**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.

Dobrý den, jmenuji se Jiří Novák a dnes bych vám rád představil svůj výzkum na kterém pracuji ve Švýcarsku. Na univerzitě v Curychu studuji druhý doktorát, jsem členem výzkumného týmu na Universitě Severozápadního Švýcarska a na začátku roku jsem s kolegy založil Švýcarské kompetenční centrum pro anonymizaci dat za účelem poskytování služeb v této oblasti.

Ve svém výzkumu se zaměřuji na bezpečnou diseminaci datových souborů a v našem grantu se snažíme o vytvoření rámce pro longitudinální data. Demografické proměnné v datových souborech tohoto typu jsou vysoce citlivé na opětovnou identifikaci, což znamená, že případný útočník je může snadno použít pro získání nových informaci o určitém respondentovi. A to je naprosto nepřijatelné.

Z tohoto důvodu nelze mnoho datových souborů šířit, přičemž je v obecném zájmu, aby aspoň vědecká komunita k nim měla přístup. A proto přichází na řadu metody statistické ochrany důvěrnosti a datové anonymizace.

Pomocí těchto metod dokážeme upravit data tak, aby stále přibližně odrážela skutečná data, ale již nemohla být zneužita. Nejen, že jsme ze zákona povinni chránit citlivá data, ale je také zásadní, abychom chránili pověst národních statistických úřadů a agentur.

Tradiční metody statistické důvěrnosti nejsou úplně vhodné k řádné ochraně longitudinálních dat při zachování jejich užitečnosti. Proto zde zkoumáme generování syntetických dat, což znamená vytvoření umělého souboru dat se zachovanými všemi základními statistickými rysy originálu.

V tomto ilustrativním příkladu jsem využil data z Panelové studie dynamiky příjmů a k simulaci demografických proměnných jsem použil regresní strom s algoritmem XGBoost. Pro vyhodnocení jednorozměrné užitečnosti jsem vám zde vybral proměnou věku a můžete zde vidět box ploty v jednotlivých letech a dolní části jejich histogram. Boxploty jsou vedle sebe téměř identické a histogramy původních a syntetických dat se mi prakticky překrývají. To vyhodnocujeme úspěch pro evaluaci užitečnosti. Na tomto plátně to úplně ale neoceníte.

Závěrem bych Vás chtěl pozvat na diskuzi, hlavně pokud máte zkušenost s longitudinálními daty.

Děkuji.