Implementing Differential Privacy for the 2020 Census

Simson L. Garfinkel
U.S. Census Bureau

Usenix Enigma
February 1-3, 2021
Oakland, CA
https://www.usenix.org/conference/enigma2021



The views in this presentation are those of the author, and not those of the U.S. Census Bureau.

Shape your future START HERE >

United States Census 2020

Abstract

Differential Privacy was invented in 2006 to protect the privacy of people who respond to a national census. The U.S. 2020 Census of Population and Housing will mark the first time that differential privacy will be used for its existential purpose.

Bringing leading-edge privacy technology from the lab into practice required significant scientific and technical development, and presented organizational challenges to one of the world's largest statistical organization.

In three years the Census Bureau assembled a team, developed a reference implementation, transitioned that implementation to Amazon Web Services, redesigned the implementation's framework, created a system that made it easy to perform experiments, performed an end-to-end test, used the implementation to re-release data from the 2010 Census, developed new algorithms to address the data quality concerns of stakeholders, and released multiple reference implementations of the code base.

Acknowledgments

This presentation incorporates work by:

John Abowd (Chief Scientist)

Dan Kifer (Scientific Lead)

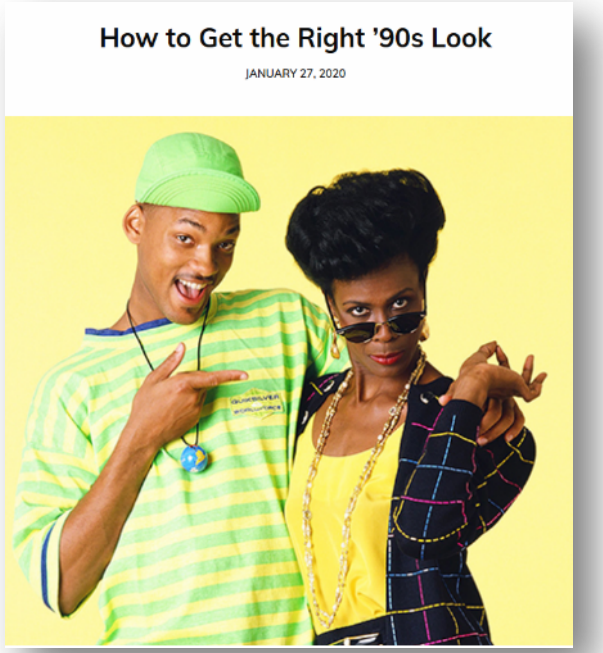
Simson Garfinkel (Senior Computer Scientist for Confidentiality and Data Access)

Rob Sienkiewicz (Chief, Center for Enterprise Dissemination)

Tamara Adams, Robert Ashmead, Stephen Clark, Craig Corl, Aref Dajani, Jason Devine, Nathan Goldschlag, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Brett Moran, Edward Porter, Sarah Powazek, Anne Ross, Ian Schmutte, William Sexton, Lars Vilhuber, and Pavel Zhuralev.

Differential Privacy was invented in 2006 (14 years ago)

Modern Public Key Cryptography was invented between 1976 and 1978.
Differential Privacy's TRL *today* is where PKI was in the 1990s.



Will Smith and Janet Hubert in
The Fresh Prince of Bel-Air (NBC)
https://www.centralcasting.com/how-get-right-90s-look/

1991 – PGP 1.0 Released (it was not secure)

