INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

DIFFERENTIAL PRIVACY FOR GOVERNMENT AGENCIES

ARE WE THERE YET?

Jörg Drechsler

NTTS-2021
March 9-11, 2021

DIFFERENTIAL PRIVACY EMBRACED BY THE INDUSTRY

Learning statistics with privacy, aided by the flip of a coin
Google Security Blog, October 30, 2014

Apple's 'Differential Privacy' Is About Collecting Your Data—But Not *Your* Data
Wired Magazine, June 13, 2016

Uber becomes the latest company to embrace differential privacy
International Association of Privacy Professionals
June 14, 2017

WHY ARE APPLICATIONS AT STATISTICAL AGENCIES SO LIMITED?

DIFFERENTIAL PRIVACY (DP) IN A NUTSHELL

A randomized function κ gives ε-differential privacy if and only if for all datasets D_1 and D_2 differing on at most one element, and for all $S \subset Range(\kappa)$,

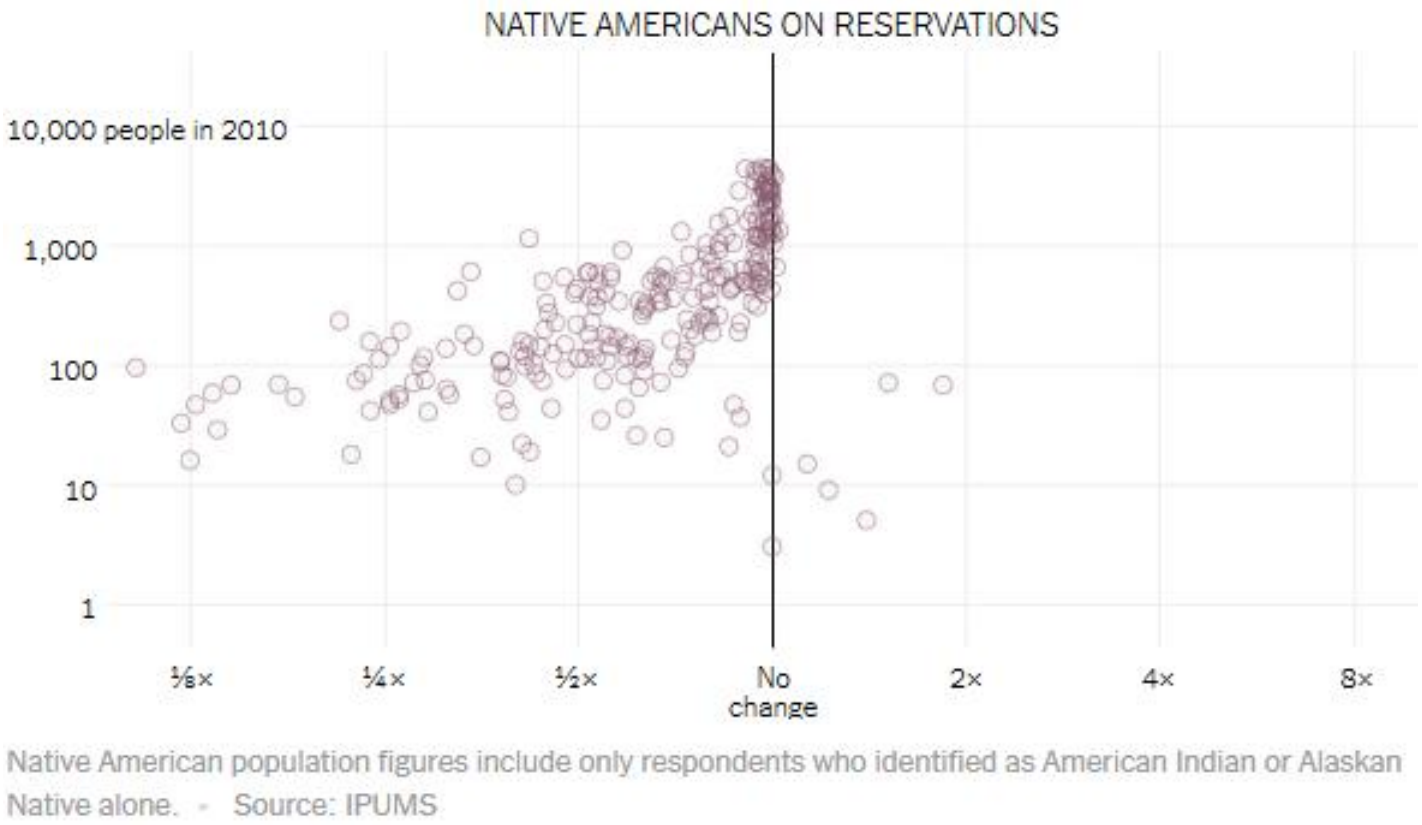
$$P(\kappa(D_1) \in S) \leq \exp(\varepsilon) P(\kappa(D_2) \in S)$$

