

Week2: Data Acquisition & NumPy

Chantri Polprasert

Dept. of ICT,AIT

Agenda

- Data acquisition
- Sources of data
- Data structure
 - Structured
 - Semi-structured
 - Unstructured
- Sample data format



Data Acquisition

The First step in Data Science process



Identify data sets

Retrieve data

Query data

What is Data Acquisition?

- Data acquisition is the processes for **bringing data that has been created by sources** inside or outside the organization, into the organization, for production use
- Data acquisition has been understood as **the process of gathering, filtering, and cleaning data** before the data is put in a data warehouse or any other storage solution.
- There are many ways to get a dataset like configuring an **API, internet, database**, etc.



Data Discovery

- Search different sources of data
- Capture structured and unstructured data



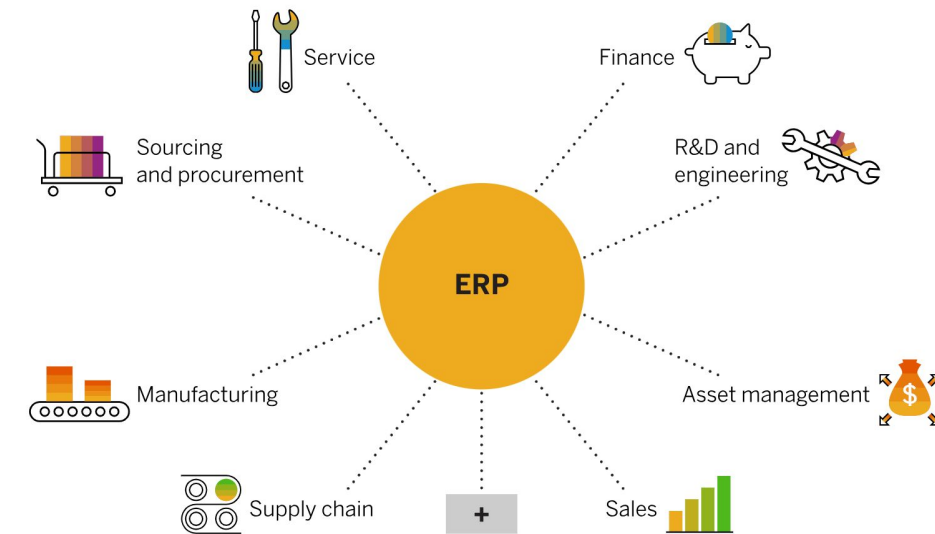
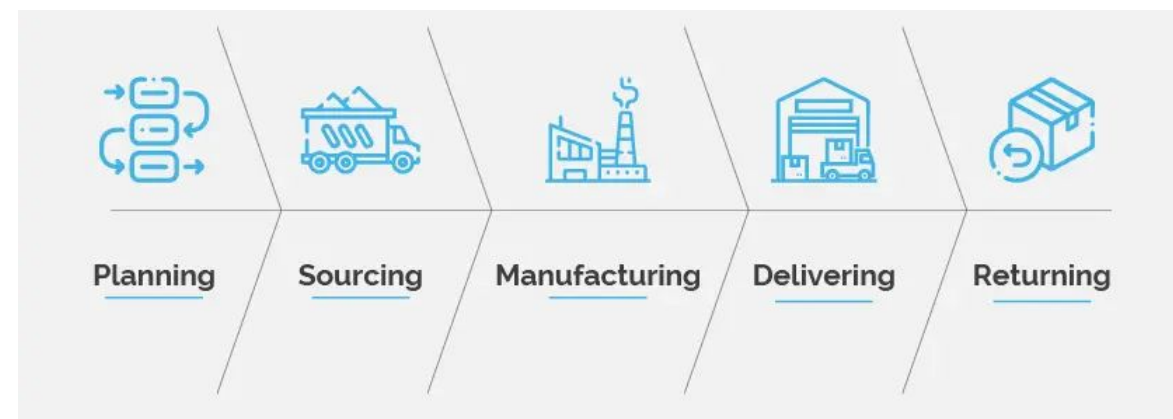
Data Preparation

- Convert data to a common format

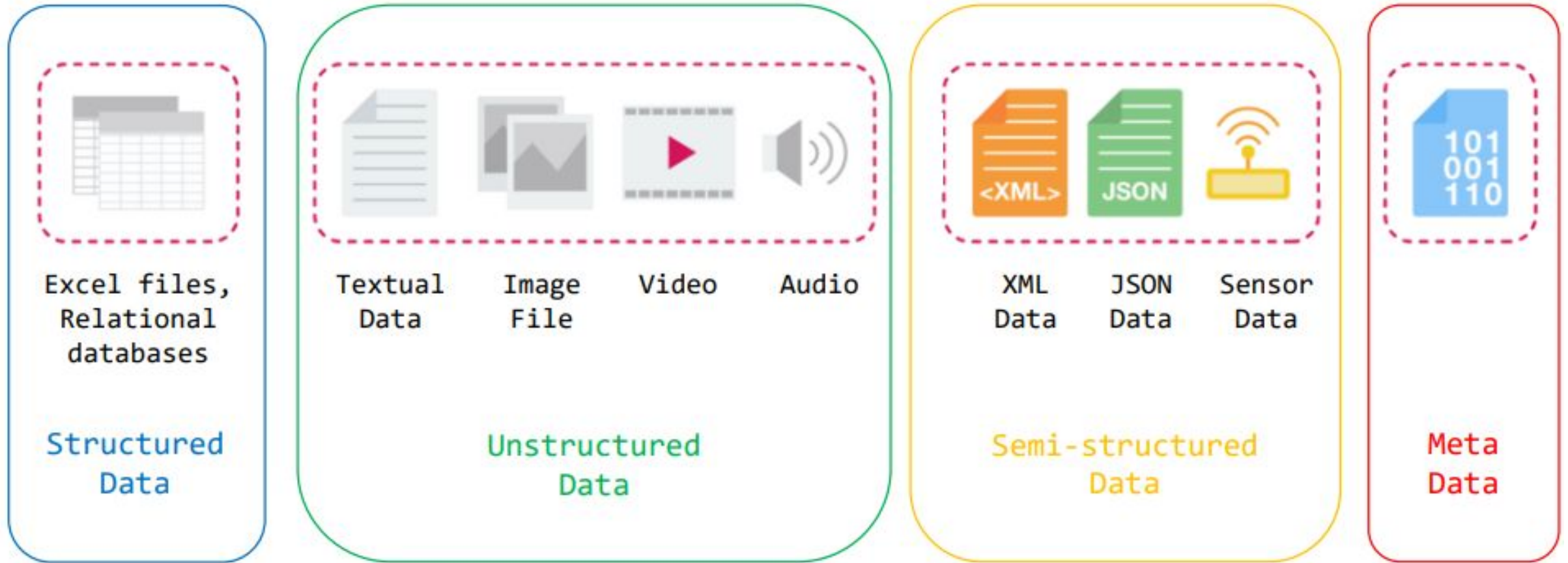
<https://intellipaat.com/mediaFiles/2015/11/Picture9.png>

Sources of data

- Enterprise Software
 - Enterprise Resource Planning (ERP) (SAP Business One, Odoo, Oracle Netsuite)
 - Supply Chain Management (SCM) (SAP, Netsuite)
 - Customer Relationship Management (CRM) (Salesforce, Dynamics 365)
- In-house Website/Application
 - Mobile App (customer activities)
 - Web Application (web log, live chat)
- External Sources
 - Social Data (FB, Twitter, Tiktok, Ig etc..)
 - Partners
 - Other Third-Party Data (Open data, weather etc..)



Three different data structures



Metadata is data about data. It provides additional information about a specific set of data.

Example, metadata of photographs could describe when and where the photos were taken.

Structured Data

- This type of data is highly organized and stored in a predefined format. The most typical examples of structured data are relational databases and spreadsheets.
- Fixed schema

Each field is discrete and can be accessed separately or jointly along with data from other fields.

	A	B	C	
1	students			
2	student_id	student_name	gpa	
3	2538	John Smith	3.5	
4	2541	Mary Sue	4	
5	2542	Tony Stark	3.8	
6				

This is an example of a Google spreadsheet where each column represents a different type of information (student ID, student name, and GPA) and each row represents one student and information related to them.

