Pratiche scientifiche ai tempi del capitalismo di piattaforma

di Emanuele Cozzo

> Questo è un mondo in cui enormi quantità di dati e la matematica applicata sostituiscono ogni altro strumento che potrebbe essere utilizzato. Fuori con [Out with] ogni teoria del comportamento umano, dalla linguistica alla sociologia. Dimenticate la tassonomia, l'ontologia e la psicologia. Chi sa perché le persone fanno quello che fanno? Il punto è che lo fanno, e possiamo monitorarlo e misurarlo con una fedeltà senza precedenti. Con dati sufficienti, i numeri parlano da soli

Così profetizzava nel 2008 Chris Anderson, all'epoca editore di *Wired*, in un noto e lungamente discusso editoriale della rivista dal titolo definitivo *La fine della Teoria: Il diluvio di dati rende il metodo scientifico obsoleto*[1]. A dieci anni di distanza, immersi come siamo nella retorica trionfale circa l'intelligenza artificiale, algoritmi e big data, la profezia si è solo apparentemente realizzata, mentre quello che si è aperto continua a essere un campo di battaglia.

«L'uomo [sic], da quando esiste, interagisce con la realtà esterna in due modi, percependola e trasformandola, ossia con l'osservazione e la tecnologia». esordiscono Alessandro Della Corte e Lucio Russo in un prezioso libretto di introduzione al metodo scientifico. «La novità essenziale che caratterizza la scienza – proseguono – […] è la creazione di "teorie scientifiche". Si crea cioè un secondo piano di discorso, "teorico", a fianco di quello che descrive gli oggetti concreti»[2].

Se assumiamo questa definizione operativa, la profezia di Anderson riguarda la morte della scienza, più che la fine della Teoria. Nel tipico stile tecnolibertario della rivista californiana, annuncia la fine dell'impero dell'autorità — della Teoria sulla realtà, degli esperti sulle persone comuni — e la nascita di un nuovo mondo che, attraverso il prisma digitale, si rivela *per ciò che è* a chi lo osserva. Ma, come sempre, quello che il profeta tace è molto di più e più oscuro di quello che dice. Nel 2009, appare su *Science* un articolo-manifesto a firma di David Lazer e altre stelle del *jet set* scientifico dal titolo *La vita in rete: la prossima era della scienza sociale computazionale*. Nell'introduzione possiamo leggere:

> La capacità di raccogliere e analizzare enormi quantità di dati ha trasformato in modo inequivocabile campi come la biologia e la fisica. L'emergere di una tale "scienza sociale computazionale" basata sui dati è stata molto più lenta, ampiamente guidata da pochi intrepidi computer scientists, fisici e scienziati sociali. Se si dovessero esaminare le principali riviste disciplinari in economia, sociologia e scienze politiche, ci sarebbe una minima evidenza di una scienza sociale computazionale emergente impegnata nella modellizzazione quantitativa di questi nuovi tipi di tracce digitali [quelle lasciate dagli utenti su social network, mail, ecommerce, eccetera. *NdA*]. Tuttavia, la scienza sociale computazionale sta avvenendo, e su larga scala, in luoghi come Google, Yahoo e la *National Security Agency*. La scienza sociale computazionale potrebbe facilmente diventare il dominio quasi esclusivo di aziende private e agenzie governative. In alternativa, potrebbe emergere un modello alla *Dead Sea Scrolls*, con un insieme privilegiato di ricercatori accademici che siedono su dati privati dai quali producono articoli scientifici che non possono essere criticati o replicati. Nessuno dei due scenari servirà l'interesse pubblico a lungo termine per l'accumulazione, la verifica e la diffusione della conoscenza[3]

In termini più prosaici e meno utopistici, questo articolo indicava un futuro molto lontano dalla realtà disvelata di Anderson su *Wired* e più vicino al campo di battaglia in cui ci troviamo oggi (il riferimento all'NSA quattro anni prima dei *leaks* di Snowden risulta premonitore).
I termini dello scontro, che come abbiamo visto sono espliciti da almeno un decennio, sono molteplici e diversi. Schematicamente, si può dire:

- esiste una questione metodologica, riassunta dalla perentoria e liberatoria affermazione "i dati parlano da soli";
- esiste una questione che si presenta come *deontologica* legata alla riproducibilità dei risultati scientifici in un mondo di dati proprietari e al consenso del soggetto umano di ricerca in un mondo di piattaforme proprietarie;
- esiste una questione *di identità* intorno alla figura del ricercatore: chi è e cosa può fare uno scienziato (generico o in particolare un *data scientist*)?

