

DOMANDA DI INVENZIONE NUMERO	102021000008552
Data Deposito	06/04/2021
Data Pubblicazione	06/10/2022

Classifiche IPC

Sezione	Classe	Sottoclasse	Gruppo	Sottogruppo
G	06	Q	30	02

Sezione	Classe	Sottoclasse	Gruppo	Sottogruppo
G	06	F	21	62

Titolo

Metodo implementato da computer per la generazione di dati sintetici
--

DESCRIZIONE dell'Invenzione Industriale dal titolo:
"Metodo implementato da computer per la generazione
di dati sintetici"

appartenente a AINDO srl, di nazionalità Italiana,
5 con sede in Località Padriciano 99, 34149 Trieste.

DESCRIZIONE

La presente invenzione ha per oggetto un metodo
implementato da computer per la generazione di dati
10 sintetici comprendente i passi di:

a) raccogliere dati reali;

b) definire la tipologia di detti dati reali
raccolti;

c) addestrare almeno un modello generativo con
15 detti dati reali;

d) provvedere la generazione di dati sintetici
mediante il detto modello generativo addestrato.

Uno dei maggiori problemi che la società attuale
deve affrontare oggi giorno è legato allo sfruttamento
20 delle informazioni contenute nei dati personali.
Questi vengono infatti raccolti, a partire dal
singolo utente, e si rivelano necessari per
promuovere l'innovazione. I dati sono infatti la
risorsa in più rapida crescita posseduta da
25 istituzioni governative e aziende.

Tuttavia, a discapito del grande numero di dati
che possono essere raccolti a partire dal singolo
utente, l'innovazione che da questi potrebbe derivare
non sta scalando allo stesso ritmo. Ciò è dovuto
30 principalmente alla limitata condivisibilità di detti
dati raccolti.

Lavorare con i dati di un individuo comporta
infatti grandi rischi legali, reputazionali e

operativi. Le informazioni statistiche contenute in detti dati sono ciò che veramente importa alle aziende, in vista di preferenze e necessità dei clienti. Le informazioni sensibili personali invece,
5 che fanno inevitabilmente parte del dato raccolto, rappresentano unicamente una responsabilità e una limitazione.

Per garantire il rispetto delle politiche relative alla privacy, non è infatti possibile
10 processare e/o condividere il dato tal quale. Un comportamento di questo tipo renderebbe disponibili a terze parti almeno pezzi di informazioni sensibili degli utenti, andando a violare le norme del GDPR.

Generalmente, per essere in grado di processare
15 il dato in vista di una sua analisi, o in vista di una sua condivisione con terze parti, si procede ad una azione denominata "anonimizzazione" del dato.

Questo processo dovrebbe, da un lato, tutelare le informazioni sensibili strettamente personali,
20 dall'altro dovrebbe far sì che il dato mantenga una sua validità ai fini di una sua analisi e/o processamento per scopi di miglioramento e sviluppo aziendale.

Questo, tuttavia, raramente si verifica. Le più
25 comuni tecniche di anonimizzazione di dati tutt'oggi disponibili, infatti, non rendono il dato anonimizzato completamente sicuro da eventuali attacchi di re-identificazione.

Per aumentare sempre più la tutela
30 dell'individuo, occorre conseguentemente anonimizzare il dato in maniera sempre più spinta. Ciò si traduce, inevitabilmente, in una perdita delle informazioni

utilizzabili che il dato porta con sé, e che sono quelle di reale e unico interesse.

Si può dire quindi che, all'aumentare della sicurezza che si può conferire ad un dato, si riduce
5 di pari passo la sua reale utilità.

In tempi molto recenti, per cercare di risolvere il problema, è stato introdotto il concetto di "dati sintetici". Questi, ad una prima analisi, sembrano essere in grado di superare le limitazioni sopra
10 citate.

Con il termine "dati sintetici", come descritto nel documento "Privacy and synthetic dataset", Steven M. Bellovin, Preetam K. Dutta, and Nathan Reiter, Columbia University, Department of Computer Science,
15 si fa riferimento al risultato di una procedura che, a partire da un insieme di dati reali, produce un insieme di dati realistici ma artificiali, che hanno le stesse proprietà statistiche dei dati reali di partenza. Questi dati, dal momento che sono
20 artificiali, possono essere analizzati e/o condivisi con terze parti, garantendo quindi da un lato una loro usabilità concreta, ed evitando dall'altro lato di mettere in pericolo possibili informazioni sensibili del dato reale di partenza.

25 Nel dettaglio il documento specifica che, per provvedere la creazione - o generazione - di dati sintetici, si prende inizialmente un dataset originale contenente informazioni sensibili. Questo dataset in ingresso viene impiegato per addestrare un
30 modello generativo abilitato all'apprendimento automatico. Il modello generativo provvede, a seguito della fase di addestramento, un set di dati realistici ma artificiali, dove detti dati

presentano, come detto, le stesse proprietà statistiche del dataset di partenza.

Nello stato dell'arte, diverse modalità di generazione di dati sintetici sono già state descritte.

L'articolo scientifico Noseong Park et al. "Data Synthesis based on Generative Adversarial Networks", PVLDB, 11 (10): 1071-1083, 2018 propone un metodo di sintesi di dati basato su reti generative avversarie (GANs, generative adversarial networks). Nello specifico, la rete generativa consiste di tre reti neurali ossia un generatore, un discriminatore ed un classificatore. Il discriminatore tenta di distinguere tra records reali e sintetici, e il generatore offusca il compito del discriminatore generando records realistici. Entrambi continuano a ripetere il gioco dell'avversario, e il generatore può così raggiungere prestazioni di generazione senza precedenti alla fine della partita a due giocatori. Il classificatore serve ad aumentare l'integrità semantica dei records sintetici. Le reti neurali vengono addestrate mediante tre tipologie di funzione obiettivo, che devono provvedere che i records sintetici abbiano le stesse caratteristiche statistiche del file records originale, e che vi sia integrità semantica. Il tempo di addestramento è inferiore a 20 minuti.

Il documento WO2020148573A1 "Using generative adversarial networks (GANs) to enable sharing of sensitive data" provvede un processo per generare dati sintetici a partire da dati raccolti da un operatore di rete. Il dato di sintesi è anonimizzato in modo da provvedere la privacy, ma allo stesso tempo garantire

le altre proprietà essenziali che lo rendono utile ai fini di analisi e ricerche. In particolare, il dato di sintesi viene generato a partire da un primo modello generativo GAN, e il dato generato deve
5 soddisfare un certo primo valore di soglia. Il dato così generato, se il valore di soglia viene soddisfatto, ha proprietà statistiche simili al dato originale. Un secondo modello generativo GAN viene poi utilizzato per garantire che nessun dato
10 sensibile venga trapelato nei dati sintetici o possa da questi essere recuperato. I dati aggiornati secondo il secondo modello generativo devono soddisfare un secondo valore di soglia. Infine, si monitora se i dati sintetici aggiornati soddisfino il
15 primo valore di soglia. Se è così, i dati sintetici possono essere condivisi con altre organizzazioni. Se così non è, si ri-eseguono il primo e secondo modello di machine learning per aggiornare i dati di sintesi fintanto che non si raggiunge il primo valore di
20 soglia.

