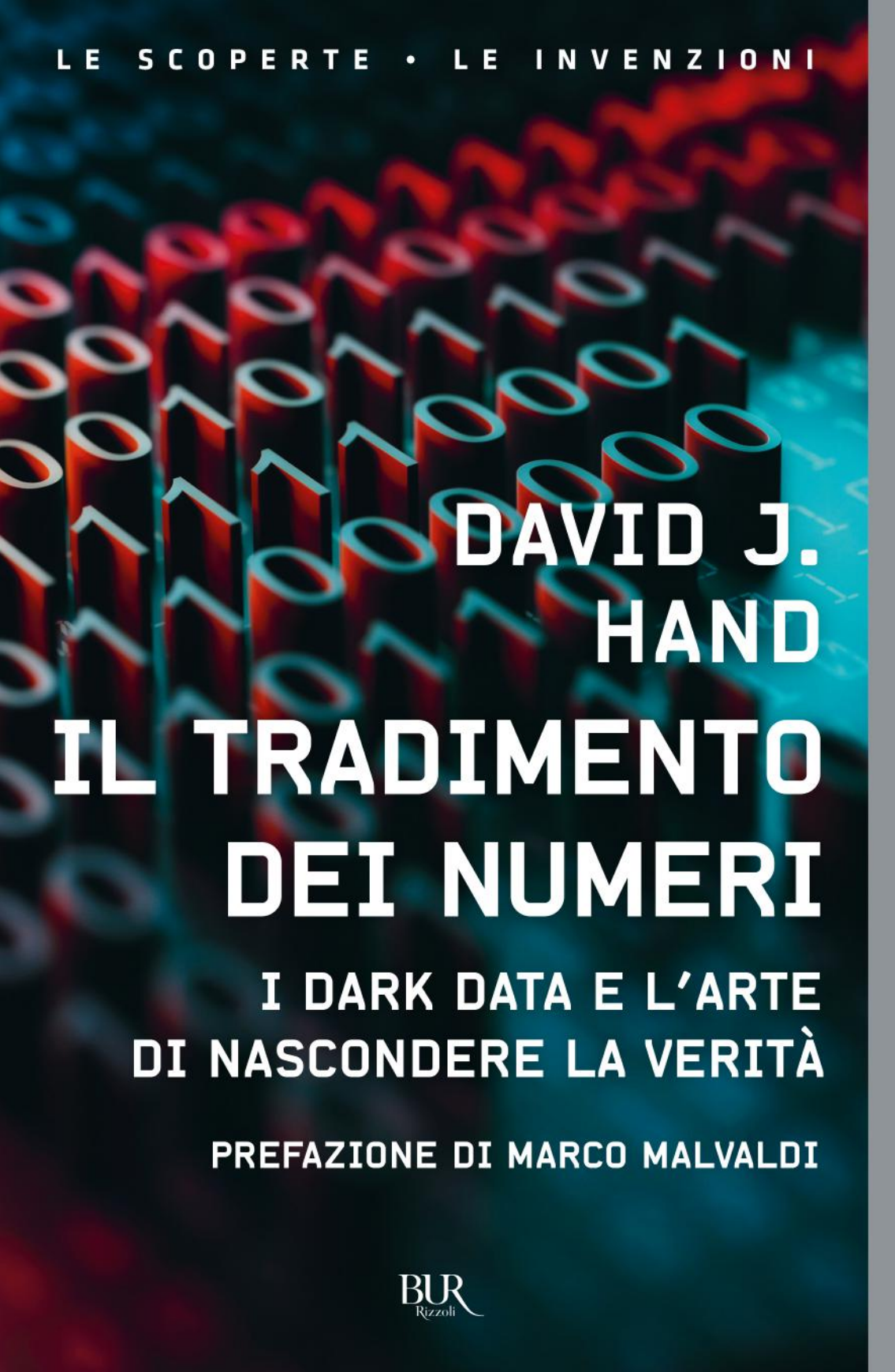
LE SCOPERTE • LE INVENZIONI



DAVID J. HAND

IL TRADIMENTO DEI NUMERI

I DARK DATA E L'ARTE DI NASCONDERE LA VERITÀ

PREFAZIONE DI MARCO MALVALDI

BUR Rizzoli

DAVID J. HAND

IL TRADIMENTO DEI NUMERI

I DARK DATA E L'ARTE DI NASCONDERE LA VERITÀ

A CURA DI MARCO MALVALDI

BUR Rizzoli

LE SCOPERTE · LE INVENZIONI

Pubblicato per

BUR Rizzoli

da Mondadori Libri S.p.A.

