## QUESTION 3: DATA DOPPELGÄNGERS

## INTRODUCTION

With rapid increase in amount and complexity of real-world data, machine learning tools are increasingly applied to integrate heterogenous data and perform predictive modelling to generate business insights (Figure 1). The usual flow of machine learning is identification of specific problem, selection of appropriate machine learning model and evaluation of model performance where a training set is randomly partitioned and fit to model and a test set is used to validate performance quantitatively, with the assumption that datasets are independent. Biological data, however, have similar genomic sequence, come from a superfamily, exhibit similar activity or structural folding that render machine learning inaccurate.

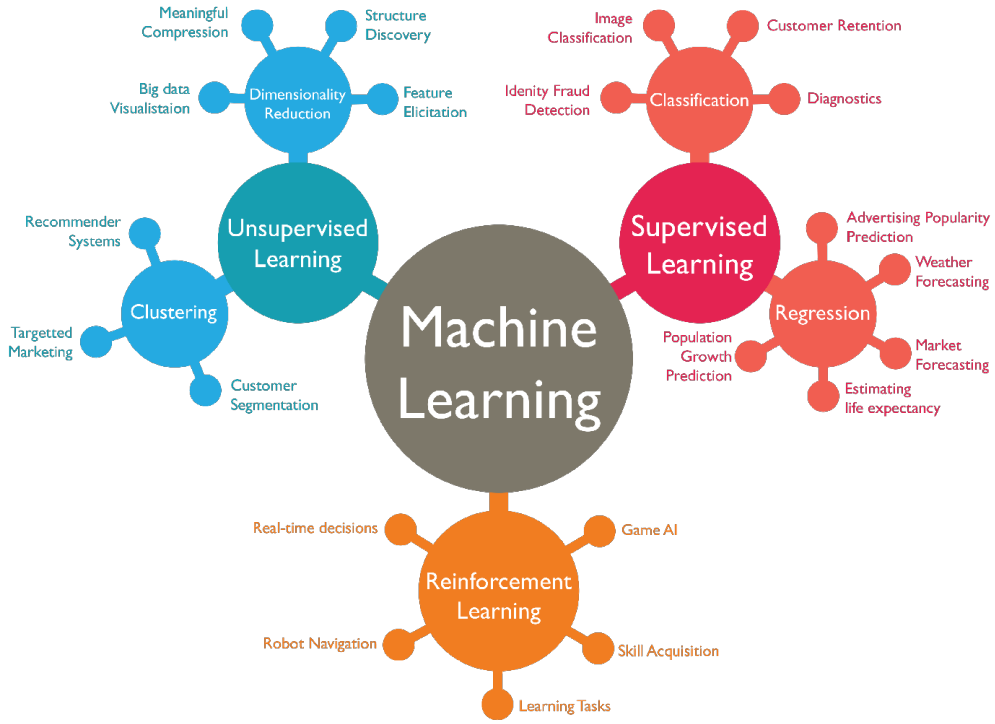


Figure 1. Applications of machine learning methods.
Image from https://wordstream-files-prod.s3.amazonaws.com/s3fs-public/machine-learning.png

### Confounding Factors

Confounders are underlying unmeasured variables or artefacts that influences dependent and independent factors to impact outcome. Error in modelling arises when confounders are excluded in modelling. Confounders may be introduced during data collection from technical artefacts. An example the type of radiology scanner in healthcare as a confounding factor in deep learning model of hip fractures (Badgeley, 2019).

A common pitfall in biomedical research with strong confounding effect when samples are processed according to outcome or treatment or disease groups, as such genomic data analysis and integration show batch effects, easily misinterpreted as outcome difference. It is therefore important to randomise sample during processing with case and control within a batch. There are also existing methods to adjust batch effects such as ComBat-Seq (Zhang, 2020) with better statistical power and false positive control.

## Data Leakage

Data leakage occurs when the training set is derived based on data from test set, or vice versa, hence the two dataset have a degree of dependence or similarity which does not provide proper validation and evaluation of model performance. Within a dataset, data leakage may establish false association between data and impairs unsupervised machine learning methods, such as clustering, k-nearest neighbour. It is important to ensure that parameters are learned only from training set and test set is independently derived.

A commonly used machine learning package in R, caret, also known as classification and regression training, is a set of functions used to create predictive model and its createPartition( ) and train( ) function learns parameters from training set (Kuhn, 2008). Such packages ensures no data leakage between training and test sets.

## Conclusion

Machine learning has great potential to generate useful biological insights, despite challenges to overcome complex systems and data doppelgängers. It is the responsibility of modellers to construct and navigate around biases and ensure true biological significance are displayed to users and readers.

COMPANY NAME

# BUDGET

## Ut vehicula nunc mattis pede

Curabitur labore. Ac augue donec, sed a dolor luctus, congue arcu id diam praesent, pretium ac, ullamcorper non hac in quisque hac. Magna amet libero maecenas justo.

| Description | Quantity | Unit Price | Cost |
|---|---|---|---|
| Item 1 | 55 | $100 | $5,500 |
| Item 2 | 13 | $90 | $1,170 |
| Item 3 | 25 | $50 | $1,250 |
| | | | |
| **Total** | | | **$7,920** |