Data Science Methodology

DigitalSkola
Uncover The World Of Digital Skills With Us

Your imagination is
your preview of life's coming attractions.

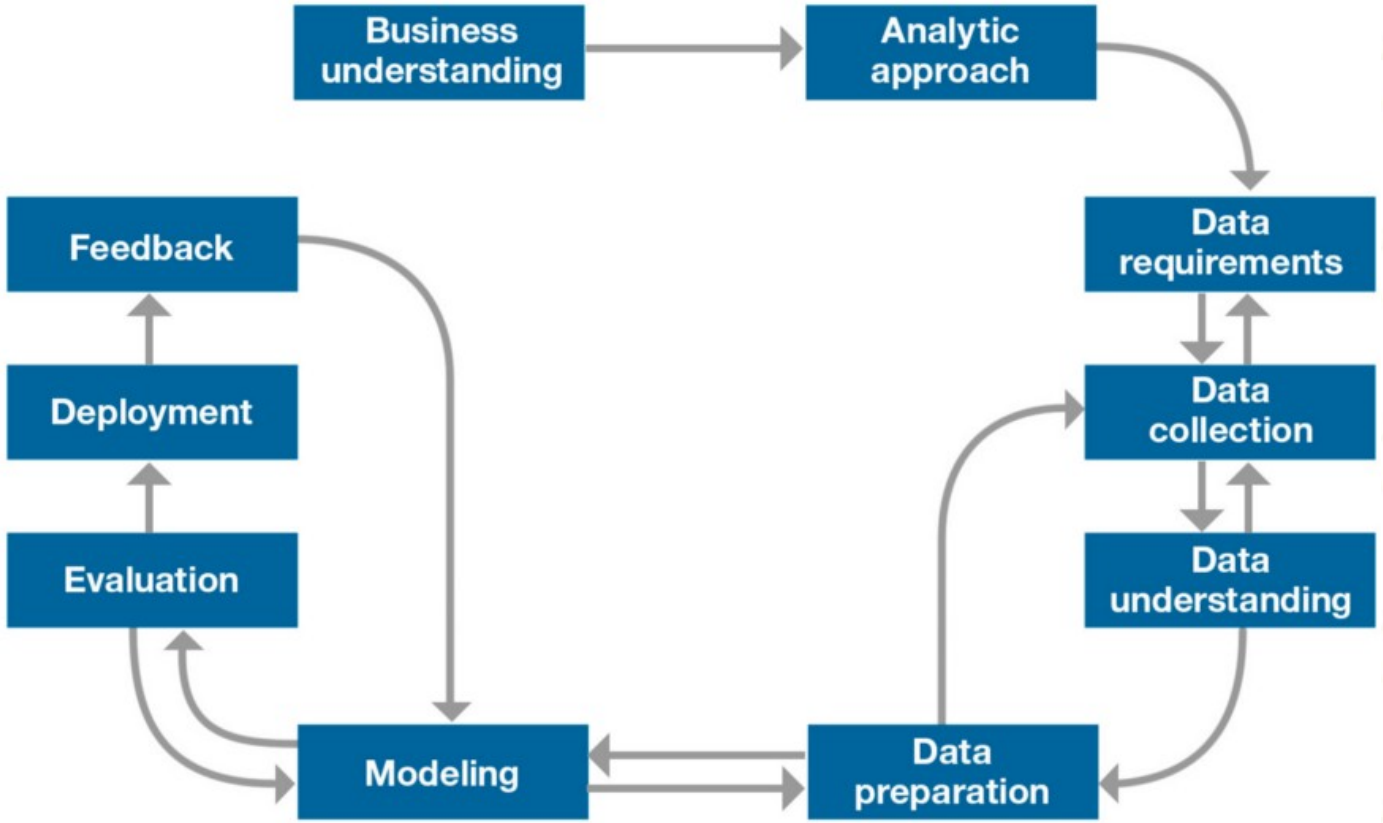
By Albert Einstein

DigitalSkola
Uncover The World Of Digital Skills With Us

Table of Content

Data Science Methodology

Pt. 1



Source: https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science

Data Science Methodology

Pt. 2

Process	Description
Business Understanding	Try to understand current situation & context
Analytics Approach	Choose the analytical approach that fit.
Data Requirements	List down required data
Data Collection	Collect all the data required
Data Understanding	Do exploration to understand the data we have.

