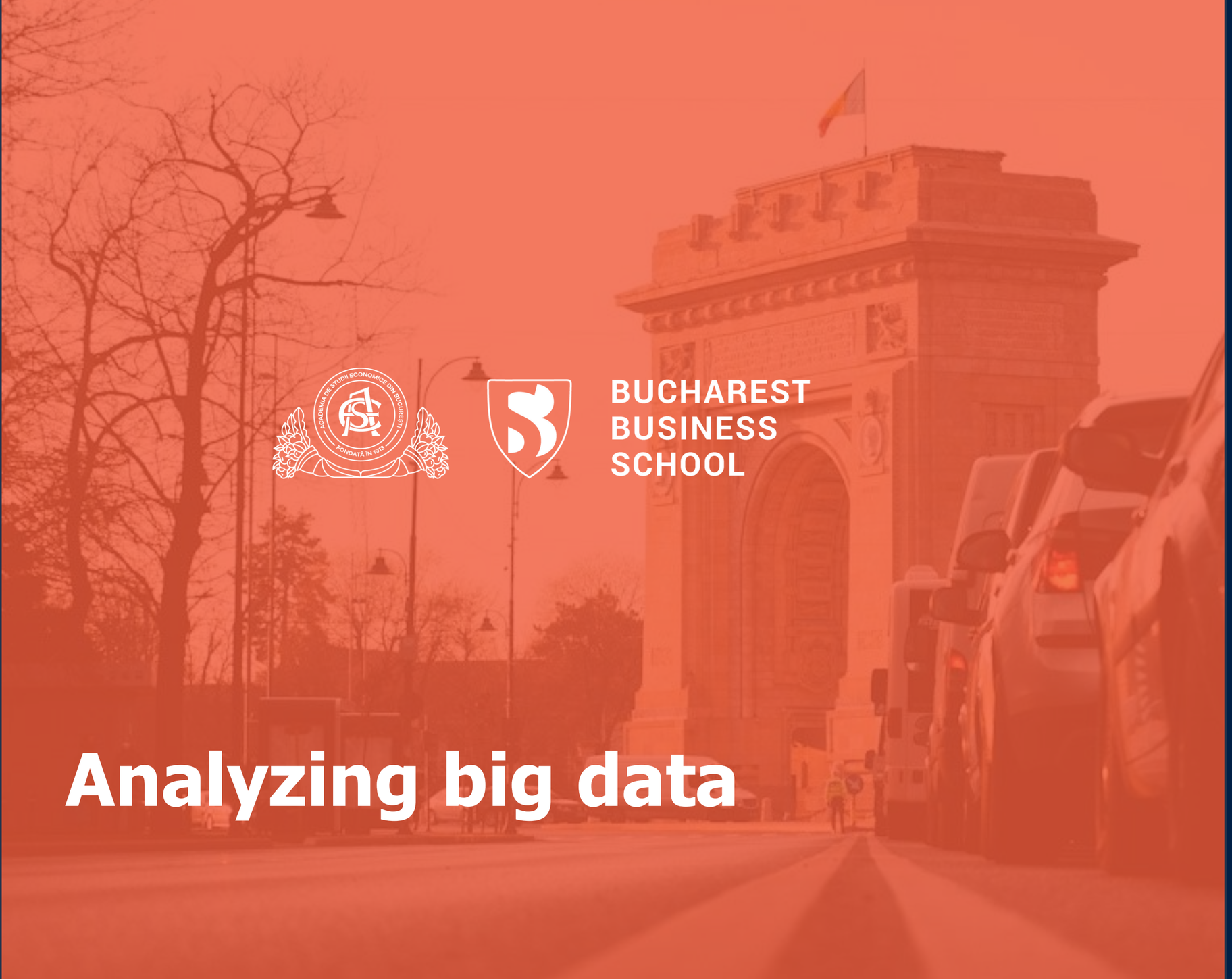


BUCHAREST BUSINESS SCHOOL

ASSOCIATION OF MBAS ACCREDITED

Leadership Through Value

Analyzing big data

DATA MINING AND BIG DATA

CATALINA CHINIE
PH.D. ASSOCIATE PROFESSOR

SUMMARY

Data Mining and Big Data

Block 2

- Structured, semi-structured & unstructured data
- Data mining
- Text analytics
- Sentiment analysis
- Predictive analysis

