Please stand by for realtime captions.Good afternoon everyone. Welcome to the F DLP Academy. We have another terrific webinar for you today. My name is Joe from GPO. I am with my colleague Cory Holder as tech-support. Today's webinar, why should librarians know about differential privacy in the 2020 census? with us today is our presenter, David Van Ripper. He is the director spatial analysis at the Minnesota population center. He is the principal investigator of an age GIS, national historical geographic information system. Which provides access to historical and contemporary small area census data. David also leads the team researching differential privacy and its impact on decennial census data. Before we get started on the webinar, I have to go through some of my usual housekeeping comments. If you have any questions you would like to ask David, or if you have any technical issues, please feel free to use the chat box which most people on desktops or laptops, it is located in the bottom right-hand corner of the screen. I will keep track of all questions that come in, and at the end of the presentation, David will respond to each of them. We are recording today's session and will email a link to the recording and slides to everyone who registers for this webinar. We will also send you a certificate of participation using the email you used to to register for today's webinar. If anyone needs additional certificates because most of people watched the webinar with you, please email FDLP out region include the title of today's webinar along with the names and email addresses of those needing certificates. Desktop computer or laptop users may zoom in on the slides being presented. Click on the full-screen button in the bottom left side of your screen. To exit the full-screen mode, mouse over the blue bar at the top of your screen so it expands, then click on the blue return button to get back at the default view. Finally, at the end of the session we will share a webinar satisfaction survey with you. We will let you know when the survey is available. The URL will appear in the chat box. We very much appreciate your feedback after the session is through today. Including comments on presentation style and value of the webinar. With that, I will hand the virtual microphone over to David who will take it from here.

Thanks very much Joe. I am super excited to present today. I was telling Joe before this that we used a lot a federal depository library materials for holding, so it is a real pleasure to get to speak to people who work with that group today. I'm very excited. My name is David Van Ripper. My contact information is on the first line. Should any of you have follow-up questions, my email address is there. And you can follow me on Twitter, D.C. Van Ripper. I tend to tweet about the senses and differential privacy as well as a few other topics. Today I want to talk about what I think is important for librarians to know about differential privacy in the 2020 census. Protecting the privacy of census respondents while publishing quality data are tool mandates for the U.S. Census Bureau. In August 20 18 the Bureau announced a major change in their approach to privacy protection. According to the Bureau, increases in computing power and access to large individual level databases mean that their traditional disclosure avoidance techniques no longer provide strong enough protection. In response, the Bureau is adopting a framework termed differential privacy 2020 census disclosure avoidance system which entails the injection of random noise into nearly all published data in order to guarantee a minimal risk of privacy laws. But this privacy protection comes at a cost. The published data will be less accurate. How did we get here? in August 2018 the chief scientist and associate director for research and methodology of the Census Bureau published a blog post describing a database reconstruction every identification attack the Bureau undertook. They essentially took the 2010 summary file one data that were published at the census block level and tract level and took the published cross

