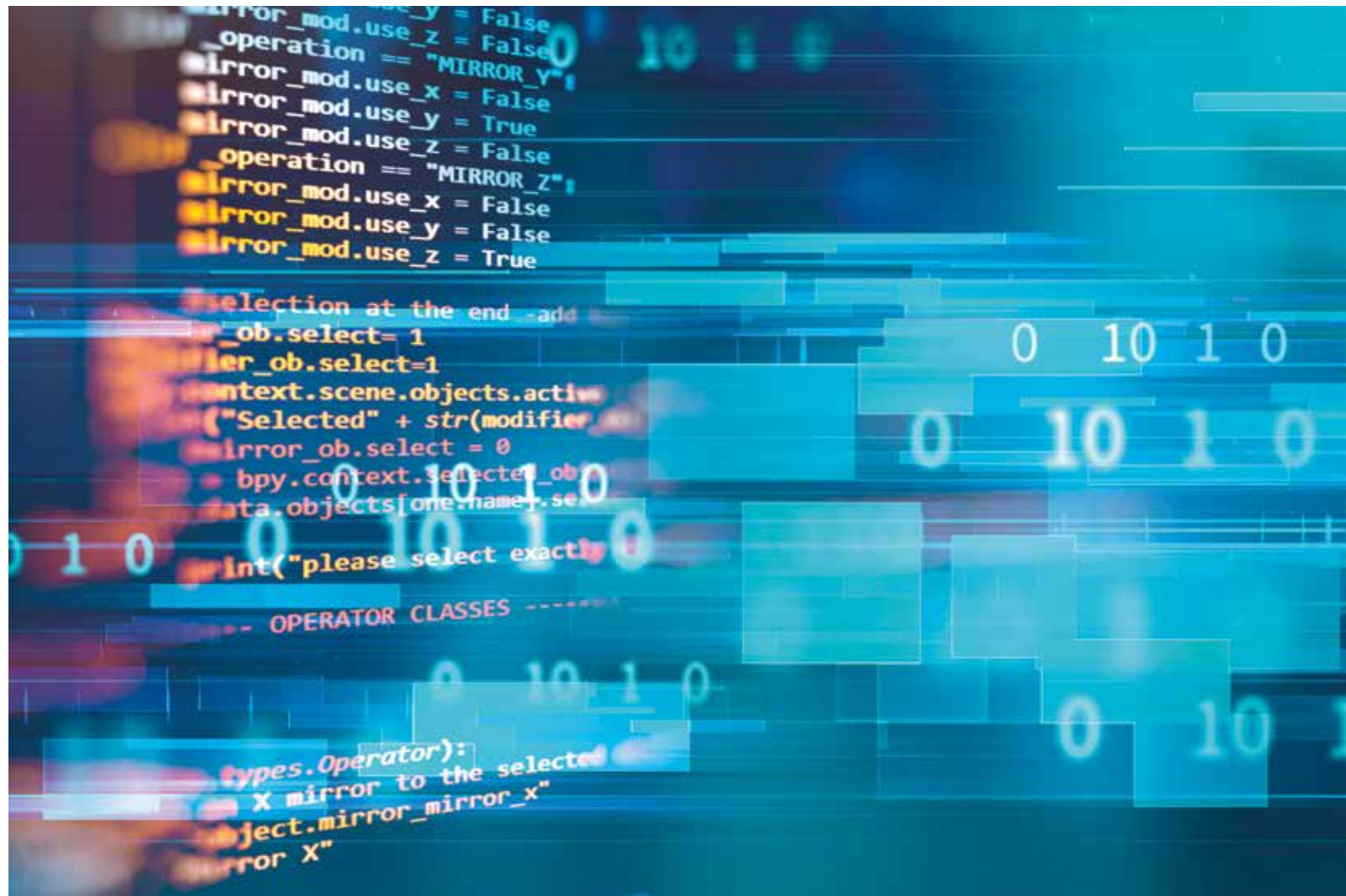


OPEN INNOVATION



I Big Data attraverso il Data Mining

Il processo di scoperta di modelli, relazioni e informazioni significative all'interno di grandi insiemi di dati



