

# 30 minuti con... Big Data



Corso di Quality Outsourcing Management

**Roberto Nai** (Dipartimento di Informatica – UNITO)



# Agenda

- L'era dei Big Data
- Cosa sono i Big Data
- Le proprietà dei Big Data
- Esempi di applicazioni Big Data
- Archiviazione e analisi dei Big Data
- Conclusioni



Materiale della lezione: <https://github.com/roberto-nai/SUISS>

 GitHub

SCAN ME

# L'era dei Big Data

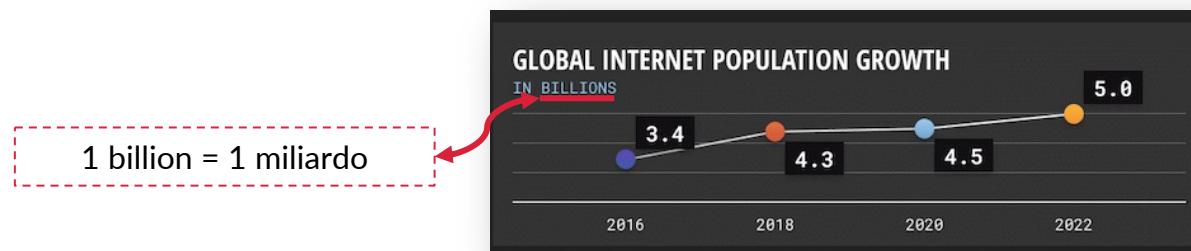
- Nel 2013, il 90% di tutti i dati del mondo era stato generato nei due anni precedenti.
- Nel 2014, l'International Data Corporation (IDC) ha previsto una crescita del volume globale di dati digitali da 4,4 ZB nel 2013 a 44 ZB entro il 2020.
- Nel 2018, IDC ha rivisto le sue previsioni: da 33 ZB nel 2018 a 175 ZB entro il 2025.
  - Per avere un'idea, si pensi che un personal computer o uno smartphone moderni hanno la capacità di memorizzare 1 TB; il volume di dati sarebbe quindi equiparabile al contenuto di 175 miliardi di personal computer o smartphone.



1 ZB (ZettaByte) =  $10^9$  TB (1 miliardo di TeraByte)

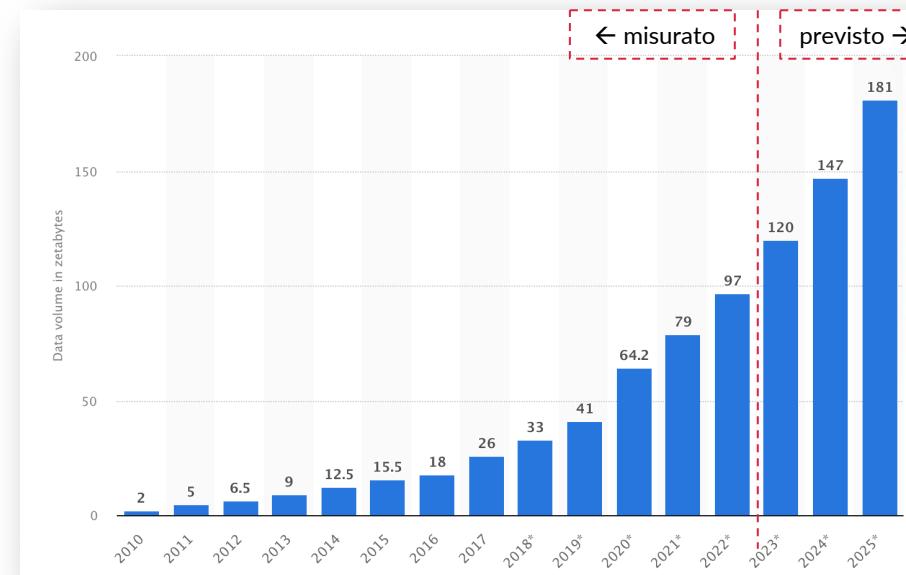
# L'era dei Big Data

- A novembre 2022 la popolazione mondiale era di circa 8 miliardi di persone;
  - fonte: Nazioni Unite, Dipartimento degli Affari Economici e Sociali, Divisione Popolazione.
- Ad aprile 2022 la "popolazione Internet" era circa 5 miliardi di persone ovvero circa il 63% della popolazione mondiale;
  - fonte: DOMO Inc, "Data never sleeps", 2022.