tabulations and converted those into individual census micro data. They then matched those individual census micro data with a large commercial database. They linked on age and sex and census block ID. That attached names to each record in the reconstructed micro data. They then went in and took that reconstructed data with names on it, they took it into their internal data system and linked it with the confidential 2010 responses, which included our names. With this reorganization attack, the Bureau we identified 17% of the 2010 decennial respondents, approximately 50 million people. They were able to re-identify them from the published tables. Census argued that prior disclosure avoidance techniques, which included whole table suppression in 1970 and 1980, data swapping and blanket in 1990 and data swapping in 2010 did not provide enough privacy protection for individual respondents. Two weeks later John published a second blog post describing these previous disclosure avoidance techniques and argued that there is a lack of transparency surrounding these techniques. He argued the differential privacy would be more transparent if census could publish details about the code and the noise injection parameters used to protect the privacy.. In February 2019, just over one year ago, the deputy director and chief operating officer of the Census Bureau made an official announcement that differential privacy would be used for the 2020 decennial census. This blog post indicated that complaints that had come out of groups like mine and others, they were going to be rejected by the Bureau, and this differentially private algorithm for protecting privacy would be used. At the time that this blog post came out, the Bureau was still planning to use a differentiated see driver seat mechanism to protect responses from the American community survey. In July he walked to that back and stated that differential privacy would not be used for protecting ACS data until 2025 at the earliest. That is because the American community survey in such a complex data set that they currently don't have algorithms that will scale to protect respondent privacy in the ACS. They will continue to use traditional disclosure techniques for the ACS. Okay, so in today's talk I want to talk about -- that sets the table. That is how we got here. The Bureau is going to use this new differential privacy based algorithm for protecting privacy. So now that that is there, what does that mean downstream essentially for data users over the long run? first I'm going to go over this new disclosure avoidance algorithm that the Bureau is using. I want to talk about what that means in the downstream effect. A few of what I call in variance. Less publicly available data for end-users. There will be lots of consistency among the publicly available data products. All of these things, I think are really important for librarians to know about. Librarians tend to work with users, they work with the general public, people come to you and ask questions about what data are available, whether or not that data is comparable to data that we are collecting in 2000 or 2010 or 1990. It is critical for librarians to understand what types of changes are going on to the census data and what that means for data users that come to you for support. The new disclosure avoidance algorithm is based on differential privacy. Before I talk about that, I want to talk a little bit about how the census handles disclosure avoidance in the past. The left column here, suppression and swaps. That is how the Census Bureau has traditionally handled disclosure avoidance. What they do is topcoat for household income size. They won't report the income values for anything over 200,000 or 250,000. They won't publish data for household sizes that exceed seven people. This provides some privacy for the people who live in those large household sizes or who have high incomes. They would suppress whole tables. If cell sizes were too small, the Bureau would suppress the entire table not publish the data, keeping information, if they keep the information from users, the users can't use the data to identify firms or individuals. The final thing they started using in 1990, and this is what they use today and ACS and was used in 2010, is what is called swapping.

The Bureau what identify unique households along dimensions such as race, ethnicity, age, household structure, household type. They would identify unique household in a geographic area. It could be a block, census block, and they would swap that household with a similar household in a different geographic unit. They might say that your household is too identifiable, we are going to take your data and essentially assign it to a different census block ID in Minnesota. Then we are going to take someone from that census block ID and assign them to where you live. And that swapping essentially protects the ability for someone to identify me. The key thing to know about swapping is that it tends to be a more absolute style of disclosure avoidance. The Bureau would try to find particular households in particular areas, and they would say we can't let these people by identified. We are going to work hard to protect the privacy of those individuals. Differential privacy is quite a different set up from that absolute guarantee of privacy protection. Differential privacy injects noise in the statistics. The swapping works at the micro data level. The Bureau will actually swap households in the data and then create the tabulations from the swapped data. And differential privacy, the Bureau will construct cross tabulation, they will count the number of white household heads who live in census tract 3003 in St. Paul Minnesota. They will produce a counter that and inject random noise into that count. That random noise is based on a particular statistical distribution which is defined by a set of policy decisions that are made by a group of individuals at the Census Bureau. The policy decisions have a great impact on the magnitude, the size of that noise. The Bureau and the committee can decide that they want to make particular statistics or particular geographic areas to be more accurate or less accurate than others. So there's a lot of policy decisions that can have large impacts on the accuracy of the output data. What all this does is provide a more relative guarantee for privacy. So instead of saying this particular household is identifiable, we are going to move them somewhere else, the differentially private mechanism provides a more relative guarantee for everybody who lives in that particular geographic unit. There is a lot of complicated mathematics that go into differential privacy, and I admit I don't understand all of the mass. Is far beyond my training in that subject, but the thing that statisticians and mathematicians and Bureau scientists like about differential privacy is that it provides is relative guarantee that can be mathematically proven. Swapping, on the other hand, wallet may provide absolute guarantee against identification, it doesn't have a mathematical proof behind it. I just want to lay out how things used to happen and how things are going to happen going forward. You must be wondering how is differential privacy actually implemented. What I'm going to use is a toy example to explain the basics of the implementation. The first thing I want you to know is that differential privacy is not an algorithm. Differential privacy is a mathematical definition. There are lots of algorithms that can be differentially private in the sense so spirit is developing one particular algorithm that implements differential privacy. Differential privacy on its own is not a single algorithm. The algorithm that I'm going to describe in this example is what I think of as the most naïve algorithm that you can use to implement differential products. Let's say we start with a micro data set that contains two variables and that we did a survey of people and asked them to questions. We asked them what is your sex and what is your school attendance and the sex variable has two different values, male and female, the school attendance variable has three possible responses, one that they never attended school, two that they are currently attending school, and three, they attended school in the past. We take the responses from the survey and create a cross tabulation from the true data. Here we have a two by three cross tabulation, each row is a sex response, male or female. Each column is a response from cool school attendance. You can see we have three males who have never attended school, 12 who are

