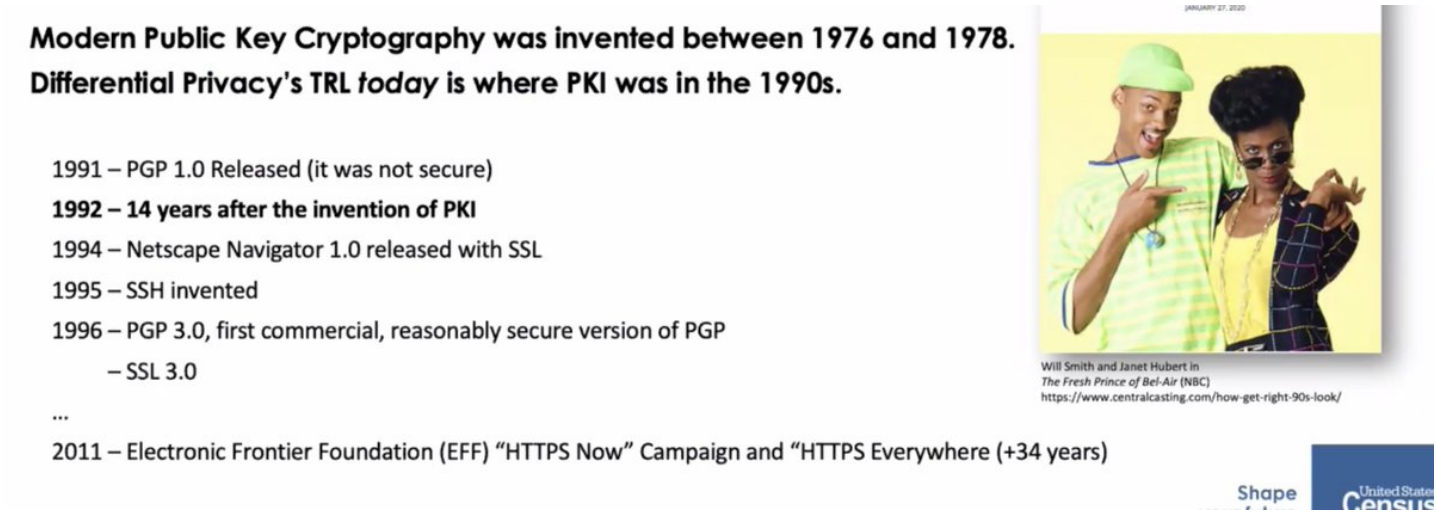
Twitter Thread by Lea Kissner

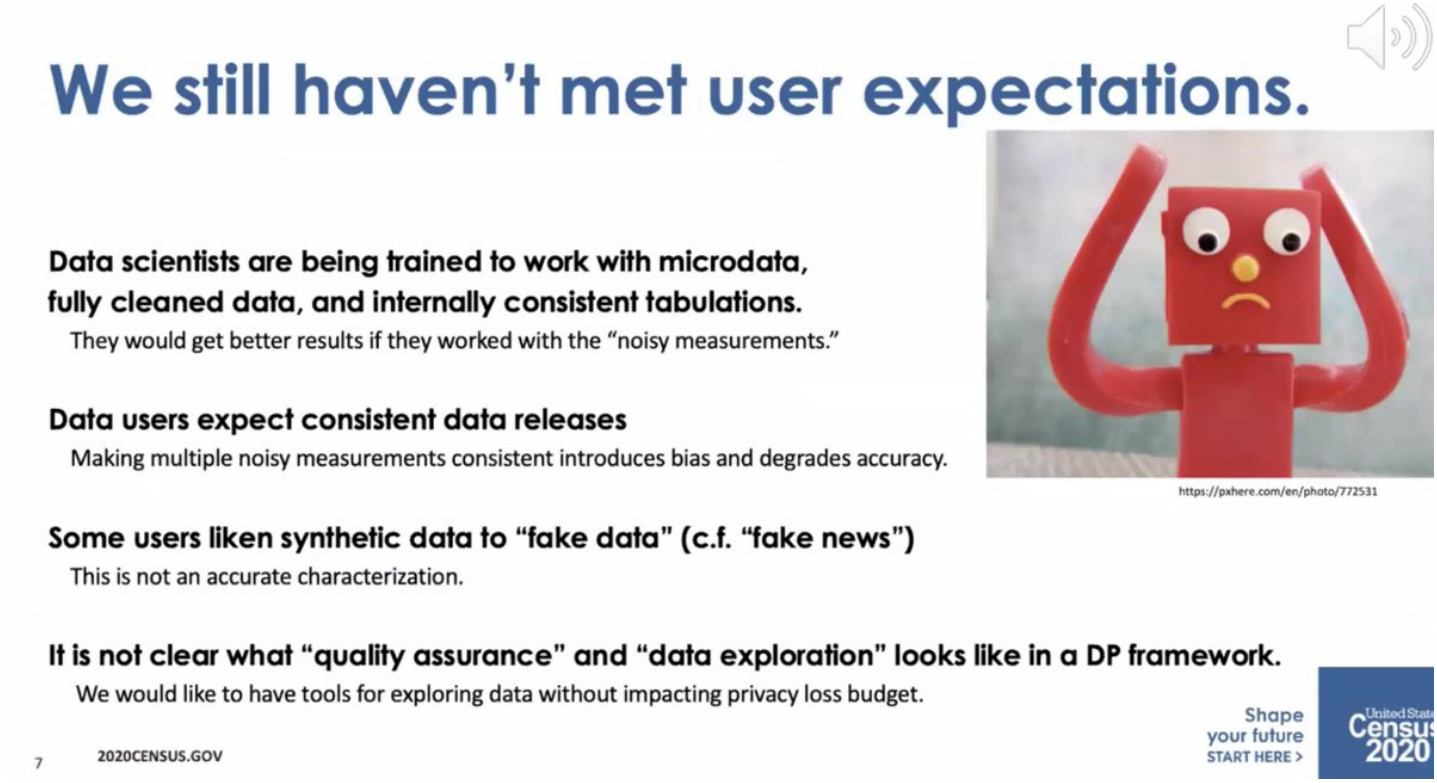
Lea Kissner
@LeaKissner

Last up in Privacy Tech for #enigma2021, @xchatty speaking about "IMPLEMENTING DIFFERENTIAL PRIVACY FOR THE 2020

Differential privacy was invented in 2006. Seems like a long time but it's not a long time since a fundamental scientific invention. It took longer than that between the invention of public key cryptography and even the first version of SSL.



But even in 2020, we still can't meet user expectations.
* Data users expect consistent data releases
* Some people call synthetic data "fake data" like
"fake news"
* It's not clear what "quality assurance" and "data exploration" means in a DP framework



We just did the 2020 US census
* required to collect it by the constitution
* but required to maintain privacy by law

But that's hard! What if there were 10 people on the block and all the same sex and age? If you posted something like that, then you would know what everyone's sex and age was on the block.

Previously used a method called "swapping" with secret parameters
* differential privacy is open and we can talk about privacy loss/accuracy tradeoff
* swapping assumed limitations of the attackers (e.g. limited computational power)

Needed to design the algorithms to get the accuracy we need it and tune the privacy loss based on that.

