# Introduzione

L'argomento di questo libro è la *scienza dei dati*, una disciplina che ha sperimentato una rapida crescita negli ultimi decenni. Proprio per questo motivo ha sollevato un grande interesse sia nei media, sia nel mercato del lavoro. Gli Stati Uniti hanno recentemente nominato il primo "esperto di scienza dei dati", *data scientist*, Dhanurjay "DJ" Patil. Questa mossa è stata sollecitata dalle società che si occupano di nuove tecnologie le quali, a dire il vero, solo recentemente hanno iniziato ad assumere veri e propri team che si occupano in modo specifico dei dati. Queste competenze sono sempre più richieste e la loro applicazione va ben oltre l'attuale mercato del lavoro.
Questo libro tenta di colmare un divario esistente in termini di competenze in ambito matematico, di programmazione e di esperienza nel dominio di applicazione della disciplina. La maggior parte delle persone, al giorno d'oggi, è dotata di un'esperienza in almeno uno (a volte due) di questi campi, ma la scienza dei dati richiede competenze in tutti e tre. Affronteremo quindi argomenti tratti da questi tre campi e risolveremo problemi complessi. Nel farlo, ripuliremo, esploreremo e analizzeremo i dati per trarre conclusioni scientifiche e accurate. Per risolvere questi complessi problemi impiegheremo tecniche di *machine learning* e di *deep learning*.

## Argomenti trattati in questo libro

Capitolo 1, *Essere uno scienziato dei dati*: un'introduzione alla terminologia di base usata dai professionisti e una panoramica sui tipi di problemi che ci troveremo ad affrontare nel corso di questo libro.
Capitolo 2, *Tipi di dati*: esamina i vari livelli e tipi di dati disponibili e i modi in cui possono essere manipolati. Questo capitolo inizia trattando le basi matematichle legate alla scienza dei dati.
Capitolo 3, *I cinque passi della scienza dei dati*: svela i cinque passi nei quali si sviluppa la scienza dei dati, compresa la manipolazione e la pulizia dei dati, utilizzando a supporto vari e dettagliati esempi.
Capitolo 4, *Basi matematiche*: aiuta a conoscere i principi di base della matematica che guidano le azioni dei data scientist, esaminando e risolvendo esempi negli ambiti del calcolo, dell'algebra lineare e molto altro ancora.

Capitolo 5, *Impossibile o improbabile: introduzione al calcolo delle probabilità*: un'introduzione semplificata al campo della teoria della probabilità e al suo uso per comprendere il nostro universo casuale.
Capitolo 6, *Approfondimenti sul calcolo delle probabilità*: applica i principi trattati nel capitolo precedente e introduce e applica alcuni teoremi, come il teorema di Bayes, con la segreta speranza di svelare il significato ultimo del mondo.
Capitolo 7, *Basi di statistica*: tratta quei tipi di problemi che l'inferenza statistica tenta di spiegare, usando gli elementi di base della sperimentazione, della normalizzazione e del campionamento casuale.
Capitolo 8, *Approfondimenti di statistica*: usa la verifica delle ipotesi e l'intervallo di confidenza per trarre conoscenze dai nostri esperimenti. È altrettanto importante per fornire la capacità di individuare il test appropriato e per interpretare i valori e i risultati.
Capitolo 9, *Comunicare i dati*: spiega in quale modo la correlazione e la causalità influenzano la nostra interpretazione dei dati. Parleremo anche dell'uso delle presentazioni per condividere con gli altri i nostri risultati.
Capitolo 10, *Quando le macchine apprendono: il machine learning*: si concentra sulla definizione di machine learning e presenta esempi realistici dei modi e delle situazioni in cui vengono impiegate tecniche di machine learning. Spiega inoltre l'importanza della valutazione dei modelli.
Capitolo 11, *Le previsioni non crescono sugli alberi… O forse sì?*: esamina modelli di machine learning più complicati, come gli alberi decisionali e le previsioni bayesiane, i quali consentono di risolvere problemi più complessi legati ai dati.
Capitolo 12, *Oltre le basi della scienza dei dati*: introduce alcune delle forze, un po' misteriose, che governano la scienza dei dati, fra le quali il bias e la varianza. Qui vengono introdotte le reti neurali, una moderna tecnica di apprendimento profondo (o *deep learning*).
Capitolo 13, *Casi di studio*: impiega vari casi di studio per aiutare il lettore a concretizzare e a far sedimentare i concetti e le idee legati alla scienza dei dati. Seguiremo il flusso di lavoro legato all'intera scienza dei dati, dall'inizio alla fine, per più volte e per vari esempi, comprendendo la previsione della quotazione delle azioni di borsa e l'interpretazione della scrittura a mano libera.

