Enabling Dynamic Linkage of Linguistic Census Data at Statistics Canada

Arnaud Casteigts, Marie-Hélène Chomienne, Louise Bouchard, Guy-Vincent Jourdan University of Ottawa, Canada

{acasteig,mh.chomienne,louise.bouchard,gjourdan}@uottawa.ca

uOttawa

Context

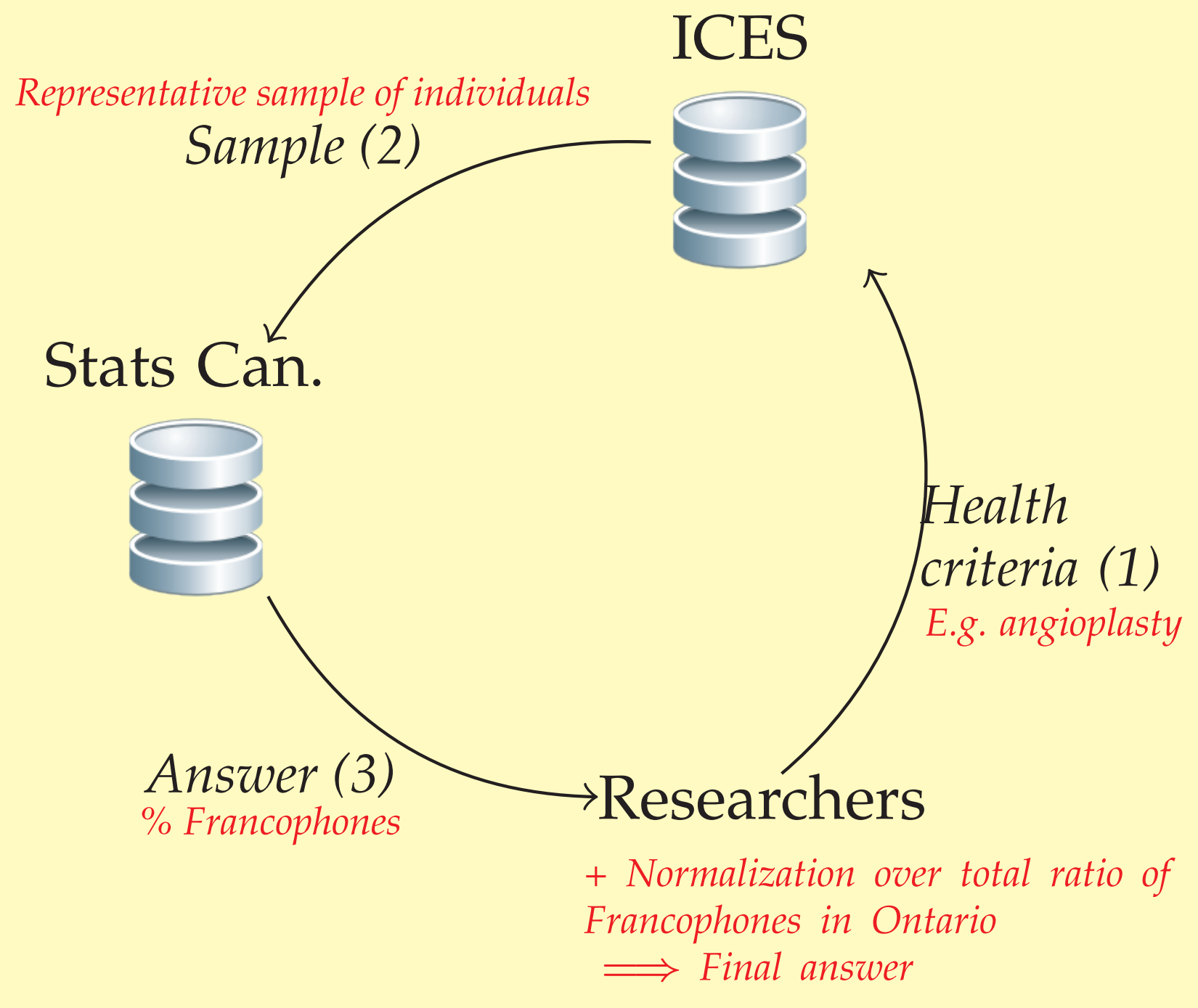
Research in population health consists in studying the impact of various factors (*determinants*) on health, with the long-term objective of yielding better policies, programs, and services. Researchers of Official Language Minority Communities (OLMCs) focus specifically on determinants related to speaking a minority language, such as English in Quebec, or French in the rest of Canada. Investigations of this type require the possibility of associating health data to linguistic information. Unfortunately, the largest health databases in Ontario, held at the Institute for Clinical Evaluative Sciences (ICES), do not contain linguistic variables. High-quality language variables however exist at Statistics Canada (SC) through the 2006 Census.