currently attending school and 33 who attended school in the past. You can see here from the responses we have a total count of 100, 100 people who responded to the survey. In order to make this data differentially private, but we do is draw a particular statistical distribution, in this case what is called a Lepage distribution. What we do for each cell in that cross tabulation, we draw a random value from this Laplace distribution. Then we add the value of that random draw to each cell. The width of this statistical distribution, and you can see here that the distribution is peaked around zero, and then it falls off fairly rapidly until we get up to values of 20 or -20. The spread of that particular distribution is based on this policy decisions I talked about before. So we can have -- policymakers who think the state are not particularly sensitive. So we can have a fairly narrow statistical distribution which means that a large fraction of a random draws are going to fall kind of clustered around zero, as you can see here. But as you notice from the set of random drawings, we do have a probability of drawings that are somewhat large. After we have done the random draws, we then add those random draws to the original sales in the cross tabulation. Males never attended, we drew a value of -1 from that distribution, so we change that value from three down to two, and the bold numbers are what would be publicly available in the public data set. In most cases you can see that the values are really not that different from the true data, but you can see for females who have never attended school the value tripled from the true count of four up to a count of 12. You can also see that the overall count in the data set, the sum, has gone from 100 up to 108. So now we no longer have a one-to-one relationship between the published data, which are the bold values here, and the true data, which are the block values that are not bolded. All of the sudden if we reconstructed the individual level responses from the noisy data we no longer have a one-to-one relationship this provides that, we were not disclosing that much information about the people that responded to the survey. But the drawback to this as you can tell is now we say we have 12 females who never attended school, if I was a school planner I maybe sing I have got 12, more people coming in, I may plan to hire next the teacher, but all of the sudden school starts and there are a few more people than I expected and I put resources towards an extra teacher to teach the students, and we did not need to do that. The public data presents a false picture of what is going on in this particular implication. So that is, as fundamental, that is really what differential privacy is, you create cross tabulations, you draw values from a particular distribution, you add those values to the cross tabs, and you publish the data. The Census Bureau algorithm is much more complicated than that, but fundamentally that is what it does. There are many implicit policy decisions that were embedded within this example. That is what I want to talk about, the major policy decisions that the Census Bureau must make before it runs and implements its differentially private algorithm. This is where I could talk for a long time about all these policy decisions, but I'm going to go over them fairly quickly. There are really three fundamental policy decisions that the Bureau has to make before they can run their algorithm. One is that they have to decide what they call and what I call the global privacy laws budget. This is often in the computer science literature denoted by the epsilon. The way to pick about the global privacy laws budget is that the larger the value of epsilon that you see, the more accurate the public data will be, the more accurate the public data are, the less privacy protected they are. The smaller the value of that epsilon, the less accurate the data will be in the less accurate the data are, the more privacy protective they are. It is this trade-off. You can protect peoples privacy more, but that is going to lead to less accurate data. You can prioritize accurate data, but that is not going to provide as much privacy protection. In the theoretical computer science literature, a value of 0.1, and epsilon value of 0.1 is kind of the largest that computer scientists would like to see. That is a very low value epsilon. A very