Data Science Methodology

Pt. 3

Process	Description
Data Preparation	Start the preparation process and feature engineering
Modelling	Apply algorithm to our data
Model Evaluation	Evaluate model performance
Model Deployment	Deploy the algorithm, thus other service can utilize it
Environment's Feedback	Gather feedback

Business Understanding

1. Before we set the objectives of future project. It's better to have a **solid understanding for current business processes**.
2. List and define business problems, then set the priority.
3. Define business objective
4. Set the success criteria

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.

Business Understanding

Case: a company has stagnant revenue in the last 1 year.

Objective (Option)

1. Increase number of users.
2. Activate churned users.

Success Criteria

1. Get 1000 new users.
2. Activate 1500 churned users.

DigitalSkola
Uncover The World Of Digital Skills With Us

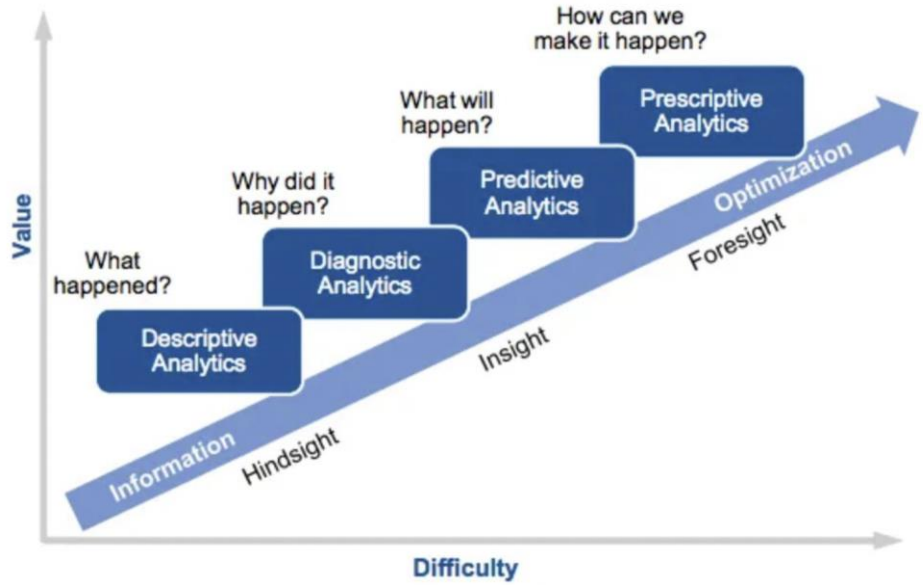
Business Metrics

Digital Marketing	Transactions	Promo
# Visitors	Revenue	Promo Disbursed
# New Visitors	# Transactions	Cost of Promo
# App Installed	# Users	

Analytic Approaches

Once we have lists of business requirements, now we need to select the best analytical approaches that fit the requirements most effectively.

1. Descriptive Analytics
2. Diagnostic Analytics
3. Predictive Analytics
4. Prescriptive Analytics



DigitalSkola
Uncover The World Of Digital Skills With Us

Analytic Approaches
Descriptive Analytics

Describe what happened in previous period.

It's associated with data visualization via reports, dashboards, and scorecards that facilitates decision makings.

Steps are need to be done:

1. State business metrics
2. Identify data required
3. Extract and prepare data
4. Analyze the data
5. Present the data

DigitalSkola
Uncover The World Of Digital Skills With Us

Analytic Approaches
Diagnostic Analytics

Describe why something happened in previous period.

The output of diagnostic analytics is the root cause of something anomaly that happened.

1. Identify the anomaly
2. Discover the root cause
3. Determine causal relationships

DigitalSkola
Uncover The World Of Digital Skills With Us

Analytic Approaches
Predictive Analytics

Predict what will happen in the future.

Utilize descriptive data accumulated over time, and use it to predict incoming events

1. Identify business outcome
2. Determine data required as training data
3. Determine type of analysis
4. Validate result
5. Test predicted data

DigitalSkola
Uncover The World Of Digital Skills With Us

Analytic Approaches
Predictive Analytics

Predict what will happen in the future.

Utilize descriptive data accumulated over time, and use it to predict incoming events

1. Identify business outcome
2. Determine data required as training data
3. Determine type of analysis
4. Validate result
5. Test predicted data

DigitalSkola
Uncover The World Of Digital Skills With Us

Analytic Approaches
Prescriptive Analytics

Predict what will happen in the future.

