



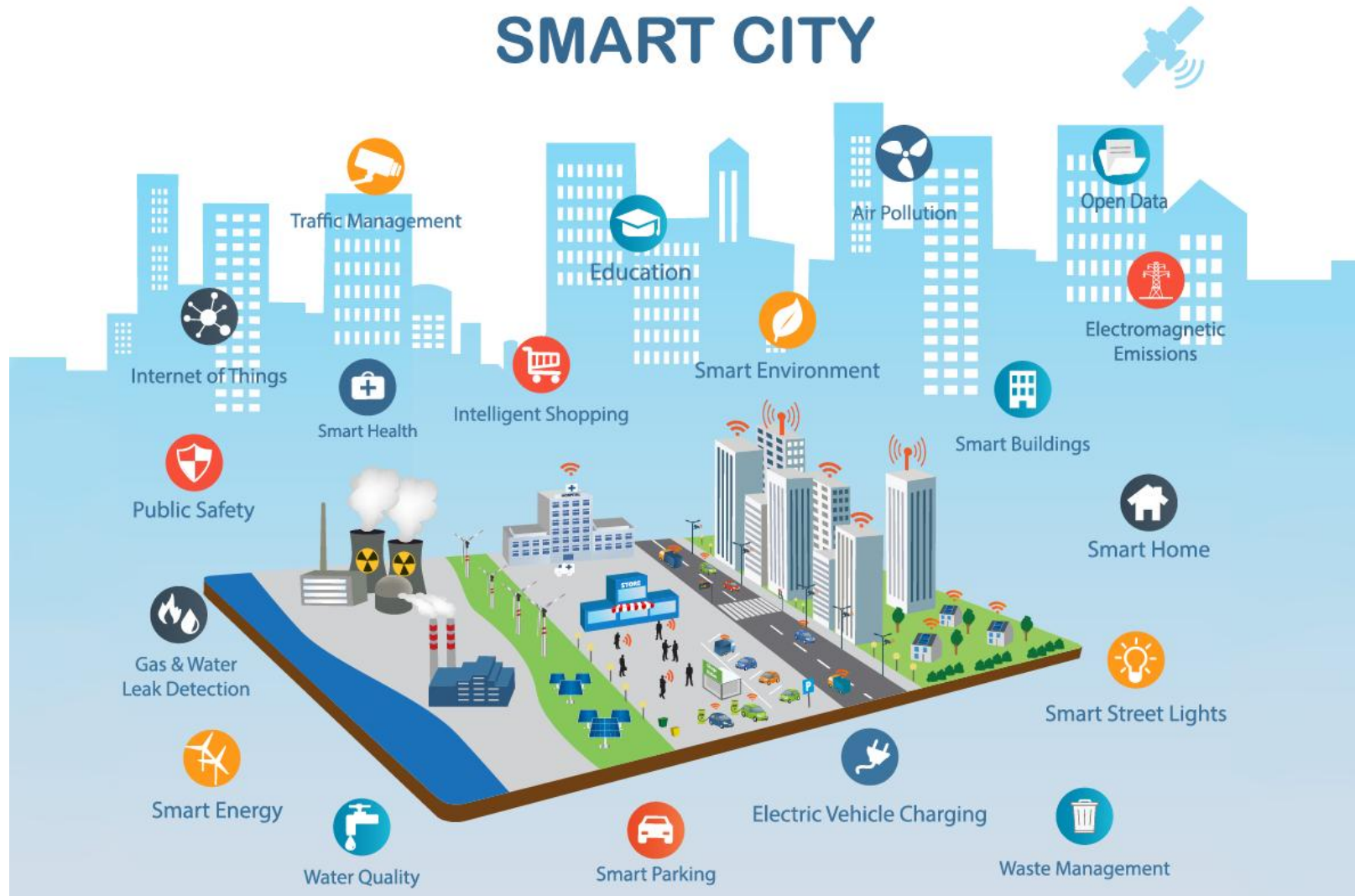
DSA5104

# Principles of Data Management and Retrieval

Lecture 0: Overview



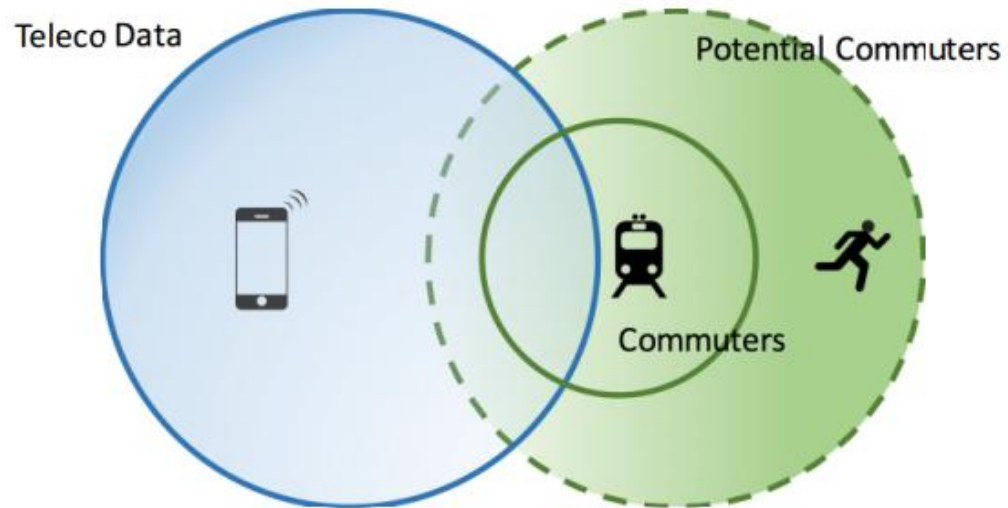
# Big Data - Data Flooding in Smart Cities



# Teleco + Metro Data → Crowd Forecasting

| User ID | Time      | Longitude  | Latitude | Action |
|---------|-----------|------------|----------|--------|
| 1001    | 8:00:00am | 103.737459 | 1.326909 | SMS    |
| 1002    | 8:00:01am | 103.737512 | 1.327108 | CAL    |
| 1003    | 8:00:03am | 103.741002 | 1.339921 | SMS    |
| 1004    | 8:00:03am | 103.738199 | 1.331231 | INT    |

Caller Detail Records (CDR) [1]