Tali questioni riguardano tutti i campi della ricerca (tecno) scientifica, — tornerò più avanti su questo termine — ma si fanno più evidenti in quello che si è venuto a chiamare "scienza sociale computazionale", che possiamo prendere come rappresentante particolare, e più sviluppato, di una tendenza generale. Quello che c'è sullo sfondo è la vecchia questione della neutralità della scienza, tanto nella sua dimensione epistemologica come in quella più direttamente materiale del finanziamento della ricerca.

Una questione di metodo: due esempi

Anche se a distanza di dieci anni forse ormai più nessuno ripeterebbe negli stessi termini *naïf* di Anderson l'affermazione sulla morte della Teoria, continua a essere vivo il dibattito metodologico intorno alla relazione tra dati e modelli. Tutto ormai è e deve essere *data driven* per essere spendibile (si legga "pubblicabile") nel mercato della ricerca scientifica.

In un'ottica della pratica scientifica come *sistematizzazione dell'esperienza,* esiste, secondo i sacerdoti dell'analisi dei big data, solo una questione di quantità. Il metodo scientifico, affermano, è un residuo di un mondo dominato dalla scarsità di dati. I modelli sono solo "rozze approssimazioni della realtà" che permettono di riempire i vuoti lasciati dai pochi dati prodotti in laboratorio o raccolti sul campo da ricercatori connettendoli causalmente. In un mondo dove ogni interazione lascia una traccia digitale o può essere digitalmente registrata, non esistono più questi vuoti per cui non c'è più nessuna necessità di approssimare la realtà.

Ciò che è necessario fare è individuare *patterns* e correlazioni. Queste non spiegano *perché* accade ciò che accade (fenomeno osservato), ma sono abbastanza per *prevedere* quando accadrà di nuovo.

Dal punto di vista accademico, invece, quella che era nata come socio-fisica[4] e si è poi trasformata in ciò che chiamiamo scienza sociale computazionale, aveva un programma ricalcato sul programma classico della fisica statistica: individuare le poche e semplici regole che reggono le interazioni tra oggetti a livello micro (l'atomo per la materia inerte, gli *agenti* per i sistemi sociali) che danno luogo a livello macro a complessi pattern emergenti.

Il caso *Google Flu Trend* (GTF) è abbastanza propedeutico a individuare i due approcci e ciò che implicano.

GFT era un progetto per la predizione dell'attività influenzale basato sull'aggregazione delle chiavi di ricerca usate dai milioni di utenti di Google lanciato dall'impresa nel 2008 e lasciato morire

a partire dal 2014. L'idea, piuttosto semplice, su cui si basava GFT è spiegata in un articolo apparso sulla rivista *Nature* nel 2009[5]: le persone con sintomi influenzali tendono a utilizzare Google per ricerche relazionate all'influenza. Poiché Google dispone di molti dati di ricerca e della capacità computazionale per processarli velocemente, può predire i picchi di influenza *sul momento*.

In una prima fase, centinaia di miliardi di chiavi di ricerca personali sono state processate per definire un insieme di chiavi di ricerca relazionate con l'attività influenzale, dove relazione va intesa come correlazione. Il numero di ricerche contenenti queste chiavi viene poi utilizzato come input per un modello che predice i casi di influenza. Qui "modello" è inteso come "modello statistico", ovvero una formula che, dato il numero di ricerche con le chiavi dell'insieme e certi parametri, restituisce la probabilità di avere un certo numero di casi di influenza.

In tutta la sua storia fino al 2013, le predizioni di GFT hanno alternato efficacia e rapidità a errori anche piuttosto grandi, che hanno richiesto diverse successive ricalibrazioni del modello statistico. Nel 2013 si verifica il crash definitivo, quando GFT sovrastima il picco influenzale per più del doppio.

In un articolo su *Science* del 2014 di nuovo a firma di Lazer e altri[6], la ricerca accademica si prende una rivincita criticando diversi punti del metodo Google che hanno portato al disastro.

Il primo bersaglio è l'affidabilità di Google Search come strumento di ricerca.

A differenza dei tradizionali strumenti di laboratorio, l'algoritmo di ricerca non è stato costruito per raccogliere dati sul fenomeno in studio, questi sono solo un sottoprodotto della sua funzione principale. Pertanto, le continue modifiche applicate all'algoritmo per migliorare la performance (restituire "buoni" risultati di ricerca), retroagiscono sul processo di produzione dei dati, portando a errori, anche grossolani, e alla necessità di ricalibrare costantemente i modelli statistici.

