The original meaning of Doppelganger is a double that is not biologically related to the real body exactly like it. In data science, a data doppelganger refers to data that are independently derived but very similar to each other. The reliability of the cross-validation results of models for machine learning built using data with such characteristics often suffers. Since the training and validation sets are highly similar, validation presents high-quality results regardless of how the model was trained (Wang et al., 2022). The phenomenon of doppelgängers has been found in biomedical data, which I believe is related to the characteristics of biomedical data itself. The main types of biomedical data are omics data (including genome, transcriptome, proteome, metabolome, methylation group, microbiome, interaction group, etc.), drug and toxic substance data, disease data (including data on symptoms, disease-related genes, etc.), imaging data, wearable device data, Internet data, patient-reported data, etc. Among them, omics data and drug and toxicity data are often used for bioinformatics analysis. These data are characterized by a large amount, high dimensionality, nonlinearity, high noise, and uneven data distribution, which may be the reason for the presence of doppelgängers in some of the data. Other types of data may also have such characteristics, for example, the dataset obtained from the web by crawling for training when constructing a shopping recommendation system may have these characteristics. Information about items and user preferences can be measured by multiple features, so the data may be high-dimensional, and nonlinearity, high noise, and uneven data distribution may occur. Therefore, I believe that the doppelganger effect may also appear in the emergence of other types of data, not unique to biomedical data.

There are dupChecker, doppelgangerIdentifier, and other methods for identifying doppelganger effects in practice. dupChecker is reasonable in terms of method design principles, but still has shortcomings in identifying true doppelgangers. doppelgangerIdentifier uses the pairwise Pearson's correlation coefficient (PPCC) principle to identify data doppelgangers (Wang, Choy, et al., 2022). In general, there is still a lot of room for improvement in the identification of data doppelgangers. If

the efficiency and accuracy of recognition can be improved, it can also provide strong support to avoid the doppelganger effect in the future. In the actual building of machine learning models, avoiding data doppelgangers should be started from each step. Firstly, when generating or collecting data, try to use high-quality data and avoid data with characteristics that can easily produce data doppelgangers. The data can be normalized, regularized, or processed with dimensionality reduction algorithms such as PCA, tSNE, etc. when data cleaning is performed. If the proportion of doppelganger data is found to be large, it may be possible to reduce the proportion of doppelganger data by increasing the amount of data. There is also a lot of room for improvement in the method of avoiding data doppelgangers. If the impact of data doppelgangers on cross-validation can be effectively avoided, the machine learning models built will be more practical.

Reference

Wang, L. R., Wong, L., & Goh, W. W. B. (2022). 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

Wang, L. R., Choy, X. Y., & Goh, W. W. B. (2022). Doppelgänger spotting in biomedical gene expression data. *IScience*, *25*(8), 104788. https://doi.org/10.1016/j.isci.2022.104788