RON S. KENETT | THOMAS C. REDMAN

LA DATA SCIENCE NELLA REALTÀ

COME TRASFORMARE I DATI IN INFORMAZIONI, DECISIONI MIGLIORI E ORGANIZZAZIONI PIÙ FORTI

Edizione italiana a cura di

Giancarlo Manzi e Silvia Salini

IUSTITIAM COLIMUS

G. Giappichelli Editore

Prefazione all'edizione italiana

Era gennaio 2020, Ron Kenett era da noi come visiting e teneva un corso sul significato della data science ai nostri studenti della laurea magistrale di Data Science and Economcis all'Università degli Studi di Milano. Eravamo nell'ufficio di Silvia, stavamo parlando del profilo del data scientist e di cosa le aziende si aspettano dai nostri studenti, quando Ron ci ha chiesto se ci avrebbe fatto piacere curare l'edizione italiana del suo libro. Silvia coordina da tre anni la laurea magistrale in inglese in Data Science and Economics dell'Università degli Studi di Milano (https://dse.cdl.unimi.it/en), Giancarlo coordina da tre anni nella stessa università il Master in Data Science for Economcis, Business and Finance (https://ceeds.unimi.it/master-data-science/); quest'ultimo è in italiano. In questi anni Silvia e Giancarlo hanno lavorato tanto, insieme ad altri colleghi che insegnano sia alla laurea magistrale che al master e che afferiscono al Data Science Research Center dell'Università di Milano (https://datascience.unimi.it), per creare solidi rapporti di collaborazione con aziende e parti sociali. Hanno lavorato anche per fare in modo che i programmi formativi della laurea magistrale e del master forniscano ai loro studenti tutti gli elementi che serviranno loro per essere competitivi sul mercato del lavoro. Sia la laurea magistrale che il master cercano di offrire agli studenti esperienze che accrescano i loro *soft skill* e le loro competenze tecniche. Sia alla laurea magistrale che al master sono previsti laboratori svolti direttamente dalle aziende che propongono ai ragazzi *hackathon*, casi studio e problemi aziendali. Il valore aggiunto di questi laboratori non è solo legato ai contenuti reali, ma c'è anche il fatto che viene mostrato ai ragazzi come funziona il lavoro in azienda, in *team* multisciplinari, con deadline ravvicinate e la necessità di comunicare i risultati a chi non è un data scientist ma deve prendere decisioni. Il libro di Ron e Thomas racconta tante cose che ci siamo spesso detti tra noi e che ogni data scientist dovrebbe leggere. Abbiamo accettato con entusiasmo.

La versione italiana è arricchita di cinque appendici. Le prime tre sono

scritte da tre colleghi che lavorano come data scientist in Italia in tre ambiti molto diversi, *fintech* il primo (AcomeA), *retail* il secondo (Oney data) e *media* e *communication* (Sky) il terzo. I Dott.ri Giuseppe Codazzi, Filippo Manfroni e Francesco Ranucci hanno svolto lo scorso anno tre laboratori molto apprezzati dai nostri studenti della Laurea Magistrale in Data Science and Economics. Quest'anno li faranno sia alla magistrale che al master. I primi due sono stati anche studenti del nostro master nella sua prima edizione ormai 3 anni fa. Abbiamo deciso di aggiungere queste appendici per poter fare raccontare a loro, che lavorano sul campo, chi è o cosa fa il data scientist in Italia oggi.

Le ultime due appendici parlano di Covid-19, di pandemia e di dati. Era inevitabile, abbiamo iniziato a tradurre il testo mentre eravamo in pieno *lockdown*. La prima appendice, scritta da un nostro studente, Andrea Cutrera, racconta come si è sentito un futuro data scientist di fronte all'*infodemia* che ci ha travolto insieme alla pandemia. La seconda racconta cosa abbiamo fatto noi, statistici e ricercatori, quando abbiamo cominciato a leggere le notizie che arrivavano dai giornali e dai *social media*. Racconta di come ci siamo ritrovati uniti sul fronte della ricerca nella data science applicata all'epidemiologia per combattere il virus. Non nascondiamo le difficoltà del lavoro di traduzione svolto nel corso dell'anno 2020, tutti impegnati nella didattica online e alle prese con le nuove piattaforme e le nuove difficoltà.

Ci auguriamo allora che l'opera tradotta possa essere apprezzata e utile per lo sviluppo di tante nuove carriere di giovani appassionati nella data science.

Vogliamo ringraziare la Dr.ssa Francesca Tozzi, responsabile editoriale della linea universitaria di Giappichelli, che ha creduto subito nel potenziale di questo testo e di questa traduzione, ci ha supportato e ci ha seguito con pazienza e cortesia. Un ultimo doveroso e affettuoso ringraziamento va alle nostre famiglie che ci hanno sostenuto anche per questa parte del lavoro in un momento così difficile per tutti. Ringraziamo anche gli autori del volume per alcuni preziosi consigli sull'impostazione della traduzione.