ISBN 978-88-17-15383-6

Titolo originale dell'opera:
DARK DATA. Why What You Don't Know Matters

Traduzione di Daniele Didero e Nicolina Pomilio

Prima edizione Rizzoli: 2019
Prima edizione BUR Le Scoperte – Le Invenzioni: ottobre 2020

Realizzazione editoriale: Netphilo Publishing, Milano

Seguici su:

www.rizzolilibri.it /RizzoliLibri @BUR_Rizzoli @rizzolilibri

Un gatto nero in una stanza buia

di Marco Malvaldi

La storia che vi vorrei raccontare è una storia di guerra. Per essere precisi, della Seconda guerra mondiale. Come fa notare il matematico Jordan Ellenberg, è anzi la tipica storia della Seconda guerra mondiale: inizia, infatti, con i nazisti che costringono qualcuno a emigrare e finisce con i nazisti che si pentono amaramente di averlo fatto.

La persona costretta a emigrare si chiama Abraham Wald. È un ebreo, che dalla propria città natale (Cluj, in Romania, che però all'epoca dei fatti si chiama Klausenburg) partirà adolescente per andare a studiare matematica in Austria, e dalla propria città adottiva (Vienna, in Austria, che però all'epoca dei fatti è in Germania) partirà adulto per andare negli Stati Uniti, per i motivi di cui si parlava sopra.

All'inizio, Wald si unisce a Oskar Morgenstern, il padre della teoria dei giochi, che lo ha chiamato in Colorado; ma dopo pochi mesi rifà le valigie e si trasferisce a New York. Tutto un altro posto. A New York c'è la Columbia University, che gli offre la cattedra di Statistica; c'è Times Square, Broadway, Central Park, che gli offrono svaghi e relax; e c'è l'SRG, lo Statistical Research Group, che gli permetterà di combattere i nazisti senza muoversi da casa.

L'SRG è uno dei tanti istituti di ricerca dove la guerra si combatte a equazioni invece che a cannoni, e non si occupa di problemi fisici (bombe atomiche, sottomarini e così via) ma di problemi matematici e statistici. E proprio uno di questi problemi verrà risolto da Wald in una maniera che scintilla di genialità.

Il problema riguardava la salvaguardia degli aerei da caccia e dei piloti dei medesimi negli scontri con i caccia della Luftwaffe. Un caccia è per definizione un velivolo leggero e agile, e non è costruito per resistere agli urti: il Savoia-Marchetti S.M.79, noto come «il gobbo maledetto» a causa dell'arma montata sul dorso ricurvo, era fatto in tubi di ferro, legno e tela cerata. Per evitare l'abbattimento, questi aerei venivano spesso corazzati con robuste lastre di ferro; ma il problema era quanto e come. Se un aereo non è corazzato, è facile da abbattere; se è troppo corazzato, è difficile da manovrare. Occorre trovare la giusta via di mezzo.

Anche perché gli aerei che tornavano dai combattimenti erano crivellati in modo non uniforme; la maggior parte dei colpi si trovava sulla fusoliera e sulle ali, un numero lievemente minore sul sistema di alimentazione, e pochissimi sul motore. Wald, per ottimizzare la distribuzione delle lastre, chiese una statistica precisa e divisa per settori della densità di colpi (numero per decimetro quadrato) sui vari settori degli aerei, e ne ottenne la seguente tabella:

Zona dell'aereo	*Numero di colpi/dm²*
Motore	0,103
Fusoliera	0,161
Alimentazione	0,143
Ali	0,167

Grazie a questa tabella, Wald fu in grado di posizionare le lastre nel punto più rischioso.

Intorno al motore. Dove i colpi erano pochissimi.

Gli ufficiali rimasero di sasso. «Professor Wald,» obiettarono «ma il numero di colpi intorno al motore è molto minore dei colpi che sono stati ricevuti sul resto dell'aereo.»

«Appunto» rispose Wald. «Chiedetevi dove sono quegli aerei che sono stati colpiti al motore.»

Gli ufficiali tacquero.

Le misure erano state fatte sugli aerei che erano tornati dalle missioni; ma un numero non trascurabile, circa il dieci per cento, non tornava affatto.

E se non tornava, era proprio perché era stato colpito.

Fin qui, la narrazione. Una narrazione comprensibile, efficace, di impatto – e assolutamente di fantasia.

Una sola cosa nella narrazione è corretta: Abraham Wald aveva a che fare con dati nascosti. E quando si ha a che fare con dati nascosti, è necessario rendersi conto che ragionare sui dati che si hanno credendo che siano tutti i dati necessari può portare a conclusioni *opposte* a quelle corrette. Non diverse, non dissimili: il con-tra-rio.

Una delle parti più divertenti del lavoro dello statistico è proprio questa: cercare di immaginarsi che valore dare ai dati mancanti; se non è possibile, cercare di fare in modo che questi dati obbediscano a una serie di vincoli, di relazioni obbligate che limitino tali ignote quantità in un intervallo di numeri, se non in un numero ben preciso. Fare delle assunzioni: ovvero, delle ipotesi su cosa può succedere e sui motivi per cui tali cose accadono, in modo da creare una catena di eventi che porti al risultato sperato.

