The Presence of The Doppelganger Effect

The doppelganger effect explains how seemingly harmless but organized events and transactions in cyberspace can have unintended consequences for natural persons. The result is distilled and clarified into an ordered singularity that, however wrong, seems plausible: misidentify an airline passenger and they will be put on a no-fly list, or a university professor with a financial grounding will be denied a big-box credit card. I think digital doppelgangers are also present in online society, and not just in biological data.

The "doppelganger effect" is a constant, replicating effect that is part of the networked society. It is not something controlled by material regulations or formal legal rules, but something to be considered in an ongoing effort to understand the impact that may continue to exist in the ever-increasing collection and dissemination of personal data. The personal identities created by databases as "perfect writing machines" may be right or wrong, and may be inaccurate in time or history, but in any case, their output has an impact on natural persons. The multiple ways in which an identity can be written to a database are completely beyond the control of a natural person. This is the heart of the doppelganger effect.

There are literally millions of instances of data exposure worldwide as a result of outright data theft by criminal effort; as high as 162 million records worldwide in 2007 (Erickson and Howard 2007; CBC 2008). Everything from stealing a wallet with credentialed identification, banking, and credit cards to sophisticated online breaches, phishing schemes, and database hacking. Such exposures trigger digital doppelgangers. Data is also exposed through public places such as Facebook, email, YouTube, and other personal and social media sites, where it can be exploited for fraud purposes as it can be incorporated into an insatiable appetite for news, gossip, and intrigue (Solove 2007). Identity exposure can manifest itself through the "doppelganger effect" -- cases of mistaken identity and damaged reputation -- exacerbated by the ability of the Internet to disseminate information, whether correctly or not.

One example concerns a breach of data security by a public sector organization in the United Kingdom. On 18 October 2007, HM Revenue and Customs (HMRC) and the National Audit Office (BBC.co.) lost two discs in transit containing 7.25 million child benefit records, involving 25 million people. HMRC and the government already have data protection policies in place to manage data whether transmitted manually or electronically, but these policies are not being followed. This example involves detailed personal information collected by the government and then collated into a multi-record database, including national insurance numbers; Name, address, and date of birth; Details of the marriage; Child's name, sex, and age and bank account details. Given the type of personal data stored in one place and its authenticity, this example would make it easier for future identity fraud to be a precursor to the doppelganger effect. This data breach represents a potential risk to data record retention due to the

total number of records stored together and the possibility of data and/or database consolidation. Database merging, that is, merging multiple records and/or multiple databases, is important for the Doppelganger effect. Consolidation also involves synthetic identity theft, in which multiple personal identity details are mixed with false facts to produce a new identity to enable identity-based fraud.

When such large amounts of data are exposed, individual instances of identity fraud will be difficult to trace, especially when multiple data records are used to compile synthetic identities. It is too early to tell the exact consequences, if any, of the death of TX and HMRC breaches, but the doppelganger effect suggests that sophisticated identity fraud can be the result. Digital doppelgangers copied from these events will contaminate individual credit reports in ways that are increasingly difficult to challenge (Solove 2007,18). The Doppelganger effect suggests that control code (such as a DBMS) will increase the coherence of fraudulent activities that result from the misuse of such a wealth of personal information.

The data exposure illustrates the destabilizing effects of the doppelganger effect. With each new data breach, it undermines data protection security by adding a layer of fact-checking and constant vigilance. Government agencies, such as the Office of Privacy and Information, and media reports tend to shift the onus of vigilance to consumers. However, individuals cannot physically interfere with data collection and storage practices and are often unaware of the impact of fraudulent activity on their credit and financial position. Thus, one result of the doppelganger effect is to consolidate data protection and security standards and encourage individual consumers to pre-emptively check their data. Both government and private sector organizations encourage consumers to check their financial and credit histories. This priority effect facilitates the verification of detailed information in private and public sector databases, which benefits the bureaucratic collection strategies of both sectors. The doppelganger effect mainly demonstrates the power of databases to replicate identities in ways beyond an individual's control.

How to Avoid Doppelganger Effects in Biomedicine

Data duplication has been observed in modern bioinformatics. In particular, data doppelganger effects can occur when some systems are evaluated on test sets with highly similar training sets.

Considering the possible confusion caused by dual effects, it is important to be able to identify whether there is a data dual effect between the training set and the validation set before validation. One logical approach mentioned in the paper is to use a sorting method (e.g., principal component analysis) or an embedding method (e.g., T-SNE), plus a scatter plot to see how the sample is distributed in a dimensionally reduced space. However, this method is not feasible because data duplication is not necessarily distinguishable in dimensionality reduction space. Earlier research on similar issues also suggested measures to identify data replicators. DupChecker is a method of

identifying duplicate samples by comparing the MD5 fingerprints of sample CEL files. The same MD5 fingerprint indicates that the sample is duplicated. Therefore, dupChecker does not detect real replicators of data as independent derived samples that are incidentally similar. Another measure paired Pearson correlation coefficients (PPCC), captures relationships between pairs of samples from different data sets. An unusually high PPCC value indicates that a pair of samples constitute a PPCC data replication. While sound and intuitive, the primary limitation of the original PPCC paper was that it never conclusively established the link between PPCC data replicators and their ability to obfuscate ML tasks.

We found that the presence of PPCC data replicators in training and validation data improves ML performance. The more doppelganger pairs represented in the training set and validation set, the higher ML performance. This indicated that there was a dose-based relationship between the number of DOPapelganger in PPCC data and the size of the doppelganger effect. This result confirms that PPCC data

Doppelgangers (based on pairwise correlations) act as functional doppelgangers (doppelgangers outcomes), producing inflationary effects similar to data leakage.

By putting all the Doppelgangers in the training set, the accuracy was reduced to 0.5, which is the expected accuracy of a model trained based on random signatures. Obviously, the Doppelganger effect is eliminated when all PPCC data are placed in the training set. This provides a possible way to avoid the Doppelganger effect. However, limiting PPCC data replicators to training or validation sets is a suboptimal solution. We should develop more comprehensive and rigorous assessment strategies based on the specific context of the data analyzed. Attempts to mitigate the Doppelganger effect with methods that do not result in a significant reduction in sample size or require large amounts of contextual data have also failed. I'm also less optimistic about pruning the data to mitigate the effects. The reason why sample pairs are highly correlated cannot simply be explained by a subset of highly correlated variables.

I also think the three tips for preventing Doppelganger effects are very important. The first is a careful cross-check using metadata as a guide. Using this information in the metadata, we were able to identify potential Doppelgangers and classify them into training or validation sets, effectively preventing the Doppelganger effect and allowing a relatively more objective assessment of ML performance. The second is the formation of data stratification. We can divide the data into different layers of similarity. Third, we can perform very robust independent validation checks involving as many data sets as possible (divergent validation). Although different validation techniques cannot directly combat data replicators, they can inform the objectivity of classifiers.

In summary, doppelgangers were fairly common in our test data, and it had a direct

inflationary effect on ML accuracy. This, in turn, reduces the usefulness of ML for phenotypic analysis and subsequent identification of potential drug cues. Therefore, to avoid performance bloat, it is important to check for potential Doppelganger in the data before classifying training and validation data ([1]).

References

[1] Robinson S J . The doppelganger effect: Spaces, traces and databases and the multiples of cyberspace. 2008.