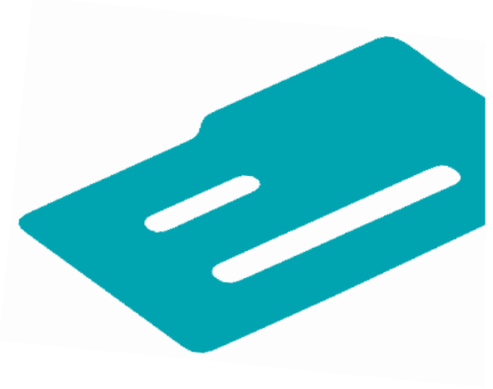
Il quarto paradigma

Pietro Greco

L'avvento del computer ha dato la possibilità di produrre nuova conoscenza scientifica. Il quarto paradigma consiste nel navigare in un mare di dati. Fra qualche anno la vera necessità non starà nella raccolta di questi dati, ma nel saperli gestire. Occorre infatti non solo imparare a conservarli ma soprattutto a saperli utilizzare



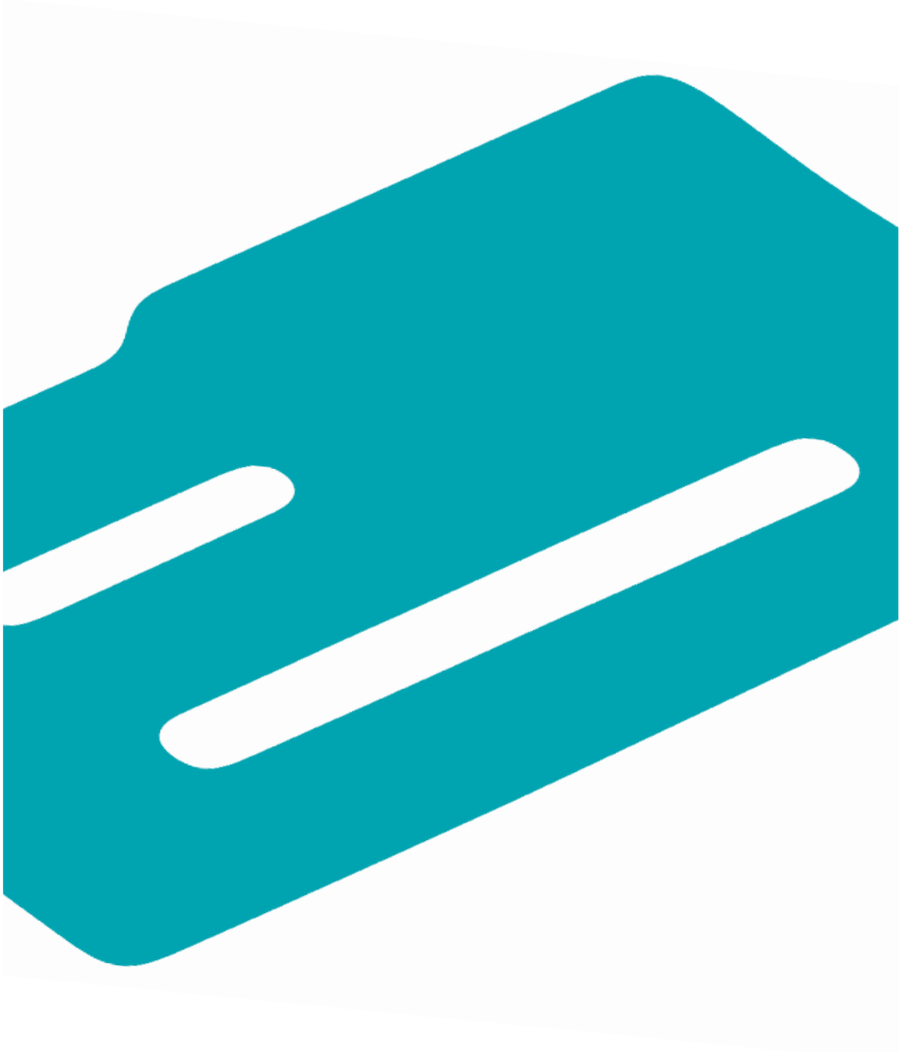
Abbattere il paradigma della segretezza. "*Open data for open science*", dati trasparenti e completamente disponibili per una scienza a sua volta trasparente e completamente aperta, invoca l'inglese *Royal Society* in un rapporto, *Science as an open enterprise*, la scienza come impresa aperta, pubblicato all'inizio del 2012: perché sarà questa la rivoluzione scientifica prossima ventura. Sarà la gestione aperta e trasparente di una quantità enorme di dati resa possibile dalle nuove tecnologie informatiche a produrre un cambio paradigma nella storia dell'epistemologia scientifica. A generare "il quarto paradigma".

