Dati aperti, privacy e aspetti etici, quali soluzioni?

Francesco Vespignani, Giulia Calignano

DPSS, Università degli Studi di Padova

Prospettive interdisciplinari nella misura di competenze e capacità linguistiche in età scolare

Università degli Studi di Padova -25 giugno 2024

Tanti dati

E' meglio un grande corpus rumoroso o un piccolo corpus controllato?

The bigger the better.

E' meglio una misura accurata svolta in laboratorio o un'ampia mole di dati raccolti online?

(who knows)

Dati rumorosi

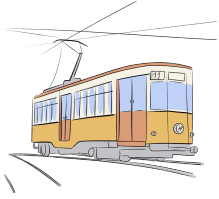
Josh de Leeuw, creatore di JsPsych (seminario online durante il covid).

I ricercatori vogliono riprodurre quanto più possibile il massimo di controllo delle variabili sperimentali ottenibile in laboratorio anche negli esperimenti online. Questo non è possibile, è importante invece sviluppare paradigmi online i cui effetti non siano sensibili alle variabili difficili da controllare online (dimensione/font/luminosità).

Nonostante questo JsPysch fornisce strumenti per massimizzare il controllo di queste variabili, p.e. virtual-chainrest

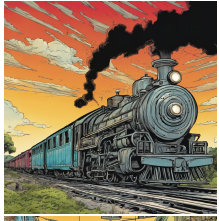
Dati diversi

Tocca il tram.



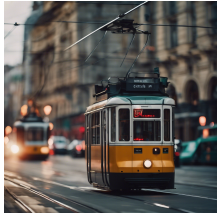
Dati diversi

Tocca il tram.



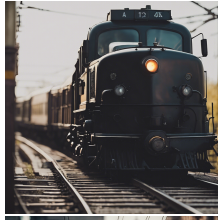
Dati diversi

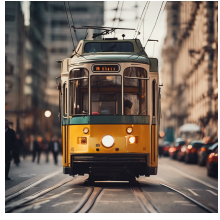
Tocca il tram.

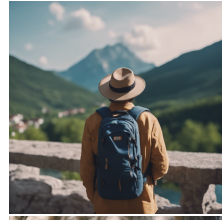


Dati diversi

Tocca il tram.







Dati diversi

Uno psicologo sperimentale (topo di laboratorio) si aspetta accuratezze diverse (in funzione dell'età) nei differenti compiti.

Se il compito è fatto online o al computer, se c'è interazione o meno può dar luogo a ulteriori differenze. E' rilevante il contesto, la stanchezza, concentrazione e motivazione (contesto/setting).

E' però anche possibile che l'età alla quale un bambino conosce la differenza fra tram e treno (e la variabilità di tale età) sia largamente indipendente da questi dettagli sperimentali.

Dati diversi

In tal senso il testing in svariate condizioni, situazioni e modalità può dar luogo a dati **più solidi e affidabili**.

A livello sperimentale la logica **multilab** nella raccolta dati e **multiverse** nella loro analisi statistica vanno in questa direzione.

Chiaramente lo studio dell'effetto di variabili accessorie non può essere condotto da un singolo soggetto con classici disegni bilanciati.

Questo tipo di domande ha bisogno di collaborazioni extra-professionali (psicologi professionisti, ricercatori, logopedisti, logogenisti, insegnanti, educatori, genitori) all'interno di una logica *citizen-science*.

Open data **VERSUS** Privacy, GDPR, norme etche e deontologiche

AIP

Codice etico AIP (art. 1.4,g): *la possibilità, se prevista, che i dati, resi completamente anonimi, siano resi disponibili per ulteriori ricerche, anche con finalità diverse rispetto alla ricerca originale, sotto la responsabilità della persona che coordina la ricerca.*

completamente anonimi

REGOLAMENTO (UE) 2016/679, considerando (26):