privacy protecting epsilon. The Census Bureau is currently looking at a minimum epsilon value of 6.0. A value that is magnitudes larger than the theoretical maximum, but the computer scientists don't always work with real data and data that are used by policymakers, so they tend to focus more on the privacy protection aspect as opposed to the data use. That global privacy laws budget must be cross tabulated and geographic levels, which cross tabulations in which geographic levels get this allocation, it indicates how accurate that will be or how inaccurate those cross tabs or geographic units will be and finally they have just decide on variance and constraints in the data. I will talk about those in a little bit. I don't want to jump the gun on that. I talked a little bit about the privacy laws. Let's talk about the fractional allocations. The Bureau, once they have set their global privacy laws budget they have to spend it. Essentially it is a value and they can allocate that budget to the geographic levels or queries. The larger that allocation is, the more accurate the data will be for a particular geographic level or query. It is fundamentally important that the Bureau takes user input into account, they take into account the legal requirements for certain data sets, whether for redistricting or to run the population estimates program, they have to take a lot of input into account is a decide how to do the fractional allocation. In a data set that the Census Bureau released in October of last year, the Census Bureau actually released a demonstration data product that was derived from the 2010 decennial census. Several file one, the Bureau constructed a private version of that and made it publicly available. Lots of people who I work with and lots of regions in the country went off and analyzed those data to see how it compared to the original data that was published in 2010. In the demonstration data that the Bureau published, they made direct fractional allocations of that privacy budget to the geographic units in red. So nationstate County, census tract block groups and census blocks in this new geographic level called tractor group, all of those received specific allocations of the privacy budget, which means that those geographic levels are going to be more accurate. All of the units that are not in red, so County subdivisions, places, ZIP Code tabulation areas, school districts, state just later districts, those received no direct allocation of the privacy budget, so their accuracy is really controlled by the accuracy of the census blocks that comprised those units. For the queries, the Census Bureau allocated direct amounts of the privacy laws budget to 13 different queries. This just lists for. One thing you need to know if you are devotees of the summary files that the Census Bureau publishes, these queries do not directly correspond to data tables that are in the summary files published by the census. These queries are simply used by the disclosure avoidance algorithm to create the final set of micro data that are run through the Bureau tabulation system that generates the summary file such as summary file one or two. So the Bureau does not take the summary file tables and work from those. Instead they work from these 13 queries that are used and in particular, I want to highlight the second one on this list, the voting age by Hispanic, by race, by citizenship query, this data were run before the Supreme Court had made a decision on the citizenship question. The citizenship question is, was taken off, but the Bureau still -- the code was already locked when they ran the data, but of course, as we all know, this question is not part of the decennial census. That received 50% of the privacy laws -- privacy loss budget for the demonstration data. That shows they redistricting use case, and the other types of topic such as detailed housing or detailed person inquiries received much less allocation. That shows you how the Bureau can privilege certain queries over others, making somethings more accurate than others. The final thing that the Bureau has to decide about is what are called invariants and constraints. Variants are accounts that received no noise injection. Those are the counts that will be published as enumerated in the 2020 decennial census. As of today, the census has not decided on a final set for the 2020 census, but they did

make some decisions about invariants for the 2010 demonstration data that we all analyzed. So in the 2010 demonstration data product there were four invariants. These were counts that were published without noise injection. The state total population was set to be invariants and at the census block level three different counts were invariant. Number of total housing units, the number of group quarters, and the number of group quarters by count. Those four counts, there was no noise injected into them. It differs quite a bit from the 2010 decennial census, the summary file one. There were six invariants in the 2010 decennial census, all of which held at the census block level, and in addition to the four I mentioned before, we also had total population at the block level, we had voting age population at the block level, and we had occupied housing units at the occupied level. Block level. Those were published as is with no swapping involved, no modification to those counts. They were published as is. You can see here that in 2010 decennial the block level pop count was 100% accurate, in the 2010 demonstration data product that no longer held, and at the census block level we are seeing counts that are hundreds of percentage points off what the value that was published in the 2010 summary files. There is a lot of noise, a lot of inaccuracy in those census block accounts. Constraints are the properties that the differentially private data must have. There are two major constraints that the Census Bureau is currently imposing. One is not negativity. By definition, differential privacy can create negative counts for geographic areas. You can draw a fairly large negative count from the distribution. Once you add that negative count to a value, it could go below zero. Well, you can't logically have negative people in a geographic area, so the Bureau is not going to publish any negative counts, so they have their algorithm take that into account and make sure that there are no negative values published. The second constraint that they will follow is what is called consistency. There is consistency both related to the geographic hierarchy and within data tables. So County populations must sum to the state populations, in the sum of males and females must equal the total population for each geographic unit. Again, under differential privacy, consistency is not maintained, so the Bureau algorithm has to go in and actually enforce the consistency constraint in the non-negativity constraint. Unfortunately, implementing an accounting for these constraints can lead to some bias and some problems in the publicly available data. I will show you that in just a second. I want to talk quickly about the noise injections to give you a sense of what the Laplace distributions look like for these counts. This is the middle case scenario for the 2010 demonstration data that was released from the Census Bureau. For counties down the blocks, for the detailed person or detailed housing unit queries, 50% of all the random drops that were taken from this distribution ranged from -29 229, and 95% of all the random draws came from -125 two positive 125. There are many blocks in the country that are less than 29 people were less than 125 people, and you have a fairly decent chance of pulling a random value that is between 29 and 125. You can fairly quickly double the number of people who are on a block or and of zeroing out again negative on that. There is a lot of noise in the data, and that can lead to inaccuracies in the final outputs. Now I to show you and give you a few examples of what happened when the Census Bureau published that 2010 demonstration data products that was using differential privacy, and what happened when we did analyses comparing the demonstration data to the original 2010 summary file one data that was published about eight years ago. There was a conference held on December 11 and December 12 by the committee on national statistics call the workshop and 2020 census data products. I've included the link here. If you follow the link you will see all the presentations, but the videos of the presentations and the slide decks that people present, and essentially what he data -- invited data users to analyze the data and let the Census Bureau know whether or not the demonstration