Milano, aprile 2021

Giancarlo Manzi
Silvia Salini

Dipartimento di Economia, Management e Metodi Qualtitativi
e Data Science Research Center
Università degli Studi di Milano

1

Una vocazione superiore

È decisamente un momento importante per la data science! L'Economist ha solennemente proclamato che i dati sono "*la risorsa più preziosa al mondo*" [1] e Hal Varian e Tom Davenport hanno più volte definito la statistica e la data science "*il lavoro più sexy del ventesimo secolo*" [2]. Cercando sul web il termine *data scientist*, si trova la seguente definizione: "*Con 'data scientist' si intende un professionista che utilizza metodi scientifici per far emergere e dare significato ai dati grezzi*" [3]. Analoghe definizioni sono state date per indicare gli statistici e gli analisti dei dati [4]. Tuttavia, noi crediamo che il lavoro del data scientist sia più complesso e vada ben oltre la semplice attività di dare significato ai dati grezzi.

Questo libro amplia e chiarisce che cosa serve per avere successo in questo lavoro, all'interno dell'ecosistema organizzativo in cui ha luogo. Si basa su anni di esperienza in un'ampia gamma di organizzazioni in tutto il mondo. Il nostro obiettivo è condividere questa esperienza e qualche conoscenza acquisita svolgendo questo lavoro sul campo. Nello specifico, proponiamo l'idea che il vero lavoro dei data scientist e degli statistici consiste nell'aiutare le persone a prendere decisioni migliori riguardo a questioni importanti a breve

[1] Copertina del numero del 6 maggio 2017.

[2] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century (Davenport e Patil 2012).

[3] http://www.datascienceassn.org/code-of-conduct.html (Data Science Association 2018).

[4] Nel corso di questo libro useremo i termini *data science*, *data analytics*, and *statistica* come sinonimi, anche se siamo consci del fatto che molte persone potrebbero trovare delle differenze di significato tra questi termini. Tuttavia queste differenze di significato non verranno considerate in questo libro.

The Real Work of Data Science, prima edizione. Ron S. Kenett e Thomas C. Redman.

Pubblicato nel 2019 da John Wiley & Sons Ltd. Sito internet: www.wiley.com/go/kenett-redman/datascience

termine e rafforzare le organizzazioni e le loro abilità a lungo termine. Con il termine "persone" intendiamo, tra gli altri, manager di organizzazioni e professionisti nel settore terziario e di produzione. Questa prospettiva è rilevante anche per gli insegnanti nelle scuole e nelle università e per i ricercatori nei laboratori e nelle istituzioni accademiche. È una "vocazione" nettamente superiore e molto più impegnativa. Per esempio, in generale non viene data la possibilità a chiunque di partecipare al processo decisionale se si ha a che fare con questioni molto importanti, a meno che non si sia considerati una persona fidata.

Dunque, il vero lavoro del data scientist richiede una partecipazione totale: aiutare a formulare i problemi e le opportunità in modo chiaro, utilizzando un linguaggio commerciale o scientifico; capire quali dati considerare e i loro punti di forza e limiti; determinare quando sono necessari nuovi dati; avere a che fare con problemi di qualità; usare dati per ridurre l'incertezza/il margine di errore; chiarire fino a che punto i dati sono necessari e quando invece deve subentrare l'intuizione; presentare i risultati in modo semplice ed efficace; riconoscere che tutte le decisioni importanti rientrano in una realtà politica; lavorare in gruppo; supportare le decisioni in modo pragmatico. Questo aspetto del lavoro non è spesso insegnato nei corsi di statistica o data science.

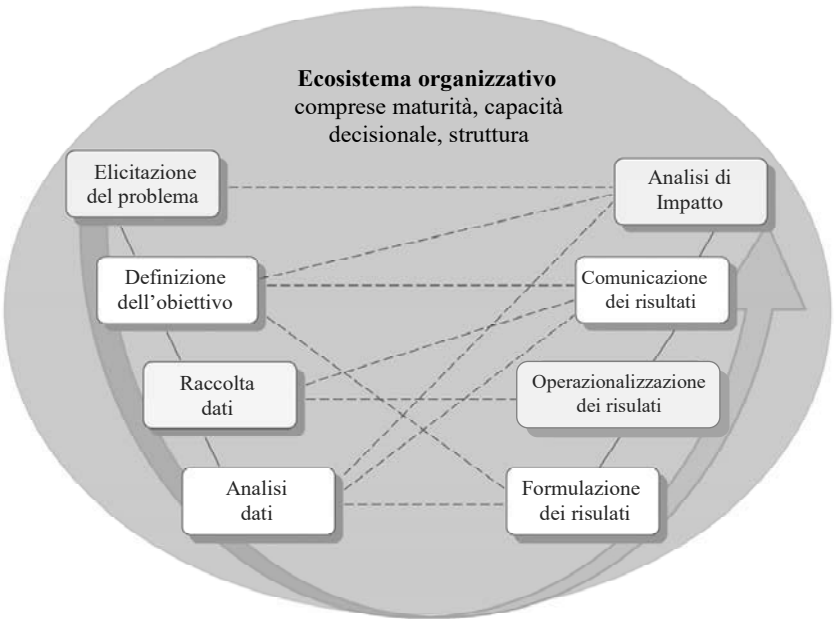

Figura 1.1. La visione del ciclo di vita dell'analisi dei dati, nel contesto dell'ecosistema organizzativo in cui si svolge il lavoro.

