Introduction to Data Mining

Irene Siragusa, PhD

CHILAB
Università degli Studi di Palermo

Outline

- Introduction
- Big Data
- Data mining processing
- Data types

Introduction

What is data mining?

The study of collecting, cleaning, processing, analyzing, and gaining useful insights from data.

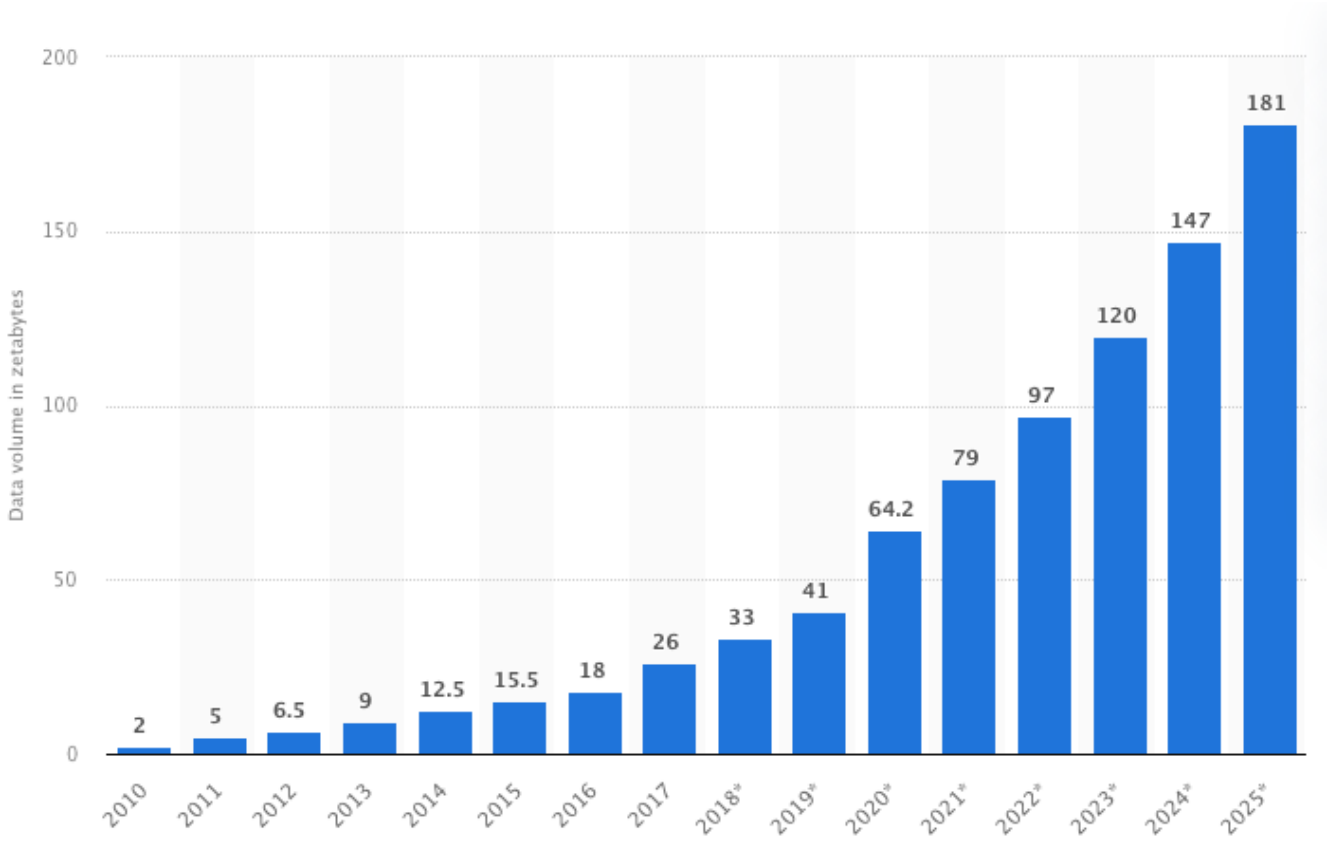
Why data is so important?

