Question 3

Data doppelgangers refers to independently derived data sets that are analogous. Machine learning models rely heavily on data sets to learn and make predictions, and for cross-validation. When there is great similarity between the data set that is used for training and the one that is used for validation, the model will be considered as well performing models even when there is biasness. (1) This phenomenon is known as doppelganger effect or functional doppelgangers. However, it is to note that the presence of data doppelgangers may not necessarily result in a doppelganger effect. (1)

It is important that we identify these data doppelgangers, so that our models are truly validated to perform well. Some methods that we can use to identify data doppelgangers are ordination methods such as principal component analysis (PCA) or embedding methods like t-SNE. (1) Together with the analysis, we can then visualize the distribution of our sample using scatterplots. That being said, data doppelgangers might not be easily identified using this method. dupChecker is another platform, which screens for duplicates in datasets and can be employed in data doppelganger identification. Nonetheless, data doppelgangers in independently derived data sets might not get picked up by this platform. (1)

Pairwise Pearson's correlation coefficient (PPCC) is one of the promising methods for identifying data doppelgangers. By calculating the Pearson correlation coefficient between every pair of variables in a dataset, a large PPCC value is an indication that the datasets are highly correlated, whereby the values are nearly identical, suggesting the presence of data doppelgangers. (1) Wang et al. created *doppelgangerIdentifier*, a code package for R, which aids in data doppelganger identification based on PPCC. (2) In their study, they demonstrated the prevalence of doppelganger effect in gene sequencing data and how data doppelgangers can be identified using *doppelgangerIdentifier.* (2) Doppelganger effect is found to be present in other data types such as proteomics. (1)

References:

1. Wang, L. R., Wong, L. S., & Goh, W. W. B. (2021). How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today*, *27*(3), 678–685. https://doi.org/10.1016/j.drudis.2021.10.017
2. Wang, L. R., Choy, X. Y., & Goh, W. W. (2022). Doppelgänger spotting in biomedical gene expression data. *IScience*, *25*(8), 104788. https://doi.org/10.1016/j.isci.2022.104788