Types of Digital Data

15CSE334- Bigdata Analytiics

A.Baskar

AGENDA

- **Types of Digital Data**

- **Structured**
  - Sources of structured data
  - Ease with structured data

- **Semi-Structured**
  - Sources of semi-structured data
  - Characteristics of semi structured data

- **Unstructured**
  - Sources of unstructured data
  - Issues with terminology
  - Dealing with unstructured data

ABOUT DATA

- Data are the values of subjects with respect to qualitative or quantitative variables.
- *Data* is information that has been translated into a form that is efficient for movement or processing.

- Data may come from homogeneous and heterogeneous sources

- Data processing requirement:

- Data → Information
- Information → Insight (Knowledge)

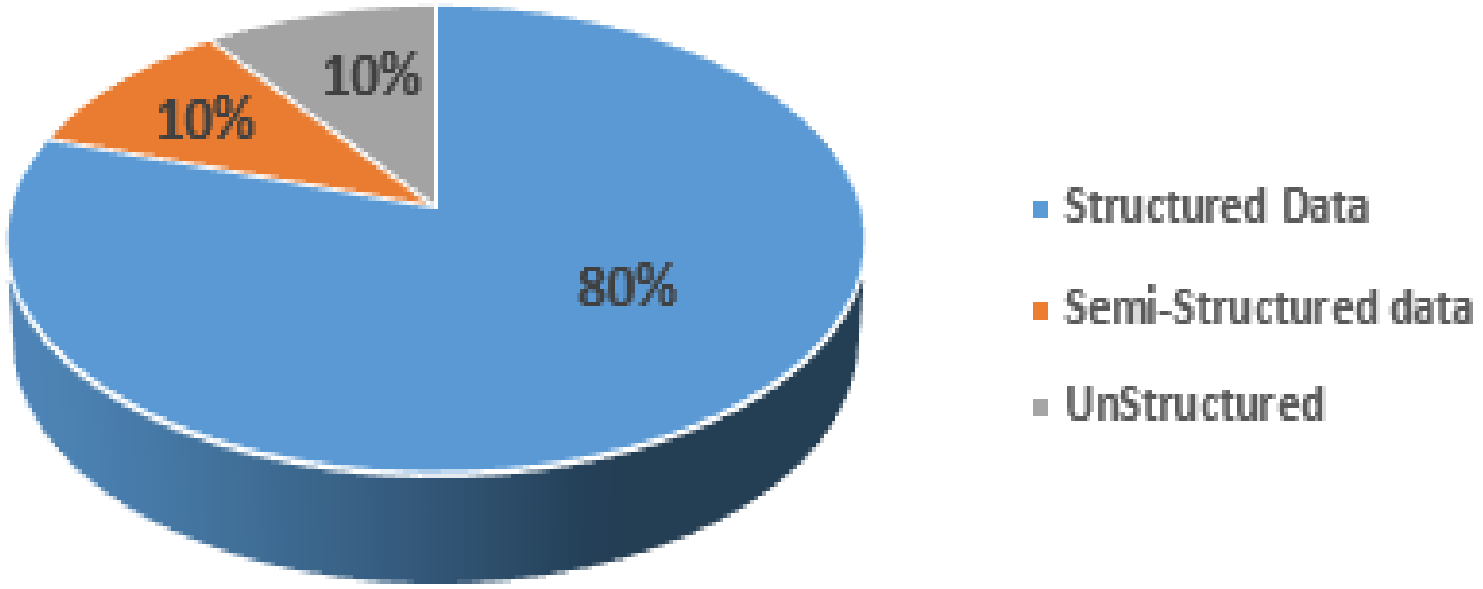
Classification of Digital data

- **Digital data is classified into the following categories:**

- **Structured data**
- **Semi-structured data**
- **Unstructured data**

Approximate Percentage Distribution of Digital Data

- **Approximate percentage distribution of digital data**



Structured Data

- This is the data which is in an organized form (e.g., in rows and columns) and can be easily used by a computer program - Relational data model
- Cardinality of Relation
- Degree of Relation
- Data type ,Constraints(Unique, Not Null)
- Relationships exist between entities of data, such as classes and their objects.
- Data stored in databases is an example of structured data.
- Example : Employee Data base

