La candela

L'ultima volta ho veramente esagerato con la lunghezza dei miei sproloqui... Potrei cercare delle attenuanti, ma preferisco rimettermi alla clemenza della corte, e mostrare il mio ravvedimento operoso: in altre parole, questa volta sarò molto più breve.

Vorrei proseguire un filone di discorso che ho già affrontato: quello dell'uso e del significato di alcune parole "popolari," ovvero "di moda." Se ricordate, ci siamo occupati una volta del termine "lineare," poi di "esponenziale." Oggi mi propongo un obbiettivo più alto: niente meno che "informazione."

Non ho bisogno di provare che si tratta di una parola di uso (e forse abuso) frequentissimo: nella pubblicistica di ogni genere, nella divulgazione, e anche nelle scienze, in particolare biologiche. Tanto per cambiare, la parola proviene dal linguaggio comune ("avrei bisogno di un'informazione ...," "monopolio dei mezzi d'informazione") ma da essa è nata addirittura una scienza (l'informatica) con relativo corso di laurea, con materie specifiche denominate ad es. "metodi per il trattamento dell'informazione"; e prima ancora una "teoria dell'informazione," ecc. ecc. Poi si è propagata nel campo che riguarda più da vicino chi mi legge: la ben nota "informazione genetica." Neanche la fisica ne è rimasta immune: per es. si discute se la "velocità di propagazione dell'informazione" possa o no superare la velocità della luce; oppure si pensa ai fenomeni di correlazione quantistica a distanza, con le fantascientifiche ipotesi di "teletrasporto" che ne conseguono...

Il primo problema che mi viene in mente è che nella maggior parte dei casi il significato di "informazione" è assai vago, e certamente non ben definito in senso scientifico. L'unica eccezione è la teoria dell'informazione, dove il significato è preciso, anche se spesso frainteso. Ne parleremo tra poco.

Prima però ragioniamo un po' sull'informatica. Questa è una parola abbastanza recente (non più vecchia di 30 anni, credo) ed è un caso raro di parola recente non importata dall'inglese americano, bensì dal francese. Sono infatti i francesi che hanno coniato "informatique," evidente calco da "mathématique," per intendere la scienza dell'elaborazione e trattamento dell'informazione, in qualsiasi forma.

L'informatica è una scienza che deve tutto a uno strumento: ovviamente il calcolatore elettronico. Prima dello sviluppo dei calcolatori esistevano dei prodromi, delle premesse; ma una visione generale, l'inclusione sotto un'unica disciplina di procedimenti attinenti a tipi d'informazione del tutto diverse (parlata, scritta, numerica, figurativa ...) non solo ha potuto nascere *dopo* lo sviluppo dei calcolatori, ma ha richiesto che queste macchine raggiungessero una certa

soglia di capacità e prestazioni, in termini di memoria e di velocità, al disotto della quale si poteva teorizzare e immaginare, ma non mettere in pratica. E la possibilità di mettere in pratica è stato un potente fattore di sviluppo anche per le strutture teoriche.

In parte per ragioni di età, ma anche per le vicende della mia carriera scientifica, ho potuto vedere da vicino questi sviluppi, e posso ricordare quanto fosse diverso, poniamo 50 anni fa, non solo il bagaglio di strumenti (calcolatori) di cui ci si poteva servire, ma proprio l'apparato concettuale, il modo stesso con cui si pensava a queste cose. Non voglio dilungarmi su un punto che non è il mio tema di oggi; ma debbo pur ricordare che quando arrivai a Pisa, chiamato appunto a far parte della piccola squadra che avrebbe progettato e realizzato il primo calcolatore interamente "made in Italy" (CEP: Calcolatrice Elettronica Pisana) i più potenti calcolatori esistenti al mondo erano molte migliaia di volte più lenti dei "personal" odierni; avevano memorie forse 10 000 volte più piccole, e ricevevano o fornivano dati mediante "periferiche" lentissime e con modestissime capacità: esclusivamente alfabetiche e numeriche. Niente schermi video, stampanti a colori ad alta risoluzione, scanner, telecamere, e via discorrendo...

