Pseudonymization of data

This document contains a description of the suggested method for pseudonymization of data for the Covid-19 Cooperation Movement.

Scope of pseudonymization

The limited amount of person-identifiable data in the three initial datasets used in the Covid-19 Cooperation Movement, reduces the variables to be pseudonymized to solely be the Danish Person-identification (PID) number, the CPR number. This scope is limited to the datasets coming from Region Capital and Zealand (Personalescreening), SSI (tests and vaccines) and DBDS (blood-donor tests). When other datasets are available from the Covid-19 Cooperation Movement, this pseudonymization document must be revisited.

Goal

The goal for the method is to ensure that the CPR number is replaced by a unique, but non-identifiable ID instead. Each research project is required to get its own unique ID, such that information across projects cannot be shared. The pseudonymization method is required to be shared across multiple data sources and needs to be available to run on other data brought in from the individual research projects as well.

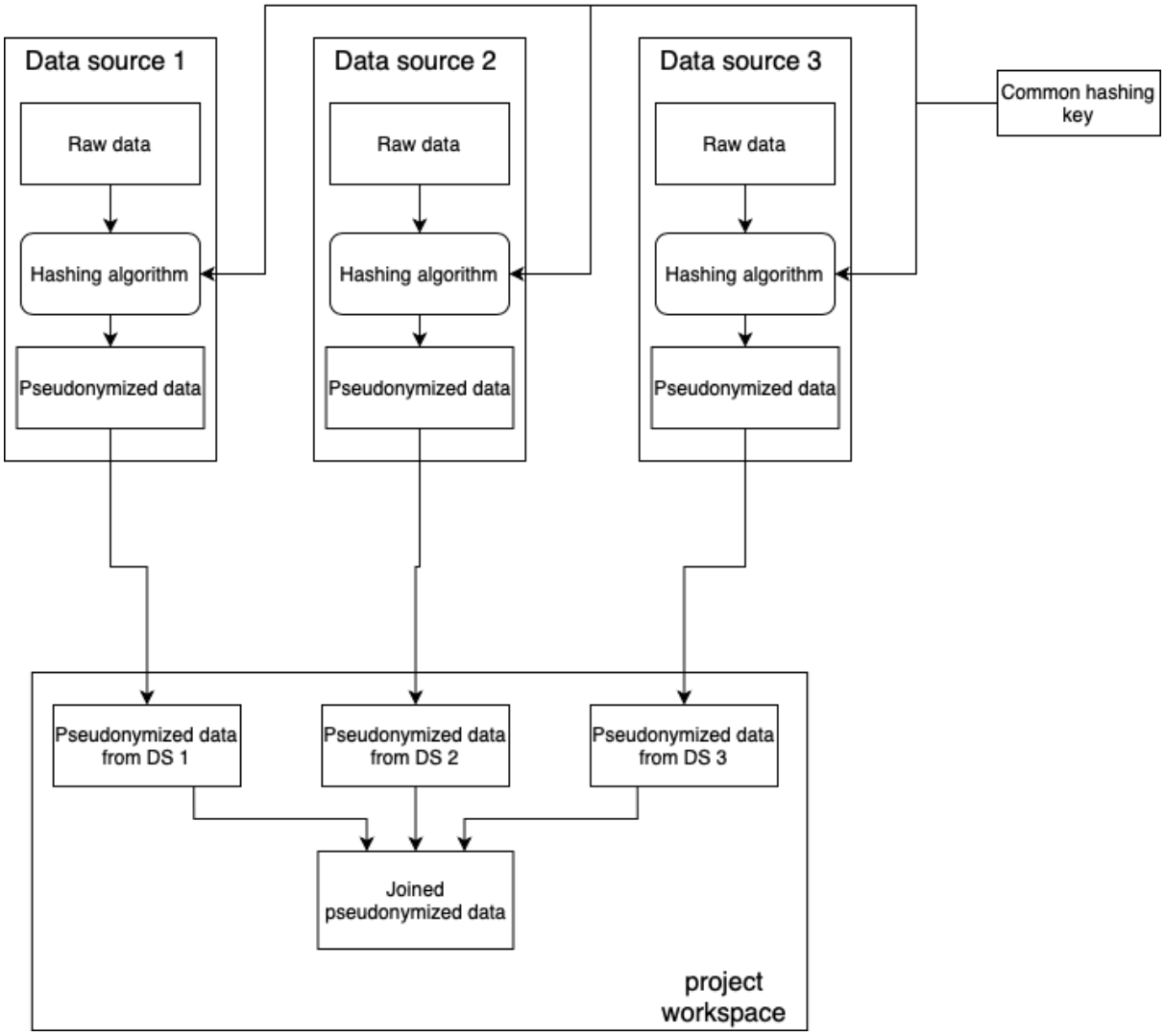
Method

Each data source stores their raw data (with CPR number) in their personal secure workspace. Each workspace will have one legal workspace owner and possibly a set of users with delegated legal rights. Only users appointed by the workspace owner will have access to the workspace and the associated data.

Pseudonymization by the data source workspace owners

When a research project is initiated at Computerome, a project-specific hashing key is generated with at least 16 digits. The hashing key will be stored encrypted in a secure environment and will not be made available for the project users. Each data source is provided the hashing key and will apply a hashing algorithm (e.g., SHA-256) on their dataset. The user performing the pseudonymization can be a Computerome employee, if authorized by data source workspace owner, and only if operating per request and by clear procedures. The pseudonymized dataset is automatically moved to the research project workspace and is ready to use. Joining/merging the data from all data sources is the responsibility of the research project workspace users.