Purpose

We are interested in enabling a form of linkage between ICES health data and SC linguistic data that could be *automated*, and yet, proven safe. To this end, we conjecture that most OLMC-related questions could be reformulated as a counting problem in given samples of patients (e.g. counting how many are Francophones), and therefore reduce the complexity of a query to its essential minimum. Two solutions are proposed based on this principle. The first one assumes a particular dataflow which preserves privacy by means of a tripartite interaction; the second discard the need for such an assumption by adding a random pertubation to the answer, which makes the collection of residual information almost impossible (we characterize the worst-case leakage precisely). Based on these results we argue that a safe exposition of linguistic data is possible, and beyond, that similar techniques could be used to enrich provincial health databases with many other census variables.

Solution 1: Tripartite interaction

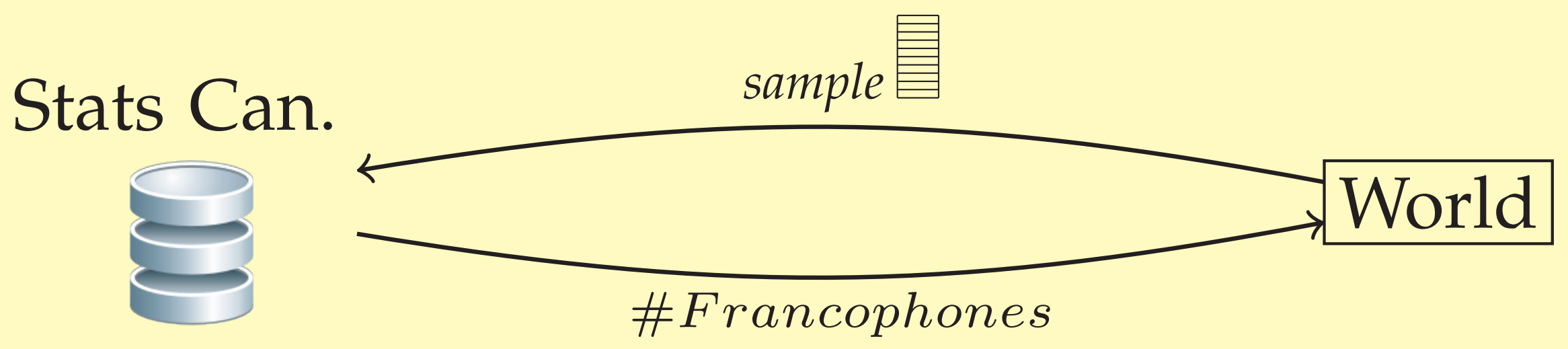
This solution consists of a circular workflow between the three involved entities: OLMC researchers, ICES, and SC. The workflow is initiated by the researcher through the submission of health criteria to ICES. A representative sample of individuals matching these criteria is then generated and sent to SC, which performs the count query. The result of the query is finally returned to the researcher.