- DP originally developed for a query response system
- Researcher submits a query
- Receives a perturbed version of the query output, which ensures DP
- DP guarantees that the probability of obtaining a specific result does not change significantly, if I change a single record in the database
- Implies that amount of information that can be obtained about a single unit is also limited

QUERY RESPONSE SYSTEM NOT AN OPTION (IMHO)

- Composition property would allow to give individual privacy budget ε^* to each user
- Still difficult to implement in a dynamic setting
  - Need to prioritize
  - Would need to know all queries in advance
  - Otherwise need to decide for each query how much of the overall budget will be spent
  - Being not restrictive enough would imply that data have to be destroyed before all questions are answered (first-come-first-served approach)
  - Being too restrictive would make query output unnecessarily inaccurate
  - What about replication studies?
- Only viable option seems to be differentially private synthetic data
- Much more difficult to preserve high level of accuracy

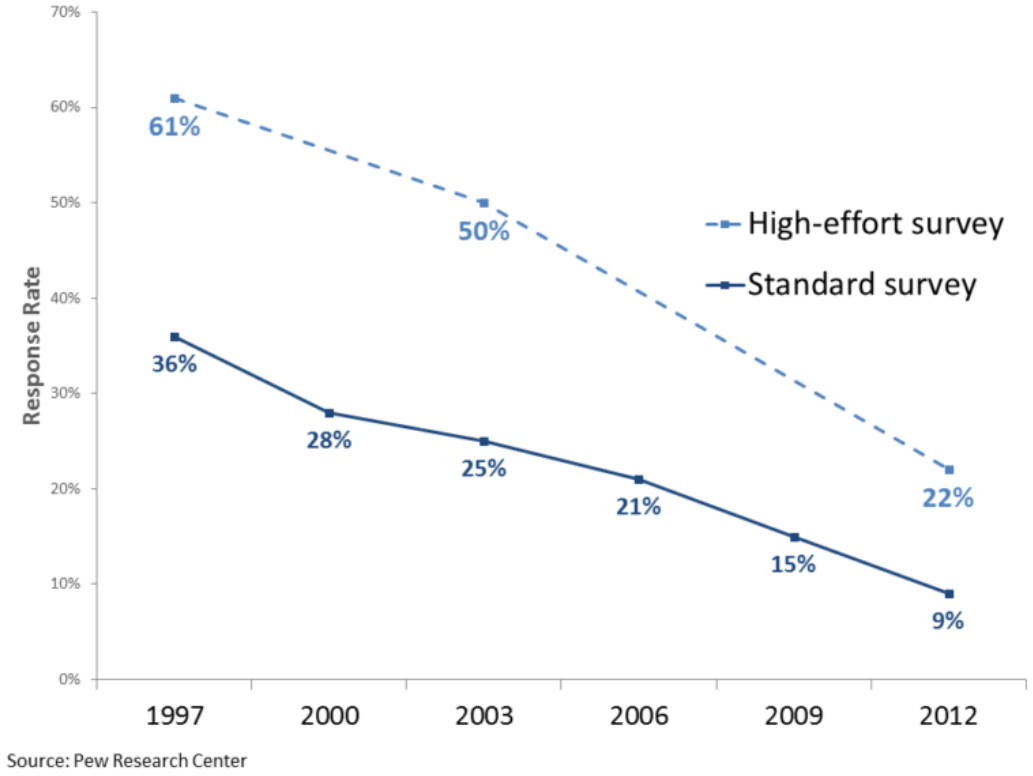
LOW ACCURACY AS HARMFUL AS LOW LEVEL OF PRIVACY