I dati personali sottoposti a pseudonimizzazione, i quali potrebbero essere attribuiti a una persona fisica mediante l'utilizzo di ulteriori informazioni, dovrebbero essere considerati informazioni su una persona fisica identificabile. Per stabilire l'identificabilità di una persona è opportuno considerare tutti i mezzi, come l'individuazione, di cui il titolare del trattamento o un terzo può ragionevolmente avvalersi per identificare detta persona fisica direttamente o indirettamente. (omissis) I principi di protezione dei dati non dovrebbero pertanto applicarsi a informazioni anonime, vale a dire informazioni che non si riferiscono a una persona fisica identificata o identificabile o a dati personali resi sufficientemente anonimi da impedire o da non consentire più l'identificazione dell'interessato. Il presente regolamento non si applica pertanto al trattamento di tali informazioni anonime, anche per finalità statistiche o di ricerca.

completamente anonimi

Più precisamente, i dati devono essere trattati in maniera tale da non poter più essere utilizzati per identificare una persona fisica utilizzando "l'insieme dei mezzi che possono essere ragionevolmente utilizzati" dal responsabile del trattamento o da altri. Un fattore importante è che il trattamento deve essere irreversibile. (0829/14/IT - WP216)

Il Codice AIP richiede la conservazione dei dati per 5 anni (Regole deontologiche per trattamenti a fini statistici o di ricerca scientifica).

Il diritto alla restituzione e alla cancellazione rende difficile la *distruzione* anticipata dei dati personali?

Dati **(completamente) anonimi** possono essere divulgati in forma non aggregata prima di 5 anni dalla raccolta?

Born open data

All'interno dello sviluppo di best practice dell'open-science una proposta è che i dati siano resi immediatamente aperti.

Abbiamo realizzato un esperimento utilizzando data-pipe (in connessione con JSPsych), che consente di raccogliere dati online utilizzando server pubblici/commerciali (github) che salava direttamente i dati in un repository OSF (open science foundation).

A garanzia della riservatezza i dati sono pseudoanonimizzati usando un codice casuale. Inoltre i dati che permettono di ricondurre il dato a una persona fisica (genere e età) sono salvati in un server Jatos del DPSS protetto da password.

Pseudo/anonimizzazione (completa)

In un momento successivo alla raccolta dati i dati sensibili (che consentono l'individuazione della persona) potrebbero essere depositati in un file aggiuntivo:

code	sex	age
FSBVXT	m	32
FIOVXT	f	32
FSPPOT	f	27
UTRVXT	m	18
KKOVXT	m	19

Chiunque sappia che una donna di 32 anni ha partecipato all'esperimento può ricondurre i dati all'individuo. I dati non sono anonimizzati semplicemente cancellando ogni dato di contatto.

Pseudo/anonimizzazione (completa)

In un momento successivo alla raccolta i dati che consentono l'individuazione della persona potrebbero essere depositati in un file aggiuntivo:

code	sex	age
FSBVXT	f	32
FIOVXT	f	32
FSPPOT	m	18
UTRVXT	m	18
KKOVXT	m	18

In questo caso i dati non permettono l'individuazione univoca (k-anonimato). Nonostante questo, se tutti i maschi di 18 anni hanno avuto la medesima performance il dato può fornire informazioni su qualsiasi maschio di diciotto anni che ha partecipato all'esperimento.

Pseudo/anonimizzazione (completa)

Alternative:

code	sex	age range
FSBVXT	NA	30-40
FIOVXT	NA	30-40
FSPPOT	NA	18-30
UTRVXT	NA	18-30
KKOVXT	NA	18-30
....	...	...

Chiaramente maggiore è il numero di soggetti più smascheramento è possibile.

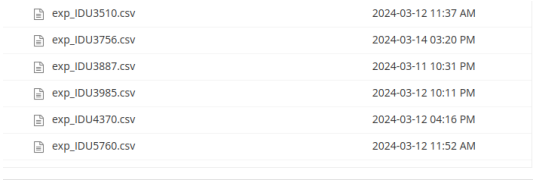
Per un overview delle tecniche di anonimizzazione (aggiunta di rumore statistico, k-anonimato, permutazioni, etc.) si veda Parere 05/2014 sulle tecniche di anonimizzazione.

Pseudo/anonimizzazione (completa)

Smascheramento progressivo.