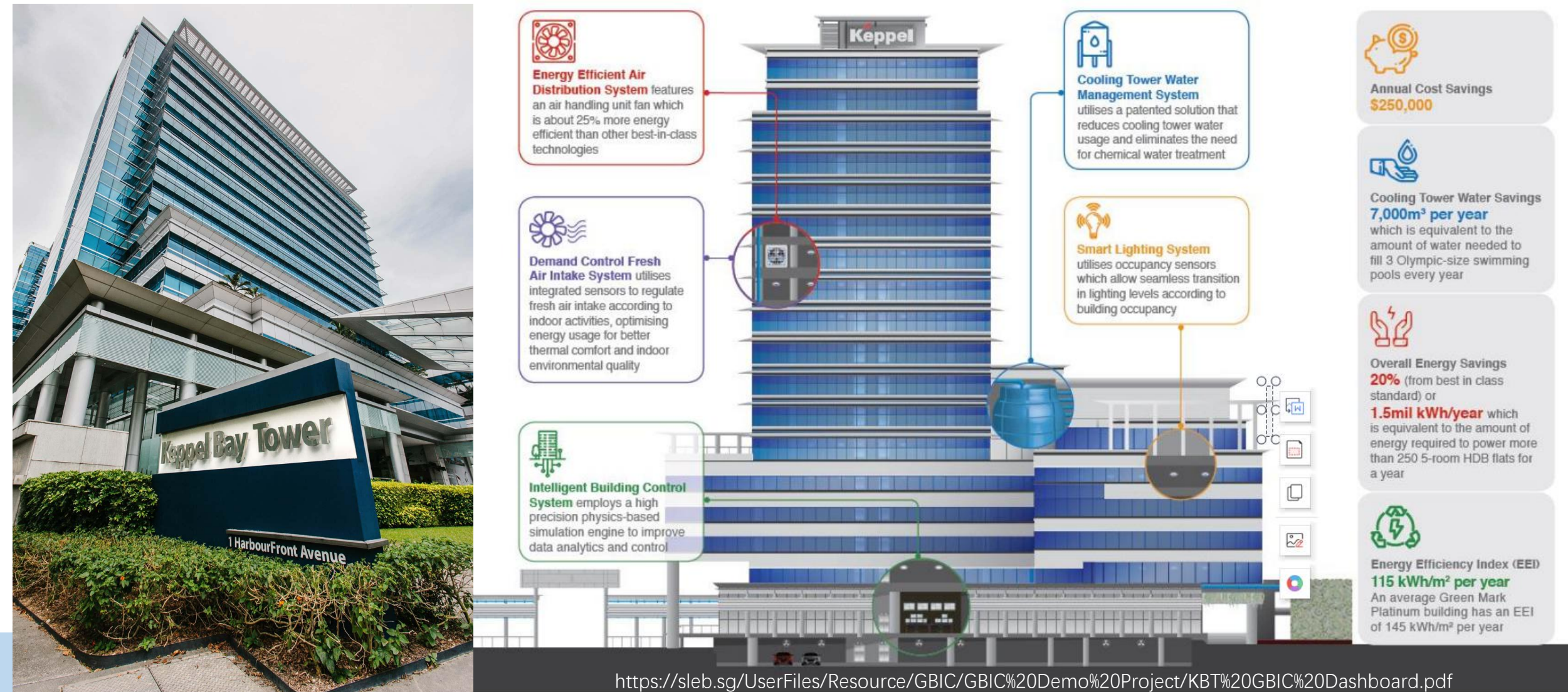


[1] Liang, Victor C., et al. "Mercury: Metro density prediction with recurrent neural network on streaming CDR data." ICDE 2016.



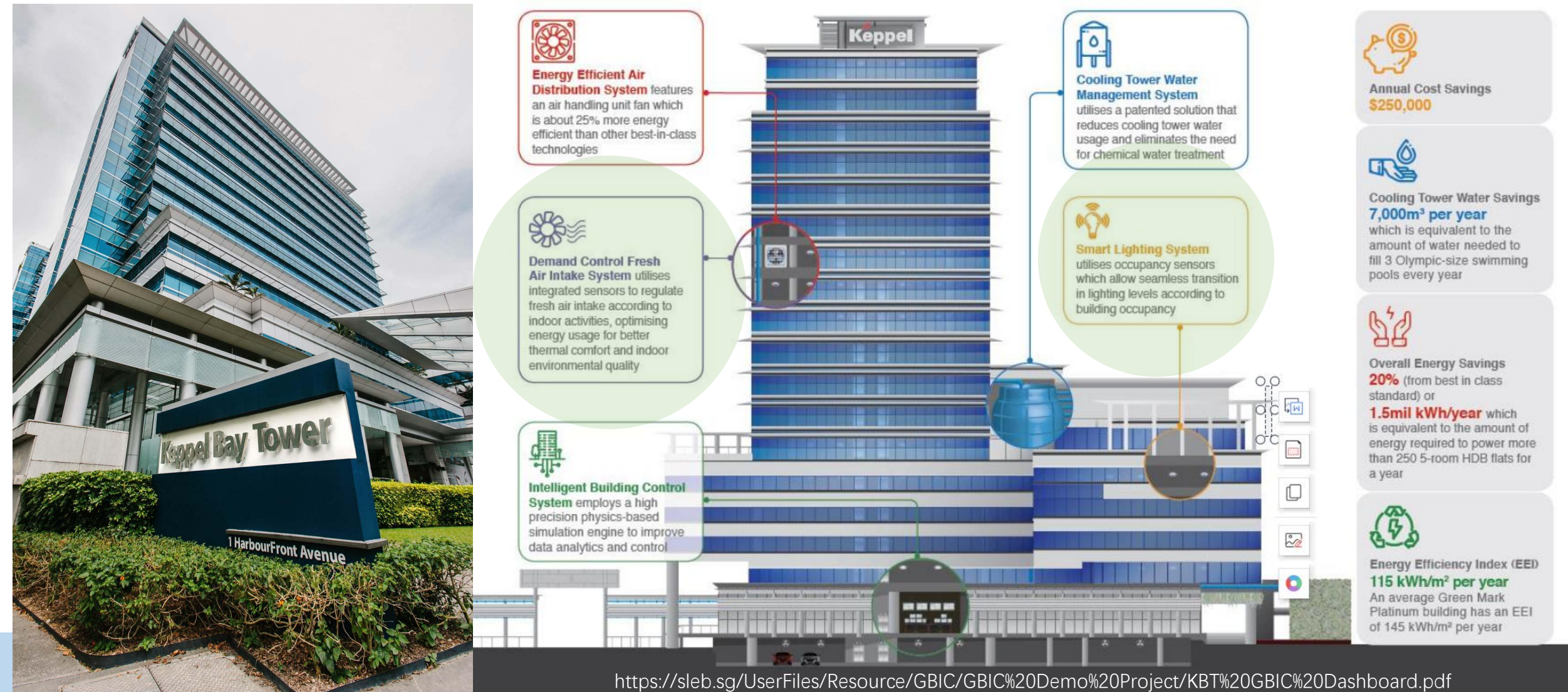


# Keppel Bay Tower - Singapore's first Green Mark Platinum (Zero Energy) commercial building





# Keppel Bay Tower - Singapore's first Green Mark Platinum (Zero Energy) commercial building



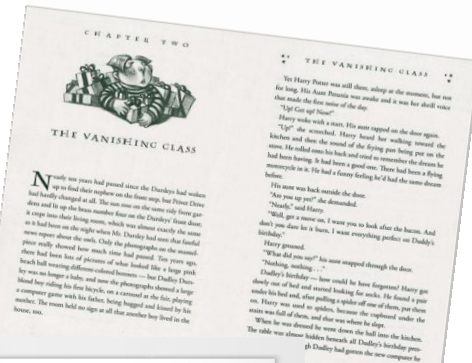


# Data Categorized by Data Types

Font size

Sample text  
The Wonderful Wizard of Oz  
Chapter 11: The Wonderful Emerald City of Oz  
Even with eyes protected by the green spectacles Dorothy and her friends were at first dazzled by the brilliancy of the wonderful City. The streets were lined with beautiful houses all built of green marble and studded over a pavement of the same green marble, and where the blocks were joined together were rows of emeralds, set closely, and glittering in the brightness of the sun. The window panes were of green glass, even the sky above the City had a green tint, and the rays of the sun were green.

There were many people, men, women and children, walking about, and these were all dressed in green clothes and had greenish skins. They looked at Dorothy and her strangely assorted company with wondering eyes, and the children all ran away and hid behind their mothers when they saw the Lion, but no harm came to them. Many shops stood in the street, and the rays of the sun were green.



## DAILY NEWS

### WORLD NEWS TODAY

World News Today  
The world is a beautiful place, full of amazing people and places. There are many different cultures and languages, and each one has its own unique way of life. We should all learn to appreciate and respect each other, and work together to make the world a better place for everyone.

## ECONOMIC NEWS

Economic News  
The economy is a complex system, and it can be difficult to understand. However, there are some basic principles that can help us understand how it works. For example, supply and demand are two of the most important factors in determining the price of a good or service. Understanding these principles can help us make better decisions about how to spend our money.



2021 in pictures | Galle...  
cnn.com



Pictures of the month: April | Reuter...  
reuters.com



350+ Free Pictures [HD] | Download Free...  
unsplash.com



Photographs vs pictures: Whatever you ...  
recordnet.com



Download - Pexels Stock Photos  
pexels.com



How to Take Good Pictures (10...  
expertphotography.com



Pictures - Home | Facebook  
m.facebook.com



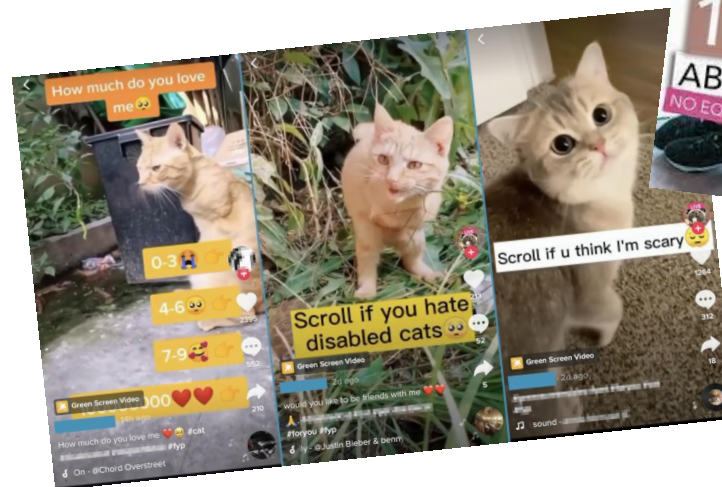
Pictures of the month: January ...  
reuters.com