Pros & Cons?

CSV (Comma-Separated Values)



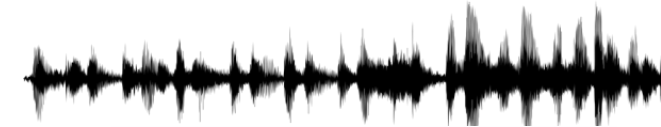
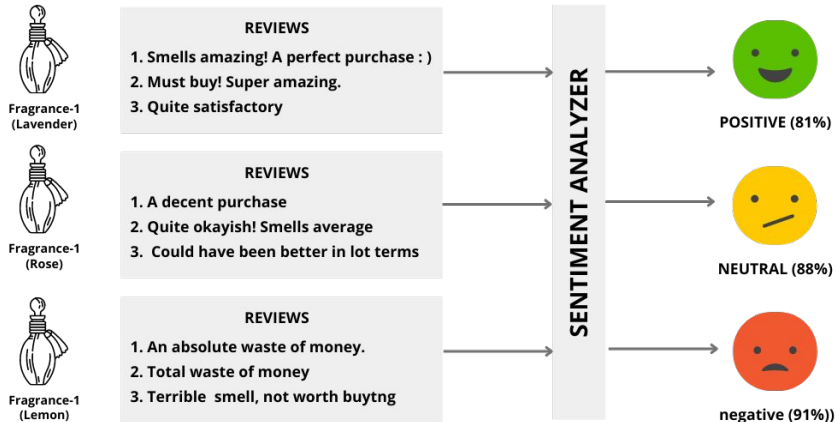
- A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values.
 - Each line of the file is a data record.
 - Each record consists of one or more fields, separated by commas.
- A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.
- Different field delimiters can be used
 - **Semicolon, tab, and space**
 - Files are often even given a .csv extension despite the use of a non-comma field separator.

A screenshot of a Notepad window titled 'contacts.csv - Notepad'. The window displays a CSV file with the following content:

```
File Edit Format View Help
First Name,Middle Name,Last Name,Title,Suffix,Initials,Web Page,Gender,Birthda
by,Billing Information,Directory Server,Sensitivity,Priority,Private,Categorie
Bob,,Smith,,,,,,,,,,,,,bob@example.com,,,,,,,,,123-456-7890,,,,,,,,,
Mike,,Jones,,,,,,,,,,,,,mike@example.com,,,,,,,,,098-7654-321,,,321 Fake Avenue,32
```

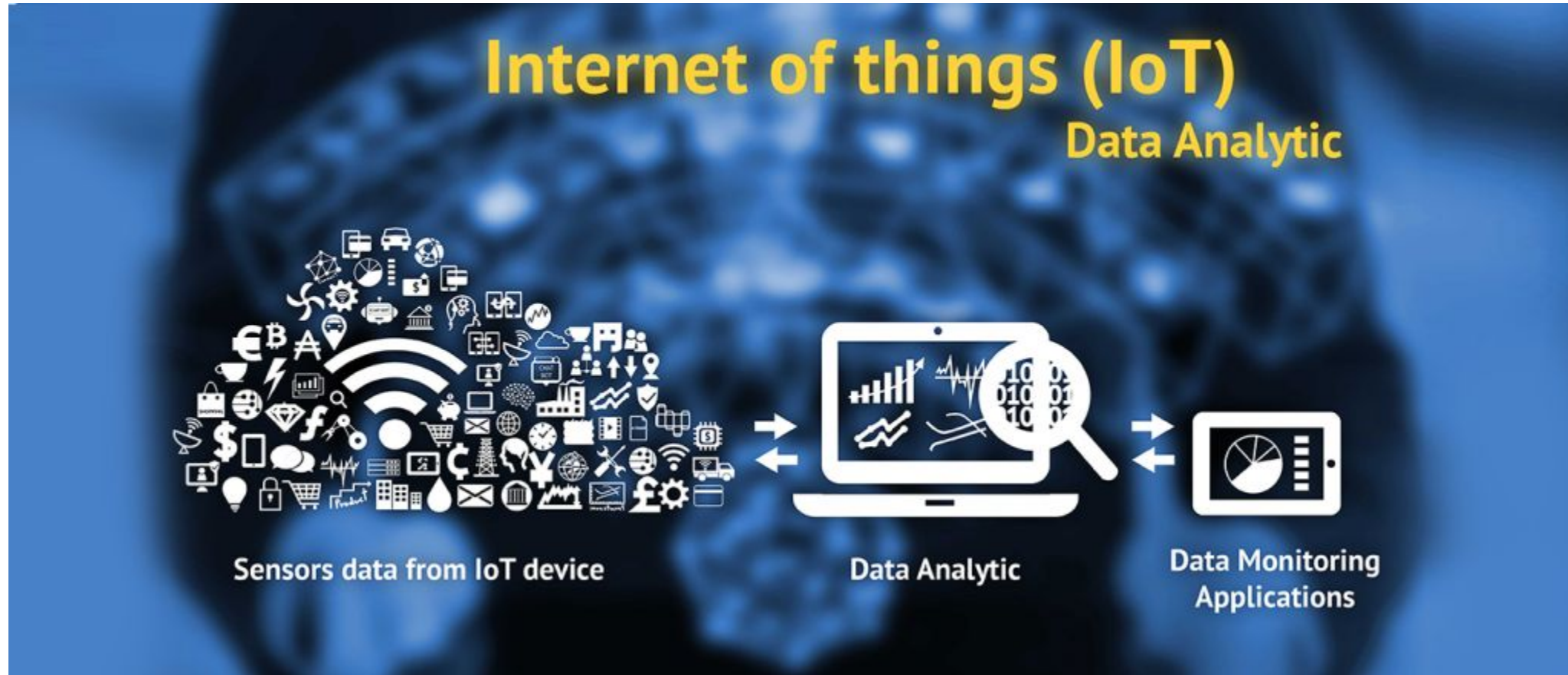

Unstructured Data

- Unstructured data is all data that isn't organized in a pre-defined manner.
 - Human generated data: text documents, emails, social media posts, images, and videos
 - Machine generated data: IoT sensor data, log, GPS
- No schema
- How to analyze them?



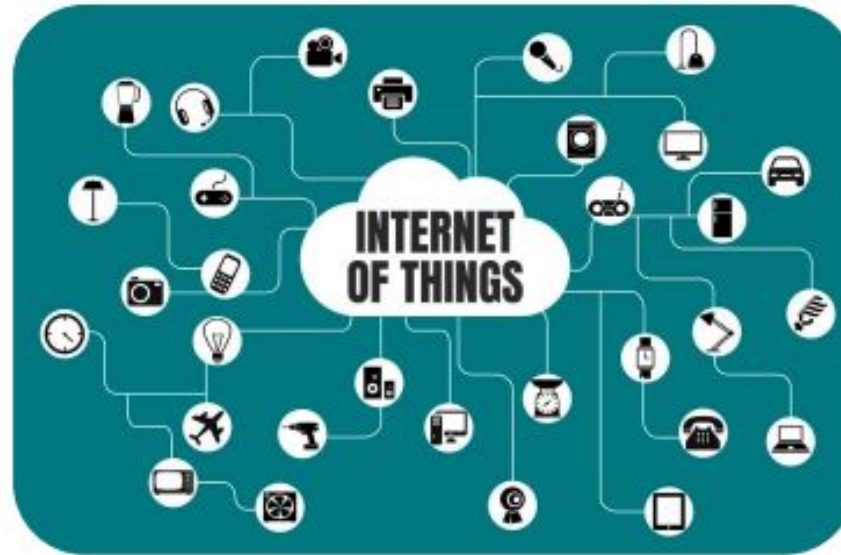
<https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/>

IoT as a Data Source



What is the Internet of Things (IoT)?

Describes the network of physical objects—“things”—that are embedded with sensors, software, and other technologies for the purpose of connecting and exchanging data with other devices and systems over the internet.

