Differential Privacy in Tripartite Interaction: A Case Study with Linguistic Minorities in Canada

A. Casteigts, M-H. Chomienne, L. Bouchard and G-V. Jourdan

Data Privacy Management (DPM)

–

Septembre 12, 2012

Institut de recherche de l'Hôpital **Montfort**

uOttawa

L'Université canadienne
Canada's university

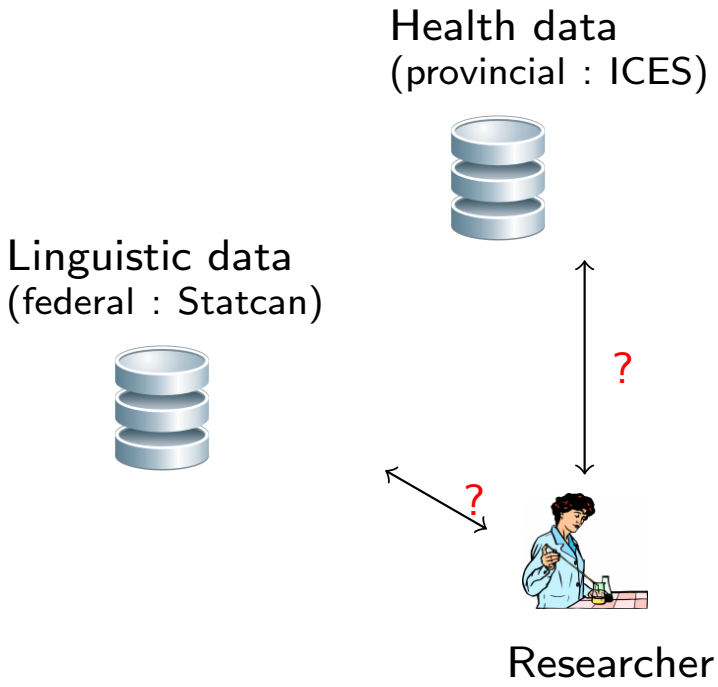
Context & Motivation

Health data
(provincial : ICES)

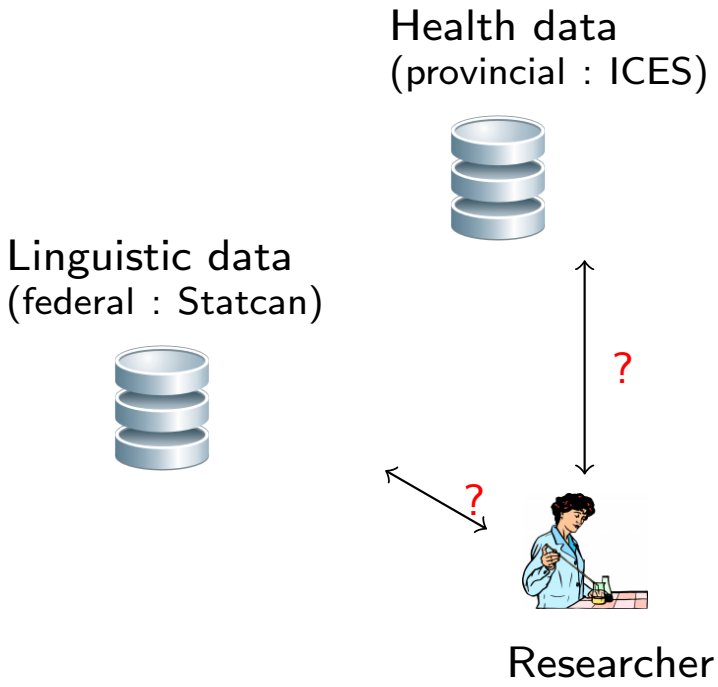
Linguistic data
(federal : Statcan)

Researcher

Context & Motivation



Context & Motivation

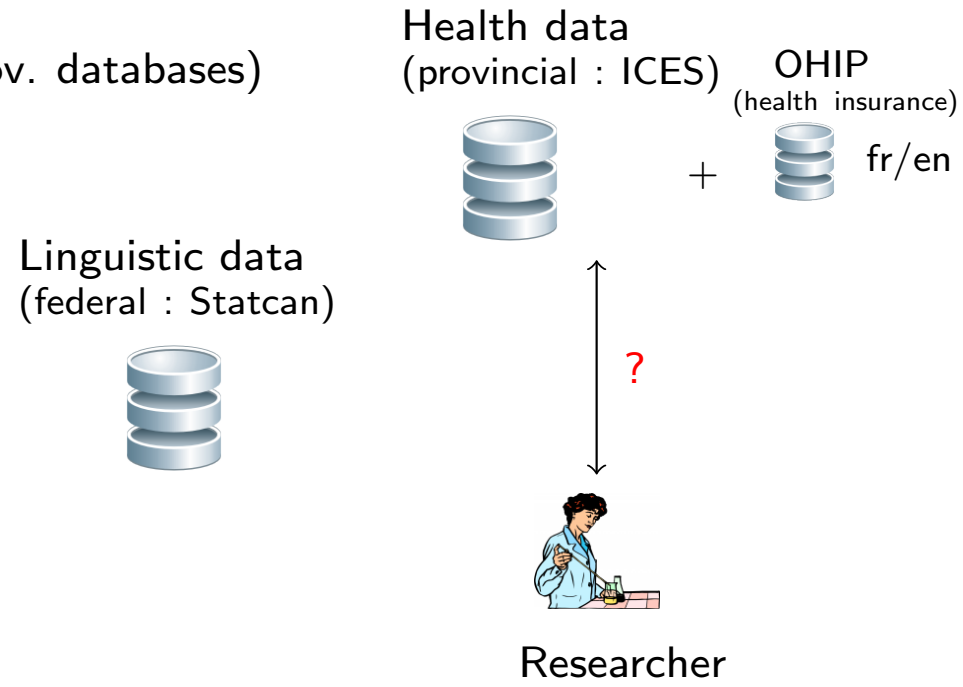


Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Context & Motivation

Option 1

(lang. variable $\rightarrow$ prov. databases)

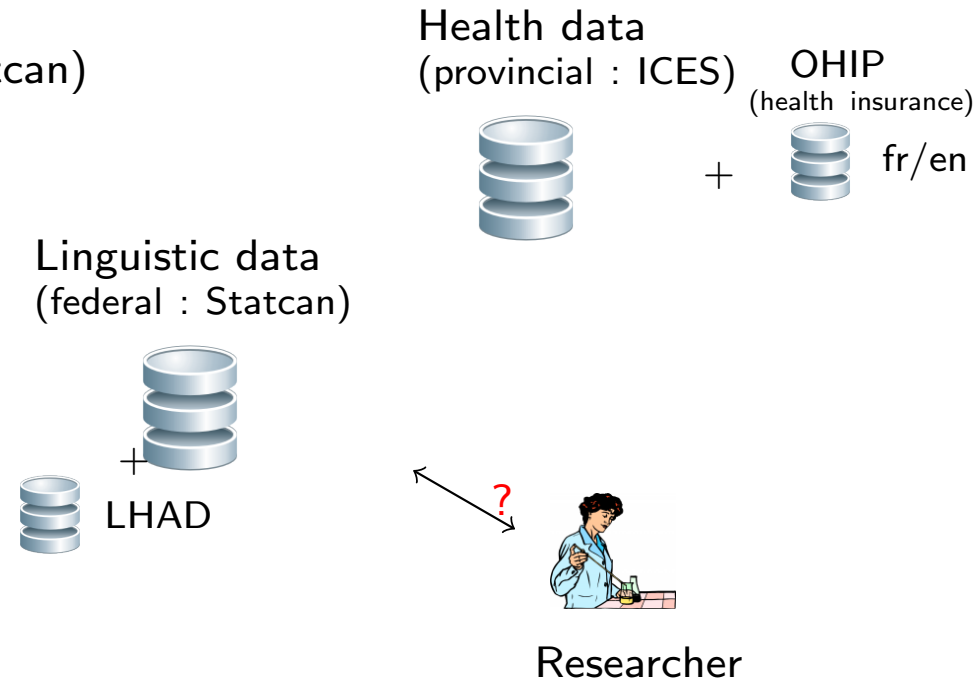


Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Context & Motivation

Option 2

(health data → Statcan)