Sources of Structured Data

Structured Data at a Glance

Characteristics of Structured Data

- High organized
- Clearly defined
- Easy to access
- Easy to analyze

Examples of Structured Data

- Name
- Age
- Gender
- Address
- Phone number
- Currency
- Dates
- Billing info

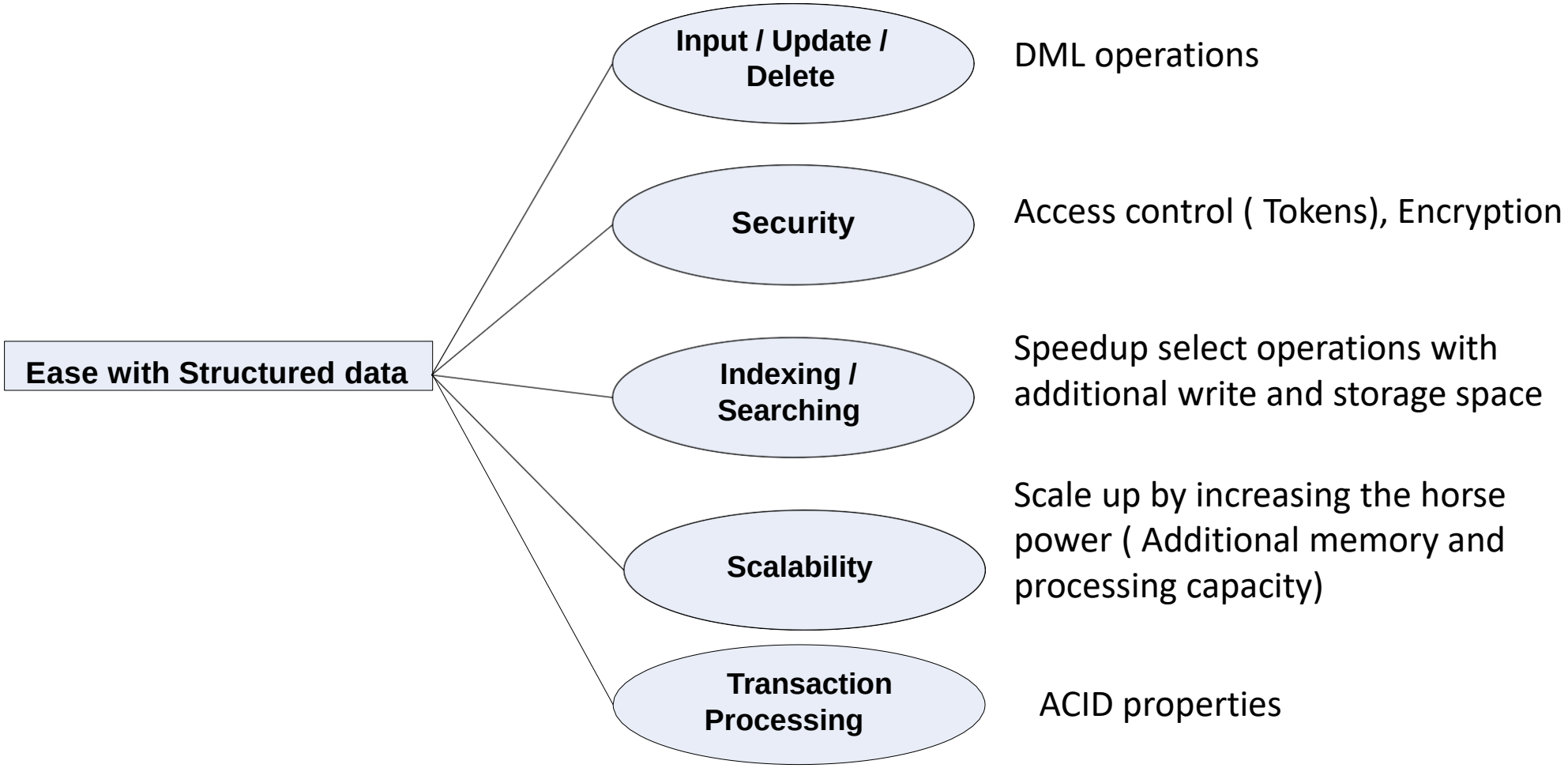
Sources of Structured Data

- SQL databases
- Spreadsheets
- Sensors
- Medical Devices
- Online Forms
- Point of Sales Systems
- Web and Server Logs

