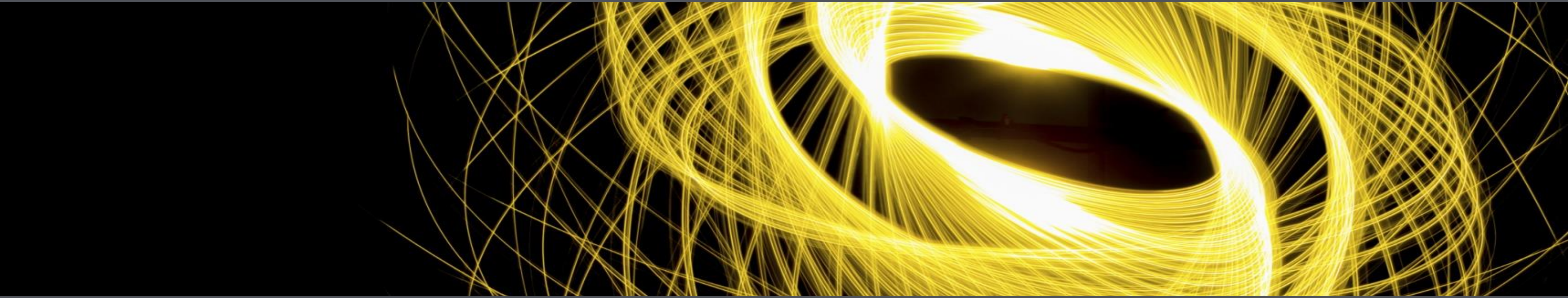
Department of Computer Science

University of Reading

CSMAD21 – Applied Data Science with Python



Data Science Introduction and Concepts

Lecture Objectives

- Understand and compare the different definitions of Data Science and the activities related to it.
- Identify and compare the different phases and activities in the Data Science Process.
- Differentiate the activities related with the Data Science Process such as Data Gathering, Data Pre – Processing, Modelling, Model Evaluation and Deployment.

Outline

- What is Data Science?
- The Data Science Process
  - Data Acquisition and Understanding
  - Data Preparation
    - Data Cleansing
    - Outlier detection
    - Data Transformation
  - Model Implementation
  - Model Evaluation
- Scikit-Learn
- Summary
- Q&A

What is Data Science?

- Data science is the process of using algorithms, methods, and systems to **extract knowledge and insights** from structured and unstructured data. It uses analytics and machine learning to help users make predictions, enhance optimization, and improve operations and decision making (IBM – Data Science, 2019).
- Data science combines multiple fields including statistics, scientific methods, and data analysis to **extract value from data**(Oracle – Data Science, 2020).
- Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to **discover hidden patterns** from the raw data (Hemant Sharma, 2020).

The Data Science Process - KDD

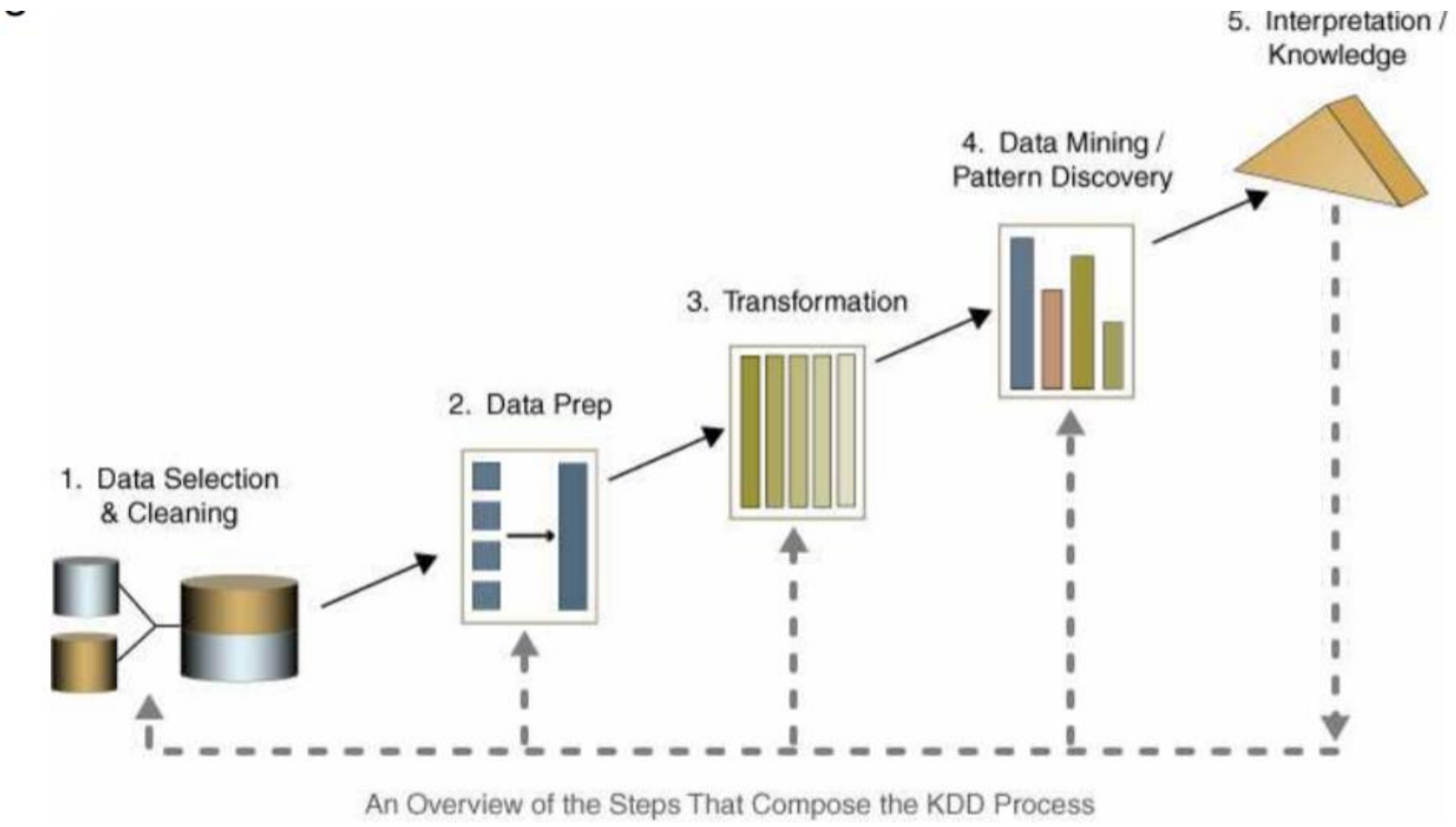

1. Data Selection & Cleaning
2. Data Prep
3. Transformation
4. Data Mining / Pattern Discovery
5. Interpretation / Knowledge
An Overview of the Steps That Compose the KDD Process

Figure 1: KDD Process in Data Mining (Rajput Abhishek , 2019)

