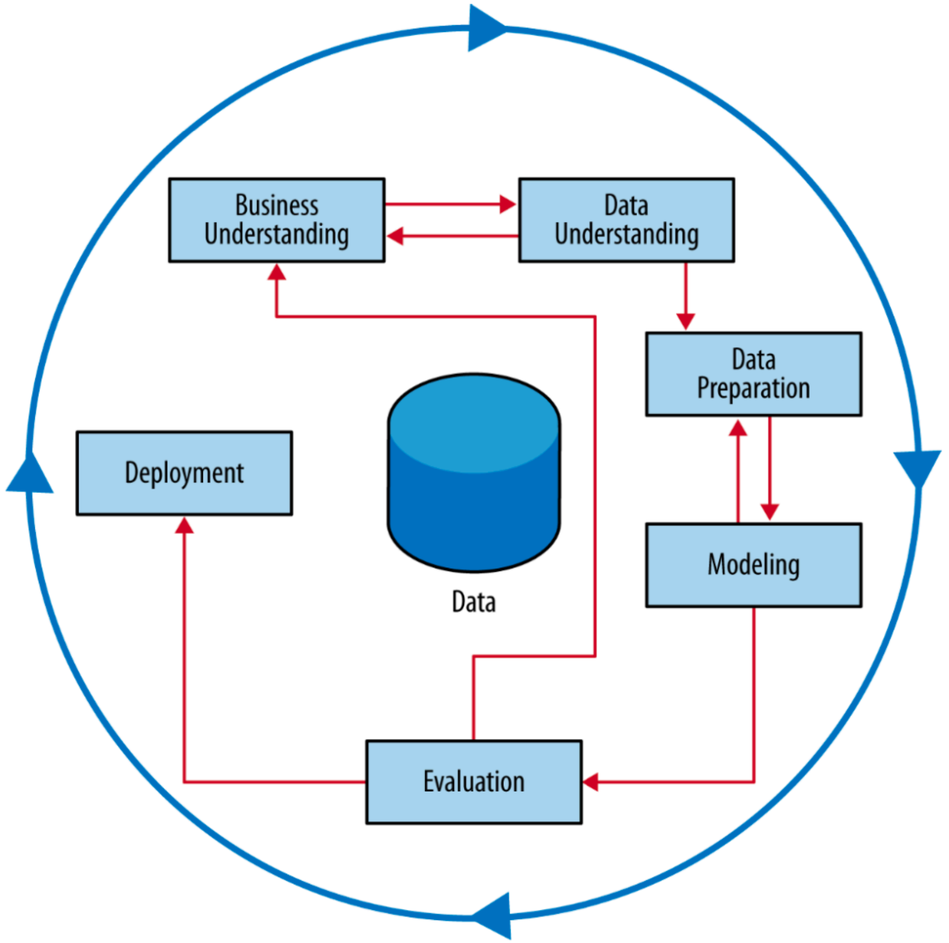
Introduction to Data Science

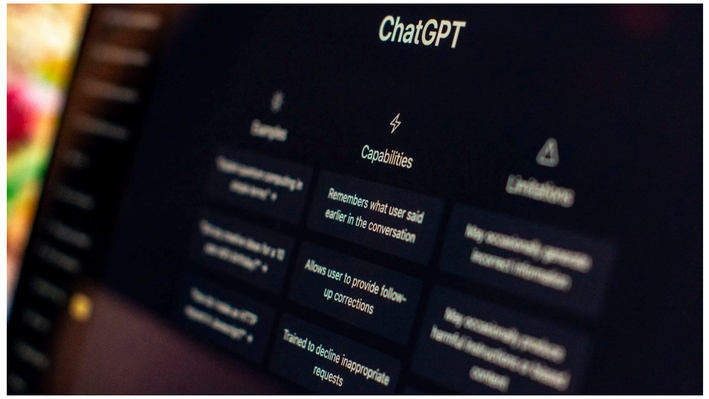
Lesson 2
Data foundations

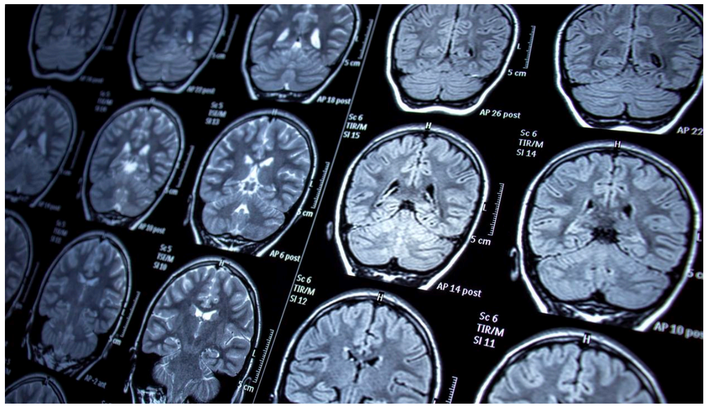
Marija Stankova Medarovska, PhD
marija.s.medarovska@uacs.edu.mk

Data science process

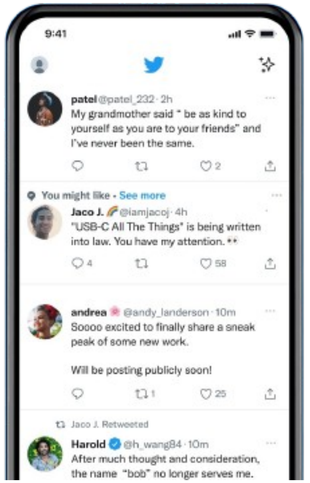


Data availability





Data availability



Textual

Audio

Images

Numerical

and more…

Video

Data generated (2023)

500 hours of video content on **Youtube** per minute

41.7 million **WhatsApp** messages sent per minute

347,222 **Instagram** stories per minute

2.4 million **Google** searches made per minute.

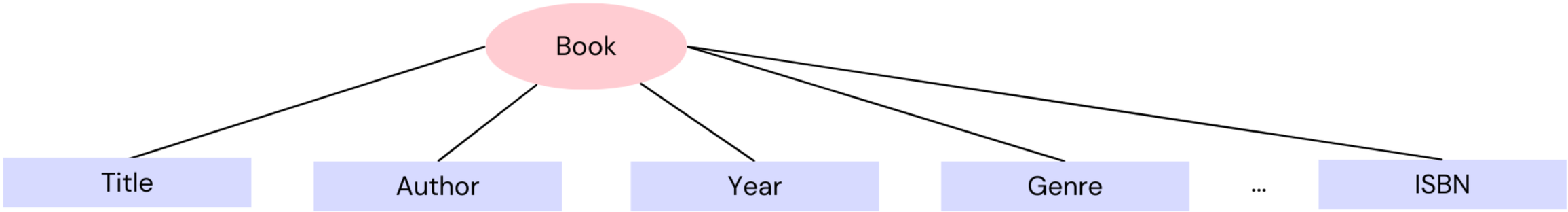
147,000 **Facebook** photos shared per minute

and more…

What is data?

- In its most basic form, a datum or a single piece of information is an abstraction of an entity (ex. person, object, event, etc).

- Each entity is typically described by a number of attributes



Example: A book entity might have the following attributes - title, author, date published, genre, International Standard Book Number (ISBN), and so on.

Dataset

- A dataset consists of the data relating to a collection of entities, with each entity described in terms of a set of attributes.

ID	Title	Author	Year	Cover	Edition	Price
1	*Emma*	Austen	1815	Paperback	20th	$5.75
2	*Dracula*	Stoker	1897	Hardback	15th	$12.00
3	*Ivanhoe*	Scott	1820	Hardback	8th	$25.00
4	*Kidnapped*	Stevenson	1886	Paperback	11th	$5.00

Example: A dataset of classic books

Dataset

- An instance (or example, entity, record) refers to a single occurrence of data within a dataset

ID	Title	Author	Year	Cover	Edition	Price
1	*Emma*	Austen	1815	Paperback	20th	$5.75
2	*Dracula*	Stoker	1897	Hardback	15th	$12.00
3	*Ivanhoe*	Scott	1820	Hardback	8th	$25.00
4	*Kidnapped*	Stevenson	1886	Paperback	11th	$5.00

Example: Each row in the table describes one book

Dataset

- An attribute (or variable, feature) refers to a property or characteristics of the data