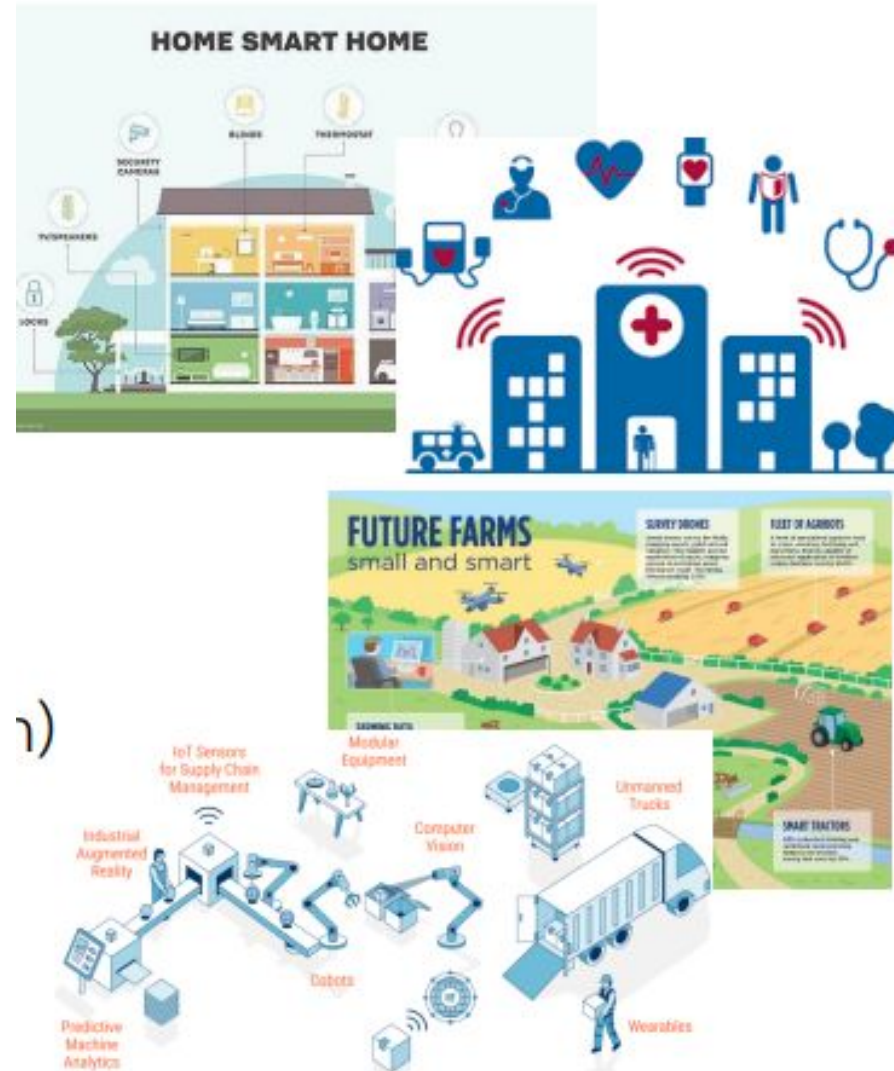


Names of related technologies:

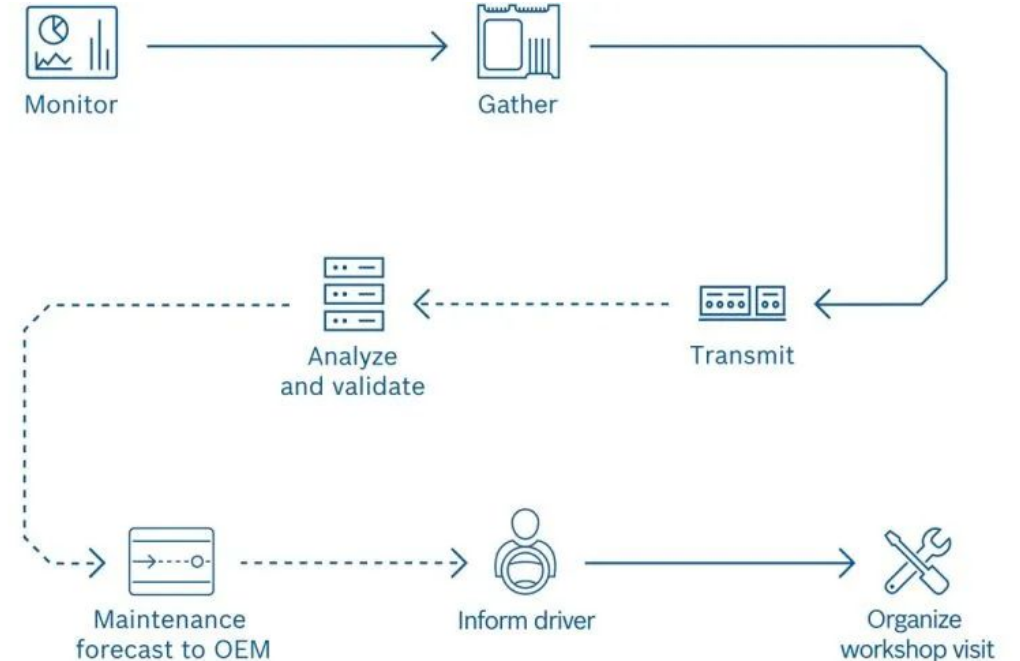
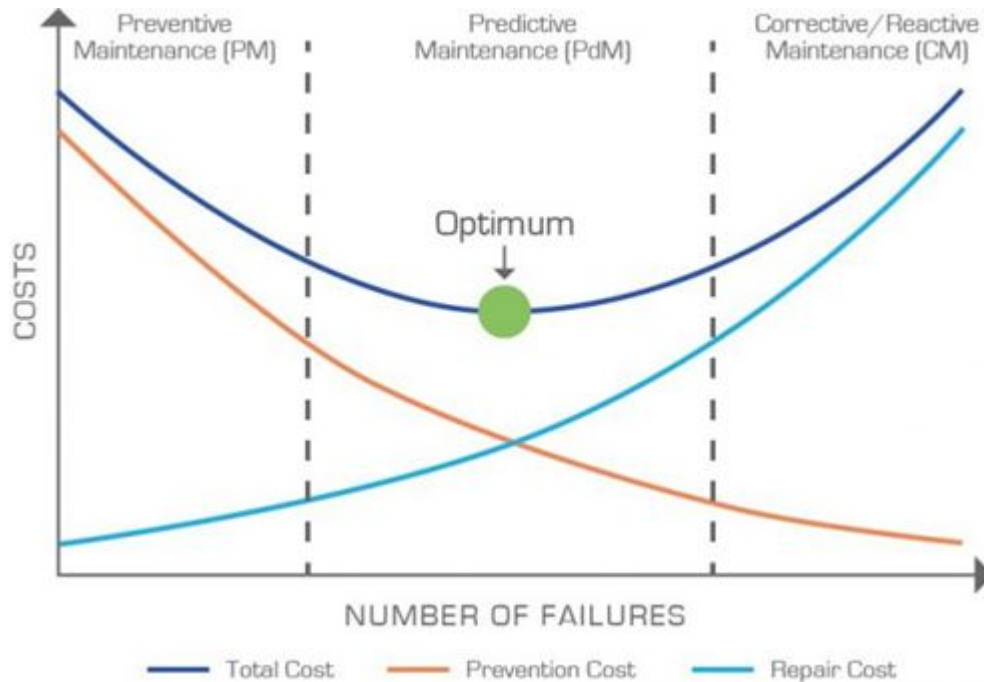
- M2M (Machine to Machine)
- Internet of Everything
- World Size Web
- Cyber-Physical System (CPS)
- Web of Things (WoT)

