

LECTURER: TAI LE QUY

# DATA SCIENCE

TOPIC OUTLINE

Introduction to Data Science

1

Use Cases and Performance Evaluation

2

Data Preprocessing

3

Processing of Data

4

Selected Mathematical Techniques

5

Selected Artificial Intelligence Techniques

6

## UNIT 4

# PROCESSING OF DATA



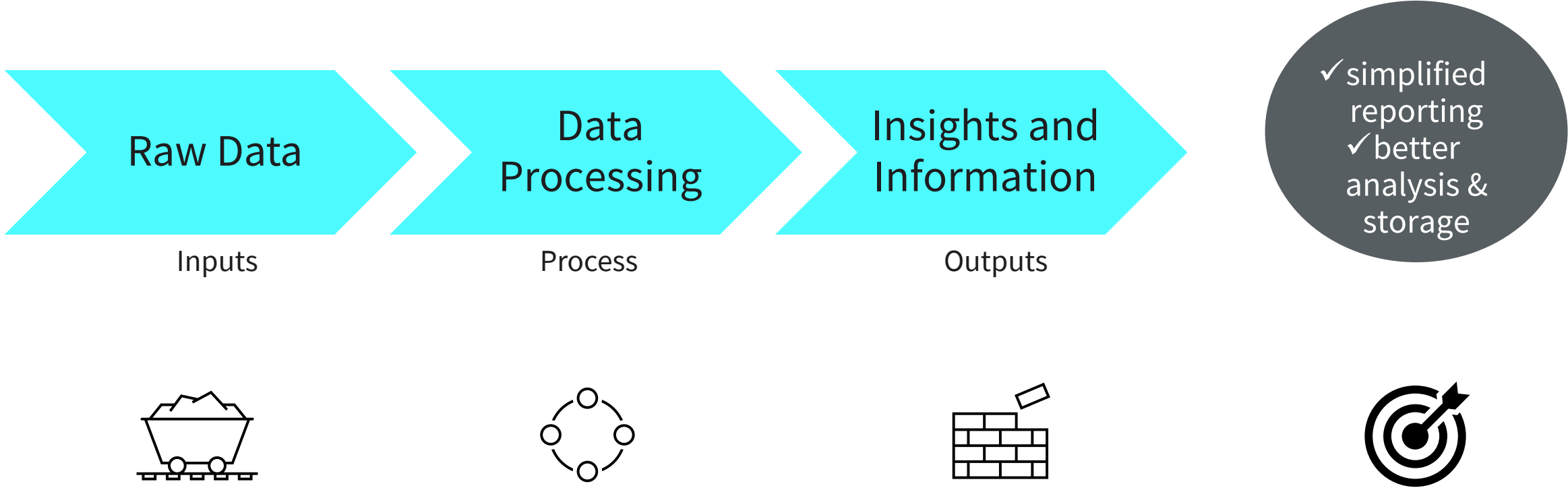
On completion of this unit, you will have learned ...

- the concepts of data, information, and data processing.
- the stages and cycles of data processing.
- the different methods and types of data processing.
- the output forms and file formats for processed data.

