Testimony of Latanya Sweeney
Commissioner, Commission on Evidence-Based Policymaking
House Committee on Oversight and Government Reform
Hearing of September 26, 2017

Chairman Gowdy, Ranking Member Cummings, and members of the Committee,

My name is Latanya Sweeney and my career mission has been to create and use technology to assess and solve societal, political and governance problems. I am a data scientist and the Director of the Privacy Lab at Harvard University. I also served as the Chief Technology Officer for the Federal Trade Commission. Thank you for the opportunity to speak before you today about the importance of the Commission's recommendations in protecting the American people's confidential data. I think it is most directly relevant for you to know that my research team and I have spent many years showing how individuals in supposedly confidential datasets on health, criminal justice, and other areas can be re-identified by using other data sources and computing power.

Let me be very clear: we have a problem protecting the privacy of confidential data today. The Commission believes we can securely increase access to confidential data for evidence building. But unless we also increase privacy protections as recommended in the Commission's report, we risk a bigger problem than barriers to data access. We risk exposing confidential data about Americans.

The Federal government collects a lot of information about individuals and businesses during the course of its daily operations. Much of that information is, and should be, open data—publicly accessible government information like weather forecasts and train timetables. The government says it will keep some of that information confidential, like the names and birth dates of social security recipients. When government pledges to keep data confidential, the data should have strong protections, and data use should generally be made known to the American public.

I'm here today to tell you about why the Commission's privacy and transparency recommendations are critical to protecting the government's confidential information. First and of utmost importance, there is great variation in how Federal agencies protect confidential data—this process should be consistent and rigorous. Second, protecting the privacy of the American people means being transparent, open, and clear about how their confidential information is being used and giving them opportunities to provide feedback.

So what happens now when a Federal agency wants to release a public use version of some confidential data it has collected? The 13 Principal Statistical Agencies routinely apply rigorous methods of data masking and seek review and approval from experts on a disclosure review board before releasing public use files. That is the current best practice and for a long time we accepted it as pretty sufficient. But the context of public use data releases has changed because the amount of information about individuals that is publicly available has grown. In addition, the technology to permit unauthorized re-identification has improved. Within the Federal government alone, the Open

Data initiative made over 150,000 datasets accessible through a single website, including many administrative datasets never before released to the public. While releasing these data can generate tremendous value, enabling entrepreneurs to produce better products and departments to understand their work better, it is important to consider how publicly available data could compromise confidentiality.

Government agencies follow their own applicable laws and regulations in providing access to their confidential data. The problem is that agencies have different policies and procedures for what it means to release data that is "not individually identifiable" (as it says in the Privacy Act). Some program agencies use the same best practice techniques as Principal Statistical Agencies to assess risk of re-identification. The Department of Education even set up a dedicated disclosure review board for its program agencies. But some program agencies do little more than remove direct identifiers such as name and address, and perhaps remove outliers before assuming the dataset is sufficiently de-identified for public release. And confidential government data collected by program agencies are often subject to FOIA with minimal redaction.

The problem is that there are so many sources of data out there today that can be matched to insufficiently de-identified confidential data to re-identify individuals or businesses. My colleagues and I just released a study showing how data on air and dust samples from 50 homes in two communities in California could be combined with data released under the Safe Harbor provisions of the Health Insurance Portability and Accountability Act (HIPAA) to "uniquely and correctly identify [in one community] 8 of 32 (25 percent) by name and 9 of 32 (28 percent) by address." Think about it: I can tell you, by name, health information about 8 people in one community from data that was released publicly as "de-identified." If those 8 people lived in your district and they learned that a Federal agency had just released their data with insufficient privacy protections, you would likely be hearing about it. This is what my colleagues and I discover every day, with many different types of data. This is a problem, and it's important that any legislation implementing the Commission's recommendations address privacy head on.

Many programs have released de-identified public-use data files for decades without being required to formally assess risk. The Commission's recommendation 3-1 will make sure that Federal agencies planning to release de-identified confidential data use state-of-the-art methods to protect individuals and businesses from privacy harm.

Next, I'd like to explain why transparency is so important to privacy. Privacy does not mean secrecy. In fact, the Commission believes that advancing beyond the status quo and achieving unparalleled transparency means first, telling the public about how government data are used for evidence building and second, regularly auditing whether the government is doing what it said it would do to protect privacy when allowing access to government data for evidence building.

As a first step, the government needs to make clear its decisions about which data are open data and which data are nonpublic, confidential data. The Commission calls for OMB to develop a public inventory of data available for evidence building, including a determination of the sensitivity level of the data. Based on data sensitivity, we are recommending that OMB establish standards for appropriate and controlled access. And this is important—agencies should use technology and statistical masking techniques to develop less sensitive versions of datasets when possible and make more information available to the public and to researchers with appropriate safeguards for evidence building.

This idea of multiple versions of datasets, or tiered access, isn't some theoretical concept. Tiered access approaches are practical and we can actually implement them today. Tiered access is an application of data minimization, which is a key privacy safeguard for evidence building. Imagine a system where each dataset is labeled based on whether versions have more or less identifiable or sensitive information. Then, researchers or the public only receive an appropriate level of access to complete their project with appropriate privacy safeguards.

This tiered access approach is a way to increase evidence building and better protect privacy at the same time. And we already have examples of how it is being done today in the Federal government. The Commission found that many PSAs implement tiered access programs that set data access and security requirements based on an assessment of dataset sensitivity. Tiered access is also taking root in Europe in response to the implementation of the General Data Protection Regulations and at home by organizations such as my own, Harvard University, to define sensitivity levels and set corresponding access and data security protections.

Recommendation 4-3 calls for OMB to develop a transparency portal that includes the data inventory with sensitivity levels, risk assessments, and descriptions of projects and approved researchers using confidential data for evidence building. Researchers like myself and my colleagues must have a way to give feedback to the government about potential risks they find. And for the public to provide feedback on data sensitivity. Therefore, the Commission's report calls for a feedback mechanism as part of the transparency portal.

Preventing bad actors from breaking into confidential data they are not authorized to use requires consistent and rigorous processes to assess the risk of release in light of all other sources of data. A disclosure review board has the expertise to determine if agencies are doing enough to protect the privacy of the American people. The Federal government needs to help the public understand how confidential data are being used and conduct regular audits to ensure compliance with privacy laws, regulations, and best-practice procedures.

Thank you for your time. I urge you to move swiftly to implement these changes and improve how the Federal government protects the confidential data it collects from the American public.