Sample IoT applications

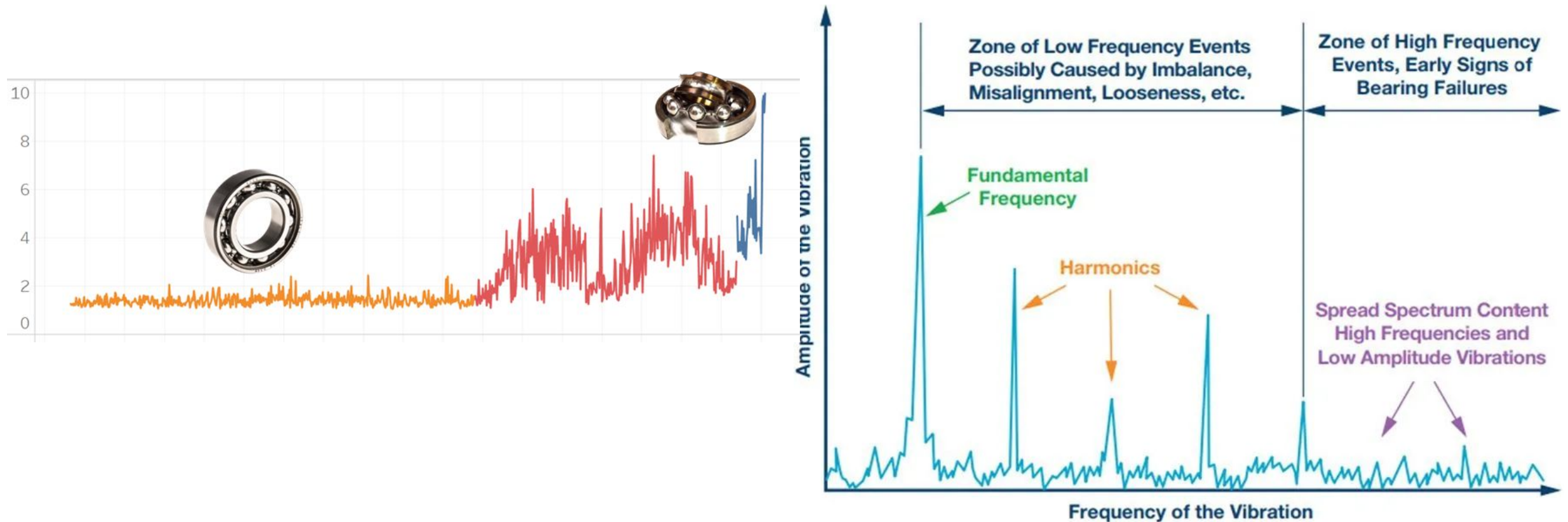
- Smart Homes
- Smart City
- Smart Farm
- Smart Health Care
- Smart Transportation
- Industry 4.0
- Others



Predictive Maintenance: Analyzing vibration signal



Predictive Maintenance: Analyzing vibration signal



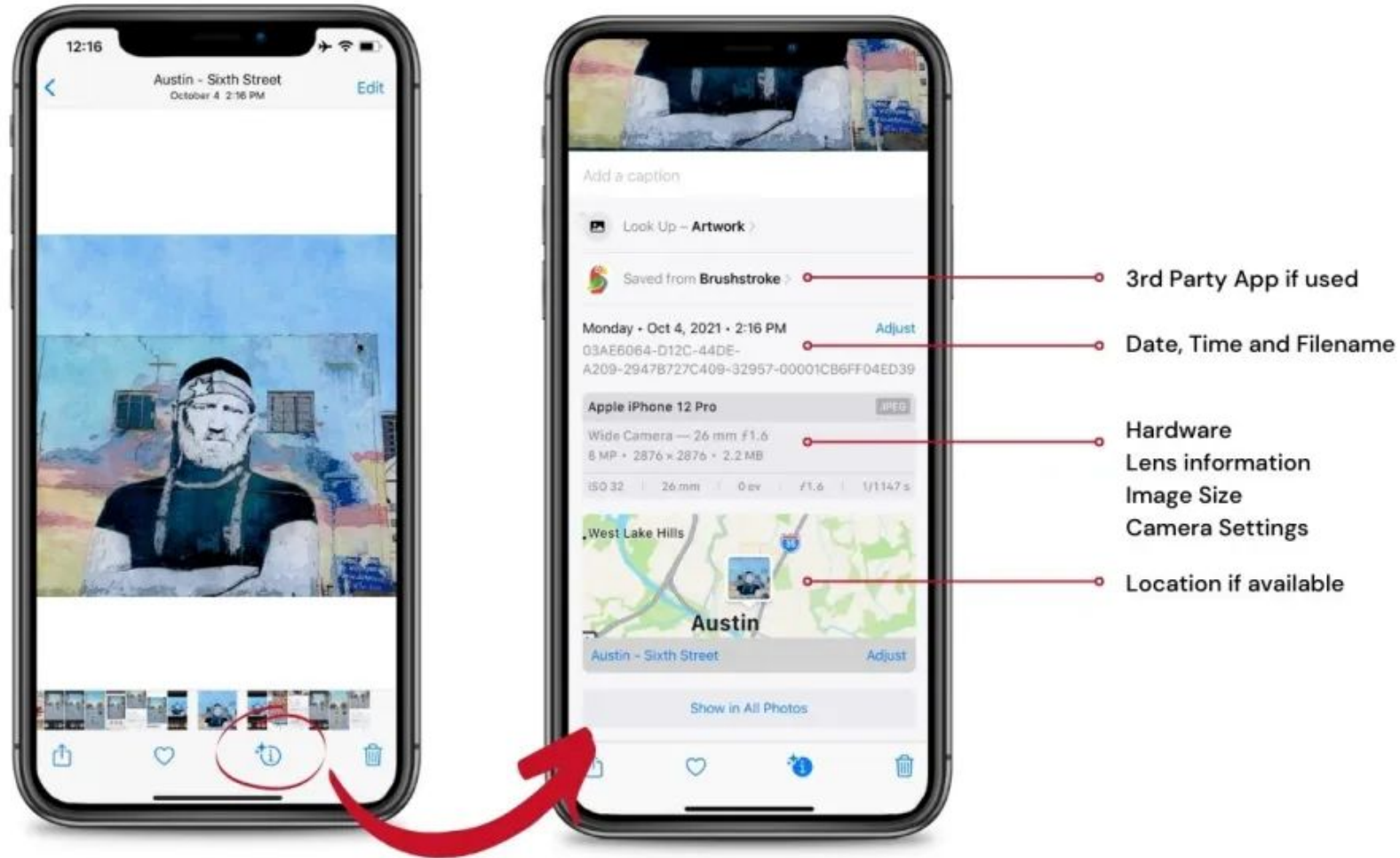
Semi-Structured Data

- This type of data falls in between structured and unstructured data.
- There are **tags or other kinds of markers** that separate semantic elements and define the hierarchy of records and fields within the data.
- Examples of semi-structured data include JSON and XML files.
- Loose schema

```
1
2 students: [
3   {
4     "student_id" : 2538,
5     "student_name" : "John Smith",
6     "gpa" : 3.5
7   },
8   {
9     "student_id" : 2541,
10    "student_name" : "Mary Sue",
11    "gpa" : 4
12  },
13  {
14    "student_id" : 2542,
15    "student_name" : "Tony Stark",
16    "gpa" : 3.8
17  }
18 ]
19
```

This is an example of a .json file. Data is represented in name-value pairs separated by commas, and curly braces indicate different objects (in this case, students) within the array

Image with metadata



Email

Original Message

Message ID	<OL8MU0S0PKU4.YEN5B2VDZRZJ3@win-645k42c5rie>
Created at:	Fri, Aug 11, 2023 at 2:27 PM (Delivered after 2 seconds)
From:	Stan Lee Williams <windenergy@pulsusglobevents.com>
To:	"chantri@ait.ac.th" <chantri@ait.ac.th>
Subject:	Proposal for being an Organizing Committee Member for the Wind and Renewable Energy 2023 Conference
SPF:	PASS with IP 103.181.21.55 Learn more
DKIM:	'PASS' with domain pulsusglobevents.com Learn more
DMARC:	'PASS' Learn more

