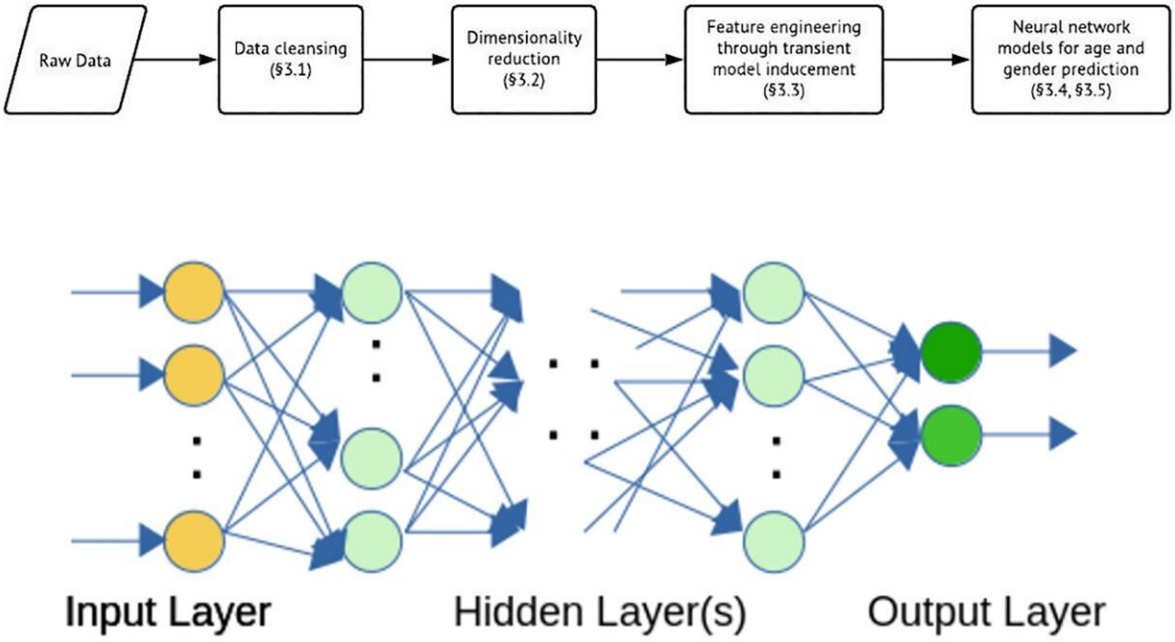
Extracting personal information from anonymous cell phone data using machine learning

October 12 2022, by Casey Moffitt



[Top] The research team's approach to modeling summary for the project. [Bottom] A feedforward neural network of how the information moves in the project. Credit: Illinois Institute of Technology

A research team at Illinois Institute of Technology has extracted personal information, specifically protected characteristics like age and gender, from anonymous cell phone data using machine learning and artificial intelligence algorithms, raising questions about data security.

The research was conducted by an interdisciplinary team of three Illinois Tech faculty including Vijay K. Gurbani, research associate professor of computer science; Matthew Shapiro, professor of political science; and Yuri Mansury, associate professor of social sciences. They were joined by Illinois Tech alumni Lida Kuang (M.S. CS '19) and Samruda Pobbathi (M.S. CS '19) who worked with Gurbani to publish "Predicting Age and Gender from Network Telemetry: Implications for Privacy and Impact on Policy" in *PLOS One*.

The researchers used data from a Latin American cell phone company to successfully estimate the gender and age of individual users through their private communications with relative ease.

The team developed a neural network model to estimate gender with 67% accuracy, which outperforms modern techniques such as decision tree, random forest, and gradient boosting models by a significant margin. They also were able to estimate the age of individual users with an accuracy rate of 78% by using the same model.

"Age and gender information does seem innocuous, but this information is used in nefarious ways by people, many times with devastating consequences," Shapiro says.

"When someone with bad intentions targets young children for anything, ranging from sales to sexual predation, it violates a number of laws designed to protect minors, such as the Children's Online Privacy Protection Act and HIPAA. At the other end of the age spectrum, seniors are targeted by sophisticated spam and phishing efforts given

their susceptibility and their access to savings."

This information was extrapolated using commonly accessible computing equipment. The team used a Linux (Fedora) operating system with 16 GB memory and an Intel i5-6200U CPU with four cores to run the neural network model.

"The laptop we used for this work is not exclusive at all," Gurbani says. "To a well-resourced adversary, there will be much more powerful machines available, including access to cluster computing, where multiple computers are configured in a cluster to provide the computer power for the AI/ML models."

The data set used to conduct the research is not publicly available, but Gurbani says an adversary could collect a similar data set by capturing data through public Wi-Fi hotspots or by attacking service providers' computing infrastructure.

"As we mentioned in our paper, such attacks unfortunately do occur and are not rare," Gurbani says. "The process to collect this data would not be easy, but it would not be impossible either."

The aim of the paper is to start a dialogue that critically examines the impact that emerging machine learning and AI techniques have on privacy regulations. There are no nationwide privacy regulations in the United States, so the researchers looked at how these techniques chip away at the European Union's General Data Protection Regulation articles, which are designed to protect consumers from the imminent threat of privacy violations.

"Machine learning and automated decision making will be a mainstream of business processes, and there is no escaping that reality," Gurbani says. "The issue at hand is how to protect individual privacy as well as

societal and economic interests from fraud using the appropriate regulatory framework."

One way to do that, Mansury says, is to provide consumers with the "opt-out option" to keep their personal information private when installing an app.

Recommendations include using synthetic data rather than user observation for machine learning models, for data holders to work with machine learning specialists to develop best practices, to build a regulatory framework that allows users to opt out of data sharing to keep personal information private, and to update existing non-compliance protocols. In other words, there is a lot more work to be done to address the policy gaps as well as the ethics of AI.

More information: Lida Kuang et al, Predicting age and gender from network telemetry: Implications for privacy and impact on policy, *PLOS ONE* (2022). DOI: 10.1371/journal.pone.0271714

Provided by Illinois Institute of Technology