2021: The year in pictures | The ...  
straitstimes.com



How to Take Sharp Photos  
photographylife.com



# Data Categorized by Data Types





# Structured Data

- Data conforms to a set schema
  - Numerical, categorical and text data with well-defined schema

Table name: *instructor*

| Column           | Data Type            |
|------------------|----------------------|
| <i>ID</i>        | varchar(5)           |
| <i>name</i>      | varchar(20) not null |
| <i>dept_name</i> | varchar(20)          |
| <i>salary</i>    | numeric(8, 2)        |

*instructor*

| <i>ID</i> | <i>name</i> | <i>dept_name</i> | <i>salary</i> |
|-----------|-------------|------------------|---------------|
| 22222     | Einstein    | Physics          | 95000         |
| 12121     | Wu          | Finance          | 90000         |
| 32343     | El Said     | History          | 60000         |
| 45565     | Katz        | Comp. Sci.       | 75000         |
| 98345     | Kim         | Elec. Eng.       | 80000         |
| 76766     | Crick       | Biology          | 72000         |
| 10101     | Srinivasan  | Comp. Sci.       | 65000         |
| 58583     | Califieri   | History          | 62000         |
| 83821     | Brandt      | Comp. Sci.       | 92000         |
| 15151     | Mozart      | Music            | 40000         |
| 33456     | Gold        | Physics          | 87000         |
| 76543     | Singh       | Finance          | 80000         |

# Semi-Structured Data

- Data with labels but no fixed schema
  - JSON, XML
  - CSV files with headers
- Usage:
  - For data exchange

```
persons.csv - Notepad
File Edit Format View Help
Family Name,Given Name,VIAF ID
Ackersdijck,Willem Cornelis,17959345
Adelung,Friedrich von,22963658
Afzelius,Arvid August,49972119
Amerling,Karel,13331054
Anton,Karl Gottlob von,183632821
Arwidsson,Adolf Ivar,8184878
Asbjørnsen,Peter Christen,116587918
Attems,Heinrich,37665468
Atterbom,Per Daniel Amadeus,46819248
Balabin,Viktor Petrovich,44473845
Banks,Joseph,46830189
Beck,Friedrich,44338671
Becker,Reinhold von,42101066
Bernhart,Johann Baptist,69674335
Bertram,Johann,32890043
Bilderdijk,Willem,14882166
Boisserée,Sulpiz,7483155
Bopp,Franz,61614118
Borovský,Karel Havlíček,100277614
Bosković,Jovan,161354270
Buslaev,Fyodor,10074560
Cenowa,Florian Stanislaw,44466031
Chomiakov,Aleksei,66492873
```

[https://e.nodegoat.net/CMS/upload/guide-import\\_person\\_csv\\_notepad.png](https://e.nodegoat.net/CMS/upload/guide-import_person_csv_notepad.png)



# Semi-Structured Data

- XML - eXtensible Markup Language

```
1  <note>
2    <to>Daddy</to>
3    <from>Jimmy</from>
4    <body>Don't forget to play with me this weekend!</body>
5  </note>
```

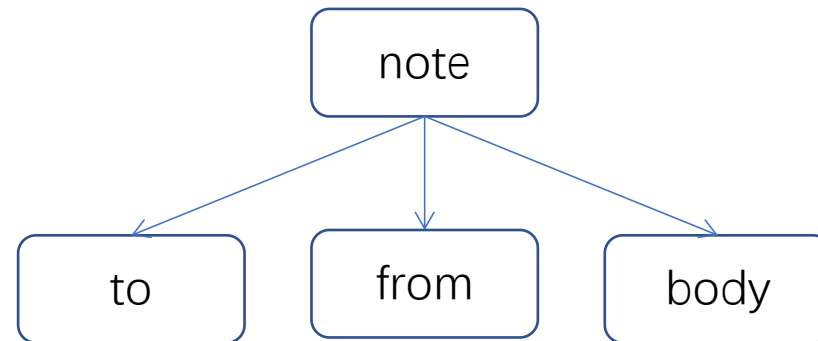
# Semi-Structured Data

- XML - eXtensible Markup Language

Tag

```
1 <note>
2   <to>Daddy</to>
3   <from>Jimmy</from>
4   <body>Don't forget to play with me this weekend!</body>
5 </note>
```

Closing tag





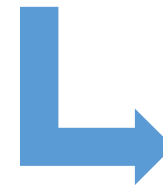
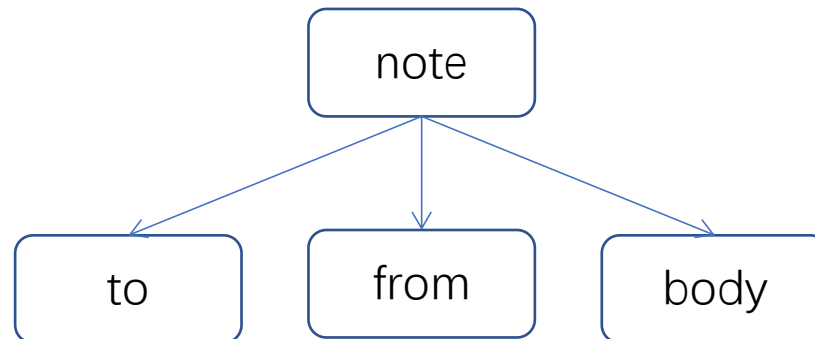
# Semi-Structured Data

- XML - eXtensible Markup Language

Tag

```
1 <note>
2   <to>Daddy</to>
3   <from>Jimmy</from>
4   <body>Don't forget to play with me this weekend!</body>
5 </note>
```

Closing tag



## Note

To: Daddy

From: Jimmy

Don't forget to play with me this weekend!

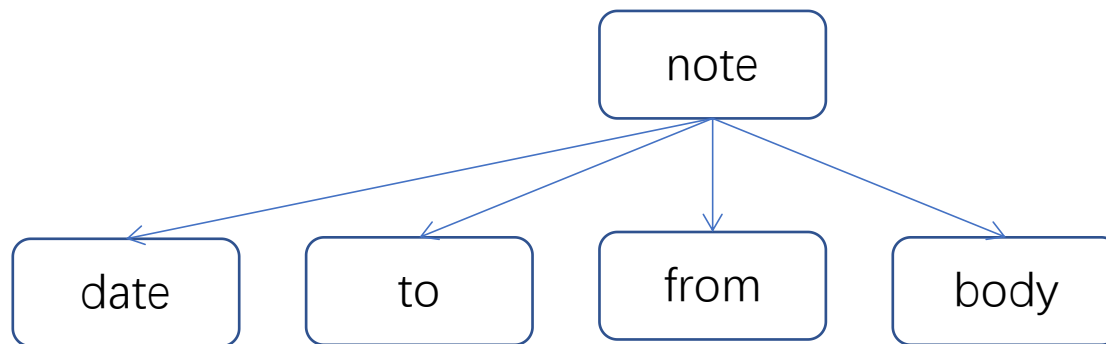
# Semi-Structured Data

- XML - eXtensible Markup Language

Tag

```
1 <note>
2   <to>Daddy</to>
3   <from>Jimmy</from>
4   <body>Don't forget to play with me this weekend!</body>
5 </note>
```

Closing tag





# Semi-Structured Data

- XML

```
<?xml version="1.0" encoding="ISO-8859-15"?>
<package destination="SU" origin="ASR" version="1.0">
  <recognized_sentence>
    <information>
      I would like the train fares from Valencia to Madrid
    </information>
    <confidences>
      <word confidence="0.47" value="I" />
      <word confidence="0.68" value="would" />
      <word confidence="0.53" value="like" />
      <word confidence="0.75" value="the" />
      <word confidence="0.64" value="train" />
      <word confidence="0.56" value="fares" />
      <word confidence="0.84" value="from" />
      <word confidence="0.93" value="Valencia" />
      <word confidence="0.78" value="to" />
      <word confidence="0.93" value="Madrid" />
    </confidences>
  </recognized_sentence>
  <grammar name="dihana.jsgf">
</package>
```