Le assunzioni che Wald dovette fare, in mancanza di dati, furono le seguenti:

1. Un aereo cade solo perché viene colpito: non precipita per mancanza di carburante o per infarto del pilota.
2. Un aereo non perde efficienza in modo progressivo a causa dei colpi subiti – se fosse un umano, diremmo che un aereo non viene ferito. Un colpo può essere fatale oppure innocuo, ma non procura danni parziali. In questo modo, Wald si risparmia di ipotizzare complicate funzioni che descrivano la perdita di efficienza del velivolo, man mano che da aliante si trasforma in colabrodo. Le cose per Wald sono discrete: o succedono o non succedono, non c'è una via di mezzo.

3. C'è un massimo di colpi che l'aereo può subire. Wald ipotizzò che questo numero di colpi massimo fosse n+1, ovvero il numero di colpi ricevuti dall'aereo più crivellato (il numero di pallottole su questo velivolo lo chiamava n) aumentato di uno.
4. Un aereo viene colpito in modo omogeneo: il numero di colpi che arrivano è indipendente dal punto dell'aereo in cui arrivano.

Ci sono due paradigmi focali, e tipici di uno statistico, nella descrizione che Wald fece del problema. Il paradigma principale è: c'è un limite a quello che può succedere. Vediamo questo modo di pensare in azione al numero 3: la probabilità che un aereo venga colpito all'infinito è zero, perché prima o poi l'aereo rientra o cade, e se cade è difficile pensare che continuino a sparargli. *Devo fissare questo limite a un numero di colpi ragionevole*. Il secondo paradigma è: ragionare per enti discreti.

Wald non cerca di calcolare la funzione di distribuzione dei colpi sull'aereo: non tratta il biplano come un insieme di coordinate, ma come un insieme di zone. Non lo fabbrica intagliandolo nel legno, ma con quattro pezzi di Lego: fusoliera, motore, alimentazione, ali. È meno accurato, ma molto più facile. E quando si hanno dati poco accurati, generati da un sistema complesso e in continua evoluzione, è il modo migliore di procedere.

È un po' quello che si fa quando si parla di squadre di calcio. Descrivere una squadra di calcio come composta da un portiere, tre difensori, cinque centrocampisti e due attaccanti non basta per vincere, certo; ma perlomeno viene schierata in modo corretto. Se descrivessimo i nostri giocatori sulla base delle zone del campo che coprono, e della velocità alla quale si muovono – funzioni continue, invece che enti discreti – rischieremmo di giocare con otto centrocampisti e tre portieri, di cui due in attacco. Commetteremmo un errore a un livello più profondo

di quello – fisiologico – che commettiamo schierando la squadra per ruoli.

Dividendo l'aereo in zone, Wald fu in grado di fare i suoi calcoli, e quindi di ottenere il primo risultato fondamentale, che non appariva chiaro agli ufficiali: *gli aerei che cadono sono stati colpiti almeno una volta al motore*. E il calcolo è un calcolo: probabilistico, ma – sulla base delle assunzioni fatte da Wald – rigoroso. Da questo risultato, il punto dove mettere le protezioni risultava chiaro.

Se notate, nelle due narrazioni c'è una differenza fondamentale: nella prima, quella romanzata, Wald prima indica dove mettere le protezioni e dopo spiega il motivo. Nella vita reale, invece, accade il contrario: Wald costruisce un sistema in cui alcune variabili, alcuni ingredienti necessari per risolvere il problema, mancano. Quindi fa le sue brave assunzioni per inserire nel sistema tali ingredienti. Da questi ricava il motivo per cui gli aerei vengono abbattuti *e quindi* gli appare ovvio dove mettere le protezioni.

La direzione della narrazione è contraria a quella del ragionamento rigoroso. Perché la narrazione deve impressionare, emozionare, stupire: è come saltare su un tappeto elastico da una piattaforma alta cinque metri. Lo sanno fare tutti. Il ragionamento invece consiste nel salire sopra a quel gradino: è un procedimento più lento, con una scala a pioli su cui arrampicarsi, ammesso che la scala già esista. Sennò va costruita.

Ecco: in questo libro, Hand vi mostra come costruire la vostra scala e dove trovare i pioli, quando per caso tale scala non c'è. Altrimenti, resterete dalla parte di quelli che le cose se le devono far raccontare dagli altri: emozionante, certo, ma sempre alla mercé di cosa gli altri decidono di spiegarci. È una scelta legittima, intendiamoci. Io cerco solo di dirvi che vi perdereste il cinquanta per cento del divertimento – scegliere da dove, e da quale altezza saltare.

Buona lettura.

Il tradimento dei numeri

A Shelley