Structured, semi-structured & unstructured data

	Structured Data	Semi-Structured Data	Unstructured Data
Definition	Organized data with a fixed schema, stored in tabular formats like rows and columns.	Data that doesn't conform to a rigid structure but contains tags or markers to separate elements.	Data lacking a predefined format or structure, making it difficult to store and analyze using traditional tools.
Examples	Spreadsheets, relational databases, customer records, transaction data.	JSON, XML, HTML files, email headers, sensor data with metadata.	Images, videos, audio files, social media posts, PDFs, emails.
Storage Systems	Relational databases (e.g., MySQL, PostgreSQL), data warehouses.	NoSQL databases (e.g., MongoDB), data lakes, document stores.	Data lakes, content management systems, file systems.
Ease of Analysis	High; easily queried using SQL and other traditional tools.	Moderate; requires specialized tools to parse and analyze.	Low; necessitates advanced tools like NLP and machine learning for analysis.
Use Cases	Business intelligence, reporting, CRM systems, inventory management.	Web data integration, IoT data processing, email analysis.	Sentiment analysis, image and video recognition, customer feedback analysis.
Flexibility	Low; changes require altering the schema.	Moderate; can accommodate changes more easily than structured data.	High; highly adaptable but challenging to manage.
Tools & Technologies	SQL, OLAP tools, relational database management systems.	NoSQL databases, data integration tools, XML/JSON parsers.	Big data platforms (e.g., Hadoop), AI/ML frameworks, NLP tools.
Searchability	Easy; data is indexed and searchable.	Moderate; searchable with appropriate parsing.	Difficult; requires content-based search techniques.
Volume in Enterprises	Approximately 10-20% of enterprise data.	Varies; often used in specific applications.	Approximately 80-90% of enterprise data.

Data mining is the process of extracting knowledge from data

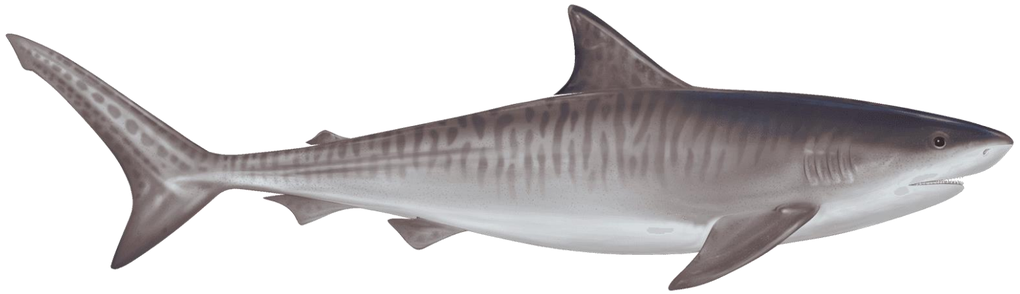
Its objective is to discover patterns, correlations, and insights from large datasets using statistical, machine learning, and analytical techniques.

When performing data mining, we must be aware of false positives.