ID	Title	Author	Year	Cover	Edition	Price
1	*Emma*	Austen	1815	Paperback	20th	$5.75
2	*Dracula*	Stoker	1897	Hardback	15th	$12.00
3	*Ivanhoe*	Scott	1820	Hardback	8th	$25.00
4	*Kidnapped*	Stevenson	1886	Paperback	11th	$5.00

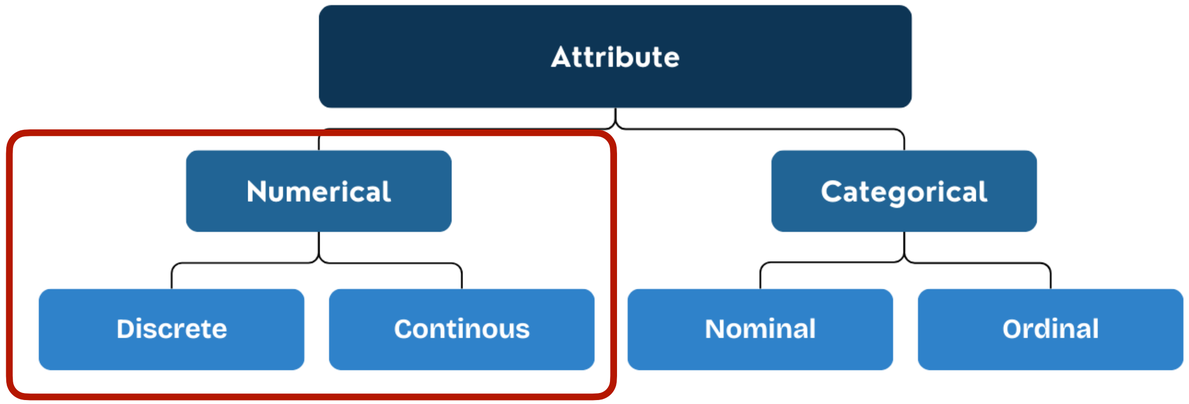
Example: Each column in the table describes a particular attribute of the book

Attributes selection

- We could have included many more attributes for each book, but, as is typical of data science projects, we need to make a choice when designing the dataset
- More data is usually better, however, it does not come without cost:
  - additional time and effort in collecting and quality checking
  - including irrelevant or redundant attributes could lead to reduced performance of the models (overfitting)
- Choosing the correct attributes is a challenge faced by all data science projects - it comes down to domain knowledge, data availability, possibly trial-and-error experiments where each iteration checks the results achieved using different subsets of attributes.

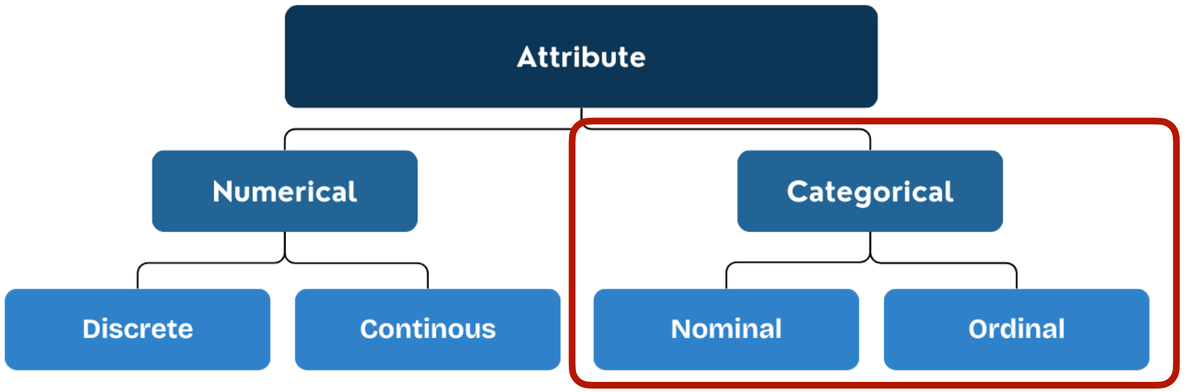
Types of attributes

- Numerical (quantitative) attributes describe measurable quantities that are represented iusing integer or real values.
  - Discrete attributes take on finite or countable values (ex. number of pages: 125, number of copies sold: 20 000)
  - Continuous attributes can take on any value within a range (ex. price: $15.99, weight: 500.25 grams)



