Discussione di

GDPR e Ricerca: tra vincoli e potenzialità

Daniele Checchi (Università di Milano)

L’informazione è un bene pubblico: costa produrla – i benefici ricadono sul produttore e anche su altri (esternalità)

Le esigenze del GDPR → lasciare all’individuo la libertà di scelta nel rivelare o meno informazioni che lo riguardano e che possono essere frutto di scelte personali

Le esigenze della ricerca → caratterizzare nel modo più preciso possibile il comportamento degli individui, facendo leva sul numero massimo di informazioni disponibili.

Esiste un conflitto intrinseco tra queste due esigenze, che può essere composto da un sistema di regole chiare.

L’interpretazione italiana del GDPR e la sua applicazione del Garante sono sbilanciate a favore della tutela dell’individuo e ignorano le esigenze della ricerca.

Tre esempi:
① carriere scolastiche e lavorative → identificativi individuali → effetti fissi
② incroci banche dati → rischio di reidentificabilità
③ banche dati amministrative → procedure ed errori

① carriere scolastiche o lavorative

Ieri al CNEL presentazione e discussione indagine PIAAC.

"*Come curare la potenziale endogeneità ?*
① la soluzione migliore sarebbe misurare le competenze da piccoli (prima delle scelte scolastiche – per esempio test invalsi di terza media) e cercarne gli effetti vent'anni dopo.

Fattibilità ? tecnicamente possibile (primo Invalsi 2009-10) – ricerca degli stessi individui negli archivi ANS (anagrafe nazionale studenti – MIM gestito da Cineca), COB (comunicazioni obbligatorie nel momento dell'assunzione - Ministero del Lavoro) o (contribuzione previdenziale Uniemens - INPS)".

Cosa osta ?
⇨ consenso individuale – impossibile su larga scala
⇨ finalità amministrativa – monitoraggio assolvimento obbligo scolastico ? NEET ?
⇨ collaborazione interministeriale – molteplicità di fornitori

Per contro perché permetterebbe un salto negli studi sulle carriere ?

Modelli longitudinali (lo stesso individuo osservato ripetutamente) permettono di tener conto della eterogeneità non osservabile (capacità innata, intelligenza, carattere, ecc) caratterizzando i comportamenti al netto della stessa → come se gli individui diventassero tutti uguali → esattamente l'opposto della privacy individuale.

Esempio: se posso osservare la retribuzione individuale mensile per **molti** anni di **molti** individui in **molte** aziende, posso stimare la media retributiva individuale delle persone e delle aziende.

Al netto di questi effetti fissi individuali, il resto sono fenomeni comuni a tutti gli individui, che sono l'interesse dei ricercatori:
⇨ rendimento dell'istruzione
⇨ profili di carriera
⇨ discriminazione

Ma richiede di processare moltissima informazione formalmente protetta da privacy.

Via d'uscita: permettere accesso remoto all'informazione pseudonomizzata.
Ma chi controlla il ricercatore ? ispezione manuale degli outcome statistici ?

② incrocio banche dati

Incrociare banche dati permette la profilazione dei comportamenti (esempio: chi prende brutti voti diventa disoccupato).

Regola d'oro della protezione della privacy → impedire la reidentificazione in ciascuna banca dati originaria assicurando la presenza di almeno un numero congruo di individui identici (quanti "cloni" sono necessari per assicurare la protezione: 3 ? 5 ? 10 ?)

Esempio ricerca sui gemelli: dalla retribuzione giornaliera al decile decennale nella distribuzione del reddito.

Ma il ricercatore perde ricchezza informativa enorme.

Come salvare capra e cavolo ?
Fidandosi della deontologia professionale del ricercatore e sanzionando collettivamente i comportamenti scorretti (sospensione dell'accesso ai dati del dipartimento o dell'università di appartenenza del ricercatore scorretto piuttosto che sanzioni monetarie).

③ banche dati amministrative

Le banche dati amministrative non sono ideali per la ricerca:
⇨ raccolgono solo le informazioni utili alla gestione della procedura che normalmente è sconosciuta nei dettagli al ricercatore.
⇨ costruite da informatici molto diversi che usano codifiche diverse e protocolli diversi.
⇨ non hanno informazione sui non compliers

Per contro la potenza della dimensione: dai campioni alla popolazione (quasi) intera.

Occorrerebbero dei "mediatori" interni (data scientists) che
- ✓ conoscono le procedure amministrative
- ✓ capiscono le domande di ricerca

e redigano dei codebook.

A partire dalla pulitura delle banche dati amministrative emergono tipicamente errori → le amministrazioni tipicamente non apprezzano questo feedback → si autoproteggono negando accesso ai dati.

Come superare questo empasse ? Costruendo file standard corrispondenti all'attività amministrativa svolta e mettendoli a disposizione dei ricercatori.

Il file standard può rappresentare una mediazione tra amministrazione e ricercatore con la "benedizione" del DPO (purchè illuminato).

La sua costruzione richiede
a) individuazione e selezione delle variabili rilevanti
b) definizione della codifica delle variabili selezionate
c) la calibratura del grado di copertura (5% - 10% - intera popolazione)
d) definizione di forme di feedback tra amministratori e ricercatori
e) definizione di domande di ricerca concordate tra amministratori e ricercatori che autorizzino il rilascio dei dati.

Esempi:
1) file standard della VQR
2) file standard imprese di INPS