data met fitness for use criteria. I'm going to talk about three different results. This first plot here with by a presentation by a person named Matt Spence who actually works at the Census Bureau. When he did is looked at county level differences in total populations. Essentially what Matt did is took the summary file one population, or he took the differentially private total population for each county and subtracted the summary file one count for each of those counties, and then he did a simple box plot after he grouped the counties in the different bins. The key take away here for counties that fall between zero and 7500 people the differentially private counts consistently were higher than the true summary file one counts. That box plot is all above the zero line. For counties that were quite large, 100,000 or more, the differentially private cow was lower than the summary file one counts. What we are seeing is the counts for large areas are systematically lower than they should be in the true data, in the counts for small counties are systematically higher. The magnitude of the count difference is not very high for those large population counties. So the median is like 180 people. On a percentage basis, that is pretty small if you are minimum to the imitator is 100,000 people. If you look at the small counties, looking at a median round about 75 people, that is a larger percentage value of the counties total population count, but the key take away is a systematic bias in the differences by County populations sizes. This mind is from Randy, a professor at UCLA school of public policy. What Randy looked at was the count of American Indians on American Indian reservations in the U.S. He did the same thing as Matt did. He took the differentially private count of loan person from the differentially private data. He subtracted the summary file one population counts and plotted those differences. And then along the X axis is the count of AIA and people. Each dot is a reservation and the X axis is the number of people on that reservation from summary one and 2010. What you can see here is that the differentially private data systematically undercounted the number of American Indians alone persons on a reservation. So in every case the value was lower than, in a differentially private data set, than it was in summary file one from 2010. You might be wondering why that is the case. The alone population in the U.S. is not very large. The reason this happens is because on reservations the alone population is the largest population group on those reservations. And differentially privacy, those large, the large groups, groups of large magnitude, those tend to get pulled down by differential privacy. So because it is the largest population group, it will get pulled down and other race groups, white alone, Asian loan, black alone, those counts would be higher in the differentially private data on reservations than they were on, in the 2010 census. This is a real issue for American Indian groups because often the federal funding they receive for projects on reservations are tied to the number of AI a letter and persons. If we're undercounting the number of people on those reservations, they will get less money from their projects and it will increase or make the inequalities that already exist on AIAN Lowndes, it will make them even worse and help them persist over time because of this differentially private algorithm. The final thing I want to talk about is a presentation that I gave along with Seth Spillman from the University of Colorado. We were looking at what the impacts of differential privacy makes on the geographic hierarchy. So what we did is took the total population from summary file one in the total population from the demonstration data, and we looked at the percent difference between the two counts. What we did is for each geographic level, state, County, tracker block group, we subdivided each of those geographic entities and their decile, we took the total publishing counts for 2010 and simply assigned them to a population decile, decile one is the smallest population count, decile two is the largest population count. We then simply looked at the percentage of units within each decile that differed by 10 percentage points or more. 10 percentage point difference is not, there is no rhyme or reason to