Utilize descriptive data accumulated over time, and use it to predict incoming events

1. Identify business outcome
2. Determine data required as training data
3. Determine type of analysis
4. Validate result
5. Test predicted data

DigitalSkola
Uncover The World Of Digital Skills With Us

From Business Understanding
To Analytics Approach

Business Understanding

What is the problem we're trying to solve?
Or
What is the question we're trying to answer?

Analytics Approach

How can we use data to achieved our goals?

DigitalSkola
Uncover The World Of Digital Skills With Us

Data Requirements

At this stage, 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 later.

DigitalSkola
Uncover The World Of Digital Skills With Us

From Data Requirements
To Data Collection

Imagine we're a chef in a restaurant and plan to prepare fried rice for a dinner. List of items in ingredient are similar to data requirements. Once it has been listed, our job is to collect the data required to proceed to next process

If we want to create a prediction on whether or not a new user will do repurchase next month,
1. We need to have data of users that doing transaction for **more than 1 month.**
2. We need to have data of **what products that bought by them**.
3. Other supporting data e.g. complaint data etc.

Data Collection

In the Data Collection Stage, data scientists **identify the available data resources** relevant to the problem domain, **all the data** resources in all forms such as structured, unstructured and semi structured **will be collected.**

To retrieve data, we can do web scraping on a related website, or we can use repository with premade datasets ready to use or consume the data directly using API.

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.

DigitalSkola
Uncover The World Of Digital Skills With Us

Data Collection
Sources

Internal

DB Production	DB Events Tracker	Documents
Data Transaksi	Data User Click	File Excel
Data User	Data User Page View	Notes
Data Product	Data User Scroll	

External

Data Public	Data 3rd Party	Scraping
Open Data	Data Survey	File Excel
Data Repository	Data Vendor	Notes
Dashboard		

Data Understanding

In the Data Understanding stage, data scientists try to understand more about the data we've collected previously.
We have to check the type of each data and to learn more about the attributes and their names. We also need to check missing data and anomaly.

Data understanding encompasses **all activities related to constructing the data set.** Essentially, the data understanding section of the data science methodology answers the question: **Is the data that you collected representative of the problem to be solved?**

Data Preparation

Once we understand the attribute, the missing data points and anomaly of the data. We need to do Data Preparation that consist of **data cleaning process, i.e. managing missing data, deleting duplicates, changing the data into a uniform format, etc.**

The expected output of this process is data has no error and has been stored in the correct format for further data exploration.

*only the data needed to solve the problem is retained to make the model run smoothly with minimal errors.

This process also include the **feature engineering** process.

Data Preparation

Data Cleansing
Data duplicated
Missing Data
Different Format

Formatting + Feature Engineering
Check data type
Data Manipulation

Loan ID	User ID	Gender	Martial Status	Children	Education	Job	Salary	Other Info	Loan Term	Loan Amount	Status
N25005	001	Male	Single	0	S1	PNS	10.000.000	+	60	1.000.000	Approved
N25001	002	Male	Single	0	S1	Marketing	8.000.000	+	60	1.000.000	Approved
N23013	003	Female	Married			PNS	5.000.000	+	60	800.000	Rejected
N23013	003	Female	Married			PNS	5.000.000	+	60	800.000	Rejected
N24011	004	Male		Tidak Punya	S1	PNS	10.000.000	+	60	1.000.000	Rejected

Data Preparation

Prepared Data for Modelling

User ID	Name	Third Party Data Valid	OJK Flag Sehat	Emergency Contact Fraud	Total Loan Requester per Region	Flag as Fraud
001	Bayu	1	1	0	S1	0
002	Adhi	1	1	0	S1	0
003	Susi	0	0	11		1
003	Susi	0	0	11		1
004	Aryo	1	1	1	S1	1

