I dati sono un bene scarso?
Nessun problema, li creiamo con l'intelligenza artificiale

Creati a immagine di quelli "reali", i dati sintetici favoriranno performance più accurate dei modelli di intelligenza artificiale. Secondo Gartner, nel 2024 il 60% dei dati utilizzati in progetti AI sarà generato sinteticamente (era solo l'1% nel 2021).

Questa tecnologia può essere applicata a tutti i campi che necessitano di molti dati per migliorare i processi: assicurazioni, finanza, energia, telecomunicazioni, mobilità urbana, retail – per citarne alcuni.

Società come Amazon, American Express, John Deere li stanno già utilizzando. Per cosa? Dalla gestione delle frodi al training di sistema di riconoscimento del linguaggio di Alexa, fino alle simulazioni nel Metaverso.

A cura di Shalini Kurapati, Co-Founder e CEO di Clearbox AI
e Piergiorgio Stano Director, Head of Data & Analytics Italy presso BearingPoint

L'adozione dell'AI continua a essere in costante aumento: il 56% delle persone intervistate nella Global Survey 2021 di McKinsey sull'intelligenza artificiale riferisce che la **propria azienda usa l'intelligenza artificiale in almeno una funzione**, rispetto al 50% del 2020. Tuttavia, numerose analisi di mercato concordano che ancora nel 2022 la maggior parte delle iniziative di intelligenza artificiale (nell'ordine del 60%-80%) **non entra in produzione**. Perché? Principalmente per problemi legati ai **dati**: in particolare, perché non **si ha accesso alle informazioni decisive**. Infatti, per alimentare ed addestrare un sistema di AI servono tantissimi dati, di buona qualità e non viziati da pregiudizi.

È chiaro anche solo da queste poche informazioni che, in un mondo in cui l'intelligenza artificiale diventa sempre più pervasiva, **trovare dati di qualità diventa una priorità**. È un problema, quello dei dati, che porta con sé già la sua soluzione: se i dati veri sono difficili da reperire, scarsi, viziati da errori, non utilizzabili per motivi di privacy, parziali o alterati dai *bias* di chi li ha sviluppati… allora **possiamo usare la stessa intelligenza artificiale per simularli**.

È qui che entrano in gioco i **dati sintetici**: trattasi di **informazioni artificiali** che riproducono in maniera fedele sotto il profilo matematico e statistico i dataset del mondo reale. Utili in caso di carenza, costi eccessivi, tempi stretti, limiti normativi o cattiva qualità di quelli disponibili. È solo da alcuni anni che se ne parla, ma ancora in pochi sanno come e in che occasioni vengano utilizzati.

Perché scegliere i dati sintetici? E quali vantaggi offrono?

Oggi i data scientist perdono l'80% del loro tempo a selezionare, ordinare e pulire i dati (Osservatorio Big Data del Politecnico di Milano). Con quelli sintetici potrebbero invertire la rotta e dedicare la maggior parte del loro tempo all'analisi vera e propria, che sta alla base della creazione degli algoritmi.

Oltre a semplificare il lavoro dei data scientist, i dati sintetici portano una lunga serie di vantaggi.

Un primo beneficio è quello di evitare di incorrere in problematiche legate alla **lesione della privacy** delle persone. Come? Pensiamo a un ospedale o a una clinica privata che deve **fornire a una società informatica dati medici** per addestrare un sistema di diagnosi del cancro basato sull'intelligenza artificiale. Con i dati sintetici, gli sviluppatori dispongono di set di informazioni di qualità da utilizzare durante la progettazione e la compilazione del sistema, senza che vengano scambiate le informazioni sensibili delle persone reali: così la rete ospedaliera non corre il rischio di mettere in pericolo la privacy dei pazienti.

Un secondo vantaggio è quello di poter **accelerare e rafforzare lo sviluppo dei modelli di intelligenza artificiale**: la raccolta dei dati dal mondo reale può richiedere molto tempo perché le informazioni devono essere abbondanti, devono anche essere selezionate, classificate, elaborate e sottoposte a controlli di conformità. Con i dati sintetici, l'intero processo si accorcia perché si possono creare sin da subito dei dati puliti, ordinati e conformi. È quello che stiamo facendo con la collaborazione tra **BearingPoint e Clearbox AI** al fine di creare un sistema di intelligenza artificiale più efficiente ed affidabile di quelli esistenti per identificare le frodi finanziarie.

I dati sintetici permettono inoltre di simulare **scenari futuri**: uno dei problemi dei dati reali è che sono *storici*, permettono di valutare solo eventi già accaduti e possono quindi diventare obsoleti. Per esempio, il COVID ha impattato in maniera rilevante le abitudini delle persone: pensiamo agli spostamenti in auto per andare a lavoro, con il relativo formarsi di code in città o ai caselli. L'utilizzo estensivo del *remote working* ha cambiato pesantemente i flussi di veicoli sulle strade, così **tutti i dati storici relativi agli spostamenti delle persone hanno perso una parte significativa del loro valore predittivo**.

I dati sintetici possono essere utili anche per **testare se le intelligenze artificiali hanno dei pregiudizi** (o *bias*): se può sembrare strano che una "macchina" possa avere un pregiudizio, bisogna ricordare che i sistemi di AI imparano immagazzinando grandi quantità di informazioni. Ma i dati storici possono essere viziati da pregiudizi sociali del tempo a cui si riferiscono. Testare le intelligenze artificiali con i dati sintetici può invece aiutare ad identificare e neutralizzare tali pregiudizi nascosti e potenzialmente fuorvianti.