that count. We just assume that you would not want to be off by 10%. The width of those gray bars is the fraction of units that have a 10% discrepancy. So if you are on the geographic hierarchy, the center that hierarchy, they each received a direct allocation, you can see that in general very few units have a 10% discrepancy. That is good. That looks fine. The minute we move off the geographic hierarchy like County subdivisions and places, we see the percentage of the discrepancy skyrockets. So for the first population decile, over 75% of County subdivisions have a 10% discrepancy or more. Even when we move into the fourth population decile, approximately one quarter of all County subdivisions differ by 10% or more between summary file one in the demonstration data. You can see this as well if we look at the fifth population decile for block groups in County subdivisions. They each have the same mean population, 1200 persons, but you can see the block groups, no units differ by 10% or more in that fifth population decile, but for County subdivisions almost 10% of County subdivisions in the U.S. have a discrepancy of 10 percentage points or more. The reason this is an issue is because in the 2010 summary file one total population was invariant. We know that is the true count. I've been talking about this as a difference or discrepancy, in effect, it is error introduced by the differentially private algorithm. And this persists as you subdivide the population into its characteristics. This is the same bar chart, but here we are looking the Hispanic or Latino population. Now we see that even for units like counties or census tracts we are seeing that large fractions of those geographic units within the decile differ by 10 percentage points or more for the Hispanic or Latino population, even for large counties in the 6 decile, over 10% of those counties have Hispanic or Latino populations that differ by 10% or more. Finally, for my democracy friends in the crowds, we look to age and sex does divisions, and again, this is by Matt Spence. Lyon County Minnesota is the meeting County by population in the U.S., he took the data and made essentially population pyramids for Lyon County. The pink is female, the blue is male. The height of the blue, the light blue bars are the count of persons by age, bisects in the differentially private data. The black see-through outlines are the counts of people by age and sex from the SF 120 10 data. As you can see, the differentially private counter very spiky. For women age 23 the count is 500 people in the differentially private data and 250 in the summary file one. So they're estimating there are two times as many women age 23 and the differentially private data than they counted in the summary file one data. Even if we aggregate those 25 year age Vince, we still see great discrepancy. So if you look at the women age 75 to 79, the differentially private data are indicating account of approximately 200 and the summary file one had a count of approximately 400. So if you're doing things like trying to plan for nursing home capacity or compute COVID-19 rates in Lyon County, and their using the differentially private data as your denominator, you are going to get a much different disease incidence rate than you would get if you use the summary file one data from 2010. I'm going to skip this next slide because it just reinforces the point that the 2010 demonstration data was not really fit for use for A lot of use cases. Not going to talk a little bit about the less publicly available data that will come out of this differentially private algorithm. Here I'm going to direct you. I'm not going to show this in detail, but there is a crosswalk that the Bureau has put up discussing their census data products for 2020 and how they differ from 2010. What I want to show you is that there is going to be a lot less data published. The slide is for a set of tables, the P 35 tables, this was available at the census block level in the 2010 decennial census. In 2020 they are proposing to not publish this data table. Going from publishing it in the 2010 data to not publishing it in the 2020 data at the block level. Here we see the family type by age of children by raising 2010. These tables were also published at the block level. In 2020 they are proposing to publish it only

at the county level. They are really coarsening the geographic detail available in a number of these data sets. So the users who are used to having detailed data on family type at the blocker track level would no longer have that available. Why are they doing this? it protects the privacy of the respondents and the data is inaccurate at levels below County that they just would not publish the information. In general, race was available at the block level in 2010, race will be available at the block level in 2020, although detailed race and ethnicity may not be available at the blocker track level. Households with counts at the block level, 2020 the proposed geographic level is the County. Families, we got 2010 data at the block level, in 2020 we don't know yet what it's going to be, but I would expect potentially track or County is the finest unit available. Group quarters were available at the blocker track level in 2010, in 2020 they are proposing to make county or state the smallest unit we can get the data for. A lot of this is moving, you know, the Bureau is continually making improvements are trying to make improvements to their algorithm. This may change, but based on the crosswalks they have put out, we are looking at a lot less data that would be available. Finally, I will and with a short discussion of less consistency among products. The Bureau has identified groups of products for the 2020 census. The first group, what they're calling group 1, 94 data, the Democratic and housing characteristics file, summary file one, the profile and the congressional district data set. These will all be consistent within one another. So the counselor from the PL 94 171, if you look up the equivalent count in the demographic and housing characteristics file, you would find the same values. The track that I live in may have 3000 people in it, it would be the same in both of those products. The group 2 products which they are currently targeting for a detailed race and at the city file, American Indian and Alaska native summary file in person household join files counts from those would not be consistent with the data that were published in the group 1 files. So if you have a count of American Indians by County in the group to file, that value would not be the same as the American Indian count from the demographic and housing characteristics file. This is because they are going to be using a separate algorithm to produce the group 2 files for a number of technical reasons, and because the noise injection is random, the noise values that get infused into these group 2 products will not be the same as the group 1 products. So you can't necessarily blend data from group 2 products in group 1 products to do analysis. So what is coming next? the Census Bureau is hard at work modifying its algorithm to try to fix the issues found in the 2010 demonstration data. They are soliciting feedback. They have gotten a lot of feedback, but they're working under an incredibly short timeline. The announce this in 2018, in the first data did not come out less than a year from now. Truthfully, I'm not sure the census has time to address all of the issues and create the usable data that all the users have come to expect. The one thing that the Bureau has said thus far is that they do not plan to release another demonstration data set for us to analyze to see how algorithm improvements are impacting the quality of the data. I, along with others, are arguing we need another demonstration data set to make sure that the Bureau is going to be creating usable data. That is the last line I have. I want to point you to a few resources. The Census Bureau maintains the disclosure avoidance page, that is the first link. They talk about the enhancements they are making to their algorithm. We have been maintaining a differential privacy page where you can find links to a lot more talks that I have given on this topic. You can find more technical talks there. there will be some papers that I have written. You can also find an editorial I published in February with a staff writer at the New York Times talking about the impact that their privacy algorithm they have on counts, particularly in rural areas. It is a fairly light introduction to the algorithm and the type of impact

