Capitolo 1

Le sette insidie dei dati

Devi darti il permesso di essere umano.
– *Joyce Brothers*

Insidie dei dati. Chiunque abbia lavorato con i dati ne è stato vittima, molte volte. Io di sicuro. È come se avessimo usato i dati per lastricare la strada per un futuro migliore, ma la strada che abbiamo tracciato sia piena di crateri che sembra non riusciamo a vedere, finché non ci ritroviamo sul fondo, con il naso all'insù. A volte ci caschiamo senza nemmeno saperlo e scoprirlo molto più avanti può essere molto umiliante. Se avete già lavorato con i dati, conoscete questa sensazione. Date una presentazione importante, i vostri dati sono istruttivi oltre ogni aspettativa, i diagrammi e i grafici sono impeccabili, perfettamente conformi a tutte le indicazioni di Richard Tufte, il ragionamento che porta alla conclusione grandiosa è a prova di bomba e lascia tutti a bocca aperta. Poi un tizio in fondo alla sala (quello con le braccia incrociate e l'aria corrucciata) aspetta fino alla fine e vi chiede se siete consapevoli che state lavorando con un database fondamentalmente viziato, togliendovi il tappeto da sotto i piedi e facendovi finire in fondo all'ennesima trappola. C'è di che far venire i sudori freddi a un povero esperto dei dati.

La natura delle insidie dei dati è che siamo particolarmente inclini a non vederle. Ha senso, se ci si pensa. La specie umana non ha mai dovuto lavorare con miliardi di record di dati sotto forma di zero e uno fino alla seconda metà del secolo scorso. Solo un paio di decenni dopo, però, la nostra epoca è caratterizzata da un'abbondanza sempre crescente di dati e da una

In questo capitolo

- **Sette tipi di insidie dei dati**
- **Evitare le sette insidie**
- **"Sono caduto e non riesco a tirarmi su"**

schiera, in continuo ampliamento, di strumenti di incredibile potenza. Per molti versi, il nostro cervello non si è ancora messo al passo.
Queste insidie dei dati però non condannano al fallimento ogni nostro sforzo, tutt'altro. Abbiamo fatto grandi cose, in questa nuova era dei dati. Abbiamo mappato il genoma umano e abbiamo iniziato a comprendere la complessità del cervello, come i suoi neuroni interagiscono rendendo possibile la cognizione. Abbiamo riportato sulle carte astronomiche grandi galassie lontane, e abbiamo raggiunto una migliore comprensione dei fenomeni geologici e meteorologici proprio qui sul nostro pianeta. Anche nelle attività più semplici come gli acquisti per le feste, i motori di raccomandazione sui siti di e-commerce si sono evoluti, diventando incredibilmente utili. I nostri successi con i dati sono troppo numerosi per poterli elencare.
Stanno aumentando però anche i nostri scivoloni. Il cattivo uso dei dati ha causato gravi danni e perdite. Dal colossale fallimento degli analisti quantitativi e dei loro modelli nella crisi finanziaria dei prim anni del nuovo secolo alla parabola dei Google Flu Trends e a quello che ci ha insegnato sull'arroganza indotta dai dati (vedi `http://gking.harvard.edu/files/gking/files/0314policyforumff.pdf`), il nostro uso dei dati non è sempre tanto positivo; anzi, a volte è decisamente disastroso.
Perché? Semplicemente perché abbiamo la tendenza a ripetere continuamente certi tipi di errori. Notarli nelle prime fasi di un processo è molto facile, sempre che a commetterli sia qualcun altro. Quando sono io a prendere un abbaglio, sembra che non sia in grado di scoprirlo fino a che quel tizio in fondo alla sala non lancia la sua bomba.
Come i nostri buoni amici e colleghi, siamo tutti molto bravi a individuare gli errori degli altri, vero? Ne ho avuto una testimonianza precoce: durante una mostra di scienze alle scuole medie, un piccolo gruppo di studenti, scienziati in erba, ebbe la possibilità di andare in giro con i giudici e di spiegare i diversi progetti presentati, mentre gli altri aspiranti vincitori ascoltavano. I giudici, volendo stimolare il dialogo e lo spirito critico, incoraggiavano gli studenti anche a porre domande dopo ogni presentazione. Il loro intento era lodevole, ma noi fondamentalmente usavamo quell'occasione per individuare le falle nei metodi e nelle analisi degli altri concorrenti. I ragazzi sanno essere crudeli.
Non mi dedico più ai progetti di scienze per le mostre scolastiche, a differenza di altri genitori dei compagni di scuola dei miei figli, ma lavoro molto con i dati. E lavoro con altri che lavorano molto, a loro volta, con i dati. Con tutto il mio manipolare, remixare, analizzare, visualizzare e dedurre dai dati, ho notato che sulla strada che porta al paradiso dei dati si incontrano alcuni tipi specifici di insidie. In effetti, in base alla mia esperienza, queste insidie si possono raggruppare in sette categorie.