1992 – 14 years after the invention of PKI

1994 – Netscape Navigator 1.0 released with SSL

1995 – SSH invented

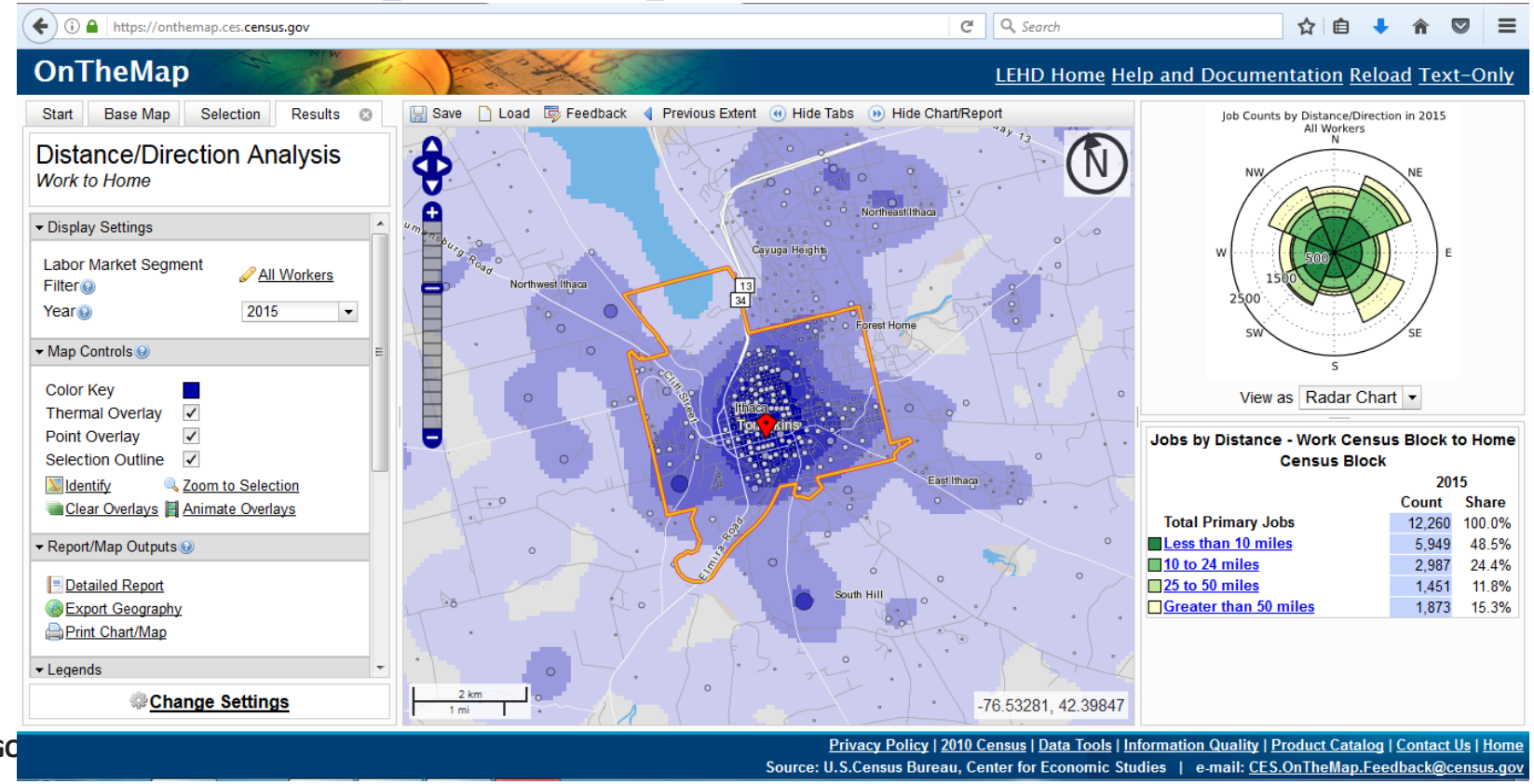
1996 – PGP 3.0, first commercial, reasonably secure version of PGP

– SSL 3.0

…

2011 – Electronic Frontier Foundation (EFF) "HTTPS Now" Campaign and "HTTPS Everywhere (+34 years)

The Census Bureau deployed the first DP implementation in the world for OnTheMap (2008)



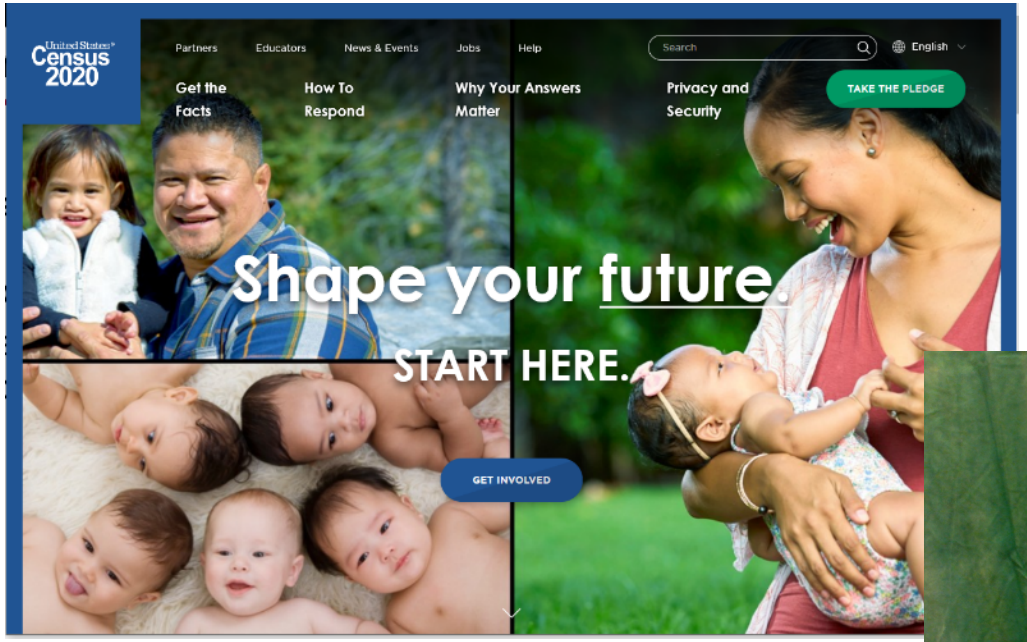
(2018 screenshot)