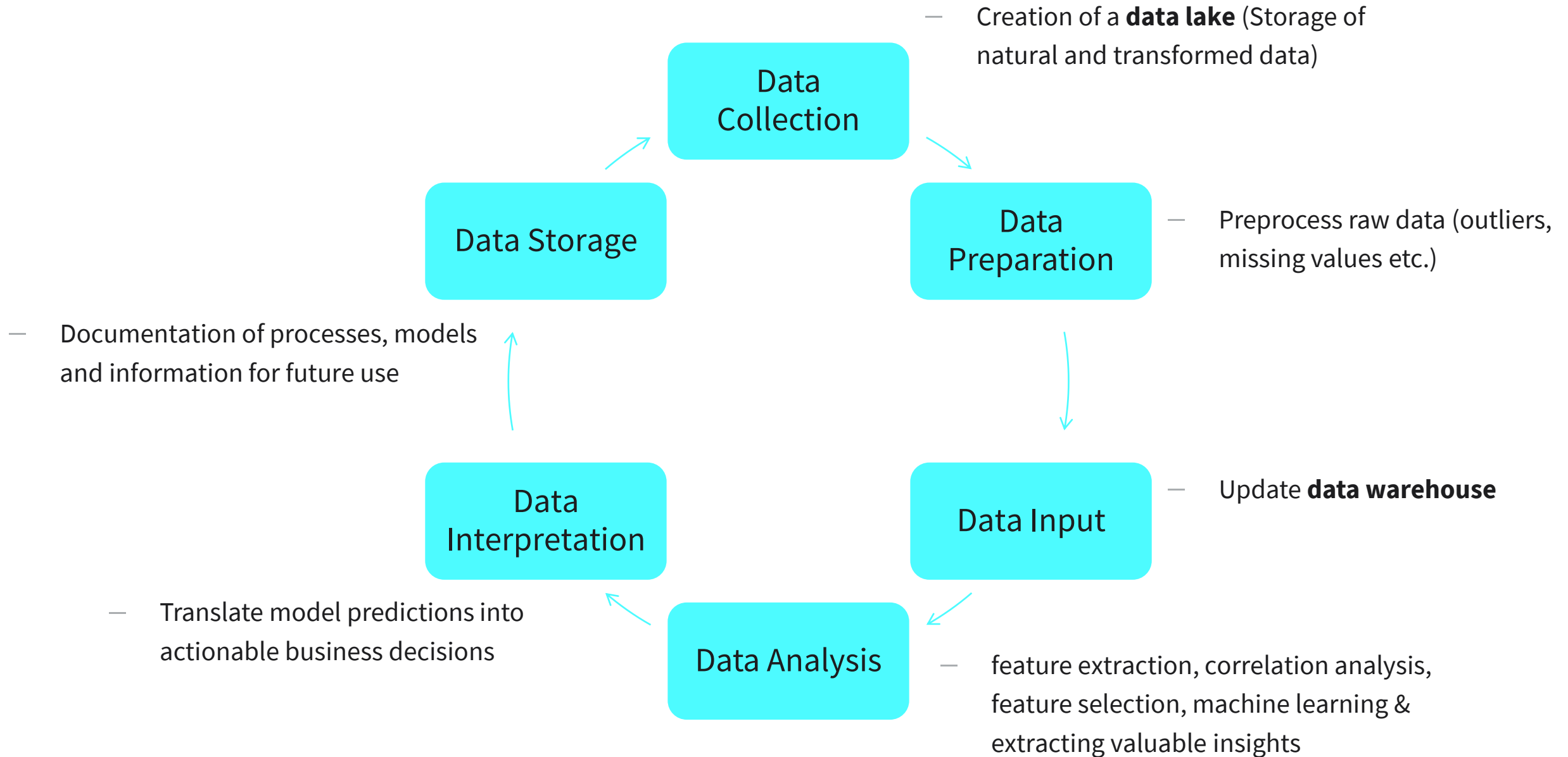


1. Explain why data processing is important.
2. In what way benefit data science projects from data processing.
3. Describe the five types of electronic data processing.

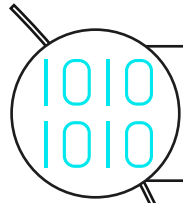
DATA PROCESSING INTRODUCTION



## DATA PROCESSING CYCLE



## ELECTRONICAL DATA PROCESSING



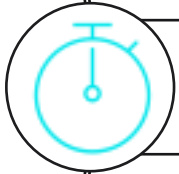
### **Batch**

split data into batches to permit sequential processing (mostly offline)



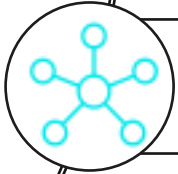
### **Online**

make use of internet connections



### **Real-time**

immediate response to requests



### **Distributed**

multiple remote workstations connected to a large server



### **Time-sharing**

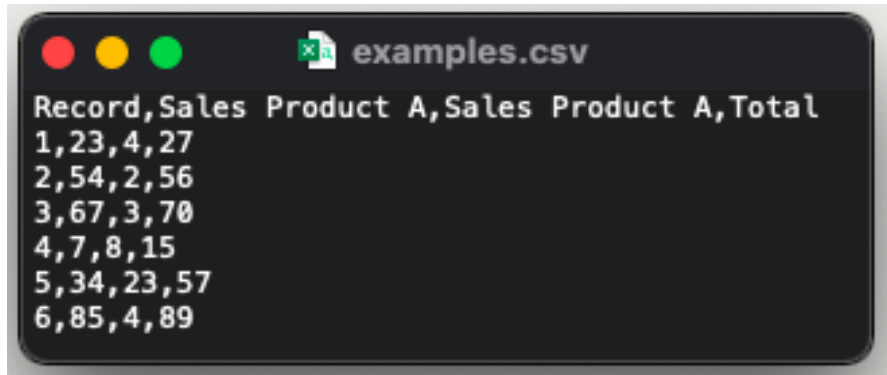
computing unit is utilized by multiple users