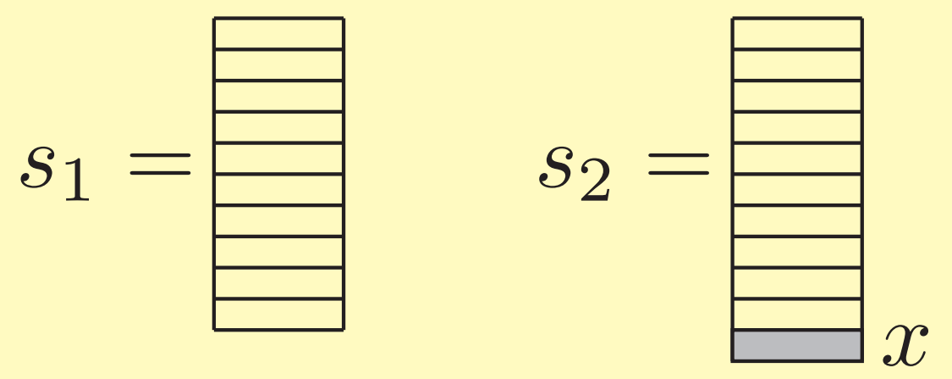
Example: *what is the average angioplasty rate among Francophones? (vs. Anglophones).* Researchers submit the criteria to ICES, and gets an answer from SC. By normalizing this answer over the global ratio of Francophones in Ontario, one can answer the initial question.

Privacy in this mechanism stems from the fact that i) researchers do not know the sample details, ii) SC does not know the health criteria that were used to generate that sample, and iii) ICES does not know the final answer. Therefore, none of the entity can learn something it is not supposed to know.

Drawback: This scheme assumes that no additional exchange occurs between the entities. In particular, the distinction between ICES and the researchers may be debatable from the point of view of SC, which generally considers anything from outside as one and a single entity.



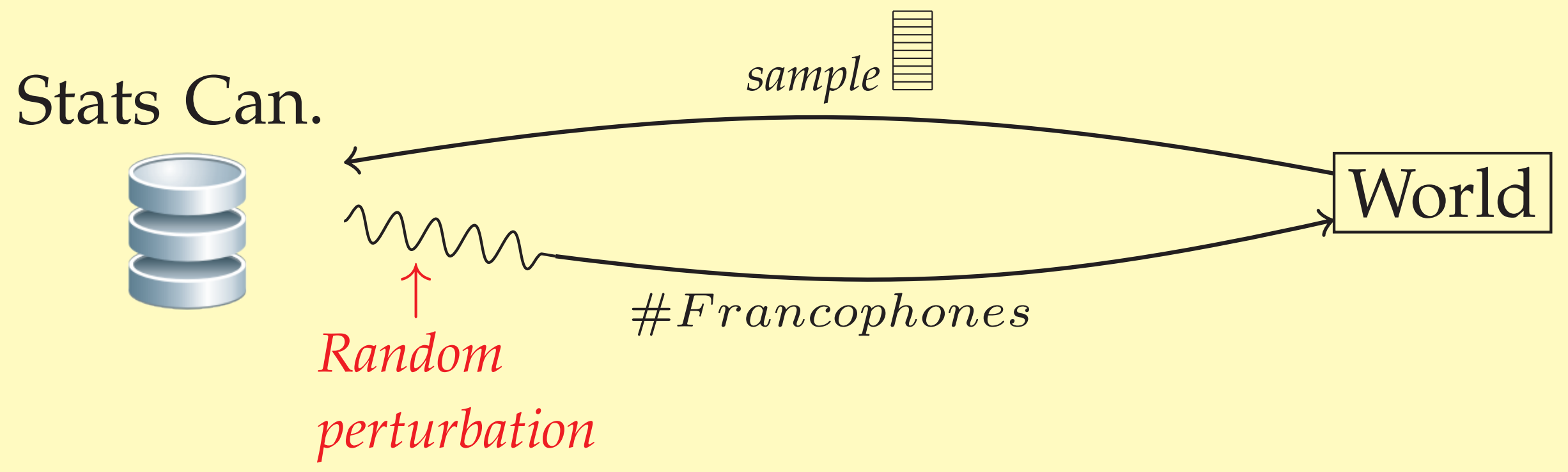
Here, a malicious use of count queries, if unsupervised, could make it possible to identify the language of a given individual (say, Madame x). Consider the following attack: making a query with a sample s_1 that does not contain x; then making a second query with the same sample, plus Madame x.