Sette tipi di insidie dei dati

Insidia 1: errori epistemici, ovvero come pensiamo ai dati

Che cosa possono dirci i dati? Forse ancora più importante: che cosa *non possono* dirci? L'epistemologia è quell'ambito della filosofia che si occupa della teoria della conoscenza: che cosa è una convinzione ragionevole anziché una semplice opinione. Spesso affrontiamo i dati con una mentalità e presupposti sbagliati, che portano a errori lungo tutta la strada, indipendentemente dal tipo di grafico che scegliamo, come i seguenti.

- Assumere che i dati che usiamo rispecchino perfettamente la realtà.

- Trarre conclusioni sul futuro sulla base solo di dati storici.
- Cercare di usare i dati per verificare una convinzione già formata, anziché sottoporla a prova per vedere se non sia in realtà falsa.

Evitare gli errori epistemici, e assicurarsi di avere le idee chiare su ciò che è ragionevole e ciò che non lo è, è un fondamento importante per un'analisi dei dati efficace.

Insidia 2: sconfinamenti tecnici, ovvero come elaboriamo i dati

Una volta deciso di usare i dati per risolvere un particolare problema, dobbiamo raccoglierli, memorizzarli, unirli con altri set di dati, trasformarli, pulirli e metterli nella forma giusta. Ciò che si ottiene può presentare molti problemi.

- Dati "sporchi", con livelli categoriali non corrispondenti ed errori ortografici nell'inserimento dei dati.
- Unità di misura o campi di data che non sono coerenti o non sono compatibili.
- Unione di set di dati non omogenei, record nulli o duplicati che distorcono l'analisi.

Possono essere aspetti complicati e fonte di confusione, ma l'accuratezza dell'analisi dipende dal fatto che siano trattati correttamente. A volte la verità contenuta nei dati viene "persa nella traduzione" ed è possibile andare avanti a testa bassa e prendere decisioni senza nemmeno rendersi conto che si sta lavorando con un insieme di dati seriamente difettoso.

Insidia 3: errori matematici, ovvero come calcoliamo i dati

Lavorare con i dati comporta quasi sempre effettuare dei calcoli, eseguire operazioni matematiche sui dati quantitativi a nostra disposizione.

- Si sommano i dati a vari livelli di aggregazione.
- Si calcolano percentuali o rapporti.
- Si lavora con proporzioni o percentuali.
- Si tratta con unità di misura diverse.

Questi sono solo alcuni esempi di come prendiamo dei campi di dati già esistenti e da quelli creiamo nuovi campi di dati. Come alle scuole elemntari, è possibilissimo commettere errori nei calcoli, errori che possono avere un costo enorme: un errore di questo tipo ha causato la perdita nel 1999 di un Mars orbiter da 125 milioni di dollari (`http://www.cnn.com/TECH/space/9909/30/mars.metric.02/`). In quel caso, è stato più come finire in un buco nero che cascare in una trappola.

Insidia 4: lapsus statistici, ovvero come confrontiamo i dati

"Esistono le bugie, le maledette bugie e la statistica." Questa battuta di solito implica che qualcuno stia manipolando i numeri per trarre in inganno altri, ma nel campo della statistica capita altrettanto spesso che mentiamo a noi stessi. Che si parli di statistica descrittiva o inferenziale, le insidie abbondano.

- Le misure di tendenza centrale o di variazione che usiamo ci portano fuori strada?
- I campioni che usiamo sono rappresentativi della popolazione che vogliamo studiare?

- Gli strumenti di confronto che usiamo sono validi e statisticamente corretti?

Queste insidie sono numerose e particolarmente difficili da individuare, perché hanno a che fare con un modo di pensare che può indurre all'errore, a volte, persino gli esperti. Un "semplice campione casuale" può essere tutto fuorché semplice da selezionare, e provate qualche volta a chiedere a un esperto di dati che vi spieghi in termini comprensibili a un non addetto ai lavori che cos'è un "valore p".

