’s understand this by connecting this concept with two analogies.
  - One is to wash freshly picked vegetables and second is only taking the wanted items to eat in the plate during the buffet.
  - First analogy:
    - Washing of vegetables indicates the removal of dirt i.e. unwanted materials from the data. Here noise removal is done.
  - Second analogy:
    - Taking only eatable items in the plate is, if we don’t need specific data then we should not consider it for further process.
    - This whole process includes transformation, normalization etc.

# From Understanding to Preparation

- Now that the data collection stage is complete, data scientists use descriptive statistics and visualization techniques to understand data better.
- Data scientists, explore the dataset to understand its content, determine if additional data is necessary to fill any gaps but also to verify the quality of the data.
- In the **Data Understanding** stage, data scientists try to understand more about the data collected before.
- We must check the type of each data and to learn more about the attributes and their names.

# Example codes:

```
# get all columns from a dataframe and put them into a list
attributes = list(dataframe.columns.values)

# then we check if a column exist and what is its name.
print([match.group(0) for attributes in attributes for match in
[(re.compile(".*(column_name_keyword).*")).search(attributes)] if
match])
```

- In the **Data Preparation** stage, data scientists prepare data for modeling, which is one of the most crucial steps because the model has to be clean and without errors.
- In this stage, we have to be sure that the data are in the correct format for the machine learning algorithm we chose in the analytic approach stage.
- The dataframe has to have appropriate columns name, unified boolean value (yes, no or 1, 0).
- We have to pay attention to the name of each data because sometimes they might be written in different characters, but they are the same thing; for example (WaTeR, water), we can fix this making all the value of a column lowercase.
- Another improvement can be made by deleting data exceptions from the dataframe because of their irrelevance.

# Example codes:

```
# replacing all 'yes' values with '1' and 'no' with '0'
dataframe = dataframe.replace(to_replace="Yes", value=1)
dataframe = dataframe.replace(to_replace="No", value=0)

# making all the value of a column lowercase
dataframe["column"] = dataframe["column"].str.lower()
```

# Modelling

- Modelling decides whether the data prepared for processing is appropriate or requires more finishing and seasoning.
- This phase focuses on the building of predictive/descriptive models.

# Evaluation

- Model evaluation is done during model development.
- It checks for the quality of the model to be assessed and also if it meets the business requirements.
- It undergoes diagnostic measure phase (the model works as intended and where are modifications required) and statistical significance testing phase (ensures about proper data handling and interpretation).

# From Modeling to Evaluation

- Once data are prepared for the chosen machine learning algorithm, we are ready for modeling.
- In the **Modeling** stage, the data scientist has the chance to understand if his work is ready to go or if it needs review.
- Modeling focuses on developing models that are either descriptive or predictive, and these models are based on the analytic approach that was taken statistically or through machine learning.
- **Descriptive modeling**
  - It is a mathematical process that describes real-world events and the relationships between factors responsible for them.
  - For example, a descriptive model might examine things like: if a person did this, then they're likely to prefer that.

- **Predictive modeling**
- It is a process that uses data mining and probability to forecast outcomes;
- For example, a predictive model might be used to determine whether an email is a spam or not.
- For predictive modeling, data scientists use a **training set** that is a set of historical data in which the outcomes are already known.
- This step can be repeated more times until the model understands the question and answer to it.
- In the **Model Evaluation** stage, data scientists can evaluate the model in two ways: Hold-Out and Cross-Validation.
- In the Hold-Out method, the dataset is divided into three subsets:
  - a **training set** as we said in the modeling stage.
  - a **validation set** that is a subset used to assess the performance of the model built in the training phase.
  - a **test set** is a subset to evaluate the likely future performance of a model.

Here is an example of modeling and evaluation:

```
# select dataset and training field
data = pd.read_csv("student-mat.csv", sep=";")
data = data[["G1", "G2", "G3", "studytime", "failures", "absences"]]
predict = "G3" # select field to predict

x = np.array(data.drop([predict], 1))
y = np.array(data[predict])

# split the dataset into training and test subsets
x_train, x_test, y_train, y_test =
sklearn.model_selection.train_test_split(x, y, test_size = 0.1)

linear = linear_model.LinearRegression() # create linear regression
model

linear.fit(x_train, y_train) # perform the training of the model
acc = linear.score(x_test, y_test) # calculate the accuracy
print("Accuracy: ", acc) # print the accuracy of the model
```

# Deployment

- As the model is effectively evaluated it is made ready for deployment in the business market.
- Deployment phase checks how much the model can withstand in the external environment and perform superiorly as compared to others.

# Feedback

- Feedback is the necessary purpose which helps in refining the model and accessing its performance and impact.
- Steps involved in feedback define the review process, track the record, measure effectiveness and review with refining.

# From Deployment to Feedback

- Data scientists should make the stakeholders familiar with the tool produced in different scenarios, so once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test.
- The **Deployment** stage depends on the purpose of the model, and it may be rolled out to a limited group of users or in a test environment.
- A real case study example can be for a model destined for the healthcare system; the model can be deployed for some patients with low-risk and after for high-risk patients too.

- The **Feedback** stage is usually made the most from the customer. Customers after the deployment stage can say if the model works for their purposes or not.
- Data scientists take this feedback and decide if they should improve the model; that's because the process from modeling to feedback is highly iterative.
- When the model meets all the requirements of the customer, our data science project is complete.

UTHM
Universiti Tun Hussein Onn Malaysia
1993-2023
AMANAH • PROFESIONAL • INOVATIF

# Thank you