DI VINCENZO GERMANO*

L'era digitale ha visto una crescita esplosiva nella generazione di dati, un fenomeno che ha portato alla creazione del concetto di "Big Data". Come analizzato nell'articolo precedente (si veda *Il Giornale dell'Ingegnere* n. 8/2023 ndr.), i Big Data rappresentano un insieme di dati generati da sorgenti eterogenee come possono essere sensori, social media, transazioni online e molto altro ancora. Essi abbracciano le "3V", ossia volume, velocità e varietà dei dati, che oggi permeano ogni aspetto delle nostre vite digitali. La velocità riguarda il flusso continuo e rapido con cui vengono generati i dati, il volume riguarda la loro mole massiccia e la varietà, poiché possono essere strutturati, semi-strutturati o non strutturati. Questa classificazione si riferisce alla modalità con cui le informazioni sono organizzate e possono influenzare notevolmente la gestione e l'analisi dei dati. Generalmente i dati strutturati sono organizzati in un formato tabellare tipo "righe e colonne", ogni colonna rappresenta un attributo o una caratteristica e ogni riga un record con dati specifici. Basti pensare a database relazionali, tipo tabelle in un database che rappresentano un esempio comune di dati strutturati, con colonne chiave, dati numerici e dati di testo organizzati in righe. Un altro esempio possono essere i fogli di calcolo, con colonne e righe che

definiscono chiaramente i dati. Passando ai dati semi-strutturati, essi non seguono uno schema tabellare rigido come i dati strutturati, tuttavia, contengono elementi di struttura che consentono di categorizzare o organizzare i dati in modo gerarchico. Un esempio possono essere i documenti XML (*eXtensible Markup Language*), ossia un "linguaggio di marcatura" che consente di definire tag personalizzati per strutturare i dati in modo gerarchico. Un altro esempio può essere il JSON (*JavaScript Object Notation*), che viene utilizzato comunemente nei servizi web, grazie al quale si riescono a organizzare i dati in coppie "chiave-valore", consentendo una flessibilità di struttura. Infine, i dati non strutturati non seguono uno schema predefinito e non sono organizzati in colonne o righe, infatti possono essere di natura testuale, multimediale o di altro tipo. Gli esempi sono molteplici, come testi liberi, quali documenti, articoli, e-mail e altri contenuti di testo senza una struttura specifica. File multimediali come immagini e video, che contengono informazioni senza una disposizione tabellare. File audio, come registrazioni che non seguono una struttura specifica come i dati testuali. Risorse multimediali che a loro volta incorporano elementi di dati strutturati, semi-strutturati e non strutturati, dato che possono contenere metadati (strutturati), tag (semi-strutturati) e dati di contenuto (non strutturati). Basti

pensare a foto digitali con metadati, che includono informazioni strutturate come data e ora, ma anche tag semi-strutturati e dati non strutturati nell'immagine stessa. Oppure file video che possono avere descrizioni strutturate ma il contenuto effettivo può essere non strutturato. Da ciò si può capire come la classificazione dei dati in queste categorie rifletta la diversità delle informazioni digitali che incontriamo quotidianamente. La capacità di gestire e analizzare dati con diversi livelli di strutturazione è essenziale nell'era dei Big Data, poiché molte fonti di informazioni forniscono una miscela di dati eterogenei. Le tecnologie e gli strumenti di gestione dati devono essere in grado di adattarsi a questa varietà per consentire una comprensione completa e approfondita delle informazioni a nostra disposizione. Affrontare questa enormità di dati richiede approcci innovativi e soluzioni tecnologiche avanzate. Nel cuore di questa enorme quantità di informazioni sorge una disciplina cruciale, il *Data Mining* (da ora in poi riportato con la sigla *DM*), una pratica essenziale per estrarre valore significativo dai Big Data.