La proposta è davvero significativa. In primo luogo perché viene avanzata da un'accademia antica e prestigiosa, la *Royal Society*, che ha dato un formidabile contributo, nel Seicento, alla creazione di un moderno sistema di comunicazione della scienza. Che è anche un sistema di comunicazione di dati. E poi perché la proposta cade a cinquant'anni esatti dalla pubblicazione del libro *The Structure of Scientific Revolutions*, con il quale Thomas S. Khun introduce il concetto di «cambio di paradigma» e lo assurge a motore primario della scienza.

Ma prima di affrontare il problema nei suoi termini epistemologici conviene dare un'occhiata ai fatti che dovrebbero consentire il nuovo modo di produrre conoscenza scientifica. La *Royal Society* ci propone l'esempio dell'EBI, l'*European Bioinformatics Institute*, un centro di ricerca dell'*European Molecular Biology Laboratory* (EMBL) che ha sede proprio in Gran Bretagna. Non è un esempio casuale: il centro è stato fondato per fornire dati e, più in generale, servizi informatici all'intera comunità scientifica in maniera libera e del tutto gratuita. In pratica, l'EBI raccoglie, conserva e poi distribuisce i dati ottenuti dai ricercatori che lavorano in tutti i settori delle scienze della vita: dalla biologia molecolare alla medicina all'ecologia. Ebbene, questa opera di raccolta dell'EBI sta ottenendo un grosso e crescente successo, tanto che nell'anno 2010 ha accumulato per 4.000 terabyte, quattro volte di più rispetto ai mille che possedeva nel 2008. A mero titolo di paragone: la *Library of Congress* degli Stati Uniti, a Washington – che, con i suoi 28 milioni di libri e 50 milioni di manoscritti, è una delle più grandi biblioteche del mondo – racchiude una quantità di informazione pari a circa 20 terabyte. L'archivio dell'EBI contiene, dunque, informazioni equivalenti a 200 biblioteche del Congresso degli Stati Uniti.

Una grande quantità di dati, certo. Ma che impallidisce di fronte ai 10.000 terabyte raccolti in appena un anno, il 2010, da CSM, uno solo dei tanti esperimenti che i fisici del CERN, il Centro europeo di fisica delle alte energia, stanno realizzando a Ginevra con il *Large Hadron Collider* (LHC), la macchina più grande mai costruita al mondo. La collaborazione CSM si è imbattuta nella prima difficoltà in cui si imbatte chi produce una simile quantità di dati: la loro gestione. A Ginevra l'hanno risolta distribuendo il loro archivio in cinquanta siti diversi, sparsi per il mondo.

