Università di Genova | MaLGa

# Introduzione alla Data Science

Nicoletta Noceti

UniGe | MaLGa

# Informazioni organizzative sul corso

# Chi siamo

Nicoletta Noceti
nicoletta.noceti@unige.it

Vito Paolo Pastore
Vito.paolo.pastore@edu.unige.it

Se volete più informazioni potete visitare **https://ml.unige.it**

# Regole del gioco

- ~24 ore di lezione
  - 4 ore a settimana (martedi 11-13, mercoledi 9-11)
  - Di solito 2 ore di “teoria” + 2 ore di laboratorio in Python (escluso la prima settimana)
- Progetto finale
  - Dettagli sui possibili progetti nelle ultime due ore di lezione
  - Incontri regolari organizzati durante le 4 ore del corso per monitorare il progresso
- Esame finale con due possibili modalità
  - Discussione del progetto a fine corso + versione ridotta dello scritto
  - Consegna del progetto una settimana prima dell’esame + versione completa dello scritto

UniGe | MaLGa

# Partiamo dalle definizioni

# Partiamo da qualcosa che sapete già: un sistema informativo

## Compiti

- Raccogliere i dati
- Conservare i dati raccolti, archiviandoli
- Elaborare i dati, trasformandoli in informazioni
- Distribuire l'informazione agli utilizzatori

## Componenti

- Strumenti
- Procedure
- Strutture

***Definizione indipendente dal grado di automazione***

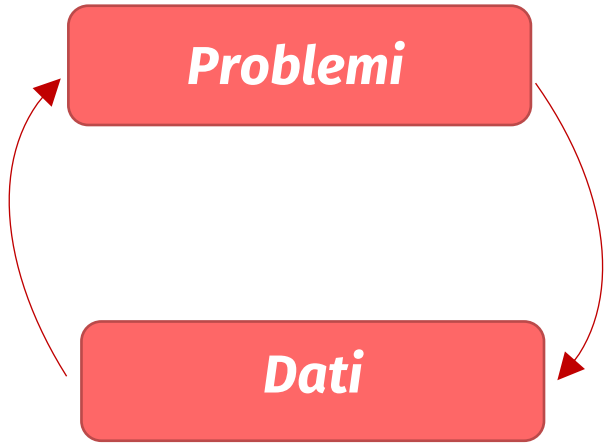
# Cosa sono le informazioni?

- Tutto ciò che **produce variazioni nel patrimonio cognitivo di un soggetto**, ossia chi percepisce l'informazione
- L'informazione deve essere **utile** per (=**comprensibile** da) il percettore
- Un sistema informativo deve fornire una chiave di lettura mediante cui **interpretare** l'informazione che gestisce

# Cos’è la data science?


https://www.youtube.com/watch?v=X3paOmcrTjQ

# Cos'è la Data Science?




## Un paio di definizioni

*Un campo multi-disciplinare che usa metodo scientifico, processi algoritmi e sistemi per estrarre conoscenza da dati strutturati e non strutturati*

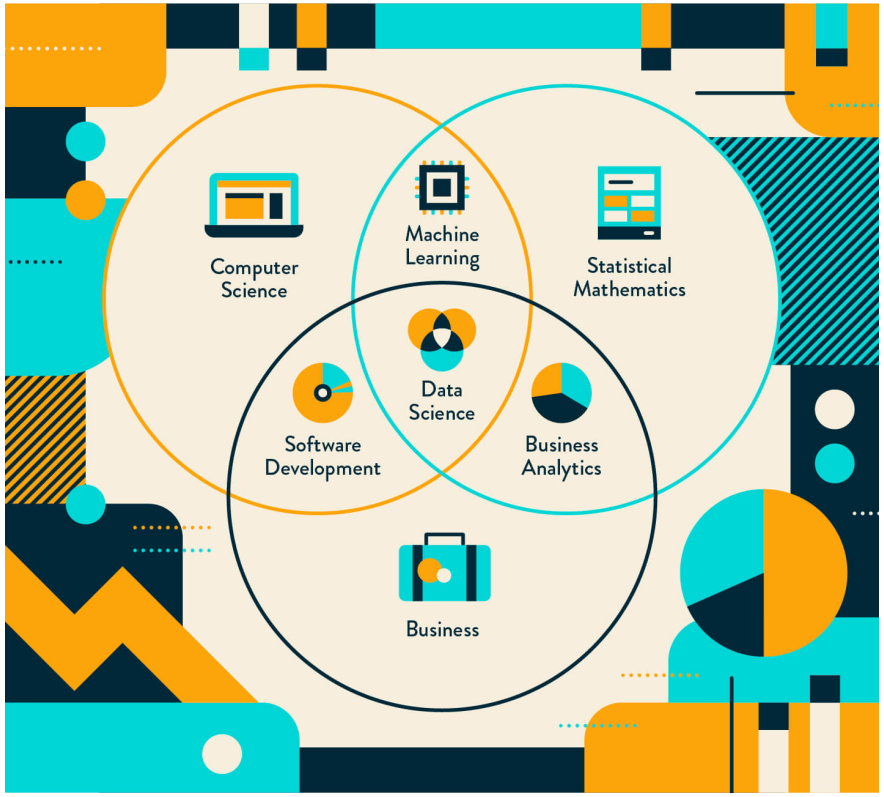
... oppure...