-----iiLIEP4g1+facscf1UhEog==

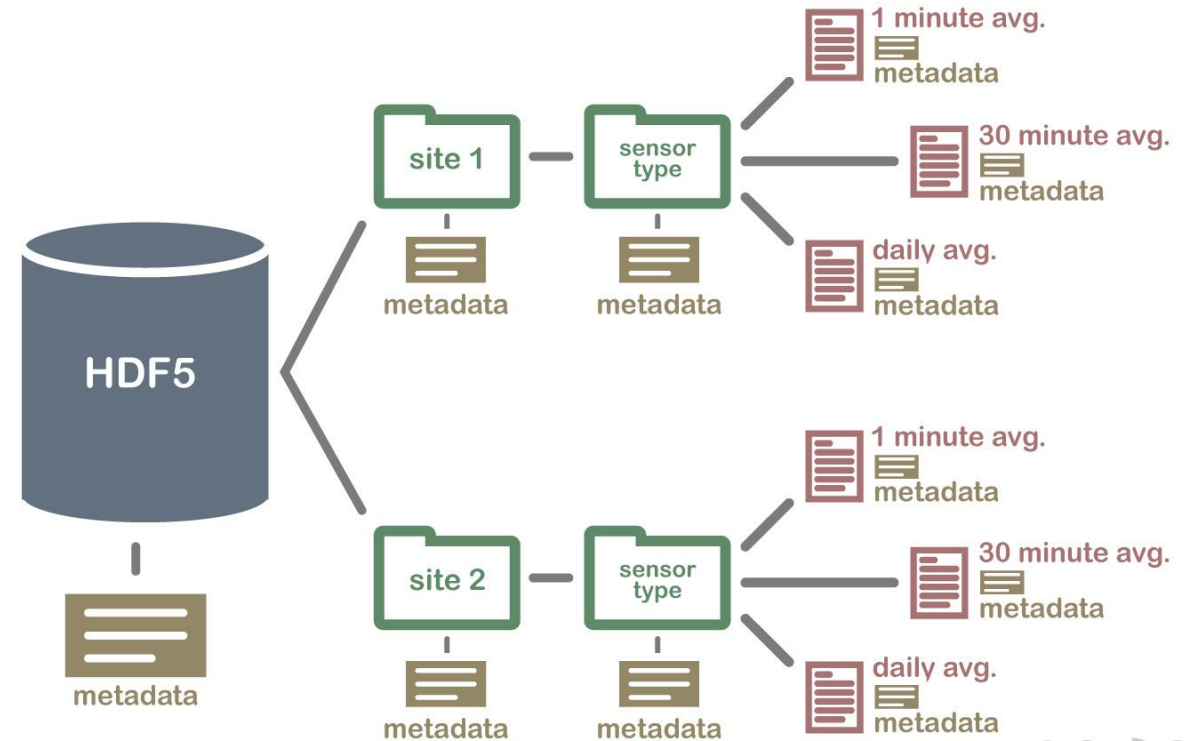
Content-Type: text/plain; charset=utf-8

Content-Transfer-Encoding: quoted-printable

Emailidea TemplatesDear Dr. Chantri PolprasertI hope this email finds you well. I am excited to reach out to you with an opportunity to play a vital role in organizing our upcoming Wind & Renewable Energy 2023. Your passion for Renewable Energy and your experience in Renewable Industry make you an ideal candidate to be a part of our esteemed Organizing Committee.Wind & Renewable Energy 2023 aims to bring Industry leaders, Experts together, and we believe that with your expertise and dedication, we can create a truly impactful and unforgettable experience for all participants.As a member of the Organizing Committee, you will have the opportunity to contribute to the event's planning, decision-making, and execution. Your responsibilities may include:Collaborating with fellow committee members to develop event strategies and action plans.Chair a Session on the topic of your interest.Reviewing the Speaker's Abstracts for the conference Assisting in the selection of=

Data Meant To Be Read by Machines

- Text File (text, log, csv, json, xml)
- Binary Data format
 - Excel File
 - Hierarchical Data Format file (HDF)
 - Hierarchical
 - HIGH PERFORMANCE
 - EXTENSIBLE
 - PORTABILITY
 - Web API (JSON, XML)
 - Database



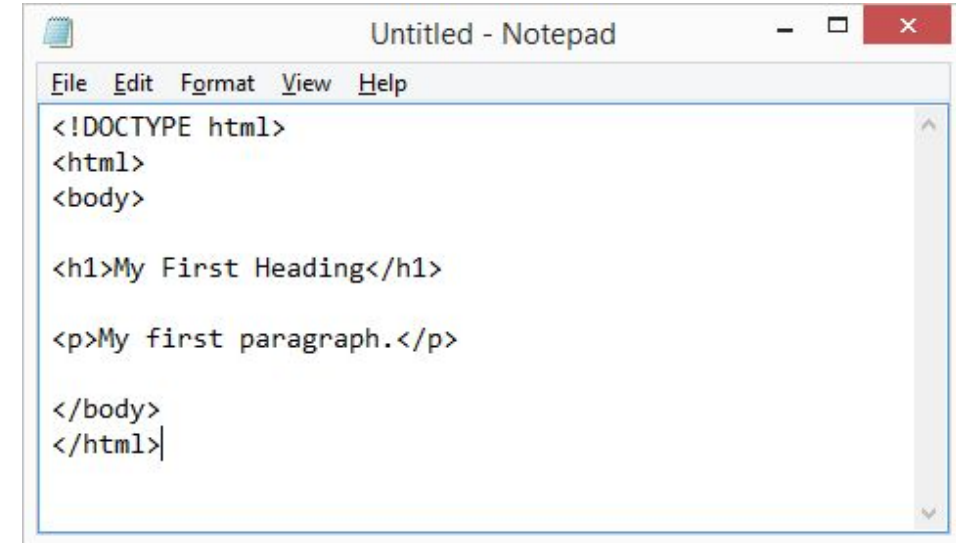
How to obtain a summary of the 2024 Summer Olympic Medal Table?

2024 Summer Olympics medal table^{[157][B]}

Rank ↕	NOC ↕	Gold ↕	Silver ↕	Bronze ↕	Total ↕
1	 United States	40	44	42	126
2	 China	40	27	24	91
3	 Japan	20	12	13	45
4	 Australia	18	19	16	53
5	 France*	16	26	22	64
6	 Netherlands	15	7	12	34
7	 Great Britain	14	22	29	65
8	 South Korea	13	9	10	32
9	 Italy	12	13	15	40
10	 Germany	12	13	8	33
11–91	<i>Remaining NOCs</i>	129	138	194	461
Totals (91 entries)		329	330	385	1044

HTML

- Stands for Hypertext Markup Language
- Computer language used to create web pages
- HTML file = text file containing markup tags such as `<p>`, `</p>`
- Tags tell Web browser how to display a page
- Can have either `*.htm` or `*.html` file extension

A screenshot of a Notepad window titled "Untitled - Notepad". The window has a menu bar with "File", "Edit", "Format", "View", and "Help". The text area contains the following HTML code:

```
<!DOCTYPE html>
<html>
<body>

<h1>My First Heading</h1>

<p>My first paragraph.</p>

</body>
</html>
```

HTML Elements

- Tags are the elements that create the components of a page
- Tags surrounded by angle brackets < >
- Usually come in pairs
 - Example: Start tag <p> and end tag </p>
- Stuff between is called “element content”
- Tags are not case sensitive

Structure of a Web Page

- All Web pages share a common structure
- All Web pages should contain a pair of <HTML>, <HEAD>, <TITLE>, and <BODY> tags

```
<HTML>
<HEAD>
<TITLE> Example </TITLE>
</HEAD>
<BODY>
    This is where you
    would include the text
    and images on your Web
    page.
</BODY>
</HTML>
```


Sample HTML code

```
<!DOCTYPE html>
<html>
<body>

<h1>Welcome to CS, AIT.</h1>
<p>This field of study fosters high-level teaching and re
  One focus is on educating educators who can, in turn, e
  with the faculty particularly active in artificial inte
  in computer architectures, object orientation, neural n
</p>

</body>
</html>
```