it might have. That is the end. Joe, I don't know if you want to hop back on and we can start handling questions.

Thank you David. Excellent webinar. Really appreciate it. let's check the questions here. Tonya asks, what is the best epsilon number for getting the most money with the most privacy protection? I think I am a bit confused, although this session is good.

We don't know yet. When I first started, I mentioned that the Bureau had identified, re-identified 17% of the population in the original summary file one data from 2010. The demonstration data that they published used a privacy laws budget of 6.0, and when they did their re-identification attack on that, they still identified 6%. They were able to identify 6% of the population. They went from 17% down to 6% while impacting the usability of the data. The Bureau is currently doing an experiment where they are running their algorithm with 25 different values of epsilon from 2 up to 100. And they are trying to see at what level of epsilon do they need to get to to get data that looks very similar to what they published in 2010. Once they establish that, they will then do a re-identification attack and see how many people that they can identify. That is an empirical question. You need to do the experiments to figure that out. 6.0 is higher than I thought they were going to go with, but after seeing the results of the data, I think they should think about doubling or tripling that value.

Thank you. What was the epsilon value for the 2010 demonstration data?

---

that was sent to 6.0, and that 6.0 was then allocated for.0 was allocated to the person tables and to .0 was allocated to the housing unit tables.

Okay. Michael asks, when did the Census Bureau start suppressing and swapping the data of high income earners, and why did they adapt this policy?

---

they started top coding income in the published data in 1940 when the income was first asked on the census. They were top coding the data at that time already so the individuals for high income cannot be identified in the data. If they undo that, you can actually find people like Bill Gates are high paid individuals in the data. And they did it to protect the privacy of respondents.

Barber asked, how will the change methodology affect comparability of published data 2010 220 20?

That is an excellent question Barbara. I actually have not -- that is my next thing to do. To look into that comparability. I think for geographic units that have less than 1000 people, which is a lot of cities and county subdivision and entities, I think it will really impact -- you could measure differences, changes, 10% to 30% between those two years, and that difference is solely due to this new mechanism. And so, I need to go use the 2010 data and do an experiment with 2000 and see how the change over time metrics happen. But I think will have a major impact on that. That is actually -- one of the things that I'm trying to get ready for it how I explained to users that the changes that they are observing in the data may be solely due to the algorithm and not due to actual population change.

Thank you. Carol asks, any suggestions for how folks will be able to use census data for local planning and mapping, the laws of block group data is huge.

Yeah, I think the local planning agencies are really in a tough spot. I think Census Bureau has heard loud and clear that planning agencies are one of the most, one of the biggest users of these data sets and that they need to do as much as they can to produce usable data for them. At this point in time, what we have been say to people is respond to Census Bureau inquiries, start talking and having plenty agencies talk to congressional leaders, talk to their senators about the potential impact this is going to have. I don't know how much leverage people have, but it is definitely going to be an issue. I think the track level data I have seen looks fairly reasonable for lots of applications, so I think you're going to have to start moving to that geographic unit, and I totally a group -- totally agree.

Michael says, slide 28, what any of this affect voting redistricting or are the differences not enough to count in the aggregate?

For things like U.S. congressional seats, where the population is 800,000 people for the House of Representatives see, you are right, in the aggregate the differences net out and average out to be near zero. I think it is going to be an issue for any districts that are on the edge of majority minority. Any kind of voting rights act types of districts. This could be affected. I think state legislature districts could be much more problematic because state legislative districts tend to have smaller population counts. I fully expect this to get litigated after the data comes out in the first data set of districts are drawn, and I would not be surprised if all, if we see Supreme Court case in 2022 related to this.

Thank you. Mark asks, will margin of error still be included in the tables and would it be meaningful?