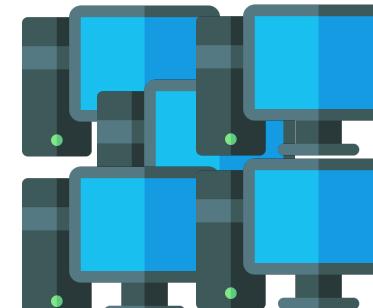
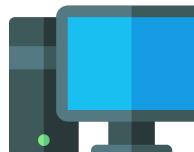


# L'era dei Big Data

- Il sito web Statista ha stimato che nel 2022 la quantità totale di dati *consumati* a livello globale è stata di 97 ZB.
  - Circa 19,7 TB a persona (per chat, e-mail, social, streaming, ecc.).

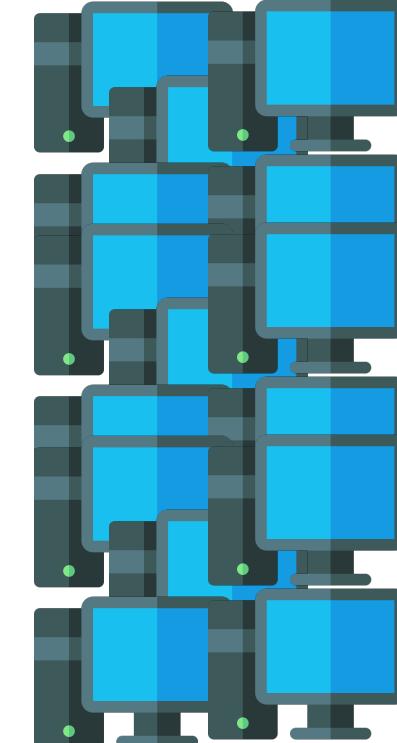


# L'era dei Big Data



Un personal computer (PC)  
moderno ha la capacità di  
1 TB

Un utente medio  
ogni anno *consuma* dati  
per l'equivalente di 19 PC  
(19 TB)



A livello mondiale,  
nel 2022 sono stati *consumati* dati  
per l'equivalente di 97 miliardi di PC  
(97 ZB)

# L'era dei Big Data

- Da dove proviene questa marea di dati?
  - Ricerche sul Web (Google, Bing, ecc.).
  - Social (Facebook, Instagram, Twitter, ecc.).
  - App di messaggistica istantanea (Whatsapp, Telegram, ecc.).
  - eCommerce (Amazon, Alibaba, eBay, ecc.).
  - Streaming (YouTube, Spotify, Netflix, ecc.).
  - Esperimenti scientifici su larga scala (Large Hadron Collider o LHC del CERN, ecc.).
  - Dispositivi dell'Internet delle cose (IoT) (sanità elettronica, casa intelligente, città intelligenti, ecc.).

# L'era dei Big Data

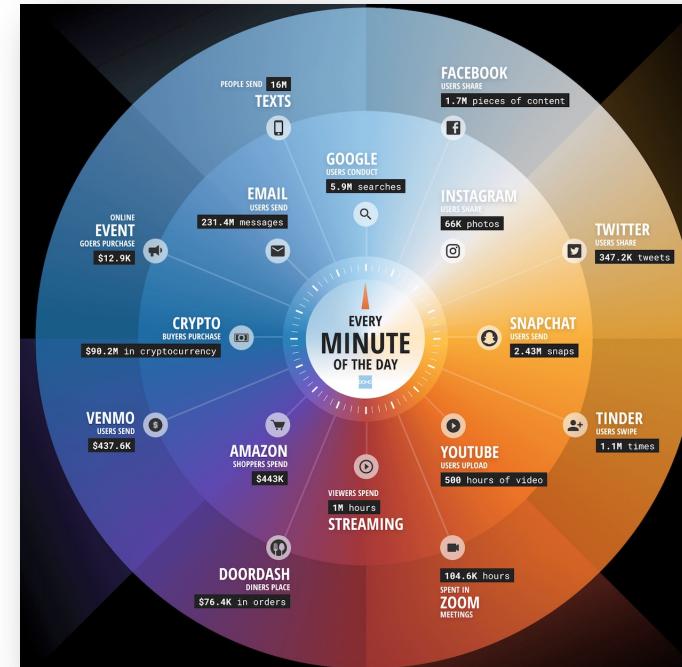
- Alcune statistiche:
  - la Borsa di New York (New York Stock Exchange) genera circa 4-5 TB di dati al giorno;
  - Facebook ospita più di 240 miliardi di foto, con una crescita di 7 PB al mese;
  - nel 2014, l'Internet Archive ha immagazzinato più di 18 PB di dati;
  - l'LHC del CERN produce circa 30-50 PB di dati all'anno;
  - CISCO ha stimato che 50 miliardi di dispositivi IoT erano connessi a Internet nel 2020
    - nel 2020 l'IoT ha creato 40 TB di dati al giorno (stimato).