L'amara verità è che la maggior parte delle aziende ottengono solo una parte del valore che i loro dati, la data science e la statistica offrono (si veda, per esempio, Henke et al. 2016). I data scientist e i loro manager, *chief analytics officer* (CAO) compresi, i chief data scientist, i dirigenti del reparto data science, e altri professionisti che assumono data scientist[5], devono imparare come gestire le difficoltà che ostacolano il percorso. Quindi, il vero lavoro comprende anche quello di fare in modo che tutti siano in grado di condurre analisi semplici e di comprendere anche quelle più complesse, di capire le potenzialità dei dati, capire la variabilità e integrare i dati con l'intuizione; saper dare i giusti incarichi ai data scientist e agli statistici; educare i dirigenti senior a considerare le potenzialità dei dati; aiutarli a diventare buoni consumatori della data science; insegnare loro i ruoli che ricopriranno in modo che l'analisi progredisca; creare le strutture organizzative necessarie per far sì che tutto ciò avvenga in modo effettivo e (ragionevolmente) efficiente. Di questo tratterà questo libro.

Per fornire il valore aggiunto di cui stiamo parlando è necessario avere una mente aperta. La Figura 1.1 rappresenta il ciclo vitale dell'analisi dei dati nel contesto di un'organizzazione il cui obiettivo è quello di ricavare profitto dalla data science (adattato da Kenett 2015). Come si evince dall'illustrazione nella figura, si tratta di un lavoro altamente iterativo (per sapere di più su questo processo, si veda Box 1997).

Il ciclo di vita

La prospettiva del ciclo di vita è pensata per aiutare i data scientist ad aiutare i *decision maker*. Consideriamo il ciclo passo per passo.

Elicitazione del problema: capire il problema

Si tenga presente cosa succede quando si va dal dentista: si fornisce al dentista un'idea dei sintomi, si viene invitati ad accomodarsi su una poltrona, il dentista osserva i denti, presenta una diagnosi e (si spera) risolve il problema, e dice quando tornare, tutto in meno di un'ora.

Un data scientist esperto fa di meglio. Descriveremo in dettaglio l'attività del data scientist nel capitolo 2. I data scientist ascoltano attentamente e inda-

[5] Anche in questo caso si potrebbero fare delle distinzioni sottili tra questi ruoli, ma anche in questo caso useremo questi termini in modo interscambiabile.

gano, mantenendo i clienti (i decision maker) concentrati e ottenendo dettagli rilevanti per capire quali siano le loro esigenze. Potrebbe trattarsi di un *operations manager* che sta affrontando ingenti spese a causa di una rilavorazione, un *marketing manager* che sta cercando di entrare in un nuovo mercato, o un responsabile delle risorse umane che vuole ridurre il *turnover* dei dipendenti.

Il data scientist esperto è anche in grado di interpretare il linguaggio del corpo del cliente e ottenere informazioni senza che questi apra bocca: il cliente ha secondi fini? Sta cercando di mettere qualcuno in cattiva luce? O vuole trovare supporto per una controversia? Come molti altri non lo ripeteremo mai abbastanza: bisogna capire a fondo il vero problema se si spera di risolverlo. La qualità del lavoro di analisi dipende da questo (Kenett e Shmueli 2016a). Maggior dettagli su questo verranno forniti nei capitoli 3 e 4.

Formulazione dell'obiettivo: chiarire gli obiettivi a breve e a lungo termine

Non ci si deve aspettare che il decision maker fornisca in modo chiaro i termini del problema. Bill Hunter, un famoso statistico della Università del Wisconsin-Madison, racconta la storia di due farmacisti che chiesero il suo aiuto. Quando fu chiesto loro di descrivere il problema, si immersero in una lunga discussione con Bill che li portò a riformulare il loro problema. Questa formulazione era molto più semplice da risolvere rispetto alla prima formulazione e quindi non ebbero più bisogno dell'aiuto di Bill. Lasciarono il suo ufficio dopo averlo ringraziato profusamente (Hunter 1979). Nonostante sembri che l'aiuto di Bill sia stato minimo, in realtà fu essenziale! Il punto è che per capire in profondità il problema è necessario capire a fondo il contesto in cui esso ha luogo, contesto che comprende anche l'obiettivo principale. Maggiori dettagli verranno forniti nel capitolo 4.