Il documento US20180165475 "Methods and apparatus for transforming and statistically modeling relational databases to synthesize privacy-protected anonymized data" fa riferimento a un metodo,
25 apparato, sistema per la trasformazione e modellazione statistica di database relazionali, per sintetizzare dati anonimi che siano protetti dal punto di vista della privacy. I dati di sintesi seguono le stesse proprietà matematiche dei dati
30 originali, quindi una loro analisi è valida quanto quella sui dati reali. Il metodo/sistema/apparato fa riferimento a "Synthetic Data Vault" (SDV), che provvede tecniche trasformative per modellare

database relazionali e sintetizzare dati in maniera automatica, in maniera tale che detti dati di sintesi siano virtualmente indistinguibili dai dati originali. I dati di sintesi hanno le stesse
5 proprietà matematiche dei dati originali, gli stessi intervalli, distribuzione di valori e covarianze. SDV coinvolge quattro passi generali; inizialmente, un database relazionale originale viene trasformato, ad esempio ristrutturato, secondo un
10 formato/organizzazione che facilita una modellazione generativa. Successivamente, gli aspetti strutturali del database relazionale ristrutturato vengono specificati e inclusi nel database ristrutturato stesso (ad esempio vengono fornite alcune
15 informazioni di base sulla struttura del database trasformato per facilitare la modellazione dei dati e le interrelazioni delle rispettive tabelle nel database). Ancora successivamente, il database trasformato/ristrutturato viene modellato
20 statisticamente (generativamente) e, infine, il modello può essere utilizzato per sintetizzare dati che statisticamente mimino i dati originali nel database.

Le soluzioni sopra citate, tuttavia, falliscono
25 nel trattare dati tabulari che rappresentano entità ed eventi, dove per di più detti dati possono risultare raggruppati per entità, e dove gli eventi possono inoltre presentare una relazione temporale tra di loro.

30 Ancora oggi, di conseguenza, si sente la necessità di un metodo implementato da computer, e di un relativo sistema, per la generazione di dati sintetici a partire da dati tabulari che

rappresentano entità ed eventi, detti dati essendo raggruppati per entità, e dove gli eventi possono presentare una relazione temporale tra di loro.

La presente invenzione mira a superare gli
5 inconvenienti dell'arte nota tipica del settore per mezzo di un nuovo metodo implementato da computer per la generazione di dati sintetici come definito all'inizio, detto metodo essendo caratterizzato dal fatto che i dati reali sono dati tabulari e il
10 modello generativo comprende almeno un autoencoder variazionale (VAE).

Con il termine dati si fa qui riferimento a dati di qualsivoglia natura, che possano essere suscettibili di un processamento al fine di ricavarne
15 qualsivoglia informazione statistica e/o rilevante, essendo detti dati in forma tabulare, ossia comprendendo almeno una riga e colonna.

Preferibilmente, i detti dati in forma tabulare comprendono una pluralità di righe e colonne.

20 Preferibilmente, con il termine dati si fa qui riferimento a dati medici di pazienti.

Secondo una forma di attuazione, il passo relativo alla raccolta dei dati reali consiste di una fase di upload di una pluralità di dati, detta fase
25 essendo eseguita preferibilmente da un utente.

Preferibilmente, detta fase di upload di una pluralità di dati viene eseguita da un utente mediante una interfaccia web.

Il detto utente può avere accesso alla detta
30 interfaccia web mediante qualsivoglia dispositivo elettronico dotato di uno schermo per la visualizzazione di informazioni e contenuti, detto

dispositivo risultando abilitato ad una connessione internet.

Detto dispositivo elettronico può fare riferimento, in maniera esemplificativa ma non
5 limitativa, ad un computer o ad uno smartphone.

Detta interfaccia web fa riferimento ad una interfaccia del modello generativo che provvederà la generazione dei dati sintetici a partire dal detto insieme di dati reali, detta interfaccia web
10 risultando accessibile ad un utente per mezzo di un dispositivo elettronico come definito.

Conseguentemente, detto passo relativo al raccogliere dati reali consiste preferibilmente di una fase di upload di dati tabulari eseguita da un
15 utente mediante una interfaccia web, dove detto utente visualizza detta interfaccia web da un dispositivo elettronico come definito.

Secondo una forma di attuazione, detto passo di raccolta di dati reali può comprendere differenti
20 metodi di upload della pluralità di dati.

Detti differenti metodi di upload di dati possono preferibilmente variare in relazione alla dimensione dei dati stessi.

Conseguentemente l'utente può, a seconda della
25 dimensione dei dati scelti, impostare una particolare modalità di upload dei detti dati mediante l'interfaccia web che visualizza dal dispositivo.

Secondo una forma di attuazione, il metodo oggetto della presente domanda di brevetto provvede,
30 a seguito della raccolta di dati reali realizzata dall'utente, una visualizzazione grafica dei detti dati reali tabulari mediante l'interfaccia web sul dispositivo elettronico.

Detta visualizzazione grafica risulta accessibile all'utente.

Il metodo oggetto della presente domanda di brevetto, a seguito del passo relativo alla raccolta
5 e visualizzazione grafica di dati reali tabulari, provvede una definizione della tipologia di detti dati tabulari raccolti.

Secondo una forma di attuazione, detta definizione della tipologia dei dati tabulari
10 raccolti viene fornita dall'utente, che precedentemente ha provveduto l'upload dei dati da sintetizzare.

Preferibilmente, detta fase di definizione della tipologia di detti dati raccolti viene eseguita
15 dall'utente mediante l'interfaccia web.

Conseguentemente detto utente può, mediante il dispositivo elettronico, provvedere una definizione relativa alla tipologia dei dati da lui forniti.

Come detto l'utente, a seguito dell'upload dei
20 dati tabulari da sintetizzare, visualizza sull'interfaccia web i detti dati organizzati secondo almeno una riga e una colonna, preferibilmente secondo una pluralità di righe e colonne.

Secondo una forma di attuazione preferita, in
25 detta fase di definizione della tipologia dei dati raccolti, l'utente provvede almeno una definizione per ciascuna riga e colonna secondo la quale i detti dati reali risultano organizzati.

Una volta fornita una definizione relativa alla
30 tipologia di dati raccolti, il metodo proposto provvede una fase di addestramento di almeno un modello generativo sulla base dei detti dati tabulari

reali raccolti, detto modello generativo comprendendo almeno un autoencoder variazionale VAE.

I modelli generativi sono comunemente definiti come strumenti in grado di apprendere, a partire da
5 un certo insieme di dati in ingresso - e a seguito di un periodo detto appunto di "addestramento" su detti dati - come detti dati sono strutturati da una prospettiva statistica e, successivamente, sono in grado di generare nuove istanze di tali dati che
10 riproducono i modelli appresi.

Esempi di altri modelli generativi fanno ad esempio riferimento a reti neurali, quali le reti generative avversarie o GANs.

Per convenzione, nel seguito della descrizione,
15 si farà riferimento ad un autoencoder variazionale VAE ogni qual volta in cui si vorrà fare riferimento al modello generativo per la generazione di dati sintetici a partire dall'insieme di dati tabulari reali.

20 La detta interfaccia web permette al detto utente di interagire con il detto autoencoder variazionale VAE.

Secondo il metodo proposto, la detta fase di addestramento del detto autoencoder variazionale VAE
25 viene eseguita, preferibilmente, in maniera automatica una volta fornita, al detto autoencoder, una definizione della tipologia di dati per ciascuna riga e colonna secondo le quali i dati tabulari raccolti risultano organizzati.

30 L'autoencoder variazionale VAE, secondo il metodo oggetto della presente domanda di brevetto, provvede quindi, a seguito di detta fase iniziale di addestramento, una generazione di dati sintetici.

Secondo una forma di attuazione, detta generazione di dati sintetici viene eseguita in maniera automatica dal detto autoencoder variazionale a seguito della fase di addestramento con i dati
5 tabulari reali raccolti.

Secondo una prima forma di attuazione, il detto autoencoder variazionale VAE può fare riferimento ad un autoencoder variazionale β , o β -VAE.

Secondo una forma di attuazione, detti dati
10 tabulari sono identicamente indipendentemente distribuiti e detto autoencoder variazionale VAE consiste di almeno un autoencoder variazionale Beta β -VAE comprendente un encoder ed un decoder.

Detto β -VAE consiste di una pluralità di livelli
15 o strati di una rete neurale artificiale (aNN) interconnessi tra loro.

