EDUCAZIONE CIVICA digitale

I Big Data

Quando si parla di **dati**, ci si riferisce alla misurazione primaria di un fenomeno che si è interessati a osservare; dall'elaborazione dei dati si ottengono delle **informazioni**, che ci permettono di aumentare lo stato di conoscenza relativo al fenomeno osservato.

La qualità dell'informazione che otteniamo a partire dai dati può condizionata da fattori di varia natura:

- **fattori quantitativi**: maggiore è la quantità di dati a disposizione, migliore sarà l'informazione e la conoscenza del fenomeno;
- **fattori qualitativi**: tra tutti i dati che caratterizzano un certo scenario, occorre saper individuare quelli effettivamente necessari ai propri bisogni distinguendoli da quelli ininfluenti o fuorvianti;
- **fattori elaborativi**: riguardano le tecniche e gli algoritmi usati per attribuire un significato ai dati;
- **fattori temporali**: l'informazione ha spesso una validità limitata nel tempo e diventa vitale ottenerla il più velocemente possibile.

Figura 1 Piramide DIKW (Data, Information, Knowledge, Winsdom).

Attualmente molte aziende private e pubbliche producono e acquisiscono grandi quantità di dati, che vengono raccolti per essere analizzati e capitalizzati. Le informazioni estratte servono sono necessarie per rimanere competitivi sul mercato (**Figura 1**).

Con il termine *Big Data* si identificano grandi collezioni di dati (che vanno oltre le caratteristiche dei classici strumenti di database) e le tecnologie usate per estrarre da queste le informazioni desiderate. Questo fenomeno si è evoluto in funzione della rapida crescita del numero e della varietà di dati resi disponibili da Internet:

- ***Web and Social media data***: dati *clickstream* (cioè la sequenza di click di un utente su una pagina internet) e interazione con social media, per esempio Facebook, Twitter, Instagram, LinkedIn ecc.
- ***Machine to machine data***: letture di sensori, misuratori e altri dispositivi facenti parte del sistema *Internet of Things* (IoT), in cui oggetti «intelligenti» interconnessi si scambiano dati in automatico.
- ***Big transaction data***: dati economici-finanziari, di tipo sanitario o di telecomunicazioni.
- ***Biometric data***: impronte digitali e facciali, scansioni della retina, dati genetici, calligrafia, ecc.
- ***Human-generated data***: grandi quantità di dati (spesso non strutturati) come email, registrazioni vocali, note di call center, documenti digitalizzati, sondaggi, cartelle cliniche ecc.

Nelle attività quotidiane su tecnologie ICT, lasciamo, più o meno consapevolmente, tracce digitali dei nostri interessi, affari, relazioni, acquisti, comunicazioni, movimenti. Le nostre ricerche su Internet, l'uso del GPS, i messaggi vocali, le e-mail, i tweet o i post su qualsiasi social network generano dati che *seminiamo* dietro di noi: «*siamo tutti dei pollicini digitali*» (D. Pedereschi – UniPi) (**Figura 2**) . In questo senso, i Big Data diventano un vero e proprio osservatorio sociale, in grado di registrare, misurare e prevedere i bisogni di mobilità, di risorse economiche o energetiche, la diffusione di opinioni, crisi, pandemie ecc.