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

DIFFERENTIAL PRIVACY IN THE SURVEY CONTEXT

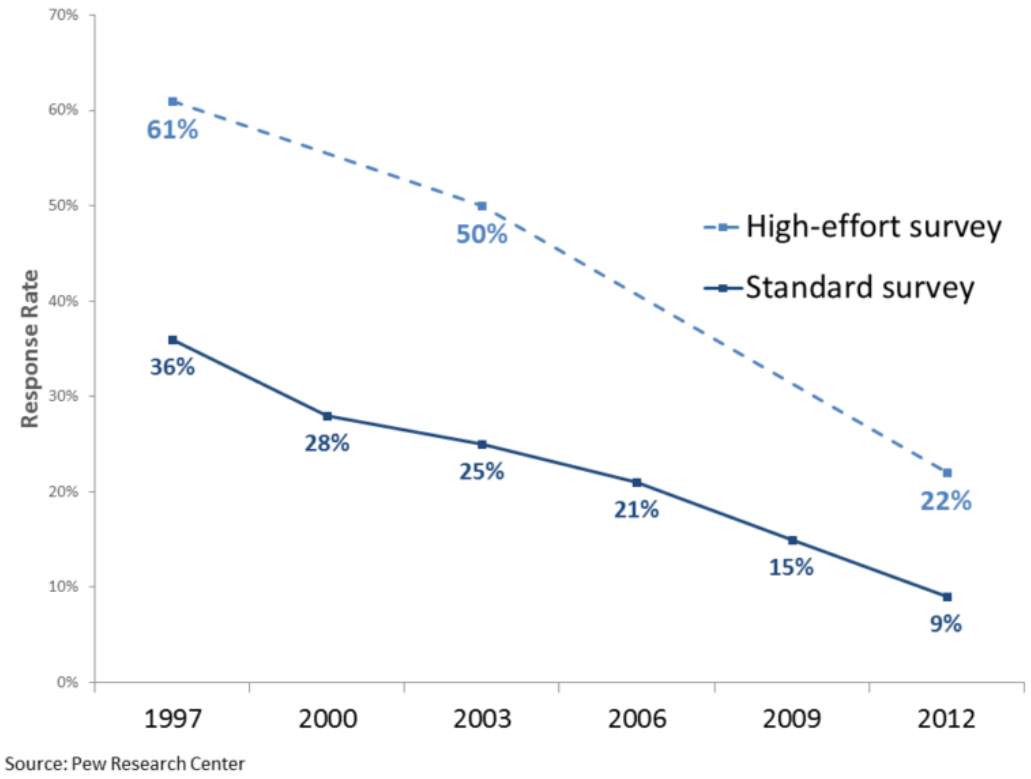
- Benefits from sharing the data obvious in the industry context
- Benefits from survey participation far less obvious
- Response rates in surveys are constantly declining



https://verstaresearch.com/blog/response-rates-fall-to-new-low/

DIFFERENTIAL PRIVACY IN THE SURVEY CONTEXT

- Guaranteeing DP should have positive effects
  - Response rates might increase
  - Quality of responses might increase
- Guaranteeing DP might have negative effects
  - Benefits from participation even less clear



OTHER ASPECTS I DON'T HAVE TIME TO TALK ABOUT

- For surveys sample sizes are typically small, but amount of information is large
- Data need to be available over a long period of time
- Goal is typically to make inference regarding an underlying population → Difficult to account for extra uncertainty from data protection
- Unclear how to deal with weighting, imputation, and data editing
- Difficult to interpret the value of ε
  - What is the level of data protection offered?
  - What are the impacts on analytical validity?
- How do we set the level of ε?

CONCLUSIONS

- Differential privacy very attractive as a concept
- Sensible implementation in practice much more difficult
- Situation for government agencies substantially different from previous applications
- Progress has been made, but many open questions remain
- Interesting area for research

- Full paper available at: http://arxiv.org/abs/2102.08847

CONTACT

Jörg Drechsler

joerg.drechsler@iab.de