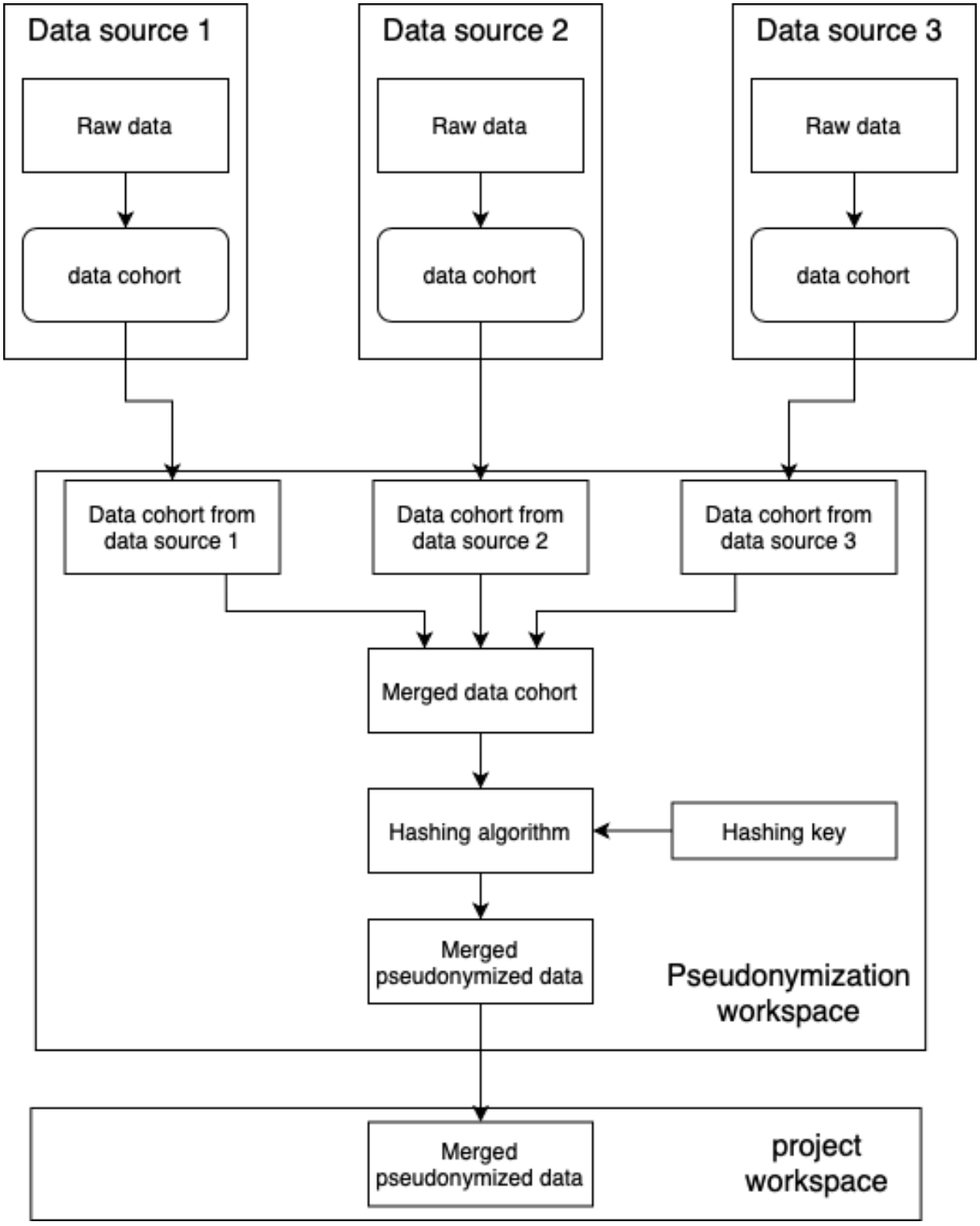
Below is a diagram of the process shown.



Pseudonymization by Computerome

When a research project is initiated at Computerome, a project-specific hashing key is generated with at least 16 digits. The hashing key will be stored encrypted in a secure environment and will not be made available for the project users. Each data source is prepared by the workspace owner according to request by applicant, and transferred to a dedicated pseudonymization workspace, accessible only by Computerome personnel. Computerome personnel will merge the dataset, apply a hashing algorithm (e.g., SHA-256) on all the merged data using the hashing key. The pseudonymized dataset is automatically moved to the research project workspace and is ready to use.

Below is a diagram of the process shown.



Merging with own research data

If the research project has its own data that needs to be joined with data from the Covid-19 Cooperation Movement, then either of the above methods can be extended to include this data. In both cases the data will need to be moved to a secure server, where a trusted person will be able to pseudonymize and potentially merge it along with the other data sources.

Consequences

The methods above are 1-way pseudonymization by design. However, due to the limited variety in CPR numbers, if you gain access to the hashing key it would be possible to brute force the CPR numbers from the hashes. Therefore, the hashing keys need to be stored encrypted in a secure environment. The advantage with storing the hashing key is that it allows for adding additional data during the project stages without having to reload all data again.

Both methods require the raw data to be stored at Computerome. These datasets will need to have tight access control about who is granted access, a process that is already under full control and subject to audits, cf. the Computerome Security Overview and related documentation.

The suggested methods are pseudonymizations and *not* full anonymization. This means it may still be possible to re-identify individuals based on the full, merged dataset. The risk of re-identification is especially high if a research project merges the Covid-19 Cooperation Movement datasets, as the researcher would be able to compare their original data with the pseudonymized data.