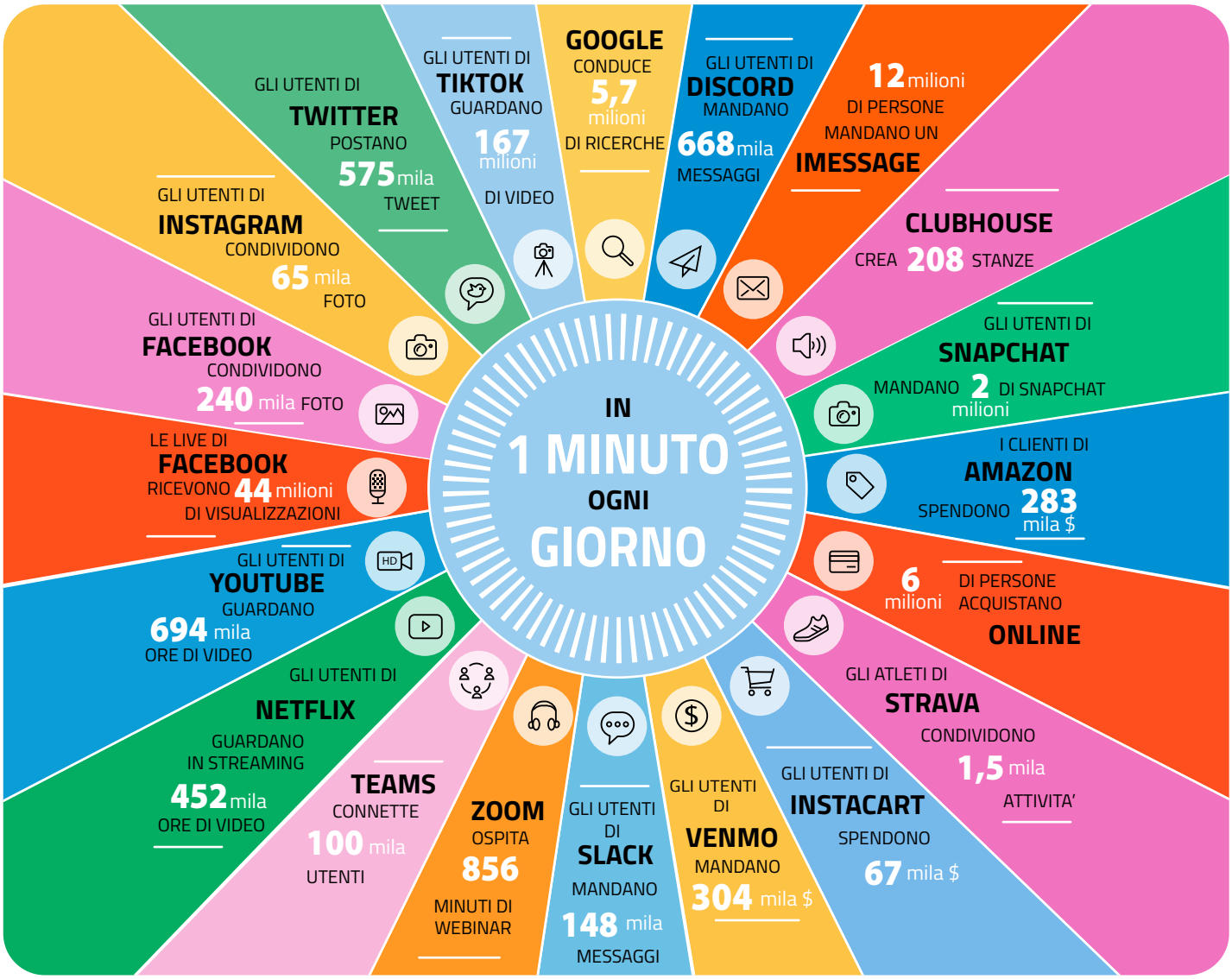

Figura 2 Stima della quantità di dati generati ogni minuto dall'uso dei servizi internet.

1. Big Data e *Cloud Computing*

Con il termine *Cloud Computing* o semplicemente *Cloud* ci si riferisce a un insieme di tecnologie e di servizi offerti da aziende che con i loro *data-center* favoriscono la fruizione e l'erogazione di applicazioni informatiche, di capacità elaborativa e di stoccaggio dati via web (**Figura 3**).

La sinergia tra Big Data e Cloud è assodata: dati e contenuti sensibili per svariate necessità fluiscono ininterrottamente nel Cloud che vede così, giorno dopo giorno, aumentare il proprio valore informativo.