## Dotazione software necessaria

Questo libro utilizza Python per tutti gli esempi di codice. Dovrete necessariamente avere in dotazione un computer (Linux, Mac o Windows) con l'accesso a un terminale in stile Unix e dotato di Python 2.7. È consigliata anche l'installazione della distribuzione Anaconda, in quanto è richiesta dalla maggior parte dei pacchetti usati negli esempi.

## A chi è rivolto questo libro

Questo libro è rivolto a tutti coloro che stanno cercando di comprendere e utilizzare le tecniche di base della scienza dei dati, a prescindere dal dominio di applicazione.
Il lettore dovrebbe avere un'infarinatura dei concetti base della matematica (algebra e un po' di calcolo delle probabilità) e non dovrebbe avere problemi a interpretare brevi frammenti di codice R/Python e di pseudocodice. Tuttavia non è affatto necessario che

il lettore abbia una precedente esperienza professionale nel campo dell'elaborazione dei dati; quello che conta davvero è che sia animato di un sincero desiderio di apprendere e di applicare le tecniche presentate in questo libro, provandole sui propri dataset o su quelli che si trova a utilizzare.

## Convenzioni utilizzate

Per uniformità con lo stile Python e in concordanza con l'edizione originale, nel presente volume sono state adottate le convenzioni inglesi in termini di virgola decimale, la quale è rappresentata da un punto (.).
In questo libro, troverete vari stili di testo per distinguere diversi tipi di informazioni. Ecco alcuni esempi degli stili utilizzati e del loro significato.
Gli elementi di codice nel testo, i nomi di tabelle di un database, i nomi di cartelle, i nomi di file, le estensioni di file, i percorsi, gli indirizzi URL e gli account Twitter sono tutti scritti utilizzando il carattere monospaziato, per esempio: "Per questi operatori, considerate l'impiego del tipo di dati `boolean`".
Un blocco di codice ha il seguente aspetto:

```
tweet = "RT @j_o_n_dnger: $TWTR now top holding for Andor, unseating $AAPL"
words_in_tweet = first_tweet.split(' ')       # lista delle parole del tweet
```

Quando è necessario attrarre l'attenzione su una specifica parte del codice, le righe o gli elementi in questione saranno indicati in grassetto:

```
for word in words_in_tweet:                  # Per ogni nuova parola della lista
    if "$" in word:                          # se la parola ha un "cashtag"
    print "THIS TWEET IS ABOUT", word        # avverti l'utente
```

**NOTA**

Avvertimenti e note importanti sono evidenziate con questo tipo di nota.

## Scarica i file degli esempi

Il codice degli esempi è disponibile su GitHub all'indirizzo: `https://github.com/PacktPublishing/Principles-of-Data-Science`.
In caso di problemi è anche possibile scaricare un archivio ZIP del codice dal sito dell'editore originale inglese, Packt Publishing. Per farlo è necessario registrarsi gratuitamente all'indirizzo `https://www.packtpub.com/register`. Quindi andate sulla scheda del libro (`http://bit.ly/packt-pods-book`) e fate clic su *Code Files*.
I lettori interessati possono anche scaricare un file PDF contenente una versione a colori delle immagini presenti nel libro. Le immagini a colori potranno esservi utili per comprendere meglio alcuni dettagli sulle variazioni nell'output. Potete scaricare questo file all'indirizzo `http://bit.ly/packt-pods-img`.