Di qui a qualche anno assisteremo, però, ad una nuova fase di transizione nella capacità di raccogliere e nella necessità di gestire grandi quantità di dati. Fra meno di dieci anni, ad esempio, dovrebbe diventare operativo SKA, lo *Square Kilometre Array*. Un grande radiotelescopio distribuito su migliaia di chilometri quadrati tra l'Africa del Sud e l'Australia,



che ha la missione di raccoglie un milione di terabyte di informazione al giorno. In pratica, SKA sarà chiamato a raccogliere, immagazzinare e gestire ogni giorno una quantità pari a cento volte quella raccolta da CSM in un anno e a cinquantamila della *Library of Congress*. Ma non ci sono solo questi esempi. Ormai nel mondo lavorano oltre sette milioni di ricercatori. Certo non tutti raccolgono dati con la medesima cupidigia dei loro colleghi dell'EBI, del CERN o di SKA. Ma è anche vero che quasi tutti, grazie soprattutto alle nuove tecnologie digitali, producono una quantità di informazione sconosciuta in altri tempi. La somma è difficile da calcolare, ma certo nuove *Library of Congress* si aggiungono ogni giorno al grande archivio della scienza. Creando i presupposti per un nuovo modo di produrre conoscenza. Qualcuno potrebbe obiettare: ma l'informazione non è, di per sé, conoscenza. E avrebbe ragione.

Da sola, l'Ebi ha raccolto quasi 4.000 terabyte di dati, rispetto ai 20 presenti nella *Library of Congress* di Washington

O, almeno, avrebbe avuto ragione in passato. Perché, come spiegano Tony Hey, Stewart Tansley e Kristin Tolle, nel libro *The Forth Paradigm. Data-Intensive Scientific Discovery* (Il Quarto Paradigma. La scoperta scientifica fondata sull'intensità di dati), pubblicato dalla Microsoft nel 2009, oggi il salto di quantità nella produzione di informazione è tale da costituire di per sé un salto di qualità. L'idea è che l'informazione raccolta in quantità mastodontica possa trasformarsi facilmente, quasi automaticamente, in nuova conoscenza. E che, dunque, i dati raccolti ogni giorno dagli oltre sette milioni di ricercatori di tutto il mondo possano trasformarsi in nuova conoscenza scientifica. Anzi, in un nuovo modo di produrla. Per questo Hey e colleghi parlano di una nuova transizione epistemologica e di un "quarto paradigma" nella storia della scienza. Non è un'idea astratta. Gli esperti della *Royal Society* propongono l'esempio concreto della *Biobank*, la banca biologica inglese che conserva i campioni di sangue, di urina e di saliva di 500.000 persone che hanno dato un consenso alla produzione e alla diffusione delle conoscenze che possono derivare dal loro studio. Ebbene, moltiplicando queste cifre per il numero

di cellule e tenendo conto che ogni cellula contiene una molecola di Dna costituita da 3 miliardi di unità d'informazione (le basi nucleiche), possiamo avere un'idea della quantità di dati biochimici e, poi, clinici conservati. Questa quantità mastodontica di dati potrebbe generare, sostiene il rapporto della *Royal Society*, un'autentica svolta nelle conoscenze su un'intera costellazione di malattie: dal cancro agli infarti, dal diabete alla depressione. Per passare dalla potenza all'atto, occorre solo imparare a conservarla e, soprattutto, ad analizzarla quella mole di dati.
La soluzione tecnologica del problema costituisce di per sé una grande sfida. Come sostiene da tempo la *National Science Foundation* – l'Agenzia federale che coordina e finanzia la ricerca scientifica pubblica negli Stati Uniti – che già nel 2007 consigliava alla comunità scientifica (e politica) del suo Paese di creare le infrastrutture per la gestione di grandi quantità assumendo una visione cibernetica. Perché è questo che consentirà di portare avanti le frontiere della scoperta nel XXI secolo nei più svariati settori: dall'ecologia al clima, dalla fisica delle particelle alla sociologia. Questo tipo di sfida tecnologica è stato colto dagli scienziati e dai tecnici di SKA, il radiotelescopio che segnerà lo sbarco della "*big science*" in Africa. Il progetto prevede la realizzazione di una rete di computer che dovrà gestire il database del radiotelescopio che, come abbiamo detto, sarà senza precedenti. Se la sfida informatica avrà successo, da qui a dieci anni avremo quanto di più simile a quell'intelligenza evocata all'inizio del XIX seco-