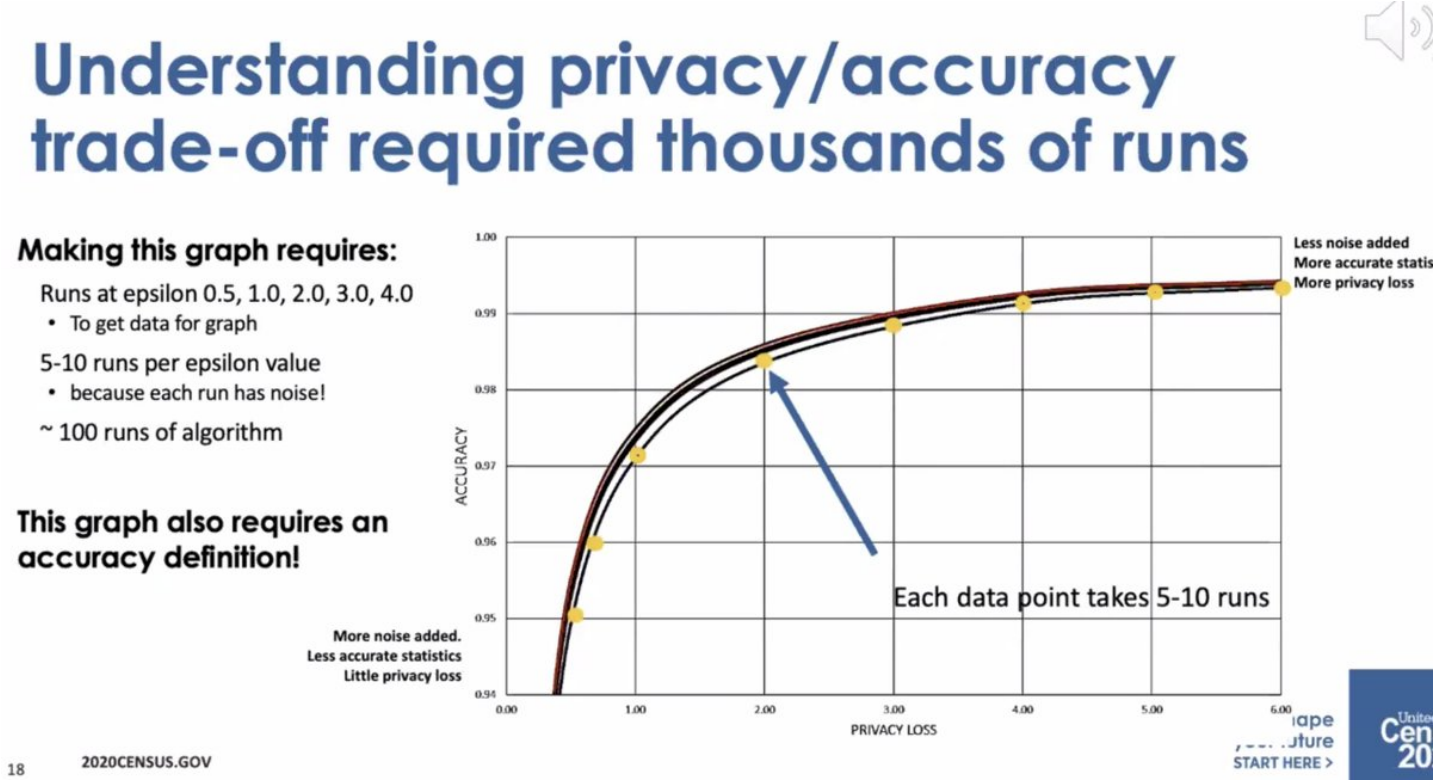
Change in the meaning of "privacy" as relative -- it requires a lot of explanation and overcoming organizational barriers.

By 2017 thought they had a good understanding of how differential privacy would fit -- just use the new algorithm where the old one was used, to create the "micodata detail file".

Surprises:
* different groups at the Census thought that meant different things
* before, states were processed as they came in. Differential privacy requires everything be computed on at once
* required a lot more computing power

* differential privacy system has to be developed with real data; can't use simulated data to do this because the algorithms in the literature weren't designed for dats anything like as complex as the real data (multiracial people, different kinds of households, etc)