Insidia 5: aberrazioni analitiche, ovvero come analizziamo i dati

L'analisi è al centro di ogni lavoro sui dati. È l'insieme degli strumenti grazie ai quali traiamo conclusioni e prendiamo decisioni. Esistono molte persone che hanno, nella loro qualifica, il termine "analista", ma in realtà l'analisi dei dati è un'attività che praticamente chiunque finisce per svolgere, prima o poi. L'analisi dei dati ha raggiunto nuove vette, ma si può finire anche in fondo a qualche nuovo baratro, come nei casi seguenti.

- Quando si ottiene un iper-adattamento dei modelli ai dati storici.
- Quando si trascurano nei dati segnali importanti.
- Quando si estrapola o si interpola in modi che non hanno alcun senso.
- Quando si usano metriche che in realtà non hanno alcuna importanza.

Era davvero ragionevole ipotizzare che le tendenze di ricerca potessero consentirci di prevedere accuratamente il numero delle persone che si sarebbero ammalate di influenza, anche sapendo che gli algoritmi di ricerca cambiano costantemente e che la popolazione che effettua ricerche reagisce a input come le notizie pubblicate e pompate dai media e le raccomandazioni del motore di ricerca?

Insidia 6: gaffe grafiche, ovvero come visualizziamo i dati

Sono gli errori che si notano più facilmente e di cui si parla di più. Perché? Perché sono quelli visuali. Sono lì in bella vista e tutti possono rimirarli orripilati. Sapete di che cosa parlo: splendidi grafici a torta con decine di fette, istogrammi fuorvianti con assi y che partono dalla metà del valore massimo. Per fortuna queste insidie sono ben documentate e si possono identificare ponendosi poche domande.

- Abbiamo scelto un tipo di grafico adatto allo scopo?
- Se si deve evidenziare un punto, è mostrato chiaramente o bisogna cavarsi gli occhi per riuscire a vederlo?
- Usiamo delle regole empiriche senza esserne indebitamente limitati?

Certo, trovare il tipo di grafico perfetto è del tutto inutile, se siamo caduti in una delle prime cinque insidie, ma che peccato quando si fanno tutte le cose giuste fino a questo punto e si casca proprio sul traguardo.

Insidia 7: difetti di design, ovvero come abbelliamo i dati

Siamo esseri umani e apprezziamo il buon design. Andiamo al lavoro in automobili ben progettate, con tutti i controlli al posto giusto, ci sediamo alla scrivania in poltroncine ergonomiche che si adattano alla perfezione al nostro corpo. Perché sedere comodi e

poi aprire il browser per guardare un'infografica pacchiana o un goffo cruscotto di dati? Il design conta.

- La scelta dei colori confonde chi guarda o aiuta a rendere le cose più chiare?
- Abbiamo usato la nostra creatività per abbellire sapientemente i grafici, o ci siamo persi una buona occasione per includere componenti estetici che aggiungono valore?
- È facile interagire con gli oggetti visuali che abbiamo creato, oppure disorientano l'utente?

Mettere gli elementi di design giusti al posto giusto può fare la differenza: il nostro pubblico presterà attenzione al nostro messaggio, oppure ci ignorerà totalmente e presterà attenzione invece a qualche altra cosa?
Queste sette insidie sono come i sette peccati capitali: ciascuna può vanificare tutti i nostri sforzi di trattamento dei dati. Non ha senso però temerle; meglio imparare come tirarsi fuori rapidamente, se si cade in qualcuna o, meglio ancora, imparare a evitarle del tutto. Come si fa?

Evitare le sette insidie

Quando si incontra un'insidia su un particolare sentiero nel mondo reale, vorremmo che ci fosse un bel cartello segnalatore, che ci avverta utilmente e ci metta in guardia dal pericolo, come quello che si trova sul sentiero di Coal Creek Falls vicino a casa mia a Bellevue, Washington (Figura 1.1).

Figura 1.1 Un minaccioso cartello avverte di un pericolo sul sentiero che porta a Coal Creek Falls, vicino a Bellevue, Washington.