Raccolta dei dati: identificare le fonti rilevanti e raccogliere i dati

Cobb e Moore (1997) sottolineano che "la statistica esige un modo di pensare differente, perché i dati non sono semplici numeri, sono numeri all'interno di un contesto preciso". Il contesto aiuta ad identificare le fonti rilevanti da cui ricavare dati e la loro interpretazione.

Per comprendere meglio, si consideri questa storia proveniente dalla Danimarca raccontata da Kenett e Thyregod (2006). Riguarda un esercizio presente in un testo scolastico di quarta elementare e dimostra l'importanza del

contesto e il modo in cui i numeri si trasformano in dati. In questo esercizio, i numeri presentati nella Figura 1.2 riportano la quantità di gelati venduti ogni giorno, senza indicare però il giorno della settimana. A luglio, è stato molto caldo per nove giorni consecutivi. Venne chiesto agli studenti di (i) identificare i giorni in cui aveva fatto caldo e (ii) determinare quali giorni erano domeniche. Il grafico presenta soltanto 31 numeri, ma gli studenti danesi sanno che i loro genitori sono più inclini ad offrire loro del gelato nel fine settimana o durante i giorni più caldi. Conoscendo il contesto, non fu difficile per gli studenti completare il compito. Il contesto viene rivelato nel luogo stesso in cui i dati sono generati, che sia l'officina, il laboratorio o l'ambiente dei social media. I data scientist devono capire questo contesto e identificare i dati pertinenti al problema. Maggiori dettagli verranno forniti nel capitolo 5.

Analisi dei dati: utilizzare metodi descrittivi, esplicativi e predittivi

Qui si tratta di fare il lavoro di "estrarre il significato dai dati", "separare il segnale dal rumore", "trasformare i dati in informazioni" e così via. Ci sono letteralmente migliaia di esempi. Prendiamo le aste di eBay. Quando si vuole vendere un oggetto su eBay, viene chiesto di specificare un "prezzo di riserva", ovvero stabilire il valore da cui avrà inizio l'asta.

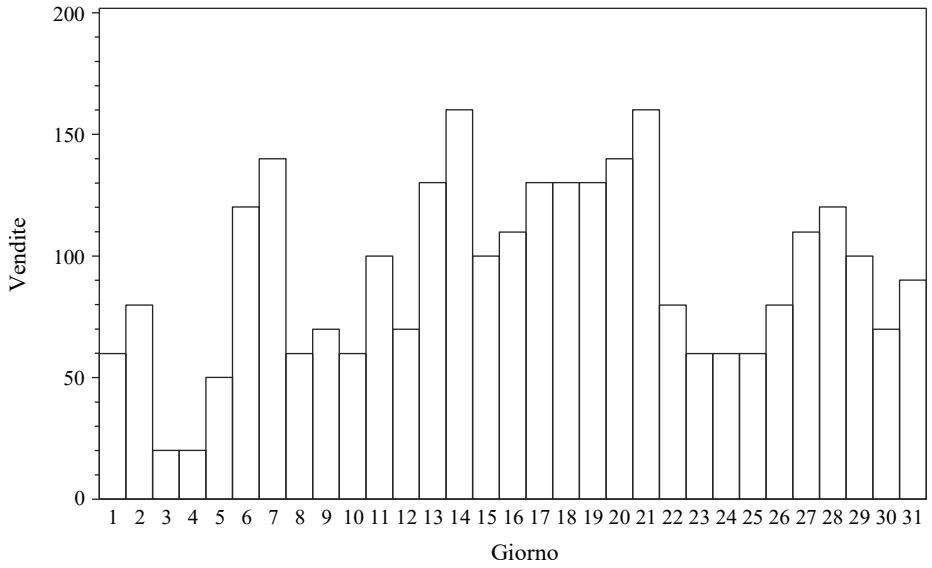

Figura 1.2. Il numero di gelati venduti in una località danese, dati giornalieri di luglio.

Se il prezzo finale non è superiore al prezzo di riserva, l'asta viene annullata e l'oggetto resta invenduto. Su eBay i venditori possono scegliere se impostare un prezzo di riserva pubblico, dunque visibile agli offerenti, o un prezzo di riserva segreto (in questo caso gli offerenti sanno soltanto che un prezzo di riserva è stato impostato, ma non sanno a quanto ammonta).

Katkar e Reiley (2006) hanno indagato gli effetti di questa scelta. Hanno ottenuto i loro dati attraverso un esperimento, che consistette nel vendere 25 paia di carte Pokémon identiche. Ogni carta venne messa all'asta due volte: prima con un prezzo di riserva pubblico e poi con un prezzo di riserva segreto. Utilizzarono quindi le informazioni di tutte e 50 le aste. Applicarono modelli di regressione lineare e test di significatività per quantificare il possibile effetto del prezzo di riserva pubblico/segreto sul prezzo finale. Arrivarono alla conclusione che "un oggetto all'asta con un prezzo di riserva segreto viene venduto in media per 0,63 dollari in meno", una semplice dichiarazione che chiunque può comprendere.