# Semi-Structured Data

- XML

A prolog defines the XML version and the character encoding

```
<?xml version="1.0" encoding="ISO-8859-15"?>
<package destination="SU" origin="ASR" version="1.0">
  <recognized_sentence>
    <information>
      I would like the train fares from Valencia to Madrid
    </information>
    <confidences>
      <word confidence="0.47" value="I" />
      <word confidence="0.68" value="would" />
      <word confidence="0.53" value="like" />
      <word confidence="0.75" value="the" />
      <word confidence="0.64" value="train" />
      <word confidence="0.56" value="fares" />
      <word confidence="0.84" value="from" />
      <word confidence="0.93" value="Valencia" />
      <word confidence="0.78" value="to" />
      <word confidence="0.93" value="Madrid" />
    </confidences>
  </recognized_sentence>
  <grammar name="dihana.jsgf">
</package>
```

# Semi-Structured Data

- XML

A prolog defines the XML version and the character encoding

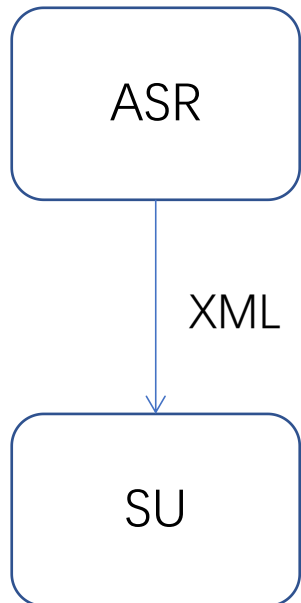
```
<?xml version="1.0" encoding="ISO-8859-15"?>
<package destination="SU" origin="ASR" version="1.0">
  <recognized_sentence>
    <information>
      I would like the train fares from Valencia to Madrid
    </information>
    <confidences>
      <word confidence="0.47" value="I" />
      <word confidence="0.68" value="would" />
      <word confidence="0.53" value="like" />
      <word confidence="0.75" value="the" />
      <word confidence="0.64" value="train" />
      <word confidence="0.56" value="fares" />
      <word confidence="0.84" value="from" />
      <word confidence="0.93" value="Valencia" />
      <word confidence="0.78" value="to" />
      <word confidence="0.93" value="Madrid" />
    </confidences>
  </recognized_sentence>
  <grammar name="dihana.jsgf">
</package>
```

Element *package* has three attributes: destination, origin and version.



# Semi-Structured Data

## ■ XML



A prolog defines the XML version and the character encoding

```
<?xml version="1.0" encoding="ISO-8859-15"?>
<package destination="SU" origin="ASR" version="1.0">
  <recognized_sentence>
    <information>
      I would like the train fares from Valencia to Madrid
    </information>
    <confidences>
      <word confidence="0.47" value="I" />
      <word confidence="0.68" value="would" />
      <word confidence="0.53" value="like" />
      <word confidence="0.75" value="the" />
      <word confidence="0.64" value="train" />
      <word confidence="0.56" value="fares" />
      <word confidence="0.84" value="from" />
      <word confidence="0.93" value="Valencia" />
      <word confidence="0.78" value="to" />
      <word confidence="0.93" value="Madrid" />
    </confidences>
  </recognized_sentence>
  <grammar name="dihana.jsgf">
</package>
```

Element *package* has three attributes: destination, origin and version.

# Semi-Structured Data

- JSON

```
{ "menu": {  
  "id": "file",  
  "value": "File",  
  "popup": {  
    "menuitem": [  
      { "value": "New", "onclick": "CreateNewDoc()" },  
      { "value": "Open", "onclick": "OpenDoc()" },  
      { "value": "Close", "onclick": "CloseDoc()" }  
    ]  
  }  
}}
```

# Semi-Structured Data

- JSON

key/value pair

```
{ "menu": {  
  "id": "file",  
  "value": "File",  
  "popup": {  
    "menuitem": [  
      { "value": "New", "onclick": "CreateNewDoc()" },  
      { "value": "Open", "onclick": "OpenDoc()" },  
      { "value": "Close", "onclick": "CloseDoc()" }  
    ]  
  }  
}
```



# Semi-Structured Data

- JSON

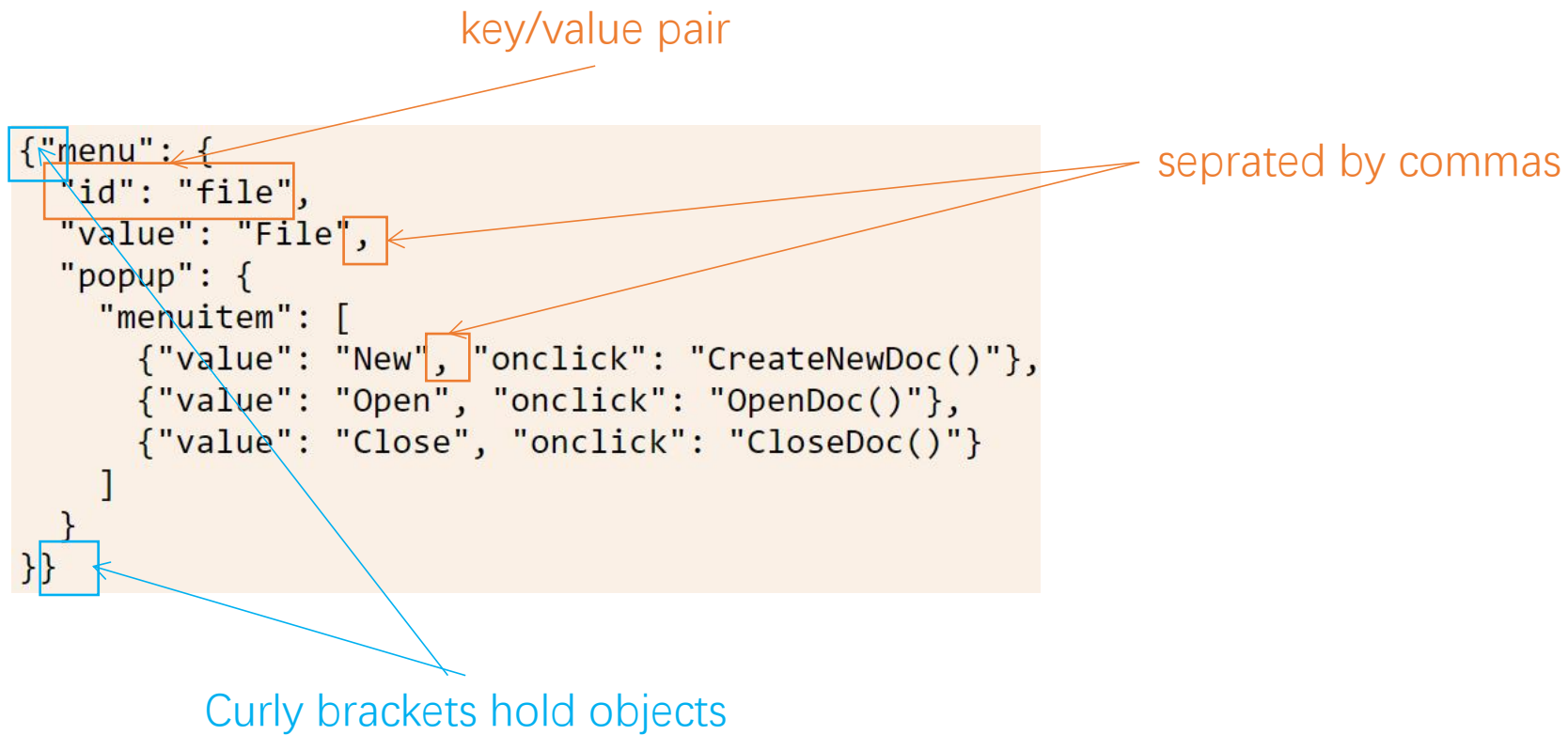
key/value pair

separated by commas

```
{ "menu": {  
  "id": "file",  
  "value": "File",  
  "popup": {  
    "menuitem": [  
      { "value": "New", "onclick": "CreateNewDoc()" },  
      { "value": "Open", "onclick": "OpenDoc()" },  
      { "value": "Close", "onclick": "CloseDoc()" }  
    ]  
  }  
}
```

# Semi-Structured Data

- JSON



The diagram illustrates the JSON syntax using an example object. The JSON text is: 

```
{ "menu": { "id": "file", "value": "File", "popup": { "menuitem": [ { "value": "New", "onclick": "CreateNewDoc()" }, { "value": "Open", "onclick": "OpenDoc()" }, { "value": "Close", "onclick": "CloseDoc()" } ] } } }
```

