

Discussant Remarks

Kunal Talwar

CSAC Member

March 2019 Meeting

Reconstruction

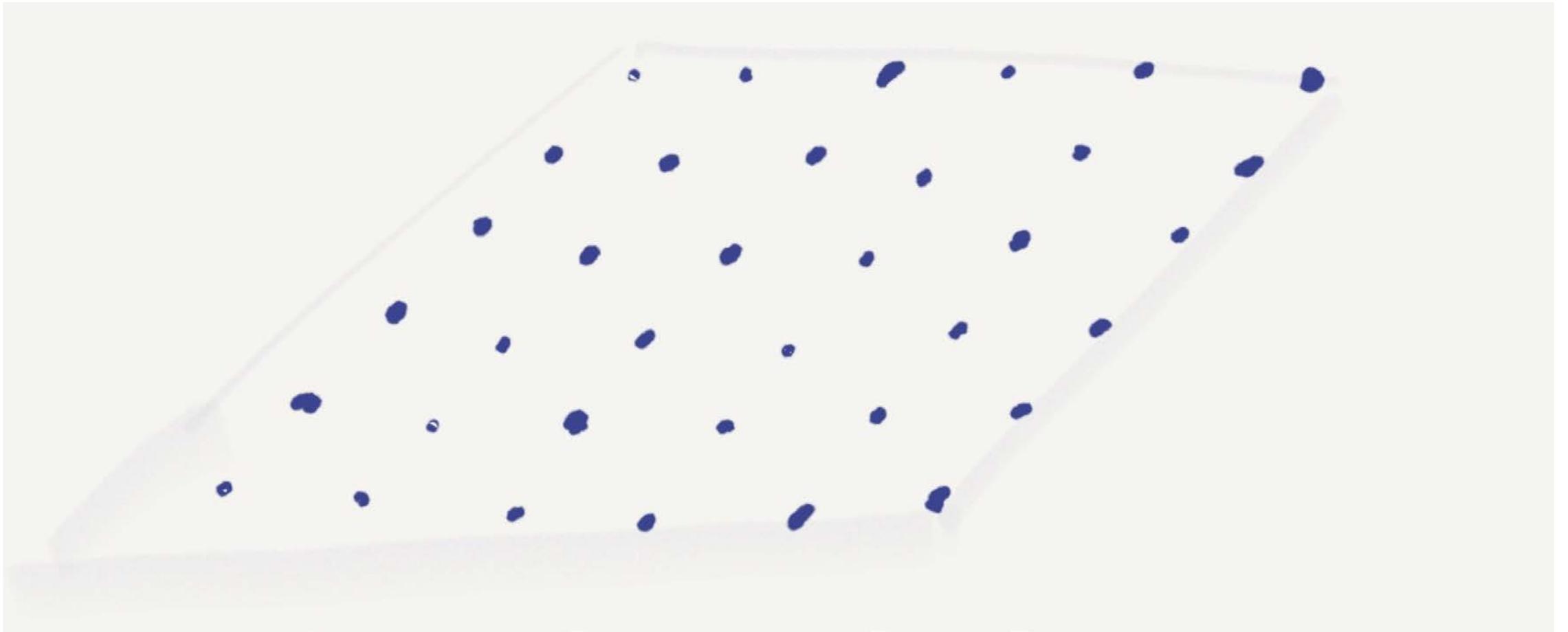
- The fact that the (relatively simple) reconstruction attack works so well is worrying
- The 52 million people correctly identified is a lower bound. More sophisticated attacks will almost surely identify more.
- This emphasizes the need for modernizing disclosure limitation