The Data Science Process - IBM

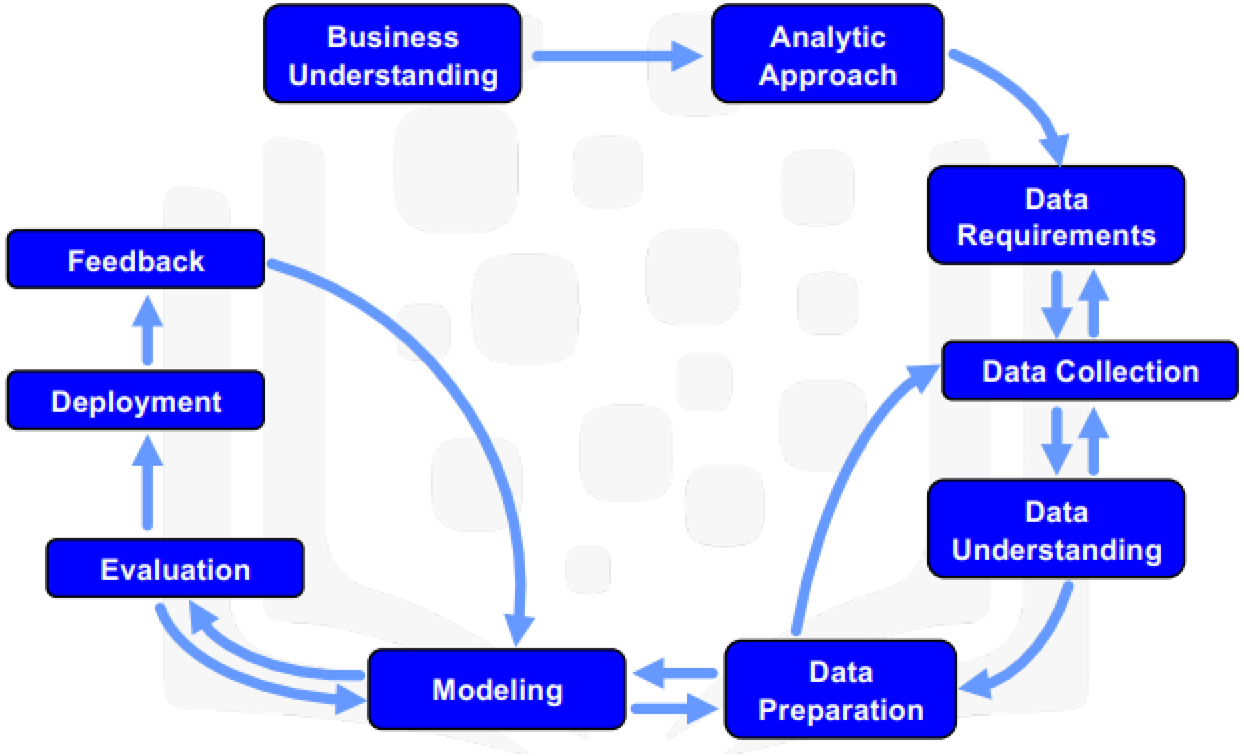
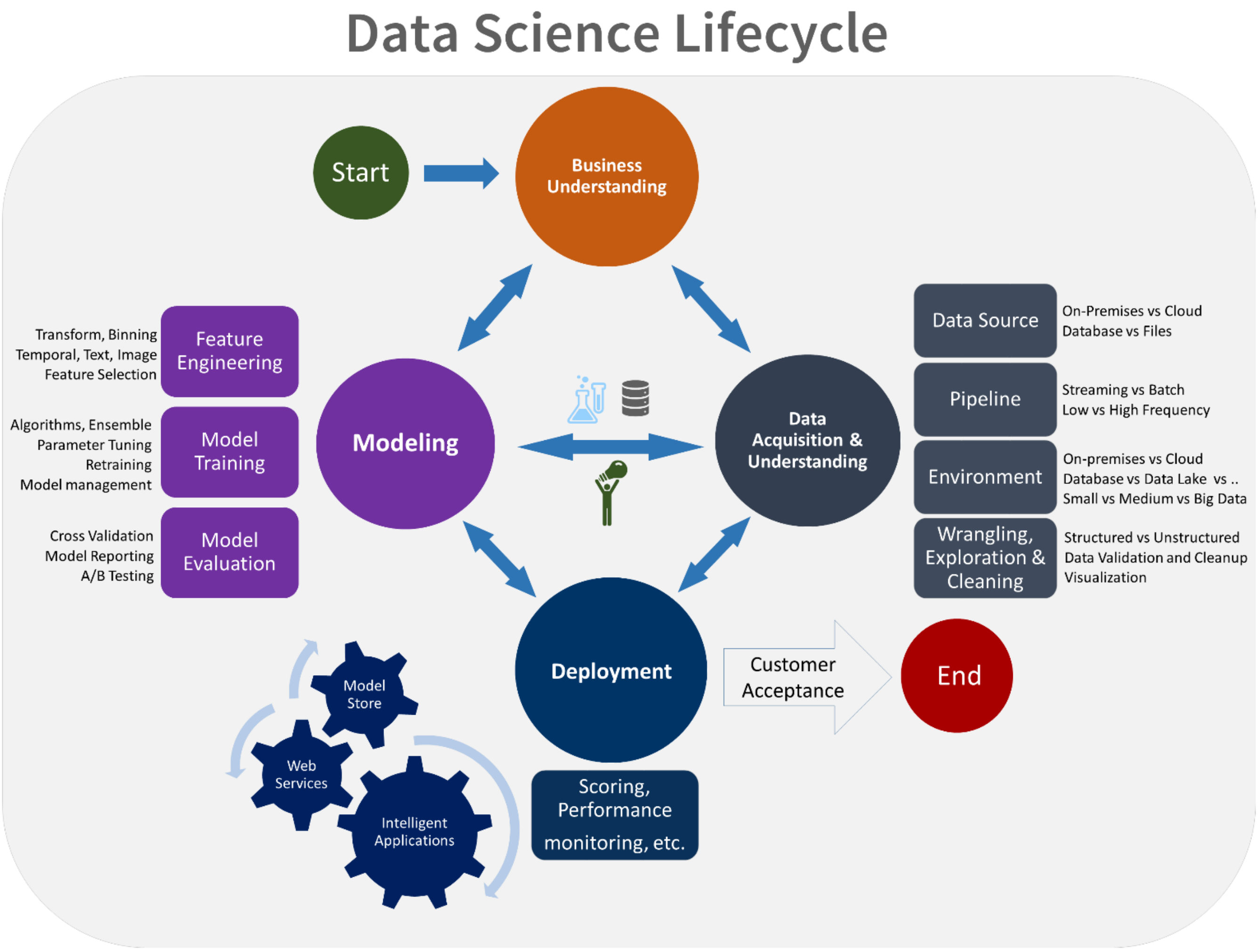


Figure 2: CRISP-DM Methodology diagram (Lin Polong, 2019)

The Data Science Process - Microsoft



Data Science Lifecycle

Start

Business Understanding

Data Acquisition & Understanding

Data Source

On-Premises vs Cloud
Database vs Files

Pipeline

Streaming vs Batch
Low vs High Frequency

Environment

On-premises vs Cloud
Database vs Data Lake vs ..
Small vs Medium vs Big Data

Wrangling, Exploration & Cleaning

Structured vs Unstructured
Data Validation and Cleanup
Visualization

Modeling

Feature Engineering

Transform, Binning
Temporal, Text, Image
Feature Selection

Model Training

Algorithms, Ensemble
Parameter Tuning
Retraining
Model management

Model Evaluation

Cross Validation
Model Reporting
A/B Testing

Deployment

Scoring, Performance monitoring, etc.

Model Store

Web Services

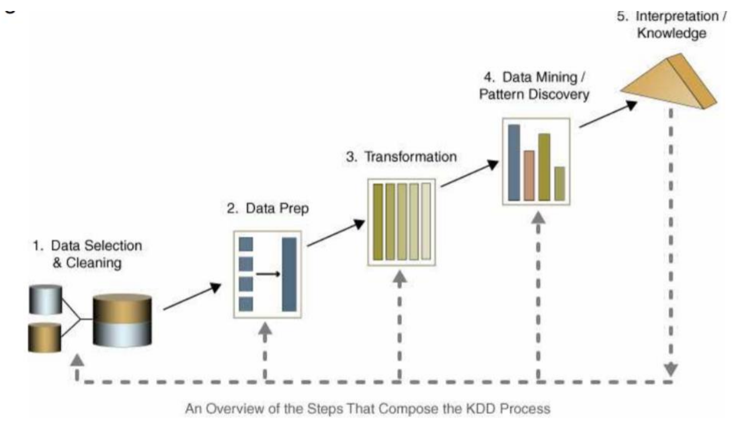
Intelligent Applications

Customer Acceptance

End

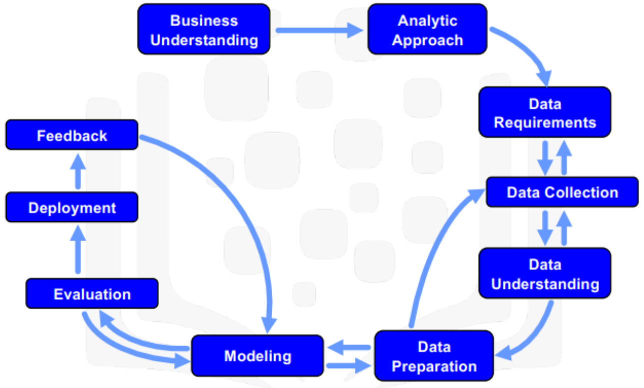
Figure 3: Data Science Lifecycle (Microsoft, 2020)

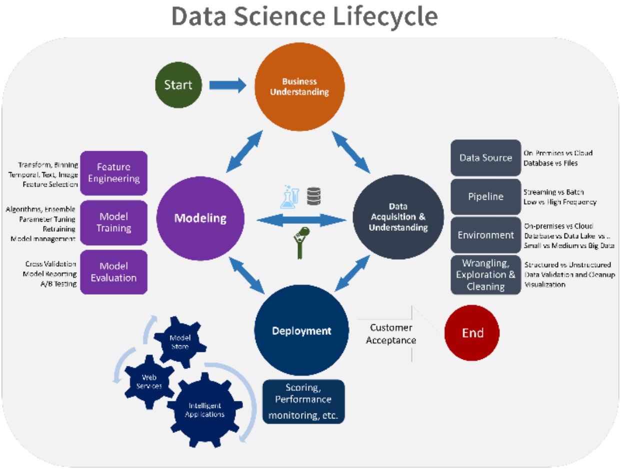
The Data Science Process

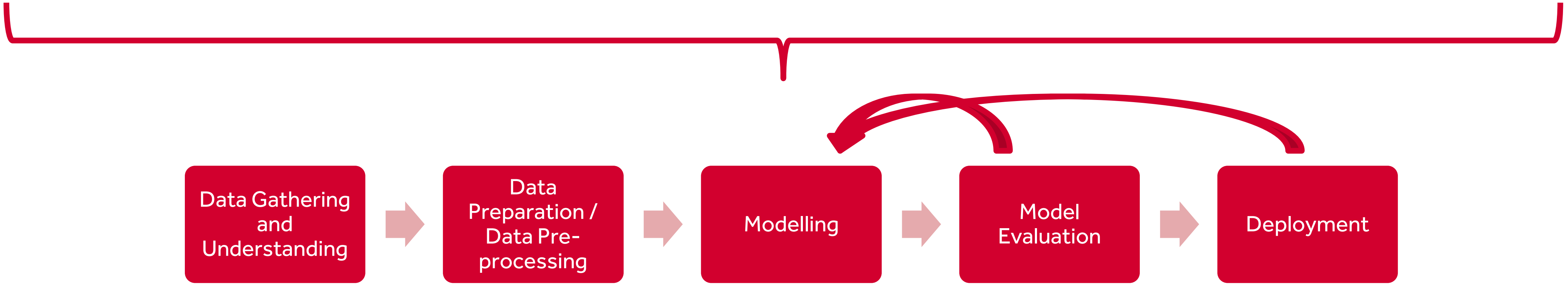


5. Interpretation / Knowledge
4. Data Mining / Pattern Discovery
3. Transformation
2. Data Prep
1. Data Selection & Cleaning