Tutto ciò non è però esente da critiche. Per il leader del movimento per il software libero, Richard Stallman, l'idea di utilizzare datacenter ubicati chissà dove, è solo un ulteriore tentativo dei colossi dell'ICT di ingabbiare gli utenti, poiché l'uso delle applicazioni web comporta *la perdita del controllo* dei propri dati. Egli propone quindi l'uso di network decentralizzati che permettano di navigare anonimamente senza lasciare tracce delle proprie attività. Per esempio, la rete TOR (*The Onion Router*) utilizza circuiti virtuali crittografati a strati per fare in modo che quando una pagina web arriva a destinazione, abbia un indirizzo IP diverso da quello di partenza.

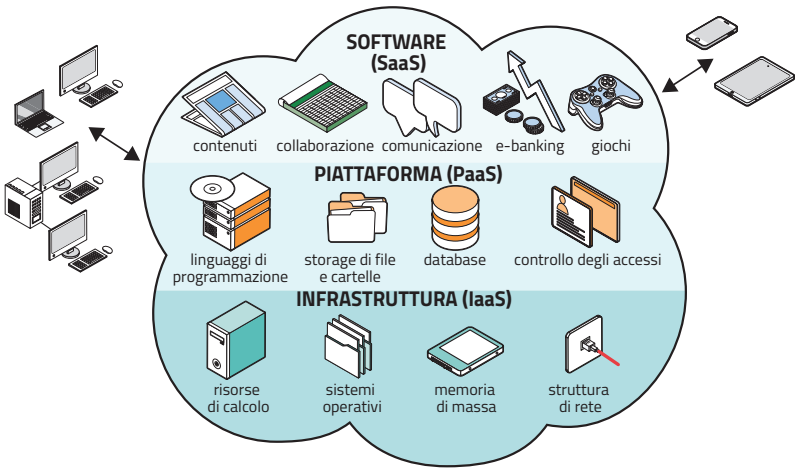

Figura 3 Cloud computing.

2. Big Data e Psicostoriografia

Lo scrittore di fantascienza Isaac Asimov, nel romanzo *Cronache della Galassia* del 1951, immaginò che un suo personaggio, lo scienziato Hari Seldon, nell'ambito dell'ecosistema Galaxia, ideasse la psicostoriografia, cioè una teoria matematica in grado di prevedere, su un periodo temporale molto lungo, i comportamenti di un gruppo numerosamente grande di individui (centinaia di miliardi) se sottoposti a determinati stimoli. Tale approccio si basa sull'idea che gli eventi storici si ripetono ciclicamente e quindi possono essere predetti tramite equazioni matematiche applicate ai profili di un numero elevato di soggetti, a patto che questi rimangano all'oscuro dell'analisi psicostoriografica stessa.

Questa tecnica non può essere applicata all'attuale popolazione terrestre perché non è composta da centinaia di miliardi di individui, ma è possibile applicarla ai Big Data che la popolazione umana produce. In questo senso la psicostoriografia si concretizza permettendo, attraverso i dati, lo studio dei comportamenti umani in termini matematici.

È facile trovare significative analogie tra la psicostoriografia e l'analisi dei Big Data oppure tra il progetto Galaxia e l'IoT: essi perseguono gli stessi obiettivi di implementare una conoscenza globale per lo sviluppo delle attività umane.

3. Big Data Analytics

Per la loro natura di insiemi estremamente grandi di dati, i Big Data non si prestano a essere gestiti con metodi tradizionali perché possono essere anche fortemente inconsistenti. Occorre quindi realizzare collegamenti multipli tra diversi set di dati per individuare correlazioni e gerarchie che ne permettano un adeguato controllo in tempi rapidi. Sviluppare le capacità di gestire e armonizzare tutti questi aspetti è fondamentale per dare un senso ai dati disponibili e sfruttarli al meglio. Per fare ciò si coinvolgono analisti, utenti aziendali e dirigenti in grado di fare ipotesi e porre le domande giuste per individuare tendenze e modelli previsionali.