2020CENSUS.GO

Shape your future START HERE >

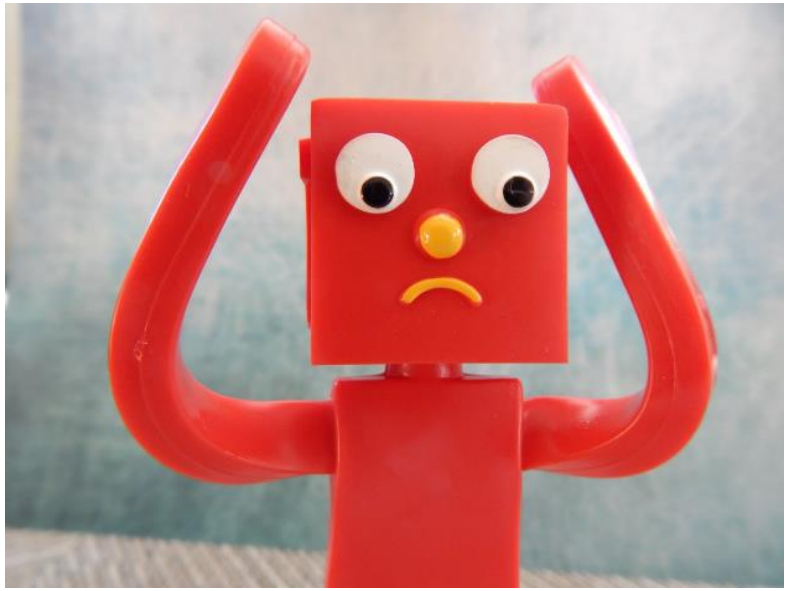
United States Census 2020

We solved many technical challenges deploying DP for the 2020 Census!





We still haven't met user expectations.



https://pxhere.com/en/photo/772531

Data scientists are being trained to work with microdata, fully cleaned data, and internally consistent tabulations.

They would get better results if they worked with the "noisy measurements."

Data users expect consistent data releases

Making multiple noisy measurements consistent introduces bias and degrades accuracy.

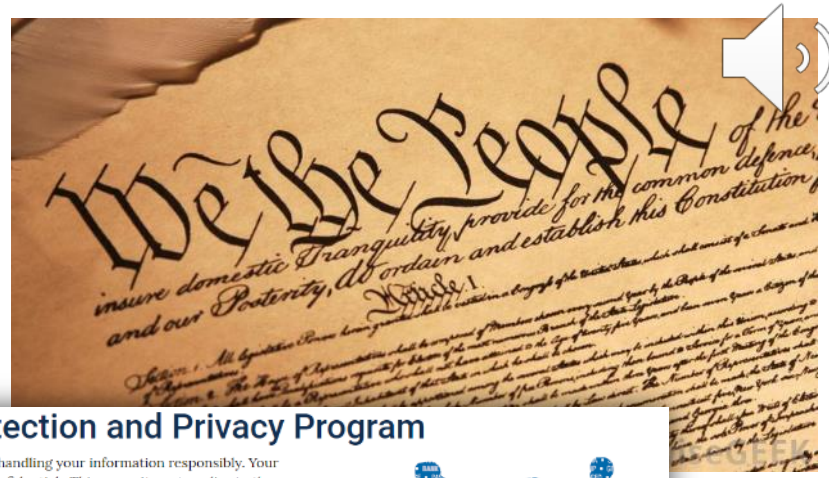
Some users liken synthetic data to "fake data" (c.f. "fake news")

This is not an accurate characterization.

It is not clear what "quality assurance" and "data exploration" looks like in a DP framework.

We would like to have tools for exploring data without impacting privacy loss budget.

The Decennial Census!



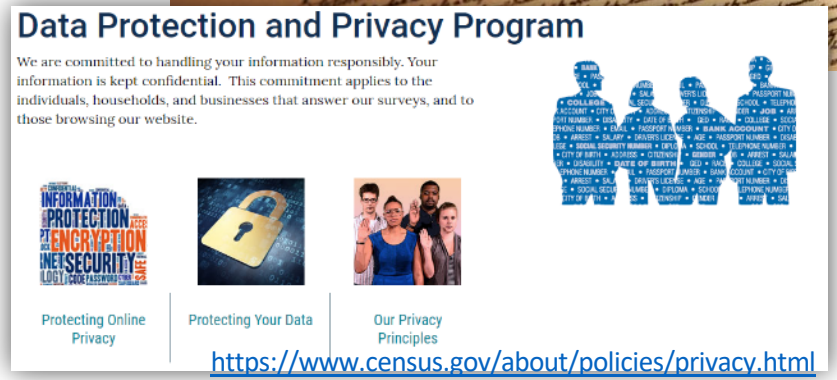
Mandated by Article I of the US constitution

Data used for:

Apportioning the House of Representatives (U.S. Constitution)

Distributing $675 billion in FY2015 (internal estimate)

Distributing $1.5T in FY2020 (external estimate)



Title 13 Section 9 mandates data confidentiality

(a) Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, or local government census liaison, may, except as provided in section 8 or 16 or chapter 10 of this title or section 210 of the Departments of Commerce, Justice, and State, the Judiciary, and Related Agencies Appropriations Act, 1998 or section 2(f) of the Census of Agriculture Act of 1997—

(1)use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or

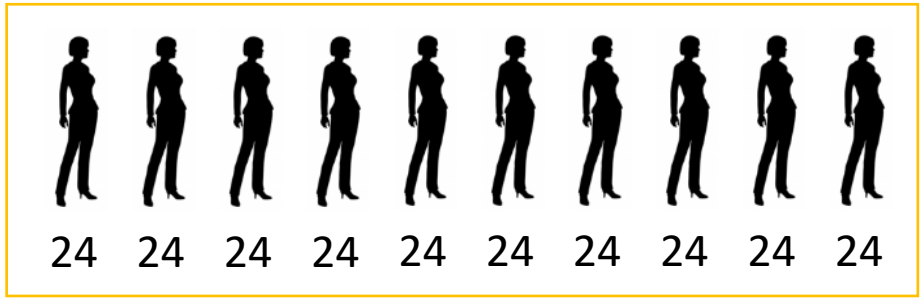
(2)make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or

(3)permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports.