An Overview of the Steps That Compose the KDD Process





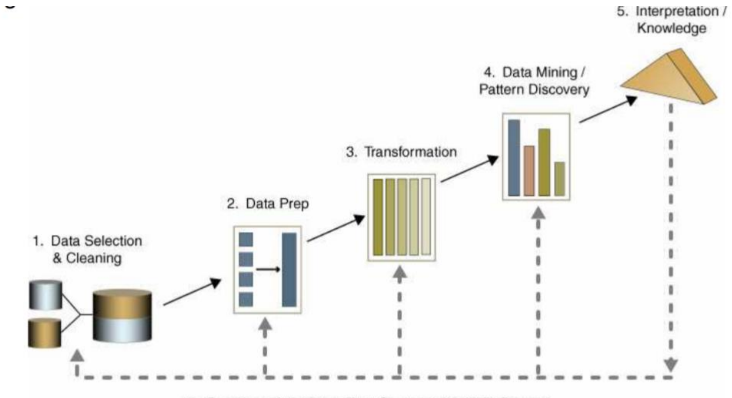


Data Gathering and Understanding → Data Preparation / Data Pre-processing → Modelling → Model Evaluation → Deployment

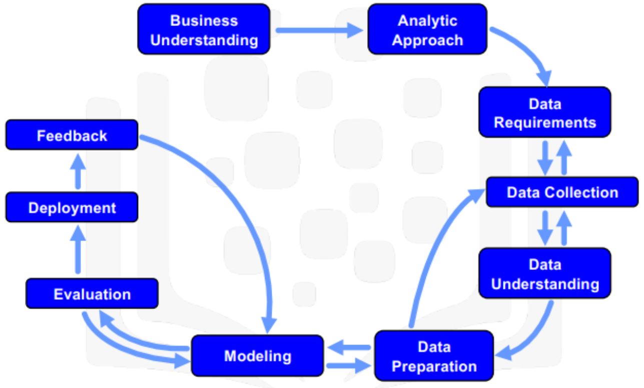
Data Gathering and Understanding

- Define the process to move/access the data source to the environment where the analysis is going to be preformed.
- What is the problematic?
- What is the project objective/goal?
- How to measure a success the project?
- How the data looks like (distributions, data types)?
- What about the data quality?
- Define an initial strategy of the solution (data pre-processing needed, machine learning algorithm/s to implement, test strategy, etc.)

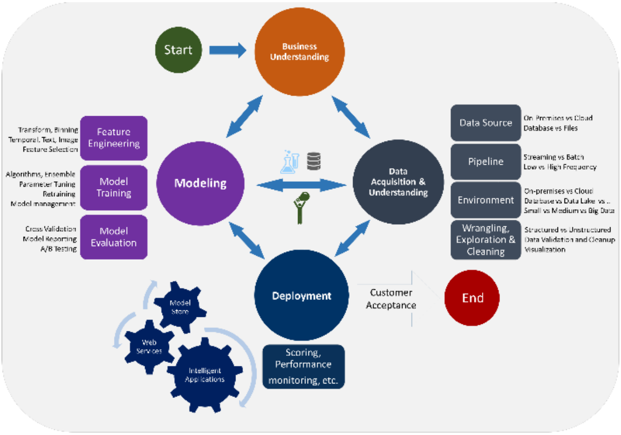
The Data Science Process

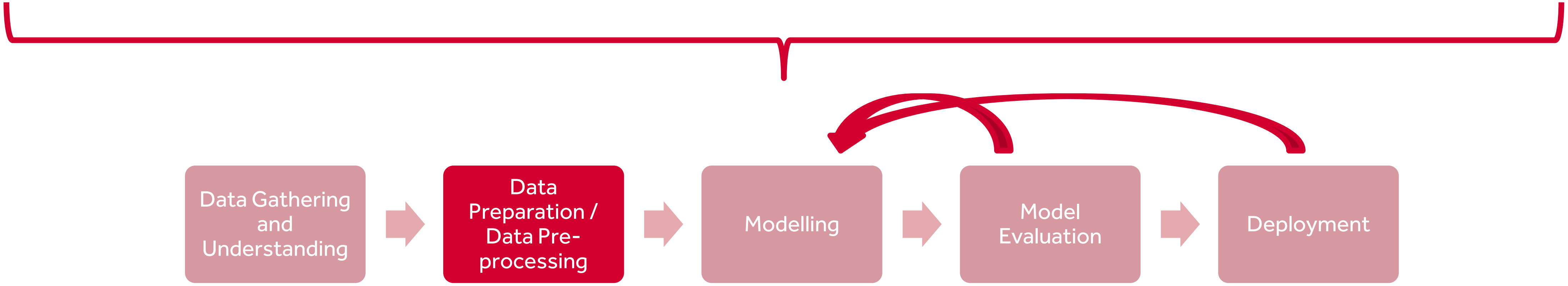

5. Interpretation / Knowledge
4. Data Mining / Pattern Discovery
3. Transformation
2. Data Prep
1. Data Selection & Cleaning

An Overview of the Steps That Compose the KDD Process



Data Science Lifecycle





Data Gathering and Understanding

Data Preparation / Data Pre-processing

Modelling

Model Evaluation

Deployment

Data Preparation/ Data Pre-processing

- Define and Select the amount of data for the analysis (scope).
- Data Quality.
- Data Reduction and Transformation

Data Preparation/ Data Pre-processing

- Data Quality. Is the planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it **is fit for consumption** and meet the needs of data consumers. The data must be **consistent and unambiguous** (DAMA, 2020). Some of the data anomalies are (Jiawei Han, 2011):
  - **Noise**. Refers to modification of original values. E.g. Distortion of a person’s voice during a phone call.
  - **Outliers**. Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set.
  - **Duplicated Values**. Records that appears more than one in the dataset (in full or partially). e.g. person with multiple emails.
  - **Missing Values**. Data was not collected, or the attributes are no applicable to all cases.

Data Quality - Missing Values

- Some of the most common strategies to deal with null values are (Jiawei Han, 2011):
  - Eliminate Data Objects
  - Ignore the Missing Values During the Analysis
  - Estimate Missing Values:
    - Default Value
    - Mean or Median
    - Most Common Value

Data Reduction and Transformation

Correlation.

Transformation.

- Numeric Data
  - Normalisation
    - Min and max
    - Z normalisation
- Categorical Data
  - Ordinal Encoding
  - One – hot Encoding
  - Dummy Variable Encoding

Data Reduction and Transformation

- Correlation.
  - Correlation tells us whether there is a relationship between two variables. One measure of correlation, known as **Pearson's correlation coefficient**, is calculated for a sample of values x1,...,xn and y1,...,yn as:

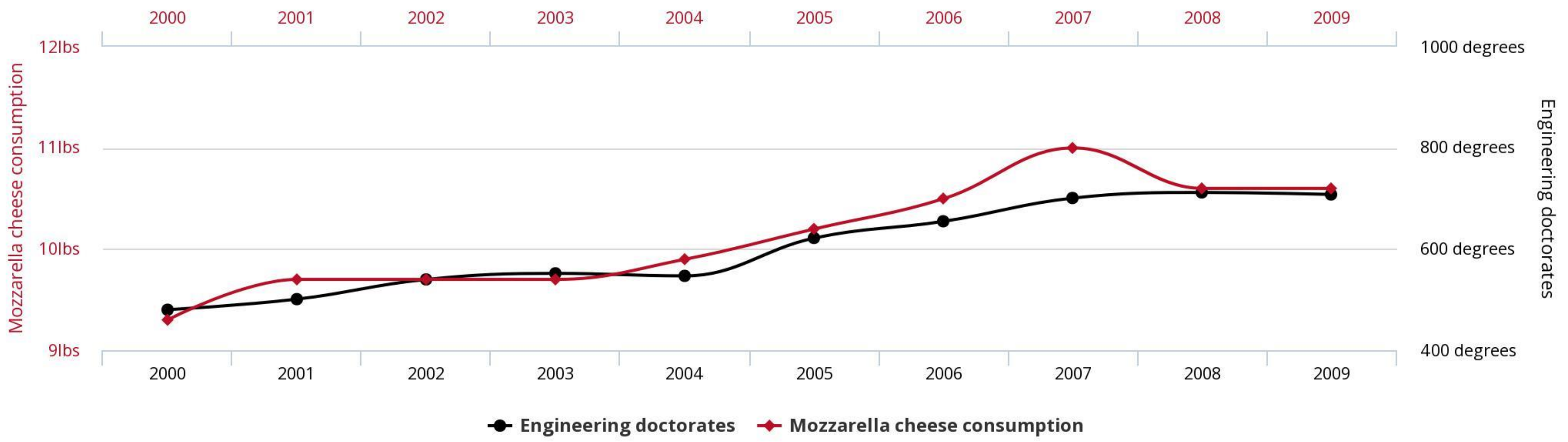
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

  - where sx and sy are the sample standard deviations for x and y respectively. It takes a value between −1 and 1, where **positive values mean a positive correlation**, **and negative values correspond to anti-correlation**. In the case where the correlation is equal to 1, this means there is a linear equation linking the values of x and y.