 Annotations include: an orange arrow pointing to the opening curly brace of the 'menu' object with the label 'key/value pair'; an orange arrow pointing to the comma after 'value: \"File\"' with the label 'seprated by commas'; an orange arrow pointing to the comma after the first menu item object with the label 'seprated by commas'; and a blue arrow pointing to the closing curly brace of the outermost object with the label 'Curly brackets hold objects'.

```
{ "menu": {  
  "id": "file",  
  "value": "File",  
  "popup": {  
    "menuitem": [  
      { "value": "New", "onclick": "CreateNewDoc()" },  
      { "value": "Open", "onclick": "OpenDoc()" },  
      { "value": "Close", "onclick": "CloseDoc()" }  
    ]  
  }  
}
```

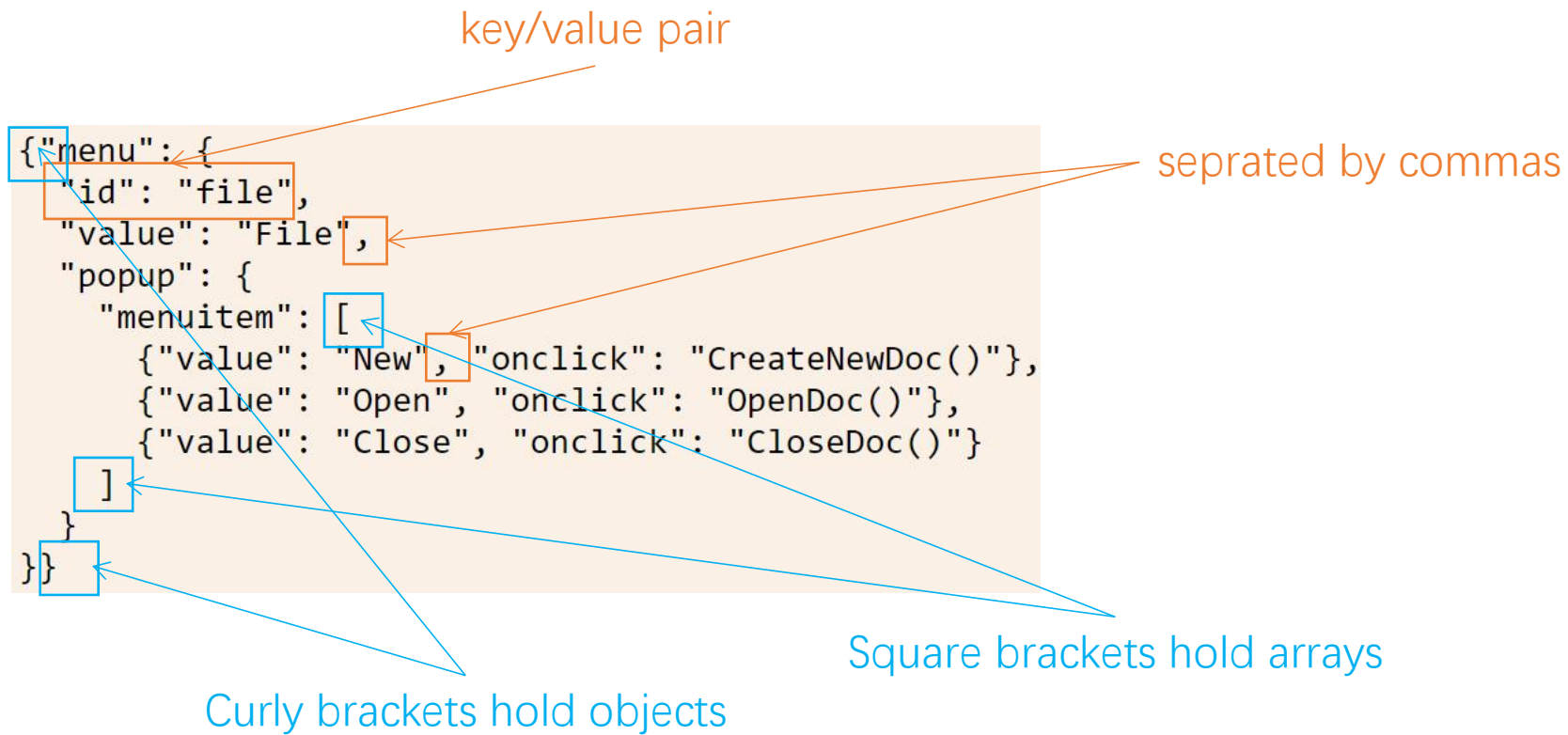
key/value pair

seprated by commas

Curly brackets hold objects

# Semi-Structured Data

- JSON





# Semi-Structured Data

- JSON

```
{ "menu": {  
  "id": "file",  
  "value": "File",  
  "popup": {  
    "menuitem": [  
      { "value": "New", "onclick": "CreateNewDoc()" },  
      { "value": "Open", "onclick": "OpenDoc()" },  
      { "value": "Close", "onclick": "CloseDoc()" }  
    ]  
  }  
}}
```

- XML

```
<menu id="file" value="File">  
  <popup>  
    <menuitem value="New" onclick="CreateNewDoc()" />  
    <menuitem value="Open" onclick="OpenDoc()" />  
    <menuitem value="Close" onclick="CloseDoc()" />  
  </popup>  
</menu>
```

# Load JSON File Using Pandas

data.json

```
1 {
2   "Duration":{
3     "0":60,
4     "1":60,
5     "2":60,
6     "3":45,
7     "4":45,
8     "5":60
9   },
10  "Pulse":{
11    "0":110,
12    "1":117,
13    "2":103,
14    "3":109,
15    "4":117,
16    "5":102
17  },
18  "Maxpulse":{
19    "0":130,
20    "1":145,
21    "2":135,
22    "3":175,
23    "4":148,
24    "5":127
25  },
26  "Calories":{
27    "0":409,
28    "1":479,
29    "2":340,
30    "3":282,
31    "4":406,
32    "5":300
33  }
34 }
```

# Load JSON File Using Pandas





# Load JSON File Using Pandas

data.json

```
1 {
2   "Duration":{
3     "0":60,
4     "1":60,
5     "2":60,
6     "3":45,
7     "4":45,
8     "5":60
9   },
10  "Pulse":{
11    "0":110,
12    "1":117,
13    "2":103,
14    "3":109,
15    "4":117,
16    "5":102
17  },
18  "Maxpulse":{
19    "0":130,
20    "1":145,
21    "2":135,
22    "3":175,
23    "4":148,
24    "5":127
25  },
26  "Calories":{
27    "0":409,
28    "1":479,
29    "2":340,
30    "3":282,
31    "4":406,
32    "5":300
33  }
34 }
```

Jupyter Notebook

```
In [1]: import pandas as pd
        df = pd.read_json('data.json')
```

```
In [2]: df
```

Out[2]:

|   | Duration | Pulse | Maxpulse | Calories |
|---|----------|-------|----------|----------|
| 0 | 60       | 110   | 130      | 409      |
| 1 | 60       | 117   | 145      | 479      |
| 2 | 60       | 103   | 135      | 340      |
| 3 | 45       | 109   | 175      | 282      |
| 4 | 45       | 117   | 148      | 406      |
| 5 | 60       | 102   | 127      | 300      |

# — Unstructured Data

“I’m Hilda. I was born in 1990.”

“My name is Max. I’m turning 20 this year.”

# Unstructured Data - Natural Language Text

## Discharge Summary

**Patient Name:** *Russell Johnson*  
**Medical Record Number:** *123456789*  
**Admission Date:** *08/01/14*  
**Discharge Date:** *08/05/14*  
**Attending Physician:** *Dr Gary Marshall*  
**Dictated by:** *Dr Gary Marhsall*

**Primary Care Physician:** *Dr Dianna Miller*  
**Referring Physician:**  
**Consulting Physician(s):** *Dr Gary Marshall - hospitalist*  
**Condition on Discharge:** *stable*

**Final Diagnosis:** *RLL pneumonia, COPD exacerbation, mild CHF, osteoarthritis*

**Procedures:** *none*

**History of Present Illness** *72 year old thin white male presented to emergency on 8/1/14 with shortness of breath, weakness and dehydration. Chest X-ray showed right lower lobe infiltrate, ABGs unremarkable. Pulse ox on RA was 79%.*

- 1) Pneumonia: treated with ceftriaxone and azithromycin iv. Switched to PO after 72 hours.*
- 2) Exacerbation of COPD: patient treated with inhaled and oral steroids, O2 at 2l/nc. On RA at time of discharge*
- 3) Weakness and dehydration: secondary to pneumonia and COPD. Responded well to strengthening with PT and regular meals.*