## OUTPUT FORMATS OF PROCESSED DATA

### CSV (comma-separated-value)

- row-based: every line represents one record
- features are separated by comma

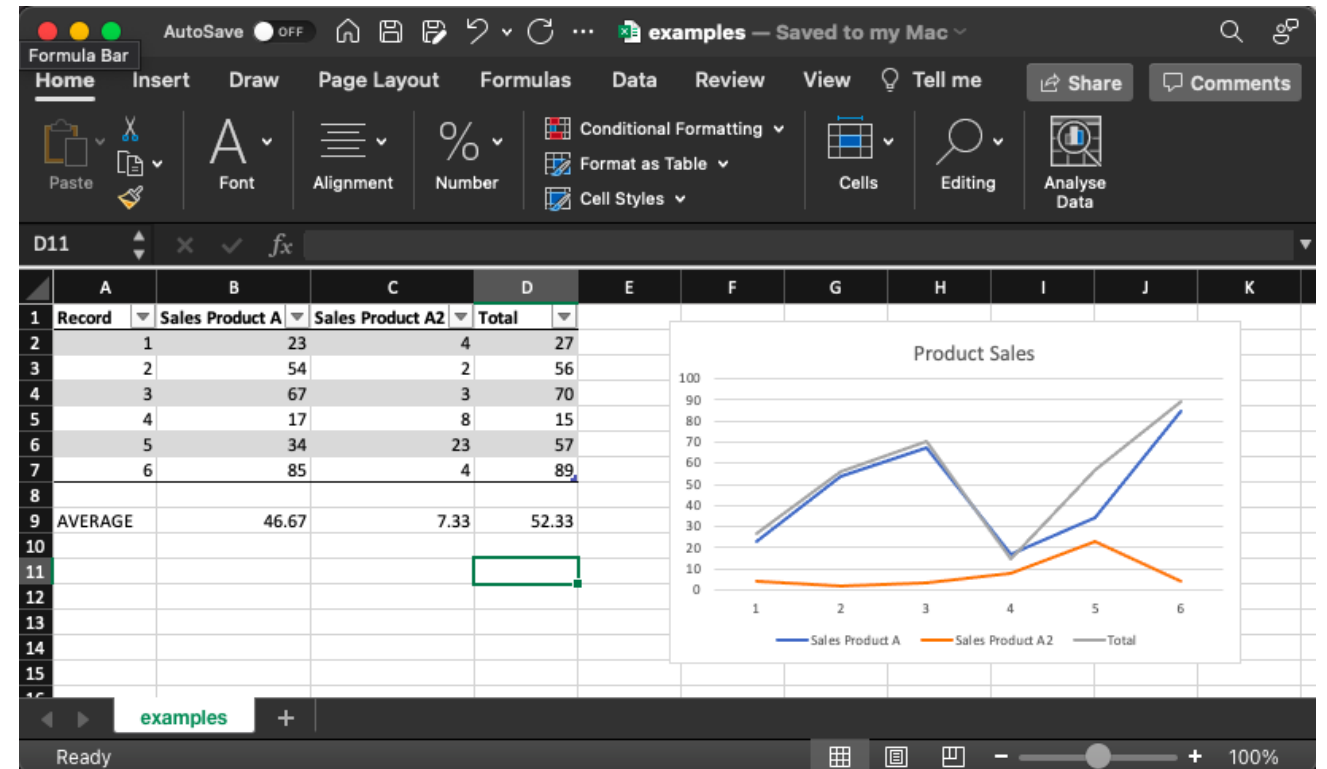


The screenshot shows a text editor window titled 'examples.csv'. The content is a CSV file with 6 rows of data. The first row is a header: 'Record,Sales Product A,Sales Product A,Total'. The subsequent rows contain numerical data. The last row is an average calculation: 'AVERAGE,46.67,7.33,52.33'.

Record	Sales Product A	Sales Product A	Total
1	23	4	27
2	54	2	56
3	67	3	70
4	7	8	15
5	34	23	57
6	85	4	89
AVERAGE	46.67	7.33	52.33

### XLS (Excel spreadsheet)

- tabular format of records and features
- possibility to add graphs, computations