Data Reduction and Transformation

- **Correlation vs Causation**
  - Although causation implies correlation, just because two variables are correlated, it does not mean that there is a causal relationship between them.
  - It is quite easy to find correlations between variables that are clearly not related, especially if you make many tests.

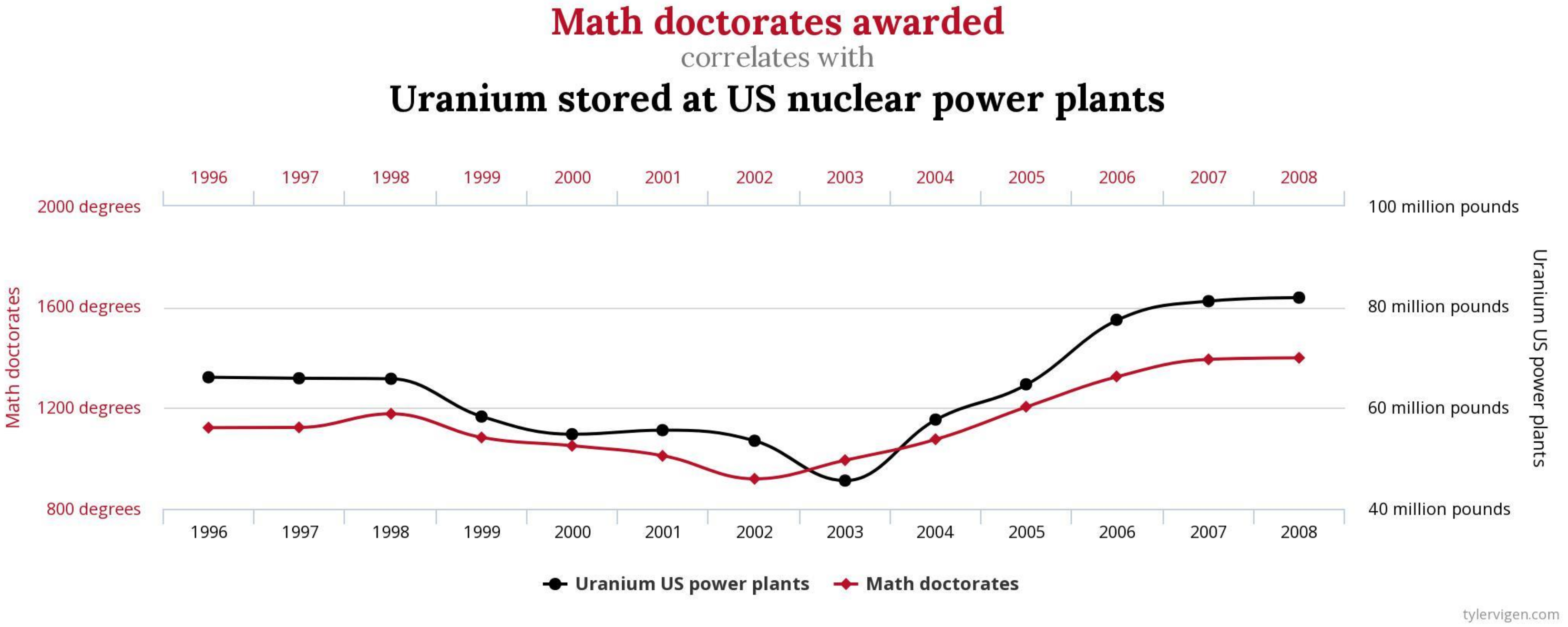
Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded



tylervigen.com

Data Reduction and Transformation

- **Correlation vs Causation**
  - Correlation can occur purely by chance, or it could be the case that there is some other underlying variable that has a causal relationship with each of the measured variables.



More examples here: http://tylervigen.com/spurious-correlations

Data Reduction and Transformation

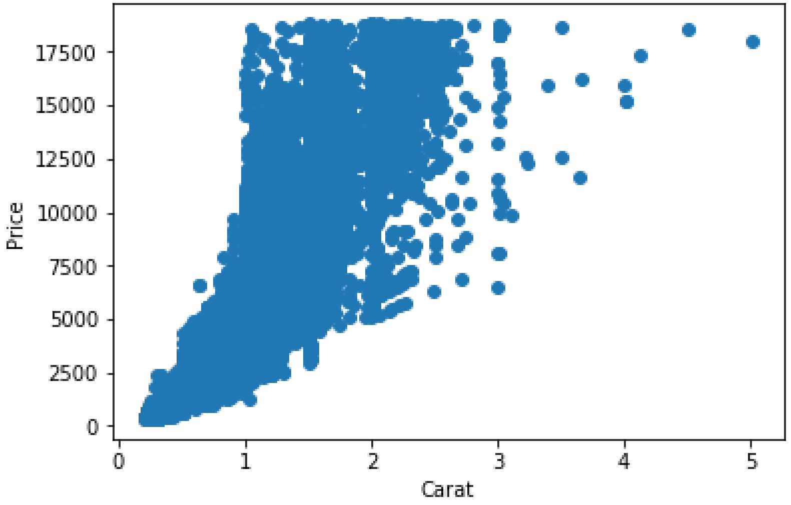
- Transformation – Numerical Data.
  - Normalisation. The objective of the normalisation/standardisation is to **transform the scale of the variables**. Different scales in the data can bias the algorithms (Jiawei Han, 2011). This process is also beneficial to process faster the data specially in Support Vector Machines (SVM), Neural Networks and Regression algorithms.
    - Min and Max
    - Z Normalisation

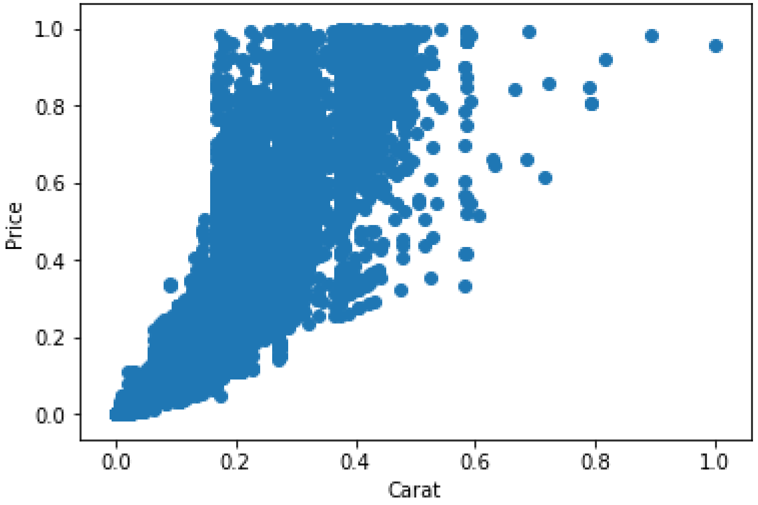
Data Reduction and Transformation

- Normalisation.
  - **Min – Max**. One of the most common ways to normalise data. The min value in the data is given 0 and the maximum gets 1. The remaining data is transformed between this range.

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

  - Downside. Does not handle outliers very well.



