[Slide 1] [Before I start, I want to remind members of the audience that, while I am appearing in my official capacity as the Chief Scientist of the U.S. Census Bureau, I am presenting a summary of research findings. The views expressed in this talk are my own, not those of the Census Bureau.]

Staring Down the Database Reconstruction Theorem

[Slide 2] The 2020 Census will be the safest and best-protected ever. This is not nearly as easy as it sounds.

Throughout much of the history of the decennial census, our country has struggled with two challenges:

1) collect all of the data necessary to underpin our democracy;
2) protect the privacy of individual data to ensure trust and prevent abuse.

The first obligation derives directly from the Constitution, of course. As for the privacy requirement, Section 9 of the Census Act (Title 13 of the U.S. Code) prohibits making "any publication whereby the data furnished by any particular establishment or individual under this title can be identified." In fact, the Census Bureau is about the only organization operating under a blanket U.S. legal requirement never to release data that can be tied back to individuals or companies no matter what.

The Census Bureau has always been committed to meeting both of its obligations; that is, providing population statistics needed by decision-makers, scholars, and businesses while also protecting the privacy of census participants.

A paper by Laura McKenna (2018), who supervised the confidentiality protection systems used by the Census Bureau for more than 15 years, catalogued the public information about the technical systems used for protection of publications from decennial censuses since 1970.

As McKenna noted, beginning with the 1990 Census, the primary confidentiality protection method employed was household-level swapping of geographic identifiers—moving an entire household from one location to another—prior to tabulating the data. The goal was to introduce uncertainty about whether households allegedly re-identified from the published data were correct.

Essentially the same methods were used for the 2000 and 2010 Censuses but with refinements that recognized the changing external environment.

The discipline of statistics has evolved over the last century. So too has the widespread availability of data. With each new development, the Census Bureau must ask how the current state of affairs will affect the production of the statistical products that it releases to the public so as to be both useful and privacy-preserving.

Sixteen years ago, two computer scientists, Irit Dinur and Kobbi Nissim (2003), wrote a seminal article proving a "database reconstruction theorem," which is also known as the "fundamental law of information recovery."

Three years later, Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith (2006) provided a mathematical foundation for what we now call "differential privacy." In short, they explained how to quantify the limits on the accuracy of answers to queries based on the confidential data and the privacy-loss to the entities in those data, when the queries are answered publicly. More importantly, they provided a technique for enhancing privacy that goes far beyond the swapping approach that many statisticians have been using for years.

[Slide 3] The full implications of database reconstruction were not understood in 2003, but over the next several years a scientific consensus emerged in the data privacy community that:

- **Too many statistics**, published too accurately, expose the confidential database with near certainty (Dinur and Nissim 2003).
- **A necessary condition for controlling privacy loss** against informed attackers is to add noise to every statistic, calibrated to control the worst-case disclosure risk, which is now called a privacy-loss budget (Dwork, McSherry, Nissim and Smith 2006; Ganta, Kasiviswanathan, and Smith 2008).
- **Transparency about methods helps rather than harms,** Kerckhoff's principle, applied to data privacy, says that the protections should be provable and secure even when every aspect of the algorithm and all of its parameters are public. Only the actual random number sequence must be kept secret (Dwork, McSherry, Nissim, and Smith 2006).

If you curate confidential data, then you can use those data for two competing goals:

- You can publicly and precisely answer statistical queries about the data.
- You can preserve and protect the privacy of those whose information is in the data.

You can do some of both.

[Slide 4] But if you do all of one, you can't do any of the other.

Period.

This trade-off is one of the hardest lessons to learn in modern information science. It is a lesson about data generally, not about counting people. And it is a mathematical theorem, not an opinion or implementation detail.

[Slide 5] This transformation in the fields of statistics and computer science is truly mind-blowing. It's at the heart of the science that we're here to celebrate. Cryptographers usually study the safety of methods for encrypting information about private data. Now their insights show us safe ways to publish information from private data. The cryptographic approach shows that some new methods can provably protect privacy, and some old methods provably do not. But the safe methods only work if we accept the inherent limitations on the accuracy of those publications that the cryptographers have highlighted.

Specifically, technical advances revealed a new vulnerability, allowing people to reconstruct data from tables that were previously assumed to be privacy-preserving, given the available computing resources. But other technical advances have also enabled a new form of privacy protection that is not only more sophisticated but also mathematically grounded in a way that allows statisticians to fully understand the limits of what they can make available and what kind of privacy they can provably offer. This dual breakthrough is transforming how we protect data today.