Communicating Vulnerabilities due to Invariants

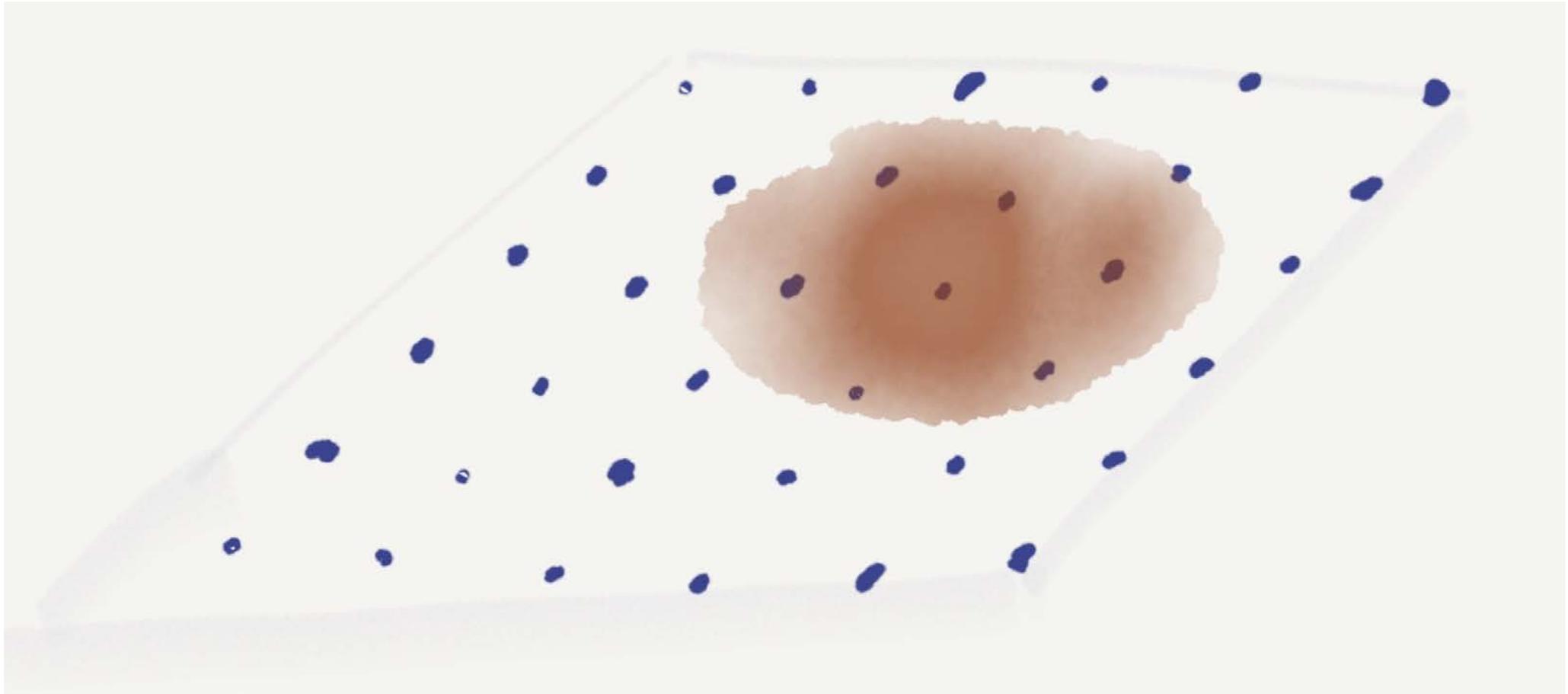
How should the Census Bureau communicate the vulnerabilities that invariants produce while trying to eliminate them from the publications?

Invariants put strong constraints on the set of feasible datasets. These strong constraints make the release mechanism vulnerable.