No department, bureau, agency, officer, or employee of the Government, except the Secretary in carrying out the purposes of this title, shall require, for any reason, copies of census reports which have been retained by any such establishment or individual. Copies of census reports which have been so retained shall be immune from legal process, and shall not, without the consent of the individual or establishment concerned, be admitted as evidence or used for any purpose in any action, suit, or other judicial or administrative proceeding.

Title 13 Section 9 limits what the Census Bureau can publish.

Consider a block with 10 people on it, all the same sex and age:



The Census Bureau cannot publish an exact statistic, e.g. that there are 10 women on the block, all with the same age without offering some demonstration that the statistic protects confidentiality. If it did and you knew that Alison lived on the block:

You would know that:

Alison is a woman

Alison is 24



Alison

We use the term *disclosure avoidance* to describe the process of protecting confidential data, because it is designed to prevent an improper disclosure.

We had strong reasons for moving to DP



I **alled "swapping" with secret parameters as our privacy mechanism.**

ency and accountability.

eflecting the move to open source in the community.

I **s/accuracy trade-off**

losure Avoidance Research" had set the swap rate.

With DP, we made setting the trade-off a **policy decision** made by **policy makers**.



DP privacy is "future-proof"

Our previous privacy mechanism made assumptions about the availability of external data and methods

DP's mathematical basis assumes an attacker with unbounded computational power and unlimited access to external information

Each of these "advantages" ended up having complications within our organization!

We had strong reasons for moving to DP

Previously we used a method called "swapping" with secret parameters as our privacy mechanism.

- Many people within the Census Bureau thought secrecy should remain an important part of the privacy protection mechanism.
- Many people were concerned about publishing the differential privacy source code.

DP let us control the privacy-loss/accuracy trade-off

- We needed to develop tools to help policy makers make this decision.
- We needed to provide policy makers with *context* and *training* to understand the decision they were making.
- We needed to design the algorithm so that accuracy would be *where it was needed* so that privacy loss could be minimized.

DP privacy is "future-proof"

- DP is "future-proof" because its privacy definition and guarantees are *relative*, not absolute.
- This required significant explanations and additional training.

We expected that there would be organizational barriers that needed to be overcome.

The need to develop and tune the DP algorithm was unlike anything Decennial Response Processing had done before….





… and that was before the pandemic!

By 2017, we thought we had a good understanding of how DP would fit.



Enumerators (1790-)
Self-response surveys (1960-)
Administrative data (2020-)

Swapping started in 1990

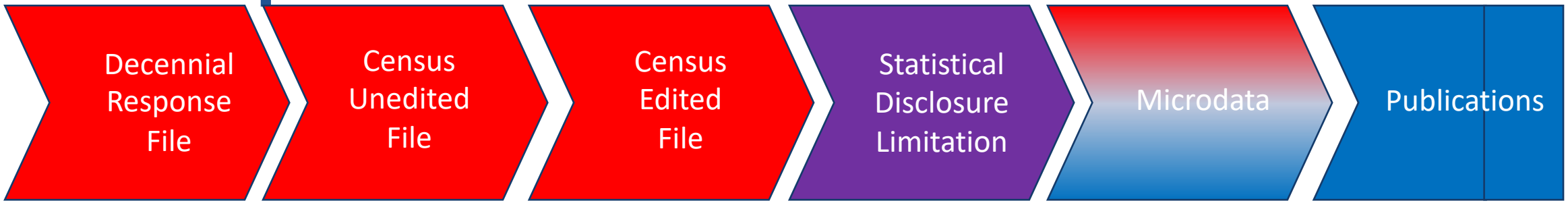
"Hundred Percent File" (HDF) Confidential 1990, 2000 and 2010

Decennial Response File → Census Unedited File → Census Edited File → Statistical Disclosure Limitation → Microdata → Publications

Move to DP in 2020

"Microdata Detail File" (MDF) Privacy-protected 2020

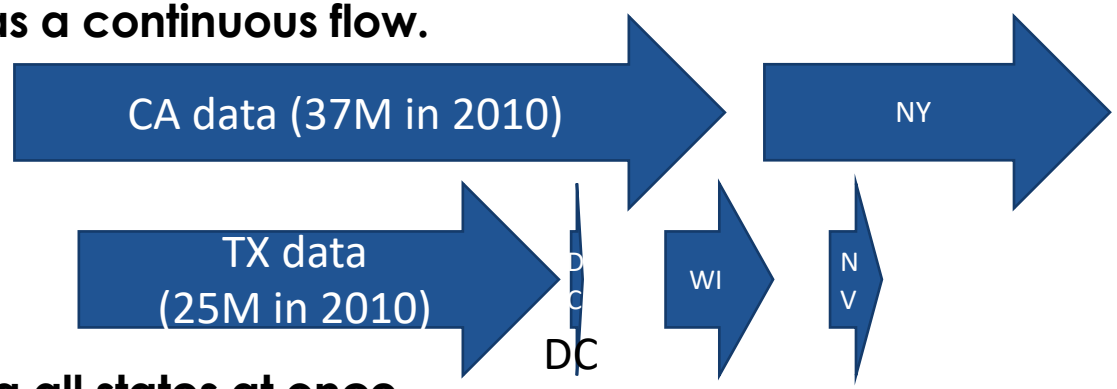
Different groups at USCB had different assumptions about the data flow.