Good science and real privacy protection turn out to be partners, not competitors, in the efforts to modernize the methods data analysts use. For this reason, we have seen many companies, like Google, Microsoft, and Apple, turn to differential privacy to secure data and make guarantees about the privacy of

statistical tables. But it was actually the Census Bureau who first recognized the power of this method at scale.

[Slide 6] In 2008, the Census Bureau implemented an early version of differential privacy on data that display the commuting patterns of people based on where they live and work (Machanavajjhala et al. 2008; U.S. Census Bureau 2019).

Working with statisticians and computer scientists, we have collectively advanced the state of differential privacy such that we are going to implement it at scale as part of the 2020 Census. While I will talk about what that looks like in more detail tomorrow at 8:00AM, today I want to explain why we absolutely must implement differential privacy in order to protect the privacy of those participating in the census.

Starting in 1972, researchers began highlighting how it was possible to combine statistical tables and use differencing techniques to identify which census respondents provided the associated data (Fellegi 1972). As the market for detailed data grew and evolved, researchers also began highlighting how combining commercial data with census tables could introduce new vulnerabilities. While external users could not provably know whether or not their reconstructions were accurate, the Census Bureau recognized that it was critical to know the potential vulnerability of census data.

We acted proactively, as the Census Bureau has done for many decades. We designed our own internal research program to assess the current state of this vulnerability without waiting for a specific external threat. I'm now going to explain what we found.

[Slide 7] Here are the steps we followed:

- Using only published contingency tables (summary statistics), we applied the database reconstruction theorem to construct record-level images for all 308,745,538 persons enumerated in the 2010 Census. A record-level image is a row in the reconstructed database with the same variables that were used in publications from the confidential database. There is no traditional PII (personally identifiable information) on these reconstructed records.

- Using only the information in the reconstructed data records, we linked those records to commercial databases to acquire name and address information. This information would have been available to an external attacker, circa 2010.
- When the record linkage operation is successful, the PII from the commercial data are attached to the reconstructed census record. We call the reconstructed record, now laden with PII, "putatively re-identified," which means that an attacker might think that the attack was successful.
- We then compared the putatively re-identified census records to the real confidential census records. When this comparison matched on all variables, including the PII and those variables not available in the commercial data, we called this a "confirmed re-identification."
- The harm from such re-identifications, in the 2010 Census, is that the attacker learns the self-reported race and ethnicity on the confidential census record. Those data are not available in identifiable form to any commercial or governmental agency except the Census Bureau.

[Slide 8] Here are the basic results:

- In the reconstructed data, certain variables are always correctly reconstructed—meaning that the value in the reconstructed variable always matches its value in the confidential data. The census block, where the person lived on April 1, 2010, is always correctly reconstructed. This is true for every one of the 6,207,027 inhabited blocks in the 2010 Census.
- All the variables we studied: block, sex, age in years, race, and ethnicity are exactly correct in the reconstructed records for 46% of the population (142 million of 308,745,538 persons)—meaning that the reconstructed record exactly matches the confidential record on the value of all five variables. This result is salient because in the confidential data, more than 50% of the records are unique in the population—the only instance of this combination of values observed in the census (the exact percentage is confidential). If we allow the age to vary by plus or minus one year, then the number of reconstructed records that match the confidential data on these five variable rises to 71% (219 million of 308,745,538 persons).
- When we use the reconstructed block, sex and age to link each reconstructed record to the records harvested from commercial data

acquired at the time of the 2010 Census, we putatively re-identify 45% of the total population (138 million of 308,745,538 persons). That means that we were able to attach a unique name and address to 45% of the reconstructed records from the 2010 Census. The match is exact for block and sex. Age is allowed to vary by plus or minus one year.

- When we compared the unique name, block, sex, age, race, and ethnicity on the putative re-identifications to the same variables on the 2010 Census confidential data, we confirmed 38% of these matches (52 million of 308,745,538 persons, or 17% of the total population).

The putative re-identifications probably have a recall rate (or sensitivity) of at least 45%. Neither the attacker nor the Census Bureau have PII on all 308,745,538 persons enumerated in the 2010 Census, so the correct recall rate denominator is certainly less than the total population.

The precision of the record linkage is 38%, which means that the attacker would be correct between one-quarter and one-half of the time.

And both of these estimates (45% putatively re-identified; 38% of which are correct) are really lower bounds for other reasons: our experiments didn't use all of the information that the Census Bureau published from the 2010 Census. For example, we didn't use any information on household composition, which means that potential harm from discovering other features of households, like same-sex unions and adoptions, is still unquantified. We also made no use of the 2010 Public-Use Microdata Sample.