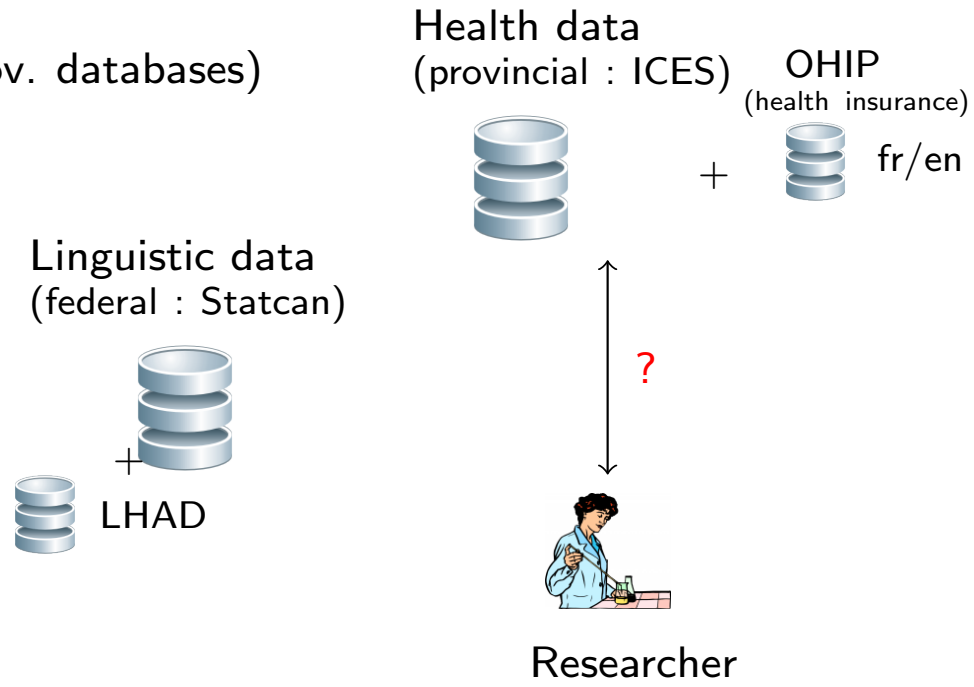


Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Context & Motivation

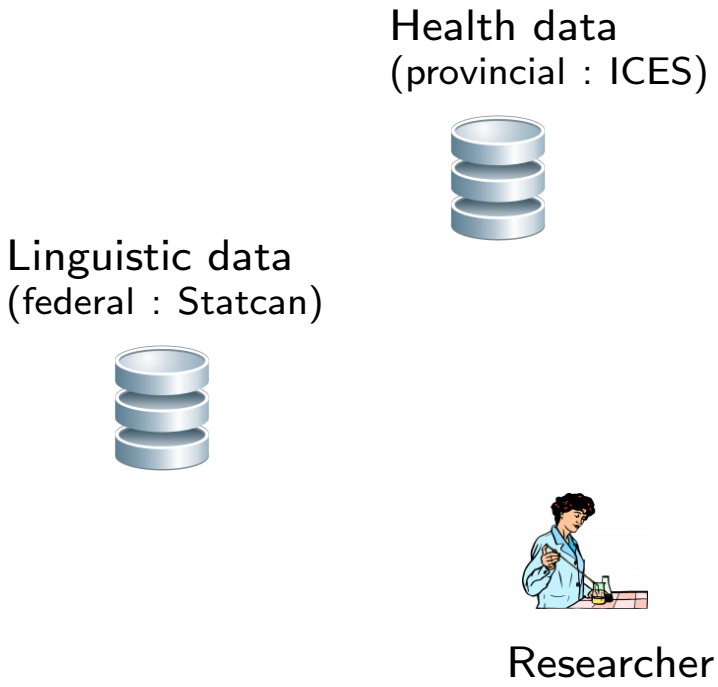
Option 1

(ling. variable $\rightarrow$ prov. databases)



Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

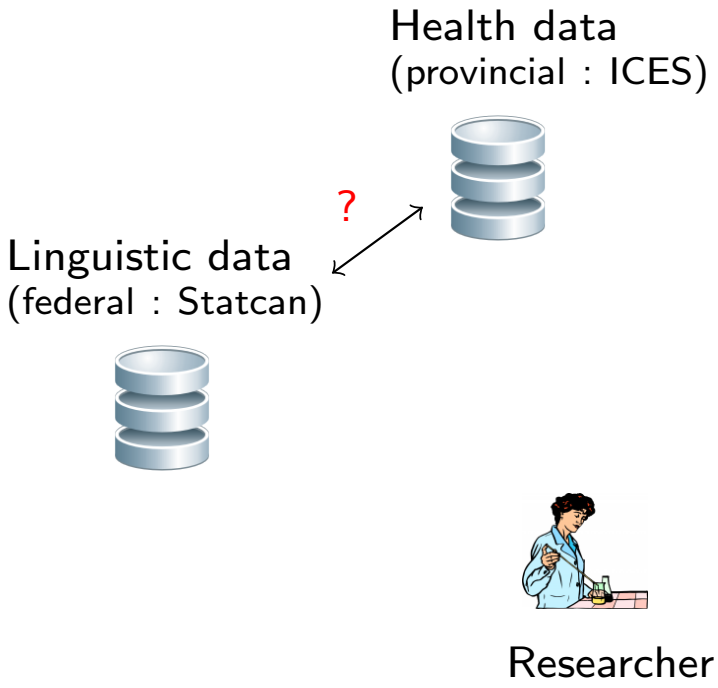
Context & Motivation



Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Context & Motivation

Option 3
(Data linkage)



Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Context & Motivation

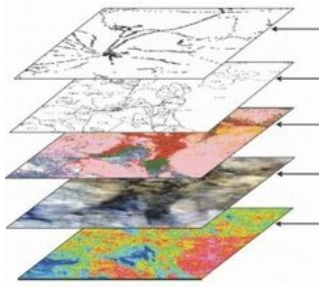
Option 4

(Geographical correlations ?)

Health data
(provincial : ICES)

Linguistic data
(federal : Statcan)

Researcher



Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Context & Motivation

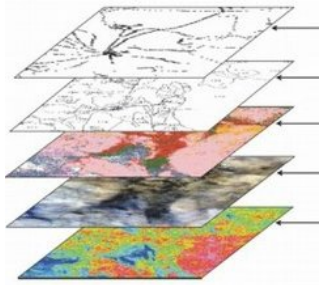
Option 4

(Geographical correlations ?)

Health data
(provincial : ICES)

Linguistic data
(federal : Statcan)

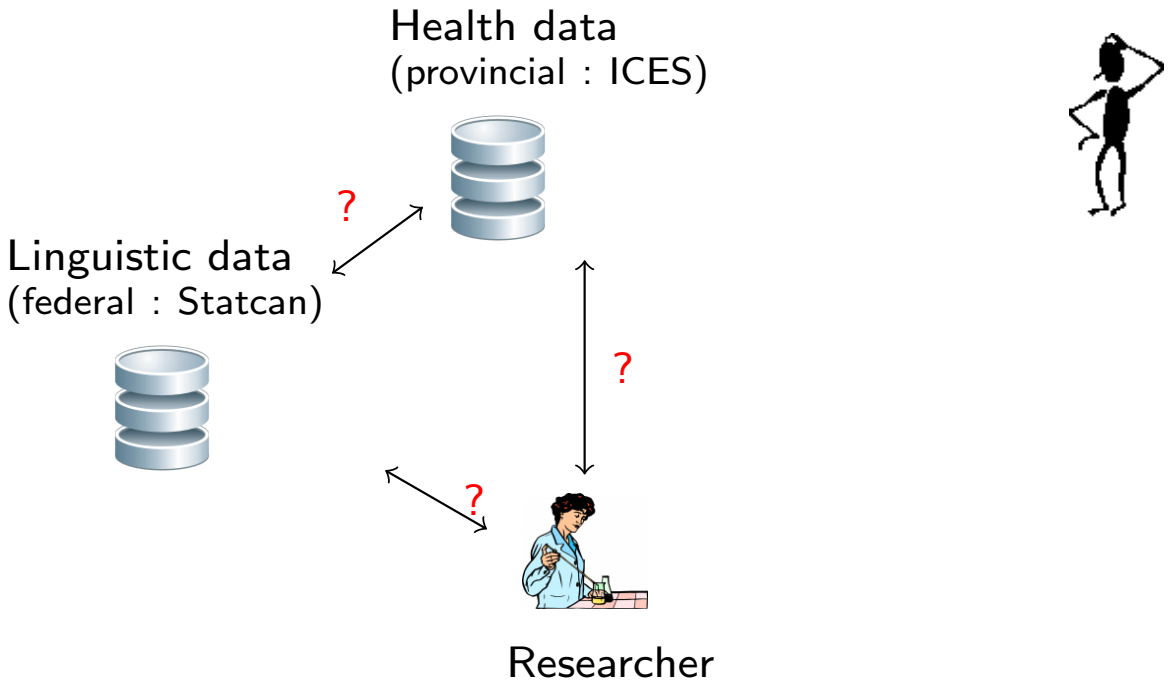
Researcher



Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Ex. 2 : Rate of angioplasty ? ($\Longrightarrow$ not enough density...)