Quando velocità e memoria sono aumentate, è stato possibile pensare non solo a calcoli più complicati, ma a immagini grafiche, a musica; e poi a collegamenti a distanza: prima fra un calcolatore centrale e una serie di "terminali," poi fra calcolatori distanti, in città diverse... Così è nata Internet, e poi il WWW che oggi tutti conoscono e trovano naturale più o meno come il telefono. Da qui la possibilità (e la necessità) di trasmettere quantità d'informazioni sempre maggiori, e quindi la ricerca di una sempre maggiore efficienza e sicurezza nella trasmissione. E quando dico "sicurezza" intendo due cose: che la trasmissione sia esente da errori, e che sia protetta da intrusioni.

E da qui le varie tecniche di compressione e di crittografia, oggi di uso corrente, che hanno avuto origine da antichi lavori (di cui riparlerò tra poco) ma sono state fortemente stimolate dalle esigenze pratiche che ho detto.

* * *

La nascita della teoria dell'informazione viene fatta concordemente risalire a un lavoro del 1948 di Claude Shannon, intitolato "The Mathematical Theory of Communication." Ecco un brano dell'introduzione:

> "Il problema fondamentale della comunicazione è quello di riprodurre in un luogo, esattamente o approssimativamente, un messaggio scelto in un altro luogo. Frequentemente i messaggi hanno *significato*; essi si riferiscono cioè (o sono correlati secondo un qualche sistema) a certe entità fisiche o concettuali. Questi aspetti semantici della comunicazione sono irrilevanti per il problema dell'ingegnere. L'aspetto significativo è che il messaggio effettivo è uno *scelto tra un insieme* di messaggi possibili. Il sistema dev'essere progettato in modo da operare per ogni

possibile selezione, e non solo per quella che avrà realmente luogo, dal momento che questa non è nota al momento del progetto."

Mi piace proporre ogni tanto citazioni da "classici" (quale può essere definito a buon diritto il lavoro di Shannon) perché spesso un classico della scienza è tale non solo perché ha prodotto idee che sarebbero rimaste nella storia, ma anche per la sintetica chiarezza con cui sa esprimerle. Nel caso in esame, il passo più importante per la nostra discussione è l'osservazione che gli aspetti semantici sono *irrilevanti*. Shannon dice "per il problema dell'ingegnere," è vero; ma in realtà la questione è più profonda, perché tutto ciò che segue (la misura della quantità d'informazione, le tecniche di codifica, la trasmissione in presenza di rumore) fa uso in modo determinante di quest'idea: la definizione dell'informazione è tale da ignorare il carattere semantico del messaggio.

Il lettore avvertito già capisce in che misura questo approccio avrà peso quando ci accosteremo agli usi biologici del concetto d'informazione... Ma non anticipiamo. E più in generale, il medesimo lettore avrà colto una fonte di equivoci o di abusi, che possono nascere dal dimenticare quella prescrizione, gravando invece il concetto d'informazione di un carico semantico, di vario genere.

* * *

Il primo problema che Shannon si pone è quello di definire una misura della quantità d'informazione. Seguendo idee già avanzate da altri autori, opta per una misura *logaritmica*, secondo un criterio assai semplice: prima di tutto, la quantità d'informazione associata a un particolare messaggio non dipende affatto dal suo contenuto (significato) ma solo dal fatto di essere stato scelto in un insieme, ed è tanto maggiore quanto più ampio era quell'insieme. Sia dunque N il numero di messaggi possibili: si potrebbe prendere lo stesso N come misura dell'informazione trasmessa, ma non sarebbe ragionevole.