*La Data Science è estrazione di conoscenza dai dati attraverso un processo di scoperta, formulazione di ipotesi e verifica delle ipotesi*

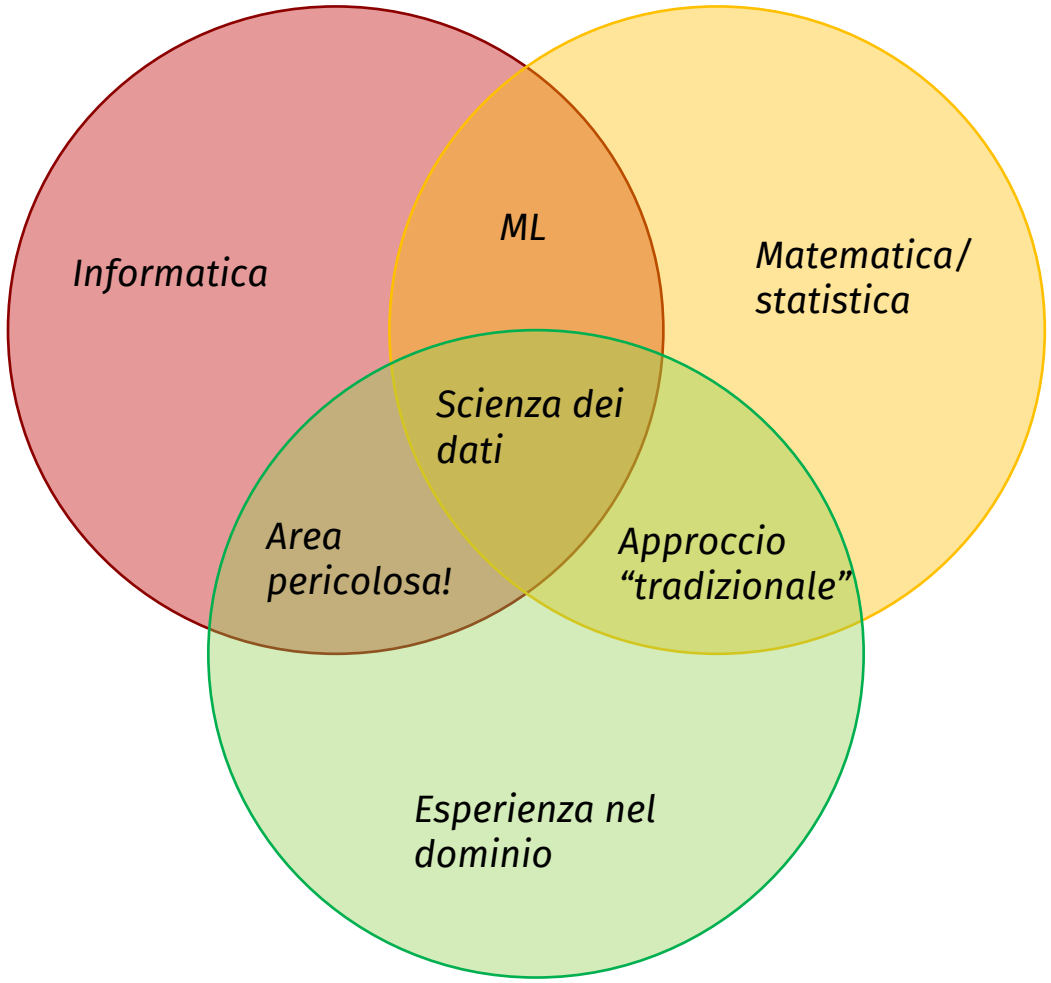
... ma anche ...