Types of attributes

- Categorical (qualitative) attributes represent categories or labels rather than numerical values
  - Nominal attributes take categories that do not have any ordering (ex. genre: biography/romance…, cover: paperback/hardback…)
  - Ordinal attributes have categories that have a meaningful order or ranking,(ex. condition: new/like new/good/fair/poor, user ranking)



Approximation of reality

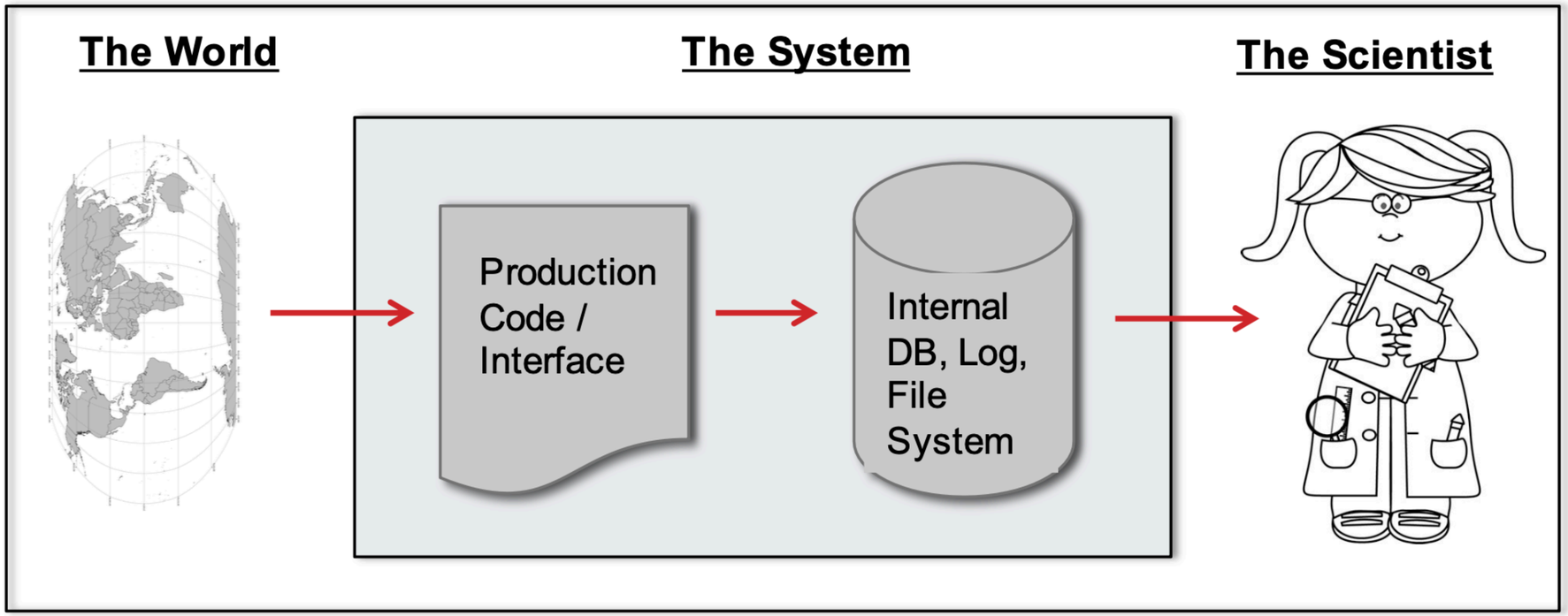
- Data are generated through a process of abstraction, where choices have been made with regard to what to abstract from and what categories or measurements to use in the abstracted representation.

- In other words, the data we use for data science are not a perfect representation of the real-world entities and processes we are trying to understand, but if we are careful in how we design and gather the data that we use, then the results of our analysis will provide useful insights into our real-world problems.