Le tecnologie e le metodologie di *Big Data Analytics* hanno tre scopi di fondo:

- **descrittivo**: ottenere le informazioni necessarie a trasformare la conoscenza derivante da esse in vantaggio competitivo;
- **predittivo**: stimare i risultati da conseguire. Per esempio, predire la tendenza all'acquisto di un prodotto per particolari tipi di clienti, o rilevare elementi comuni in reati come le frodi per poterli prevenire;
- **prescrittivo**: valutare l'effetto di future decisioni per consigliare possibili scelte in merito. Si mira non solo a prevedere che cosa accadrà, ma anche a spiegarne il perché e, quindi, a fornire consigli sulle azioni da adottare.

Esistono varie tecniche di analisi dei Big Data. Le principali si rifanno al *Business Intelligence* (BI) e al *Machine Learning* (ML).

- **Business Intelligence**: raccoglie dati grezzi su cui vengono poi usati strumenti ETL (*Extract, Transform and Load*) per manipolare, trasformare e classificare i dati in database strutturati detti *data warehouse*. Con strategie di *data mining* si esplorano i dati usando analisi OLAP (*On Line Analytical Process*) e KDD (*Knowledge Discovery in Database*), quindi con *dashboard* semplificate di visualizzazione si rendono le informazioni accessibili a chi deve analizzare e comprendere le prestazioni passate per pianificare le future strategie di miglioramento dei KPI (*Key Business Indicators*) aziendali.
- **Machine learning**: la prima importante differenza rispetto alla BI è che il ML usa l'intelligenza artificiale per rilevare i modelli in milioni di dati. A questa differenza, si possono aggiungere tre aspetti:
  - invece dei dati aggregati, il ML utilizza dati individuali, a livello di singola istanza, così che migliaia di variabili possono essere usate per definire modelli dei fenomeni esaminati;
  - il ML offre analisi predittive, e non descrittive;
  - gli ETL sono sostituiti da applicazioni basate su algoritmi che, in automatico, apprendono costantemente dai dati rilevati per integrarne i modelli e fornire capacità predittive.

Supponiamo, per esempio, che un'azienda di *e-commerce* effettui un'analisi del comportamento dei propri clienti per sapere quanti ne potrebbe perdere nel breve periodo.

Un approccio basato sulla BI utilizzerebbe dati di mesi o anni precedenti, insieme ad altre variabili come le tendenze del mercato o il numero di clienti attuale rispetto al passato. Verrebbero così creati dashboard sulla variazione percentuale di clienti prevista, e da questa si potrebbero realizzare campagne di marketing mirate per acquisire nuovi clienti.

In un approccio ML, invece, si userebbe l'intero database di clienti persi con i loro profili, per cercare modelli di comportamento e determinare quali di loro mostravano segnali di disaffezione e perché.

I dati da usare sarebbero i dettagli degli acquisti storici di tutti i clienti, i loro dati anagrafici, i dati dei prodotti in catalogo (codici identificativi, categorizzazioni, prezzi), i dati delle promozioni o delle campagne di marketing.
Mentre la BI fornisce un'analisi delle tendenze con la percentuale di clienti che probabilmente verranno persi, il ML effettua previsioni cliente per cliente così da intraprendere azioni personalizzate per riconquistarne l'interesse.

4. Big Data e Smart City

Secondo la definizione dell'Unione Europea «*una smart city è un luogo in cui le reti e i servizi tradizionali sono resi più efficienti con l'uso di soluzioni digitali a beneficio dei suoi abitanti e delle imprese. Una città intelligente va oltre l'uso delle tecnologie digitali per un migliore utilizzo delle risorse e minori emissioni. Significa reti di trasporto urbano più intelligenti, impianti di approvvigionamento idrico e di smaltimento dei rifiuti migliorati e modi più efficienti per illuminare e riscaldare gli edifici. Significa anche un'amministrazione cittadina più interattiva e reattiva, spazi pubblici più sicuri e un migliore soddisfacimento delle esigenze di una popolazione che invecchia*» (**Figura 4**).
In una smart city, Big Data e Data Analysis consentono di rilevare e interpretare quantità sempre più importanti di dati strutturati (presenti nei database) e non strutturati (flussi di messaggi, documenti di testo, foto, video, segnali GPS, file audio e social media ecc.) in maniera funzionale per la gestione della città.