*La Data Science è arte e scienza che consiste nel trarre conoscenza dai dati*

# Cos'è la Data Science?



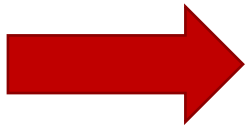

*Source: https://morioh.com/p/ef484a5ec282*



# I molti ruoli del Data Scientist

Raccogliere
Organizzare
Pulire

dati

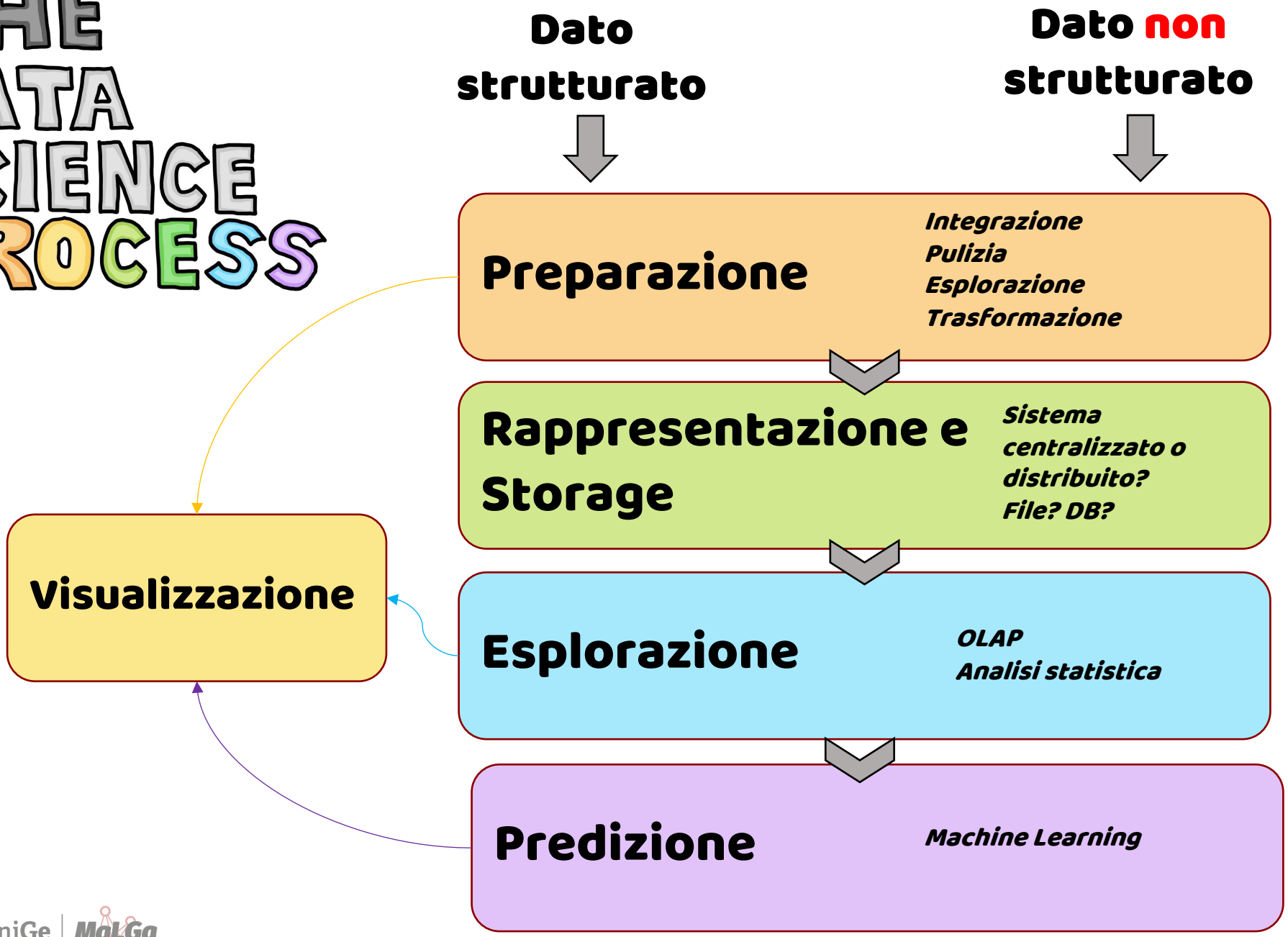


Visualizzare (comunicare)

Analizzare

Derivare modelli (Matemica, Machine Learning)

# THE DATA SCIENCE PROCESS



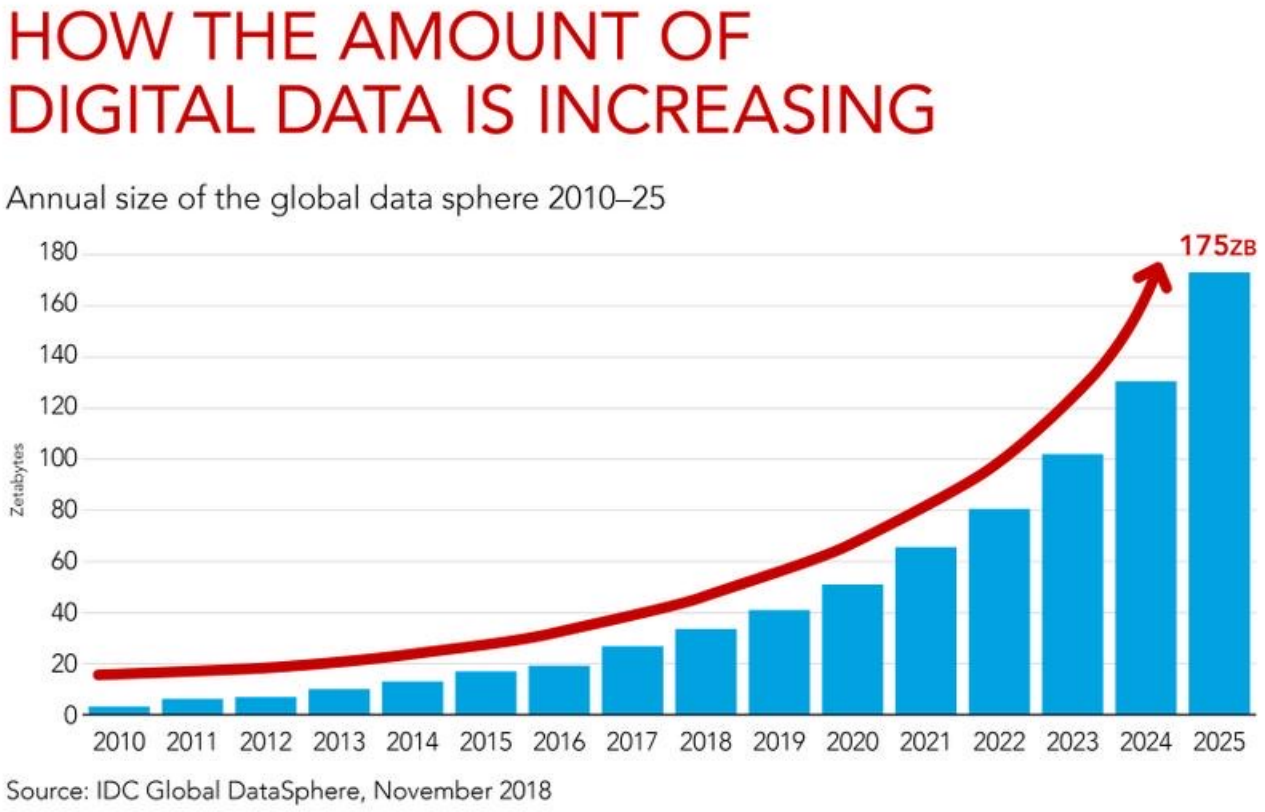
# Data deluge e Big Data


I Big data sono enormi dataset – caratterizzati dalle 4 V, volume-varietà-velocità-variabilità – che richiedono architetture scalabili per storage, manipolazioni ed analisi efficienti