Modelling

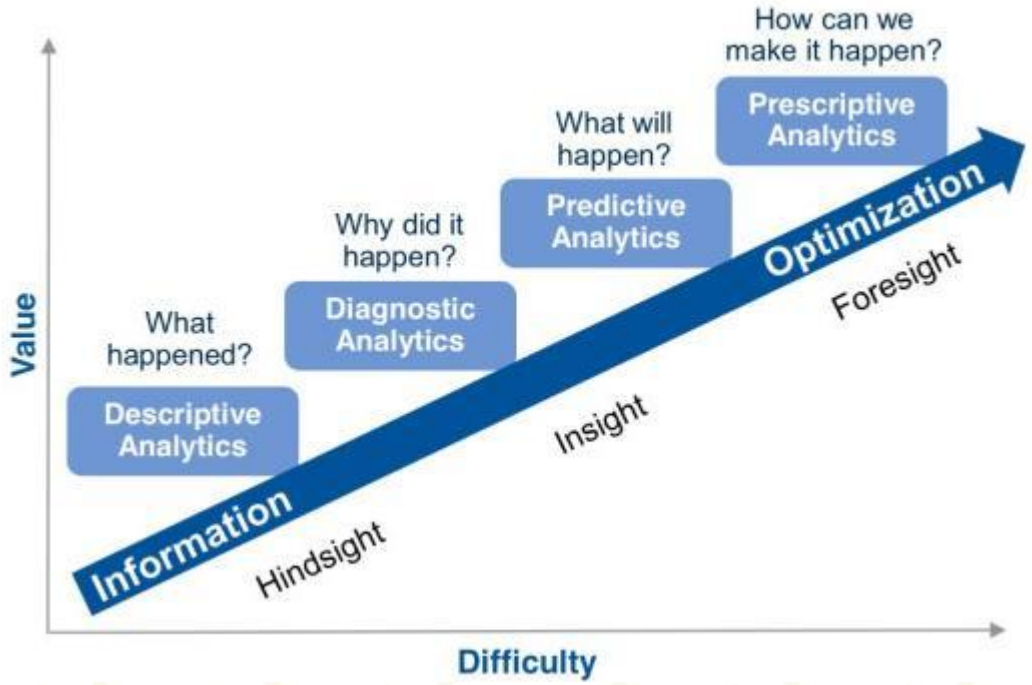
The dataset that has been passed the preparation process are being used in modelling process. Modeling focuses on developing models that are either descriptive or predictive, and these models are based on the analytic approach that was taken statistically or through machine learning. (**Descriptive modeling** is a mathematical process that describes real-world events and the relationships between factors responsible for them, **Predictive modeling** is a process that uses data mining and probability to forecast outcomes)

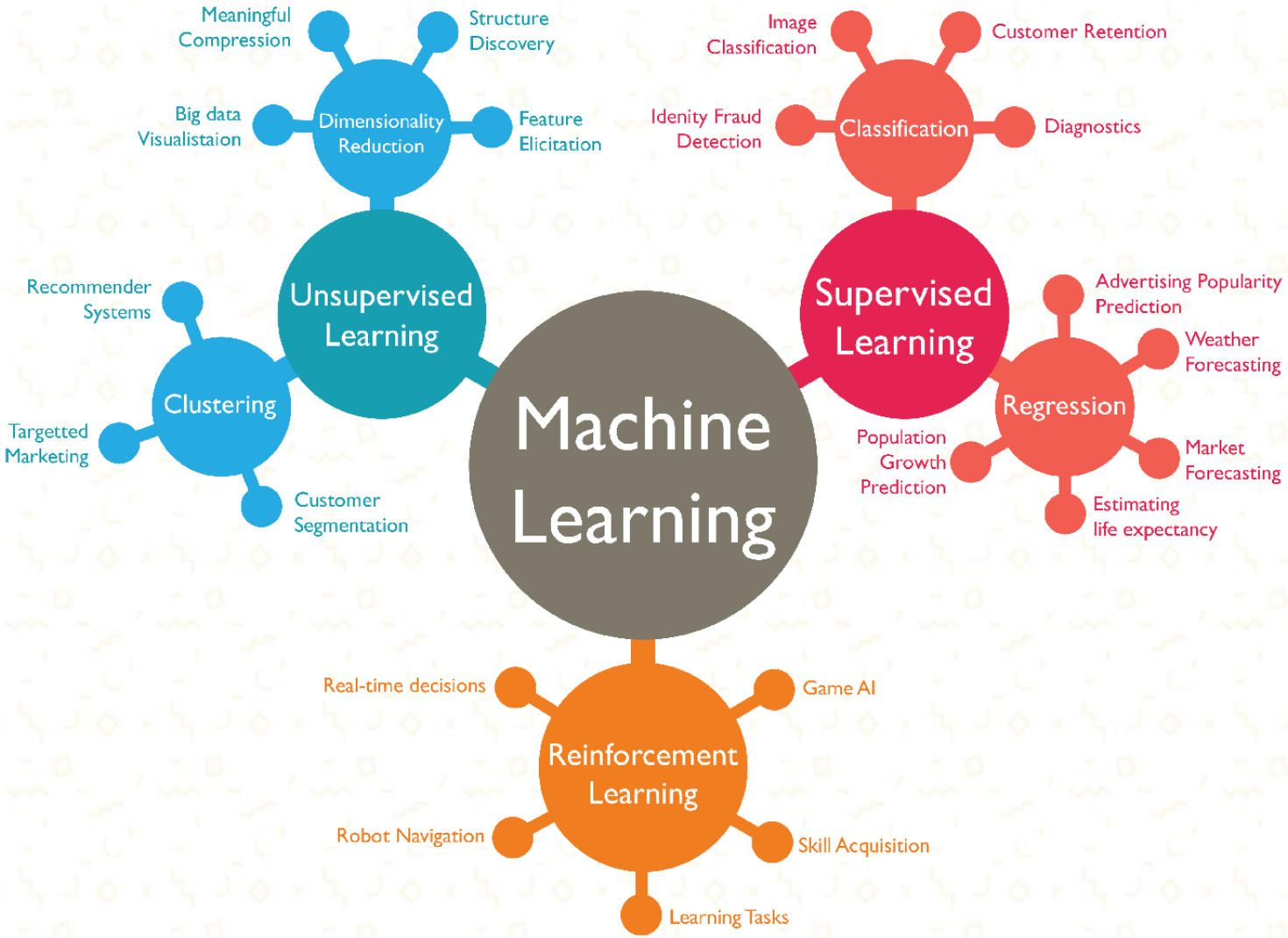
This is one of **the most iterative processes in the methodology as the data scientist will use multiple algorithms** to arrive at the best model for the chosen variables