We are in the Era of Big Data,

and big, is not just an adjective

Introduction

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

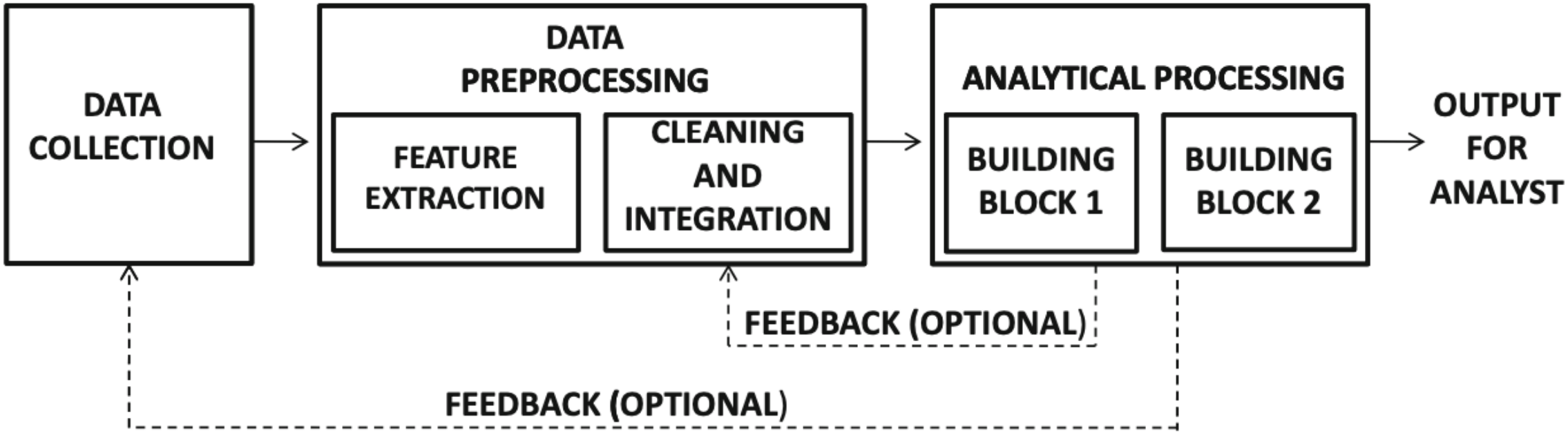


1 Zb = 1e9 TB = 1e12 GB

Big Data

- Volume in terms of memory of the data
- Velocity (speed) of data's production and update
- Variety, internal organization and structure of data
- Veracity, truthfulness of the produced data
- Value, actual value of raw or aggregated data

Data mining process



Data mining process – Data collection

- Data can be extracted from heterogenous sources: sensors, system logs, documental corpora, biomedical data …
- Managing structured or semi-structured data
- Volume and truthfulness of collected data
- Data storage: database, datawarehouse, data lake, HDFS and database NoSQL