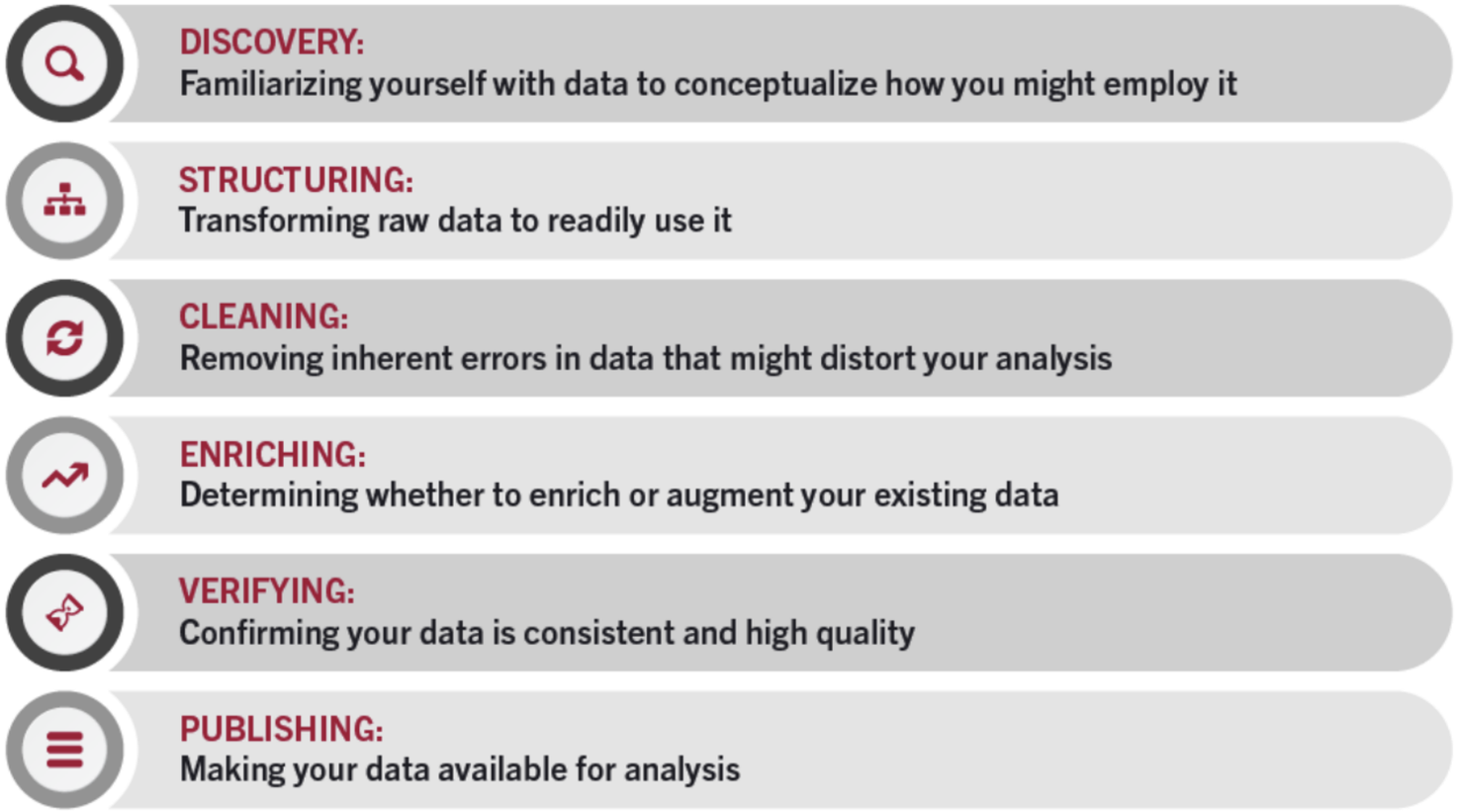
Think about:

- The role of chance

- Correlations vs causality



Data wrangling is a process that prepares data for analysis



Transformation of data

- Structurally manipulate and combine the data using Joins and Unions.
- Normalize data, that is, clean the database of unused and redundant data.
- Denormalize data, that is, combine data from multiple tables into a single table so that it can be queried faster.

Data cleaning

- Removing inherent errors in data that might distort your analysis:
  - Uncover quality issues,
  - Identify and fix outliers,
  - Fix missing values, duplicate data, irrelevant data, inconsistent formats, syntax errors.