Context & Motivation

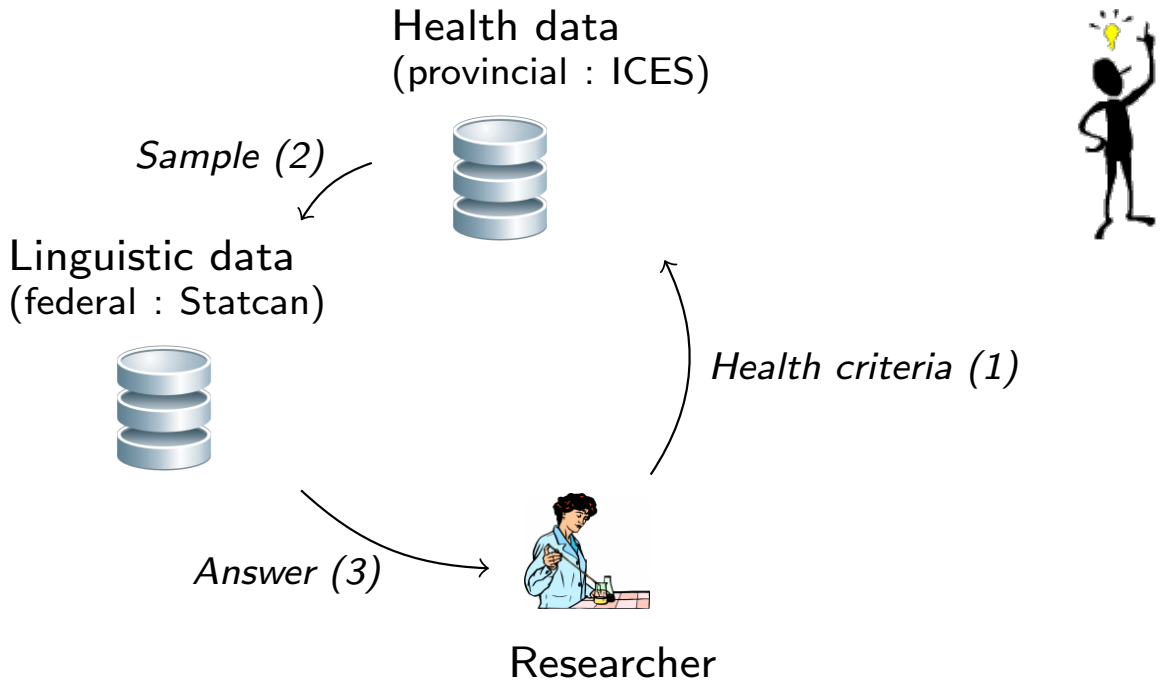


Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Context & Motivation

Suggestion 1

(Tripartite tabulation)

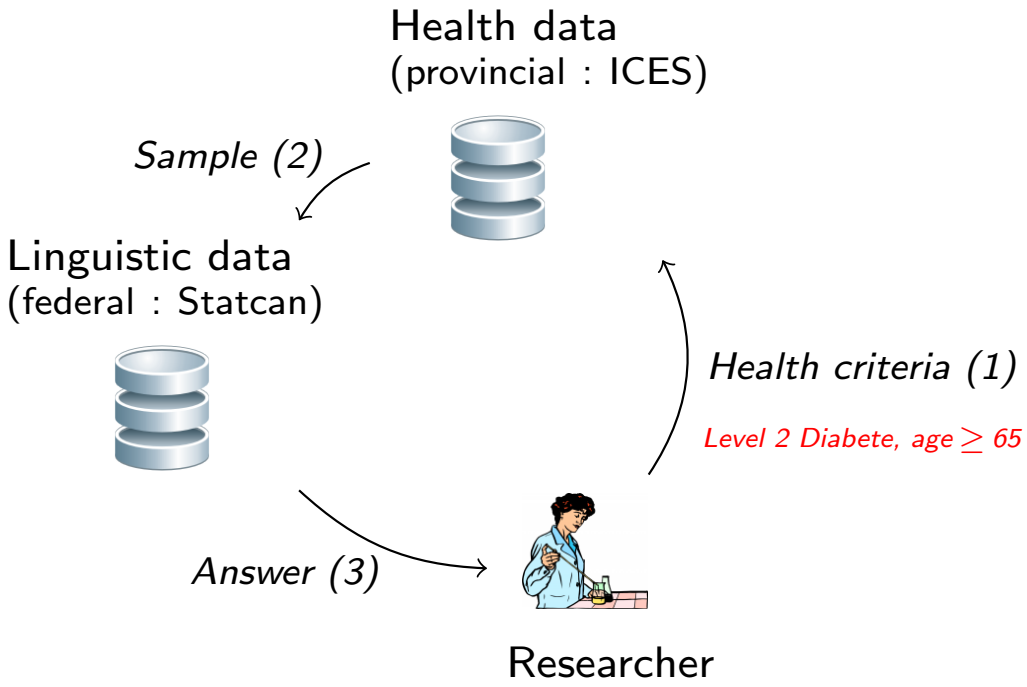


Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Context & Motivation

Suggestion 1

(Tripartite tabulation)

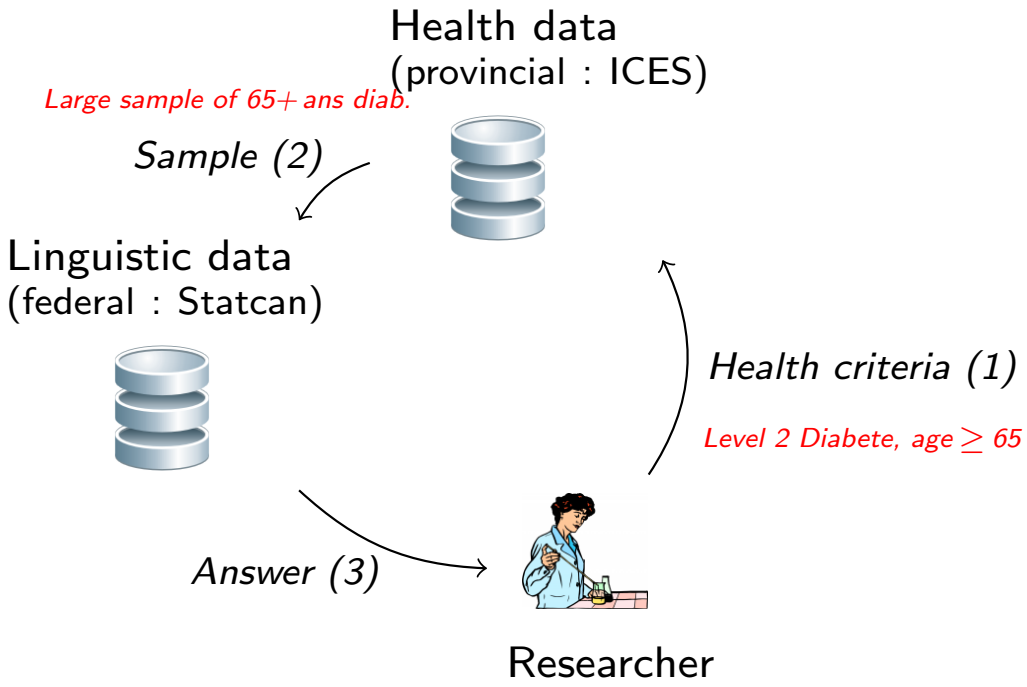


Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Context & Motivation

Suggestion 1

(Tripartite tabulation)

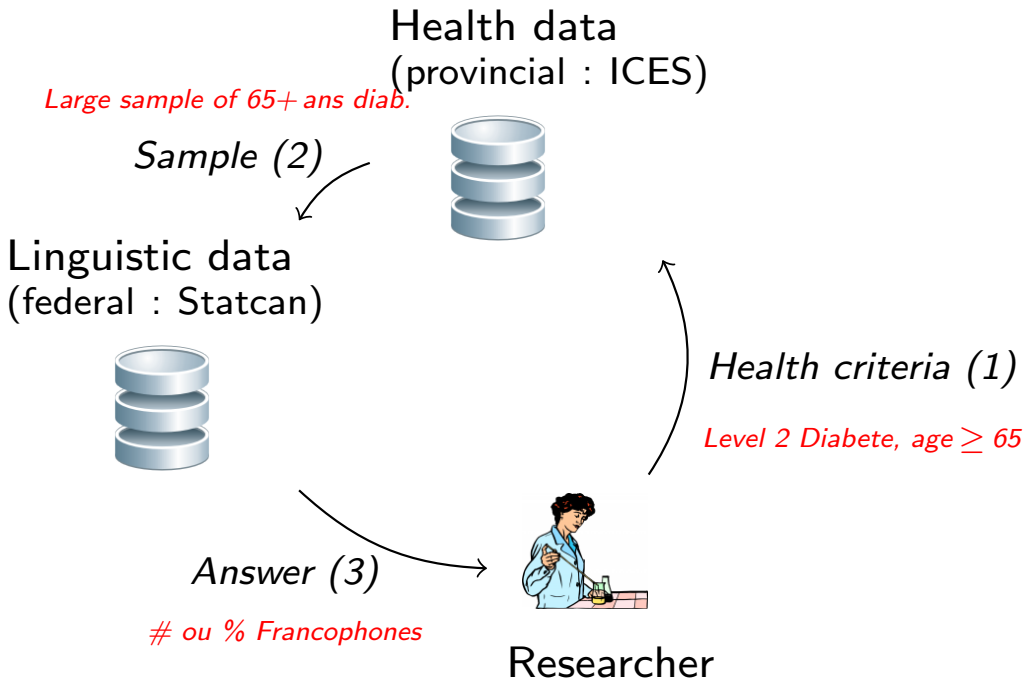


Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Context & Motivation

Suggestion 1

(Tripartite tabulation)