Il secondo bersaglio è la trasparenza e quindi la riproducibilità dei risultati di GFT.
Google non ha mai rivelato i termini di ricerca inclusi nell'insieme usato per le predizioni, pertanto viene a mancare la possibilità di replicare indipendentemente l'esperimento.
Un altro grande caso di dibattito è stato l'esperimento condotto da Facebook sul contagio emozionale i cui risultati sono apparsi sulla rivista PNAS in un articolo del 2014[7]. La novità, in questo caso, è che la piattaforma sociale Facebook viene usata direttamente come setting sperimentale e non solo come produttrice di dati.
La "teoria del contagio emozionale" afferma che gli stati emozionali possono contagiarsi da persona a persona per la sola esposizione, come un virus, generando quello che viene chiamato "effetto di rete": soggetti legati da una relazione sociale tendono a mostrare stati emozionali correlati. Facebook disegna un esperimento per dirimere la controversia circa questa teoria e andare oltre la semplice misurazione di correlazioni. Per farlo, modifica opportunamente il *News Feed* di più di 600.000 utenti, mostrando ad alcuni prevalentemente post positivi, ad altri post prevalentemente negativi e misurando poi il contenuto emozionale di questi, confrontandoli tra loro e con un gruppo di controllo a cui viene mostrato il loro normale *News Feed*. Così facendo, è in grado di *confermare* l'esistenza del contagio emozionale.
L'esperimento generò un gran dibattito intorno a due punti principali: di nuovo quello della riproducibilità e quello etico del consenso informato.
L'impossibilità di riprodurre i risultati dell'esperimento di Facebook è evidente: nessuno, oltre ai ricercatori *della casa*, ha la possibilità di condurre lo stesso esperimento ed è impossibile anche solo pensare di disegnare una piattaforma *ad hoc* che riproduca le stesse condizioni sperimentali sulla stessa scala.

La questione etica è tanto presente che la stessa rivista accompagnò l'articolo con "un'espressione editoriale di preoccupazione" che concludeva:

> Ottenere il consenso informato e consentire ai partecipanti di rinunciare sono buone pratiche nella maggior parte dei casi ai sensi delle linee di condotta del Dipartimento per la salute e dei servizi umani degli Stati Uniti per la protezione dei soggetti umani di ricerca (la *Common Rule*). L'aderenza alla *Common Rule* è la linea di condotta di *PNAS*, ma come società privata, Facebook non aveva alcun obbligo di conformarsi alle disposizioni della *Common Rule* quando ha raccolto i dati utilizzati dagli autori, e la *Common Rule* non preclude il loro uso dei dati. Sulla base delle informazioni fornite dagli autori, gli editori di *PNAS* hanno ritenuto opportuno pubblicare l'articolo. È comunque motivo di preoccupazione che la raccolta dei dati da parte di Facebook possa aver coinvolto pratiche che non erano pienamente coerenti con il principio di ottenere il consenso informato e di consentire ai partecipanti di rinunciare[8]

Il dibattito intorno alla questione etica metteva anche in luce un altro problema molto sentito dai ricercatori accademici che, dovendo sottostare a regole sul consenso informato, sono evidentemente svantaggiati rispetto ai ricercatori *commerciali* per i quali queste stesse regole non valgono[9].
Con il caso dell'esperimento di Facebook emerge, però, anche un altro aspetto fondamentale: quello dell'*anarchia metodologica*. Infatti, se da una parte la retorica sull'analisi dei big data si basa sull'affermazione per la quale *correlation is enough*, e quindi sulla mera analisi statistica dei dati raccolti *in natura*, dall'altra i *data scientists* delle compagnie commerciali non hanno nessuna remora a ricorrere a "vecchie pratiche sperimentali", disegnate per scoprire nessi causali, quando ce ne è bisogno.

Appunti a mo' di conclusioni

Dal punto di vista delle scienze sociali computazionali, l'era dei big data è, in realtà, l'era del capitalismo di piattaforma tecnosociale (Facebook, Twitter, Airbnb…). Così come la radioastronomia, la microbiologia o altre scienze, la scienza sociale computazionale è una scienza possibile solo a partire da un dispositivo: le piattaforme sociali in rete. Queste, in ultima istanza, ricoprono tanto il ruolo di strumento di osservazione quanto di fenomeno osservato. Come dicevo, non è una specificità di questa disciplina, piuttosto rappresenta un caso particolare di una tendenza generale che alcuni autori, come Marcello Cini, hanno già chiamato "la svolta tecnoscientifica"[10].