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

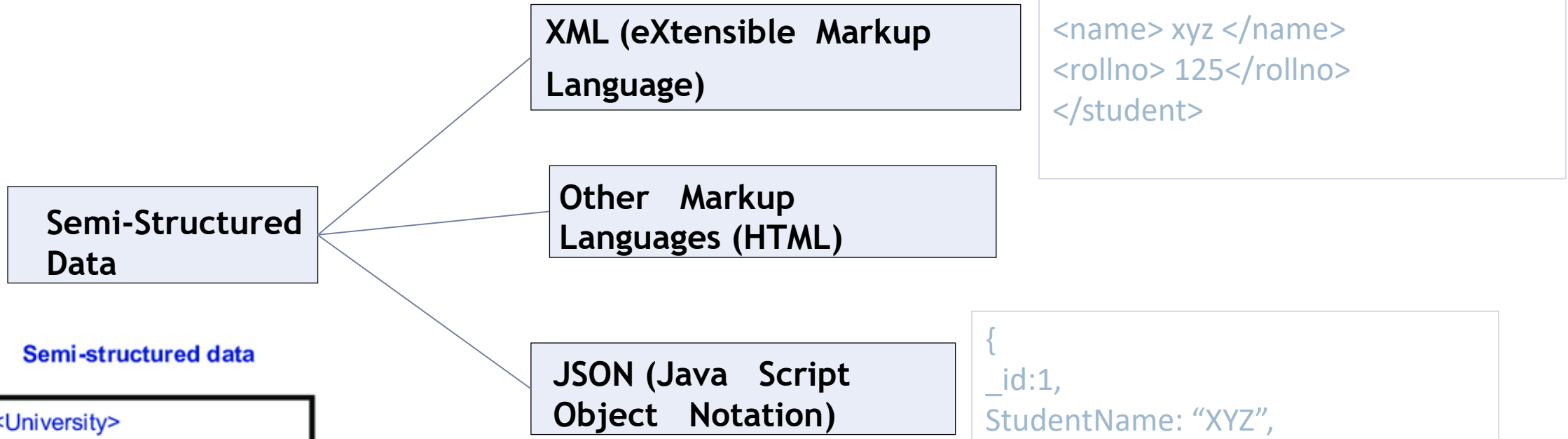
Ease with Structured Data



Semi-structured Data

- This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program.
- It uses tags to separate semantic elements and to enforce hierarchies of records and fields within data.
- No separation between schema and data.
- Metadata for this data is available but is not sufficient.
- Example: emails, XML, mark-up languages like HTML, etc.