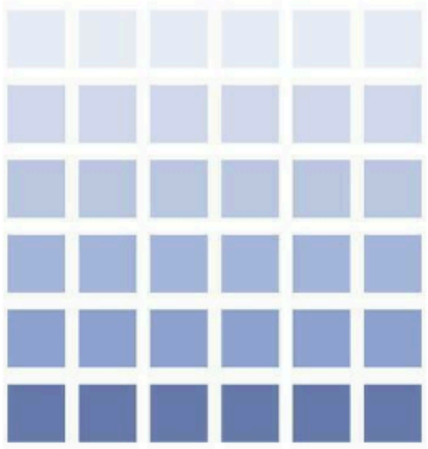
- “Garbage in, garbage out”: if the inputs to a computational process are incorrect, then the outputs from the process will be incorrect.

.

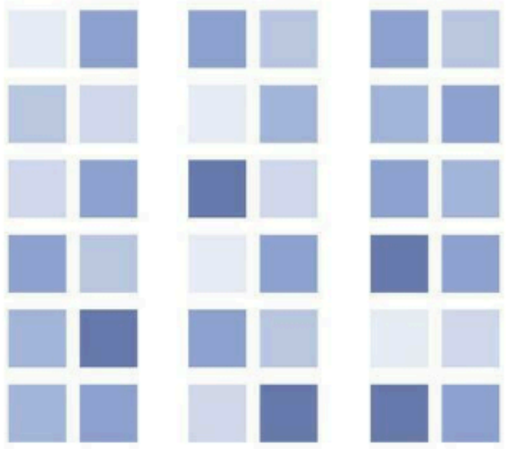
Approximation of reality



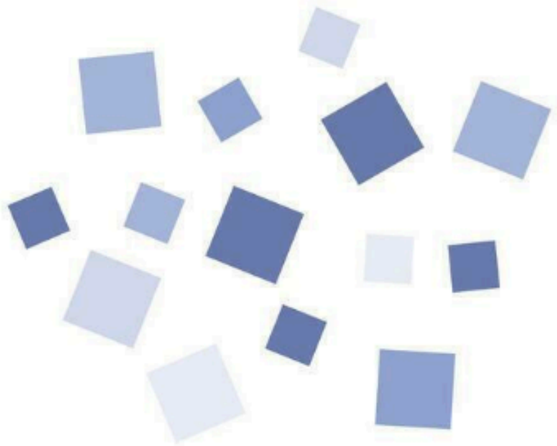
Structure of data



Structured data



Semi - structured data

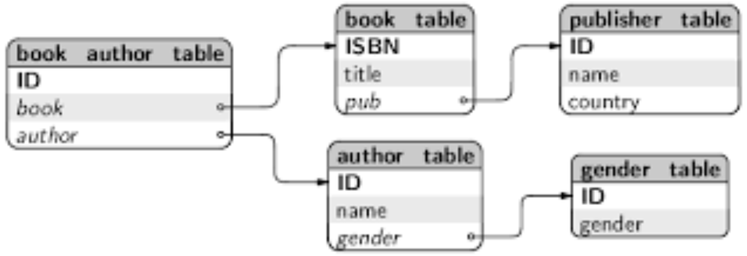


Unstructured data

Structured data

- Structured data are highly organized and easily searchable, typically stored in relational databases or spreadsheets with a fixed schema. In many cases they can be represented in a tabular manner with rows and columns.

Spreadsheets

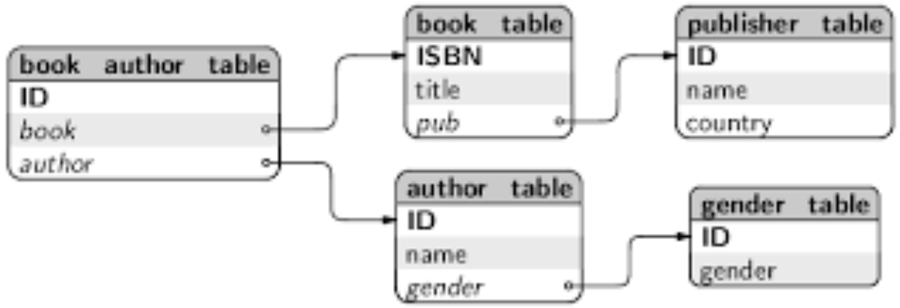


Relational databases

Other sources include Online Transaction Processing (or OLTP), Online forms, Sensors such as Global Positioning Systems (or GPS) and Radio Frequency Identification (or RFID) tags; Network and Web server logs etc.