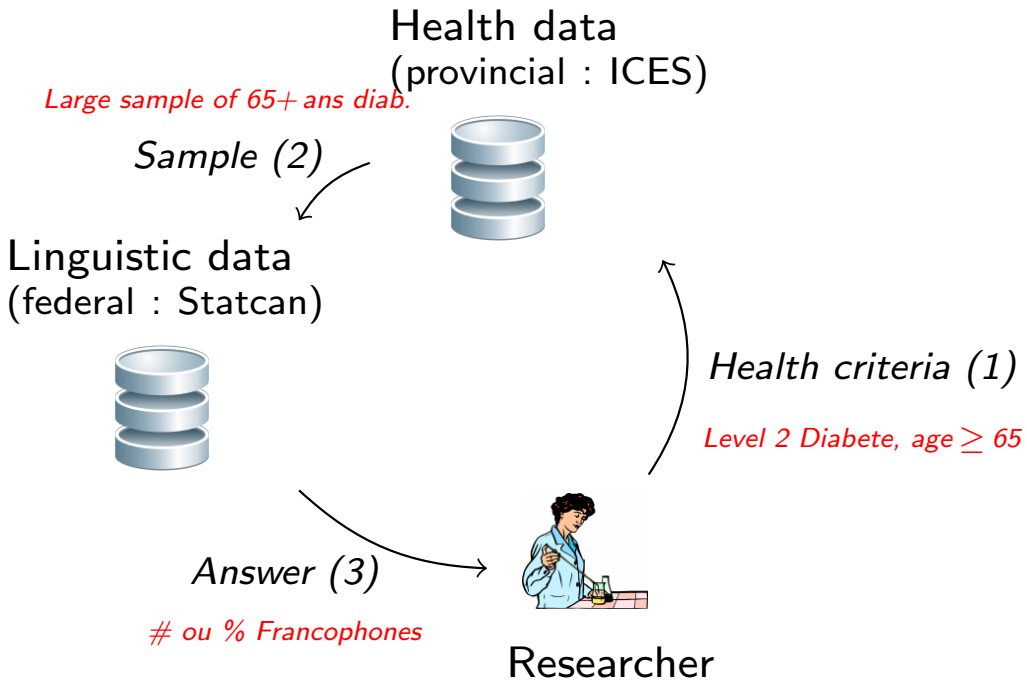


Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Context & Motivation

Suggestion 1

(Tripartite tabulation)

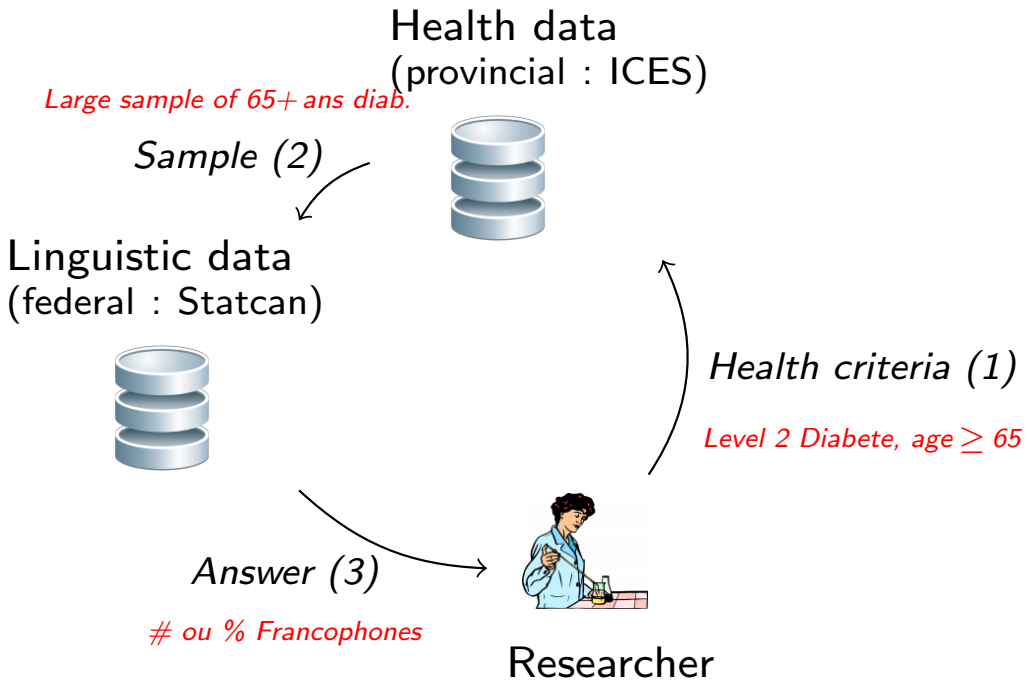


Normalization over # Franco. in Ontario $\Longrightarrow$ *Answer*

Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Context & Motivation

Suggestion 1
(Tripartite tabulation)



Normalization over # Franco. in Ontario $\Longrightarrow$ *Answer*

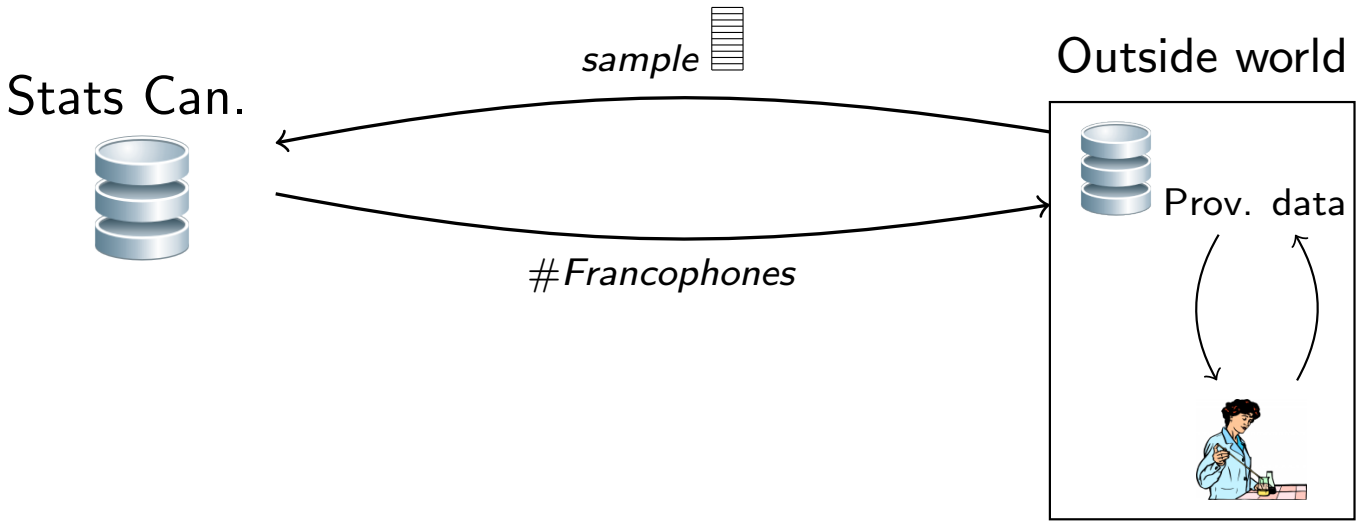
Ex : Rate of type-2 diabete among 65+ Francophones (vs. 65+ Anglophones) ?

Unfortunately..

"There is no such thing as a *tri*-partite tabulation !"
(from Statcan point of view)

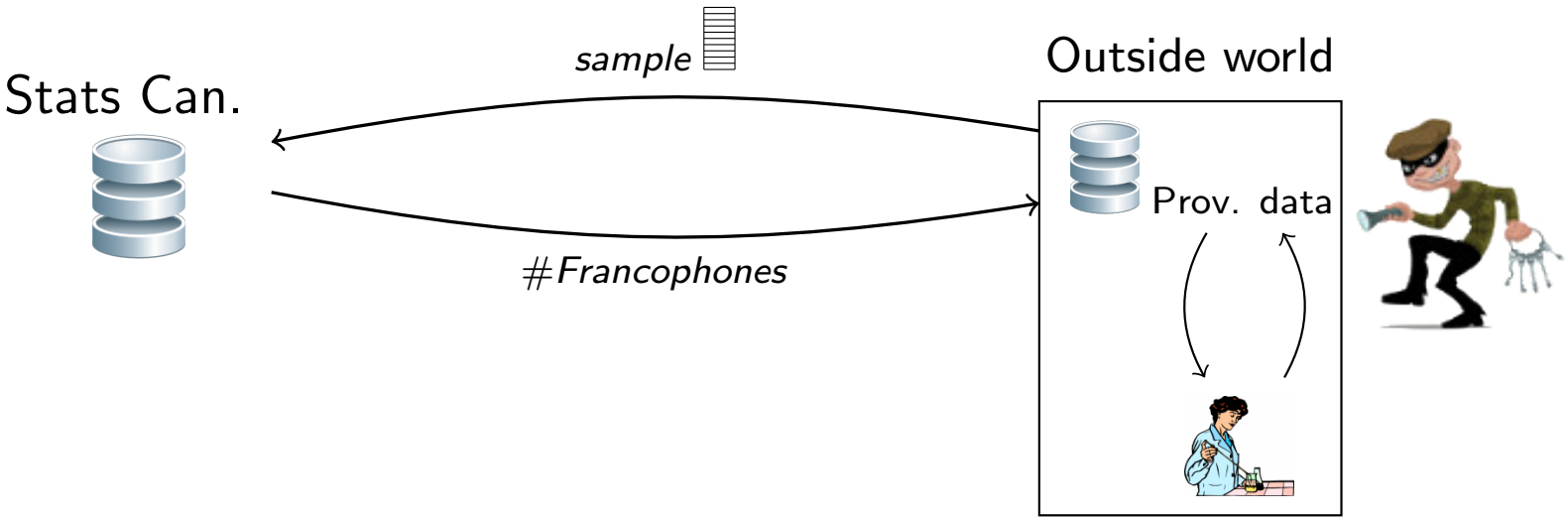
Statistics Canada point of view

This is a type of *tabulation*.