Data mining process – Data preprocessing

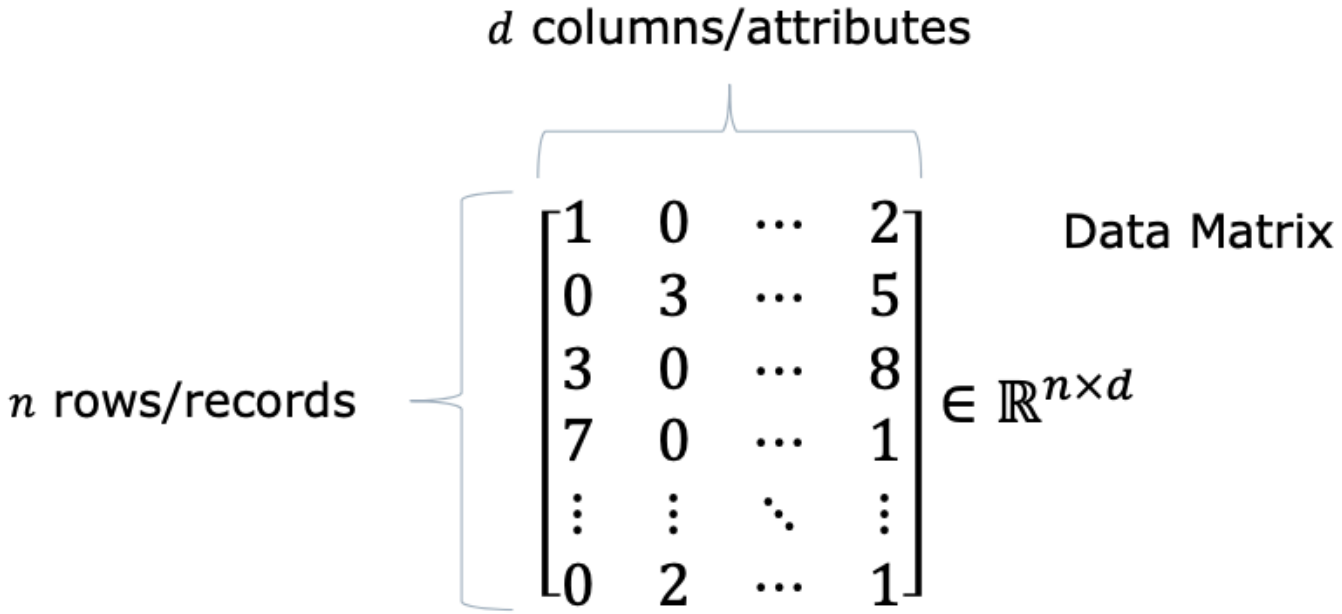
- Feature extraction
  - Feature is a valuable characteristics of data (field of interest in a record)
  - Most important feature identification
  - Transformation from raw data to a suitable format for analysis algorithms (multi-dimentional vectors, time series, binary or categorical data)

Data mining process – Data preprocessing

- Data cleaning and integration
  - Handle missing and erroneous values
  - Integrate data from multiple sources
  - Horizontal and vertical operation on data and data field
  - This process can be done relying on problem-related knowledge
  - Cured data are stored accordingly

Data mining process – Analytical processing

Cured data are organized in a database $\mathcal{D}$ with n records and d features, that can be represented by a data matrix $\mathcal{X}$ with n row vectors $\boldsymbol{x} \in \mathbb{R}^d$



Data mining process – Analytical processing

We can look for:

- Relations among $\mathcal{X}$ columns
  - Recurrent structures among features of single data point that are correlated with another (target) feature
  - e.g. Association pattern mining, Data classification

- Relations among $\mathcal{X}$ rows
  - Similarity among data points, thus making some data points more similar to other with respect to some criteria
  - e.g. Clustering, Outlier analysis

Data mining process – Analytical processing

- Association Pattern Mining
  - Originally defined in the context of *sparse binary databases*
  - e.g. customer transactional database

    each column in the data matrix corresponds to an item, and a customer transaction represents a row, $x_{i,j}$ is 1, if customer transaction i contains item j as one of the items that was bought.

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \in \{0,1\}^{5\times 4}$$

Data mining process – Analytical processing

- Frequent Pattern Mining

  Given a binary $n \times d$ data matrix D, determine all subsets of columns such that all the values in these columns take on the value of 1 for at least a fraction s of the rows in the matrix.