To further put these results in context, the last time the Census Bureau released results for a re-identification study, which did not use database reconstruction (Ramachandran et al. 2012), the putative re-identification rate was 0.017% (389 persons of 2,251,571) and the confirmation rate was 22% (87 of 389).

[Slide 9] All of us—the entire scientific community—have an obligation to examine the methods we use in light of the cryptographic critique of the privacy protections those methods offer. We must also recognize that these developments are sobering to everyone.

This is not just a challenge for statistical agencies or Internet giants, although those institutions have been in the vanguard of this movement.

It's a challenge for Internet commerce, because recommendation systems expose private data.

It's a challenge for bioinformatics, because summaries of genomes expose private data.

It's a challenge for commercial lenders, because benchmark risk assessments expose private data.

It's a challenge for nonprofit survey organizations, because their research reports expose private data.

Regardless of what anyone says, people want to be assured that their data are private. They want to know that we can't use statistical magic to re-identify information that they thought was private. They want to know that statistical tables can't come back to haunt them.

That's why I'm so grateful that the data we are showing today aren't the end of the story. They simply show that we cannot accept the status quo. We cannot presume that what worked a decade ago will work again in 2020. We have to innovate. And that's what we are doing.

In 2016, the Census Bureau acknowledged that database reconstruction was a vulnerability of the methods traditionally used to protect confidentiality in decennial census publications.

What we showed today is that we have a clear understanding of how it's possible to reconstruct 2010 Census data for block, sex, age, race and ethnicity. But this understanding isn't in vain. This understanding gave us the information we needed to develop techniques to make sure this isn't possible in 2020.

We are going into the 2020 Census confident that we can protect the privacy of all who participate. We have to make some important decisions about what statistics should be made available and how to weigh public data interests with our commitment to keep individual data private from reconstruction. But we know where the vulnerabilities are and we have the tools to make certain that what I showed today can't happen in the future.

The publications of the 2020 Census will be protected by differential privacy because it's imperative above all else that we ensure the trust of the American people.

The exact algorithms, and all parameters, will also be publicly released well in advance of the tables because it is imperative that we be accountable to the scientific community and the public at large.

[Slide 10] Statistics has evolved significantly over the last century. I'm honored to be a part of a statistical agency with a long tradition of implementing cutting-edge knowledge on the behalf of the American people. And I'm deeply grateful to the amazing team at the Census Bureau for identifying the challenges we face and ensuring that we can meet those challenges.

I promise the American people that they will have the privacy they deserve.

For those who would like to know more about how we are implementing differential privacy in the 2020 Census, please join me tomorrow at 8:00 AM where I will present our methods in more detail.

References

Dinur, Irit and Kobbi Nissim. 2003. Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '03). ACM, New York, NY, USA, 202-210. DOI: 10.1145/773153.773173.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. In Halevi, S. & Rabin, T. (Eds.) Calibrating Noise to Sensitivity in Private Data Analysis Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings, Springer Berlin Heidelberg, 265-284, DOI: 10.1007/11681878_14.

Fellegi, Ivan P. 1972. On the Question of Statistical Confidentiality. Journal of the American Statistical Association, Vol. 67, No. 337 (March):7-18, stable URL http://www.jstor.org/stable/2284695.

Ganda, Srivatsava, Shiva Kasiviswanathan and Adam Smith. 2008. Composition Attacks and Auxiliary Information in Data Privacy. In Knowledge, Discovery and Datamining, Las Vegas, NV, doi:10.1145/1401890.1401926.

Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. Privacy: Theory Meets Practice on the Map, International Conference on Data Engineering (ICDE) 2008: 277-286, doi:10.1109/ICDE.2008.4497436.

McKenna, Laura. 2018. Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing, Working Papers 18-47, Center for Economic Studies, U.S. Census Bureau, Handle: RePEc:cen:wpaper:18-47.

Ramachandran, Aditi, Lisa Singh, Edward Porter, and Frank Nagle. 2012. Exploring Re-Identification Risks in Public Domains, Tenth Annual International Conference on Privacy, Security and Trust, IEEE, doi:10.1109/PST.2012.6297917.

U.S. Census Bureau. 2019. LEHD Origin-Destination Employment Statistics (2002-2015) [computer file]. Washington, DC: U.S. Census Bureau, Longitudinal-Employer Household Dynamics Program [distributor], accessed on February 15, 2019 at https://onthemap.ces.census.gov.