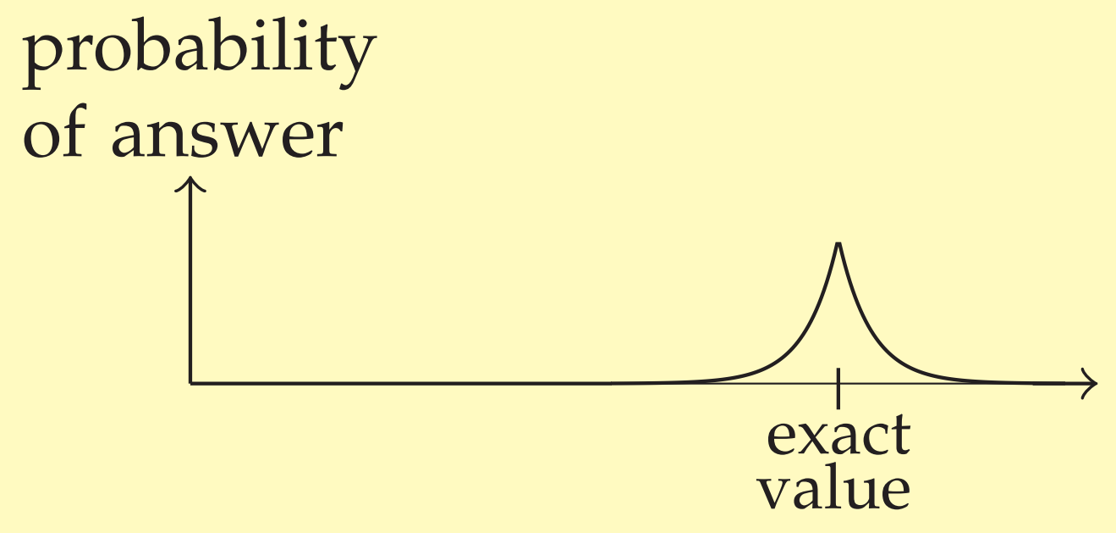


Obviously, if $answer(s_2) > answer(s_1)$, then Madame x is Francophone. A malicious adversary could actually build more complex constructs with similar effects.