  - Relative frequency of a pattern is referred to as its *support*, s is the *minimum support*, and patterns that satisfy the minimum support requirement, are *frequent patterns* or *itemset*

Minimum support	Frequent patterns	Support
2/5	{2,3}	3/5
	{1,4}	2/5

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \in \{0,1\}^{5\times 4}$$

Data mining process – Analytical processing

- Frequent Pattern Mining
  - Frequent Pattern Mining is a class of association patterns.
  - Other definitions of relevant association patterns are possible that do not use absolute frequencies but use other statistical quantifications such as the χ^2 measure.
  - These measures often lead to generation of more interesting rules from a statistical perspective.
  - Association pattern mining was originally proposed in the context of association rule mining, where confidence of the rule is considered

Data mining process – Analytical processing

- Confidence of the rule
  - $\mathrm{conf}(A \Rightarrow B)$ is the fraction of transactions containing A, which also contains B.
  - Confidence is obtained by dividing the support of the pattern $A \cup B$ with the support of pattern A

$$\mathrm{conf}(A \Rightarrow B) = \frac{\mathrm{supp}(A \cup B)}{\mathrm{supp}(A)}$$

If A appears, then B also appears

Data mining process – Analytical processing

- Association Rules
  - Let A and B be two sets of items. The rule $A \Rightarrow B$ is said to be valid at support level s and confidence level c, if the following two conditions are satisfied:

1. The support of the item set A is at least s

$$\text{supp}(A) \geq s$$

2. The confidence of $A \Rightarrow B$ is at least c

$$\text{conf}(A \Rightarrow B) \geq c$$

Data mining process – Analytical processing

- Classification
  - Prediction of a discrete label (particular feature) for a given data point
  - Supervised approach, labels are known for the training data set
  - Predict the correct label for new data, by learning the relationships of the provided features in the data with respect to the label

Data mining process – Analytical processing

- Clustering
  - Looking for similarity groups among data
  - Unsupervised approach, groups are not known (numbers or structure)
  - e.g. customer segmentation, data summarization

Data mining process – Analytical processing

- Outlier analysis
  - Outlier can be identified as a data point that highly differs from the others, and it can arise a suspect that it may be generated differently
  - Outliers can be detected via clustering analysis
  - Relevant example can be:
    - Intrusion detection
    - Financial Froud detection
    - Unexpected patterns from sensors (failure prevention)
    - Unusual patterns in medical imaging (disease)
    - Weather and environmental forecasts

Data types

Dependency is a feature related to the origin of the observed phenomenon from which data are collected

- Nondependency-oriented Data
  - Data records do not have any specified dependencies between either the data items or the attributes
  - e.g. set of demographics records

- Dependency-oriented Data
  - Implicit or explicit relationships may exist between data items
  - e.g. social networks data or time series

Nondependency-oriented Data

- Multidimendional Data (vectors)
  - A multidimensional data set $\mathcal{D}$ is a set of n records, $\overline{X_1}, \ldots, \overline{X_n}$ such that each record $\overline{X_i}$ contains a set of d features denoted by $(x_i^1, \ldots, x_i^d)$

Name	Age	Gender	Race	ZIP code
John S.	45	M	African American	05139
Manyona L.	31	F	Native American	10598
Sayani A.	11	F	East Indian	10547
Jack M.	56	M	Caucasian	10562
Wei L.	63	M	Asian	90210

Nondependency-oriented Data

- Multidimendional Data (vectors)
  - $\overline{X_i}$ is a record, data point, instance, example, transaction, entity, tuple, object, feature-vector (sample)
  - x_i^k is a field, attribute, dimension, feature.