Decennial Response Processing traditionally was a continuous flow.

Each state was processed independently.

- This allowed for pipelining.
- Each system was developed with synthetic data.
- Each system was developed and tested small, then scaled up.



The "TopDown" mechanism required computing all states at once

- Created a bottleneck in decennial response processing (it "stalled the pipeline") because of the US-to-US solve
- Required significantly more computing power than traditional tabulations because of the 800,000 blockgroup-to-block solves.

Shape

United States Census

The DP system had to be developed with real (Title 13-protected) data.

Most systems for the 2020 Census were developed using simulated data.

"Simulated data" not based on protected respondent data.

Significantly lowers the cost of working with the data in a government environment.

Little DP work in the academic literature had been done with complex confidential data.

We needed to develop algorithms that would work with data as complex as the United States!

Clusters by age and race ● Multi-racial households ● Same-sex unions ● Small Group quarters ● Large group quarters



2020CENSUS.GOV









Shape your future
START HERE >

United States Census 2020

Understanding privacy/accuracy trade-off required thousands of runs

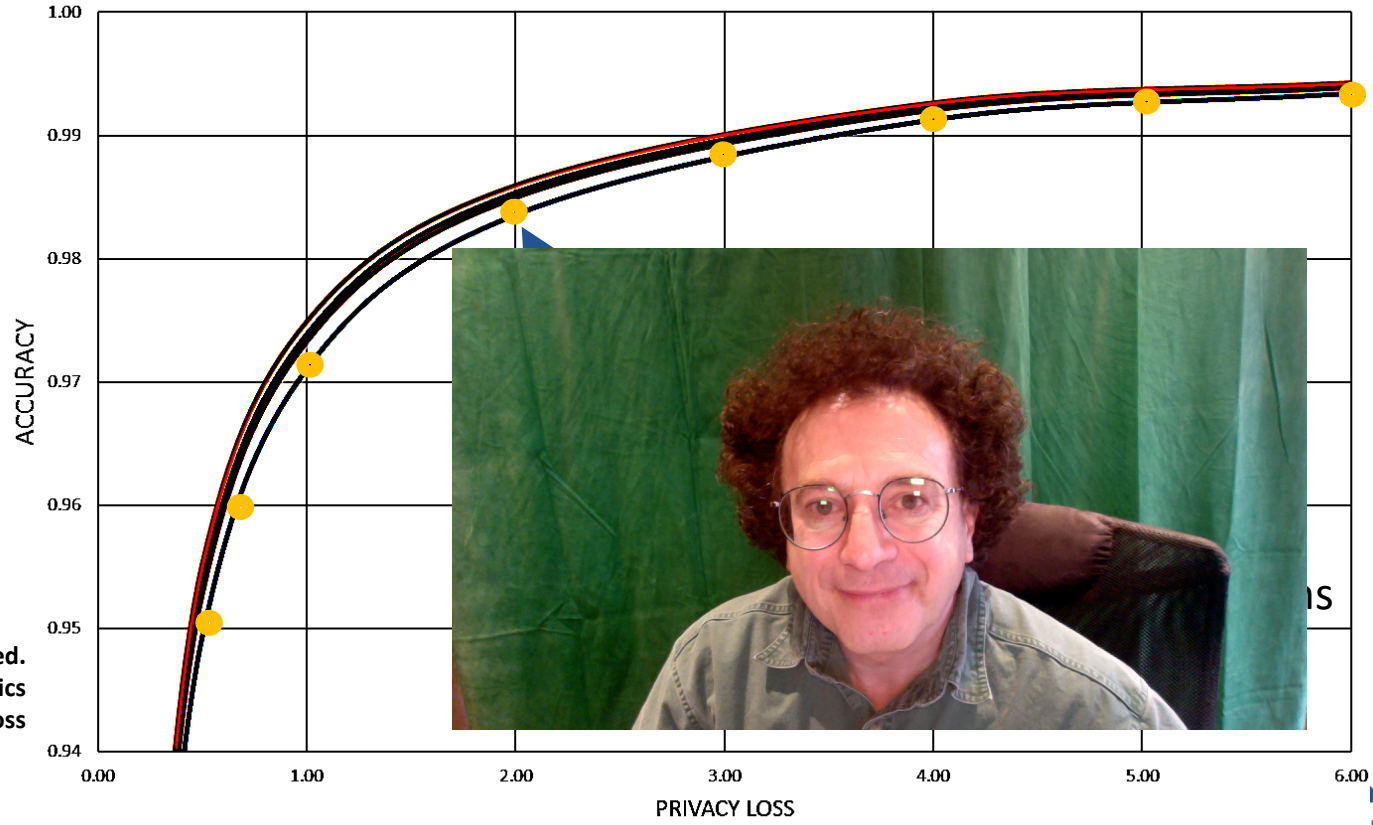
Making this graph requires:

Runs at epsilon 0.5, 1.0, 2.0, 3.0, 4.0
- To get data for graph

5-10 runs per epsilon value
- because each run has noise!