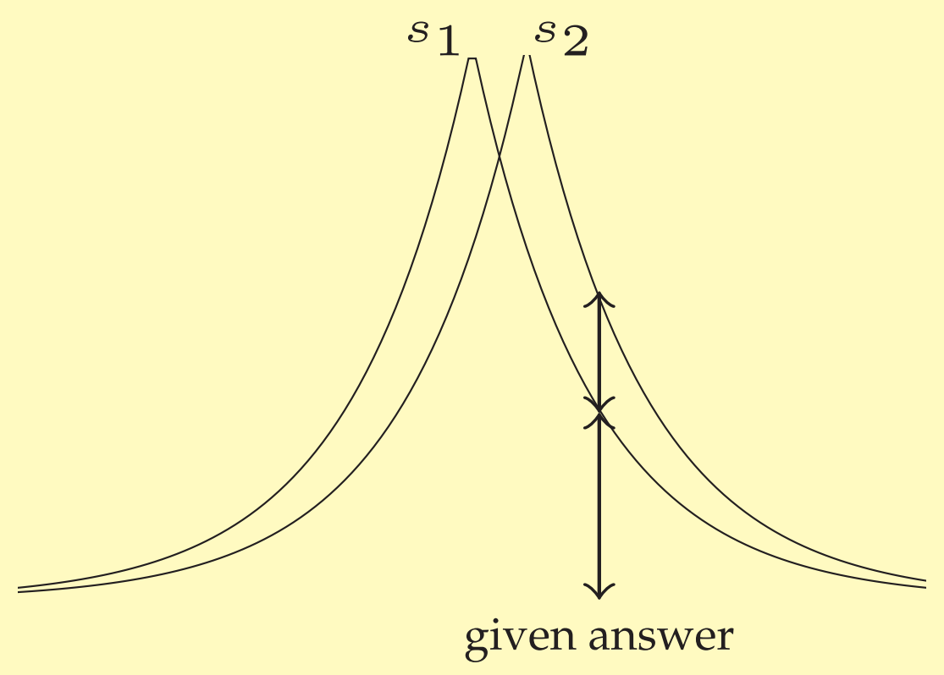
Solution 2: Noised count queries

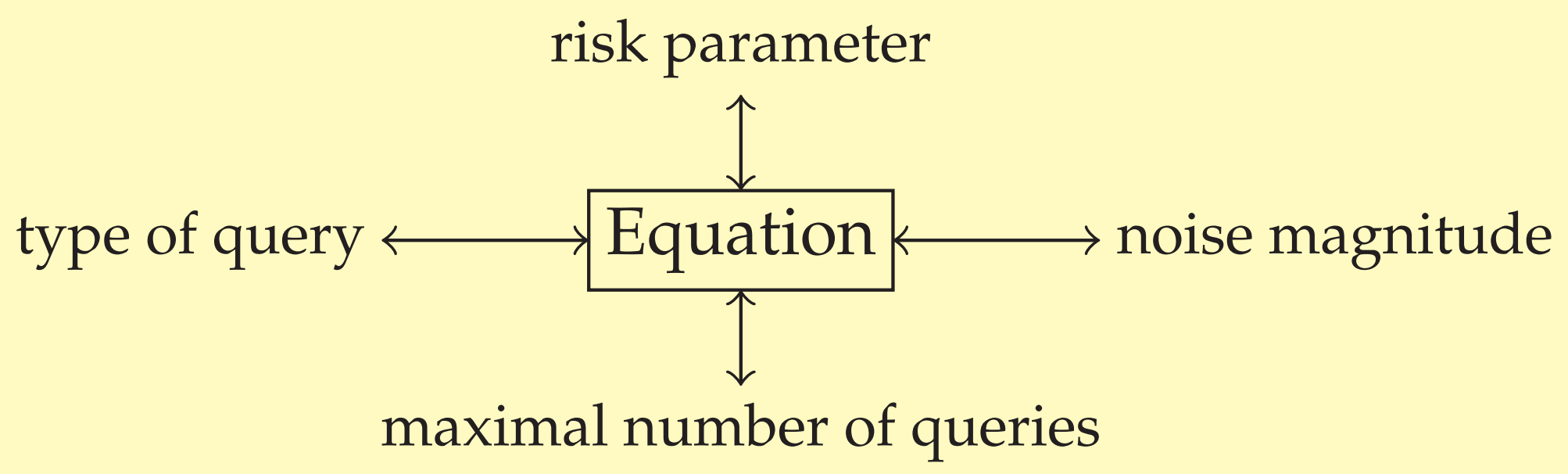
One way to prevent (or strongly limit) the collection of residual information over several queries is to pertub the answers, *i.e.*, drawing a random number – positive or negative – to be added to the exact value.