There are four key types of data analytics

- Descriptive, which answers the question, “What happened?”
- Diagnostic, which answers the question, “Why did this happen?”
- Prescriptive, which answers the question, “What should we do next?”
- Predictive, which answers the question, “What might happen in the future?”

Data mining applications

Statistical analysis

Descriptive statistics consists of methods for organizing and summarizing information (Weiss, 2012). Descriptive statistics form a basis for all quantitative analysis and are a precursor for inferential statistics (sciencedirect.com).

Inferential statistics consists of methods for drawing and measuring the reliability of conclusions about a population based on information obtained from a sample of the population (Weiss, 2012).

Data mining techniques

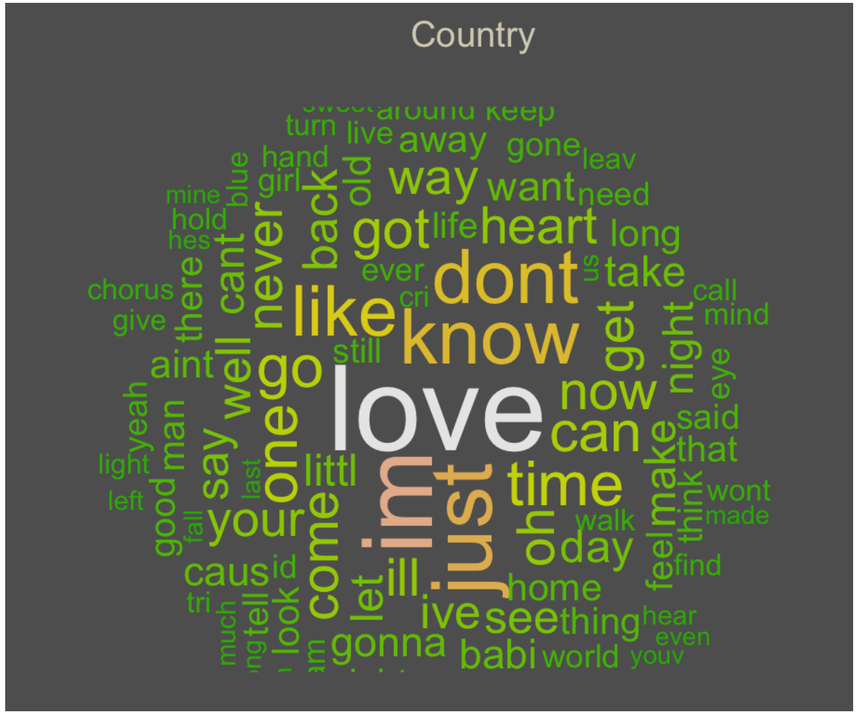
- **Regression** - used for predicting a continuous-valued attribute based on the values of other attributes
- **Classification** - categorizes data into predefined classes (Models are trained on a labeled dataset)
- **Clustering** - groups a set of objects so that objects in the same group (or cluster) are more similar to each other than to those in other groups (Ezugwu et al., 2022).
- **Association** - identifies relationships or correlations among a large set of data items, often used to discover which items are frequently bought together (Schmueli et al., 2017).

Data mining techniques

- **Sequential patterns** - discovers statistically relevant patterns where the values or events are delivered in a sequence, which is essential in customer behavior analysis and temporal databases (Olson & Delen, 2008).
- **Affinity grouping** - identifying sets of items or behaviors that frequently occur together across transactions or users, often used in retail for shelf placement and recommendation systems (Gorunescu, 2011).
- **Decision trees** - used for both classification and regression. They partition the data into subsets based on the value of input features, leading to decisions at the leaves (Maimon & Rokach, 2014).