Relational database*

- A relational database (RDB) is **a way of structuring data in tables, rows, and columns**. An RDB has the ability to establish links—or relationships–between information by joining tables, which makes it easy to understand and gain insights about the relationship between various data points.



Example: Entity-Relationship (ER) diagram for a book relational database

```sql
SELECT book_table.title AS title, author_table.name AS author
FROM book_author_table
JOIN book_table ON book_author_table.book = book_table.ISBN
JOIN author_table ON book_author_table.author = author_table.ID;
```

Example: Structured Query Language (SQL) query for getting the title and author of a book

Semi-structured data

- Semi-structured data have some organizational properties but lack a fixed or rigid schema. Semi-structured data cannot be stored in the form of rows and columns as in databases.

eXtensible Markup Language (XML)

JSON

Other sources could include: E-mails, Binary executables, TCP/IP packets, Zipped files, Integration of data from different sources. etc.

eXtensible Markup Language (XML)*

- XML (eXtensible Markup Language) is a structured markup language used for representing data and is widely used for data exchange between systems.
- Uses a hierarchical structure with opening and closing tags
- Attributes can be used to add metadata to elements

```
<book>
    <bookId>99</bookId>
    <author>A.C. Weisbecker</author>
    <title>Cosmic Banditos: A Contrabandista's Quest for the Meaning of
Life</title>
    <publicationYear>1988</publicationYear>
    <users>
        <user>
            <userId>u1936734</userId>
            <catalogueDate>2009-06</catalogueDate>
            <rating>0.0</rating>
            <tags>Literature, American Literature</tags>
        </user>
        <user>
            <userId>u0871476</userId>
            <catalogueDate>2008-12</catalogueDate>
            <rating>0.0</rating>
            <tags>Fiction, Humor</tags>
        </user>
    </users>
</book>
```

Example: eXtensible Markup Language (XML)

JavaScript Object Notation (JSON)*

- JavaScript Object Notation (JSON), is a lightweight and human-readable format for storing and exchanging data
- In JSON , data is primarily stored in two structures: objects and arrays
- Objects in JSON are collections of key/value pairs enclosed in curly braces {}
- Arrays are ordered lists of values, enclosed in square brackets []

```
{
  "book": {
    "bookId": 99,
    "author": "A.C. Weisbecker",
    "title": "Cosmic Banditos: A Contrabandista's Quest for the Meaning of
Life",
    "publicationYear": 1988,
    "users": [
      {
        "userId": "u1936734",
        "catalogueDate": "2009-06",
        "rating": 0.0,
        "tags": ["Literature", "American Literature"]
      },
      {
        "userId": "u0871476",
        "catalogueDate": "2008-12",
        "rating": 0.0,
        "tags": ["Fiction", "Humor"]
      }
    ]
  }
}
```

Example: JavaScript Object Notation (JSON)

Unstructured data

- Unstructured data do not have an easily identifiable structure and, therefore, cannot be organized in a mainstream relational database in the form of rows and columns. It does not follow any particular format, sequence, semantics, or rules.

Images

Documents

Other sources could include web pages, social media feeds, video and audio files, media logs; and surveys.