1 PB (PetaByte) =  $10^3$  TB (1000 TeraByte)

# L'era dei Big Data

- Dati generati in un minuto secondo l'articolo "Data never sleeps" del 2022.



# L'era dei Big Data

Attuale ordine  
di grandezza per i volumi  
di dati in Internet



Prefissi del Sistema internazionale di unità di misura

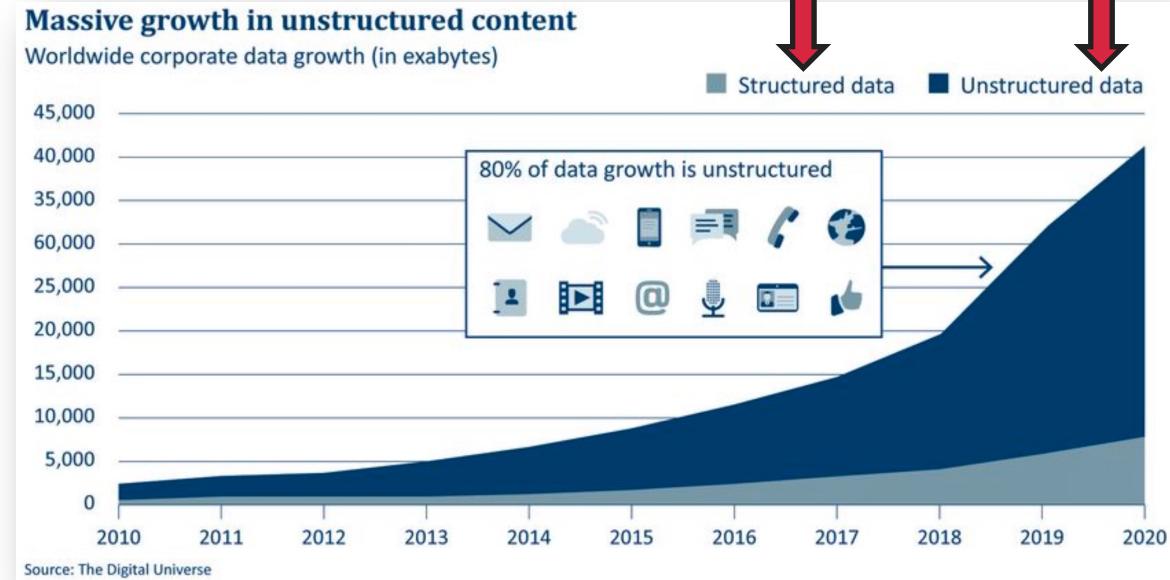
Prefisso	Simbolo	Notazione scientifica	Numero decimale	Scala lunga [note 1]	Scala corta [note 2]	Adozione [note 3]
quetta	Q	$10^{30}$	1 000 000 000 000 000 000 000 000 000 000	Quintilione	<i>Nonillion</i>	2022 <sup>[1]</sup>
ronna	R	$10^{27}$	1 000 000 000 000 000 000 000 000 000 000	Quadriliardo	<i>Octillion</i>	2022 <sup>[1]</sup>
yotta	Y	$10^{24}$	1 000 000 000 000 000 000 000 000 000 000	Quadrilione	<i>Septillion</i>	1991 <sup>[2]</sup>
zetta	Z	$10^{21}$	1 000 000 000 000 000 000 000 000 000 000	Triliardo	<i>Sextillion</i>	1991 <sup>[2]</sup>
exa	E	$10^{18}$	1 000 000 000 000 000 000 000	Trilione	<i>Quintillion</i>	1975 <sup>[3]</sup>
peta	P	$10^{15}$	1 000 000 000 000 000	Biliardo	<i>Quadrillion</i>	1975 <sup>[3]</sup>
tera	T	$10^{12}$	1 000 000 000 000	Bilione	<i>Trillion</i>	1960 <sup>[4]</sup>
giga	G	$10^9$	1 000 000 000	Miliardo	<i>Billion</i>	1960 <sup>[4]</sup>
mega	M	$10^6$	1 000 000	Milione	<i>Million</i>	1960 <sup>[4]</sup>
chilo	k	$10^3$	1 000	Mille	<i>Thousand</i>	1795
etto	h	$10^2$	100	Cento	<i>Hundred</i>	1795
deca	da	$10^1$	10	Dieci	<i>Ten</i>	1795
–		$10^0$	1	Unità	<i>One</i>	–

# L'era dei Big Data

- Che tipo di dati vengono generati?