Un primo gruppo di livelli o strati viene definito encoder o codificatore, e viene sfruttato per addestrare i parametri di una distribuzione di
20 probabilità (il cosiddetto "code", o codice) che modella variabili latenti nei dati in ingresso.

Un secondo insieme di livelli o strati, definito decoder o decodificatore, viene addestrato per decodificare campioni dalla detta distribuzione in
25 istanze di dati realistici.

Questo può essere ottenuto riducendo al minimo la probabilità logaritmica dei dati osservati derivanti dal codice.

Preferibilmente, la somiglianza tra il codice e
30 la distribuzione reale di probabilità sottostante dei dati iniziali può essere ottenuta impostando una soglia in relazione alla loro divergenza di Kullback-

Leibler, o divergenza KL, ai fini di una ottimizzazione. Partendo conseguentemente da

$$\begin{aligned} & \max_{\phi, \theta} \mathbb{E}_{X \sim D} [\mathbb{E}_{X \sim D} (Z|X) \log p_{\theta}(x|z)] \\ 5 \quad & \text{s.t. } D_{KL}(q_{\phi}(z|x) \| p(z)) < \delta \end{aligned}$$

e impiegando le condizioni necessarie Karush-Kuhn-Tucker (KKT) in relazione ai problemi di ottimizzazione vincolati, le fasi di addestramento
10 del detto VAE e successiva generazione di dati sintetici secondo il metodo proposto portano all'ottenimento della seguente funzione Lagrangiana

$$F = \mathbb{E}_{Z \sim q_{\phi}(Z|X)} \log p_{\theta}(x|z) - \beta [D_{KL}(q_{\phi}(z|x) \| p(z)) - \delta]$$

15

dove massimizzare una delle due funzioni equivale a massimizzare l'altra.

La funzione obiettivo del detto β -VAE può essere ideata minimizzando la negazione del limite inferiore
20 della funzione Lagrangiana come segue

$$\mathcal{L}(\phi, \beta) = -\mathbb{E}_{Z \sim q_{\phi}(Z|X)} \log p_{\theta}(x|z) + \beta D_{KL}(q_{\phi}(z|x) \| p(z))$$

Dove il moltiplicatore Lagrangiano β fa
25 riferimento ad un iper-parametro o fattore di peso.

Preferibilmente β viene impiegato dall'autoencoder variazionale nelle dette fasi di addestramento e generazione del metodo proposto nella presente domanda di brevetto per bilanciare un
30 adeguato quantitativo di anonimizzazione dei dati e la qualità dei dati sintetici generati in termini di loro distribuzioni, affinché i dati sintetici

generati siano il più simili possibile ai dati tabulari originali.

Preferibilmente, il detto β -VAE può inoltre risultare implementato mediante due reti neurali, dove dette reti risultano configurate per modellare l'encoder o codificatore e il decoder o decodificatore.

Secondo detta forma di attuazione, l'encoder risulta configurato per ricevere in ingresso una dimensione n_x , che risulta essere il numero di colonne dell'insieme di dati tabulari.

A seguito della fase di addestramento, il detto β -VAE implementato provvede un vettore y di dimensione n_x , che dipende dalla differente tipologia di dati contenuti nella tabella dei dati reali.

Il detto β -VAE implementato, una volta che il detto termine n_x viene calcolato, tratta la dimensionalità n_z del codice (anche definito spazio latente) e il numero L di livelli o strati come iperparametri.

La progettazione delle dette reti neurali usate per modellare l'encoder è riportata di seguito:

$$\left\{ \begin{array}{l} y = f(x) \\ h_1 = \text{ReLU}(FC_{n_x \rightarrow n_{h_1}}(y)) \\ h_2 = \text{ReLU}(FC_{n_{h_1} \rightarrow n_{h_2}}(h_1)) \\ \dots \\ h_L = \text{ReLU}(FC_{n_{h_{L-1}} \rightarrow n_{h_L}}(h_{L-1})) \\ \mu = FC_{n_{h_L} \rightarrow n_z}(h_L) \\ \sigma^2 = \exp(FC_{n_{h_L} \rightarrow n_z}(h_L)) \\ q_\phi(z|x) \sim \mathcal{N}(\mu, \sigma) \end{array} \right.$$

25

La progettazione delle dette reti neurali usate per modellare il decoder è riportata di seguito:

$$\left\{ \begin{array}{l} \hat{h}_1 = \text{ReLU} (FC_{n_z \rightarrow n_{h_1}} (z)) \\ \hat{h}_2 = \text{ReLU} (FC_{n_{h_1} \rightarrow n_{h_2}} (\hat{h}_1)) \\ \dots \\ \hat{h}_L = \text{ReLU} (FC_{n_{h_{L-1}} \rightarrow n_{h_L}} (\hat{h}_{L-1})) \\ y = FC_{n_{h_L} \rightarrow n_x} (\hat{h}_L) \\ p_i \sim f_i(y_i) \\ p_\theta(x|z) \sim \prod_{i=1}^{n_x} p_i(x_i|z) \end{array} \right.$$

Secondo una seconda forma di attuazione,
 5 alternativa o combinabile con la precedente, il detto
 autoencoder variazionale VAE può fare riferimento ad
 un autoencoder variazionale introspettivo, o Intro-
 VAE.

Secondo una forma di attuazione, detti dati
 10 tabulari sono identicamente indipendentemente
 distribuiti e detto autoencoder variazionale VAE
 consiste di almeno un autoencoder variazionale
 introspettivo Intro-VAE comprendente almeno un
 termine avversario.

15 Detto autoencoder variazionale introspettivo può
 consistere di una architettura ibrida, comprendente
 un autoencoder variazionale standard e almeno un
 termine avversario di una rete generativa avversaria
 GAN.

20 Detto autoencoder variazionale introspettivo
 vantaggiosamente combina i vantaggi di un autoencoder
 variazionale standard, quali una fase di
 addestramento stabile ed un manifold latente
 coerente, con i vantaggi derivanti dall'utilizzo di
 25 una rete generativa avversaria GAN, quali campioni
 generati che risultano essere altamente realistici.

Vantaggiosamente, inoltre, l'aggiunta di almeno un termine avversario tipico di architetture GAN provvede una minimizzazione dell'errore semantico.

Sempre vantaggiosamente, l'architettura Intro-VAE non richiede un discriminatore aggiuntivo, in quanto il modello di interferenza può fungere esso stesso da discriminatore.

Durante le fasi di addestramento e generazione del metodo oggetto della presente domanda di brevetto, l'autoencoder variazionale introspettivo combina due obiettivi: da un lato minimizza la discrepanza tra i campioni osservati $q_\phi(z|x)$ e il precedente termine $p_\phi(z)$; dall'altro, la probabilità posteriore dei campioni generati $q_\phi(z|G(z))$ è costretta a divergere da $p_\phi(z)$.

Al contrario, il generatore risulta essere incoraggiato a decodificare campioni $G(z')$ che, quando codificati, mostrano una probabilità posteriore $q_\phi(z|G(z'))$ differente dal precedente termine $p_\phi(z)$.

Preferibilmente, il detto Intro-VAE può prevedere la seguente implementazione.

Come detto, l'autoencoder variazionale introspettivo mira a separare la codifica di campioni reali z , ottenuti codificando i dati reali $\text{Enc}(x)$, sia dalla codifica dei nuovi campioni x_p ottenuti mediante campionamento dalla probabilità posteriore $q_\phi(z|x)$ sia dalla codifica del dato $\text{Enc}(x_r)$ ricostruito.