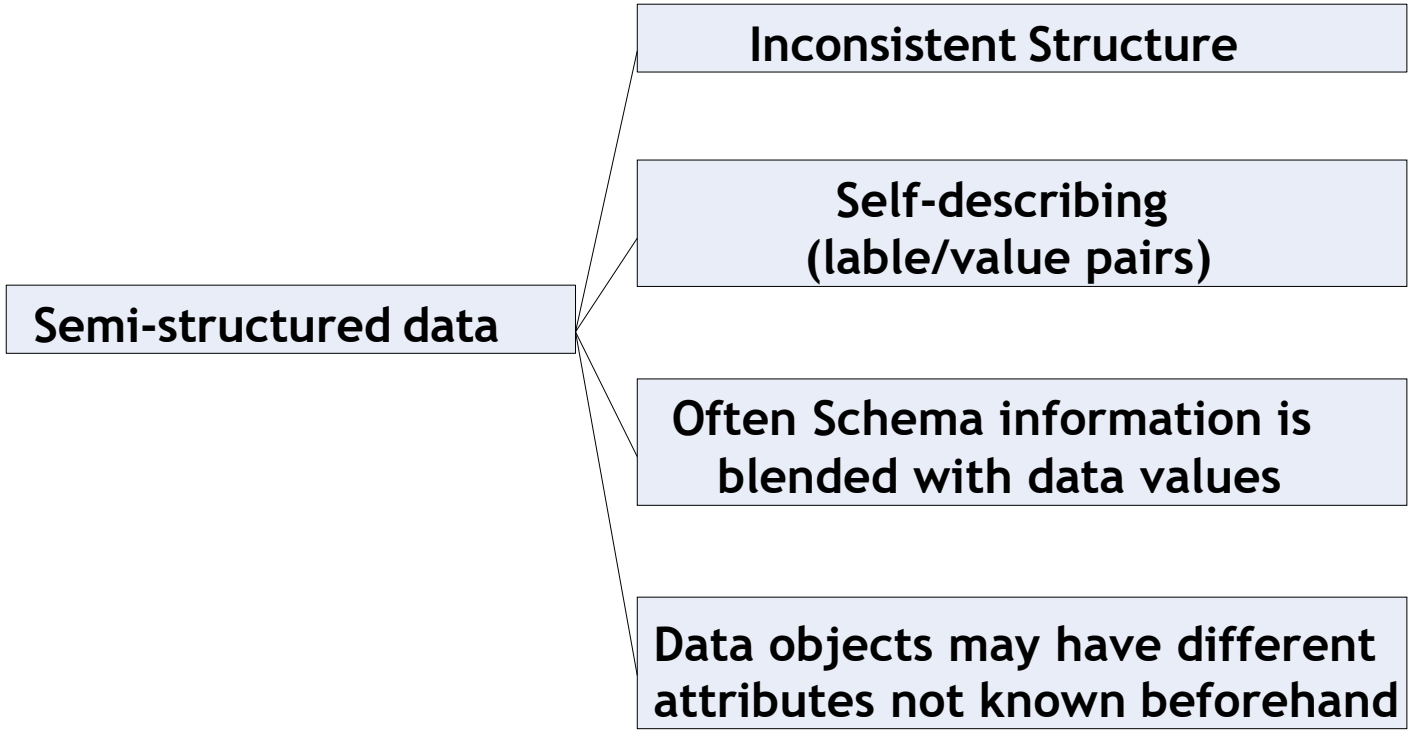
Sources of Semi-structured Data



Semi-structured data

```
<University>
 <Student ID="1">
   <Name>John</Name>
   <Age>18</Age>
   <Degree>B.Sc.</Degree>
 </Student>
 <Student ID="2">
   <Name>David</Name>
   <Age>31</Age>
   <Degree>Ph.D. </Degree>
</Student>
....
</University>
```