~ 100 runs of algorithm

This graph also requires an accuracy definition!



So how did we do it?



https://pxhere.com/en/photo/457316

Shape your future START HERE >

United States Census 2020

Census Bureau leadership was 100% behind the move to Differential Privacy



John Thompson
Director
2013-2017



Ron Jarmin
Performing non-exclusive functions and duties of director 2017-2019
Deputy Director and Chief Operating Officer 2019-



John Abowd
Chief Scientist
2016-

Team and Reference Implementation

2016-2017 – Penn State Professor Daniel Kifer takes sabbatical at US Census Bureau to build initial implementation of top-down algorithm



Decides to use Spark for "big data" manipulation, Gurobi Optimizer for the TopDown implementation.

Team members Robert Ashmead and Phil Leclerc from Research & Methodology Directorate

2017 – Expand team with additional government employees & contractors

Simson Garfinkel joins Census Bureau. Creates "DAS Framework."

- Framework is a "system for building disclosure avoidance systems"
- Plugin-ins for Reader, Engine, Validator, Writer.
- Driven by *hierarchical configuration file* to make it easy to experiment with parameters and new code.

2018 – End-to-End Test (Providence, RI)

We put source code release in specs as an official data product

Because source code requires "Census Edited File" (confidential data) we also created a reader that could load the data from 1940 Census (which is now publicly released.)

We used AWS Elastic Map Reduce for development.

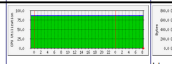
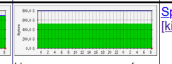
Original development was an “on-prem” Linux Cluster.

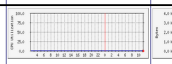
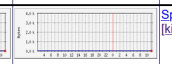
Developing in AWS with Title 13 data required policy changes.

Over time, the DAS Team took over management of the EMR clusters.

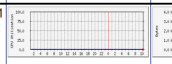
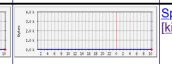
We built as “dashboard” to track:

- Running Clusters (we were using EMR in non-standard ways)
- Each run of the TopDown algorithm.

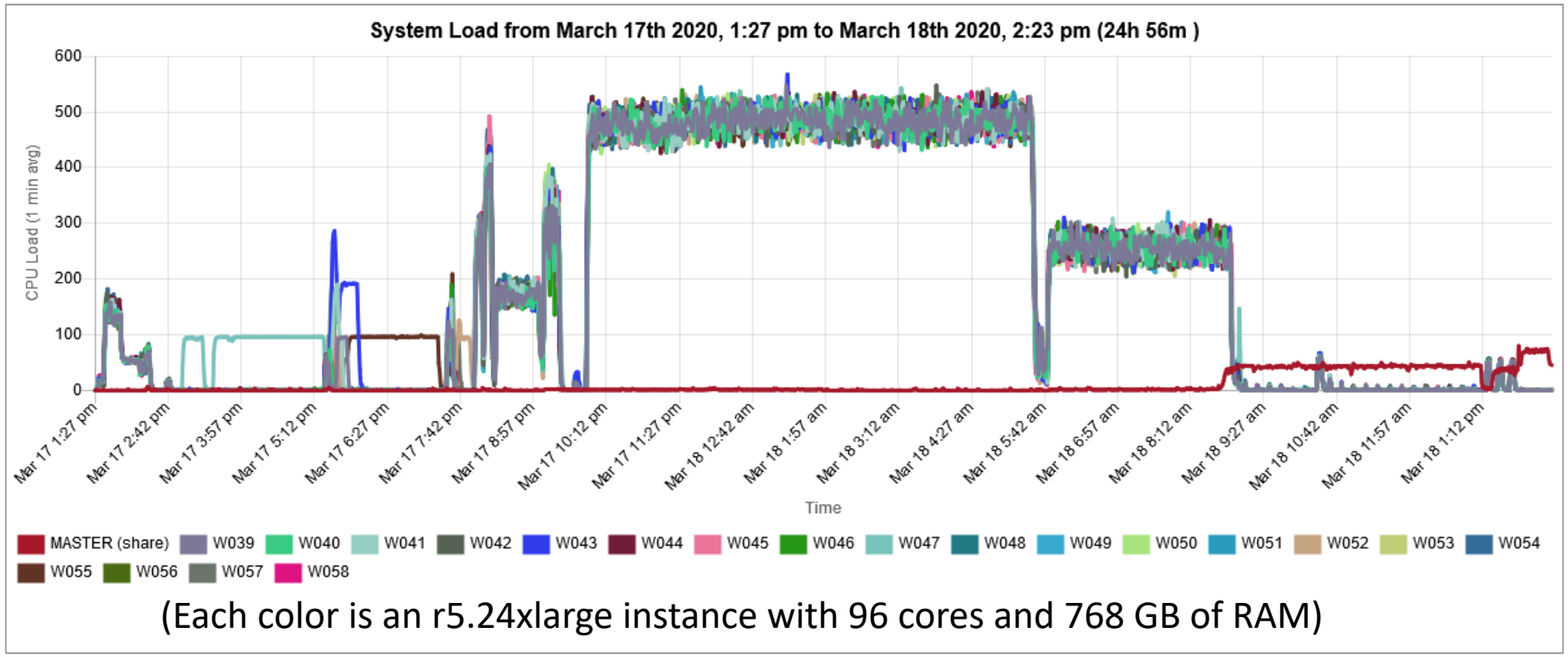
Name	ClusterId	Type	ipaddr	Load	freeGB	swap_free	JBID: #proc	CPU Load (day)	RAM Load (day)	History SparkUI
cloff show	j-308HU6D0TOH8W 2020-11-16 15:53:59 Workers: 4 Total RAM: 3072 Total vCPU: 384 awsgui cluster logs	r5.24xlarge emr-5.25.0 REL 009	10.252.44.100 instance log	1.22	722	0	heine008: 61 garfi303: 2 hodge357: 1 beggs005: 3		 blue: max memory; green: free memory.	Spark [kill cluster]
	current mission									
	last das_log	2020-11-16 14:18:26: HOLIDAY_CANCEL finished								
	dev chat	15 h /mnt/gits/das-vm-config/bin/backup-homedirs 42039 finished								