Conseguentemente, le funzioni di loss totali per il generatore G e il discriminatore D sono le seguenti

$$\begin{aligned}
L_E &:= L_{reg}(z) + \alpha \sum_{s=r,p} [m - L_{reg}(z_s)]^+ + \beta L_{err}(x, x_r) \\
&:= L_{reg}(Enc(x)) + \alpha \sum_{s=r,p} \left[m - L_{reg}(Enc(ng(x_s))) \right]^+ + \beta L_{err}(x, x_r) \\
L_G &:= \alpha \sum_{s=r,p} L_{reg}(Enc(x_s)) + \beta L_{err}(x, x_r)
\end{aligned}$$

5

L'implementazione provvede uno spostamento del parametro β davanti al termine di regolarizzazione L_{reg} .

10 Detta implementazione vantaggiosamente permette al detto autoencoder di rassomigliare la struttura dell'architettura β -VAE.

La modifica descritta non provvede un danneggiamento delle performance, essendo β un fattore di peso.

15 I detti dati sintetici provvisti dall'autoencoder variazionale VAE risultano vantaggiosamente privi delle informazioni sensibili presenti nei dati tabulari originali e, al contempo, presentano le stesse proprietà statistiche
20 dell'insieme di detti dati reali di partenza.

Il detto metodo provvede quindi la generazione di un insieme di dati sintetici che rassomiglia, in termini di proprietà statistiche, l'insieme dei dati reali di partenza. Il detto metodo provvede la
25 generazione di un insieme di dati sintetici che risultano essere immuni da eventuali attacchi volti a ricavare, dal detto dato sintetico, le informazioni sensibili contenute nel dato reale di partenza.

30 Nel caso in cui i dati reali tabulari di partenza facciano riferimento a dati medici di

pazienti, questi possono presentare dipendenze, ad esempio possono fare riferimento a risultati di molteplici esami fatti da uno stesso soggetto, quindi possono descrivere differenti eventi che possono fare riferimento ad una medesima entità.

Il metodo oggetto della presente domanda di brevetto provvede vantaggiosamente la possibilità di trattare anche questa casistica, essendo il detto modello generativo un VAE multilivello come definito di seguito.

Il metodo oggetto della presente domanda di brevetto provvede quindi una generazione di una pluralità di entità o utenti sintetici.

Secondo una forma di attuazione quindi, con il termine "dati tabulari" si fa anche riferimento, in maniera non limitativa, ad una pluralità di entità ed eventi, detti dati essendo raggruppati per entità quando più eventi possono essere ricondotti ad una medesima entità, ossia quando sono presenti una o più dipendenze.

Secondo una forma di attuazione detti dati tabulari sono non identicamente indipendentemente distribuiti e rappresentano entità ed eventi, detti dati essendo raggruppati per entità.

Detti dati possono fare riferimento ad una tabella come la seguente:

Dati di utenti (u)			Dati evento (e)			
Nome	Sesso	Età	Oggetto	Materiale	Colore	Prezzo
Marc	M	45	Pantalone	Jeans	Blu	16
Marc	M	45	Maglia	Cotone	Bianco	9
Hanna	F	32	Gonna	Lana	Rosso	15
Hanna	F	32	Giacca	Pelle	Nero	42

La tabella provvede righe o record nella forma $x = (u, e)$ dove u = dato utente ed e = dato evento.

Il detto autoencoder variazionale VAE del metodo oggetto della presente domanda di brevetto agisce
5 come se ciascun utente fosse indipendentemente tratto da una distribuzione $p_u(u)$ e, dato l'utente, i suoi eventi fossero campionati in maniera indipendente da una distribuzione $p_e(u|e)$.

Conseguentemente, ciascun campione deriva da una
10 distribuzione $p(x) \sim p_u(u)p_e(e|u)$.

Considerando $\mathbb{X}_u = [(u, e_1), \dots, (u, e_{n_u})]$ come tutti i record per un certo utente u , allora la distribuzione totale dell'utente $\mathbb{X}_u$ viene modellata dall'autoencoder variazionale secondo l'equazione

15

$$\mathbb{P}(\mathbb{X}_u) \sim p_u(u) \prod_{i=1}^{n_u} p_e(e_i|u)$$

Il metodo oggetto della presente domanda di brevetto sfrutta una distribuzione precedente sopra
20 uno spazio di codice latente $z = (c, s)$, dove c è il contenuto e codifica gli utenti, mentre s è lo stile e codifica gli eventi.

Data una certa distribuzione di probabilità a priori sullo spazio $z = (c, s)$, la detta distribuzione
25 viene approssimata come segue

$$p(x) \sim p(c)p_u(u|c) \times p(s)p_e(e|c, s)$$

L'autoencoder sfrutta due reti di codifica, che
30 danno $p(c, s|x)$ in un caso o $p(c|u)$ e $p(s|e)$ nell'altro.

Dato un gruppo $G \in \hat{G}$ dove $|G| = i$ e dove i rappresenta la variabile indipendente, l'autoencoder fa collassare il codice c_i per ciascun punto dati in G su un singolo codice C_G secondo le seguenti equazioni

5

$$q(C_G = c | X_g = x_g; \phi_c) \propto \prod_{i \in G} q(C_G = c | X_i = x_i; \phi_c)$$

$$p(\mathbb{X}_u) \sim p_u(u) \prod_{i=1}^{n_u} p_e(e_i | u)$$

L'autoencoder campiona poi le distribuzioni di
 10 probabilità a priori $p(c)$, $p(s_i)$ per $i=1, \dots, n_u$ dove n è il numero di eventi per utente u .

L'autoencoder inizialmente ha il problema di sapere quanti eventi campionare per utente, a causa dell'esistenza di possibili correlazioni.

15 Il problema viene superato provvedendo l'aggiunta di una colonna ai dati utente u , detta colonna essendo finalizzata a raccogliere informazioni circa il numero di eventi per utente.

Successivamente il metodo provvede
 20 l'eliminazione dell'ipotesi secondo la quale, dato un utente, ciascun evento sia estratto in maniera indipendente.

Secondo una forma di attuazione, detti eventi presentano una relazione temporale tra di loro e il
 25 detto modello generativo utilizza schemi ricorrenti per codificare e decodificare detti eventi.

L'autoencoder considera ciascun utente come indipendente, ma per ciascun utente $p_e(e_1, \dots, e_{n_u} | u)$ non può essere scritta come il prodotto sugli eventi.

30 Ciò è reso possibile impiegando una rete neurale ricorrente combinata al detto autoencoder variazionale VAE.

Detta rete neurale ricorrente permette di rappresentare una distribuzione $p(e_1, \dots, e_{n_u} | c, s)$, e il detto modello generativo comprendente l'autoencoder variazionale e la detta rete neurale ricorrente
5 permette e provvede la generazione di insiemi di dati relazionali molto complicati, comprendenti ad esempio relazioni temporali.

Preferibilmente, il detto modello generativo comprendente almeno un autoencoder variazionale e
10 almeno una rete neurale ricorrente provvede una fase di isolamento dei fattori indipendenti ed una fase di costruzione di uno spazio latente che rifletta questi ultimi.

Il modello generativo provvede reti che
15 riflettano la struttura delle relazioni tra variabili differenti, o tabelle nominative, e impieghino schemi ricorrenti per codificarle/decodificarle quando non sono estratte in maniera indipendente.

Secondo una forma di attuazione, il detto metodo
20 può inoltre comprendere un ulteriore passo come segue:

e) condividere detti dati generati sinteticamente.

Secondo una forma di attuazione preferita, ma
25 non limitativa, detto ulteriore passo relativo alla condivisione di detti dati generati sinteticamente consiste del download di detti dati ad opera dell'utente che ha originariamente fornito i dati reali tabulari.

30 Secondo il metodo qui proposto conseguentemente, l'utente che inizialmente ha provveduto ad eseguire l'upload dei dati reali e a fornire una definizione per ciascuna riga e colonna in cui detti dati sono

stati organizzati in maniera automatica, può provvedere al download dell'insieme dei dati sintetici generati a partire dai dati reali.