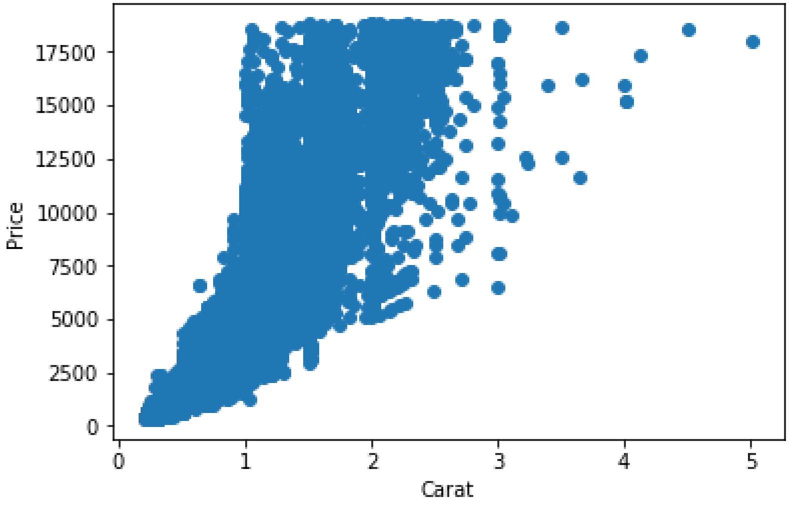


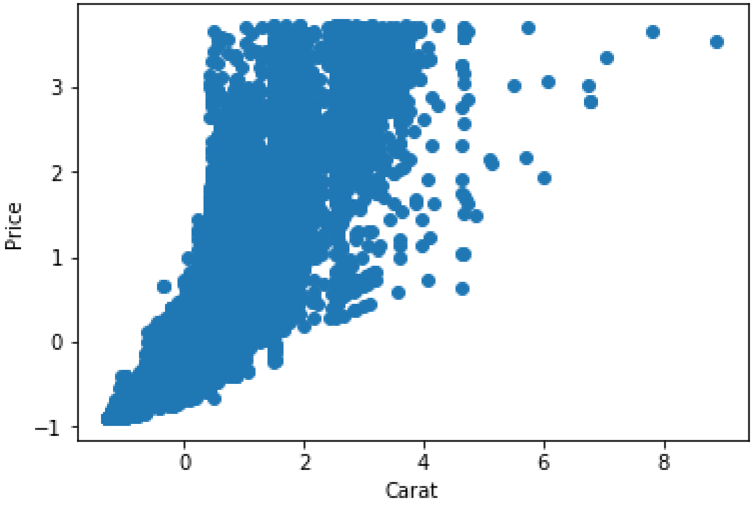
Data Reduction and Transformation

- Normalisation.
  - **Z-Score**. It is defined by the following formula:

$$z = \frac{x - \mu}{\sigma}$$

  - Data in the mean value will be transformed to 0. Values above the mean will be positive and those below the mean will be transformed to negative values.
  - Downside. Features might not be on exact same scale.





Data Reduction and Transformation

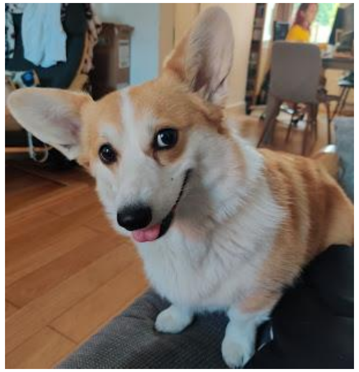
- Transformation – Categorical Data.
  - There are cases where the machine learning algorithm can work directly with categorical data (like decision trees) and there might be cases were this data needs to be transformed (VanderPlas Jake, 2017).
    - Ordinal Encoding.
    - One-Hot Encoding.
    - Dummy Variable Encoding.

Data Reduction and Transformation

- Transformation – Categorical Data (VanderPlas Jake, 2017).
  - **Ordinal Encoding**. Assign an integer value to a particular label. It works better when the values have a natural ordered relationship (E.g. first, second and third).
  - **One-Hot Encoding**. For variables where no ordinal relationship exist(E.g. dog, cat, lizard) , is better to use this method to increase the performance. It transform the labels into binary array where the value 1 is representing the label.
  - **Dummy Variable Encoding**. One-hot encoding includes redundancy. Dummy variable encoding represent C categories with C-1 binary variables.

Data Reduction and Transformation

- Transformation – Categorical Data (Example).
- We have the attribute **Species** that has the values **Dog**, **Cat** and **Lizard.**
- The data will be transformed…

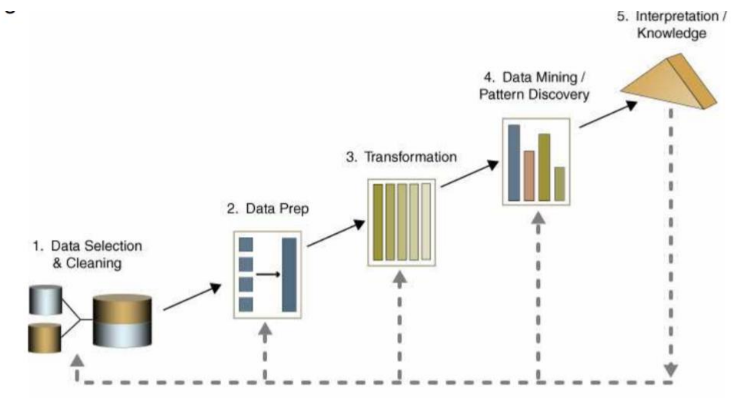


- Ordinal Encoding: 1
- One-Hot Encoding: [1,0,0]
- Dummy Variable Encoding: [1,0]

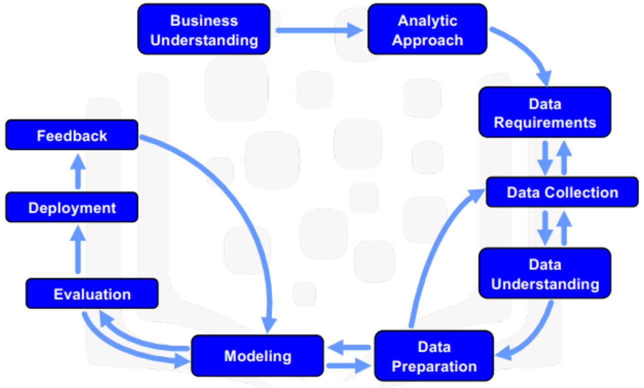
- Ordinal Encoding: 2
- One-Hot Encoding: [0,1,0]
- Dummy Variable Encoding: [0,1]

- Ordinal Encoding: 3
- One-Hot Encoding: [0,0,1]
- Dummy Variable Encoding: [0,0]

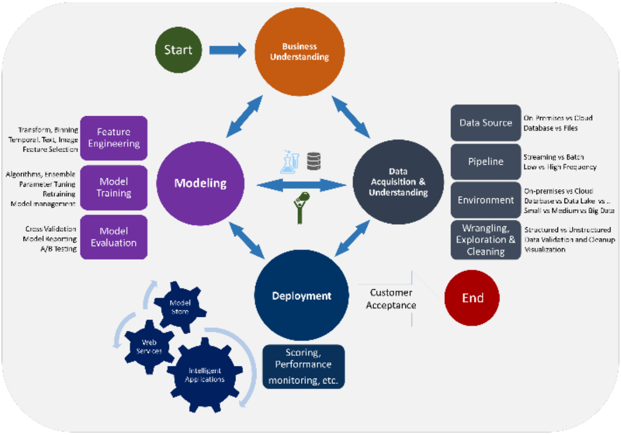
The Data Science Process

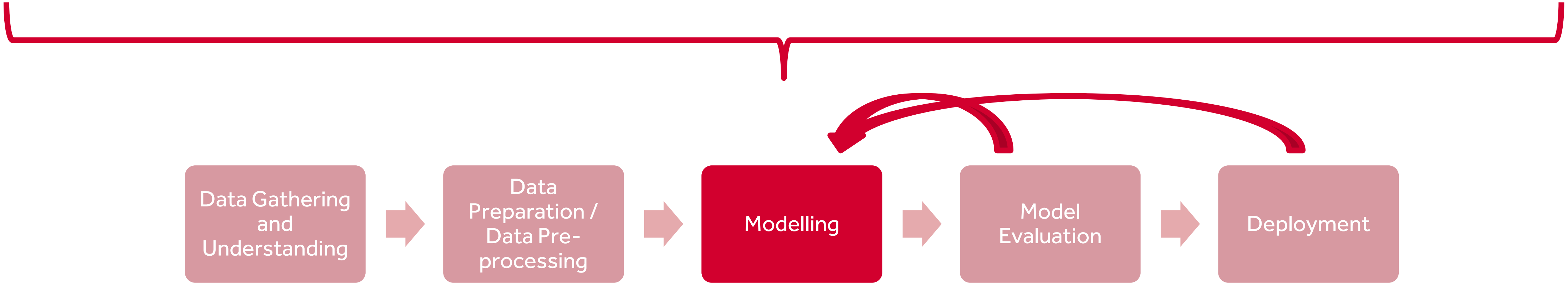

5. Interpretation / Knowledge
4. Data Mining / Pattern Discovery
3. Transformation
2. Data Prep
1. Data Selection & Cleaning