Nel caso delle insidie dei dati, però, questi utili segnali di pericolo non esistono, no? Sta a noi conoscere bene queste insidie cognitive, procedurali e comunicative e sapere perché è facile cascarci. La chiave è la consapevolezza. Se non conosciamo queste brutte trappole, che aspetto hanno, come si fa a individuarle, quali sono i loro tratti identificativi, è molto più probabile che ci cadiamo. Questo è ovvio.
Ma semplicemente conoscerle spesso non basta. Anche il più saggio degli esperti dei dati ogni tanto cade in questi tranelli ben nascosti. Abbiamo bisogno di qualche consiglio adatto e di qualche guida fidata che ci aiuti lungo la strada.
A partire dal prossimo capitolo, dedicato agli errori epistemici, cominceremo a raccogliere qualche suggerimento pratico che ci aiuterà a evitare ciascuna delle sette insidie, in modo da poter rimanere sulla strada, diritta e stretta, dei dati. Dopo aver analizzato, nel Capitolo 8 dedicato ai pericoli del design, tutti i sette tipi di insidie, avremo una lista di contollo completa, che potremo usare come una sorta di mappa dei sentieri per il nostro tragitto.

"Sono caduto e non riesco a tirarmi su"

Il fatto è, però, che spesso non abbiamo il tempo per scorrere una lista di controllo completa, prima di procedere nel nostro viaggio fra i dati. Le richieste urgenti del business e gli ambienti dai ritmi accelerati in cui operiamo in genere ci presentano scadenze dietro l'angolo e l'esigenza di generare idee dai dati in tempi più brevi di quelli di cui avremmo realmente bisogno.
In questi casi, probabilmente non abbiamo altra alternativa che continuare ad andare avanti, ma almeno possiamo usare la "Lista di controllo per evitare le insidie dei dati", che presenteremo nell'ultimo capitolo come un "esame autoptico" utile a identificare le nostre peculiari inclinazioni, e per capire in quali insidie finiamo per cadere continuamente.
E succederà di sicuro. Vi garantisco che in un futuro molto vicino cadrete nell'una o nell'altra di queste trappole. E lo stesso succederà ai vostri colleghi. E anche a me. Probabilmente sono caduto in più di una anche in questo stesso libro. Come specie, stiamo ancora imparando come modificare il nostro modo di pensare per adeguarci alla relativa novità di questo medium.
Pensando ai tempi lunghi dell'evoluzione, il dover interagire con grandi fogli di calcolo e database non è solo qualcosa di nuovo, è una novità dell'ultima ora. I primi esseri umani moderni dal punto di vista anatomico compaiono nella documentazione fossile circa 195.000 anni fa in Africa e Alan Turing, pioniere fra gli scienziati informatici, ha proposto l'idea del computer moderno in un influente articolo del 1936 (`https://www.virginia.edu/robins/Turing_Paper_1936.pdf`), circa 80 anni fa. Questo significa che ci stiamo acclimatando all'era informatica da non più di uno 0.04 per cento di tutta la storia della specie umana, che corrisponde, sulla scala di una giornata, agli ultimi 35 secondi, fra le 23:59:25 e le 24:00:00.
Certo, poi succederà. Ma come reagiamo a quel punto? Dovremmo considerare questi errori un passaggio ineludibile nel processo di sviluppo di un senso acuto dell'orientamento. Avete mai sentito parlare del salasso, quella pratica terribile, per cui si cavava sangue da un paziente per trattare le malattie? Nelle aule di tutto il mondo, i giovani di oggi inorridiscono alla follia di questa pratica barbarica, quando ne sentono parlare, ma per 2000 anni è stata una tecnica medica comune, dall'antichità fin verso la fine del XIX secolo.

Proprio come i nostri antenati, anche la nostra generazione commette regolarmente errori stupidi che le generazioni future troveranno sconcertanti. Spero che il cadere nelle insidie dei dati sarà fra quelle inclinazioni umane che le future generazioni considereranno incomprensibili.

Che cosa succede, dunque, quando finiamo sul fondo di una brutta trappola dei dati? Che cosa facciamo? Per naturale inclinazione, tenderemmo a far finta che non sia mai successo, a mascherare l'errore, nella speranza che nessuno lo scopra, ma è proprio il contrario di quello che dovremmo fare.

- Per prima cosa, cercare di uscire dalla trappola: risolvere l'errore.
- Secondo, mettere un segno di spunta sulla lista di controllo, vicino alla trappola in cui si è caduti.
- Terzo, raccontare a tutti quello che è successo.

Questo processo, per quanto possa sembrare un modo per autofustigarsi, ci aiuterà a sviluppare la capacità di lavorare efficacemente con i dati. Per assicurarci che anche gli altri possano seguire questo processo, dobbiamo astenerci dal denigrare quanti cadono nelle insidie dei dati. Ricordate, non solo avrebbe potuto capitare a voi, ma di sicuro *capiterà* a voi, prima o poi, lungo la strada.