La presente domanda di brevetto fa inoltre
5 riferimento ad un sistema parzialmente operabile da un utente per la generazione di dati sintetici a partire da dati reali comprendente:

a) un dispositivo elettronico abilitato ad una interfaccia web;

10 b) almeno una unità di elaborazione per il processamento dei dati reali,

in cui la detta unità di elaborazione è configurata per il processamento dei detti dati reali di tipo tabulare e la detta unità di elaborazione
15 comprende almeno un autoencoder variazionale VAE.

Secondo una forma di attuazione, il detto sistema provvede un upload di un insieme di dati reali tabulari selezionati da un utente verso almeno una unità di elaborazione comprendente almeno un
20 autoencoder variazionale VAE.

Preferibilmente, detto upload è reso possibile mediante la detta interfaccia web, che risulta accessibile all'utente sul detto dispositivo elettronico.

25 La detta interfaccia web fa riferimento ad una interfaccia grafica della detta unità di elaborazione, detta interfaccia risultando accessibile all'utente sul detto dispositivo elettronico.

30 Preferibilmente, la detta interfaccia web fa riferimento ad una interfaccia grafica dell'almeno un autoencoder variazionale VAE compreso nella detta unità di elaborazione per il processamento dei dati.

Secondo una forma di attuazione, la detta unità di elaborazione può comprendere, in aggiunta ad almeno un autoencoder variazionale VAE, sistemi di reti neurali, quali ad esempio reti generative avversarie o GANs e/o altro.

Secondo una forma di attuazione il sistema provvede, a seguito dell'upload della pluralità di dati tabulari eseguita da un utente, un processamento di detti dati mediante la detta unità di elaborazione, dove la detta unità di elaborazione provvede un addestramento di almeno un autoencoder variazionale VAE eseguito su detti dati tabulari raccolti e una successiva generazione di una pluralità di dati sintetici a partire dal detto insieme di dati tabulari di partenza.

Secondo una prima forma di attuazione, la detta unità di elaborazione comprende un autoencoder variazionale β , o β -VAE.

Secondo una forma di attuazione, detti dati tabulari sono identicamente indipendentemente distribuiti e la detta unità di elaborazione comprende almeno un autoencoder variazionale Beta β -VAE comprendente un encoder ed un decoder.

Detto β -VAE consiste di una pluralità di livelli o strati di una rete neurale artificiale (aNN) interconnessi tra loro.

Opzionalmente, detta unità di elaborazione può prevedere implementazioni del detto autoencoder variazionale β come definite in precedenza.

Secondo una seconda forma di attuazione, la detta unità di elaborazione comprende, in alternativa o in combinazione con il caso precedente, un autoencoder variazionale introspettivo, o Intro-VAE.

Secondo una forma di attuazione, detti dati tabulari sono identicamente indipendentemente distribuiti e la detta unità di elaborazione comprende almeno un autoencoder variazionale
5 introspettivo Intro-VAE comprendente almeno un termine avversario.

Detto autoencoder variazionale introspettivo può consistere di una architettura ibrida, comprendente un autoencoder variazionale standard e almeno un
10 termine avversario di una rete generativa avversaria GAN.

Opzionalmente, detta unità di elaborazione può prevedere implementazioni del detto autoencoder variazionale introspettivo come definite in
15 precedenza.

Secondo una forma di attuazione, la detta unità di elaborazione del sistema può comprendere una pluralità di nodi HDFS, detti nodi HDFS risultando configurati per conservare, in modo sicuro, i detti
20 dati raccolti.

Vantaggiosamente ciò garantisce, per il sistema oggetto dell'invenzione, una possibile scalabilità verso insiemi di dati anche di grandi dimensioni.

Il sistema può inoltre comprendere un sistema
25 coordinatore, configurato per contenere tutte le informazioni dell'insieme di dati raccolti.

Detto sistema coordinatore risulta inoltre configurato per contenere tutte le informazioni eseguite dall'utente sul detto insieme di dati.

30 Secondo una forma di attuazione, detto sistema coordinatore può sfruttare, in maniera esemplificativa e non limitativa, un sistema "job queue" (JQS, Job queue system).

Il sistema, secondo una forma di attuazione, può comprendere una pluralità di nodi di elaborazione.

Detti nodi di elaborazione hanno preferibilmente accesso diretto a detti nodi HDFS.

5 Secondo una forma di attuazione preferita, detti nodi di elaborazione sono impiegati per addestrare il detto autoencoder variazionale VAE.

10 Il sistema oggetto della presente domanda di brevetto provvede che, a seguito del completamento del detto addestramento dell'autoencoder variazionale VAE, il detto autoencoder venga mantenuto su detti nodi HDFS in vista di un possibile utilizzo futuro.

15 Vantaggiosamente ciò provvede la possibilità di ripristinare il detto autoencoder variazionale VAE addestrato per generare una qualsiasi quantità di dati sintetici in un qualsiasi momento futuro, anche a seguito della cancellazione dei detti dati originali.

20 Secondo una forma di attuazione, detti dati tabulari rappresentano entità ed eventi, detti dati essendo raggruppati per entità.

25 Secondo una forma di attuazione, la detta unità di elaborazione può comprendere, oltre al detto autoencoder variazionale, almeno una rete neurale ricorrente.

30 Vantaggiosamente detta unità di elaborazione così costituita può processare dati costituiti da una pluralità di entità ed eventi, dove detti dati risultano raggruppati per entità quando più eventi possono essere ricondotti ad una medesima entità.

 Secondo una forma di attuazione infatti, con il termine "dati tabulari" si può fare riferimento, in maniera non limitativa, ad una pluralità di entità ed

eventi, detti dati essendo raggruppati per entità quando più eventi possono essere ricondotti ad una medesima entità, ossia quando sono presenti una o più dipendenze.

5 Secondo una forma di attuazione, detti eventi presentano una relazione temporale tra di loro e la detta unità di elaborazione è configurata per utilizzare schemi ricorrenti per codificare e decodificare i detti eventi.

10 La detta unità di elaborazione comprendente almeno un autoencoder variazionale VAEe almeno una rete neurale ricorrente può provvedere il processamento di insiemi di dati relazionali molto complicati, comprendenti ad esempio relazioni
15 temporali.

Il sistema oggetto della presente domanda di brevetto, a seguito dell'addestramento del detto modello generativo, provvede una generazione di un insieme di dati sintetici.

20 Secondo una forma di attuazione, il sistema oggetto della presente domanda di brevetto può infine risultare configurato per provvedere il download dei detti dati sintetici generati, detti dati sintetici risultando conseguentemente accessibili al detto
25 utente mediante l'interfaccia web accessibile dal detto dispositivo elettronico.

Queste ed altre caratteristiche e vantaggi della presente invenzione risulteranno più chiaramente dalla seguente descrizione di alcuni esempi esecutivi
30 illustrati nei disegni allegati in cui:

la fig. 1 illustra una rappresentazione delle principali problematiche legate all'usabilità di un dato raccolto in relazione alla protezione delle

informazioni sensibili contenute nel dato stesso a seguito di processo di anonimizzazione;

la fig. 2 illustra un diagramma di flusso del metodo oggetto della presente domanda di brevetto
5 secondo una forma di attuazione;

la fig. 3 illustra una rappresentazione schematica visiva del sistema oggetto della presente domanda di brevetto.

La fig. 4 illustra l'architettura dell'unità di elaborazione per il processamento dati del sistema.
10

Si specifica che le figure allegate alla presente domanda di brevetto illustrano una possibile forma esecutiva del metodo implementato da computer e sistema per la generazione di dati sintetici oggetto
15 della presente invenzione, per meglio comprenderne i vantaggi e le caratteristiche descritte.

Tale forma esecutiva è dunque da intendersi a puro scopo illustrativo e non limitativo al concetto inventivo della presente invenzione, ossia quello di
20 provvedere un metodo implementato da computer ed un relativo sistema per la generazione di dati sintetici come definiti.

