

# Business Intelligence per Big Data

*Progetto di analisi di dati*

DBG  
MG



**Politecnico  
di Torino**

AA 2021-2022 Politecnico di Torino

# Datasets



Politecnico  
di Torino

- Ogni datasets contiene 4000 reviews effettuate da utenti Amazon
- In ogni dataset sono presenti 4 diverse categorie di prodotti recensiti (1000 reviews per ogni categoria)
- Ciascun gruppo dovrà utilizzare solo un sotto-campione di questi dati, in modo da poter eseguire le analisi in tempi ragionevoli

# Datasets



Politecnico  
di Torino

## ■ Caratteristiche di ogni dataset:

- file .csv contenente 4000 reviews effettuate da utenti Amazon
- 10 attributi:
  - *reviewerID*: codice identificativo dell'utente
  - *asin*: codice identificativo del prodotto recensito
  - *reviewerName*: nome dell'utente
  - *helpful*: valutazione sull'utilità della recensione (formato: [a, b])
    - b rappresenta il numero di volte totale in cui la recensione è stata valutata
    - a rappresenta il numero di volte totale in cui la recensione è stata valutata utile
  - *reviewText*: testo della recensione
  - *overall*: valutazione sul prodotto (max 5)
  - *summary*: riassunto della recensione
  - *unixReviewTime*, *reviewTime* : data della recensione
  - *label*: categoria a cui il prodotto appartiene

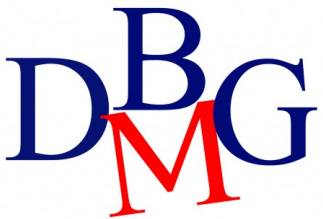
# Datasets



Politecnico  
di Torino

## ■ Categorie del prodotto recensito:

- ogni dataset contiene 1000 recensioni per categoria
- in tutto esistono 8 categorie (ma solo 4 all'interno di ciascun dataset)
- ogni categoria è identificata da un numero intero:
  - 1: cd e vinili
  - 2: casa e cucina
  - 3: video games
  - 4: kindle store
  - 5: sports & outdoors
  - 6: film & TV
  - 7: cellulari e accessori
  - 8: salute e cura della persona



# Regole



**Politecnico  
di Torino**

- Gruppi di 2 persone:
  - Registrarsi sul google sheet
  - [https://docs.google.com/spreadsheets/d/1mc\\_jsYyg0vnxBeVIFJZvANxqQhzvehuwwkYUFMHkhk/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1mc_jsYyg0vnxBeVIFJZvANxqQhzvehuwwkYUFMHkhk/edit?usp=sharing)
- Ogni gruppo deve:
  - Caratterizzare il dataset
  - Effettuare diverse sessioni di analisi su un dataset utilizzando il tool RapidMiner e/o altri tools noti al gruppo di studenti
  - Analizzare i risultati e sintetizzarli in grafici
  - Discutere come sfruttare la conoscenza estratta in un'applicazione di business

# Regole



Politecnico  
di Torino

- Preparare una breve ma completa presentazione sulle attività svolte:
  - Caratterizzazione del dataset
  - Analisi effettuata (e.g., configurazione ottimale dell'algoritmo selezionato)
  - Risultati migliori ottenuti e comparativa (di performance e qualità della conoscenza) tra algoritmi diversi
- Presentare i risultati in 15 minuti
  - 5 minuti di presentazione a testa e 5 di domande

DBG  
MG