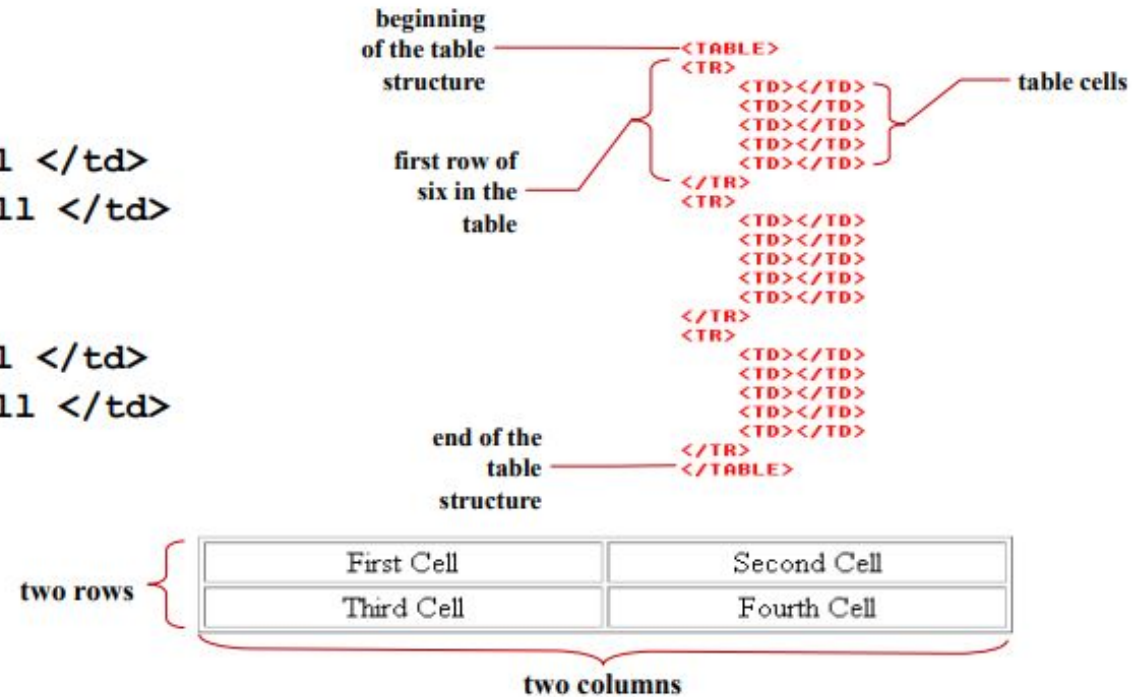
Welcome to CS, AIT.

This field of study fosters high-level teaching and research in computer science and aims to meet the growing regional demand for persons skilled in various aspects of computing. One focus is on educating educators who can, in turn, effectively disseminate knowledge and skills to more people. The core curriculum in computer science covers all aspects of computing, with the faculty particularly active in artificial intelligence, software engineering, networking, and information systems. The program also endeavors to enhance teaching and research activities in computer architectures, object orientation, neural networks, multimedia, and other rapidly-evolving areas in computer science.

HTML Tables Using the <table>, <tr>, and <td> Tags

- Graphical tables are enclosed within a two-sided <table> tag that identifies the start and ending of the table structure.
- Each row of the table is indicated using a two-sided <tr> (for table row).

```
<table>
  <tr>
    <td> First Cell </td>
    <td> Second Cell </td>
  </tr>
  <tr>
    <td> Third Cell </td>
    <td> Fourth Cell </td>
  </tr>
</table>
```



- Within each table row, a two-sided <td> (for table data) tag indicates the presence of individual table cells.



NBP Web API

Currency exchange rates and gold prices in the XML and JSON formats

```
▼<ArrayOfExchangeRatesTable xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  ▼<ExchangeRatesTable>
    <Table>A</Table>
    <No>155/A/NBP/2024</No>
    <EffectiveDate>2024-08-09</EffectiveDate>
    ▼<Rates>
      ▼<Rate>
        <Currency>bat (Tajlandia)</Currency>
        <Code>THB</Code>
        <Mid>0.1123</Mid>
      </Rate>
      ▼<Rate>
        <Currency>dolar amerykański</Currency>
        <Code>USD</Code>
        <Mid>3.9604</Mid>
      </Rate>
      ▼<Rate>
        <Currency>dolar australijski</Currency>
        <Code>AUD</Code>
        <Mid>2.6101</Mid>
      </Rate>
      ▼<Rate>
        <Currency>dolar Hongkongu</Currency>
        <Code>HKD</Code>
        <Mid>0.5080</Mid>
      </Rate>
      ▼<Rate>
        <Currency>dolar kanadyjski</Currency>
        <Code>CAD</Code>
        <Mid>2.8841</Mid>
      </Rate>
      ▼<Rate>
        <Currency>dolar nowozelandzki</Currency>
        <Code>NZD</Code>
        <Mid>2.3865</Mid>
      </Rate>
```

<http://api.nbp.pl/api/exchangerates/tables/A>

XML

- XML stands for eXtensible Markup Language.
- A **markup language** is developed in 1998 to meet the challenges of electronic publishing
- Tags are added to the document to provide the extra information.
- HTML tags tell a browser how to display the document but the XML tags give readers meaning of data
- Applications:
 - Data transfer between web servers (RSS)
 - Web Search
 - Computer applications (docx, pptx)
 - Web applications

```
<studentsList>
  <student id="1">
    <firstName>Greg</firstName>
    <lastName>Dean</lastName>
    <certificate>True</certificate>
    <scores>
      <module1>70</module1>
      <module12>80</module12>
      <module3>90</module3>
    </scores>
  </student>
  <student ind="2">
    <firstName>Wirt</firstName>
    <lastName>Wood</lastName>
    <certificate>True</certificate>
    <scores>
      <module1>80</module1>
      <module12>80.2</module12>
      <module3>80</module3>
    </scores>
  </student>
</studentsList>
```

Example of an XML document

```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
  <book category="children">
    <title>Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>
  <book category="web">
    <title>Learning XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
  </book>
</bookstore>
```

- An element can contain:
 - Text
 - Attributes
 - Other elements
 - Mix of the above
- `<title>`, `<author>`, `<year>` and `<price>` have text content because they contain text (e.g. 39.95)
- `<bookstore>` and `<book>` have **element** contents because they contain **elements**.
- `<book>` has an **attribute** (category = "children")

Difference between HTML and XML

XML	HTML
<ul style="list-style-type: none">• User Defined Tags• Data Driven• Different for different applications	<ul style="list-style-type: none">• Predefined tags• Presentation driven• Fixed meaning and aware by the web browser
<pre data-bbox="529 878 945 1068"><catalog> <book>MyBook </book> </catalog></pre>	<pre data-bbox="1567 886 2084 1049"><FORM> <input type=text> </form></pre>

```
{
  "numRecordings": "10704",
  "numSpecies": "761",
  "page": 1,
  "numPages": 22,
  "recordings": [
    {
      "id": "665584",
      "gen": "Pellorneum",
      "sp": "nigrocapitatum",
      "ssp": "",
      "group": "birds",
      "en": "Malayan Black-capped Babbler",
      "rec": "Jelle Scharringa",
      "cnt": "Thailand",
      "loc": "Khao Nor Chu Chi, Krabi",
      "lat": "7.9212",
      "lng": "99.2664",
      "alt": "70",
      "type": "song",
      "sex": "",
      "stage": "",
      "method": "field recording",
      "url": "https://xeno-canto.org/665584",
      "file": "https://xeno-canto.org/665584/download",
      "file-name": "XC665584-a04a.mp3",
      "sono": {
        "small": "https://xeno-canto.org/sounds/uploaded/KDTMIWOGNC/ffts/XC665584-small.png",
        "med": "https://xeno-canto.org/sounds/uploaded/KDTMIWOGNC/ffts/XC665584-med.png",
        "large": "https://xeno-canto.org/sounds/uploaded/KDTMIWOGNC/ffts/XC665584-large.png",
        "full": "https://xeno-canto.org/sounds/uploaded/KDTMIWOGNC/ffts/XC665584-full.png"
      }
    }
  ]
}
```



xeno-canto

Sharing wildlife sounds from around the world

[About](#)
[Explore](#)
[Upload Sounds](#)
[Forum](#)
[Mysteries](#)
[Articles](#)

Recordings from deep inside DR Congo



XC824126



0:00

0:16



Black-collared Lovebird *Agapornis swindernianus* · call
Deville Tanguy

Deville Tanguy spent nine months working on the **LuiKotale Bonobo Project** in a remote location in DR Congo. He managed to get a **good set of recordings** from an area from which XC still has very little material. Among these was a first recording of Black-collared Lovebird. Tanguy is also an accomplished photographer, as the picture, taken from the top of a tree, shows. There are also some **mystery recordings** in this set. So help out if you can!