Con particolare riferimento alla Figura 1, viene illustrata una rappresentazione delle principali
25 problematiche legate all'usabilità di un dato raccolto in relazione alla protezione delle informazioni sensibili contenute nel dato stesso a seguito di un processo di anonimizzazione. In particolar modo le tecniche di anonimizzazione
30 presenti nello stato dell'arte provvedono, come si può vedere nel grafico riportato, al crescere della protezione delle informazioni sensibili contenute nel dato stesso, una usabilità del dato che decresce

proporzionalmente. Conseguentemente ne deriva che, se si vuole provvedere una protezione delle informazioni sensibili mediante un processo di anonimizzazione, allora il detto dato anonimizzato non risulterà più di alcuna utilità reale. Le informazioni statistiche di interesse, infatti, vengono compromesse sempre più al progredire del livello di anonimizzazione fornito ai dati stessi. Come si evince dalla Figura, questa problematica viene superata grazie ai dati sintetici (corrispondenti a persone sintetiche), che garantiscono, allo stesso tempo, una usabilità del dato ai fini di analisi statistiche e/o altro e una adeguata protezione totale delle informazioni sensibili in rispetto alle normative relative alla privacy.

Figura 2 illustra un diagramma di flusso del metodo oggetto della presente domanda di brevetto secondo una forma di attuazione. Nel dettaglio, la figura illustra un metodo 1 comprendente un primo passo 10 relativo alla raccolta di un insieme di dati reali, essendo detti dati reali in forma tabulare, un secondo passo 11 relativo al provvedere una definizione relativa alla tipologia dei detti dati reali raccolti, un terzo passo 12 che fa riferimento ad un addestramento di un modello generativo comprendente almeno un autoencoder variazionale VAE e, infine, un quarto passo 13 che provvede la generazione di un insieme di dati sintetici a partire dal detto insieme di dati reali raccolti al passo 10.

Preferibilmente, la fase di raccolta di detto insieme di dati reali viene eseguita da un utente, che provvede ad un upload dei dati per mezzo di un dispositivo elettronico abilitato a visualizzare una

interfaccia web del modello generativo impiegato per la generazione dei dati di sintesi, come meglio illustrato nella Figura seguente.

Preferibilmente, a seguito della fase di
5 raccolta di dati mediante upload, il metodo 1
provvede una rappresentazione visiva e/o grafica di
detti dati tabulari, ossia sotto forma di una
pluralità di righe e/o colonne, detta
rappresentazione visiva risultando accessibile al
10 detto utente mediante la detta interfaccia web, come
meglio illustrato nella seguente Figura.

Il metodo provvede poi l'esecuzione del passo
11, che risulta essere eseguito dall'utente mediante
la detta interfaccia web accessibile da un
15 dispositivo elettronico, come meglio illustrato nella
successiva Figura.

Una volta definita la tipologia di dato, ossia
una volta fornita una definizione per ciascuna riga
e/o colonna secondo le quali i dati tabulari raccolti
20 sono organizzati, il metodo provvede una fase di
addestramento del detto modello generativo
comprendente almeno un autoencoder variazionale VAE,
detta fase di addestramento risultando finalizzata
alla generazione di un insieme di dati sintetici.

25 Detto insieme di dati sintetici rassomiglia, in
termini di proprietà statistiche, l'insieme dei dati
reali di partenza. Il detto metodo provvede la
generazione di un insieme di dati sintetici che
risultano essere immuni da eventuali attacchi volti a
30 ricavare, dal detto dato sintetico, le informazioni
sensibili contenute nel detto dato reale di partenza.

Detti dati di sintesi possono inoltre fare
riferimento ad una pluralità di dati utente e dati

evento, dove possono essere presenti più dipendenze, tra cui dipendenze temporali.

Il detto metodo può inoltre comprendere un ulteriore passo relativo al condividere detti dati generati sinteticamente.

Detto ulteriore passo relativo alla condivisione di detti dati generati sinteticamente consiste del download di detti dati ad opera dell'utente che ha originariamente fornito i detti dati reali tabulari, come meglio illustrato nella successiva Figura.

Figura 3 illustra una rappresentazione schematica visiva del sistema oggetto della presente domanda di brevetto. Nel dettaglio, il sistema 2 per la generazione di dati sintetici, che risulta essere parzialmente operabile da un utente, comprende un dispositivo elettronico 4 abilitato ad una interfaccia web 3 e almeno una unità di elaborazione e processamento di dati reali 5, detta unità di elaborazione comprendendo almeno un autoencoder variazionale VAE.

Secondo una forma di attuazione, la detta unità di elaborazione comprende un autoencoder variazionale β , o β -VAE, detto β -VAE consistendo di una pluralità di livelli o strati di una rete neurale artificiale (aNN) interconnessi tra loro.

Secondo una forma di attuazione, la detta unità di elaborazione comprende, in alternativa o in combinazione con il caso precedente, un autoencoder variazionale introspettivo, o Intro-VAE, detto autoencoder variazionale introspettivo consistendo di una architettura ibrida, comprendente un autoencoder variazionale standard e almeno un termine avversario di una rete generativa avversaria GAN.

L'utente (non visualizzato in Figura) provvede, ad un istante T_1 , ad un upload di un insieme di dati reali in forma tabulare tramite una interfaccia web 3 dell'unità di elaborazione 5 verso l'unità di elaborazione stessa, che provvederà alla creazione di un insieme di dati sintetici a partire da detto insieme di dati reali mediante l'almeno un autoencoder variazionale.

L'utente visualizza detta interfaccia web 3 mediante un dispositivo elettronico ad esempio uno smartphone 4 configurato per una connessione internet.

Una volta eseguito l'upload dei dati reali, l'unità di elaborazione 5 provvede una rappresentazione grafica X_{reale} di detti dati tabulari che risulta accessibile all'utente mediante la detta interfaccia web 3 sullo smartphone 4. La detta rappresentazione grafica X_{reale} dei dati tabulari comprende almeno una riga ed una colonna, preferibilmente una pluralità di righe e colonne, consistendo quindi di una tabella.

L'utente deve a questo punto provvedere a specificare, per ciascuna riga e colonna, la tipologia di dati rappresentati.

L'unità di elaborazione 5 provvede quindi una fase di addestramento dell'almeno un autoencoder variazionale VAE, sia esso un β -VAE e/o un Intro-VAE e/o VAE multilivello, che provvede alla generazione di un insieme di dati sintetici $X_{sintesi}$ a partire dal detto insieme di dati reali X_{reale} .

A seguito della generazione dei dati di sintesi, il sistema provvede, in un secondo istante T_2 , il download dei detti dati sul dispositivo 4

dell'utente. L'utente è in grado di visualizzare ed avere accesso all'insieme di dati sintetici X_{sintesi} mediante la detta interfaccia web 3.

Figura 4 illustra l'architettura dell'unità di
5 di elaborazione per il processamento dati del sistema. In particolare detta architettura include un encoder o codificatore, un code o codice, un decoder o decodificatore e livelli di pre-processamento di un generatore di dati sintetici. Nel dettaglio:

- 10 (a) rappresenta il livello di ingresso dei dati;
- (b) identifica un livello di pre-processamento, in cui ciascuna tipologia di dato viene processata in maniera differente prima di essere fornita ad una rete neurale artificiale - aNN;
- 15 (c) identifica il livello di embedding per variabili categoriche;
- (d) identifica un autoencoder variazionale VAE, così come una rete generativa avversaria GAN;
- (e) identifica i livelli di de-embedding per
20 variabili continue;
- (f) identifica un livello di de-processamento;
- (g) identifica i dati sintetici generati.

Mentre l'invenzione è suscettibile di varie modifiche e costruzioni alternative, alcune forme di
25 realizzazione preferite sono state mostrate nei disegni e descritte in dettaglio.