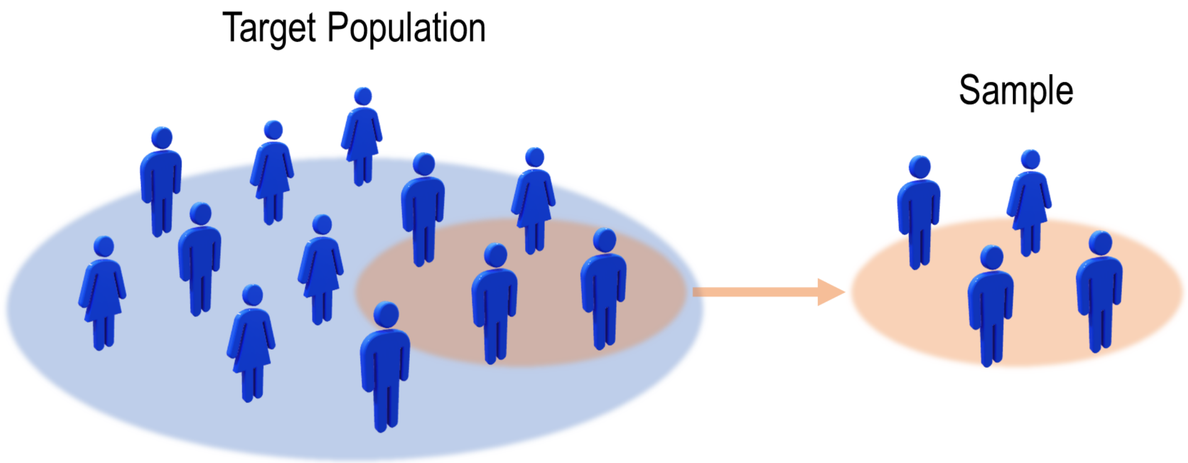
Challenges with data collection

- Data integration from different systems
- Privacy issues
- Data quality issues: missing values, inaccuracies, outliers…
- Sampling bias
- Drift

Sampling

- Sampling is the process of selecting a subset of data from a larger dataset (population) in order to analyze and draw conclusions about the entire population

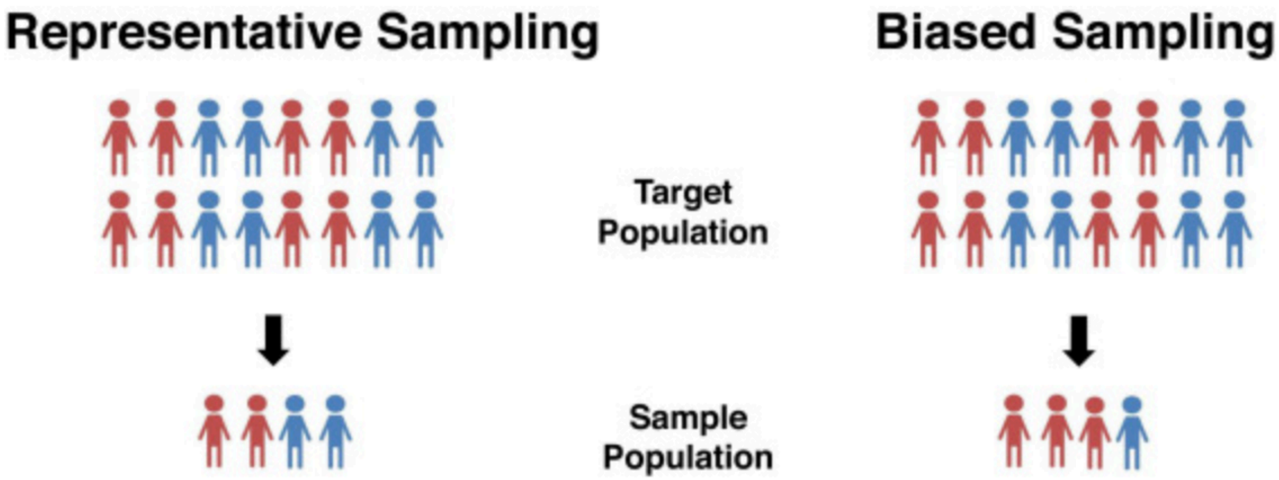
- Used when dealing with large datasets or when collecting data from an entire population is impractical or expensive



Sampling bias

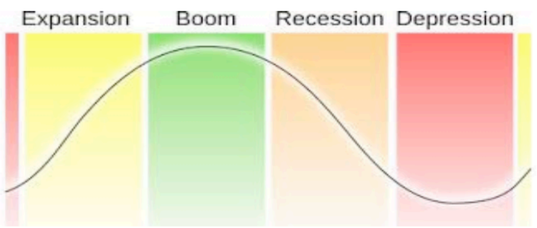
- Sampling bias occurs when some members of a population are more likely to be selected than others, resulting in an unrepresentative sample.

- Can lead to incorrect or misleading insights and predictions



Data drift

- Data drift (or non-stationarity of data) refers to changes in the distribution of the data over time



Example: Changes in the statistical properties of retail data could be due to factors such as seasonality, sales and promotions or economic cycles

Appendix