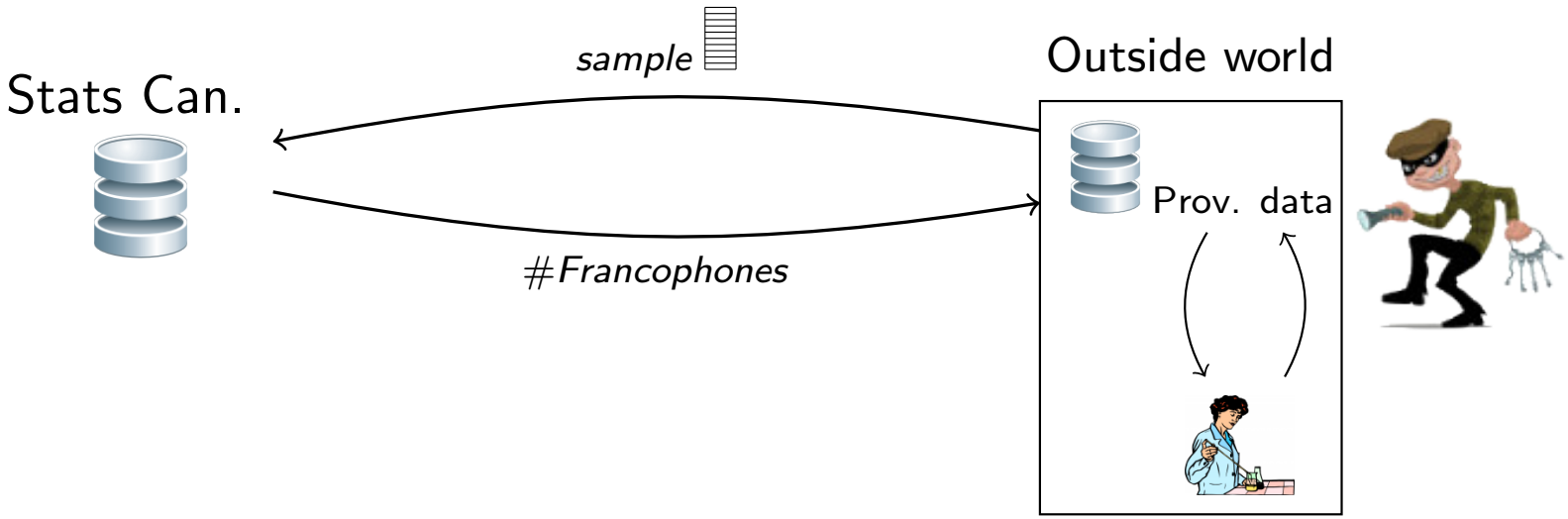
Statistics Canada point of view

This is a type of *tabulation*.



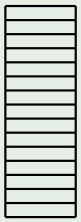
Statistics Canada point of view

This is a type of *tabulation*.



Problem..

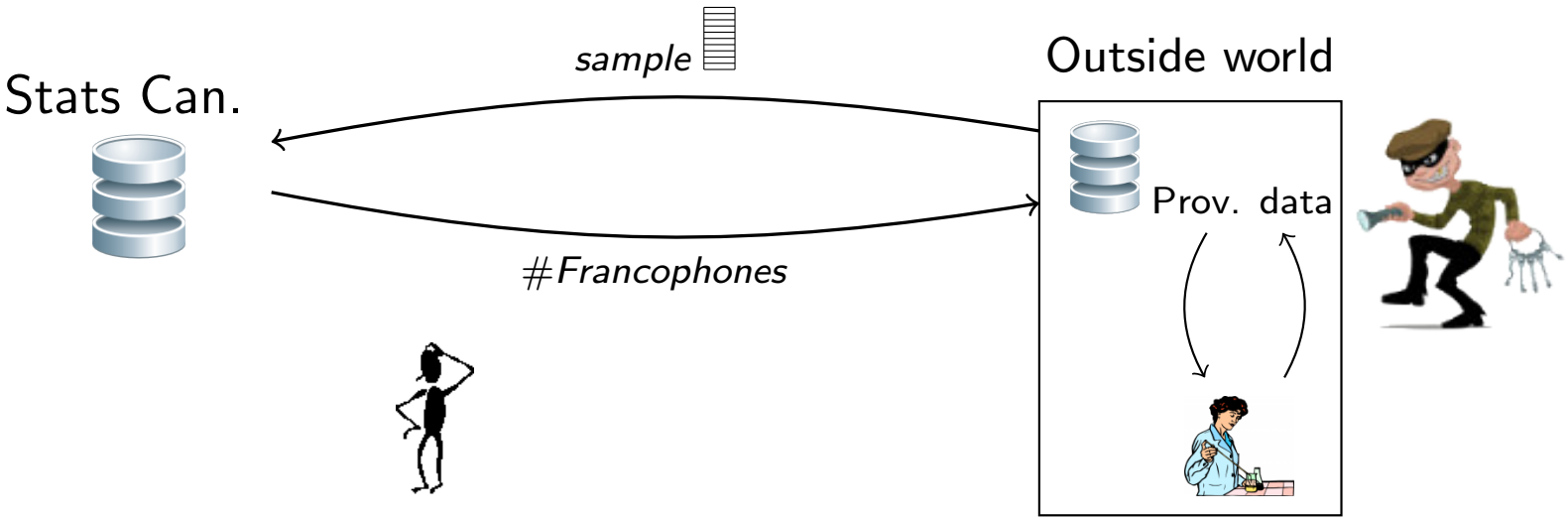
Collection of residual information by an adversary possible, e.g. :

$s_1 =$  $s_2 =$  x

$$answer(s1) - answer(s2) \implies language(x)$$

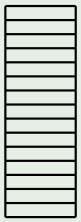
Statistics Canada point of view

This is a type of *tabulation*.



Problem..

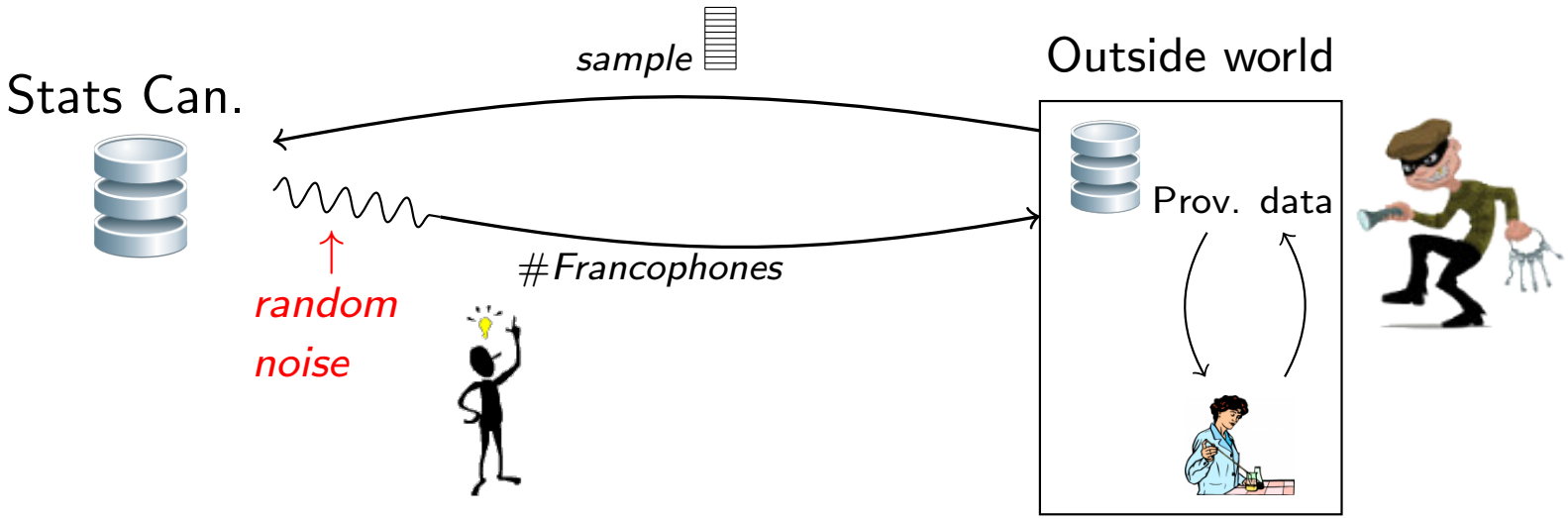
Collection of residual information by an adversary possible, e.g. :

$s_1 =$  $s_2 =$  x

$$answer(s1) - answer(s2) \implies language(x)$$

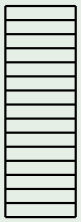
Statistics Canada point of view

This is a type of *tabulation*.



Problem..