# Produzione di dati digitali del mondo

*$10^{21}$ bytes*

*(as a reference a Terabyte is $10^{12}$ bytes)*



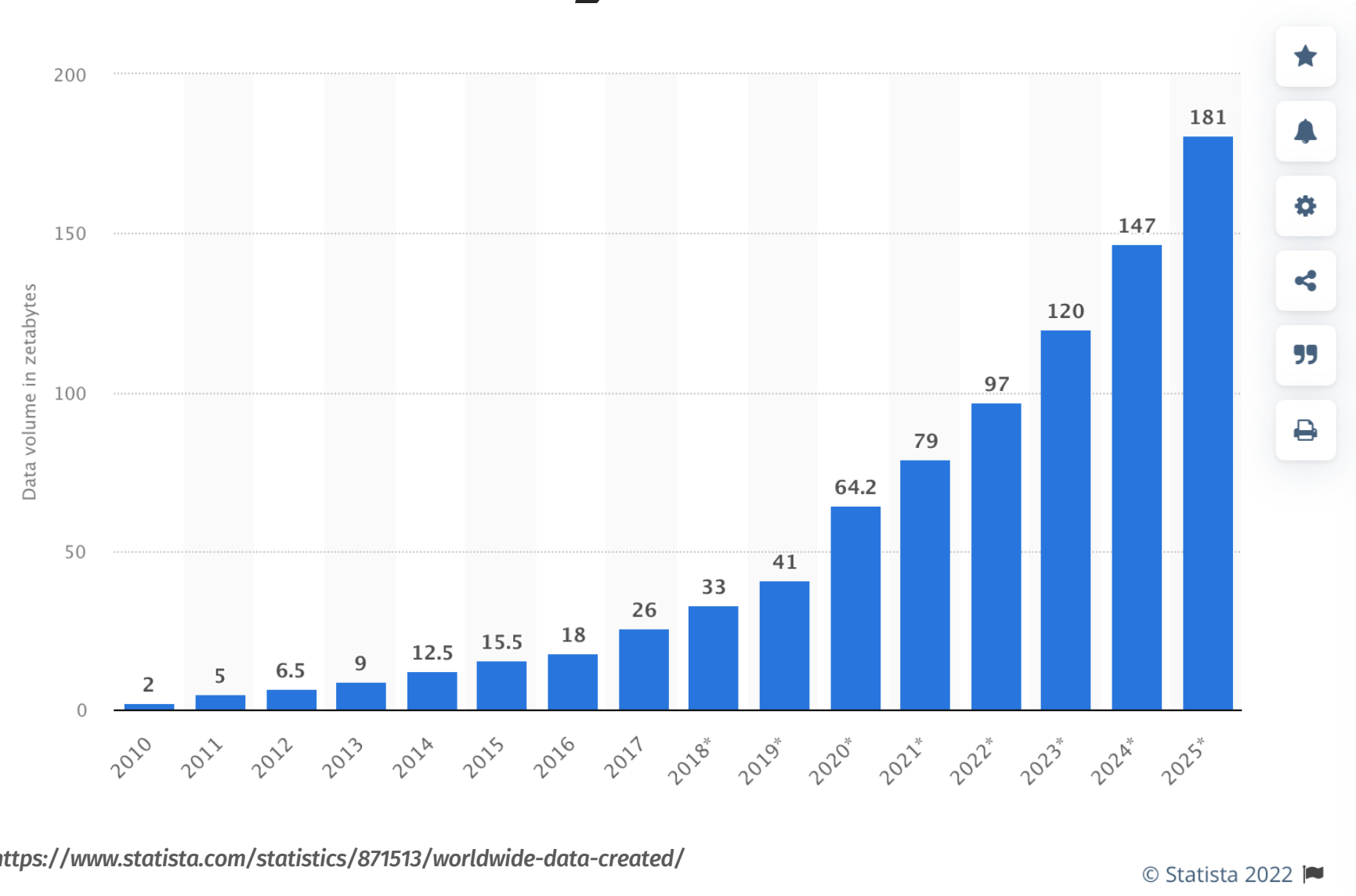

It is estimated that around

**90%**

of all the digital data in the world has been created since 2016

Think Ahead

# Produzione di dati digitali del mondo




*https://www.statista.com/statistics/871513/worldwide-data-created/*

UniGe | MaLGa

# Dati

# La necessità di fare ordine

- Troppi dati!! Un umano non sarebbe in grado di analizzarli in modo approfondito e affidabile
- Tante forme di dato ottenuti da diverse fonti
- I dati possono essere mancanti, incompleti, sbagliati...
- I dati possono avere scale di misurazione diverse, bisogna renderli confrontabili

# Una classificazione

- Dati strutturati o non strutturati (organizzati o non organizzati)
- Dati qualitativi o quantitativi
- I 4 livelli dei dati

# Dati strutturati e non strutturati

- **Dati strutturati (organizzati): sono dati tratti da osservazioni di caratteristiche, normalmente organizzati in formato tabulare (righe e colonne)**
  - Esempio: osservazioni scientifiche registrate da ricercatori che vengono conservate in modo molto ordinato
- **Dati non strutturati (non organizzati): dati che esistono come entità libere e che non seguono alcuna organizzazione standard o gerarchia**
  - Esempi: dati che hanno una natura testuale (es. Fie log dei server, post di Twitter); sequenze genetiche (es. ACGTATTGCA)

# Dati strutturati e non strutturati

- I dati strutturati sono considerati più facili da elaborare e analizzare
- Circa il 90% dei dati in circolazione è NON strutturato
- Servono tecniche di pre-analisi e pre-elaborazione (preprocessing) per dare una struttura ai dati non strutturati

# Esempio: rappresentazione di un Tweet

*This Wednesday morn, are you early to rise? Then look East. The Crescent Moon joins Venus & Saturn. Afloat in the dawn skies.*

Abbiamo bisogno di ottenere descrizioni numeriche del testo... partiamo con il conteggio delle parole

| | |
|---|---|
| This | 1 |
| Wednesday | 1 |
| Morn | 1 |
| ... | ... |

# Esempio

- Possiamo anche contare la presenza di caratteri speciali:

  ? 1

  ! 0

  & 1

- Adesso consideriamo la lunghezza del testo e la relazione con la lunghezza media dei Tweet
  - Il tweet è lungo 121 caratteri, la lunghezza media dei Tweet è di 30 caratteri

    121/30 = 4.03 → Il tweet è lungo 4.03 volte più della media, e questa può diventare una ulteriore informazione nella rappresentazione

- Altre informazioni da aggiungere alla descrizione?? Può trattarsi di informazioni derivate in modo diretto o indiretto dal testo (es. L'argomento)

# Dati quantitativi e qualitativi

- Dati quantitativi: dati che possono essere descritti tramite numeri e su cui è possibile /ha senso eseguire operazioni matematiche
- Dati qualitativi: tutto il resto, di solito è possible descriverli usando linguaggio naturale

ESEMPIO: supponiamo di dover elaborare le osservazioni effettuate nelle caffeteria di una grande città...