Text analytics

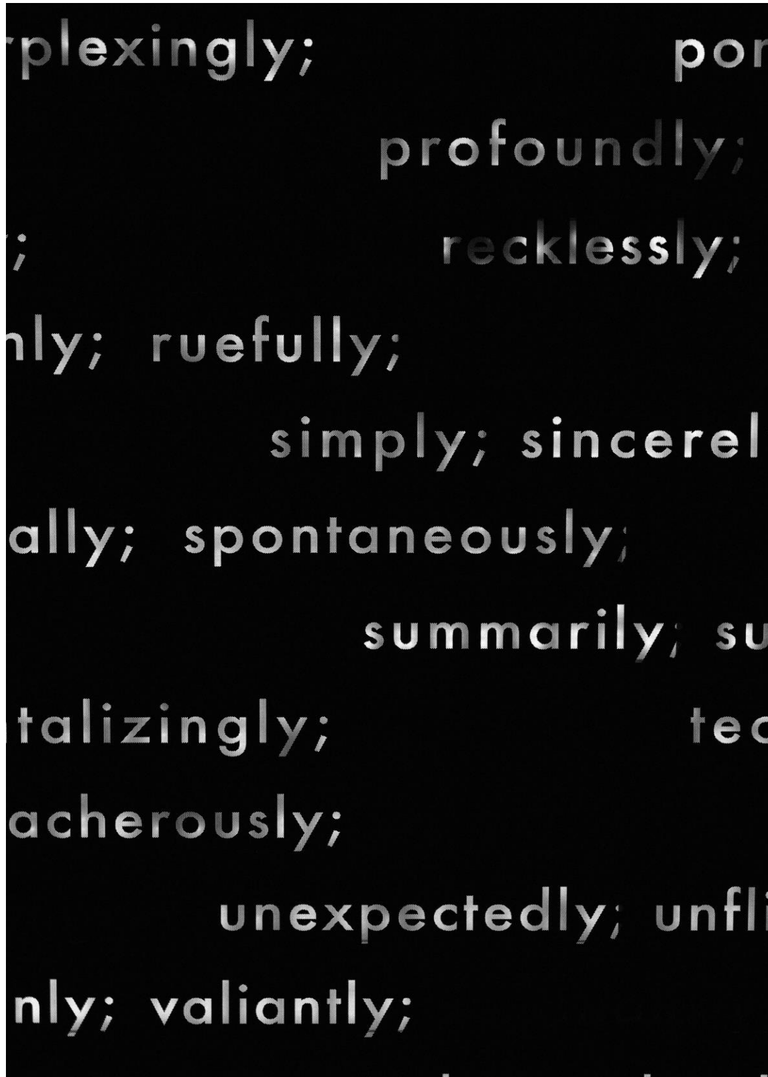
Stage 1—Data gathering
Stage 2—Data preparation
Tokenization - segregating the raw text into multiple parts that make semantic sense.
Part-of-speech tagging - assigns grammatical tags to the tokenized text.
Parsing - establishes meaningful connections between the tokenized words with English grammar. It helps the text analysis software visualize the relationship between words.
Lemmatization - a linguistic process that simplifies words into their dictionary form, or lemma. For example, the dictionary form of *visualizing* is *visualize*.
Stop words removal - are words that offer little or no semantic context to a sentence, such as *and*, *or*, and *for*.
Stage 3—Text analysis
Text classification - assigning tags to the text data that are based on rules or machine learning-based systems.
Text extraction - identifying the presence of specific keywords in the text and associating them with tags.
Stage 4—Visualization



[Text analytics & topic modelling on music genres song lyrics]

Sentiment analysis

Sentiment analysis, also known as **opinion mining**, is a natural language processing (NLP) technique used in data mining to automatically detect, extract, and classify subjective information in text—typically opinions, attitudes, or emotions toward a topic, product, or service.



Predictive analytics

Predictive analytics is a branch of data mining and machine learning that focuses on using historical data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes. It is grounded in the idea of making data-driven predictions based on patterns and trends observed in past data (Larose, 2015).

BUCHAREST
BUSINESS
SCHOOL