**Discharge Medications** *Zithromycin daily until gone, inhalers #of puffs,*

**Discharge Instructions:** *no activity restriction, regular diet, follow up in two to three weeks with regular physician.*

# Unstructured Data - Natural Language Text

## Discharge Summary

**Patient Name:** *Russell Johnson*  
**Medical Record Number:** *123456789*  
**Admission Date:** *08/01/14*  
**Discharge Date:** *08/05/14*  
**Attending Physician:** *Dr Gary Marshall*  
**Dictated by:** *Dr Gary Marhsall*

**Primary Care Physician:** *Dr Dianna Miller*  
**Referring Physician:**  
**Consulting Physician(s):** *Dr Gary Marshall - hospitalist*  
**Condition on Discharge:** *stable*

**Final Diagnosis:** *RLL pneumonia, COPD exacerbation, mild CHF, osteoarthritis*

**Procedures:** *none*

**History of Present Illness** *72 year old thin white male presented to emergency on 8/1/14 with shortness of breath, weakness and dehydration. Chest X-ray showed right lower lobe infiltrate, ABGs unremarkable. Pulse ox on RA was 79%.*

- 1) Pneumonia: treated with ceftriaxone and azithromycin iv. Switched to PO after 72 hours.*
- 2) Exacerbation of COPD: patient treated with inhaled and oral steroids, O2 at 2l/nc. On RA at time of discharge*
- 3) Weakness and dehydration: secondary to pneumonia and COPD. Responded well to strengthening with PT and regular meals.*

**Discharge Medications** *Zithromycin daily until gone, inhalers #of puffs,*

**Discharge Instructions:** *no activity restriction, regular diet, follow up in two to three weeks with regular physician.*



# Unstructured Data - Multimedia Content



# Image2Caption

Image2Caption

Back

Select New Image

Analyse

Image2Caption

Beam Search Size 1-5



Beam Search Size 1

A cat is standing in the air.

Beam Search Size 2

A cat is standing in the air.

Image2Caption

Back

Select New Image

Analyse

Image2Caption

Beam Search Size 1-5



Beam Search Size 1

A man with a blue mask is looking at a table.

Beam Search Size 2

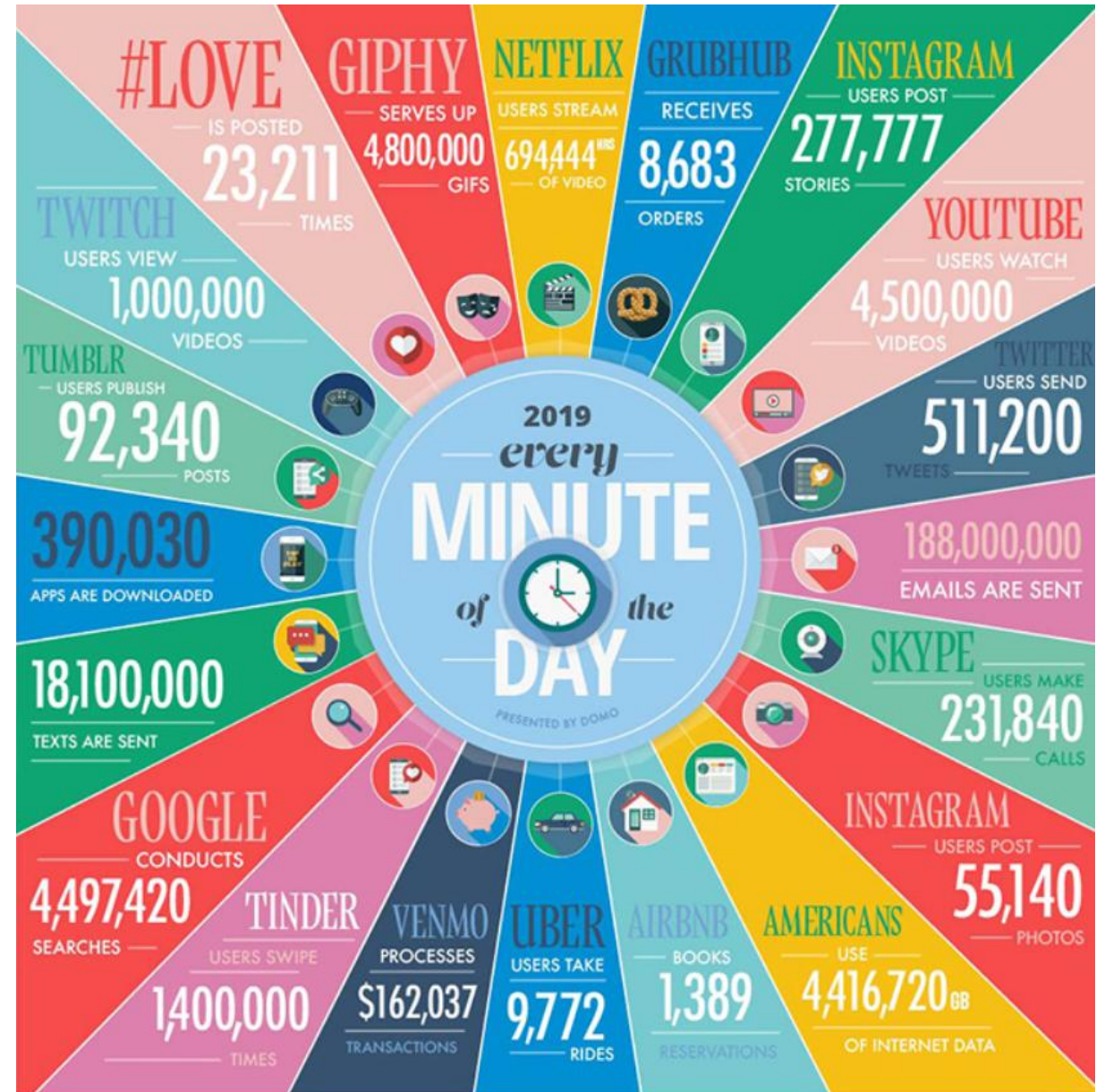
A man in a blue shirt is looking at a table.

<https://image2caption.pascalperle.de/>

# What is Big Data?

- Big data sets are too large or complex to be processed by traditional methods.

Consider that in a single minute (2019) there are:

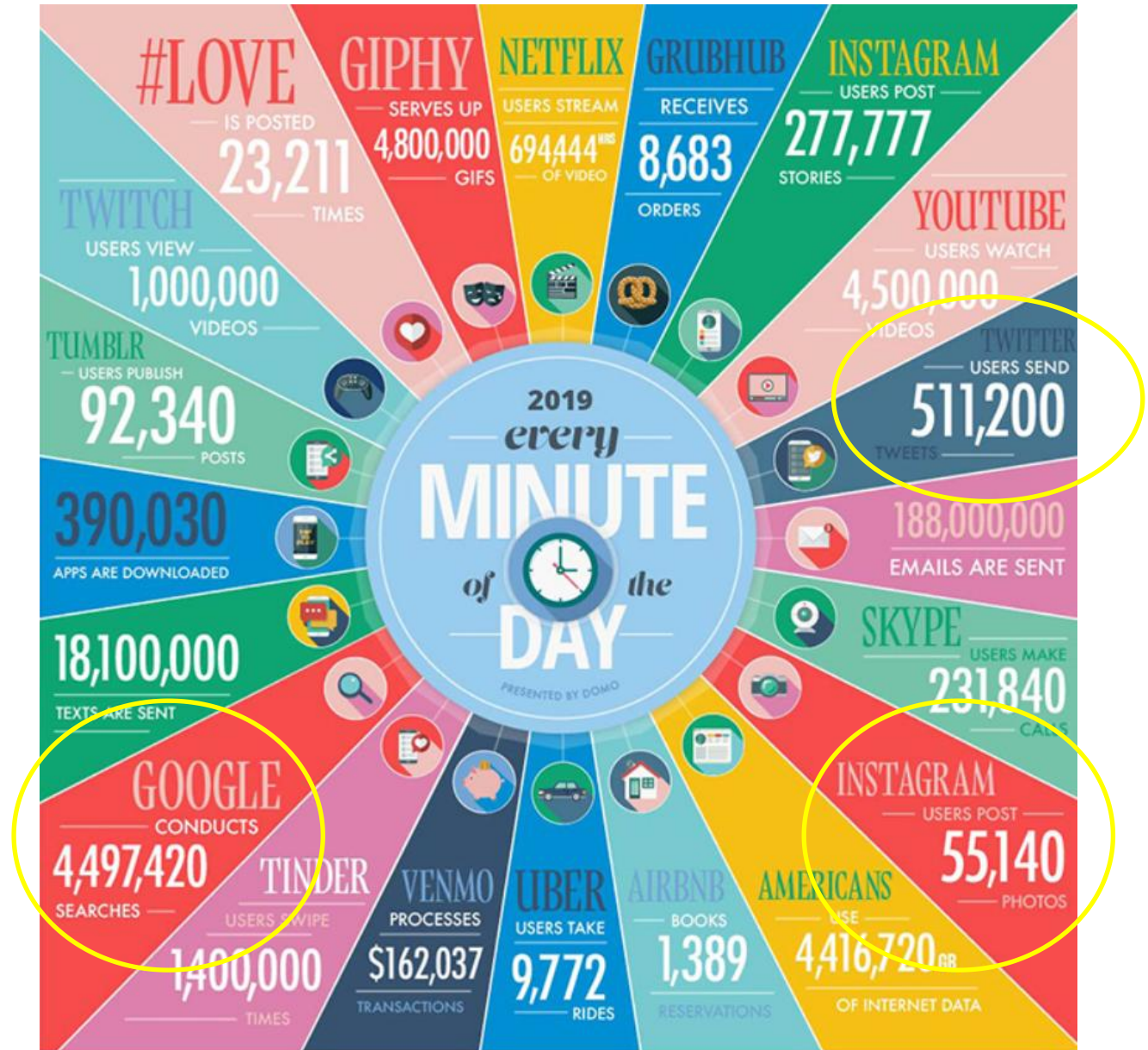




# What is Big Data?

- Big data sets are too large or complex to be processed by traditional methods.

Consider that in a single minute (2019) there are:





# What is Big Data?

- Big data sets are too large or complex to be processed by traditional methods.

Consider that in a single minute (2021) there are:



# What is Big Data?

- Big data sets are too large or complex to be processed by traditional methods.

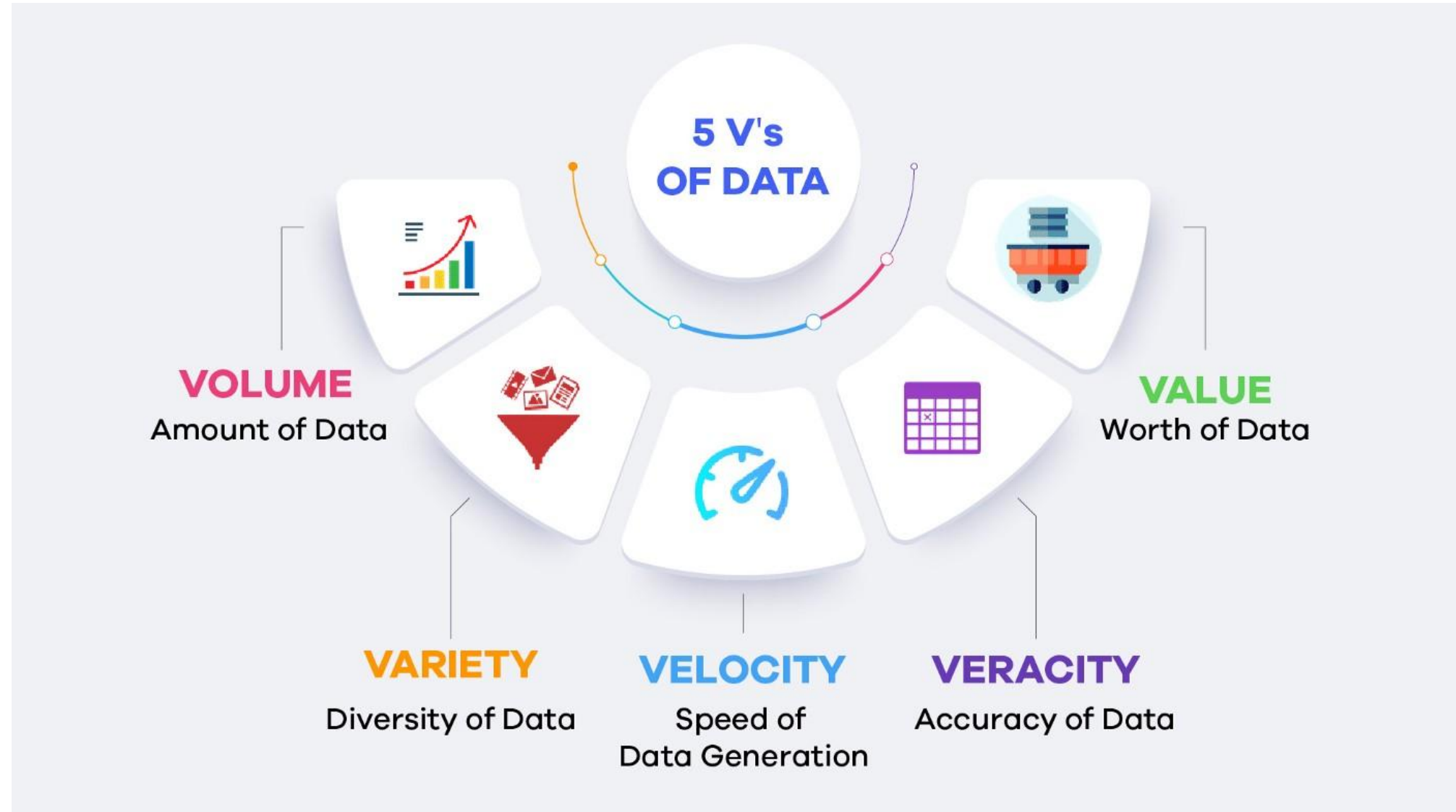
Consider that in a single minute (2021) there are:

| Activity         | 2019 | 2021 | Increase |
|------------------|------|------|----------|
| Google Search    | 4.5M | 5.7M | 26.7%    |
| Instagram Photos | 55K  | 65K  | 18.2%    |
| Twitter Tweets   | 511K | 575K | 12.5%    |