Chiaramente i dati divengono veramente anonimi solo **cancellando** i dati di contatto, i dati personali grezzi, i codici di pseudoanonimizzazione originali.

Data e ora di esecuzione dell'esperimento?



exp_IDU3510.csv	2024-03-12 11:37 AM
exp_IDU3756.csv	2024-03-14 03:20 PM
exp_IDU3887.csv	2024-03-11 10:31 PM
exp_IDU3985.csv	2024-03-12 10:11 PM
exp_IDU4370.csv	2024-03-12 04:16 PM
exp_IDU5760.csv	2024-03-12 11:52 AM

Chiunque sappia in che giorno e a che ora l'esperimento è stato fatto può individuare quali siano i dati relativi alla persona.

Pseudo/anonimizzazione (completa)

La procedura consente il rispetto della privacy e delle buone pratiche open-data ma:

- non rispettando alla lettera i codici (p.e. AIP)
- perdendo dei dati

Una cosa positiva è che **tutto cospira** verso la raccolta di grandi quantità di dati. Alla fine, se raccolgo solo 20 partecipanti, è così importante sapere se una persona ha 19, 18 o 20 anni?

I have a dream

Piattaforma condivisa

Deposito di dati anonimizzati relativi a un numero limitato di *compiti*:

- decisione lessicale
- self paced reading (o maze task)
- matching frase-figura
- denominazione di figure
- ...

Ogni operatrice può accreditarsi e ri-usare prove e compiti (sequenze di prove) creati da altre/i ma anche aggiungere materiali linguistici e iconografici propri (o crearli con AI).

Ognuno può decidere quanti e quali dati personali raccogliere.

Piattaforma condivisa

I dati (pseudo)anonimizzati sono resi disponibili (quasi) immediatamente.

subject	trial	acc	rt
HHYDTY	AA443	1	700
FIOVXT	AA473	0	1234
FSPPOT	AA679	1	956
...	...	...	...

trial	task	word
AA443	LDT	cane
AA473	LDT	acqua
AA679	LDT	gatto
...	...	...

Smascheramento progressivo

subject	age	gender	L1
HHYDTY	5-18	NA	italian
FIOVXT	5-18	0	italian
FSPPOT	5-18	1	foreign
...	...	...	...

Smascheramento progressivo

subject	age	gender	L1
HHYDTY	5-7	m	italian
FIOVXT	8-9	f	italian
FSPPOT	6-9	m	chinese
IIITEC	9-12	m	italian
POGFRC	7-8	f	italian
MIGTES	5-7	f	english
...	...	...	...

Smascheramento progressivo

subject	age	gender	L1	operator
HHYDTY	6	m	italian	psychologist
FIOVXT	8	f	italian	psychologist
FSPPOT	8	m	chinese	logopedist
IIITEC	9	m	italian	researcher
POGFRC	7	f	italian	partent
MIGTES	7	f	english	teacher
...	...	...	...	...

Smascheramento progressivo

subject	age	gender	L1	ISEE	...	operator
HHYDTY	6	m	italian	30-50k	...	psychologist
FIOVXT	8	f	italian	10-20k	...	psychologist
FSPPOT	8	m	chinese	50-80k	...	logopedist
IIITEC	9	m	italian	30-50k	...	researcher
POGFRC	7	f	italian	20-30k	...	partent
MIGTES	7	f	english	10-20k	...	teacher
...	...	...	...	...	...	...

Benefici

Alimentare il database (raccogliere e condividere i dati) rende gli strumenti più potenti.

Vantaggioso riutilizzare task e trials piuttosto che inventarne nuovi, la novità dev'essere fortemente motivata.

Complessità

Cancellazione dei dati personali entro limiti di tempo e secondo regole indipendentemente gestite dai singoli professionisti, compliance formale con codici giuridici.

Stabilire **a-priori** quali variabili smascherare per prime e in che modo (discretizzazione, rumore casuale). Lavoro complesso per gli psicometristi (difficilmente revisionabile in futuro).

Competenze per sviluppo software (setup sperimentale, database, api).

Stabilire policies di copyrigth dei dati e degli strumenti.

Si può fare!