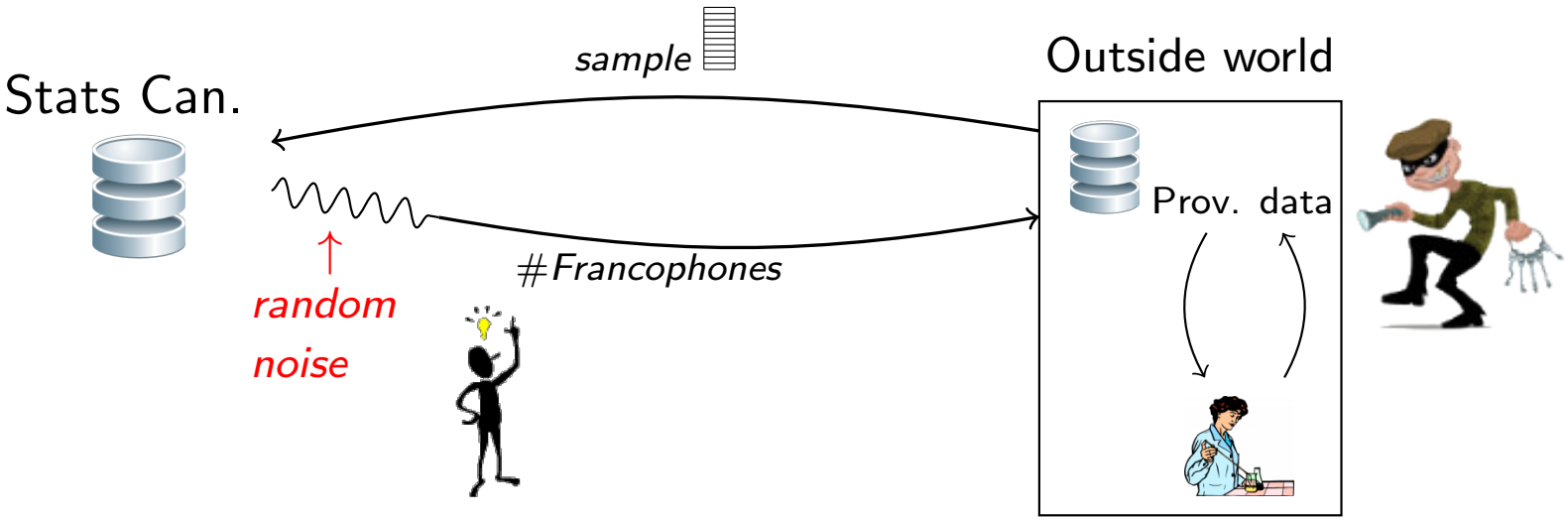
Collection of residual information by an adversary possible, e.g. :

$s_1 =$  $s_2 =$  x

$$answer(s1) - answer(s2) \implies language(x)$$

Statistics Canada point of view

This is a type of *tabulation*.



Problem..

extremely limited

Collection of residual information by an adversary ~~possible~~, e.g. :

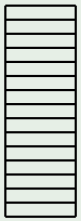
$s_1 =$ $s_2 =$ x

$answer(s1) - answer(s2) \implies P(language(x))$

Understanding the noise

Problem.. extremely limited

Collection of residual information by an adversary ~~possible~~, e.g. :

$s_1 =$  $s_2 =$  x

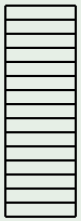
$$answer(s1) - answer(s2) \implies P(language(x))$$

Understanding the noise

Problem..

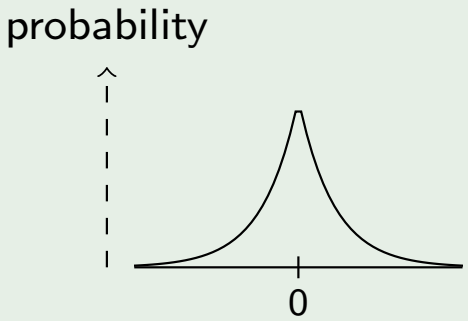
extremely limited

Collection of residual information by an adversary ~~possible~~, e.g. :

$s_1 =$  $s_2 =$  x

$answer(s1) - answer(s2) \implies P(language(x))$

Random noise



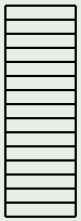
(a) Random number

Understanding the noise

Problem..

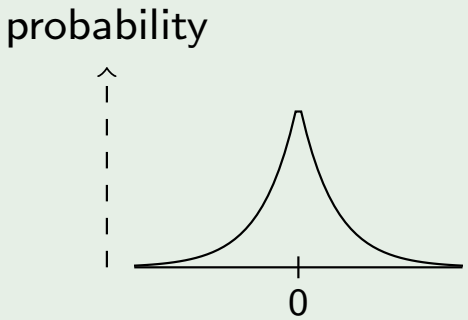
extremely limited

Collection of residual information by an adversary ~~possible~~, e.g. :

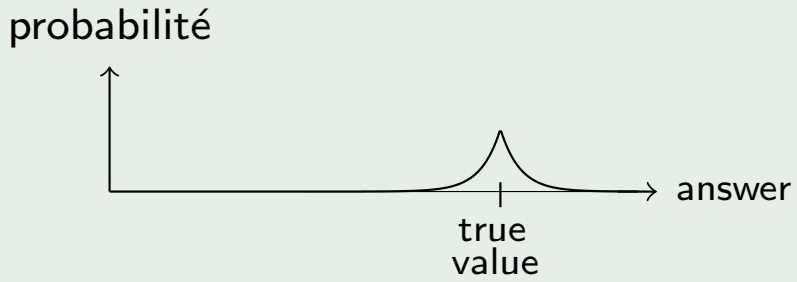
$s_1 =$  $s_2 =$  x

$answer(s1) - answer(s2) \implies P(language(x))$

Random noise



(a) Random number

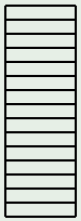


(b) Probability (answer)

Understanding the noise

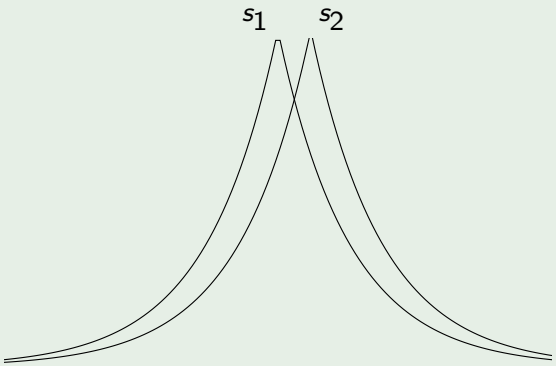
Problem.. extremely limited

Collection of residual information by an adversary ~~possible~~, e.g. :

$s_1 =$  $s_2 =$  x

$$answer(s1) - answer(s2) \implies P(language(x))$$

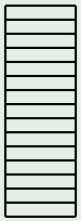
Random noise



Understanding the noise

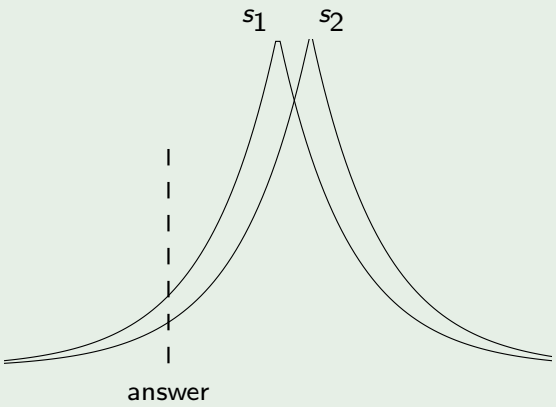
Problem.. extremely limited

Collection of residual information by an adversary ~~possible~~, e.g. :

$s_1 =$  $s_2 =$  x

$$answer(s1) - answer(s2) \implies P(language(x))$$

Random noise

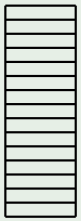


Understanding the noise

Problem..

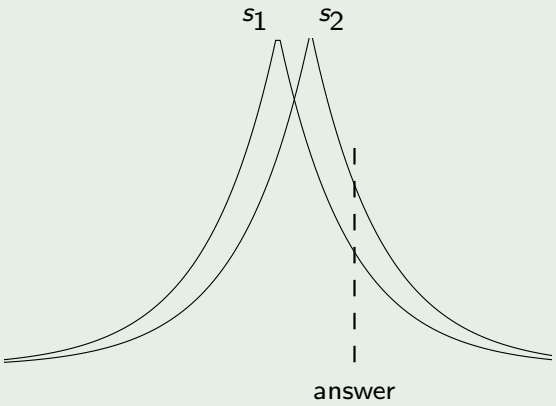
extremely limited

Collection of residual information by an adversary ~~possible~~, e.g. :

$s_1 =$  $s_2 =$  x

$$answer(s1) - answer(s2) \implies P(language(x))$$

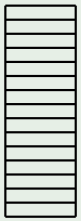
Random noise



Understanding the noise

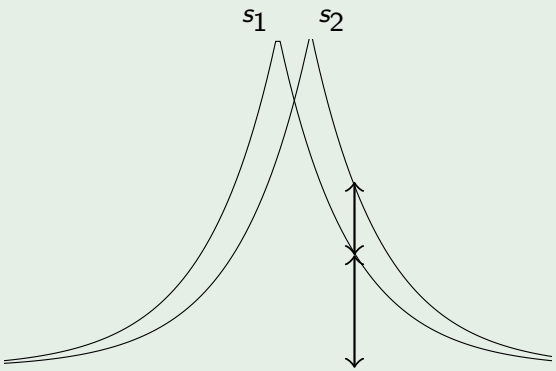
Problem.. extremely limited

Collection of residual information by an adversary ~~possible~~, e.g. :