Il motivo è semplice, e possiamo capirlo con un esempio. Supponiamo che il messaggio sia numerico, e consista di un numero di 5 cifre; dato che i possibili numeri di 5 cifre sono $100\,000 = 10^5$, avremmo qui $N = 10^5$. Ma se invece trattiamo con numeri di 10 cifre, ognuno di questi può essere pensato come due numeri di 5 cifre, ed è dunque naturale assumere che un numero di 10 cifre trasmetta una quantità d'informazione doppia di uno di 5. Invece per numeri di 10 cifre avremmo $N = 10^{10}$, che non è il doppio, bensì il quadrato di 10^5. Ci sono infatti $10^5 \times 10^5 = 10^{10}$ modi di scegliere due numeri di 5 cifre.

La soluzione è ovvia: anziché usare N, conviene usare il suo *logaritmo*. Infatti, come tutti sanno (posso dirlo?) il logaritmo di un prodotto è la somma dei logaritmi dei fattori, quindi nel nostro caso: $\log 10^{10} = \log(10^5 \times 10^5) = \log 10^5 + \log 10^5 = 2 \times \log 10^5$, e così l'uso di due numeri effettivamente raddoppia la quantità d'informazione.

C'è però un piccolo problema tecnico: l'uso di un logaritmo implica che si definisca una *base*. Ricordiamolo: si dice che x è il logaritmo di y nella base b, e si scrive $x = \log_b y$, se vale $b^x = y$. Tutti sappiamo che per i logaritmi sono in

uso prevalente due basi: la base 10, che dà luogo ai logaritmi *decimali* (es. il pH), oppure la base e, che porta ai logaritmi *naturali* o *neperiani* (pensate a tante relazioni della chimica fisica). La scelta che si è affermata è quella della base 2 (logaritmi *binari*) per una ragione pratica, connessa al fatto che i calcolatori elettronici lavorano con numeri rappresentati in base 2: infatti se pensiamo a numeri scritti in base 2, allora con L cifre si possono avere 2^L numeri diversi (scelte di messaggi) ed è proprio $\log_2 N = L$. Tutti oggi sanno che le cifre binarie si chiamano "bit," e quindi il numero di bit viene a coincidere con la quantità d'informazione associata a quel numero (messaggio).

Va da sé che non sempre i messaggi sono espressi in cifre numeriche: possono essere scritti in un alfabeto di lettere, o essere cose più complicate, come per es. un'immagine. Nasce quindi il problema di calcolare la quantità d'informazione di un messaggio di tipo più generale, il che può essere più o meno difficile a seconda dei casi.

Vediamo un esempio ancora abbastanza semplice: un messaggio alfabetico. L'alfabeto consisterà di un certo numero di "lettere," e trascuriamo per il momento che si dovrà distinguere maiuscole e minuscole. Trascuriamo pure la necessità dei segni di punteggiatura (esistono opere letterarie che deliberatamente ne fanno a meno ...) e limitiamoci a prevedere uno "spazio," indispensabile per separare tra loro le parole. Allora se il messaggio è scritto per es. in inglese richiederà in tutto 27 simboli distinti: le 26 lettere e lo spazio. Perciò un messaggio di lunghezza L farà parte di un insieme di $N = 27^L$ messaggi diversi, e la quantità d'informazione trasmessa (in bit) sarà: $\log_2 N = L \times \log_2 27 \simeq 4.75\, L$. Il che vuol dire che ogni carattere del messaggio contribuisce con 4.75 bit all'informazione totale.

Il secondo esempio, che interessa più da vicino la biologia, è per fortuna ancora più semplice. Avrete già capito che alludo al DNA. Per il momento mi occupo solo del fatto che il DNA è una catena di nucleotidi, che si distinguono tra loro solo per le 4 basi che sappiamo, indicate con A, C, G, T. Dunque se L è la lunghezza della catena, il numero di varianti (messaggi) astrattamente possibili è 4^L, e il calcolo della quantità d'informazione si fa come prima. La maggiore semplicità sta nel fatto che il logaritmo di 4 in base 2 è 2 (perché $2^2 = 4$) e quindi la quantità d'informazione è semplicemente $2L$: *ogni nucleotide contribuisce per* 2 *bit*.