# Dati quantitativi e qualitativi

Caratteristiche

- Nome della caffetteria
- Fatturato
- CAP
- Numero medio di clienti mensile
- Origine del caffè

# Dati quantitativi e qualitativi

- **Nome della caffetteria**
  - E’ esprimibile come numero? NO → QUALITATIVO
- **Fatturato**
  - E’ esprimibile come numero? SI
  - Ha senso eseguire operazioni su di esso? SI → QUANTITATIVO
- **CAP**
  - E’ esprimibile come numero? SI
  - Ha senso eseguire operazioni su di esso? NO→ QUALITATIVO
- **Numero medio di clienti mensile**
  - E’ esprimibile come numero? SI
  - Ha senso eseguire operazioni su di esso? SI → QUANTITATIVO
- **Origine del caffè**
  - E’ esprimibile come numero? NO → QUALITATIVO

# Domande che ha senso porsi su dati quantitativi (esempi)

*Ci porremo queste ed altre domande quando parleremo di* ***esplorazione dei dati***

- Qual è il valore medio?
- Se il tempo è un fattore, questa quantità cresce o decresce con il trascorrere del tempo?
- Esiste una soglia oltre la quale il valore potrebbe diventare critico?

# Domande che ha senso porsi solo su dati qualitativi (esempi)

*Ci porremo queste ed altre domande quando parleremo di* ***esplorazione dei dati***

- Quale valore è più (o meno) frequente?
- Quanti valori univoci esistono?
- Quali sono i valori univoci?

# Ancora a proposito di dati quantitativi

- **Dati discreti: possono essere contati e possono assumere solo determinati valori**
  - Esempio: numero di clienti di un caffè (non si possono avere frazioni di clienti)
- **Dati continui: devono essere misurati e possono assumere una gamma infinita di valori**
  - Esempi: peso o statura di una persona,  tempo, temperatura

# I 4 livelli dei dati

Una specifica caratteristica (colonna) dei dati strutturati può essere suddivisa in 4 livelli

- Nominale
- Ordinale
- Degli intervalli
- Dei rapporti

# I 4 livelli dei dati

I livelli dei dati sono scale che ci consentono di misurare e classificare I dati raccolti in variabili ben definite che possano essere usate per scopi diversi di analisi e comprensione

Perchè sono importanti? Ci guidano nell'analisi (cosa ha senso fare?)

# I 4 livelli dei dati

- **Nominale**: Utilizzato per classificare i dati in categorie o gruppi mutuamente esclusivi
- **Ordinale**: Utilizzato per misurare le variabili in un ordine naturale (es. una valutazione)
- **Degli Intervalli**: Utilizzato per misurare variabili con intervalli (es. la temperatura e il tempo)
- **Dei Rapporti**: Consente confronti e calcoli come rapporti, percentuali e medie

# Livello nominale

*Utilizzato per classificare i dati in categorie o gruppi mutuamente esclusivi*

E’ costituito da dati descritti unicamente per nome o categoria (talvolta da numeri) tipicamente qualitativi

Esempi:

- Nazionalità
- Classe di mammiferi
- Città di nascita
- …

# Livello nominale

Quali operazioni possiamo applicare?

- Uguaglianza. Esempio: essere un imprenditore nel campo delle tecnologie equivale ad essere nel settore tecnologico
- Appartenenza ad un insieme. Esempio: un quadrato è un rettangolo
- Calcolo della moda, una misurazione del «centro» dei dati (a cosa tendono i dati?)

# La moda

Vi ricordo che... In statistica, la moda (o norma) di una distribuzione di frequenza X è la modalità (o la classe di modalità) caratterizzata dalla massima frequenza. In altre parole, è il valore che compare più frequentemente.

Esempio: consideriamo una colonna che rappresenta lo stato europeo in cui è presente una caffetteria di una nota catena

{IT, ES, UK, UK, UK, FR, IT, FR, NE, GE, GE, FR, IT, NE, NE, NE, UK, FR, ES, ES, UK, UK, NE, ES}

Calcoliamo le frequenze:

IT 3 ES 4 UK 6 FR 4 NE 5 GE 2

# La moda

Vi ricordo che... In statistica, la moda (o norma) di una distribuzione di frequenza X è la modalità (o la classe di modalità) caratterizzata dalla massima frequenza. In altre parole, è il valore che compare più frequentemente.

Esempio: consideriamo una colonna che rappresenta lo stato europeo in cui è presente una caffetteria di una nota catena

{IT, ES, UK, UK, UK, FR, IT, FR, NE, GE, GE, FR, IT, NE, NE, NE, UK, FR, ES, ES, UK, UK, NE, ES}

Calcoliamo le frequenze:

IT 3 ES 4 UK 6 FR 4 NE 5 GE 2

# Livello ordinale

*Utilizzato per misurare le variabili in un ordine naturale*

Dati su cui è possibile definire strategie per collocare un'osservazione prima di un'altra (ma di solito continua a non essere possibile eseguire operazioni matematiche, es. sommarli o sottrarli)

Esempi:

- indice di gradimento da 1 a 5
- Indice di soddisfazione da 1 a 10
- ...

# Livello ordinale

Attenzione! La distanza tra le misurazioni può non essere sempre la stessa!

Esempio:

- Se la misura è codificata con una lista di numeri, ad es. 1 2 3, sappiamo calcolare la distanza tra 2 valori consecutive, sempre 1 in questo caso
- Se la misura è codificata in classe, come ad es. “molto soddisfatto”, “soddisfatto”, “neutrale”non riusciamo a quantificare la distanza tra di esse in modo

# Livello ordinale

Quali operazioni possiamo applicare?

- Tutte le operazioni del livello nominale
- Ordinamento
- Confronto
- Calcolo del «centro» dei dati con la mediana

# La mediana

Vi ricordo che... data una distribuzione di un carattere quantitativo oppure qualitativo ordinabile, si definisce la mediana come il valore/modalità assunto dalle unità statistiche che si trovano nel mezzo della distribuzione.

Esempio: consideriamo le risposte ad un sondaggio di gradimento del luogo di lavoro in scala da 1 a 5

{5 4 3 4 5 3 2 5 3 2 1 4 5 3 4 4 5 4 2 1 4 5 4 3 2 4 4 5 4 3 2 1}

Riordiamo:

{1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5}

# La mediana

Vi ricordo che... data una distribuzione di un carattere quantitativo oppure qualitativo ordinabile, si definisce la mediana come il valore/modalità assunto dalle unità statistiche che si trovano nel mezzo della distribuzione.