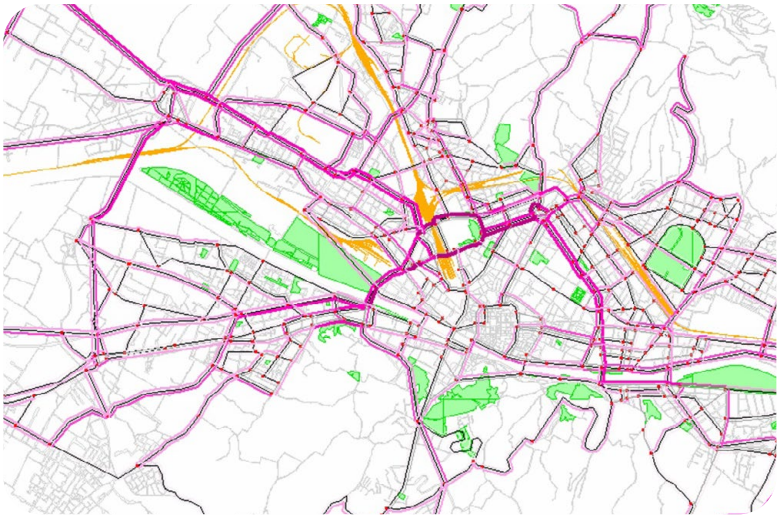

Figura 5 Flussi di traffico stimati su una rete viaria urbana in base alle tracce GPS lasciate dagli smartphone degli automobilisti.

Qui l'IoT gioca un ruolo cruciale nel fornire flussi di dati in tempo reale che permettono di:

- capire ciò che avviene sul territorio per gestirlo in modo puntuale, rapido, efficiente ed efficace;
- prevedere il verificarsi situazioni critiche per poterle prevenire (**Figura 5**).

5. Big Data e Blockchain

L'idea base della tecnologia Blockchain è quella di un registro digitale del consenso pubblico, che permette di registrare in sicurezza transazioni di vario tipo tra più soggetti senza intermediari: la fiducia reciproca viene regolamentata dalla tecnologia e dalla matematica. Per la moneta fisica e la sua trascrizione è necessaria la mediazione delle banche.

Figura 4 Trama informativa di una smart city.

Nel caso del denaro elettronico (le criptovalute), invece, esiste un registro sul quale è scritto a chi appartiene e chi lo tiene in custodia; per mezzo di trascrizioni sul registro, si conferisce il titolo di proprietà da una persona all'altra.

6. Big Data ed etica

I Big Data ci fanno trovare prodotti, auto o mezzi di trasporto per spostarci (magari scegliendo il percorso meno trafficato); tante attività, oggi, senza Big Data non sarebbero così immediate.
Ma il problema dei dati che lasciamo in rete è legato al fatto che qualcuno potrebbe farne un uso illecito o non trasparente. Per evitare che i Big Data possano diventare pericolosi è necessario fissare e rispettare regole certe.

Un esempio di problema, di natura politica, è quello di garantire che i sondaggi non siano usati per pilotare onde emotive.
Le più recenti campagne elettorali hanno dimostrato che, con la rivoluzione digitale, le elezioni si vincono anche *misurando tutto*. Per esempio, negli Stati Uniti, incrociando i dati degli elettori registrati e quelli raccolti in rete o sui social network è possibile condurre campagne elettorali mirate per far arrivare ai singoli individui messaggi sulla propensione di un candidato a prendersi cura proprio dei loro problemi (previdenziali, sanitari, lavorativi ecc.). Per far ciò gli staff elettorali dei vari candidati usano i Big Data per eseguire molte analisi al giorno con diverse decine di migliaia di simulazioni al computer. Questa nuova tecnica, interconnessa e su scala più ampia, si è quasi sempre dimostrata vincente rispetto a quella classica dei sondaggi e della statistica.