An Overview of the Steps That Compose the KDD Process



Data Science Lifecycle

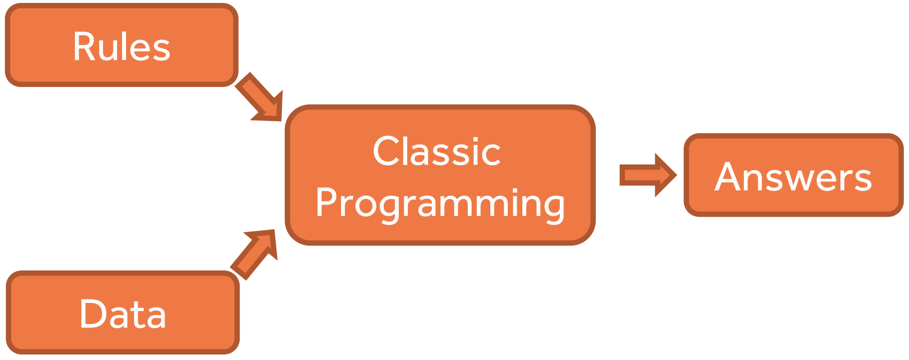




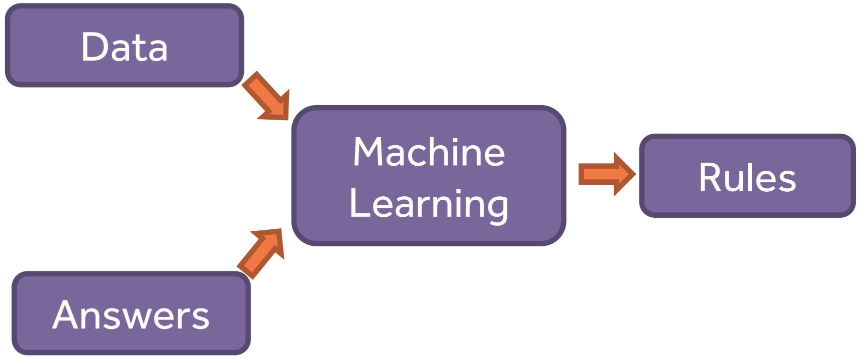
Data Gathering and Understanding → Data Preparation / Data Pre-processing → Modelling → Model Evaluation → Deployment

Modelling

- What is machine learning?
  - Machine Learning is the science (and art) of programming computers so they can learn from data (Geron Aurelien, 2019).
  - Is the field of study that gives the computers the ability to learn without being explicitly programmed. (Arthur Samuel, 1959).



Classical Programming



Machine Learning

Modelling

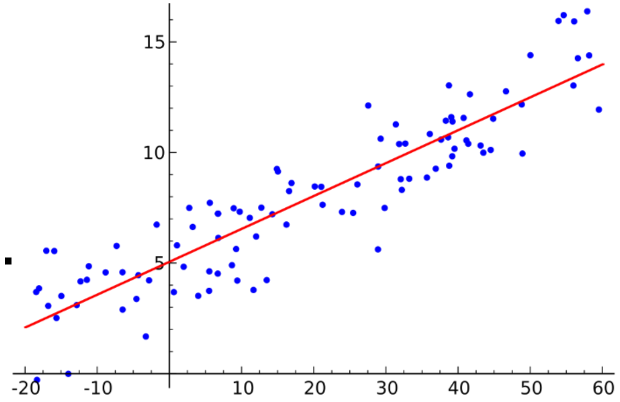
- Categories of Machine learning:
  - **Supervised Learning** (VanderPlas Jake, 2017).
    - Involves modelling the relationship between measured features of data and some **label associated with the data**; once this model is determined, it can be used to apply labels to new, unknown data.
    - The main tasks that can be achieved via supervised learning are **Regression and Classification.**
    - In Classification, the labels are discrete categories, while in regression the labels are continuous quantities.

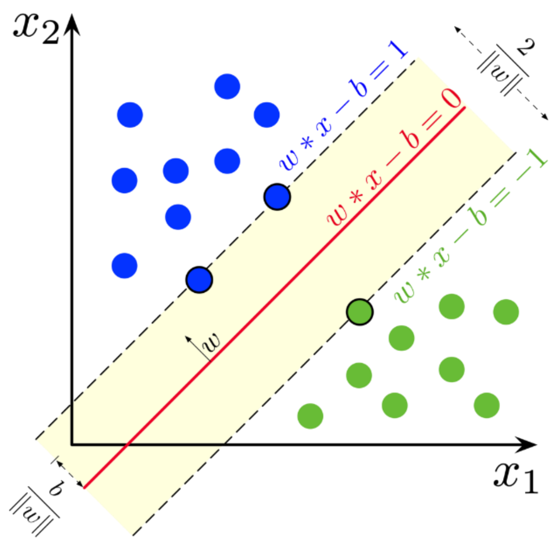
Modelling

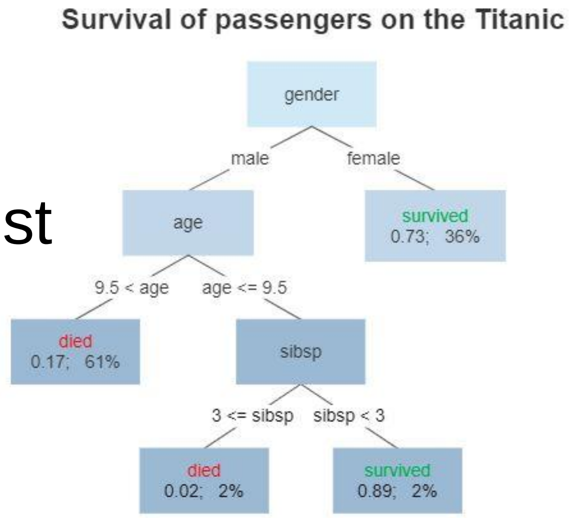
- Categories of Machine learning:
  - **Unsupervised Learning** (VanderPlas Jake, 2017).
    - Involves modelling the features of a dataset without reference to any label and is often described as **"letting the data speak for itself".** These models include tasks such as **Clustering** and **Dimensionality Reduction**.
    - Clustering Algorithms identify distinct groups of data, while Dimensionality Reduction algorithms search for more succinct representations of the data.

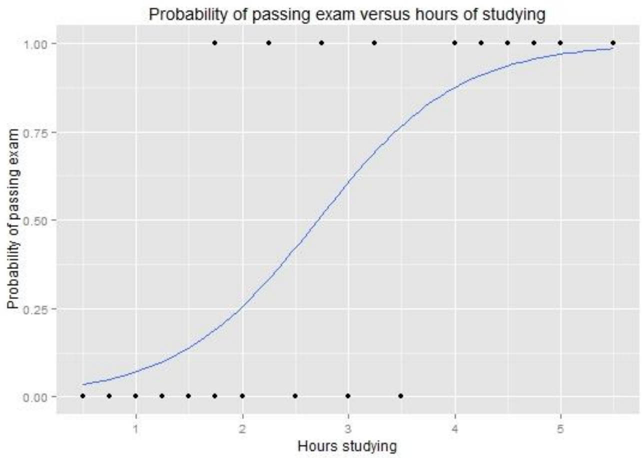
Modelling

- Categories of Machine learning:
  - **Supervised Learning Algorithms.**
    - **Logistic Regression**
    - **Linear Regression**
    - **Ridge Regression**
    - **Lasso Regression**
    - **Support Vector Machines**
    - Decision Trees and Random Forest
    - Neural Networks



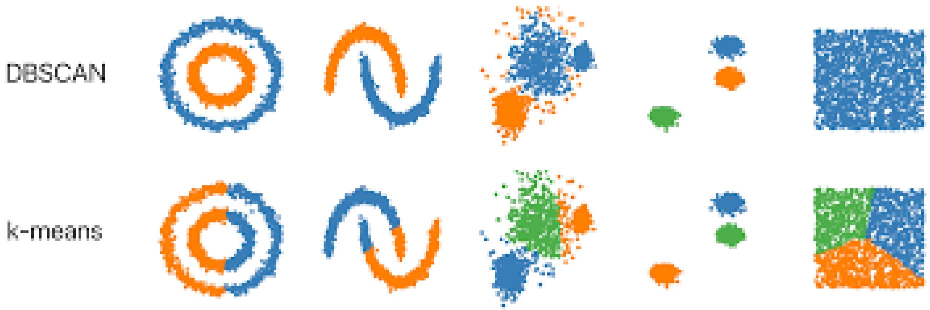


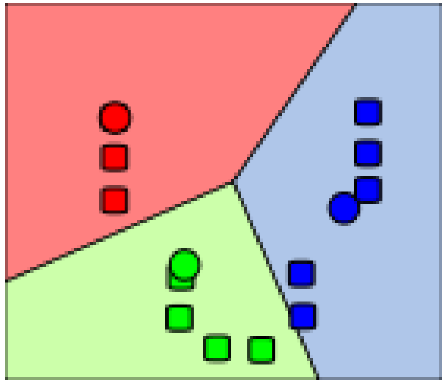


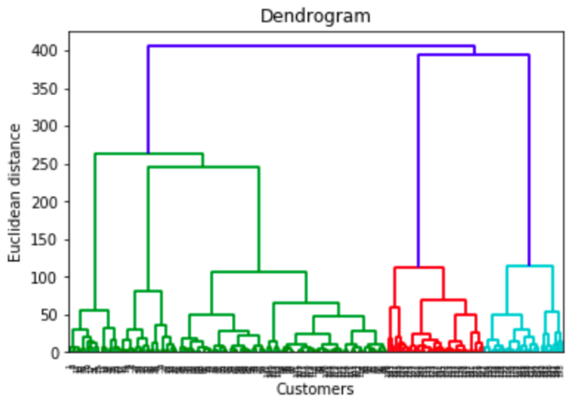


Modelling

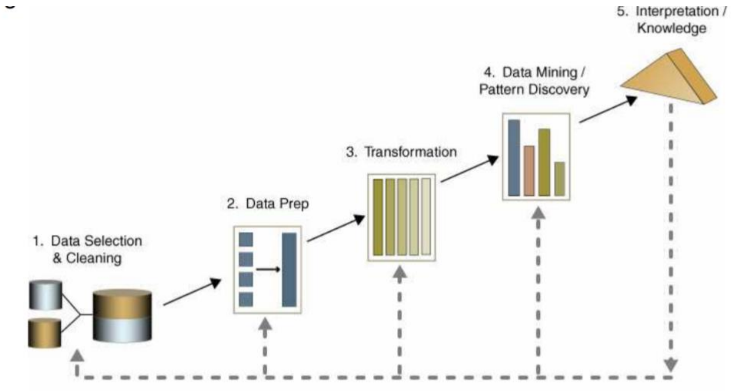
- Categories of Machine learning:
  - **Unsupervised Learning Algorithms**.
    - **K-Means**
    - DBSCAN
    - Hierarchical Cluster Analysis (HCA)
    - Principal Component Analysis (PCA)
    - **Mixture Models**
    - Associated Rule Learning - Apriori