Modelling

Descriptive Analytics: Past + Diagnostics

Predictive Analytics: To predict future data





Source: https://linkedin.com/pulse/business-intelligence-its-relationship-big-data-geekstyle

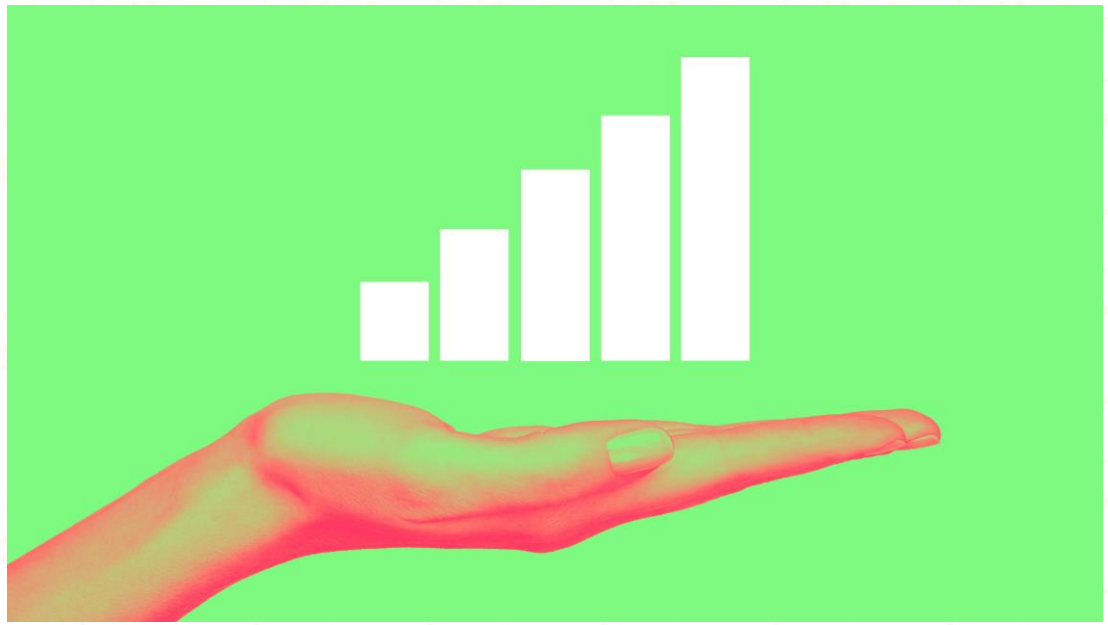
Model Evaluation

The quality of the model are being evaluated and ensured. The model need to meets all the requirements of the business problem or data scientist need to find another solution towards the data.

Deployment

Once the model have meet all the requirements.

Deploy the ML model or **present the findings** of the analytical process.



Feedback

Deploy ML Model

Maintaining and check the performance of the model.

Present Findings

Gather Feedback and discuss about the findings presented.

DigitalSkola

Uncover The World Of Digital Skills With Us

Thank YOU