Supervised learning from user-driven privatized data

General Problem Statement

At the moment many organisations – including large businesses as well as government agencies – collect sensitive personal data in centrally controlled and managed databases and data warehouses. What data are collected and how they are accessed, processed, and shared with other organisations is controlled by these organisations, but generally outside the control of individuals.

The emerging concept of Personal Online Data Stores (PODS), see: https://solidproject.org/, aims to change this centralised approach of where our personal data are being stored. Users of PODS are able to decide which of their data they share with what organisation. While this will allow fine-grained privacy settings by individuals (such as limiting the granularity of what data are shared), the resulting data sets that are made available for machine learning algorithms will potentially be of lower quality. For example, some individuals might not be willing to provide their gender to a certain project, while others are only willing to provide an age group (young, middle age, old) instead of their age in years. The benefits of individual privacy can therefore come at the costs of reduced accuracy when learning machine learning models.