

Dati storici, ibridi e sintetici e il bisogno di ludicizzazione

Senza dati, gli algoritmi – inclusa l'AI – non vanno da nessuna parte, come un motore con un serbatoio vuoto.

L'AI ha bisogno di dati per essere addestrata e pertanto di dati per applicare il suo addestramento.

Naturalmente, l' può essere estremamente esibibile: sono i dati che determinano il suo ambito di applicazione e grado di successo.

È noto che l'AI, intesa come Machine Learning (apprendimento automatico), apprende dai dati che riceve e migliora progressivamente i suoi risultati. Di solito sono necessarie enormi quantità di dati, e di regola quanti più sono i dati, tanto migliore è il risultato.

Tuttavia, recentemente l'AI è talmente migliorata che, in taluni casi, si sta passando da un'enfasi sulla **quantità** di grandi masse di dati, a volte impropriamente chiamati Big Data, a un'enfasi sulla **qualità** di insiemi di dati ben curati.

I “piccoli dati” di alta qualità costituiscono uno degli scenari futuri dell' AI

Ma sappiamo anche che l'AI può generare i propri dati. Chiamerò **sintetici** i dati interamente *generati* dall'AI.

```
> In passato, giocare a scacchi contro un computer significava giocare contro i  
migliori giocatori umani che avessero mai preso parte al gioco.  
> Ma AlphaZero, l'ultima versione del sistema di sviluppato da DeepMind, ha  
imparato a giocare meglio di chiunque altro, e in effetti di qualsiasi altro  
software, facendo adamento soltanto sulle regole del gioco, senza alcun input di  
dati da alcuna fonte esterna. Non aveva alcuna memoria storica.  
> AlphaZero ha imparato giocando contro se stesso, generando così i propri  
dati sintetici relativi agli scacchi.
```

AlphaZero ha generato i propri dati sintetici, e questo è stato sufficiente per il suo addestramento. Questo è ciò che si intende per dati sintetici.

I dati realmente sintetici hanno alcune straordinarie proprietà, sono:

- durevoli
- riutilizzabili
- rapidamente trasportabili
- facilmente duplicabili
- simultaneamente condivisibili

Sono anche:

- puliti
- affidabili
- non violano privacy o riservatezza
- se vengono persi non è un disastro perché possono essere ricreati
- sono perfettamente formattati per essere utilizzati dal sistema che li genera

Con i dati sintetici l' non è mai costretta ad abbandonare il suo spazio digitale, dove può esercitare il controllo completo su qualsiasi input e output dei suoi processi.

Tra dati storici più o meno mascherati (impoveriti attraverso una risoluzione inferiore, per esempio tramite l'anonimizzazione) e dati puramente sintetici, esiste una varietà di dati più o meno **ibridi**, che possiamo ragurare come un prodotto di dati storici e sintetici.

Un buon esempio, introdotto da Goodfellow e coautori, è fornito dalle reti generative avverse ([GANs](#)).

Il futuro dell' non risiede soltanto nei "[piccoli dati](#)" ma anche, o forse principalmente, nella sua crescente capacità di generare i propri dati.

La differenza è costituita dal processo genetico, cioè dalle regole usate per creare i dati. I dati storici sono ottenuti tramite *regole di registrazione*, in quanto sono il risultato di osservazioni del comportamento di un sistema.

I dati sintetizzati sono ottenuti tramite *regole di astrazione*, che eliminano, mascherano o oscurano alcuni gradi di risoluzione a partire dai dati storici, per esempio mediante l'anonimizzazione. Dati ibridi e realmente sintetici possono essere generati tramite regole vincolanti o costitutive.

Quando

1. un processo o un'interazione può essere trasformata in un gioco e
2. il gioco può essere trasformato in un gioco formato da regole costitutive, allora
3. l'AI sarà in grado di generare i propri dati, completamente sintetici, ed essere il miglior "giocatore" su questo pianeta