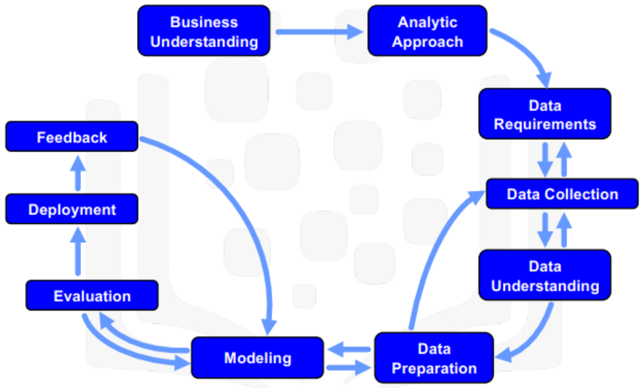


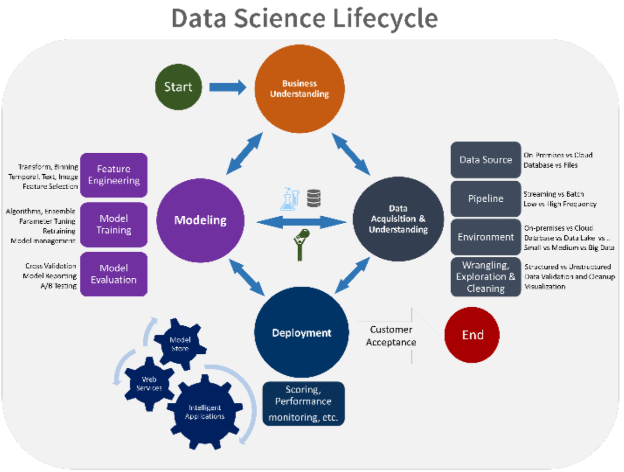


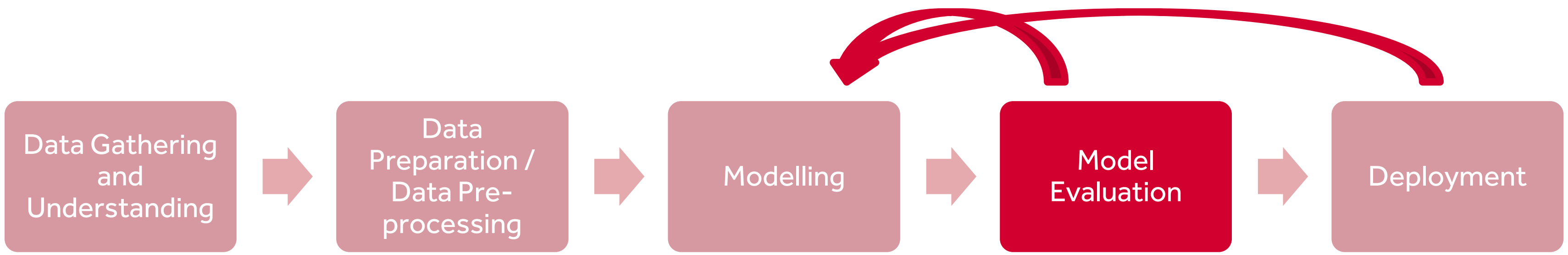
The Data Science Process


5. Interpretation / Knowledge
4. Data Mining / Pattern Discovery
3. Transformation
2. Data Prep
1. Data Selection & Cleaning

An Overview of the Steps That Compose the KDD Process







Data Gathering and Understanding

Data Preparation / Data Pre-processing

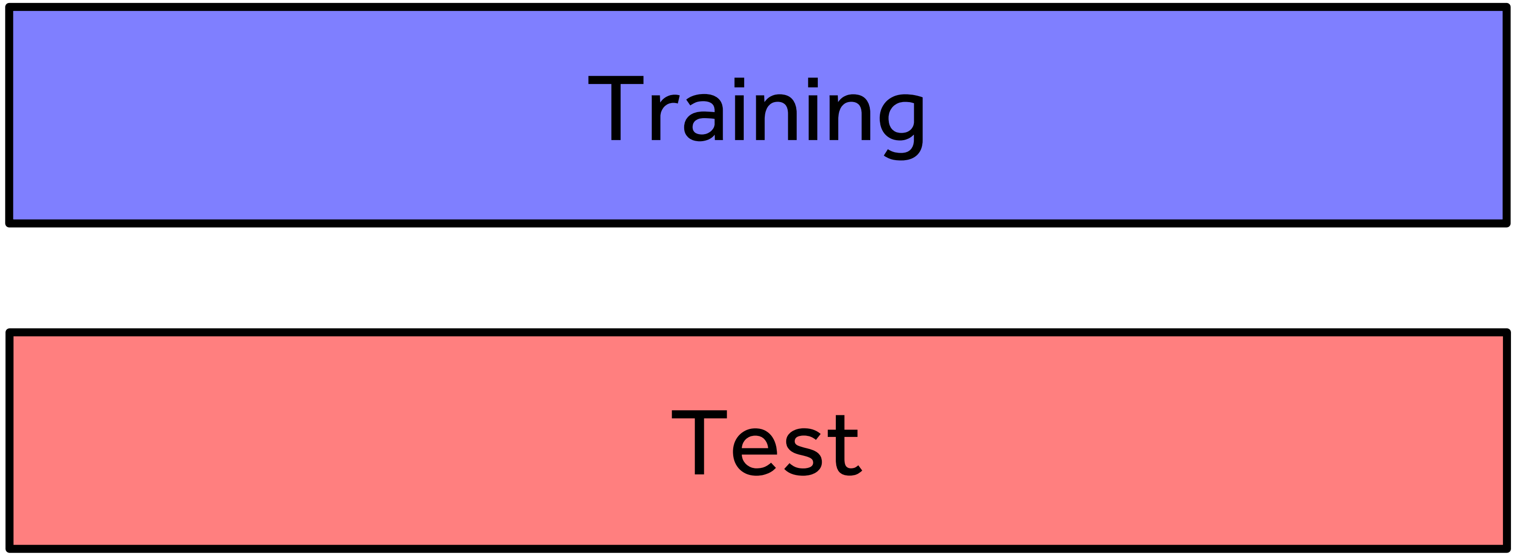
Modelling

Model Evaluation

Deployment

Module Evaluation – Testing and Training

- **Resubstituting.**
  - The model is trained and test with the same dataset.
  - It is usually avoided given that overfits the models (the model memorises the data).

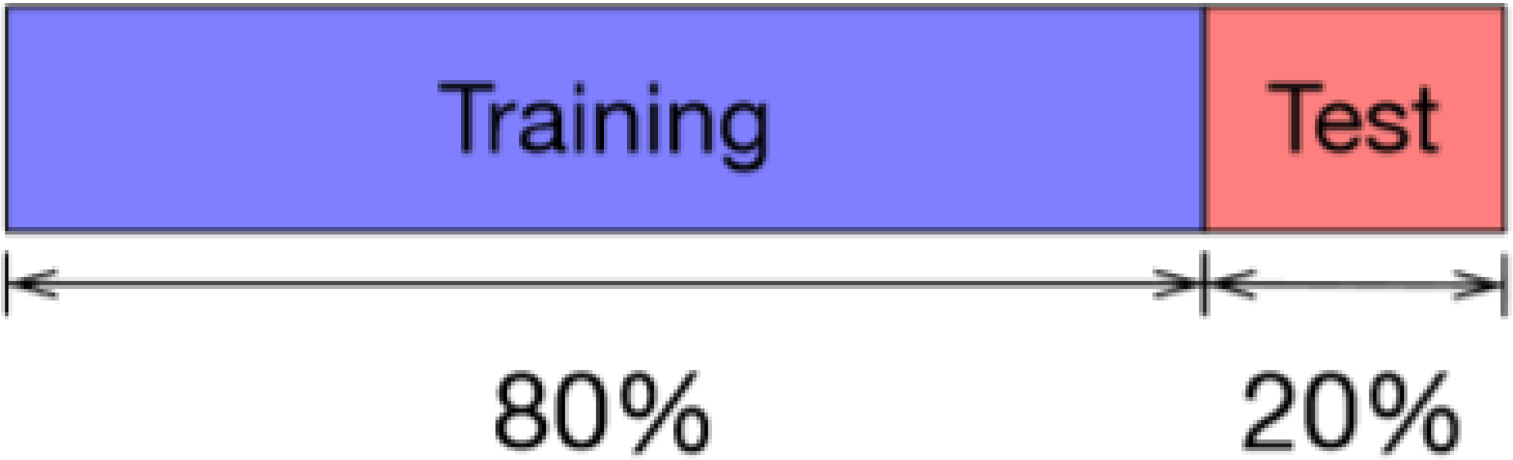


Training

Test

Module Evaluation – Testing and Training

- **Hold – Out.**
  - We can evaluate how well our model can predict the output **y**, by holding aside a subset of the data called test data. The remaining data is called training data and is used to train the model.
  - For example, we could split our credit data into a training set of 80% of the observations (rows in the data frame) and keep a test set of 20% of the data, that is not used in training the model.