Se ricordo bene, la lunghezza del DNA umano è qualche miliardo: prendiamo 2 miliardi per rendere più semplice il calcoletto che voglio fare, ma se poi sono 4 miliardi la sostanza non cambia. Dunque la quantità d'informazione associata al nostro DNA è 4 miliardi di bit. Sapete che nella pratica informatica sono in uso i "byte," che non sono altro che gruppi di 8 bit; ne segue che il DNA umano richiede 500 Mbyte: una quantità di dati che entra comodamente in un CD. Senza contare che la storia non è così semplice, e se la studiamo meglio la quantità d'informazione non può che diminuire... Vediamo subito.

Ma prima di affrontare un altro tema debbo farvi notare che di soppiatto ho introdotto una nuova idea, quando ho scritto "il DNA umano richiede 500 Mbyte." Che cosa significa quel "richiede"? L'idea che ci sta sotto è che un'informazione, qualsiasi informazione, può essere codificata e trasmessa in più modi, e che ciò che conta è la quantità d'informazione da trasmettere. Nel nostro caso, ho assunto che la lunghissima sfilza di A, C, G, T (2 miliardi) venga rappresentata da un codice binario, i bit risultanti raggruppati in byte, e trascritti in un CD, dal quale potranno essere riletti e volendo ritrasformati nella stringa originaria. È un'idea fondamentale, ed è quella che giustifica il titolo del lavoro di Shannon, che se ricordate parlava di "comunicazione." Ma di ciò diremo più avanti.

* * *

L'altro tema cui accennavo è questo: in tutto quanto precede abbiamo supposto che i messaggi potessero consistere di un certo numero L di simboli elementari, scelti in un alfabeto che contava n simboli: 27 per i messaggi propriamente alfabetici, 10 per quelli numerici, 4 per il DNA. Il numero totale di messaggi possibili era allora $N = n^L$. Ma in molti casi i diversi simboli, anche se tutti possibili, non sono però ugualmente frequenti: la cosa è evidente per i messaggi alfabetici, dove le vocali (almeno in italiano) sono più frequenti delle consonanti, e alcune consonanti sono molto più frequenti di altre. Se si tiene conto di questo vincolo, il numero totale di messaggi possibili cambia, come possiamo capire facilmente con un esempio.

Supponiamo che l'alfabeto sia *binario* (0 e 1) ma che gli zeri siano tre volte più frequenti degli uni: allora in un messaggio di 4 simboli ($L = 4$) ci sarà di regola un solo 1 e tre 0, e questo si può fare solo in 4 modi, mentre se non ci fosse stato il vincolo avremmo avuto 16 messaggi, tutti possibili. Dunque $N = 4$, il suo logaritmo in base 2 è 2, e il messaggio di lunghezza $L = 4$ tramette solo 2 bit, anziché i 4 che avrebbero potuto essere. Se $L = 8$ il calcolo si fa in modo analogo, e si trovano 4.8 bit; per $L = 16$ si ottengono 10.8 bit, ecc.

È utile considerare non i casi singoli di messaggi di lunghezza finita, ma il comportamento per L molto grande. Il solito Shannon ha dato la formula generale, che non vi cito, e che mostra che nel nostro esempio, *se il messaggio è molto lungo*, la quantità d'informazione trasmessa è $0.81L$ bit, invece di L bit. Interessa il risultato generale: se i simboli non sono equiprobabili, l'informazione trasmessa diminuisce, e diminuisce tanto più quanto più ci si allontana dall'equiprobabilità.

Se applichiamo lo stesso ragionamento al caso alfabetico, tenendo conto delle effettive frequenze delle singole lettere, ne segue che troveremo una quantità d'informazione minore dei 4.75 bit per carattere che avevamo calcolato sopra. Il dato per la lingua inglese (non ho a portata di mano le frequenze in italiano) è 4.05: non una gran differenza, ma comunque una riduzione del 15%.