Name	ClusterId	Type	ipaddr	Load	freeGB	swap_free	JBID: #proc	CPU Load (day)	RAM Load (day)	History SparkUI
curse show	j-2O0MB4W7A2CTP 2020-11-16 15:54:06 Workers: 2 Total RAM: 1536 Total vCPU: 192 awsgui cluster logs	r5.24xlarge emr-5.25.0 REL 010	10.252.47.12 instance log	0.41	734	0			 blue: max memory; green: free memory.	Spark [kill cluster]
	current mission									
	last das_log	*none*								
	dev chat	*click to edit*								

Name	ClusterId	Type	ipaddr	Load	freeGB	swap_free	JBID: #proc	CPU Load (day)	RAM Load (day)	History SparkUI
notan show	j-3LMLR778GJ51B 2020-11-16 15:54:17 Workers: 8 Total RAM: 6144 Total vCPU: 768 awsgui cluster logs	r5.24xlarge emr-5.25.0 REL 010	10.252.47.211 instance log	0.12	733	0			 blue: max memory; green: free memory.	Spark [kill cluster]
	current mission									
	last das_log	*none*								
	dev chat	*click to edit*								

Name	ClusterId	Type	ipaddr	Load	freeGB	swap_free	JBID: #proc	CPU Load (day)	RAM Load (day)	History SparkUI
wimps show	j-382EMZ2DC66MR 2020-11-16 15:54:15 Workers: 2 Total RAM: 1536 Total vCPU: 192 awsgui cluster logs	r5.24xlarge emr-5.25.0 REL 010	10.252.46.237 instance log	1.71	732	0	wang0383: 1 cumin002: 2		 blue: max memory; green: free memory.	Spark [kill cluster]
	current mission									
	last das_log	2020-11-16 15:41:35: loading table UnitData								
	dev chat	16 m MITRE test run in progress								

CPU load of 21 AWS Instances during an execution of the TopDown algorithm:



(Each color is an r5.24xlarge instance with 96 cores and 768 GB of RAM)

ipe
your future
START HERE >

United States Census 2020

We re-published the 2010 data using the 2020 TopDown algorithm.

October 29, 2019 Data and Software Release

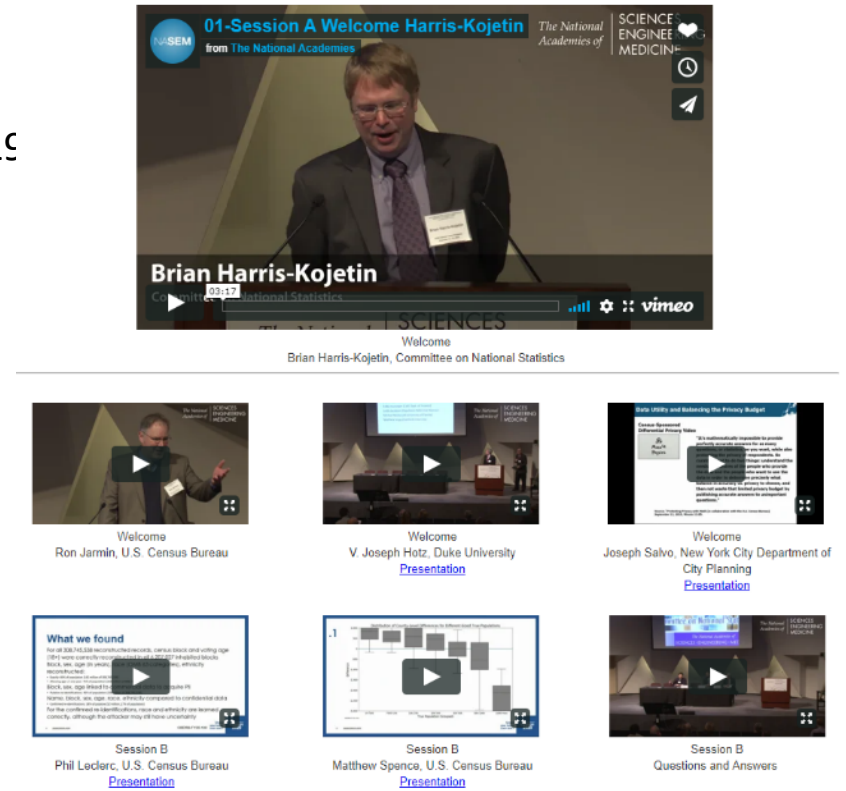
2010 Census Edited File as processed by the 2020 DAS (as of October 2019

Release of the 2020 DAS Source Code (as of October 2019)

https://github.com/uscensusbureau/census2020-das-2010ddp/

December 11-12, 2019 Workshop
National Academies of Science

https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518



Shape your future
START HERE >

United States Census 2020

“2010 Demonstration Data Products” were not well received.



