**Introduction**

Hello everyone, my name is Jiří Novák and I am grateful for the opportunity to introduce my research to you.

I am currently doing my second PhD at University of Zurich,

I am also a member of a research team at University of Northwestern Switzerland with whom I founded the Swiss Data Anonymization Center to provide services in this field.

**SDC**

In my research I focus on safe dissemination of data, which is particularly important in longitudinal data sets. Demographic variables are highly sensitive on re-identification that means you could cross reference different data sets to identify individuals and obtain new information about them. That’s unacceptable.

For this reason, many data sets cannot be disseminated and that is where my work comes in.

Using SDC and data anonymization methods, we can modify the data so that it still accurately reflects the real data but can no longer be misused.

Not only are we legally obliged to protect sensitive data, but it’s also crucial that we protect the reputation of statistical agencies and national offices.

~~Some examples of how this is achieved may include:  
value suppression (marking risk values as missing)  
category recoding (creating broader categories, like combining ages from 10 to 19)  
adding noise  
swapping data  
and so on.~~

Unfortunately, these traditional methods seem insufficient at properly protecting the data while keeping its utility. That’s why we are also exploring Synthetic data generation, that means creating an artificial data set with all essential statistical features of the original preserved.

In this illustrative example, I used XGBoost algorithm to simulate demographic variables from PSID dataset and here are the results for the age variable. The boxplots and distributions are the same that that means the univariate utility is preserved.

I invite you to reach out to learn more.

Thank you.