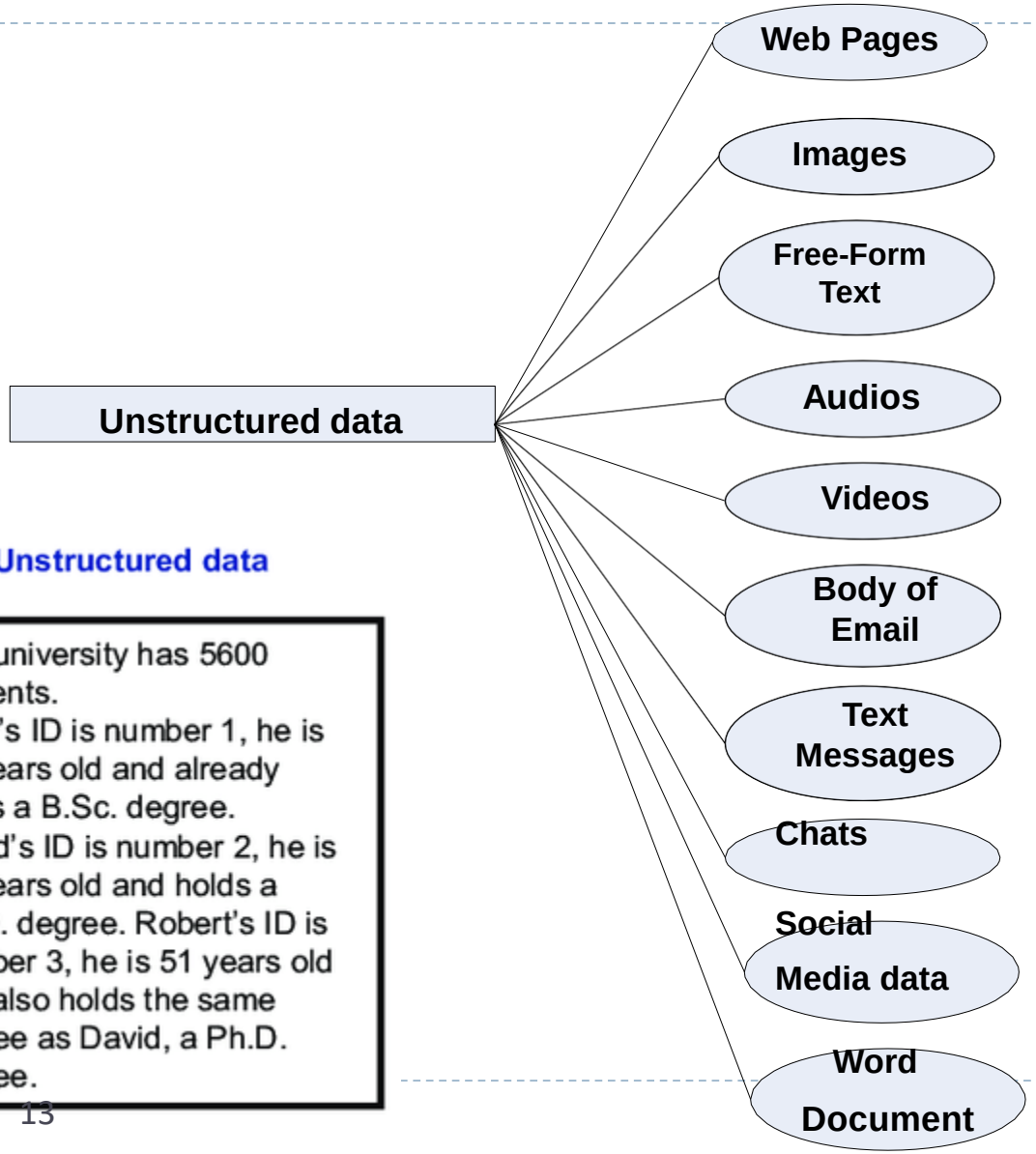
Characteristics of Semi-structured Data



Unstructured Data

- This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program.
- About 80–90% data of an organization is in this format.
- Example: memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc.

Sources of Unstructured Data



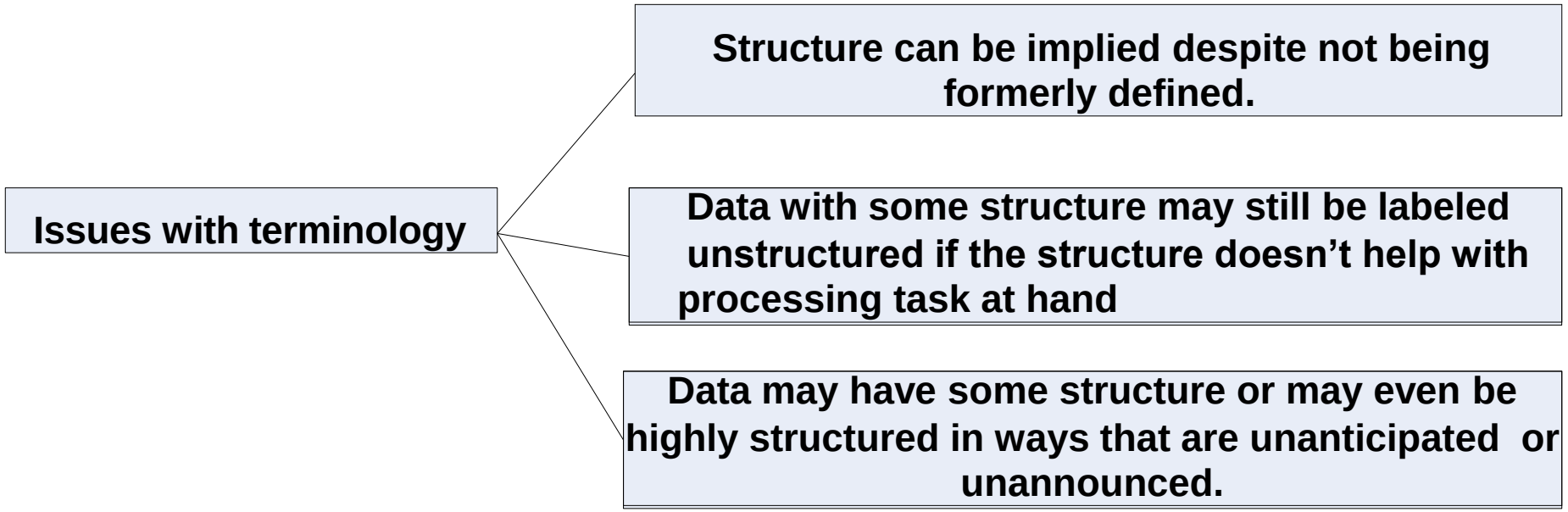
Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