Questo genere di lavoro non ci interessa molto, ma comprende un aspetto che non viene considerato abbastanza nella formazione in data science. La cruda, amara verità è che troppi dati non sono adatti all'analisi (Nagle et al. 2017) e i data scientist spendono molto più tempo risolvendo problemi riguardo alla qualità dei dati che sull'analisi stessa. Dati di alta qualità sono essenziali per ogni analisi, soprattutto per le tecnologie cognitive (Redman 2018b). Dunque i data scientist devono fare i conti con questo. Maggiori dettagli verranno forniti nel capitolo 6.

Formulazione dei risultati: fornire i risultati e le raccomandazioni

Gli analytics producono output come la statistica descrittiva, il p-value, modelli di regressione, tabelle di analisi della varianza (ANOVA), carte di controllo, alberi decisionali, foreste casuali, reti neurali, dendrogrammi e così via. Molti di questi output sono sconosciuti ai decision maker. Quindi è essenziale che i data scientist traducano i risultati ottenuti in una lingua che questi possano comprendere.

I data scientist devono esplorare le implicazioni dei loro risultati e, spesso, raccomandare modi di agire specifici. In altre parole, i data scientist non possono semplicemente buttare lì i risultati e le raccomandazioni da dare ai decision maker. Anzi, devono invece accertarsi che questi capiscano a fondo ciò che hanno scoperto e il contesto. Dal momento che, quando si tratta di decisioni importanti, molte persone sono coinvolte, questo può significare che si debba interagire con un gran numero di interlocutori e fare distinte

presentazioni ai dirigenti, così come ai funzionari ed infine ai semplici impiegati.

Concetti e notazioni tipici della statistica matematica risultano noiosi a molte persone. Grafici e illustrazioni ben fatte, invece, sono il mezzo di comunicazione più efficace. I risultati che non possono essere rappresentati attraverso un grafico, probabilmente non vale la pena che siano presentati. Ci si deve assicurare quindi che i grafici e le slide siano semplici e mantenere il "rapporto tra inchiostro e informazione" basso, evitando di utilizzare simboli e caratteri speciali (Tufte 1997). Un semplice esempio viene fornito nel capitolo 7.

Un bell'esempio è quello di un analista che si rese conto che i dirigenti responsabili non capivano i termini tecnici associati al problema posto riguardante la robustezza della rete. Dunque, rese il problema più chiaro e presentò i suoi risultati usando come esempio una celebre fiaba: "Prima di tutto dobbiamo decidere che tipo di rete vogliamo: un cucciolo, una mamma orso o un papà orso. In parole povere questo significa …". Tutti capirono.

Nonostante la decisione in sé sia presa da altri, nel modello del ciclo di vita ci si aspetta che l'analisi supporti una decisione, anche se provvisoria, come conclusione del processo.

Operazionalizzazione dei risultati: indicare chi, che cosa, dove e come

Il lavoro del data scientist non si conclude quando viene presa una decisione; il suo compito è infatti anche quello di seguire il processo attraverso cui le decisioni prese in seguito all'analisi dei dati saranno messe in atto, aiutando a definire il modo in cui i risultati saranno applicati nella pratica (ad esempio con procedure operative), rispondendo alle domande che verranno certamente poste, analizzando i nuovi dati nel momento in cui arrivano e dando consigli in situazioni oltre il limite dell'analisi originale.

Molti preferirebbero saltare questo passaggio. Ma il valore della data science si accresce solo quando l'analisi e la decisione sono messe in atto, non prima. Maggiori dettagli verranno dati nel capitolo 8.

Comunicazione dei risultati: comunicare i risultati, le decisioni e ciò che queste implicano per gli stakeholders alle parti interessate

Finora il numero di persone che abbiamo ipotizzato essere coinvolte nel processo di cui abbiamo parlato è stato relativamente piccolo. Ma le decisioni importanti hanno un forte impatto sulla vita di migliaia, addirittura milio-

ni di persone. A questo punto le scoperte devono essere comunicate a tutti coloro che potrebbero subirne le conseguenze, e questo pubblico è molto più ampio di quello coinvolto nel processo decisionale. Nonostante una grande parte di questo lavoro sia l'ambito del decision maker, il data scientist è attivamente impegnato in un ruolo di supporto.

La valutazione dell'impatto: pianificare e attuare una strategia di valutazione

Benché vada al di là del compito di supporto ai decision maker, i data scientist dovrebbero valutare l'impatto del loro lavoro. Ove possibile, si dovrebbero ottenere numeri precisi. Certo – come dimostra l'episodio di Bill Hunter – ciò non è sempre possibile. E anche quando vengono ottenuti numeri precisi, un feedback da parte dei decision maker dovrebbe essere sempre richiesto. Bisognerebbe essere poi completamente onesti nel valutare la propria performance e migliorare in futuro. Maggiori dettagli nel capitolo 9.