CSV, JSON, XML, ecc.

Audio, Video, Testi, ecc.



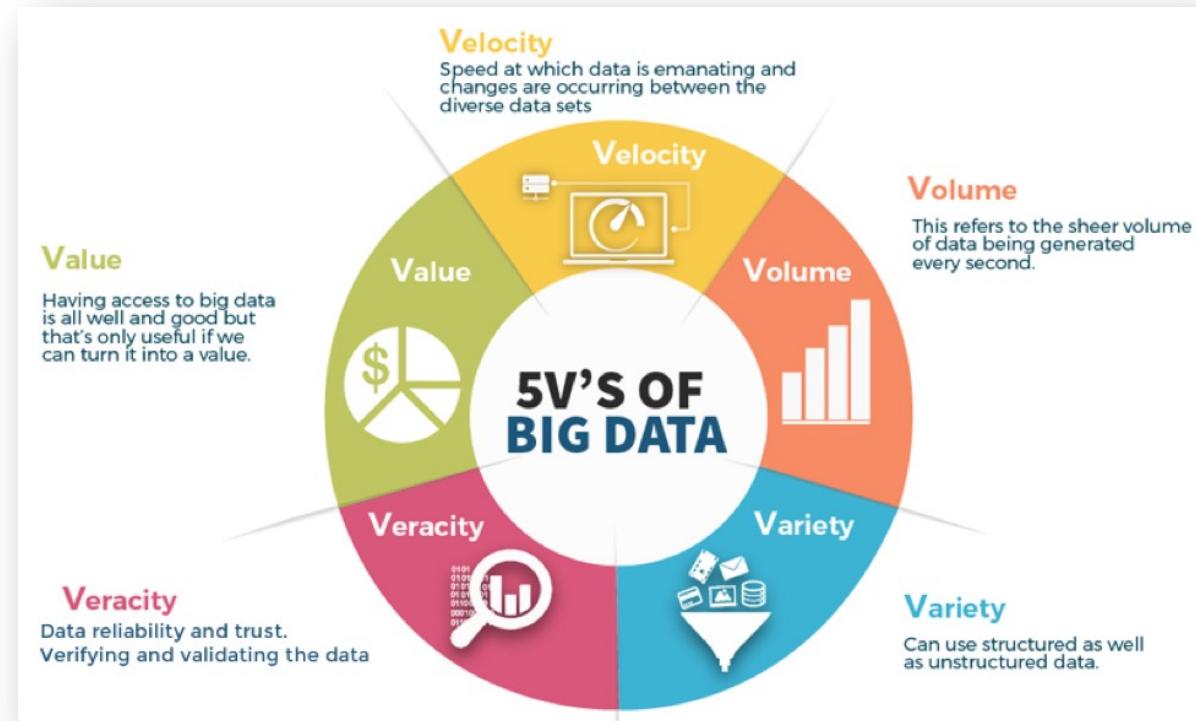
# Cosa sono i Big Data?

- Intuitivamente, i Big Data si riferiscono ad un insieme di dati così grandi e complessi che è complesso memorizzarli ed elaborarli su un singolo computer con strumenti, algoritmi e tecniche tradizionali.
- I Big Data sono difficili da acquisire, archiviare, copiare, cercare, condividere, analizzare e visualizzare.
  - I dati potrebbero dover essere acquisiti da fonti multiple e diverse.
  - Il volume dei dati potrebbe essere così grande da non poter essere contenuto in un singolo computer.
  - Il volume dei dati potrebbe essere così grande e la varietà dei dati così alta che la loro visualizzazione diventerebbe problematica.