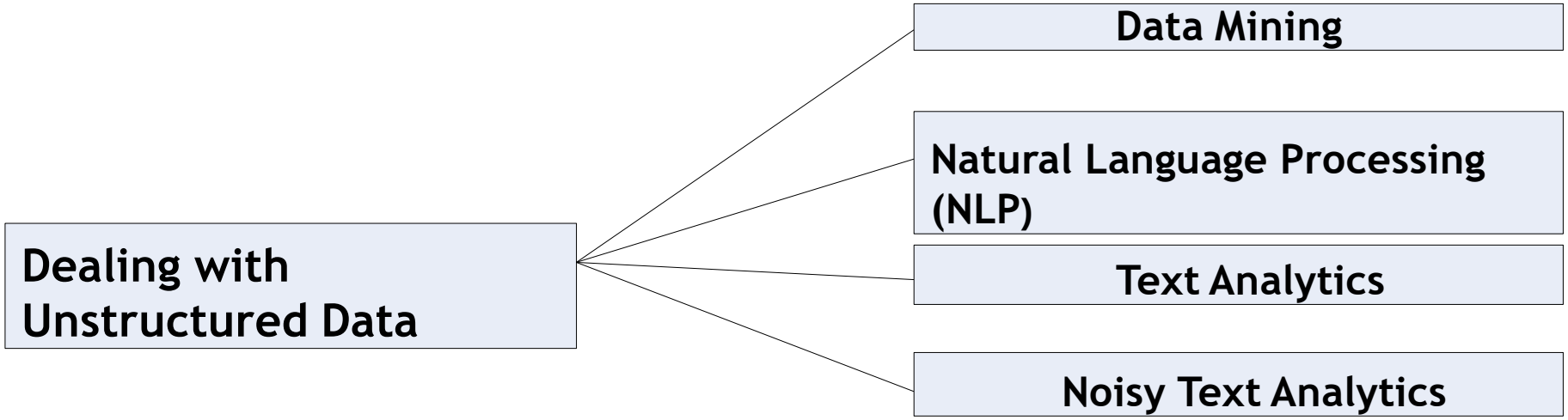
Table 1.4 Few examples of disparate unstructured data

Twitter message	Feeling miffed ☹. Victim of twishing.
Facebook post	LOL. C ya. BFN
Log files	127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I; Nav)"
Email	Hey Joan, possible to send across the first cut on the Hadoop chapter by Friday EOD or maybe we can meet up over a cup of coffee. Best regards, Tom

Issues with terminology – Unstructured Data



Dealing with Unstructured Data



7. **Unstructured Information Management Architecture (UIMA):** It is an open source platform from IBM. It is used for real-time content analytics. It is about processing text and other unstructured data to find latent meaning and relevant relationship buried therein. Read up more on UIMA at the link: http://www.ibm.com/developerworks/data/downloads/uima/

FEATURES	STRUCTURED	SEMI STRUCTURED	UNSTRUCTURED
Format Type	Relational Database	HTML, XML, JSON	Binary, Character
Version Management	Rows, columns, tuples	Not as common – graph is possible	Whole data
Implementation	SQL	Anonymous nodes	-
Robustness	Robust	Limited robustness	-
Storage Requirement	Less	Significant	Large
Applications	DBMS, RDF, ERP system, Data Warehouse, Apache Parquet, Financial Data, Relational Table	Server Logs, Sensor Output	No SQL, Video, Audio, Social Media, Online Forums, MRI, Ultrasound