L'ecosistema organizzativo

Il lavoro dei data scientist ha luogo in ambienti organizzativi complessi, che possono sia aumentare sia ridurre la sua efficacia e a volte tutte e due le cose. I data scientist e i direttori amministrativi devono esserne consapevoli e, nel tempo, migliorare le diverse componenti dell'"ecosistema organizzativo".

Il termine *data-driven* ("basato sui dati") è entrato stabilmente nei dizionari. Spesso si possono trovare definizioni anche stravaganti sul marketing data-driven, o sulla ricerca di personale data-driven, e sulle tecnologie data-driven.

Al di là delle mode, e più in profondità, nel cosiddetto data-driven c'è un nucleo concettuale potente che porta a decisioni migliori e organizzazioni più forti. In fondo, più l'organizzazione è basata sui dati, più i decisori sono esigenti nei confronti dei data scientist, più prendono sul serio analisi sofisticate e più investono in dati di alta qualità. Pertanto, i data scientist intelligenti e i CAO investono molto tempo nella formazione di se stessi e dei decision maker a tutti i livelli su questo importante approccio e collaborano per portarlo avanti nelle organizzazioni.

Discuteremo più nel dettaglio che cosa significhi *data-driven* nel capitolo 10. Non sorprende che il concetto della cosiddetta bias ("la distorsio-

ne"), in qualsiasi forma, sia diametralmente opposto al processo decisionale basato sui dati. Il primo passo per i data scientist è rimuovere i pregiudizi sul proprio lavoro – un argomento che tratteremo nel capitolo 11. Il fulcro dei capitoli 12-14 è l'istruzione. In primo luogo, il capitolo 12 consiglia ai data scientist di iniziare dalle basi quando devono interfacciarsi con i loro colleghi e gli altri responsabili delle decisioni. Il capitolo 13 prende una strada leggermente diversa. Riconosce che i clienti esigenti (ad esempio i responsabili delle decisioni) faranno tanto per promuovere una cultura basata sui dati e una scienza dei dati quanto qualsiasi altra cosa. Quindi, il capitolo fornisce un elenco di domande per aiutare i decisori a sapere cosa chiedere.

Con i big data, l'intelligenza artificiale (IA), i problemi di sicurezza, il regolamento generale sulla protezione dei dati (GDPR), la digitalizzazione e molto altro ancora nelle notizie, è difficile per i dirigenti senior vedere lo spazio dei dati in prospettiva. Il capitolo 14 considera il quadro generale, consigliando ai CAO come sviluppare una prospettiva ampia e profonda sullo spazio dei dati e di aiutare i leader più anziani della loro organizzazione a comprendere i rischi e le opportunità.

Struttura organizzativa

L'amara verità è che la posizione dei data scientist all'interno di un'organizzazione è importante e stabilisce che cosa possono effettivamente fare. Per esempio un data scientist che lavora nel reparto di manutenzione potrebbe non avere accesso a dati rilevanti del reparto operativo per la sola ragione che i dirigenti dei due reparti sono in lizza per la stessa promozione. Nonostante i data scientist credano di essere superiori a tutto, nessuno sfugge alla politica. È meglio che i data scientist e i CAO accettino la realtà e si impegnino per arrivare ognuno ad occupare la giusta posizione. Maggiori dettagli verranno forniti nel capitolo 15.

Maturità organizzativa

Infine le organizzazioni fanno usi diversi della data science, a seconda della loro maturità. Essi possono variare di molto: c'è chi ha bisogni immediati e basilari e chi necessita di analisi e predizioni che penetrino in profondità. Maggiori dettagli nei capitoli 16 e 17.

Di nuovo, il nostro obiettivo

Ora che abbiamo un'idea chiara del contesto, il nostro obiettivo è aiutare i data scientist e i CAO ad avere più successo, ovvero aiutare i data scientist a prendere decisioni migliori e i CAO a rendere più forti le organizzazioni, il tutto senza essere troppo severi al riguardo. Il materiale è organizzato in 18 capitoli, legati al ciclo di vita dell'*analitycs* e all'ecosistema, ma abbiamo anche incluso materiale che non ha a che fare con nessuno dei due. Il prossimo capitolo, per esempio, tratta la differenza tra un buon data scientist e un grande data scientist. Ogni capitolo è breve e conciso. In generale, questo libro presenta il lavoro della data science a 360°. Il nostro obiettivo è ampliare la prospettiva, spingere a pensare in modo critico e aiutare a sviluppare le proprie capacità e rendere più efficiente di quanto non lo sia uno sviluppatore o un semplice utilizzatore della data science.

2

La differenza tra un buon data scientist e un grande data scientist *

La differenza tra un buon data scientist e un grande data scientist è come la differenza tra il sole e una lampadina. Potremmo essere tutti d'accordo nel dire che sono cose diverse.