Name	Age	Gender	Race	ZIP code
John S.	45	M	African American	05139
Manyona L.	31	F	Native American	10598
Sayani A.	11	F	East Indian	10547
Jack M.	56	M	Caucasian	10562
Wei L.	63	M	Asian	90210

Nondependency-oriented Data

- Quantitative Multidimensional Data
  - Numerical in the sense that they have a natural ordering
  - Continuous, numeric, or quantitative
  - Convenient for analytical processing (mean, variance, …)

Name	Age	Gender	Race	ZIP code
John S.	45	M	African American	05139
Manyona L.	31	F	Native American	10598
Sayani A.	11	F	East Indian	10547
Jack M.	56	M	Caucasian	10562
Wei L.	63	M	Asian	90210

Nondependency-oriented Data

- Categorical Data
  - Take on discrete unordered values
  - Unordered discrete-valued Data

Name	Age	Gender	Race	ZIP code
John S.	45	M	African American	05139
Manyona L.	31	F	Native American	10598
Sayani A.	11	F	East Indian	10547
Jack M.	56	M	Caucasian	10562
Wei L.	63	M	Asian	90210

Nondependency-oriented Data

- Mixed Attribute Data
  - Combination of categorical and numeric attributes
- Binary Data
  - A special case of multidimensional categorical data, where each categorical attribute may take on one of at most two discrete values
  - A special case of multidimensional quantitative data, where an ordering exists between the two values
  - Setwise data if attribute is treated as a set element indicator

Nondependency-oriented Data

- Text Data
  - Generally referred to string, that have an order

BUT

  - Text can be represented in a vector-space representation in terms of frequencies of the words in a document, thus creating a *document-term matrix* $n \times d$ with n documents and d terms.
  - In this configuration textual data can be analyzed by distance measures (Latent Semantic Analysis)

Dependency-oriented Data

- Implicit depencencies
  - Known dependencies related to data's domain
  - Data values may be related to each other temporally or spatially

- Explicit depencencies
  - An explicitly dependencies is reported
  - Graph or network data where edges are used to specify relationships

Dependency-oriented Data

- Time-Series Data
  - Series of values generated by continuous measurement over time.
  - *Contextual attributes*
    - Define the context on the basis of which the implicit dependencies occur in the data
    - e.g. time stamp
  - *Behavioral attributes*
    - Represent the values that are measured in a particular context
    - e.g. actual value of interest

Dependency-oriented Data

- Multivariate Time-Series
  - A time series of length n and dimensionality d contains d *numeric features* at each of n time stamps
  - $(t_1; \overline{Y_1}), \ldots, (t_n; \overline{Y_n})$, where $\overline{Y_i} = (y_i^1, \ldots, y_i^d)$
- Multivariate Discrete Sequence Data
  - Discrete sequence of length n and dimensionality d contains d *discrete features* at each of n time stamps
  - $(t_1; \overline{Y_1}), \ldots, (t_n; \overline{Y_n})$, where $\overline{Y_i} = (y_i^1, \ldots, y_i^d)$

Dependency-oriented Data

- Textual Data
  - A multivariate discrete sequence with $d = 1$
  - Text is represented as vectors called *word embeddings*
    - *Noncontextual word embeddings*
    - *Contextual word embeddings*

Dependency-oriented Data

- Spatial Data
  - A d-dimensional spatial data record contains d behavioral attributes and one or more contextual attributes containing the spatial location.
  - A set of n locations associated with corresponding d behavioral attributes
  - $(L_1; \overline{X_1}), \ldots, (L_n; \overline{X_n})$, where $\overline{X_i} = (x_i^1, \ldots, x_i^d)$
- Spatiotemporal Data
  - Both spatial and temporal attributes are contextual
  - The temporal attribute is contextual, whereas the spatial attributes are behavioral (e.g. trajectory analysis)

Dependency-oriented Data

- Network and Graph Data
  - A network $G = (N, A)$ contains a set of nodes N and a set of edges A, where the edges in A represent the relationships between the nodes. In some cases, an attribute set $\overline{X_i}$ may be associated with node i, or an attribute set $\overline{Y_{ij}}$ may be associated with edge (i, j).
  - $(L_1; \overline{X_1}), \ldots, (L_n; \overline{X_n})$, where $\overline{X_i} = (x_i^1, \ldots, x_i^d)$
  - e.g. Web Graph, Social Networks, Chemical compound databases