Esempio: consideriamo le risposte ad un sondaggio di gardimento del luogo di lavoro in scala da 1 a 5

{5 4 3 4 5 3 2 5 3 2 1 4 5 3 4 4 5 4 2 1 4 5 4 3 2 4 4 5 4 3 2 1}

Riordiamo:

{1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5}

*La mediana è 4*

# Il livello degli intervalli

*Utilizzato per misurare variabili con intervalli*

Dati esprimibili attraverso metodi quantificabili e su cui è possibile eseguire formule matematiche (anche complesse)

Attenzione! A differenza del livello ordinale, la distanza nella scala della misurazione è la stessa (di solito) e ha un significato (es. la temperatura si misura a passi di 10 gradi)

Esempi:

- temperatura (se in Texas ci sono 37° e ad Istambul ci sono 27° significa che in Texas ci sono 10° in più)
- Guadagno mensile/annuale
- ...

# Il livello degli intervalli

Quali operazioni possiamo applicare?

- Tutte le operazioni del livello ordinale
- Somma
- Sottrazione
- Calcolo del «centro» dei dati con la media

# La media

Vi ricordo che... la media viene calcolata sommando tutti i valori a disposizione e dividendo il risultato per il numero complessivo dei dati

Esempio: osserviamo la temperatura misurata in gradi Fahrenheit

{31 32 32 31 28 29 31 38 32 31 30 29 30 31 26}

Media: 30.37

Mediana: 31.0

Un’altra misura utile è legata a quanto le misure sono variabili

# Deviazione standard

Definizione intuitiva (ma matematicamente non corretta!!!): distanza media di una punto dei dati rispetto alla sua media

Nell'esempio di prima: ~2.52

Perché è importante avere una misura della «dispersione» dei dati al livello di intervallo? Perché siamo tipicamente interessati a capire come i valori si distribuiscano nell'intervallo di interesse e la presenza di eventuali anomalie

# Il livello dei rapporti

*Consente confronti e calcoli come rapporti, percentuali e medie*

Contiene tutti i livelli precedenti, ha senso calcolare anche moltiplicazioni e divisioni; hanno un punto iniziale naturale o uno zero naturale, ma anche una restrizione: i valori dovrebbero essere non negativi

Esempio:

- il denaro depositato in banca si colloca al livello dei rapporti. E' possibile avere «niente denaro sul conto» (zero naturale) ed è sensato dire che 200000 euro sono «il doppio» di 100000 euro
- Altezza e/o peso di una persona
- ...

La misurazione del «centro» può essere fatta con la media

# Una tabella riassuntiva

| | Liv. Nominale | Liv. Ordinale | Liv. Degli intervalli | Liv. Dei rapporti |
|---|---|---|---|---|
| Possiamo definire un ordine | -- | si | si | si |
| Possiamo calcolare la moda | si | si | si | si |
| Possiamo calcolare la mediana | -- | si | Si | Si |
| Possiamo calcolare la media | -- | -- | Si | si |
| Possiamo confrontare variabili | -- | -- | si | si |
| Possiamo calcolare somma e differenza | -- | -- | si | si |
| Possiamo calcolare prodotto e rapporto | -- | -- | -- | si |
| Esiste uno zero assoluto | -- | -- | -- | si |

# Ancora esempi

- **Time of Day:** dawn, morning, noon, afternoon, evening, night
- **Hair Color**: Brown, Black, Blonde, Red, Other
- **Fahrenheit Temperature**.
- **IQ (intelligence scale)**
- **Type of living accommodation**: House, Apartment, Trailer, Other
- **Age**
- The Likert Scale: strongly disagree, disagree, neutral, agree, strongly agree
- **Number of children**

# Ancora esempi

| | Nom. | Ord. | Int. | Rapp. |
|---|---|---|---|---|
| Possiamo definire un ordine | -- | si | si | si |
| Possiamo calcolare la moda | si | si | si | si |
| Possiamo calcolare la mediana | -- | si | Si | Si |
| Possiamo calcolare la media | -- | -- | Si | si |
| Possiamo confrontare variabili | -- | -- | si | si |
| Possiamo calcolare somma e differenza | -- | -- | si | si |
| Possiamo calcolare prodotto e rapporto | -- | -- | -- | si |
| Esiste uno zero assoluto | -- | -- | -- | si |

- **Time of Day:** dawn, morning, noon, afternoon, evening, night
- **Hair Color**: Brown, Black, Blonde, Red, Other
- **Fahrenheit Temperature**.
- **IQ (intelligence scale)**
- **Type of living accommodation**: House, Apartment, Trailer, Other
- **Age**
- The Likert Scale: strongly disagree, disagree, neutral, agree, strongly agree
- **Number of children**