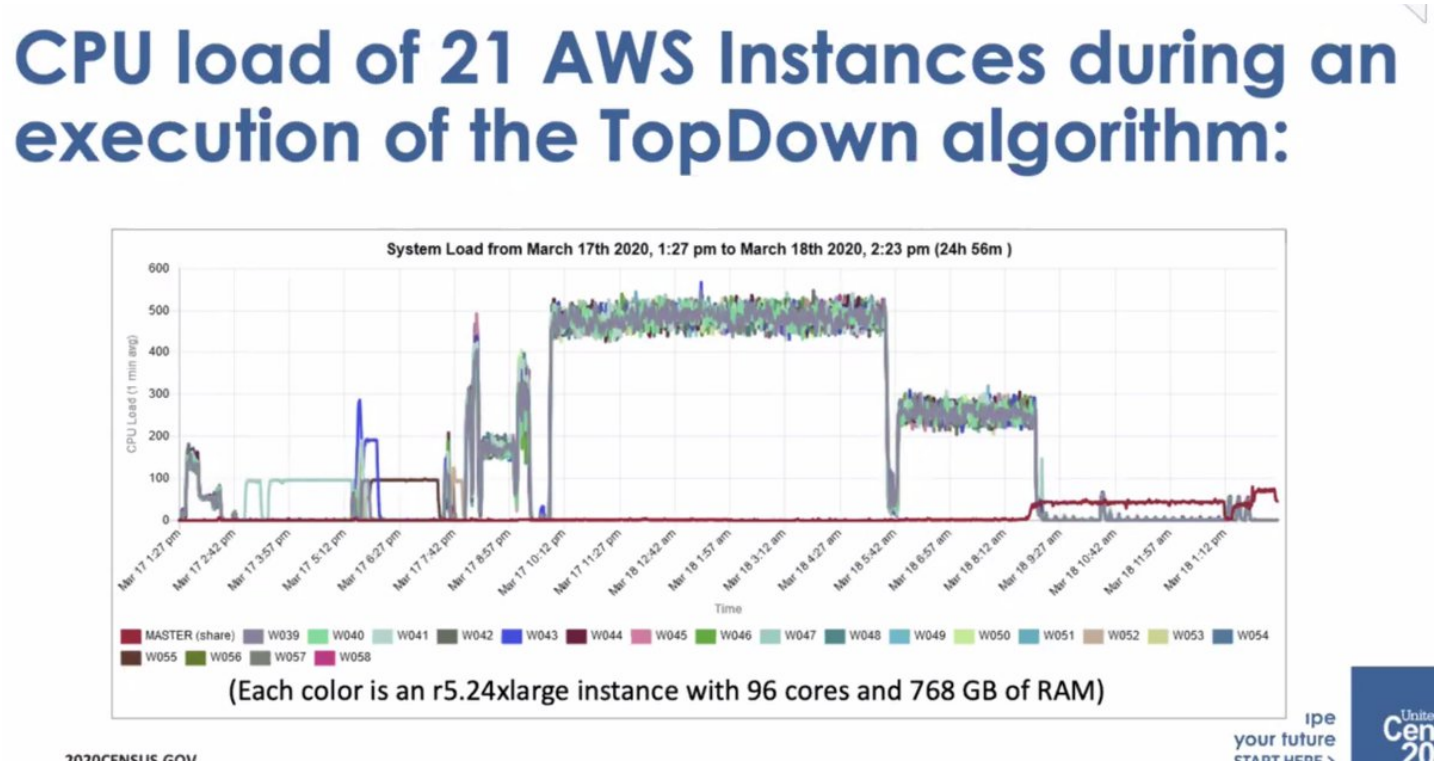
* to understand the privacy/accuracy trade-off requires a lot of runs, representing a *lot* of computer time



Census bureau was 100% behind the move

* initial implementation was by Dan Kiefer, who took a sabbatical
* expanded team to with Simson and others
* 2018 end to end test

* original development was on an on-prem Linux cluster
* then got to move to AWS Elastic compute... but the monitoring wasn't good enough and had to create their own dashboard to track execution
* it wasn't a small amount of compute



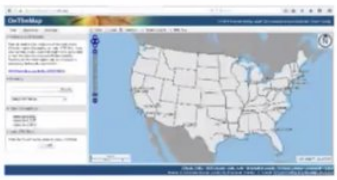
* republished the 2010 census data using the differentially private algorithm and then had a conference to talk about it
* ... it wasn't well-received by the data users who thought there was too much error

For example: if we add a random value to a child's age, we might get a negative value, which probably won't happen to a child's age.

If you avoid that, you might add bias to the data. How to avoid that? Let some data users get access to the measurement files [I don't follow]

In summary, this is retrofitting the longest-running statistical program in the country with differential privacy. Data users have had some concerns, but believe it will all come out.

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.

2020CENSUS.GOV

Shape your future START HERE >

Code is up on github and papers are up online. (@xchatty have some links?)

[end of talk]