Si deve intendere, comunque, che non vi è alcuna intenzione di limitare l'invenzione alle specifiche forme di realizzazione illustrate, ma, al contrario,
30 essa intende coprire tutte le modifiche, costruzioni alternative, ed equivalenti che ricadano nell'ambito dell'invenzione come definite nelle rivendicazioni.

RIVENDICAZIONI

1. Metodo implementato da computer per la generazione di dati sintetici comprendente i passi di:

- 5 a) raccogliere dati reali;
- b) definire la tipologia di detti dati reali raccolti;
- c) addestrare almeno un modello generativo con detti dati reali;
- 10 d) provvedere la generazione di dati sintetici mediante il detto modello generativo addestrato, caratterizzato dal fatto che i dati reali sono dati tabulari e il modello generativo comprende almeno un autoencoder variazionale (VAE).
- 15

2. Metodo secondo la rivendicazione 1, in cui in cui detti dati tabulari sono identicamente indipendentemente distribuiti e detto autoencoder variazionale VAE consiste di almeno un autoencoder variazionale Beta β -VAE comprendente un encoder ed un decoder.

20

3. Metodo secondo la rivendicazione 1, in cui detti dati tabulari sono identicamente indipendentemente distribuiti e detto autoencoder variazionale VAE consiste di almeno un autoencoder variazionale introspettivo Intro-VAE comprendente almeno un termine avversario.

25

4. Metodo secondo una o più delle precedenti rivendicazioni in cui detti dati tabulari sono non identicamente indipendentemente distribuiti e rappresentano entità ed eventi, detti dati essendo raggruppati per entità.

30

5 5. Metodo secondo la rivendicazione 4, in cui
detti eventi presentano una relazione temporale tra
di loro e il detto modello generativo utilizza schemi
ricorrenti per codificare e decodificare detti
eventi.

6. Sistema parzialmente operabile da un utente
per la generazione di dati sintetici a partire da
dati reali comprendente:

10 a) un dispositivo elettronico abilitato ad una
interfaccia web;

b) almeno una unità di elaborazione per il
processamento dei dati reali,

caratterizzato dal fatto che

15 la detta unità di elaborazione è configurata per
il processamento dei detti dati reali di tipo
tabulare e la detta unità di elaborazione comprende
almeno un autoencoder variazionale VAE.

20 7. Sistema secondo la rivendicazione 6, in cui
detti dati tabulari sono identicamente
indipendentemente distribuiti e la detta unità di
elaborazione comprende almeno un autoencoder
variazionale Beta β -VAE comprendente un encoder ed un
decoder.

25 8. Sistema secondo la rivendicazione 6, in cui
detti dati tabulari sono identicamente
indipendentemente distribuiti e la detta unità di
elaborazione comprende almeno un autoencoder
variazionale introspettivo Intro-VAE comprendente
almeno un termine avversario.

30 9. Sistema secondo una o più delle
rivendicazioni da 6 a 8 in cui detti dati tabulari
sono non identicamente indipendentemente distribuiti

e rappresentano entità ed eventi, detti dati essendo raggruppati per entità.

10. Sistema secondo la rivendicazione 9 in cui detti eventi presentano una relazione temporale tra
5 di loro e la detta unità di elaborazione è configurata per utilizzare schemi ricorrenti per codificare e decodificare i detti eventi.