# Quali sono le caratteristiche dei Big Data?

- Le 5 V che rappresentano le caratteristiche dei Big Data:
  - **Volume:** la quantità di dati generati è enorme e in continua crescita;
  - **Varietà:** i dati possono provenire da fonti diverse (es.: sensori, smartphone, social network, ecc.) e possono variare notevolmente nel tipo (ad esempio, testo, immagini, audio, video, ecc.) e nel formato (CSV, JSON, XML, TXT, MP3, MP4, ecc.);
  - **Velocità:** i dati possono arrivare a velocità diverse e si possono accumulare enormi quantità di dati in tempi molto brevi.
  - **Veridicità:** si riferisce alla qualità e all'attendibilità dei dati;
    - qualità perché i dati possono essere *rumorosi* e *incerti*, per cui le cui anomalie in essi contenute possono portare a conclusioni fuorvianti;
    - attendibilità perché i dati devono provenire da fonti affidabili.

# Quali sono le caratteristiche dei Big Data?



# Quali sono le caratteristiche dei Big Data?

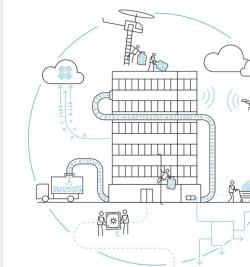
- Valore: i dati contengono *valore* e *conoscenza*; l'obiettivo finale dell'analisi dei Big Data è ottenere valore dai dati analizzati, cioè estrarre il maggior numero possibile di informazioni utili che possano fornire un beneficio misurabile.

**Neelie Kroes**

Vice-President of the European Commission responsible for the Digital Agenda

**Data is the new gold**

**Deloitte.**



**Data is the new gold.**  
The future of real estate service providers

# Esempi di applicazioni Big Data

- Esempi di applicazioni Big Data:
  - tracciamento delle epidemie in tempo reale;
  - analisi dei clienti: per aumentare la fidelizzazione dei clienti;
  - manutenzione predittiva: per rilevare le anomalie e ridurre i costi di manutenzione;
  - pubblicità personalizzata;
  - scoperte astronomiche;
  - previsione in tempo reale del mercato azionario;
  - gestione del traffico urbano;
  - web analytics: per raccogliere dati sulle modalità di accesso al Web.
  - ecc.

# Archiviazione e analisi dei Big Data

- Il processo di analisi dei Big Data



# Archiviazione e analisi dei Big Data



- Dopo aver definito la **sorgente** (ossia quali dati si vogliono reperire), si procede con la **raccolta** dei dati.
- L'obiettivo della raccolta è ottenere dati *grezzi* (*raw*) da fonti multiple e potenzialmente diverse.
  - Es.: dati in formato CSV, JSON, XML, ecc. ma anche MP3, MP4 poi trasformati in testo e/o insiemi di dati numerici.

# Archiviazione e analisi dei Big Data



- Nella fase di **preparazione**, l'obiettivo è *trasformare* e *filtrare* i dati grezzi in un formato utilizzabile dai framework di elaborazione dei dati.
  - Trasformazione: trasformare i dati dalla loro forma originale (CSV, JSON, XML, ecc.) a un'altra più adatta all'analisi.
  - Filtraggio: rimuovere o correggere dati *rumorosi*, cioè con errori, dati duplicati, valori mancanti, ecc.