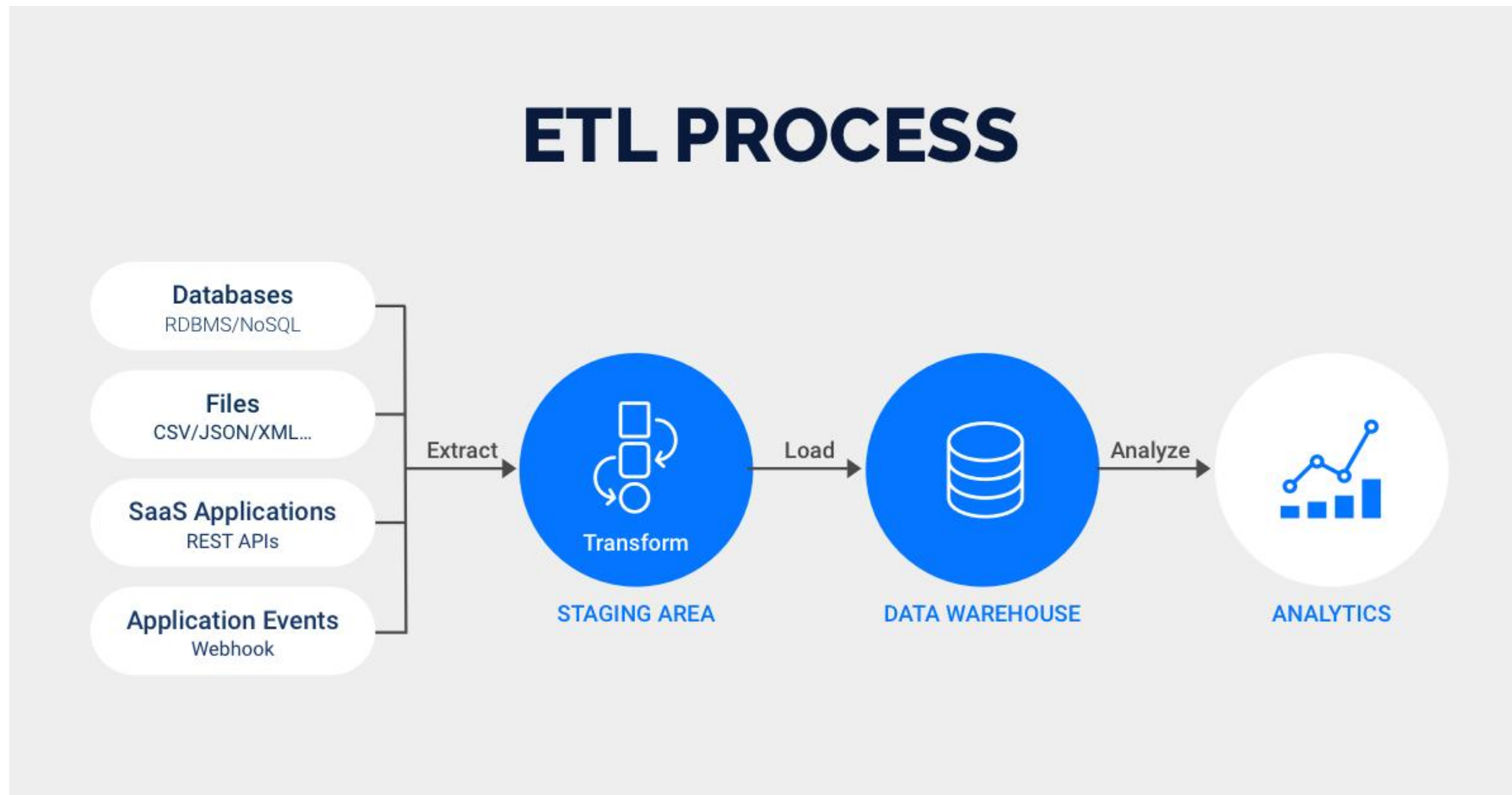




# The 5 Vs of Big Data

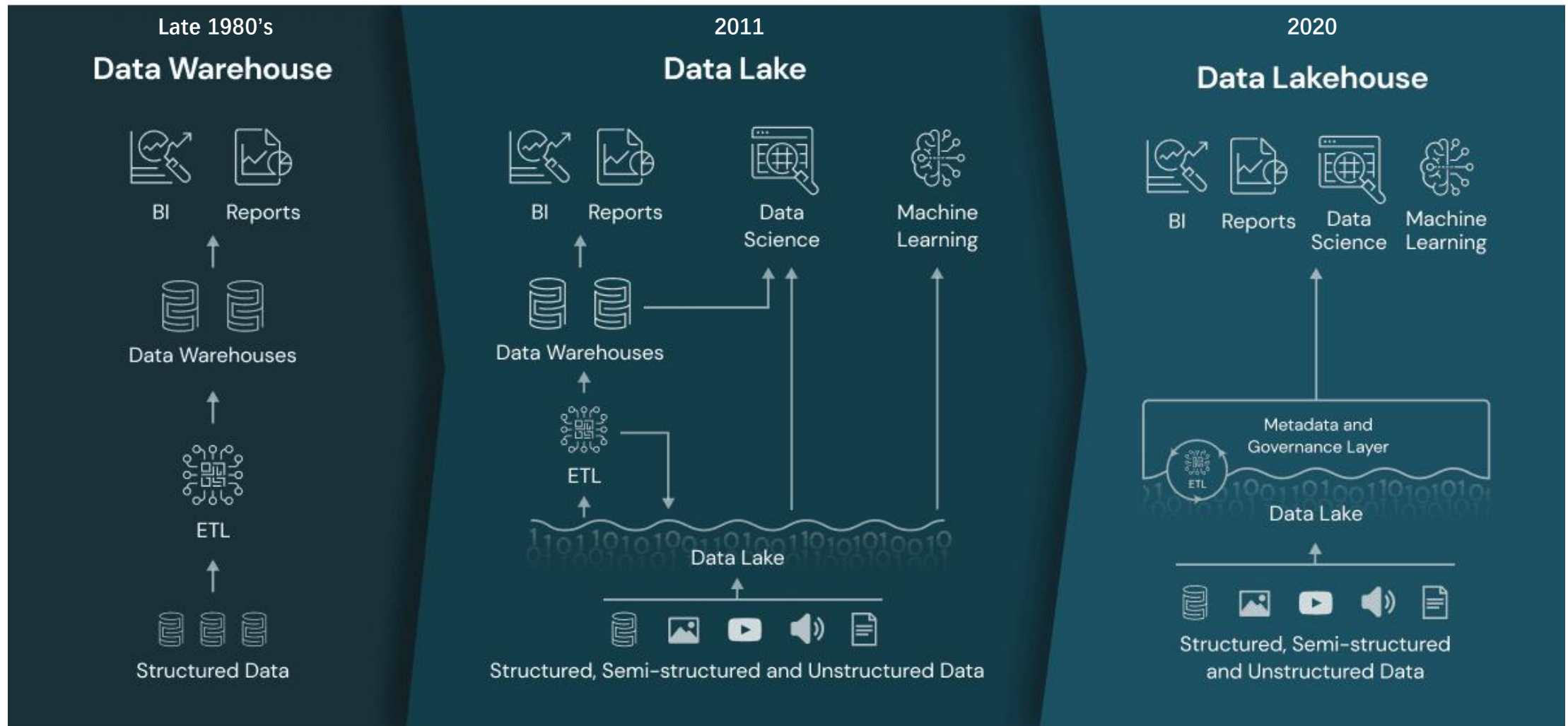


# Data Warehouse & ETL (Extract, Transform, Load)

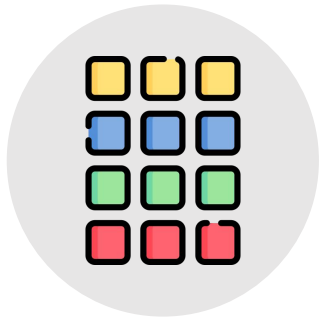




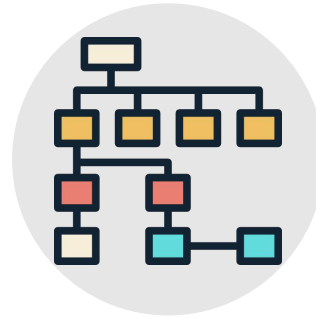
# Data Lakehouse



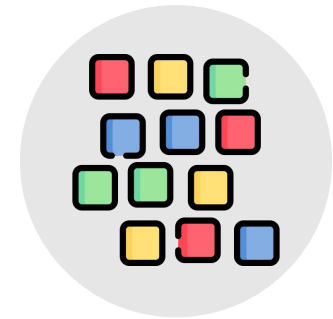
# Summary



Structured



Semi-Structured



Unstructured

|                         |  |   |   |
|-------------------------|--|---|---|
| <b>Definition</b>       | <ul style="list-style-type: none"><li>• Data with predefined schema</li></ul>  | <ul style="list-style-type: none"><li>• Data with flexible schema (e.g., XML)</li></ul> | <ul style="list-style-type: none"><li>• Data without predefined schema</li></ul>  |
| <b>Database Systems</b> | <ul style="list-style-type: none"><li>• Relational Database<ul style="list-style-type: none"><li>• MySQL, PostgreSQL</li></ul></li></ul> | <ul style="list-style-type: none"><li>• MongoDB/HBase</li></ul>                         | <ul style="list-style-type: none"><li>• No-SQL Databases<ul style="list-style-type: none"><li>• Object store (S3)</li></ul></li></ul> |
| <b>Query Language</b>   | <ul style="list-style-type: none"><li>• SQL</li></ul>  | <ul style="list-style-type: none"><li>• XPath</li><li>• XQuery</li></ul>                | <ul style="list-style-type: none"><li>• ElasticSearch for text</li></ul>  |

# Course Overview

- This course is about managing and retrieving different types of data, i.e., structured, semi-structured and unstructured data using different *database systems*
- “*Data management* is the practice of collecting, keeping, and using data securely, efficiently, and cost-effectively.” - *Oracle*
- “*Data retrieval* means obtaining data from a Database Management System (DBMS)” - Wikipedia
  - Find or extract information

# Topics - Relational Database

- Introduction to Database Systems and Relational Model
- Intermediate & Advanced SQL
- Entity-Relationship Model & Relational Database Design
- Complex Data Types
- Big Data & Data Analytics
- Indexing and Hashing
- Query Processing
- Transactions



# — Topics - Semi-Structured Data Management

- XML
- JSON
- XPath & XQuery

# — Topics - Modern Data Stack

- ETL / ELT
- ELK - Elastic search for text
- No-SQL Database Systems

# Late Policy

- Each deadline will be extended once with penalty.
- You will lose 30% of the points for a project or homework if you submit your assignment by the extended deadline.
- No submission is allowed beyond the extended deadline.
- Please contact the instructor ASAP if something comes up.

*Thank You*