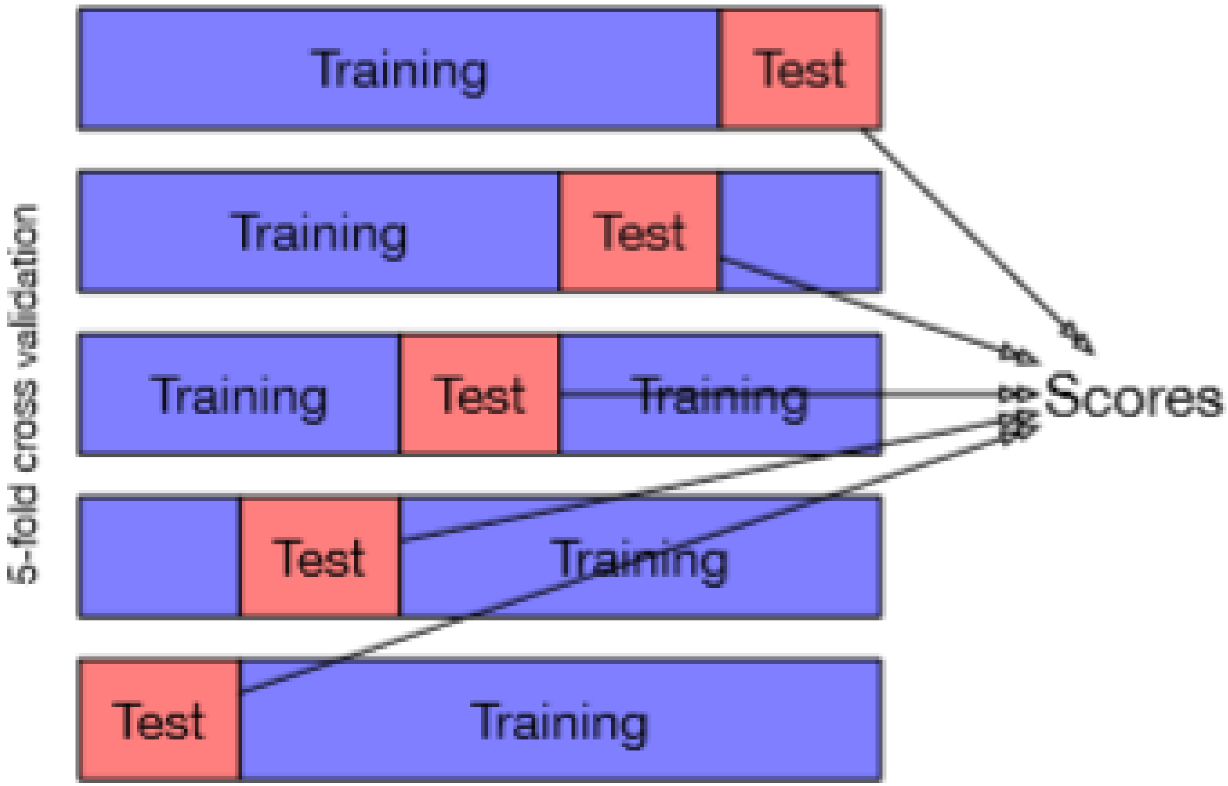


Module Evaluation – Testing and Training

- **Cross Validation.**
  - What if we just happened to choose a ‘good’ test subset of the data for the model?
  - If we repeat the training/test procedure over many different splits of the data, we might get quite different answers.
  - Cross validation aims to address this by shuffling and then splitting the data many times and combining scores or performance over all of the splits.
  - A common approach is k-fold cross validation where the data are partitioned in k different ways.
  - This can be used both for validation and for learning extra parameters of models.
  - Sometimes the training process itself can be random!

Module Evaluation – Testing and Training

- **Cross Validation.**
  - Rather than taking one partition of the data, we create several and train and score the model on all of them, collecting the scores. The score is then usually averaged to give an overall measure of performance.



Module Evaluation – Module Performance

- The machine learning algorithms can be evaluated depending on the nature of the activity they are performing.
  - **Clustering**. The most used metrics are:
    - Sum of Squared Errors (within the cluster and between clusters).
    - Akaike Information Criterion (AIC).
    - Bayesian information criterion (BIC).

  - The detail of this metrics is going to be covered in the “Data Science Applications –Clustering” Lecture.

Module Evaluation – Module Performance

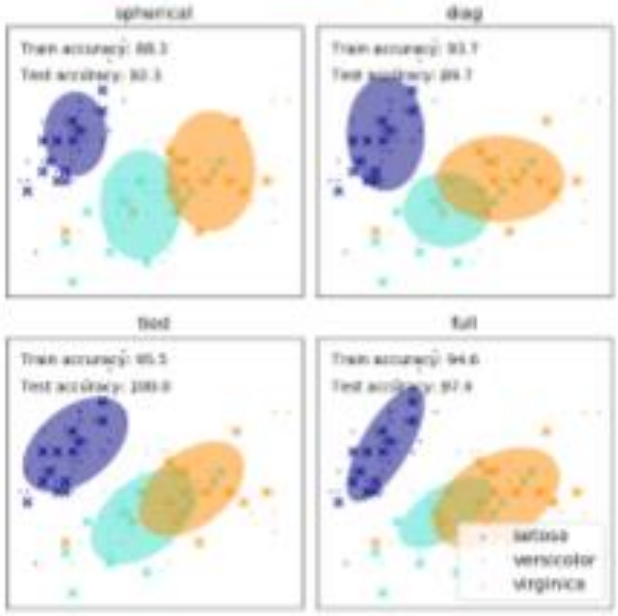
- The machine learning algorithms can be evaluated depending on the nature of the activity they are performing.
  - **Regression**. The most used metrics are:
    - R Squared
    - Mean Squared Error
    - Root mean squared Error
  - The detail of this metrics is going to be covered in the “Data Science Applications – Regression” Lecture.

Module Evaluation – Module Performance

- The machine learning algorithms can be evaluated depending on the nature of the activity they are performing.
  - **Classification**. The most used metrics are:
    - Accuracy.
    - Precision
    - Recall
    - F1 Score
  - The detail of this metrics is going to be covered in the “Data Science Applications – Classification” Lecture.

Scikit - Learn

- scikit-learn is a Python library for machine learning in Python, that is included in the Anaconda Python distribution.
- The scikit-learn workflow:
  - Create an object corresponding to the model you wish to use.
  - Fit the model to some data.
  - Make predictions or test the model.
- Many examples and API documentation can be found on the website
  - https://scikit-learn.org/stable/



Summary

- The Data Science process involves activities such as Data Gathering, Data Pre-processing, Data Transformation, Modelling and Model Evaluation.
- Must of the effort of the analysis is located in the data pre-processing stage.
- The machine learning algorithms can be divided in supervised and unsupervised learning.
- Depending on the algorithm implemented, there are different indexes/ metrics to evaluate them.
- The Scikit-Learn provides a robust framework to implement and evaluate machine learning algorithms.

Questions

References

- Rajput Abhishek (2019). KDD Process in Data Mining. Geeks for Geeks. Online at: https://www.geeksforgeeks.org/kdd-process-in-data-mining/. Accessed 23/10/2020.
- Lin Polong (2019). The Data Science Process. IBM – The Big Data University. Online at: https://www-01.ibm.com/events/wwe/grp/grp304.nsf/vLookupPDFs/Polong%20Lin%20Presentation/$file/Polong%20Lin%20Presentation.pdf. Accessed 23/10/2020.
- Azure Documentation (2020). Modeling stage of the Team Data Science Process lifecycle. Microsoft. Online at: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle-modeling. Accessed 23/10/2020.
- DAMA International (2020). Data Quality. Data Management Institute. Online at: https://www.dama.org/content/body-knowledge. Accessed 23/10/2020.
- Jiawei Han, et al. (2011). Data Mining: Concepts and Techniques. Elsevier Science & Technology.
- VandePlas Jack (2017). Python Data Science Handbook. O'REILLY.
- Geron Aurelien (2019). Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow