Challenges in Privacy Management for Big Data Systems

Yennun Huang 黃彥男
CEO, TWISC AS

The Basic Problem for Data Science

How to do de-identification to

- Enable “desirable uses” of the data while protecting the “privacy” of the data subjects?
  - Political policy
  - Academic research
  - Study drug trial
  - Security: searching for terrorists/crimials
  - Market analysis, ….

Approach 1: Encrypt the Data

Name	Sex	Blood	…
Jane	F	B	…
Perry	M	A	…
Smith	M	O	…
Ross	M	O	…
Huang	F	A	…
Chen	M	B	…

Name	Sex	Blood	…
100101	001001	110101	…
101010	111010	111111	…
001010	100100	011001	…
001110	010010	110101	…
110101	000000	111001	…
111110	110010	000101	…

Problems： Data cannot be analyzed.

Approach 2: Anonymize the Data

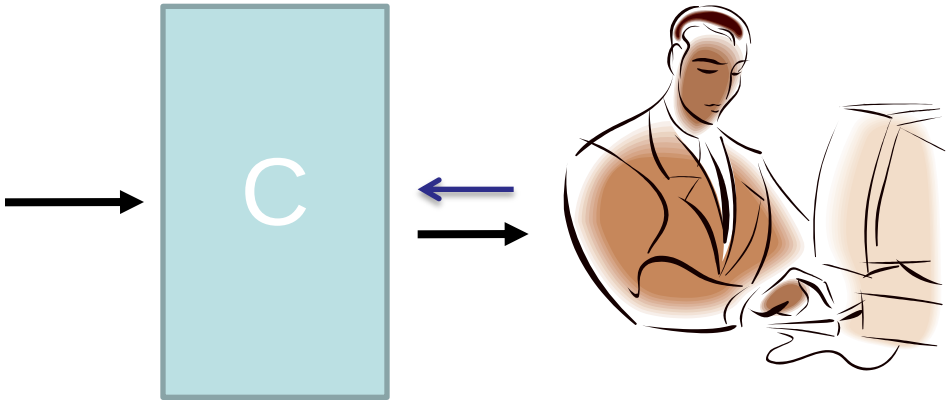
Name	Sex	Blood	…	HIV?
[redacted]n	F	B	…	Y
J[redacted]s	M	A	…	N
S[redacted]h	M	O	…	N
[redacted]s	M	O	…	Y
[redacted]	F	A	…	N
S[redacted]h	M	B	…	Y

Problems: "re-identification", linking data

[Sweeney `97]

Approach 3: Mediate Access

Name	Sex	Blood	…
Jane	F	B	…
Perry	M	A	…
Smith	M	O	…
Ross	M	O	…
Huang	F	A	…
Chen	M	B	…



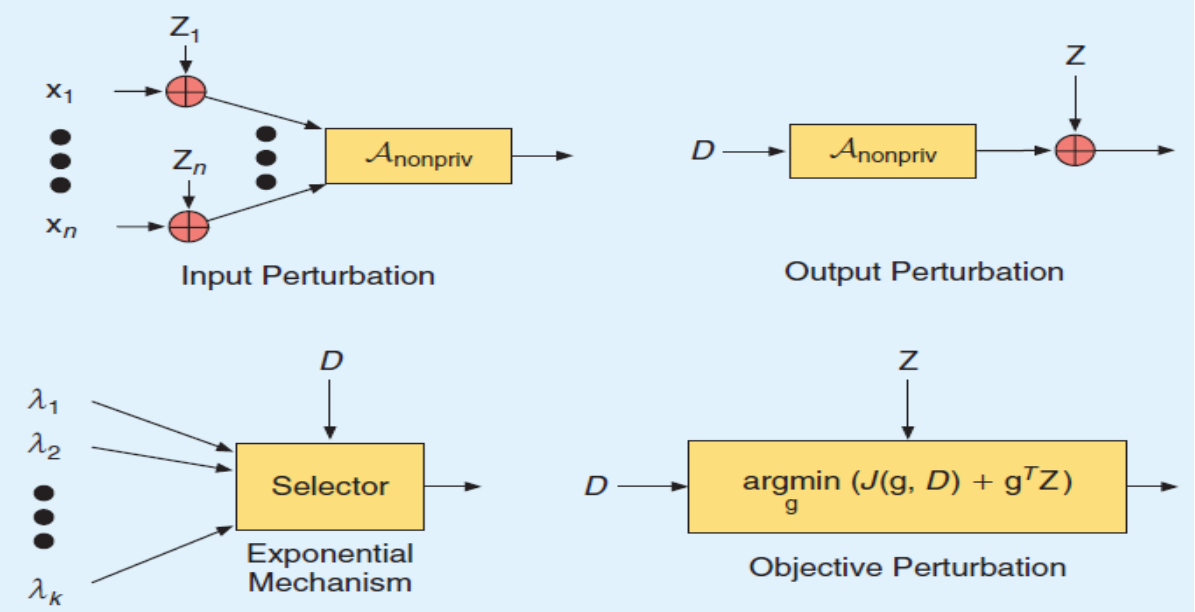
trusted
"curator"

data analysts

Problems : "aggregated" statistics can reveal individual information; query selections

Differential privacy

Adding noise to data or query: D -> D'



Synthetic data

Sex	Blood	…	Cancer?
F	B	…	Y
F	A	…	N
M	O	…	N
M	O	…	Y
F	A	…	N
M	B	…	Y

Sex	Blood	…	Cancer?
M	B	…	N
F	B	…	Y
M	O	…	Y
M	A	…	N
F	O	…	N

"fake" people

Utility: preserves statistics with *every* set of attributes!

Problem: computation time

Privacy Standards: early stage

- ISO/IEC 29100: privacy framework
- ISO/IEC 29191: requirements for partially anonymous and partially unlinkable data

Privacy Challenges in Big Data

- Linking two or more data sets
  - How to link?
  - Compromised privacy?
- Streaming data, unstructured data
- Policy and standards