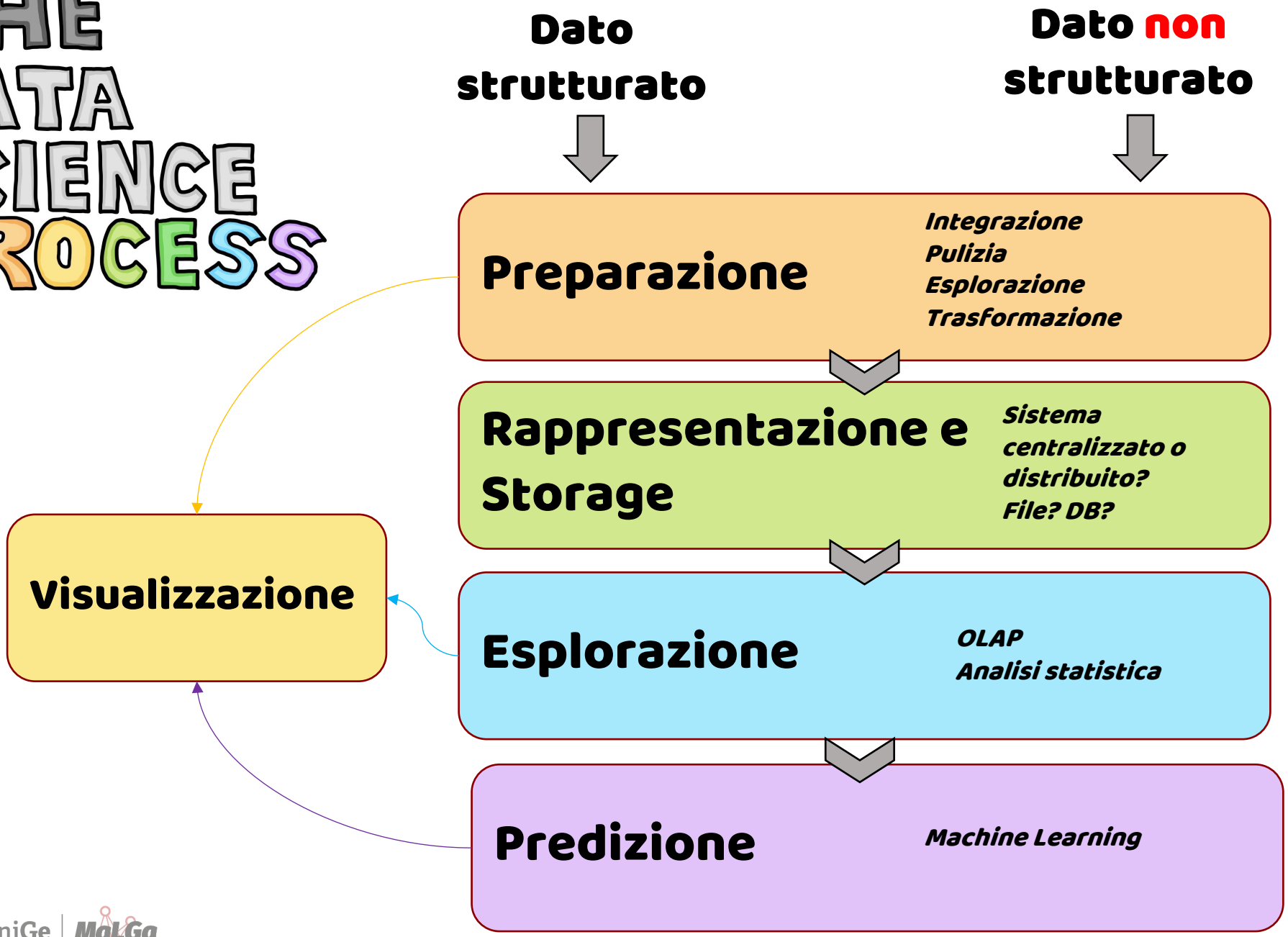
# Ancora esempi

- **Time of Day:** dawn, morning, noon, afternoon, evening, night (ORDINALE)
- **Hair Color**: Brown, Black, Blonde, Red, Other (NOMINALE)
- **Fahrenheit Temperature**. (DEGLI INTERVALLI)
- **IQ (intelligence scale)** (DEGLI INTERVALLI)
- **Type of living accommodation**: House, Apartment, Trailer, Other (NOMINALE)
- **Age** (DEI RAPPORTI)
- The Likert Scale: strongly disagree, disagree, neutral, agree, strongly agree (ORDINALE)
- **Number of children** (DEI RAPPORTI)