IL DATA MINING

Si può sintetizzare il DM in un processo di scoperta di modelli, relazioni e informazioni significative all'interno di grandi insiemi di dati, una disciplina "multidisciplinare" che, tra le altre cose, fonde principi di statistica, apprendimento au-

tomatico, intelligenza artificiale e *database management*. L'obiettivo fondamentale è trasformare i "dati grezzi" in conoscenza utile e informazioni predittive da poter poi utilizzare per lo scopo d'interesse. Il DM ha applicazioni estese in una varietà di settori: solo per citarne alcuni, lo ritroviamo nella finanza, ad esempio nell'identificazione di frodi, nella previsione di mercati finanziari e nella gestione del rischio; nella sanità, per la diagnosi predittiva, nella ricerca di nuovi farmaci e nell'ottimizzazione dei protocolli di trattamento; nel commercio, per l'analisi del comportamento dei consumatori, nella previsione della domanda e nella personalizzazione delle offerte; nel settore manifatturiero per la manutenzione predittiva, nell'ottimizzazione della catena di approvvigionamento e nel controllo della qualità. Uno degli aspetti chiave del DM è l'estrazione di modelli, infatti tali algoritmi analizzano i dati alla ricerca di correlazioni, tendenze e strutture nascoste che possono sfuggire all'occhio umano. Questo consente di identificare comportamenti ricorrenti, relazioni causali, anomale e significative, aprendo la strada a previsioni più accurate. Per permettere l'estrazione di modelli, si ha la necessità di segmentare e classificare i dati e attraverso algoritmi specifici (*clustering*), è possibile raggruppare dati simili, rivelando così sottogruppi o categorie nascoste. La classificazione, d'altra

parte, consente di assegnare nuovi dati a categorie già identificate, rendendo il processo di presa decisionale più efficiente ed efficace. Uno degli utilizzi più potenti del DM è la previsione di eventi futuri basata su modelli costruiti dai dati storici; questo approccio, noto come "modellizzazione predittiva", è ampiamente utilizzato in settori come la finanza, la sanità e il marketing per anticipare trend, comportamenti dei clienti e risultati futuri. Un altro aspetto rilevante è l'analisi dell'associazione del DM, che identifica relazioni tra variabili, ad esempio, nel settore del commercio al dettaglio, può rivelare quali prodotti sono spesso acquistati insieme, consentendo alle aziende di ottimizzare le strategie di vendita e marketing.

L'utilizzo del DM e dei Big Data offre numerosi vantaggi alle aziende, basti solo pensare che la scoperta di pattern e tendenze nascoste all'interno di grandi insiemi di dati consente di scoprire relazioni tra variabili, comportamenti dei clienti e trend di mercato che potrebbero non essere immediatamente evidenti. Ciò permette di adattare le strategie di marketing, vendite e produzione in risposta a pattern emergenti, migliorando l'efficacia delle loro operazioni. Si riescono anche a migliorare le decisioni aziendali, infatti il DM contribuisce a prendere decisioni più informate e basate su evidenze. Le informazioni estratte consentono di comprendere meglio il contesto aziendale, ottimizzando processi decisionali in settori come la pianificazione, la gestione delle risorse e la strategia aziendale, portando a risultati aziendali più efficaci, riducendo il rischio di errori strategici e contribuendo a una gestione più efficiente delle risorse. Questi sono solo alcuni dei molteplici vantaggi che il DM può portare.

CONCLUSIONI

Il Data Mining è uno strumento fondamentale che guida nell'esplorazione dei Big Data, rivelando connessioni profonde e informazioni nascoste. Attraverso l'uso di algoritmi avanzati e tecnologie innovative, il DM offre un'opportunità senza precedenti di comprendere meglio il mondo vasto e complesso dei dati digitali. Mentre continuiamo a navigare attraverso questa era di Big Data, il DM rimane una bussola affidabile per coloro che cercano di scoprire il significato intrinseco dietro le immense quantità di informazioni a nostra disposizione.

***INGEGNERE ELETTRONICO E VICECOORDINATORE COMMISSIONE INNOVATION & PROJECT MANAGEMENT ORDINE DEGLI INGEGNERI DELLA PROVINCIA DI TORINO**