# Archiviazione e analisi dei Big Data



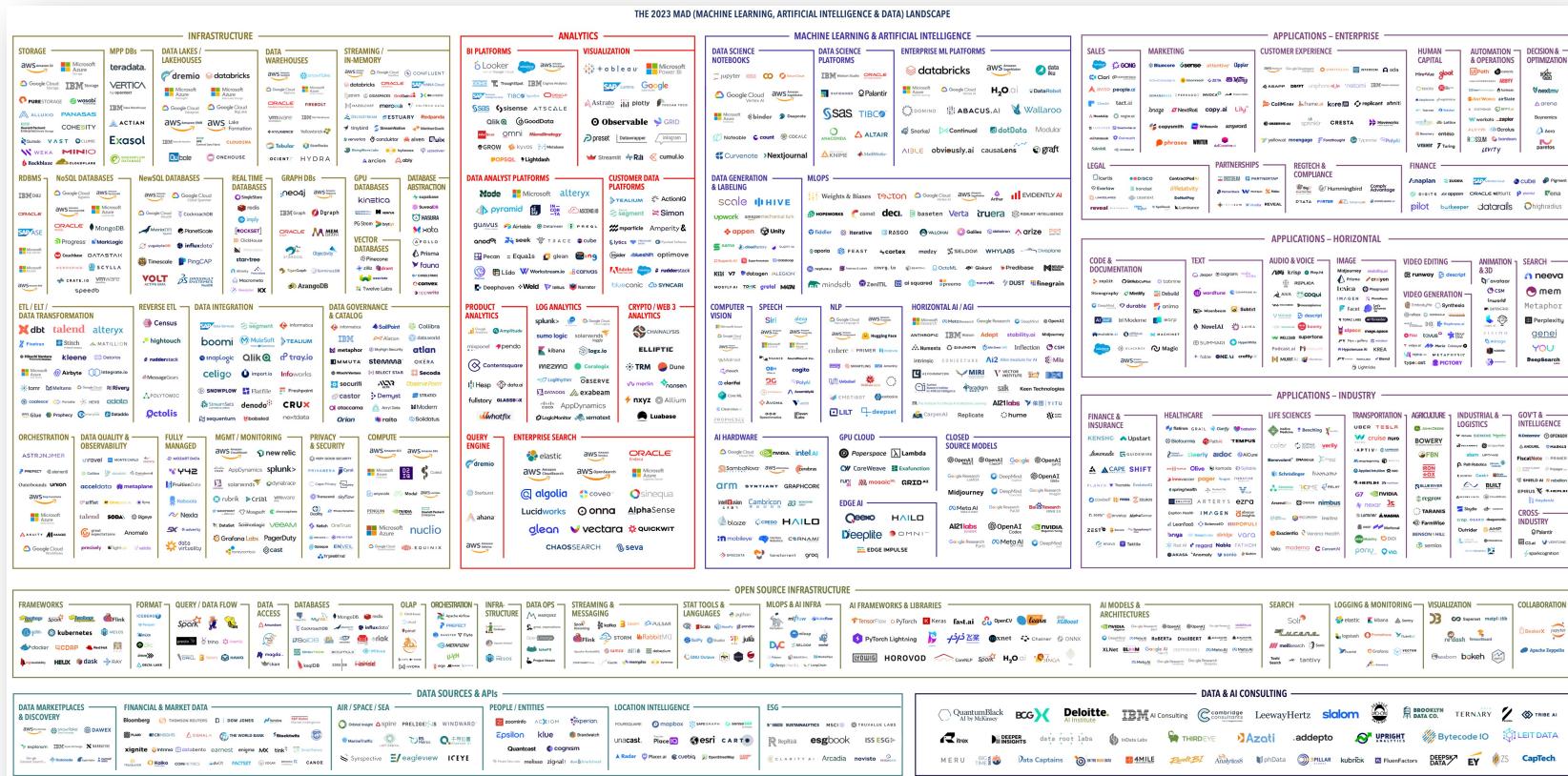
- Nella fase di **analisi**, confermare o trovare nuova conoscenza.
  - Analisi **descrittiva** risponde alla domanda: "Cosa è successo?".
  - Analisi **diagnostica** risponde alla domanda: "Perché è successo qualcosa?".
  - Analisi **predittiva** risponde alla domanda: "Cosa è probabile che accada in futuro?".
  - Analisi **prescrittiva** risponde alla domanda: "Quale azione è necessario intraprendere per evitare un determinato evento?".

# Archiviazione e analisi dei Big Data



- Nella fase di uso, l'obiettivo è *interpretare* e *utilizzare* i risultati dell'analisi svolta precedentemente per estrarre informazioni ad alto valore aggiunto e riutilizzabili.
- A tale scopo, si adottano strumenti di analisi visiva e *business intelligence*.
  - Reportistica e visualizzazione interattiva dei dati.

# Conclusioni



# Bibliografia

- International Data Corporation (IDC)
  - <https://www.idc.com>
- Nazioni Unite, Dipartimento degli Affari Economici e Sociali, Divisione Popolazione
  - <https://population.un.org/wpp>
- Data never sleeps, rapporto 2022
  - <https://www.domo.com/data-never-sleeps>
- Statista
  - <https://www.statista.com>
- The Big Data challenge at the Large Hadron Collider
  - <https://www.innovationnewsnetwork.com/big-data-challenge-large-hadron-collider/11359>

# Bibliografia

- General Conference on Weights and Measures
  - <https://www.bipm.org/en/cgpm-2022>
- Machine Learning, Artificial Intelligence and Data (MAD) Landscape
  - <https://mad.firstmark.com>

# Fine presentazione

Grazie per l'attenzione

 roberto.nai@unito.it

LinkedIn



SCAN ME