## OUTPUT FORMATS OF PROCESSED DATA

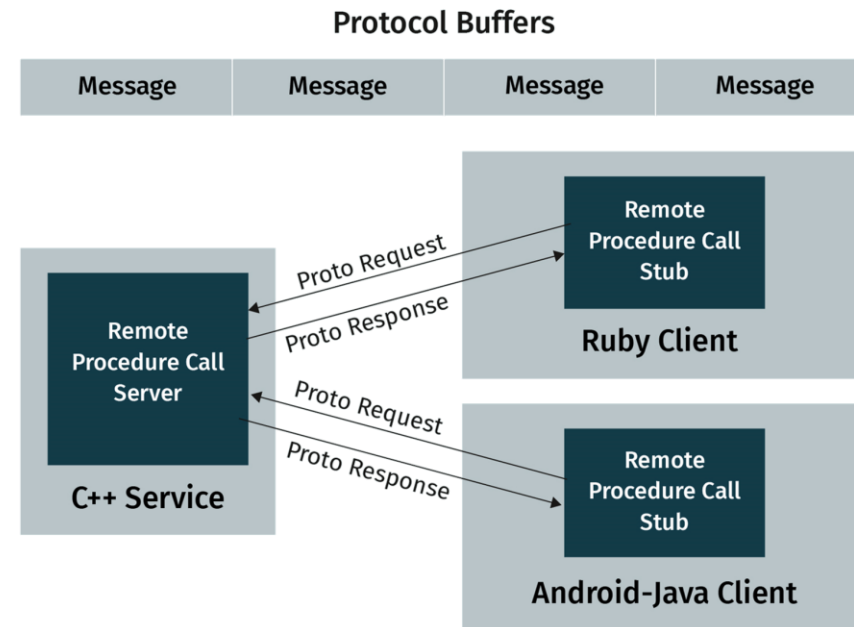
### XML (extensible markup language)

- structured, non-tabular data written as text with annotations

```
<student>
  <name>
    <firstname> John </firstname>
    <lastname> Miller</lastname>
  </name>
  <birthdate> 1999-12-12</birthdate>
</student>
<student>
  <name>
    <firstname> Alice </firstname>
    <lastname> Doe</lastname>
  </name>
  <birthdate> 1987-01-06</birthdate>
</student>
```

### Protobuf (protocol buffers)

- reduced XML version



```
// polyline.proto
syntax = "proto2";

message Point {
  required int32 x = 1;
  required int32 y = 2;
  optional string label = 3;
}

message Line {
  required Point start = 1;
  required Point end = 2;
  optional string label = 3;
}

message Polyline {
  repeated Point point = 1;
  optional string label = 2;
}
```

Example of protobuf

OUTPUT FORMATS OF PROCESSED DATA

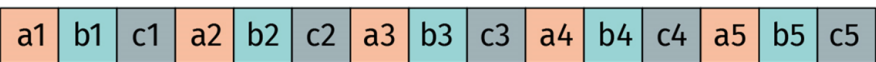
Apache Parquet

- column-based file format

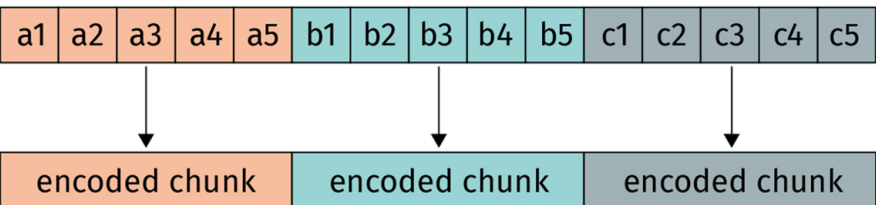
Logical table representation

a	b	c
a1	b1	c1
a2	b2	c2
a3	b3	c3
a4	b4	c4
a5	b5	c5

Row layout



Column layout



JSON (Java script object notation)

- a list of key-value pairs

```
examples.json
{
  "firstName": "John",
  "lastName": "Doe",
  "gender": "male",
  "age": 28,
  "address": {
    "streetAddress": "101",
    "city": "San Diego",
  },
  "phoneNumbers": [
    { "type": "home", "number": "7349282382" },
    { "type": "mobile", "number": "7349282382" }
  ]
}
```



You have learned ...

- the concepts of data, information, and data processing.
- the stages and cycles of data processing.
- the different methods and types of data processing.
- the output forms and file formats for processed data.

**SESSION 4**

# **TRANSFER TASK**

## **Working in group:**

- Choose your domain
- Create a framework that helps data practitioners to choose the best sub-type of electronic data processing.
  - + Which questions should they ask themselves? Eg. Data format, time requirement, etc.
  - + Potential data science tasks

**TRANSFER TASK  
PRESENTATION OF THE RESULTS**

Please present your  
results.

The results will be  
discussed in plenary.





1. In which step is data with missing values handled?
  - a) feature selection
  - b) machine learning
  - c) correlation analysis
  - d) data pre-processing





2. The data provided in this format, `<img fig="Alice.jpg" tag="Alice" />` , represents the...

- a) SQL data format.
- b) XLS data format.
- c) XML data format.
- d) CSV data format.



3. The patterns and relationships among data elements are defined as ...

- a) data.
- b) properties.
- c) information.
- d) features.

© 2021 IU Internationale Hochschule GmbH

This content is protected by copyright. All rights reserved.

This content may not be reproduced and/or electronically edited, duplicated, or distributed in any kind of form without written permission by the IU Internationale Hochschule GmbH.