I casi d'uso dei dati sintetici

Questa tecnologia può essere applicata a tutti quei campi che necessitano di molti dati per migliorare i propri processi, dal **mondo finanziario** a quello delle **assicurazioni**, dall'**energia** alle **telecomunicazioni**, dalla **mobilità urbana** al **retail**.

Sono già parecchie le grandi aziende che utilizzano i dati sintetici. **John Deere**, per esempio, impiega foto sintetiche per addestrare le proprie AI a riconoscere le piante infestanti in condizioni atmosferiche non ottimali. **Amazon** ricorre ai dati sintetici per il training di sistema di riconoscimento del linguaggio di Alexa. **American Express**, invece, impiega tali dati ai fini del riconoscimento delle transazioni fraudolente. Molte società del settore *automotive* stanno iniziando a utilizzarli per **addestrare i sistemi di guida autonoma**. È infine notizia recente un progetto che ha vinto un finanziamento della Commissione Europea nell'ambito del programma Horizon Europe e punta a sviluppare nuovi sistemi di analisi dati nell'ambito delle **malattie ematologiche**. Per farlo **Synthema**, questo il nome del progetto, usa tecniche innovative basate sull'intelligenza artificiale per rendere anonime le informazioni cliniche e biologiche dei pazienti e generare dati sintetici, nel rispetto delle norme sulla privacy, per superare la scarsità e la frammentazione delle informazioni disponibili oggi per la ricerca, in modo conforme al GDPR (General Data Protection Regulation). I dati sintetici arrivano fino nel **Metaverso** che richiede simulazioni virtuali in 3D di ambienti di gioco, sociali e aziendali. I dati sintetici possono colmare alcune lacune per creare impostazioni e oggetti realistici.

Il mercato dei dati sintetici

Un'analisi di Gartner prevede che il mercato dei dati sintetici crescerà fino a che, nel 2024, **il 60% dei dati utilizzati in progetti AI sarà generato sinteticamente**. Attualmente, i dati sintetici rappresentano solo l'1% di tutti i dati digitali. Questo aumento amplierà i casi d'uso per le applicazioni di intelligenza artificiale e, a sua volta, aumenterà i posti di lavoro nel settore dell'intelligenza artificiale. Entro il 2027 si prevede che il segmento di mercato dei dati sintetici crescerà fino a un valore complessivo di 1,15 miliardi di dollari.

Non è un caso che sempre Gartner abbia incluso i dati sintetici **tra le tecnologie più promettenti per il futuro**. Ed è singolare che oggi se ne senta parlare ancora così poco. Ma chi li conosce lo sa: i dati sintetici saranno una delle monete del domani.

Informazioni su Clearbox AI

Clearbox AI è la startup tech italiana che aiuta le aziende a lanciare progetti di AI e di Analytics attraverso la generazione di dati sintetici di alta qualità. La missione aziendale consiste nel comprendere e risolvere le sfide che le imprese incontrano nello sviluppo dei processi di Intelligenza Artificiale. Questi ostacoli sono spesso legati ai dati sensibili che sono difficili da gestire a livello di privacy e che possono non essere abbastanza rappresentativi per tutte le fasce di popolazione, o la loro quantità non è sufficiente per garantire risultati di successo. Il Data Engine di Clearbox AI è fondato su tecnologia proprietaria e agnostica basata su modelli generativi avanzati, creata anche grazie alle solide radici nel mondo della ricerca del team. La soluzione supportaqualsiasi tipo di azienda sia per incrementare la disponibilità di dati, la loro qualità e quindi

mitigando eventuali bias interni ai dati, ma anche per aiutarle in termini di policy, compliance e privacy degli stessi.
Clearbox AI è stata recentemente selezionata dalla Commissione Europeaper il progetto **Women TechEU**, che supporta le startup deep-tech guidate da donne per valorizzare il talento e favorire l'innovazione nell'ecosistema tech europeo. È inoltre vincitrice del Premio Nazionale dell'Innovazione (categoria ICT, 2019), il Seal of Excellence dell'Unione Europea, ed è stata selezionata da **Fortune Italia** come una delle migliori startup AI del paese.

Informazioni su BearingPoint
BearingPoint è una società di consulenza gestionale e tecnologica indipendente con radici europee e portata globale. L'azienda opera in tre business unit: Consulting, Products e Capital. L'area Consulting offre servizi di consulenza con un chiaro focus su aree di business selezionate. L'area Products fornisce soluzioni digitali guidate dalla proprietà intellettuale e servizi per gestire processi critici all'interno del business. L'area Capital fornisce servizi a supporto di fusioni, acquisizioni e transazioni. I clienti di BearingPoint includono molte aziende e organizzazioni leader a livello mondiale. L'azienda ha una rete di consulenza globale di oltre 13.000 persone e supporta clienti in oltre 70 paesi, impegnandosi con loro per ottenere un successo misurabile e sostenibile.

Contatti
Clearbox AI: shalini@clearbox.ai
BearingPoint: piergiorgio.stano@bearingpoint.com

Ufficio stampa ddl studio: innovationteam@ddlstudio.net
Mara Linda Degiovanni: +39 3496224812
Elisa Giuliana: +39 3386027361