La nuove tecnologie consentono non solo di studiare sistemi più complessi, ma di simularne il comportamento

lo dal marchese Pierre-Simon de Laplace, in grado, conoscendo posizione e velocità di ogni particella dell'universo, di conoscere il presente, il passato e il futuro del cosmo intero.
Il problema non è solo come acquisire e archiviare enormi quantità di dati. Il problema è come analiz-

zarli. Come trasformare l'informazione in conoscenza. Trent'anni fa, nella montagna di dati raccolti dal satellite *Solar Mesosphere Explorer* (SME), inviato in orbita per studiare l'ozono nell'alta atmosfera, era contenuta una preziosa informazione: la concentrazione di quel gas nella stratosfera stava diminuendo. Il sistema automatico di analisi e correzione dei dati allora a disposizione non riuscì a "leggere" il contenuto di conoscenza nascosto in quelle informazioni chimico-fisiche. C'è poi voluta tutta l'abilità di Paul Crutzen, Mario Molina e Sherwood Rowland per estrarla, portandoli a ricevere il premio Nobel per la chimica nel 1995. Il pagliaio delle informazioni raccolte da SME era relativamente piccolo e l'acume di Crutzen, Molina e Rowland ha potuto trovare l'ago della nuova conoscenza che vi era nascosto. Ma il pagliaio di informazioni che hanno allestito EBI o LHC e che allestirà SKA è così grande che nessun umano potrà realisticamente infilarcisi per cercare l'ago. Detta in altri termini: non possiamo contare più sull'intelligenza e sulla capacità di lavoro degli uomini. Dobbiamo affidarci alla potenza degli algoritmi e delle macchine. Perché solo il combinato disposto di una gran quantità di dati e della capacità di analizzarli, può trasformare l'informazione in conoscenza.

Ieri questo combinato disposto era difficile da ottenere. Oggi, invece, è accessibile. Per questo Tony Hey, Stewart Tansley e Kristin Tolle parlano esplicitamente di un nuovo paradigma epistemologico, il quarto. In realtà, il primo ad alludere a un quarto paradigma prodotto dalla *eScience*, dalla rivoluzione digitale in ambito scientifico, è stato Jim Gray, un informatico che ha vinto il premio Turing assegnato ai grandi matematici e che ha collaborato a lungo con la Microsoft. Sono diversi anni che Gray, con crescente successo, cerca di convincere il mondo intero che siamo entrati in una nuova era nella produzione della conoscenza scientifica. Il primo e il secondo paradigma della scienza, ricorda Gray, sono la descrizione dei fenomeni naturali e la scoperta delle "leggi della natura". Galileo alludeva a questi due paradigmi già nel XVII secolo quando parlava delle "sensate esperienze" e delle "certe dimostrazioni". Il combinato disposto di scienza sperimentale e di scienza teorica hanno prodotto un formidabile aumento delle conoscenze sulla natura per quasi quattro secoli. Negli ultimi decenni, tuttavia, è nato un nuovo modo di produrre conoscenza: la scienza computazionale. L'avvento del computer nella seconda metà del XX secolo ha consentito non solo di studiare sistemi più complessi, ma di simularne il comportamento. Grazie al computer è nata, dunque, una terza possibilità di produrre nuova conoscenza scientifica, la simulazione. In molti campi, ormai, la ricerca scientifica non riguarda più il mondo naturale, ma un mondo virtuale, riprodotto al computer in analogia a quello reale. Per esempio, le previsioni sui cambiamenti del clima o quelle sull'evoluzione di una stella a neutroni vengono realizzate mediante simulazioni al computer. C'è un indubbio svantaggio in questo modo di fare scienza: i risultati riguardano non la realtà, ma solo un'approssimazione più o meno buona della realtà. Ma c'è anche un grande vantaggio: gli esperimenti controllati si possono ripetere all'infinito, modificando a piacimento ogni parametro e scarrozzando senza limiti nello spazio e nel tempo. È grazie a questa possibilità che Edward Lorenz ha (ri) scoperto le leggi del caos e ha verificato, al computer, l'estrema sensibilità alle condizioni iniziali del sistema meteorologico: «basta un battito d'ali di una farfalla in Amazzonia per scatenare una tempesta sul Texas». Oggi la scienza simulante ha una funzione decisiva. Non ci sarebbe, per esempio, una scienza