Un buon data scientist lavora per scoprire informazioni nascoste in grandi quantità di dati spesso disparati e spesso di scarsa qualità. È un lavoro impegnativo. Tuttavia, i buoni data scientist hanno buone intuizioni sulle esigenze dei clienti, sulle cause della variabilità nei processi e sulle prestazioni delle aziende, cosa che non tutti sono in grado di fare. Non se ne trovano tanti di collaboratori del genere e il loro contributo è estremamente prezioso.

I grandi data scientist hanno un modo completamente diverso di pensare. Non sono interessati soltanto a trovare nuove informazioni nei dati. Sono interessati a sviluppare nuove conoscenze riguardo al mondo che li circonda. Attraverso i dati, ovviamente, ma non solo; utilizzano tutto ciò che può essere loro utile. Per capire meglio, consideriamo le previsioni per le elezioni presidenziali statunitensi del 2016. Fino al 7 novembre 2016 i sondaggisti prevedevano che Clinton avrebbe con grande probabilità trionfato su Trump:

Sondaggista	Probabilità della vittoria di Clinton
538 (Nate Silver):	72%
New York Times:	86%
Princeton Election Commission (PEC):	> 99%

* Questo capitolo si basa, in parte, su un paio di articoli digitali di Redman pubblicati nella rivista Harvard Business Review (2013a, 2017a).

The Real Work of Data Science, prima edizione. Ron S. Kenett e Thomas C. Redman.

Pubblicato nel 2019 da John Wiley & Sons Ltd. Sito internet: www.wiley.com/go/kenett-redman/datascience

è importante specificare che nessuno di questi sondaggisti ha condotto i sondaggi autonomamente. Hanno invece costruito modelli utilizzando dati grezzi forniti da altri. Siamo stupiti dai risultati ottenuti da Nate Silver e la società '538' anche se, bisogna riconoscerlo, Silver si era dichiarato consapevole dei limiti dei sondaggi appena prima dell'elezione. Nonostante non sappiamo chi abbia fatto le analisi, siamo certi che il Times e la PEC avessero assunto buoni data scientist. Per avere una rassegna completa dei sondaggi riguardo all'elezione, si veda Kenett *et al.* (2018).

Un ottimo data scientist va molto più a fondo e studia i sondaggi precedenti per avere un'idea dei loro punti di forza e dei loro limiti. Facendo questo potrebbe scoprire, per esempio, che molte persone mentono ai sondaggisti. In pubblico neanche una persona aveva confessato che avrebbe votato per Trump, ma privatamente molti hanno ammesso: "Voterò per Trump, ma non voglio che mia moglie (o mio marito) lo sappia".

Analogamente, qualche giornalista ha dichiarato che i raduni di Trump erano molto più entusiasmanti ed energici di quelli della Clinton. Hanno concluso che coloro che hanno detto che avrebbero votato erano più propensi a farlo effettivamente. Anche una piccola quantità di bugie o di mal riposto ottimismo sul voto potrebbe alterare i risultati dei sondaggi. Un ottimo data scientist condurrebbe qualche simulazione per saperne di più.

Inoltre esistono molti altri fattori su cui basarsi per predire il vincitore di un'elezione, come l'economia, il tasso di occupazione, il vincitore della scorsa edizione del Super Bowl e così via. Dunque, un ottimo data scientist allargherà i propri orizzonti. Ecco un esempio pratico: alcuni ritengono che gli americani tentino sempre di evitare che si stabilisca una dinastia politica; quindi dopo che un partito ha mantenuto la presidenza per due mandati, gli americani tenderanno a sostenere l'altro. Prima del 2016, delle precedenti otto elezioni rilevanti, sei furono vinte dall'"altro partito". Seguendo questa logica, le probabilità della vittoria di Trump sarebbero di 6/8 = 75%.

Un ottimo data scientist, tuttavia, non è soltanto alla ricerca dei migliori dati, della migliore spiegazione o modello. Il suo obiettivo è comprendere prospettive differenti, per capire quali si supportano a vicenda, quali sono in conflitto, la variazione che si preannuncia e qualsiasi altra cosa. Parla con tantissime persone diverse, testa nuove teorie, scarta senza scrupoli quelle che si rivelano insoddisfacenti ed è sempre alla ricerca di nuovi dati. È così che scopre come funziona il mondo!

L'appendice A elenca alcune caratteristiche di un data scientist di questo tipo.

Negli anni abbiamo avuto l'onore di lavorare con dozzine, forse addirittu-

ra centinaia, di data scientist, statistici e analisti. Qualcuno anche molto bravo. Il desiderio inesauribile di scoprire come funziona il mondo è ciò che li distingue dagli altri. I migliori hanno quattro altre caratteristiche:

1. *Maturano e traggono vantaggio da grandi reti di relazioni*. Ne hanno bisogno. Hanno moltissimi interessi e non possono essere esperti in tutti i campi. Gli ottimi data scientist coltivano relazioni con persone che hanno prospettive diverse dalla loro per esplorare il mondo, scoprire nuove fonti di dati e testare teorie anche provvisorie.