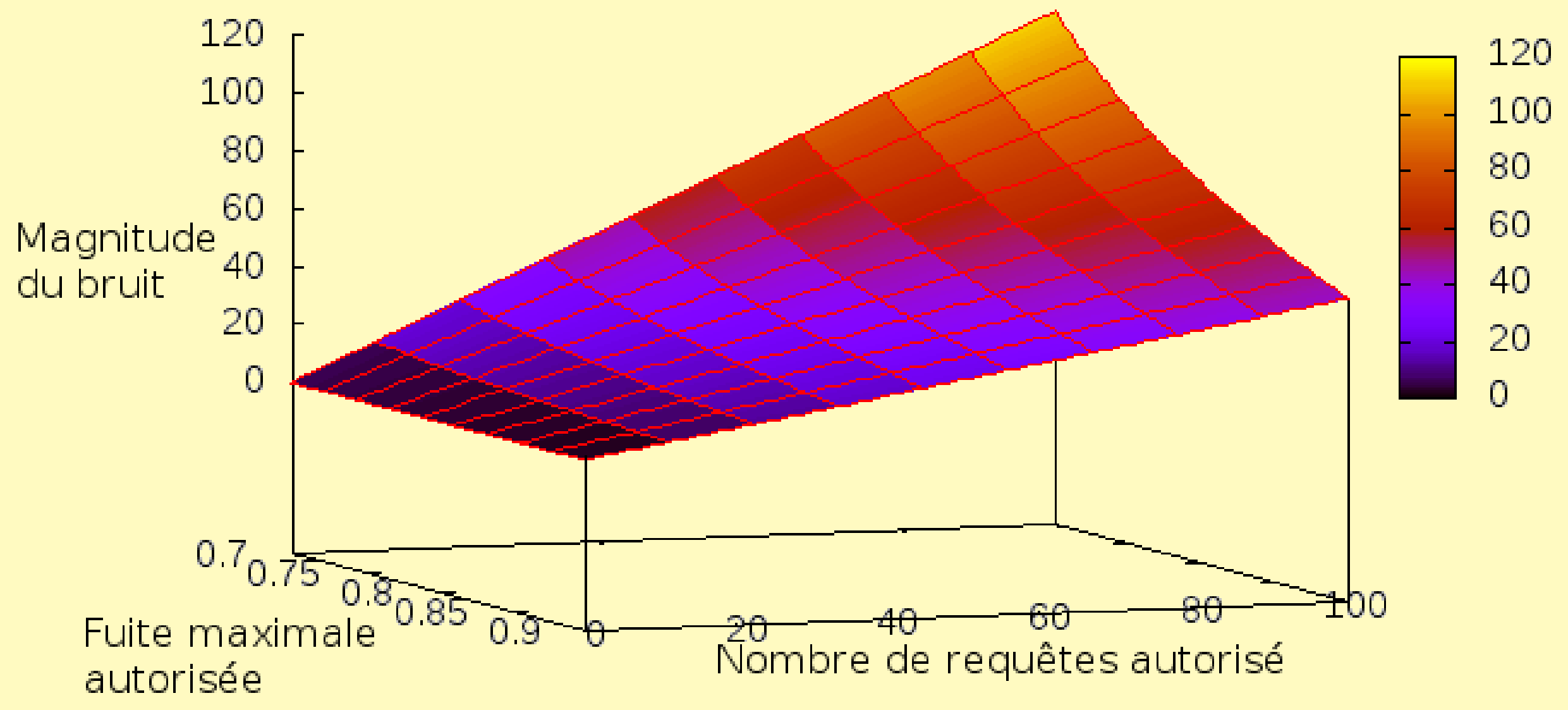


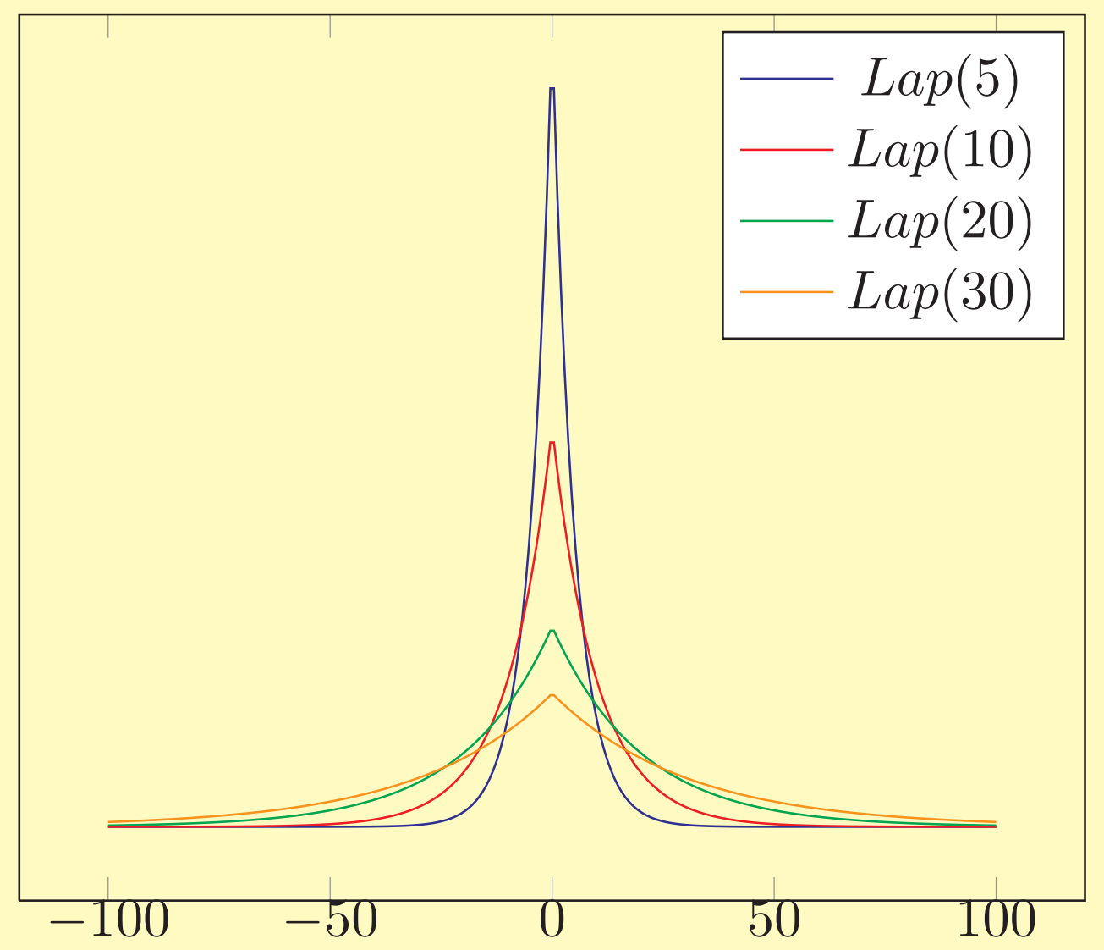
It is intuitively clear that no residual information can be collected with certainty, due to the fact that a same answer could be induced by different counts. The most an adversary can learn is *probabilities* (one would say *belief* in this case) that given individuals are Francophones or Anglophones. The amount of this belief leakage depends on how different the probabilities of answer among several samples are (neighbor samples such as s_1 and s_2 represent a worst-case scenario). This difference depends in turn on the shape and magnitude of the noise (which we assume to be a *Laplacian* one).





Recent works in the field of *private data analysis* (e.g. [1]) made it possible to understand the exact tradeoff between *leakage* and *utility* in a worst-case scenario (right picture above). Following these results and others from the same co-authors, we characterized the particular tradeoff at play in our case, i.e., queries counting the number of Francophones in the samples. Given a desired maximal belief about one's language (typically chosen by the database holder), and a maximal perturbation OLMC researchers can stand for each answer, the tradeoff tells us how many queries can possibly be done.





Conclusion

Both mechanisms enable the linkage of provincial health data with federal census data. In the framework of our concern, they enable to supplement *Ontario* health data with *linguistic* variables. However, we believe this approach is more general and could apply to other variables than language and other provinces than Ontario. The road to realization may be long, but this initial work demonstrates that technical solutions do exist and deserve to be explored.

Reference

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography (TCC'06)*, pages 265–284, 2006.

Funding

RRASFO (*Réseau de Recherche Appliquée sur la Santé des Francophones de l'Ontario*).

IRHM (*Institut de Recherche de l'Hôpital Montfort*).