Le nuove banche dati consentiranno ai ricercatori di mettere a disposizione l'intera massa dei dati raccolti

del clima, con tanto di previsioni, se i climatologi non avessero una quantità grande a piacere di pianeti Terra virtuali su cui sperimentare. La simulazione è il terzo paradigma della scienza. Ebbene, sostiene Jim Gray, ci sarà a breve un quarto paradigma: la possibilità di navigare in un mare sconfinato di dati – otte-

nuti da strumenti scientifici come LHC e SKA, dalla rete di sensori di ogni natura, forma e dimensioni sparsi per il pianeta, o generati dalle simulazioni al computer – alla ricerca (anche) di ordine e regolarità che non vediamo e che le teorie non prevedono. Si tratta di una navigazione interdisciplinare capace di generare nuova conoscenza. Nel grande pagliaio dei megadati, a cercare in maniera automatica gli aghi della nuova conoscenza saranno gli algoritmi che i matematici (e i loro computer) metteranno a punto. Questa ricerca è chiamata *eScience*. E la *eScience* – che non è né sensata esperienza, né certa dimostrazione e neppure simulazione – è il quarto paradigma della scienza.

Difficile dire se Jim Gary, Tony Hey, la NSF e la *Royal Society* hanno ragione. Se possiamo davvero parlare di una transizione epistemologica prossima ventura. È certo tuttavia che i megapagliai di dati in ogni settore esistono. Ed è certo che noi abbiamo la possibilità tecnica di entrarci dentro e di esplorarli a piacimento con quella sorta di esercito di robot cognitivi che sono gli algoritmi. Sarebbe, dunque, un peccato perdere o ridurre fortemente l'opportunità offerta dal «quarto paradigma». Sarebbe un errore se qualcuno impedisse all'esercito di robot cognitivi l'accesso ai pagliai. Ecco perché, sostiene la *Royal Society*, dobbiamo operare almeno tre scelte molto nette. Primo: aumentare il tasso di comunicazione. Significa costruire i pagliai. Fuor di metafora: tutti gli scienziati, in totale trasparenza, devono conferire a una banca dati globale ogni e qualsiasi dato in loro possesso. A questa allocazione di informazioni devono poter contribuire anche i cittadini comuni, non esperti. Secondo: aumentare il tasso di accessibilità. Significa libertà di entrare a piacimento nei pagliai. Tutti devono liberamente accedere alla banca globale e intraprendere percorsi di navigazione digitale (e non) nel mare magnum dei dati. Terzo: aumentare il tasso di risorse pubbliche. Significa che il pubblico crea e controlla l'esercito dei robot cognitivi. In altri termini le istituzioni, nazionali e internazionali, devono mettere a disposizione una quantità sufficiente di quattrini necessaria a creare le infrastrutture informatiche adatte. Non ci vuole molto, in termini economici. Alcuni governi – a iniziare da quello inglese e da quello americano – e alcune istituzioni sovranazionali, a iniziare dall'Unione Europea – stanno già dimostrando di aver colto la novità e di aver raggiunto la convinzione che sarà questa in futuro la strada principale con cui si produrrà nuova conoscenza e innovazione tecnologica. In Italia si fa e se ne discute poco. Occorre almeno iniziare a parlarne.