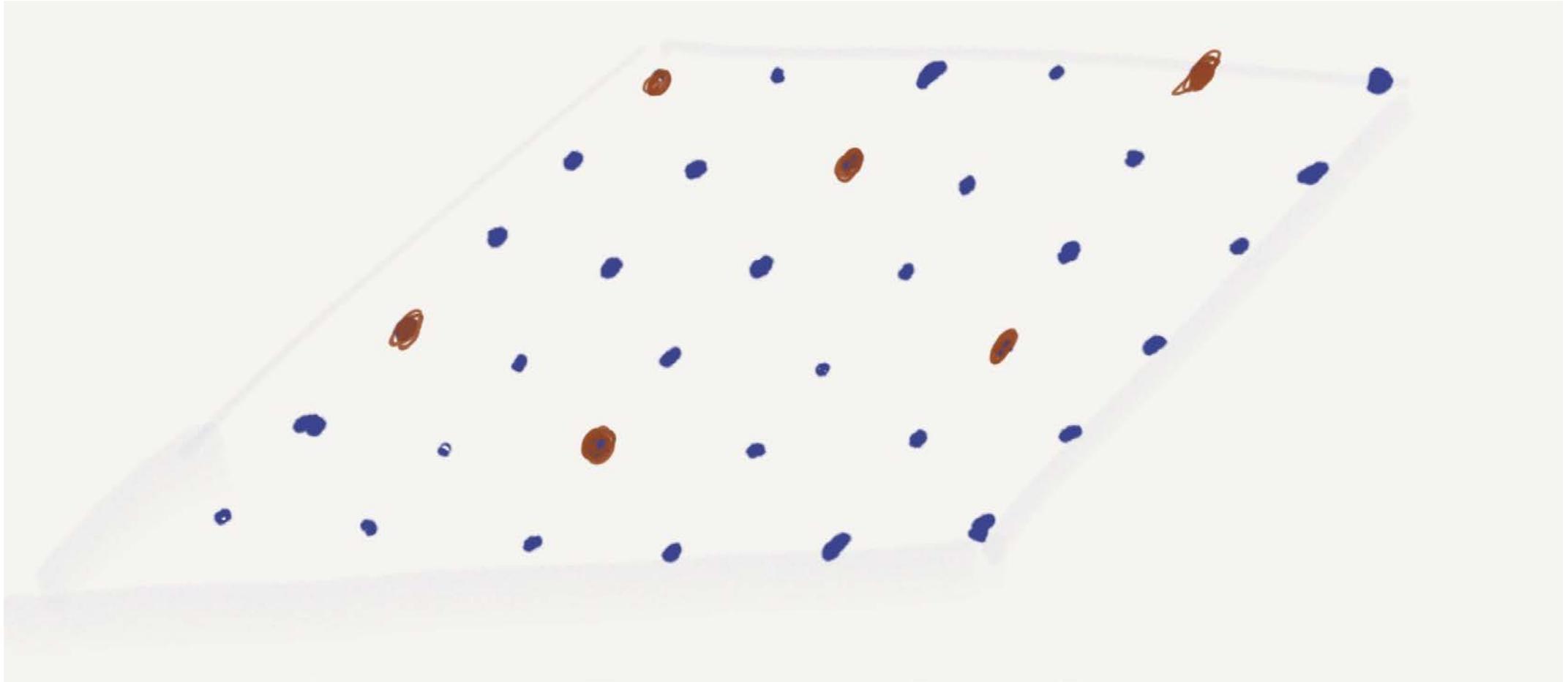
Invariants



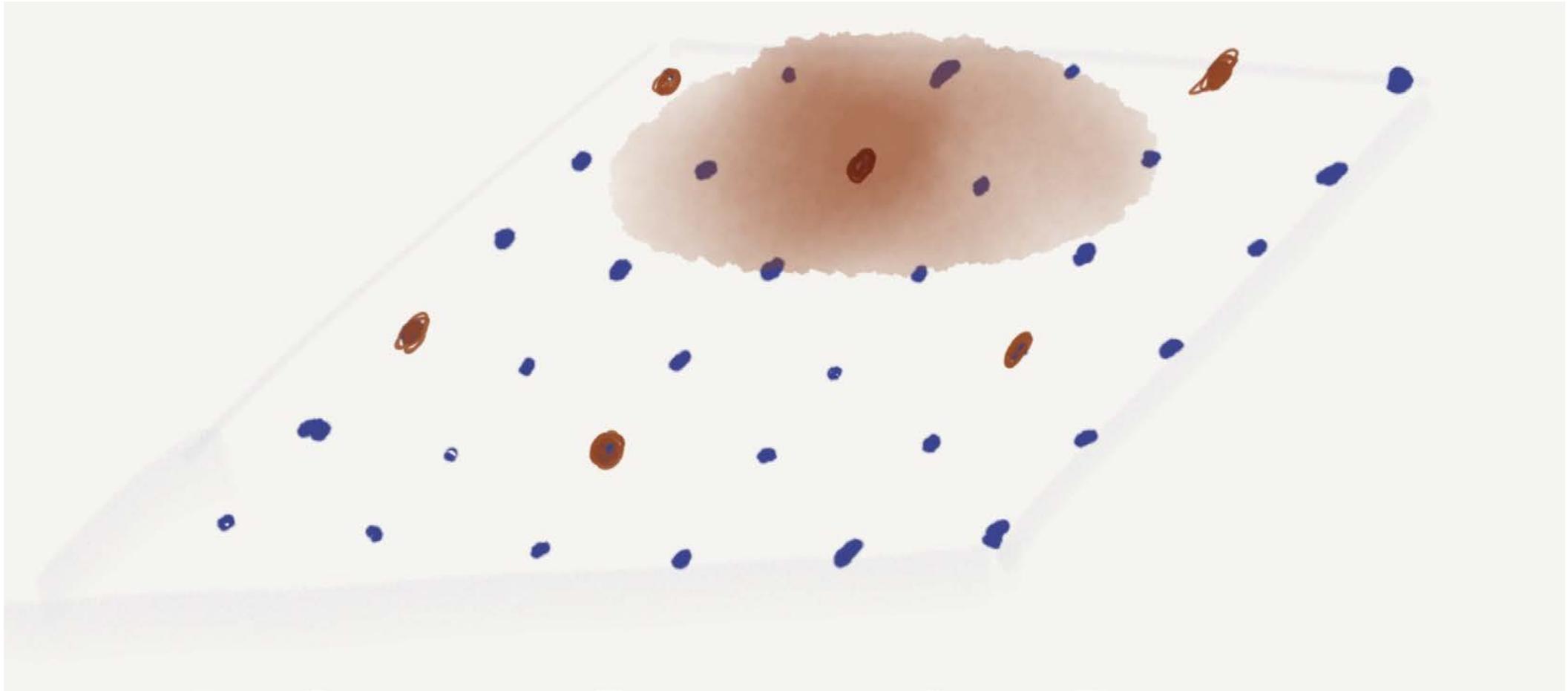
Invariants



Invariants



Invariants



Communicating Vulnerabilities due to Invariants

How should the Census Bureau communicate the vulnerabilities that invariants produce while trying to eliminate them from the publications?

Invariants put strong constraints on the set of feasible datasets. These strong constraints make the release mechanism vulnerable.

Releasing invariants causes quite some privacy harm. Should only be published if there are very strong reasons to publish them.

Complete Accuracy Infeasible

How can the Census Bureau effectively communicate to users that complete accuracy of inputs to their use cases is infeasible, and was not true historically?

- Traditional Disclosure Limitation techniques distort data in a way unknown to the user. Difficult/Impossible for user to account for that.
- DP techniques allow one to account for the noise added. Bayesian post-processing gives precise error bars.
- Some users perhaps worry that these error bars are larger than what previous techniques had. This relates to the policy question of choosing ϵ .
 - Also much smaller error bars would mean reconstruction still works
- Fundamental tension between accuracy and reconstructability. DP is the state of the art solution to this tension. For such queries, it is “nearly optimal”.

Balancing Accuracy Requirements

How can the Census Bureau best do principled balancing of the accuracy requirements of diverse use cases?

Three questions:

- Data Collection: Figure out the use cases.
- Policy: What balance do we want between them?
- Technical: Knowing the use cases and the objective, how do we optimize?

Partitioning privacy across geographic levels