---

that is a great question. The decennial data products have never had margins of error. The 100 Bissau count tables are always viewed as a census. So there are no error bounds around those counts. They are seen as -- [Indiscernible]. One thing we've been pushing census for is to generate some kind of error bound for these differentially private counts so we can get a sense of the quality of a count for a given unit. The -- as of today, the Bureau has not committed to creating those. I think they well. The one drawback is that the creation of these error bounds uses up privacy. For them to create these they actually have to use privacy budget to do that. It is trade-off, we would have to give something up. To get that we would have to give something up in the rest of the privacy laws budget, and so, I think that we need these. Even though we know a lot of data users don't necessarily know how to interpret them, it would help a lot to give people a sense of the quality, particularly for those small units and for people to know that these data are just not reliable, kind of like we are getting used to with the American community survey.

Thank you. Michael says thank you. I currently serve in a minority district in the deep South.

This is a major change. A massive change to what the Bureau has done in the past. They are trying to do this very quickly under incredible resource and time constraints with brand new technology. This will be the largest data set ever created using differential privacy and they are

really experimenting with technology on a cornerstone of American democracy. It makes me nervous that that is what is going on.

Carol makes the comment, yes, please to error bounds.

The National Academy of Sciences committee is working with the census on this. In the error bounds and we frequently argued for that.

Another comment from Barbara, a question, are there viable alternative recommendations on how to achieve a proper balance between privacy and accuracy?

Differential privacy is the only way to provide some mathematical guarantees on that balance between privacy and accuracy. All other types of disclosure control don't provide any mathematical proof. The Bureau have decided that they need that proof in order to abide title 13 of the code regarding privacy. We think it is a misinterpretation of title 13, but that is what they are going with. They have said to me that they will go to jail if they do not do differential privacy because they would be violating title 13. Any other recommendation we made to them has fallen onto fears, onto deaf years, and there is really no turning back.

Thank you. Keep the questions coming. I'm going to go into my wrap-up comments. But we have time for more questions. Keep them coming in. Cory is going to put up a survey in the chat box. Please give that a look. First I would like to thank David for a fantastic webinar. A lot of great information. Really appreciate it. I would also like to thank Corey for his great work today for tech support, keeping everything running smoothly and many thanks to the audience. I hope you enjoyed the webinars much as we did today. Don't forget the upcoming webinars. We have seven more webinar scheduled for June. A lot of webinars for the COVID-19 shutdowns. The next webinar is Thursday June 11 titled resources across the generations, government resources to cover everyone from the greatest generation to generation Z. You will receive notice of all of our upcoming webinars when they are announced and you can sign up for our events on FDLP.gov. From the FDLP Academy webpage which linked to an index section at the bottom of the homepage, you can view a calendar of upcoming webinars and other events, pass webinars from her webinar archive and links to web form to volunteer to prevent the FDLP Academy webinar. At the Cory will also put in information about our ribbon or repository. We are recording just about all of our webinars. This will be available tomorrow, and also an article for my former colleague about the FDLP Academy and all the things we do, webinars and other presentations. Let's see if we have any other questions. A lot of shout outs David. Thank you so much. Everyone is saying great webinar. Thanking you.

No problem.

Okay. Any last questions for David? we're a few minutes over, but that's okay if you have some questions. Cory, did you put the other information the chat box?

Actually, I don't have it.

Okay. Hold on. I will put it in myself. Hold on audience. Sorry I did not send that. You've probably seen this before. I have sent it before. Give that good information a look when you have a chance. Let's see. Here we go. Barbara has a comment or question. If you can say which is likely to be worse, accuracy in the 2020 decennial census or accuracy in the ACS?

I think the ACS is probably still going to be less accurate just because it is a sample. But I think for certain counts of people in the decennial they are going to be as inaccurate as things in the ACS. I think this is the type of experiment or research or analysis that the Census Bureau should've been doing. I would like to say that if this was 2015 I would be super excited for five years of awesome research into differential privacy in the census, but it is so quick. I think in general the accuracy of decennial should be a little bit better than ACS, but no one has really done any analysis yet.

Thank you David. Last questions for David? this was a great webinar. It really appreciate it. looks like you've covered everything David. I would like to thank you one last time. Thank you Corey. Thank you audience., Back to the FDLP Academy Thursday and the rest of our webinars we have in June and throughout the summer going forward. Have a great rest of the day. Thank you. [Event Concluded]