2. *Hanno il pallino per la matematica, ci sanno fare con i numeri*. Gli ottimi data scientist vedono cose che ad altri sfuggono. Uno stagista estivo (che adesso sfrutta le sue abilità analitiche in veste di dirigente di una media company) al suo secondo giorno di lavoro in una banca d'investimento scoprì di avere questo talento innato. Il suo capo gli aveva consegnato una pila di documenti da leggere e sfogliandoli trovò un errore nel calcolo di un guadagno bancario. Gli ci volle circa un'ora per verificare l'errore e correggerlo. Il punto è che quell'errore non era stato individuato da migliaia di altre persone. Per quel ragazzo fu subito ovvio, ma per tutti gli altri no. E si trattava di una banca d'investimento di prestigio. Si presume che qualche analista competente avesse letto lo stesso documento e che non avesse notato l'errore. La matematica fornisce una lingua utile e straordinariamente efficace (Einstein la descrisse come "irragionevolmente efficace") per descrivere il mondo. Un ottimo data scientist attinge a questa lingua facilmente e in modo intuitivo, e neanche i buoni data scientist hanno questa abilità.

3. *Possiedono tenacia*. Gli ottimi data scientist sono persistenti in molti modi diversi. Lo stagista dell'episodio raccontato sopra ha fatto la sua scoperta dopo un solo sguardo e l'ha confermata in un'ora. Sono rare le volte in cui ciò succede. Come amava dire Jeff Hooper, di Bell Labs: "I dati non rivelano i propri segreti facilmente. Devono essere torturati per confessare". Questo è molto importante. Anche nelle circostanze migliori troppi dati risultano poco definiti e sbagliati e la maggior parte saranno irrilevanti per risolvere il problema. Lavorare con questi dati è difficile e frustrante. Anche un buon data scientist potrebbe abbandonare e passare al problema successivo. Un ottimo data scientist va fino in fondo.

I grandi data scientist, inoltre, persistono anche quando vengono ignorati. Avere a che fare con la burocrazia può essere ancora più frustrante che lavorare con dati inesatti. Lo stagista di cui abbiamo parlato, per esempio, passò l'estate cercando di difendere la sua scoperta. Il gruppo che aveva commesso l'errore, infatti, offeso dall'accusa, lo attaccò pesantemente. Altri reagirono

con gioia alla notizia dell'incompetenza dei loro colleghi. Lo stagista dovette affrontare questa situazione da solo, ma i grandi data scientist, tenaci e persistenti, devono essere pronti a questo.

4. *Infine, la statistica scorre nelle loro vene*. È molto importante che siano in grado di analizzare i dati attraverso tutti i nuovi pacchetti e linguaggi (ciò include i metodi classici ma anche i più innovativi come il machine learning). Questi, tuttavia, possono essere facilmente appresi; è invece meno scontato essere in grado di far valere il rigore della statistica. Una spiegazione estremamente semplice è questa: ci sono due tipi di analisi – descrittiva e predittiva. Le analisi descrittive sono già piuttosto toste, ma quelle più vantaggiose comprendono la previsione, che è per sua natura incerta (Shmueli 2010). I grandi data scientist accolgono l'incertezza a braccia aperte. Capiscono al volo quando una previsione ha basi solide e quando invece si tratta di una speranza vana. Sono in grado di descrivere in modo formidabile che cosa debba accadere affinché la previsione diventi realtà, che cosa non deve assolutamente andare storto e quali sono le domande ancora senza risposta che fanno loro perdere il sonno. Ove possibile, quantificano l'incertezza e sono bravi a suggerire semplici esperimenti per confermare o smentire le ipotesi, ridurre l'incertezza, andare alla scoperta delle prossime domande ecc.

Per dirlo in modo diverso, ci sono alcuni che ritengono che, per i big data, sia sufficiente capire la "correlazione" senza entrare nella complessità della "causalità". Ci sono sicuramente alcuni problemi per i quali questo è vero. Ma non quelli veramente importanti! La comprensione del nesso di causalità porta a previsioni migliori. I grandi scienziati dei dati lavoreranno per stabilire i nessi causali.

Ciò richiede loro di generalizzare a un livello superiore. Concentrarsi solo sui dati a portata di mano può portare a un "overfitting", portando a modelli troppo complessi per un utilizzo futuro. La generalizzazione scientifica richiama la conoscenza specifica del dominio, i principi generali e l'intuizione, ben oltre le *cross-validation* o il confronto dei risultati dal training set e dal set di controllo (Kenett e Shmueli 2016a).

Per essere chiari, non è questa l'abilità matematica di cui parlavamo prima. È un'abilità inferenziale che si ottiene allenandola; è sofisticata, pervasa sia dal successo che dal fallimento. Alcuni di questi concetti fanno parte del programma dei corsi di data science (De Veaux *et al.* 2017; Coleman e Kenett 2017), ma la maggior parte non ne fanno parte.