$s_1 =$  $s_2 =$  x

$$answer(s1) - answer(s2) \implies P(language(x))$$

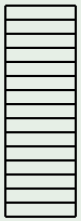
Random noise



Understanding the noise

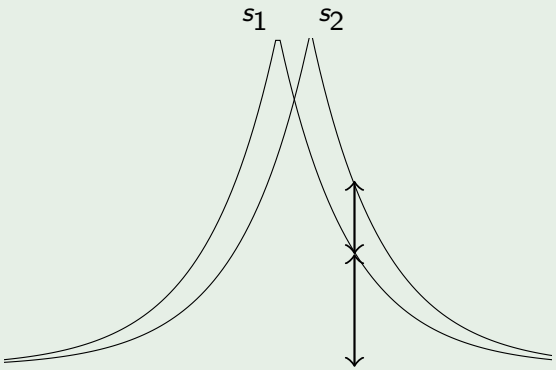
Problem.. extremely limited

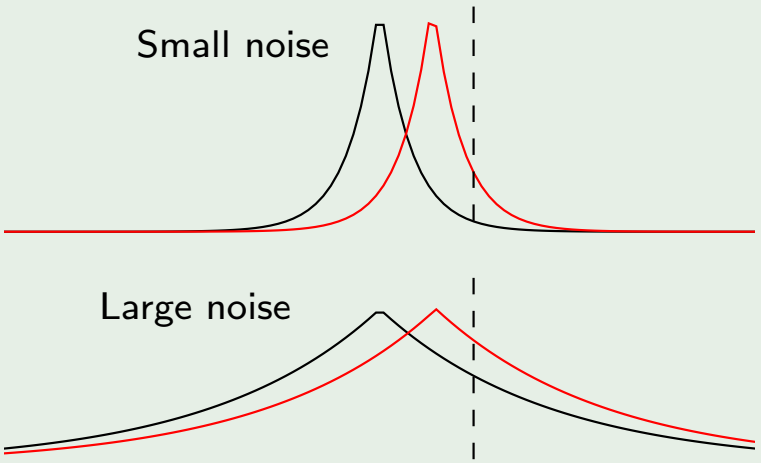
Collection of residual information by an adversary ~~possible~~, e.g. :

$s_1 =$  $s_2 =$  x

$$answer(s1) - answer(s2) \implies P(language(x))$$

Random noise

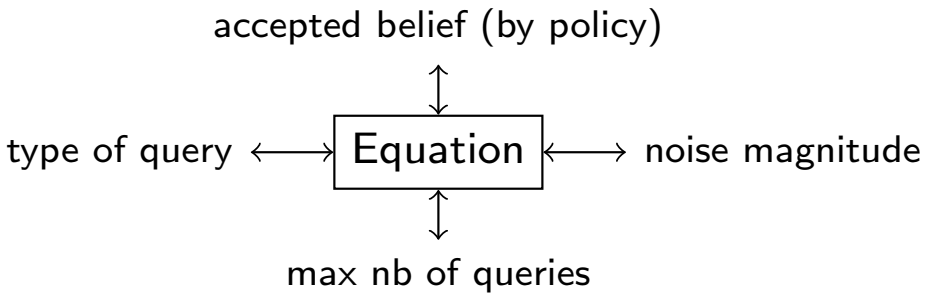




Setting up the tradeoff

"Calibrating noise to sensitivity in private data analysis."

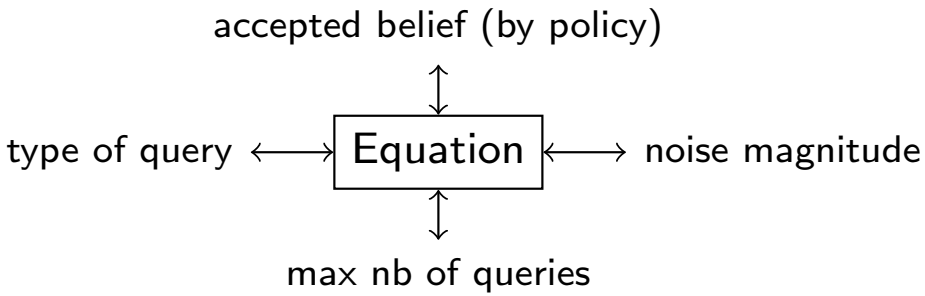
(Dwork *et. al*, Theory of Cryptography, 2006)

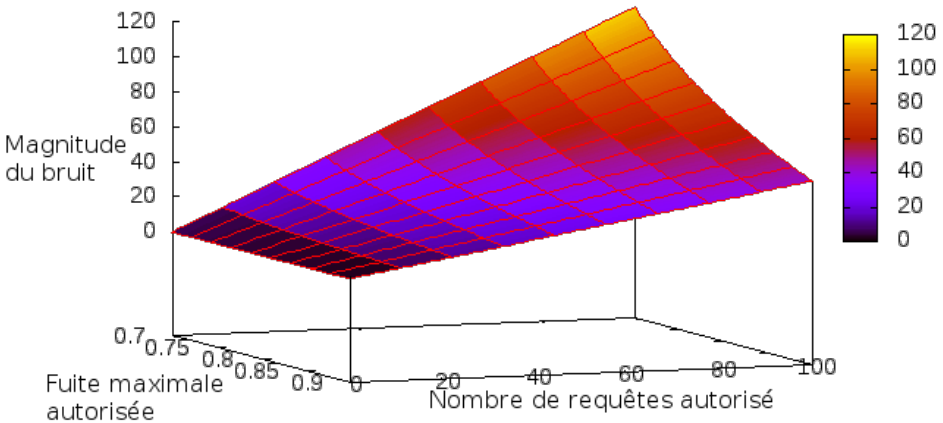


Setting up the tradeoff

"Calibrating noise to sensitivity in private data analysis."

(Dwork *et. al*, Theory of Cryptography, 2006)

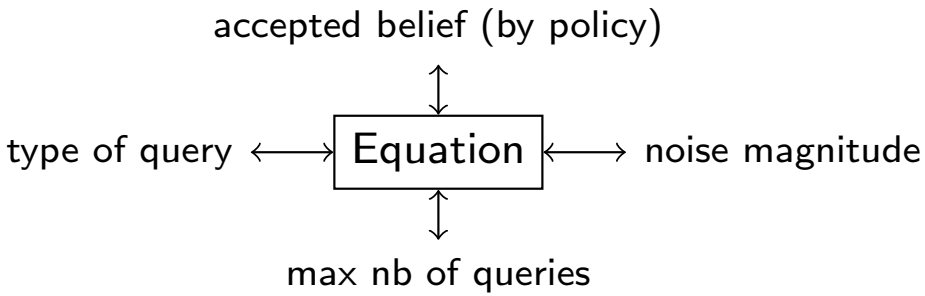


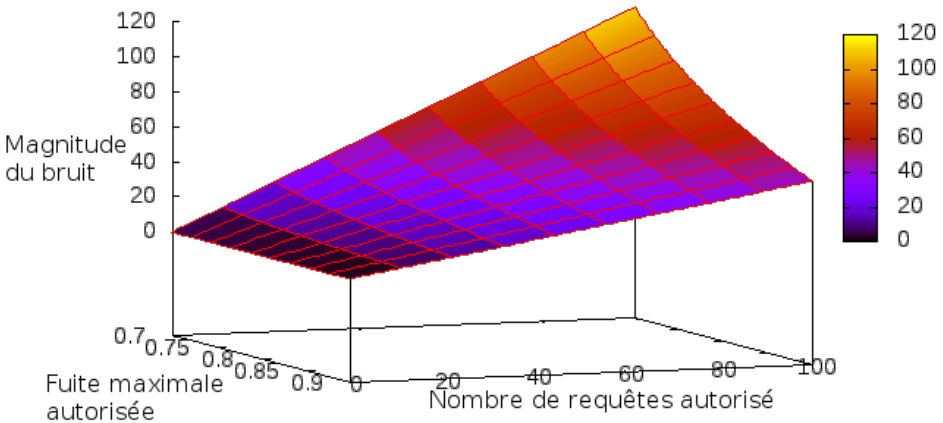


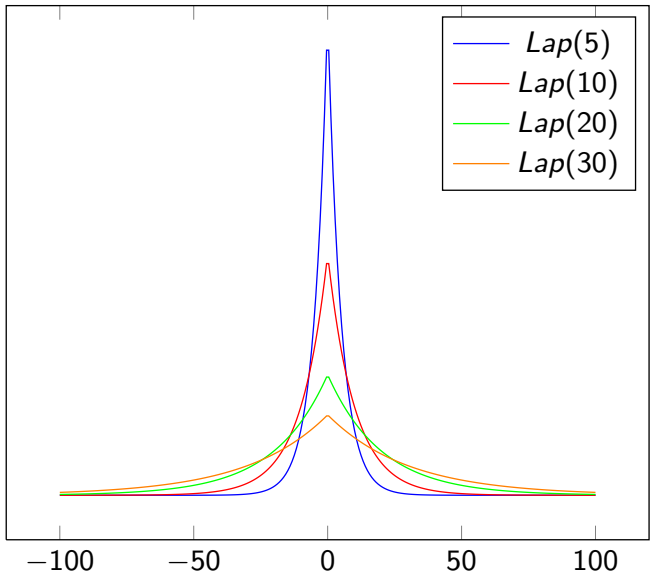
Setting up the tradeoff

"Calibrating noise to sensitivity in private data analysis."