Da una fase di "subordinazione della tecnologia alla scienza" — la seconda è ciò che permette la progettazione della prima che a sua volta stimola nuove domande scientifiche — con il nuovo secolo siamo entrati in una fase di indissolubile intreccio tra scienza e tecnologia: le tecnoscienze. In questa fase emergono sotto una nuova luce tanto questioni di ordine metodologico, come deontologico e, direttamente, politico. Cercherò di appuntarne schematicamente alcune continuando a utilizzare il prisma delle scienze sociali computazionali.

Il momento dell'analisi — dei dati prodotti su piattaforme tecnosociali proprietarie — non può essere separato dal momento della critica.

Lo strumento di osservazione, infatti, non è inerte né nel *senso deterministico classico* né nel *senso probabilistico quantistico,* ma concorre esso stesso a produrre la *realtà* osservata e, nella maggior parte dei casi, è fuori dalla portata tanto del ricercatore come del soggetto di ricerca ogni possibilità di sua modifica. Questa questione si traduce in quella della *non neutralità dell'algoritmo.*

Questa è solo una delle questioni metodologiche che hanno bisogno di essere affrontate. Se infatti la liquidazione del metodo

scientifico *á la* Anderson ha dimostrato la sua pretestuosità, affermare *á la* Lazer che nulla è cambiato e i big data sono solo uno strumento in più per continuare a fare scienza *come sempre* è un espediente retorico utile solo a mantenere una posizione privilegiata di interlocuzione con le istituzioni del potere politico ed economico.

L'obiettivo concreto di piattaforme come Facebook nel loro fare ricerca è quello di stabilire un mercato della predizione e della modifica del comportamento umano[11]. In questo quadro il soggetto umano di ricerca, l'utente, è sia risorsa da cui estrarre la materia prima dati, sia obiettivo su cui realizzare la valorizzazione dell'informazione prodotta a partire dai dati estratti. La questione della propaganda politica computazionale ne è espressione più alta, tanto è così che prima dell'esperimento riportato sopra, Facebook condusse e pubblicò un esperimento analogo sul comportamento elettorale dei suoi utenti[12].

In questo quadro, la non neutralità della scienza, nel senso della natura contestuale e socialmente determinata delle domande a cui la ricerca scientifica cerca risposte, assume una forma più immediata.

L'opposizione portata avanti dalla ricerca accademica, che, basata sulla continuità metodologica, propone il ricercatore come interlocutore delle istituzioni pubbliche in qualità di esperto, se da una parte pone un argine all'arroganza *big data* (*Big Data hubris* nei termini di Lazer e coautori), dall'altra non cambia la posizione passiva della *società di utenti*.

La questione deontologica, insomma, è direttamente politica.

Note

1) Chris Anderson, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, in *Wired*, disponibile al link:
wired.com/2008/06/pb-theory/
2) Alessandro Della Corte, Lucio Russo, *La bottega dello scienziato. Introduzione al metodo scientifico*, il Mulino, Bologna 2016.
3) David Lazer, et al., *Life in the network: the coming age of computational social science*, in *Science* n° 323, 2009.
4) Mark Buchanan, *The social atom: Why the rich get richer, cheaters get caught, and your neighbor usually looks like you*, Bloomsbury Publishing USA, New York 2008.
5) Jeremy Ginsberg, *Detecting influenza epidemics using search engine query data*, in *Nature* n° 457, 2009.
6) David Lazer, et al., The Parable of Google Flu: Traps in Big Data Analysis, 2014, Science
7) Adam D. I. Kramer, *Experimental evidence of massive-scale emotional contagion through social networks*, in *PNAS*, 2014, disponibile al link:
pnas.org/content/111/24/8788
8) Inder M. Verna, *Editorial Expression of Concern: Experimental evidence of massivescale emotional contagion through social networks*, in *PNAS*, 2014, disponibile al link:
pnas.org/content/111/29/10779.1
9) Susan T. Fiske, Robert M. Hauser, *Protecting human research participants in the age of big data*, in *PNAS*, 2014, disponibile al link:
pnas.org/content/111/38/13675
10) Massimo Cini, *Il supermarket di Prometeo. La scienza nell'era dell'economia della conoscenza*, Codice Edizioni, Torino 2006.
11) Shoshana Zuboff, *Big other: surveillance capitalism and the prospects of an information civilization*, in *Journal of Information Technology* n° 30, 2015, pp. 75–89.
12) Robert M. Bond, et al. *A 61-million-person experiment in social influence and political mobilization*, in *Nature* n° 489, 2012, p. 295-298.