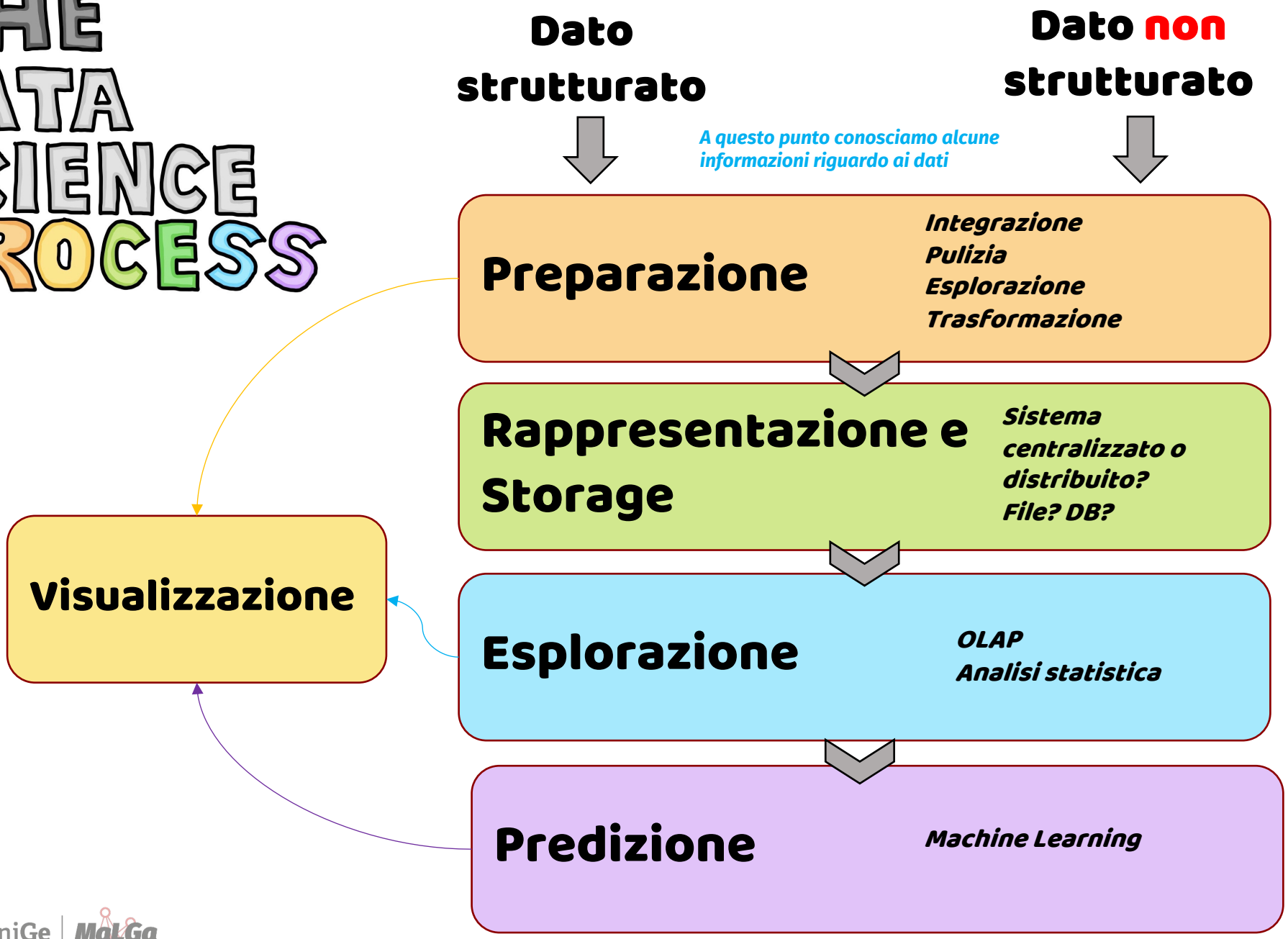
UniGe | MaLGa

# Torniamo ai passi della Data Science

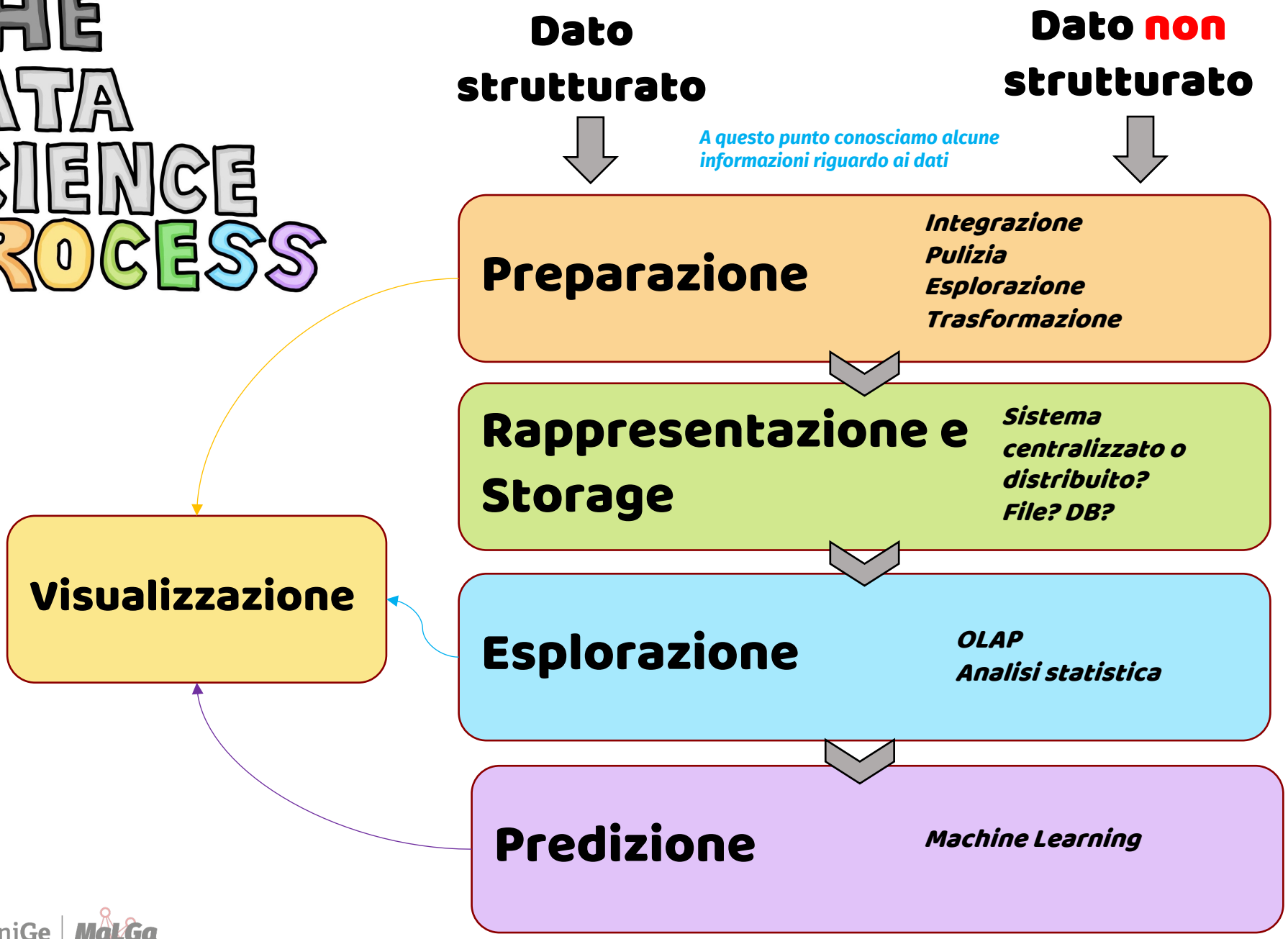
# THE DATA SCIENCE PROCESS



# THE DATA SCIENCE PROCESS



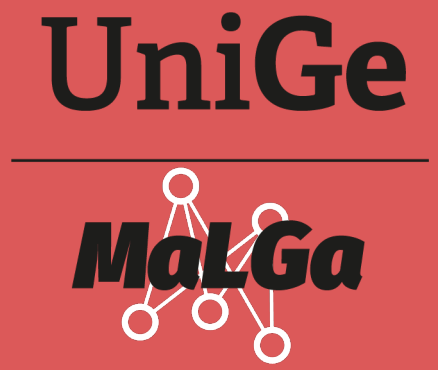
# THE DATA SCIENCE PROCESS




*Definiamo un* ***obiettivo analitico****, che ci guiderà nelle fasi successive*

*A questo livello possiamo essere abbastanza generici*

*Proseguendo nella pipeline l'obiettivo potrebbe diventare più specifico*


UniGe
MaLGa