Ed eccone una conseguenza interessante, anche questa dimostrata da Shannon: in queste condizioni è possibile "codificare" il messaggio in modo da renderlo più breve (esattamente del 15% nel nostro caso) *senza perdere alcuna informazione*, ossia con la possibilità di ripristinare in modo preciso il messaggio originario. Vediamo quindi una prima possibilità di *compressione*, che concretizza quanto avevo accennato all'inizio. Ma l'argomento è troppo importante e complesso per trattarlo di corsa, per cui lo rimando a una prossima occasione.

Ma perché la diversa frequenza dei simboli riduce la quantità d'informazione? Possiamo rendercene conto in più modi. Intanto, consideriamo un caso estremo: se in un codice binario uno dei due simboli (per es. l'1) fosse addirittura assente, tutti i messaggi consisterebbero soltanto di zeri. Ma allora non ci sarebbe alcun bisogno di trasmettere il messaggio, perché il destinatario potrebbe scriverselo da solo. Ergo: informazione zero. Se invece, in un codice alfabetico, i messaggi contenessero soltanto le lettere A e B e mai le altre, di fatto saremmo ridotti a un codice binario: invece di trasmettere 4.75 bit per carattere ne trasmetteremmo soltanto uno.

Supponiamo invece ancora un codice binario, ma con rarissimi zeri: per es. uno ogni 100 simboli. Allora per trasmettere un messaggio completo, lungo poniamo 10 000 simboli, non occorrerebbe inviare lunghissime file di 1 intervallate da rari zeri: basterebbe dire quali posti occupano i 100 zeri. Dato che ogni zero può occupare uno dei 10 000 posti disponibili, un numero di 4 cifre decimali basterebbe: in tutto 400 cifre. Ma una cifra decimale in termini d'informazione "vale" $\log_2 10 = 3.3$ bit, per un totale di circa 1300 bit, contro i 10 000 che si sarebbero avuti con 0 e 1 equiprobabili.

Si vede quindi quanto più povera sia l'informazione in un messaggio in cui uno dei due simboli è molto meno frequente dell'altro. E di passaggio ho anche indicato una possibile maniera di *comprimere* il messaggio: trasmettere 400 cifre decimali al posto di 10 000 binarie.

In termini generali, possiamo dire che simboli non equiprobabili sono (almeno in parte) più *prevedibili*: ne ricaviamo quindi che la quantità d'informazione è legata all'imprevedibilità del messaggio, e che ogni intervento o trattamento che aumenti la prevedibilità riduce l'informazione.

* * *

Per mantenere la promessa di brevità, preferisco terminare qui questa introduzione ai concetti e alla teoria dell'informazione. Vi lascio con un riassuntino, che potrà aiutare a cogliere i punti essenziali.

Dopo alcune considerazioni di carattere generale sull'informazione e sul suo trattamento, legato alla storia dei calcolatori, ho introdotto l'idea centrale della teoria di Shannon: che l'informazione vada studiata *a prescindere* dal suo significato, ma solo come un messaggio astratto, consistente di un certo numero di simboli. Abbiamo visto che una ragionevole misura dell'informazione contenuta

in un messaggio è data dal logaritmo (per es. in base 2) del numero totale di messaggi possibili, e abbiamo esaminato alcuni esempi.

Ho accennato di passaggio alle idee di *codifica* e di *compressione*, sulle quali torneremo una prossima volta; poi ho affrontato una necessaria complicazione: il caso in cui i possibili simboli non abbiano uguale frequenza (o probabilità: avete notato che ho un po' giocato sull'ambiguità tra i due concetti? Non posso ora spiegare perché...). Abbiamo appreso che uno scostamento dall'equiprobabiltà *riduce* il contenuto d'informazione.

A parte i temi che ho già citato, e qualche incursione verso ambiti più generali, come ad es. l'informazione visuale (le immagini), ci sarebbe ancora da parlare del problema della sicurezza, nelle due facce della protezione da errori e della protezione da intrusioni. Poi mi resterà da affrontare una questione centrale per l'informazione in genere, ma soprattutto per l'informazione genetica: è giusto ignorare il lato semantico? Nel caso della genetica, è corretto dire che l'informazione risiede nel DNA, e che si può misurare come Shannon ci ha insegnato? Avremo da divertirci...