<https://xeno-canto.org/api/2/recordings?query=cnt:thailand>

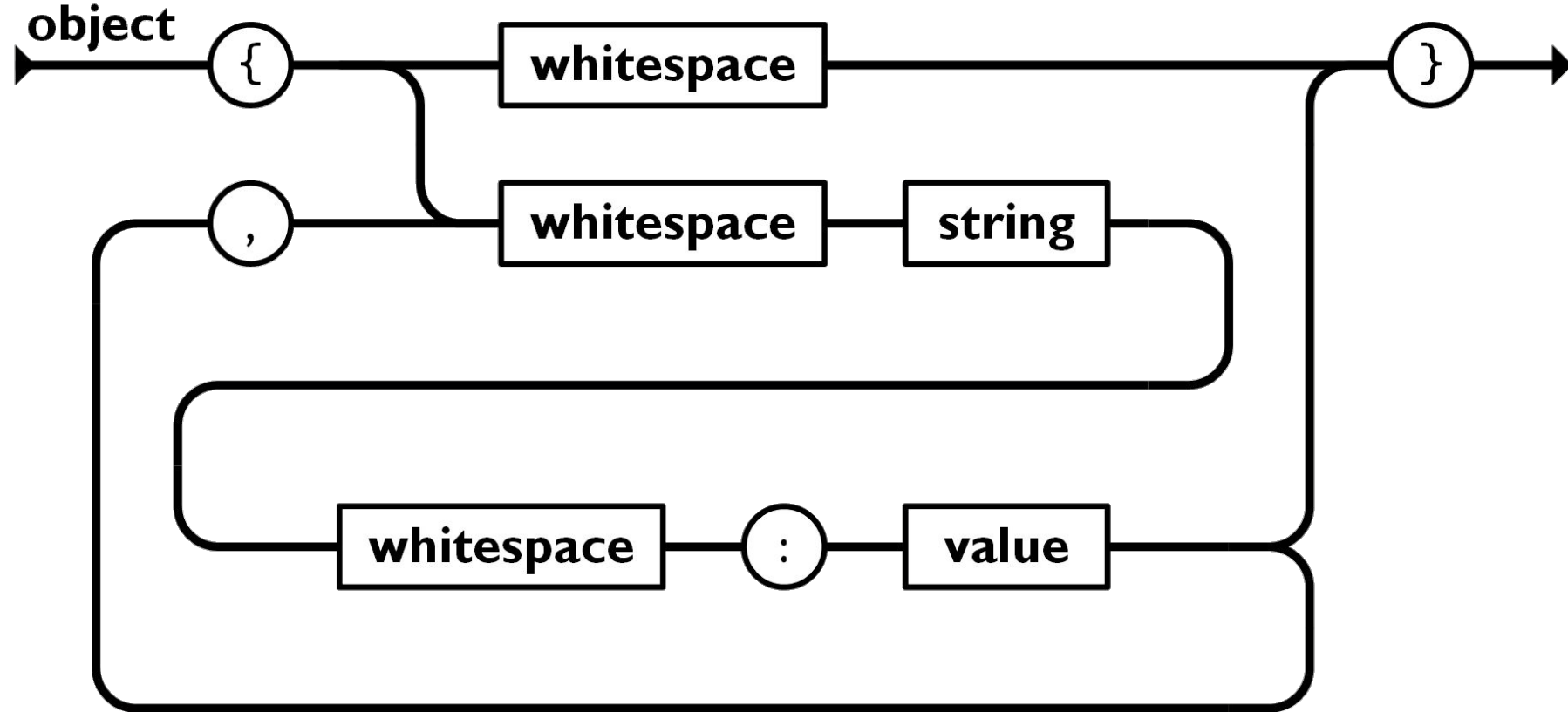
JSON (JavaScript Object Notation)

- A lightweight data-interchange format. It is easy for humans to read and write
- easy for machines to parse and generate
- Built on two structures:
 - A collection of name/value pairs. In various languages, this is realized as an object, record, struct, dictionary, hash table, keyed list, or associative array.
 - An ordered list of values. In most languages, this is realized as an array, vector, list, or sequence.
- Virtually all modern programming languages support them in one form or another.

JSON Syntax

- JSON objects start the object with “{” and end in with “}”
 - Members (properties), use pairs of “key : value”
- JSON arrays put the arrays between “[]”
- Elements put the values directly separated by commas

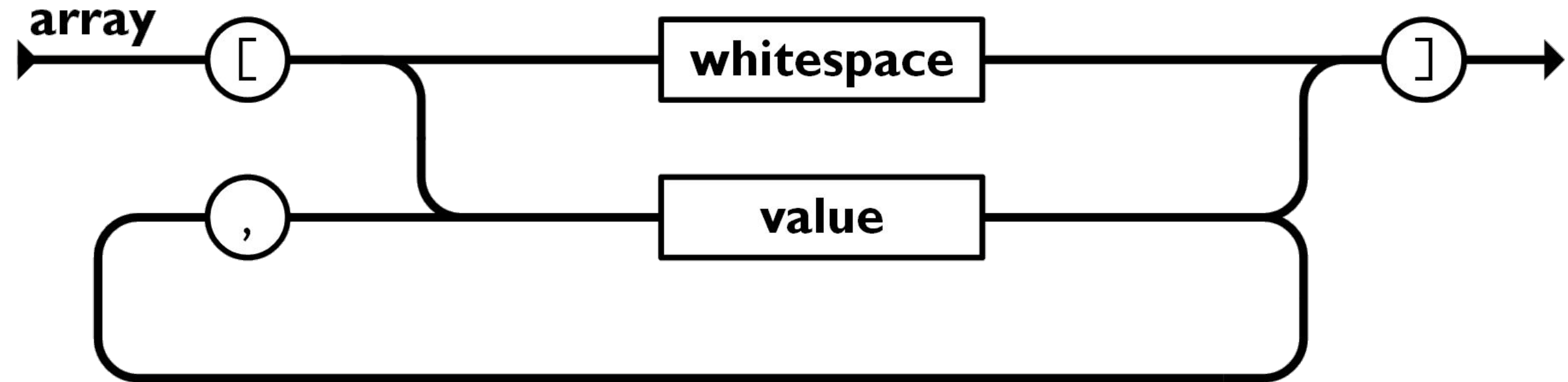
JSON Object



`{"title": "Openheimer"}`

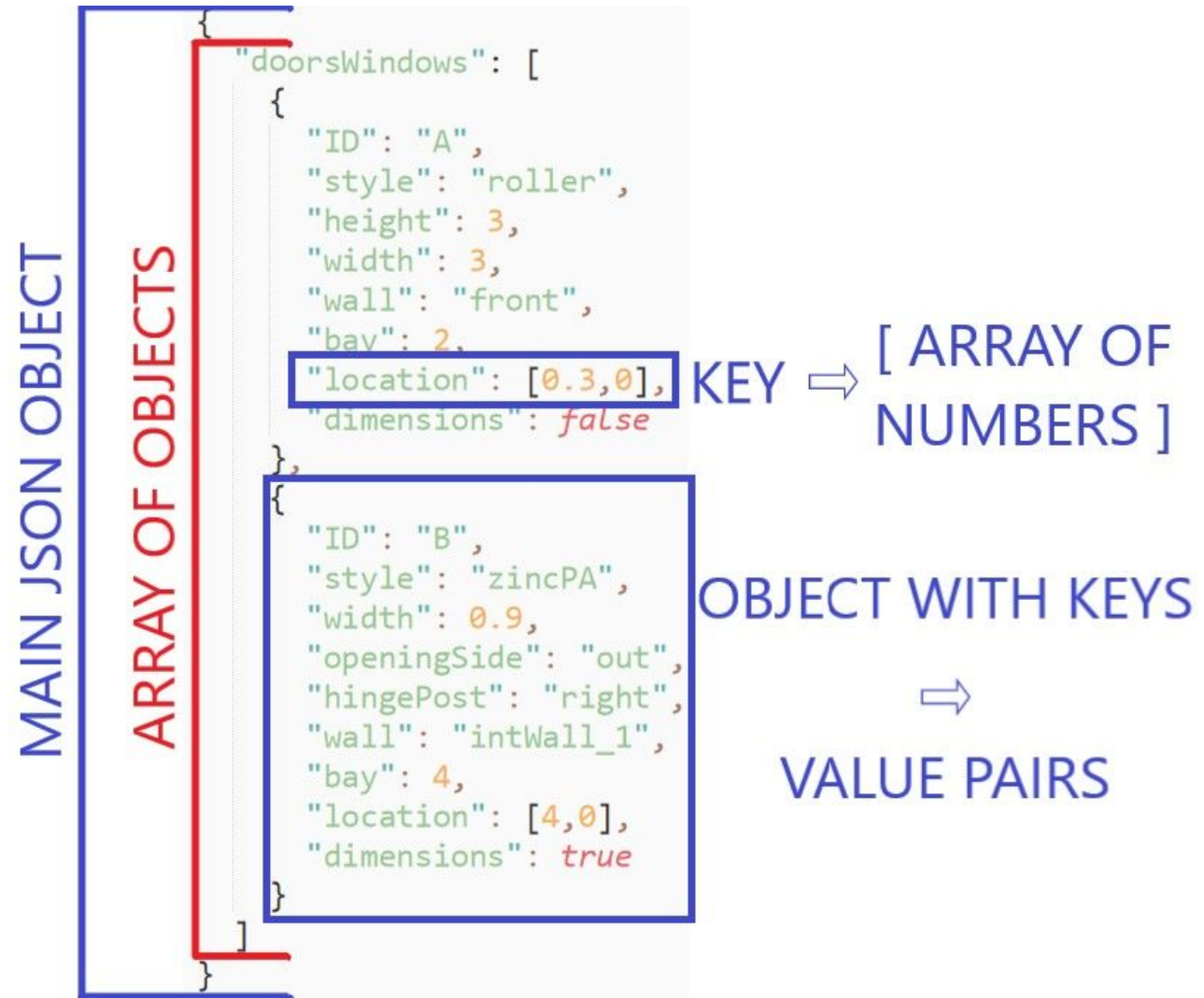
`{"id" : 0, "title": "Openheimer", "Status": "On-air"}`