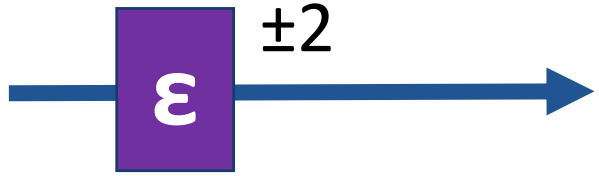
https://www.nytimes.com/interactive/2020/02/06/opinion/census-algorithm-privacy.html

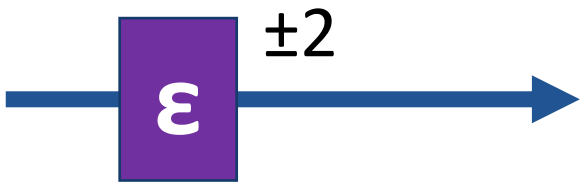
At the root of our acceptance problem: Non-negativity requirements create systematic bias and increase the size of outlier errors





Photo by mohamed hassan form PxHere



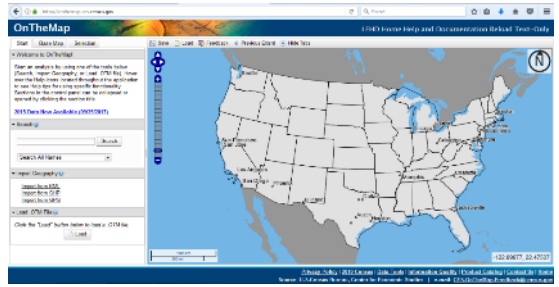




Solution is to perform analysis with the "noisy measurements"

Provide users with negative numbers and let the users create their own geographies and aggregations.

In summary:
The 2020 Census DP Timeline



2006 – Differential Privacy Invented

2008 – Census Bureau adopts DP for OnTheMap product

2010 – 2010 Census protected using data swapping (like 1990 & 2000)

2015 – Cornell University professor John Abowd named Associate Director of Research & Methodology and Chief Scientist of US Census Bureau (to start in June 2016)

2016—2017 Pennsylvania State University professor Daniel Kifer takes a year's sabbatical at Census Bureau. Goal is to develop algorithm spends year at Census

2017– Census Bureau announces to its scientific Advisory Committee that it will use DP for the 2020 Census.

2018 – Census Bureau protects test data from the 2018 End-to-End Census Test with DP

2019 – Census Bureau re-releases 2010 Census using DP for the Committee on National Statistics (CNSTAT) Workshop.

2020 – Census Bureau uses DP to protect redistricting data and other special tabulations.

The Most Technical 2020 Publications

Most recent public source code:

https://github.com/uscensusbureau/census2020-das-2010ddp

System Design Specification

https://github.com/uscensusbureau/census2020-das-2010ddp/blob/master/doc/2010-Demonstration-Data-Products-Disclosure-Avoidance-System-Design-Specification%20FINAL.pdf

Scientific paper describing mechanism:

https://github.com/uscensusbureau/census2020-das-2010ddp/blob/master/doc/20191020_1843_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf

Semantics Paper

"Effective Privacy After Adjusting for Invariants with Applications to the 2020 Census"

https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0941_Effective_Privacy_after_Adjusting_for_Constraints__With_applications_to_the_2020_Census.pdf

More Background on the 2020 Census Disclosure Avoidance System



September 14, 2017 CSAC (overall design) https://www2.census.gov/cac/sac/meetings/2017-09/garfinkel-modernizing-disclosure-avoidance.pdf?#

August, 2018 KDD'18 (top-down v. block-by-block) https://digitalcommons.ilr.cornell.edu/ldi/49/

October, 2018 WPES (implementation issues) https://arxiv.org/abs/1809.02201

October, 2018 *ACMQueue* (understanding database reconstruction) https://digitalcommons.ilr.cornell.edu/ldi/50/ or https://queue.acm.org/detail.cfm?id=3295691

December 6, 2018 CSAC (detailed discussion of algorithms and choices) https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf?#

April 15, 2019 Code base and documentation for the 2018 End-to-End Census Test (E2E) version of the 2020 Disclosure Avoidance System https://github.com/uscensusbureau/census2020-das-e2e

June 6, 2019 Blog explaining how to use the code base with the 1940 Census public data from IPUMS https://www.census.gov/newsroom/blogs/research-matters/2019/06/disclosure_avoidance.html

June 11, 2019 Keynote address "The U.S. Census Bureau Tries to Be a Good Data Steward for the 21st Century" ICML 2019 abstract, video

June 29-31, 2019 Joint Statistical Meetings Census Bureau electronic press kit
(See talks by Abowd, Ashmead, Garfinkel, Leclerc, Sexton, and others)

Thank you!