In tuning the full geographic hierarchy, which levels make the most sense to optimize for accuracy?

- Related to the previous question. Perhaps leave these as “hyperparameters” that are tuned based on the use cases.
- Hyperparameters can be tuned on previous Census data, or one can tune privately

Alternate Dissemination for join tables

If the only feasible algorithms for producing household-person join tables and detailed race, ethnicity and AIAN tables cannot deliver microdata for tabular publication, should the Census Bureau invest in a dissemination system that publishes from protected tables instead?

Are you envisioning a DP interface to the data? Or access to data in a physical or virtual protected data enclave?

This may create additional invariants?

Use cases for PUMS & restricted access

How should the Census Bureau assess the use case for PUMS and restricted-access to the confidential microdata?

Relaxing Full Consistency

Should the Census Bureau relax the requirement that all published tables be fully consistent, as other national statistical offices have done for their census publication?

This is reasonable and would make the data more useful in some ways, as forcing consistency can make the reasoning about noise harder from a computational/analytic point of view. One could imagine software (from census or external) that enforce consistency between two, or a few tables, for applications that require consistency.

Are there strong arguments for requiring full consistency?

Impact

How can the Census Bureau incorporate systems that will give a holistic perspective on the impact of these changes?

- For some common kinds of use cases of the data, provide tools/examples to do a version of the analyses that takes into account the noise added for privacy.