I dati personali devono essere trattati in termini di liceità, correttezza, trasparenza. Gli interessi di chi gestisce i Big Data devono sempre bilanciare quelli dei fornitori di dati rispettando la limitazione della finalità, la minimizzazione dei dati e la loro esattezza. In Europa, trattandosi di dati personali, devono essere gestiti in conformità con quanto previsto all'articolo 6 del GDPR – Regolamento Europeo Privacy.

FISSA I CONCETTI IMPORTANTI

1 **I Big Data consentono di**

A archiviare dati che non riguardano aspetti sociali
B raccogliere quantità enormi di dati strutturati e non strutturati derivanti da più fonti
C raccogliere quantità enormi di dati strutturati derivanti da più fonti
D raccogliere quantità consistenti di dati limitatamente ad una singola azienda

2 **Quale dei seguenti è un elemento fondamentale della tecnologia Big Data?**

A Database tradizionali
B Rete strettamente gerarchizzata con client e server ben definiti
C Big Data Analytics
D Trasparenza nell'acquisizione dei dati

3 **I Big Data sono una tecnologia**

A di raccolta dati futuribile e ancora a livello progettuale
B di raccolta dati consolidata
C di raccolta e analisi continua di dati in costante sviluppo
D di rete

4 **Dati e informazioni**

A sono la stessa cosa
B i primi derivano dalle seconde
C le seconde derivano dai primi
D non hanno nulla a che vedere gli uni con le altre

5 **La Business Intelligence**

A utilizza tecnologie di data warehousing
B utilizza tecniche di Intelligenza Artificiale
C non utilizza tecniche di data mining
D non è applicabile a grandi raccolte di dati

6 **Il Machine Learning**

A utilizza tecnologie di data warehousing
B utilizza tecniche di Intelligenza Artificiale
C non è applicabile ai Big Data
D non è in grado di supportare analisi dati di tipo predittivo

7 **Le analisi di tipo prescrittivo sui Big Data**

A sono possibili con la Business Intelligence
B sono possibili con il Machine Learning
C sono possibili sia con la Business Intelligence che con il Machine Learning
D non esistono

8 **Big Data e Cloud**

A sono la stessa cosa.
B il Big Data può contenere il Cloud.
C il Cloud può contenere i Big Data.
D non sono tra loro compatibili.

9 **Una smart-city**

A è una città con un reddito pro capite alto.
B è una città con un alto livello di istruzione dei propri abitanti
C è una città che usa soluzioni digitali a beneficio dei suoi abitanti e imprese
D è banalmente un'idea utopica.

10 **Quale dei seguenti aspetti riveste un ruolo critico nell'utilizzo dei Big Data?**

A L'integrazione della tecnologia IoT
B La protezione della sfera personale (privacy)
C La raccolta dei dati
D L'analisi dei dati.

ATTIVITÀ

Ricerca di Open Data su Internet. Ricercare siti che mettano a disposizione banche dati pubbliche, ed effettuare su una di queste delle analisi di dati.

Descrizione dello scenario

- Diversi siti pubblici su Internet offrono l'accesso a sistemi Open Data.
- È possibile eseguire analisi dei dati contenuti in essi utilizzando le dashboards predefiniti da chi offre il servizio.
- È spesso possibile effettuare il download di insiemi di dati per condurre su essi delle analisi ad hoc.

Proposta di lavoro

- Valutare i pregi e i difetti di alcune di queste offerte, anche nell'ottica della salvaguardia della liceità e trasparenza dei dati accessibili.
- Verificare e valutare la qualità e l'offerta dei formati dati scaricabili.
- Effettuare il download di un insieme di dati di interesse e utilizzare uno strumento di calcolo personale per condurre un'analisi di tipo OLAP (per esempio, ipercubi con tabelle Pivot di Excel).