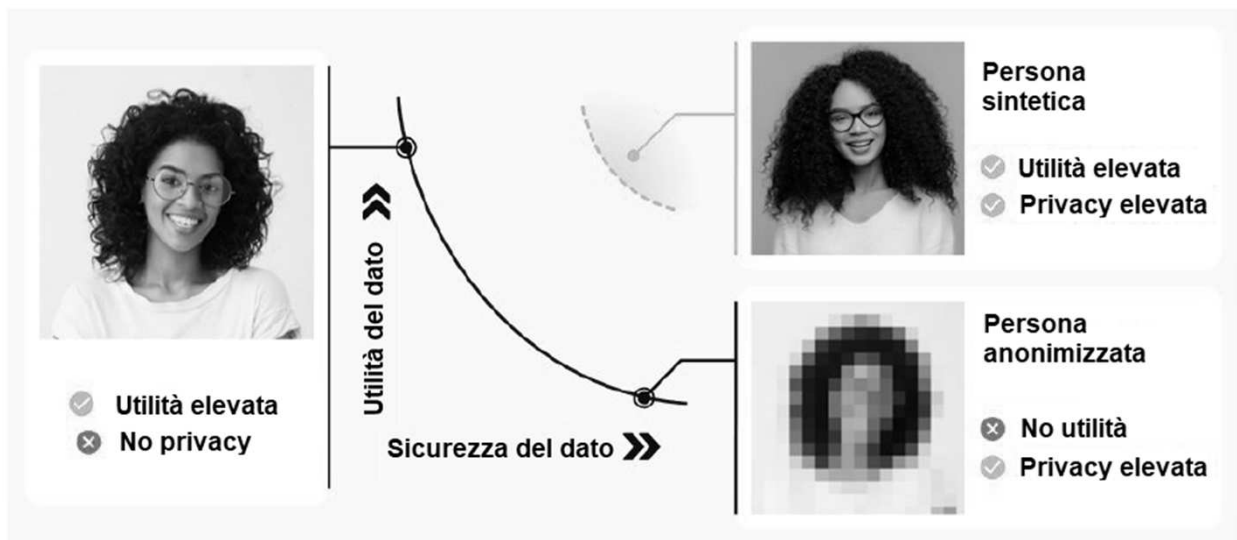


Fig. 1

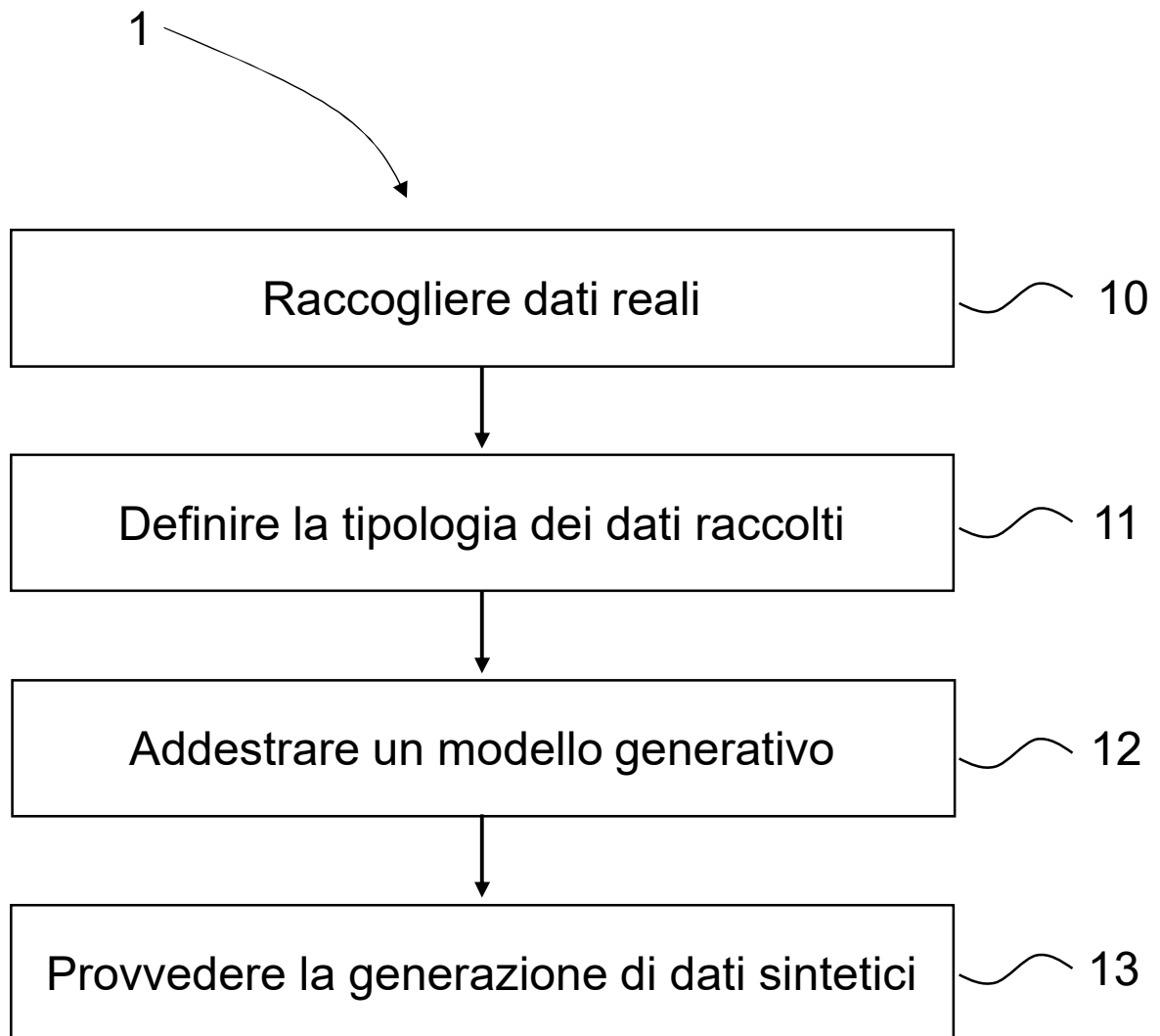


Fig. 2

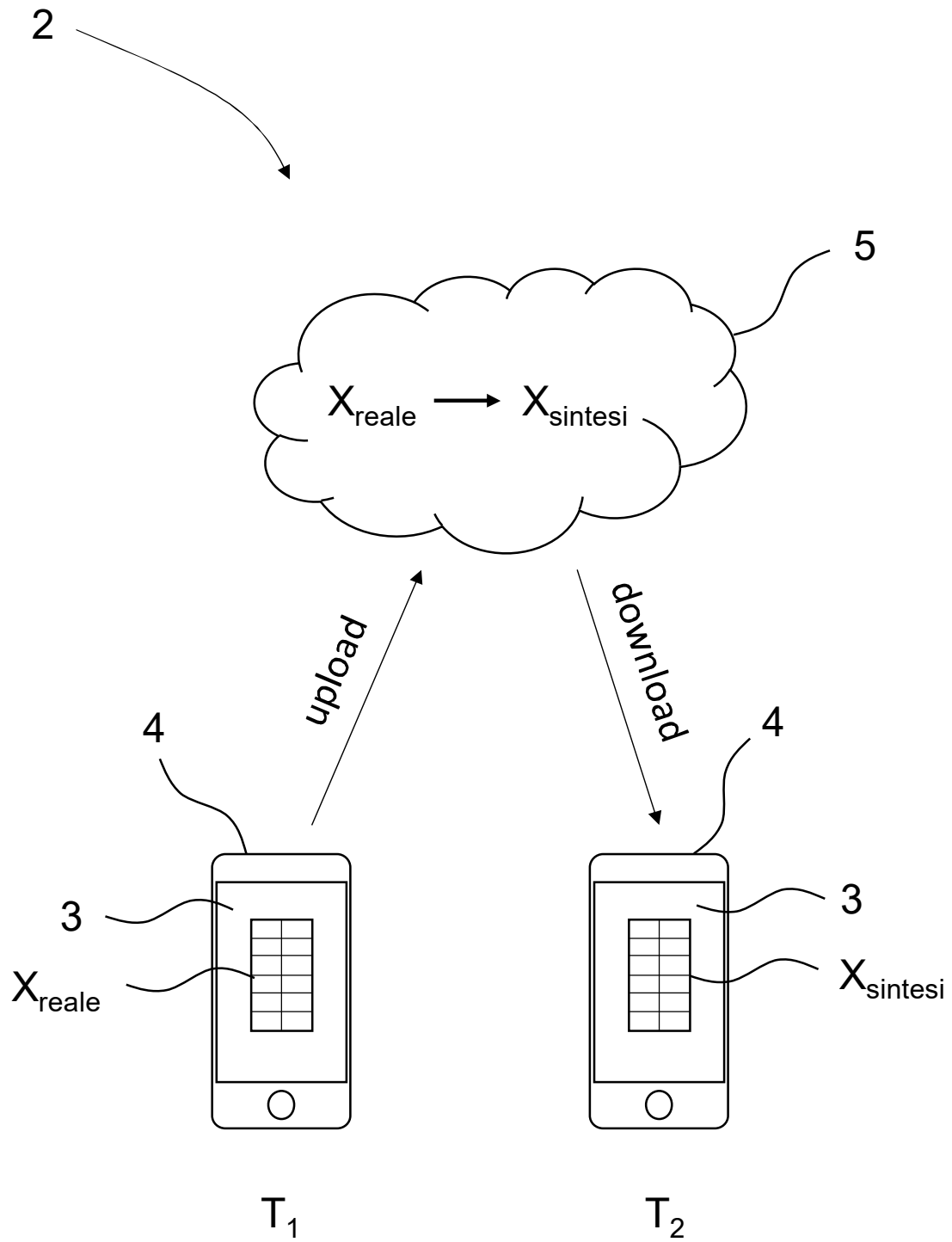


Fig. 3

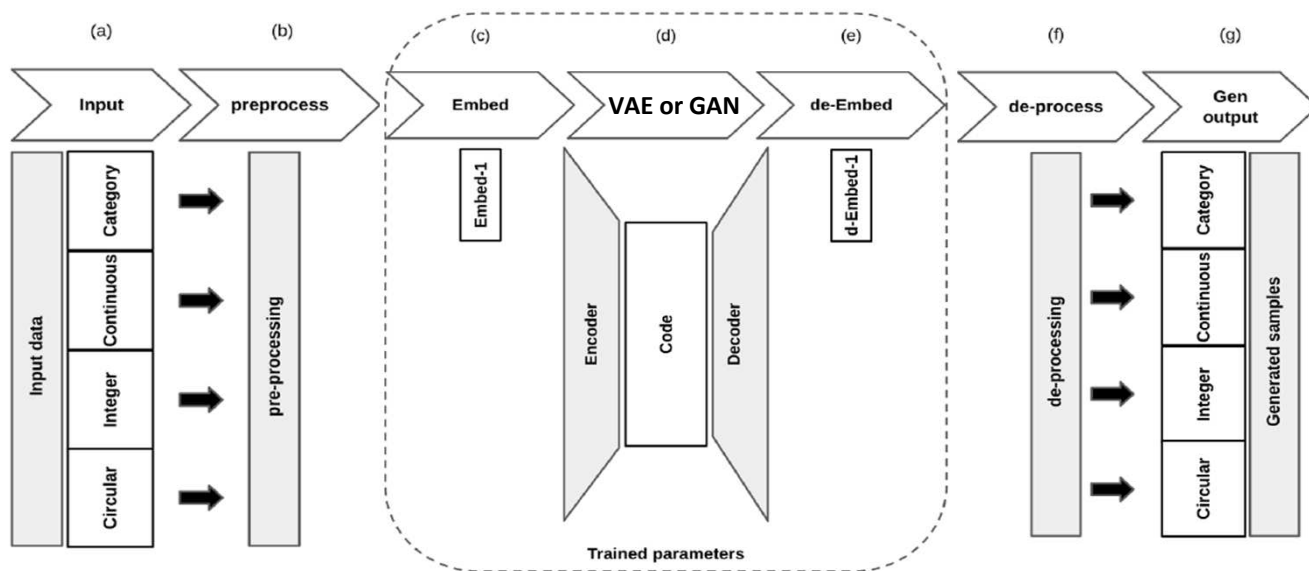


Fig. 4