(Dwork *et. al*, Theory of Cryptography, 2006)



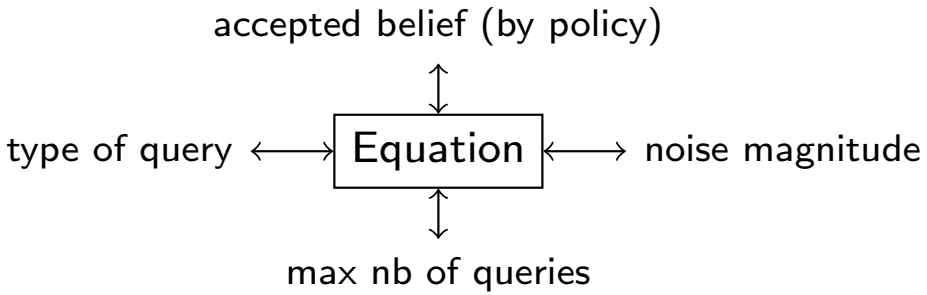


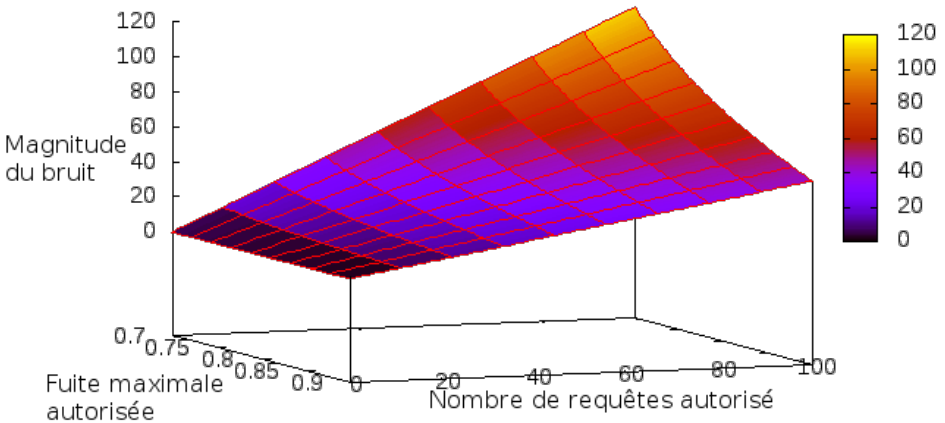


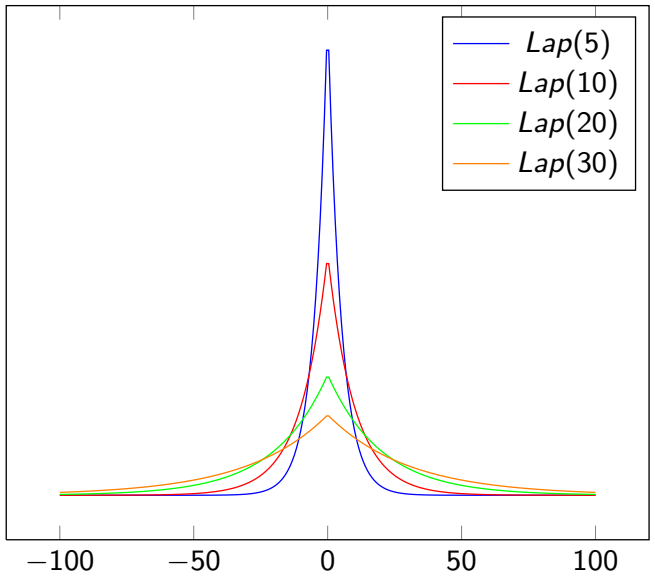
Setting up the tradeoff

"Calibrating noise to sensitivity in private data analysis."

(Dwork *et. al*, Theory of Cryptography, 2006)



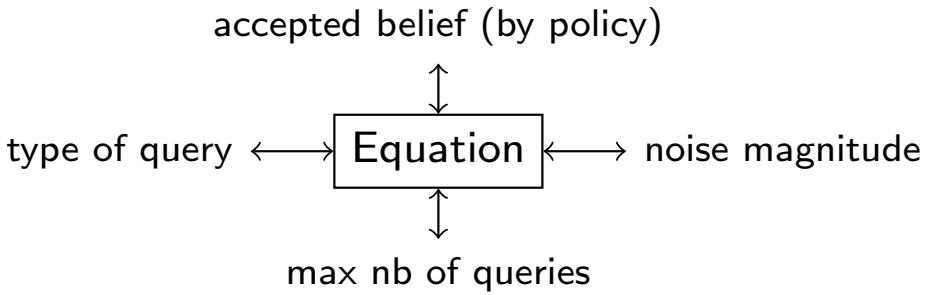


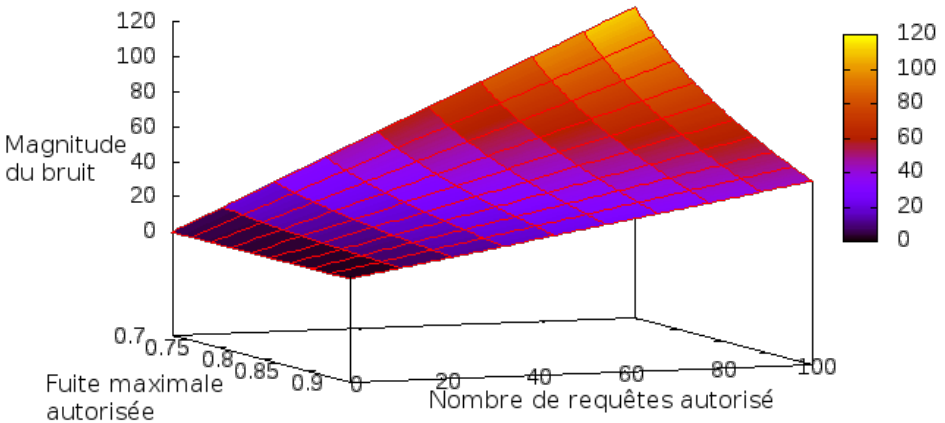


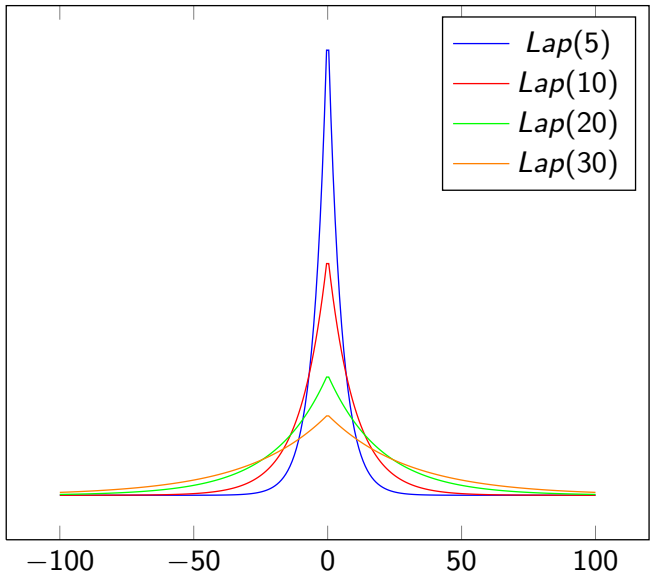
Setting up the tradeoff

"Calibrating noise to sensitivity in private data analysis."

(Dwork *et. al*, Theory of Cryptography, 2006)







Questions

Open questions

- What level of noise is acceptable to researchers ?

Questions

Open questions

- What level of noise is acceptable to researchers ?
- What leakage would be generated by a “honest” use ?

Questions

Open questions

- What level of noise is acceptable to researchers ?
- What leakage would be generated by a “honest” use ?
- Other types of queries ? (e.g. histograms)

Questions

Open questions

- What level of noise is acceptable to researchers ?
- What leakage would be generated by a “honest” use ?
- Other types of queries ? (e.g. histograms)

Perspectives

- Seamingly extending data of provincial agencies

Questions

Open questions

- What level of noise is acceptable to researchers ?
- What leakage would be generated by a “honest” use ?
- Other types of queries ? (e.g. histograms)

Perspectives

- Seamingly extending data of provincial agencies
- Other variables than language

Questions

Open questions

- What level of noise is acceptable to researchers ?
- What leakage would be generated by a "honest" use ?
- Other types of queries ? (e.g. histograms)

Perspectives

- Seamingly extending data of provincial agencies
- Other variables than language
- How could such a mechanism be implemented ?

References

This work :

Enabling Dynamic Linkage of Linguistic Census Data at Statistics Canada.

A. Casteigts, M.-H. Chomienne, L. Bouchard, G.-V. Jourdan

Technical Repport, RRASFO, 2011

The theoretical result we have used :

Calibrating Noise to Sensitivity in Private Data Analysis

C. Dwork and F. McSherry and K. Nissim and A. Smith

Theory of Cryptography, 2006

Thank you !

Institut de recherche de l'Hôpital **Montfort**

uOttawa

L'Université canadienne
Canada's university