JSON Array



```
[{"id" : 0, "title":"Openheimer", "Status":"On-air"},  
{"id" : 1, "title":"MI7", "Status":"Off"}]
```

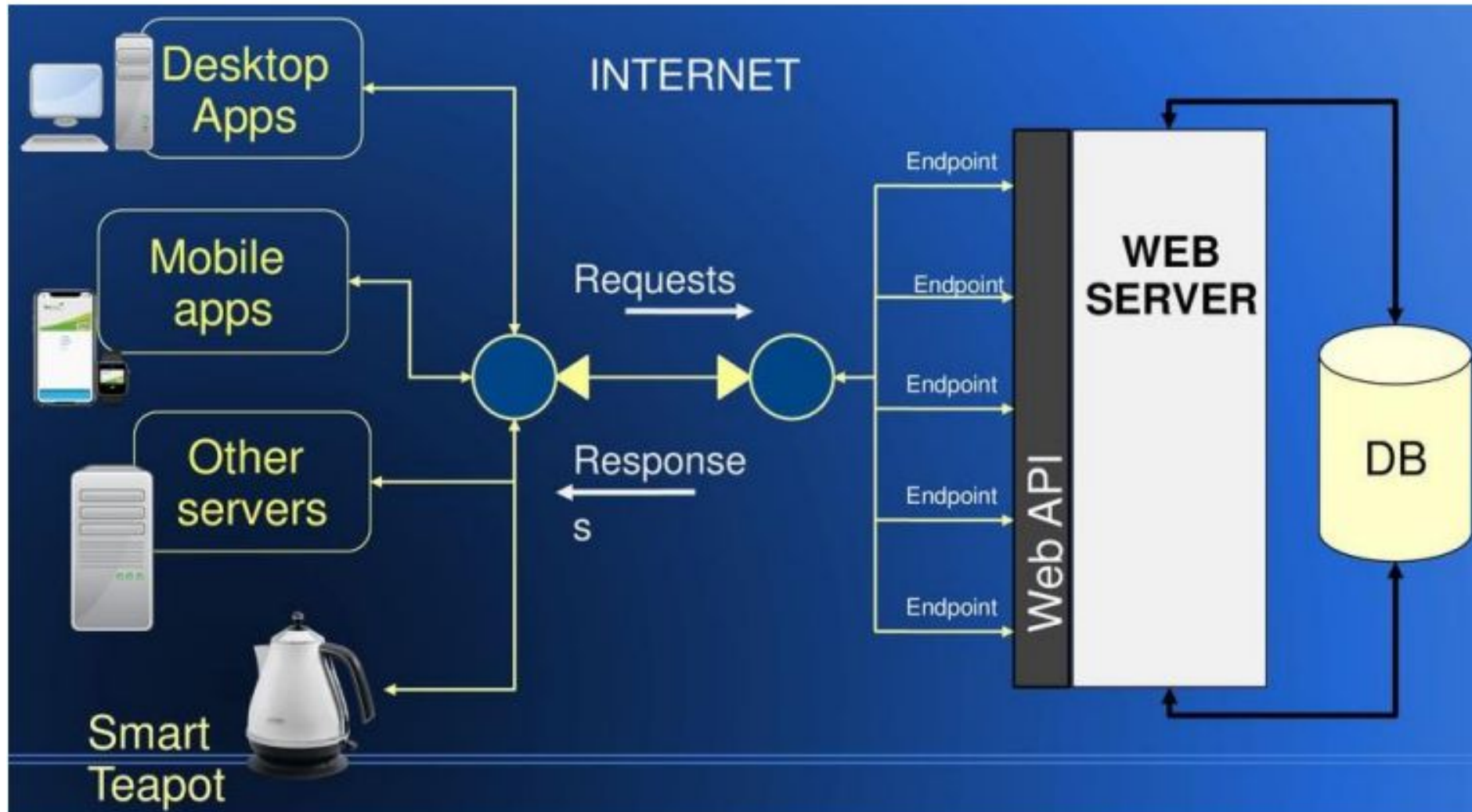
Mix of JSON objects and JSON arrays



What is Web API?

- Back-end application(server side) where actual functionality to call service/database call is happening to store and retrieve the data.
- An API over the web which can be accessed using HTTP protocol.
- A server-side programmatic interface allows the client application to communicate with the server
 - Enables external systems to use the business logics implemented in your application
- Based on one or more publicly exposed endpoints and a defined request-response message system.
 - Uses URLs in requests and helps obtain results in the JSON or XML formats
- It can be called interchangeably with the term **Web Service**.

Web API (or Web Service) Structure



More detail in FSAD (Every Friday)

Full Stack Application Development (2024) 🔍



This course familiarizes students with the principles and best practices necessary for the analysis, design, development, deployment, and maintenance of modern full stack software applications. Major design concepts and patterns are introduced then analyzed in the context of their implementations in modern full stack, front end, and back-end application development frameworks. Students put the concepts to practice by planning and executing a complete application development project in a team over the course of the semester. Students completing the course will be competent full stack application developers with the perspectives necessary to plan and execute application development projects in any enterprise.

This course requires either an undergraduate background in computer science, computer engineering, or information technology, or equivalent programming experience.

Tools

- Required Libraries

- **NumPy**: Provides a fast numerical array structure and helper functions.
- **pandas**: Provides a DataFrame structure to store data in memory and work with it easily and efficiently.
- **scikit-learn**: The essential Machine Learning package in Python.
- **matplotlib**: Basic plotting library in Python; most other Python plotting libraries are built on top of it.
- **Seaborn**: Advanced statistical plotting library.
- **watermark**: A Jupyter Notebook extension for printing timestamps, version numbers, and hardware information.

- Python

- Conda
 - <https://www.anaconda.com/download/>
- `conda install numpy pandas scikit-learn matplotlib seaborn`
- `conda install -c conda-forge watermark`

References

- [1] <https://www.firstsanfranciscopartners.com/blog/defining-data-acquisition-importance/?cnreloaded=1>
- [2] Lyko K., Nitzschke M., Ngonga Ngomo AC. (2016) Big Data Acquisition. In: Cavanillas J., Curry E., Wahlster W. (eds) New Horizons for a Data-Driven Economy. Springer, Cham.
https://doi.org/10.1007/978-3-319-21569-3_4
- [3] <https://www.forbes.com/sites/adrianbridgwater/2018/07/05/the-13-types-of-data/#48212ddf3362>
- [4] edX: Python for Data Science, UCSanDiegoX (DSE200x)
- [5] DataCamp: Introduction to Python
- [6] Microsoft: DAT203.1x Data Science Essentials
- [7] Carnegie Mellon University: CMU 15-388/688 Practical Data Science <http://www.datasciencecourse.org/>