Chapter 2
Types of Digital Data

Digital Data

- Today, data undoubtedly is an invaluable asset of any enterprise (big or small). Even though professionals work with data all the time, the understanding, management and analysis of data from heterogeneous sources remains a serious challenge.
- In this lecture, the various formats of digital data (structured, semi-structured and unstructured data), data storage mechanism, data access methods, management of data, the process of extracting desired information from data, challenges posed by various formats of data, etc. will be explained.
- Data growth has seen exponential acceleration since the advent of the computer and Internet.

Digital Data

In fact, the computer and Internet duo has imparted the digital form to data.

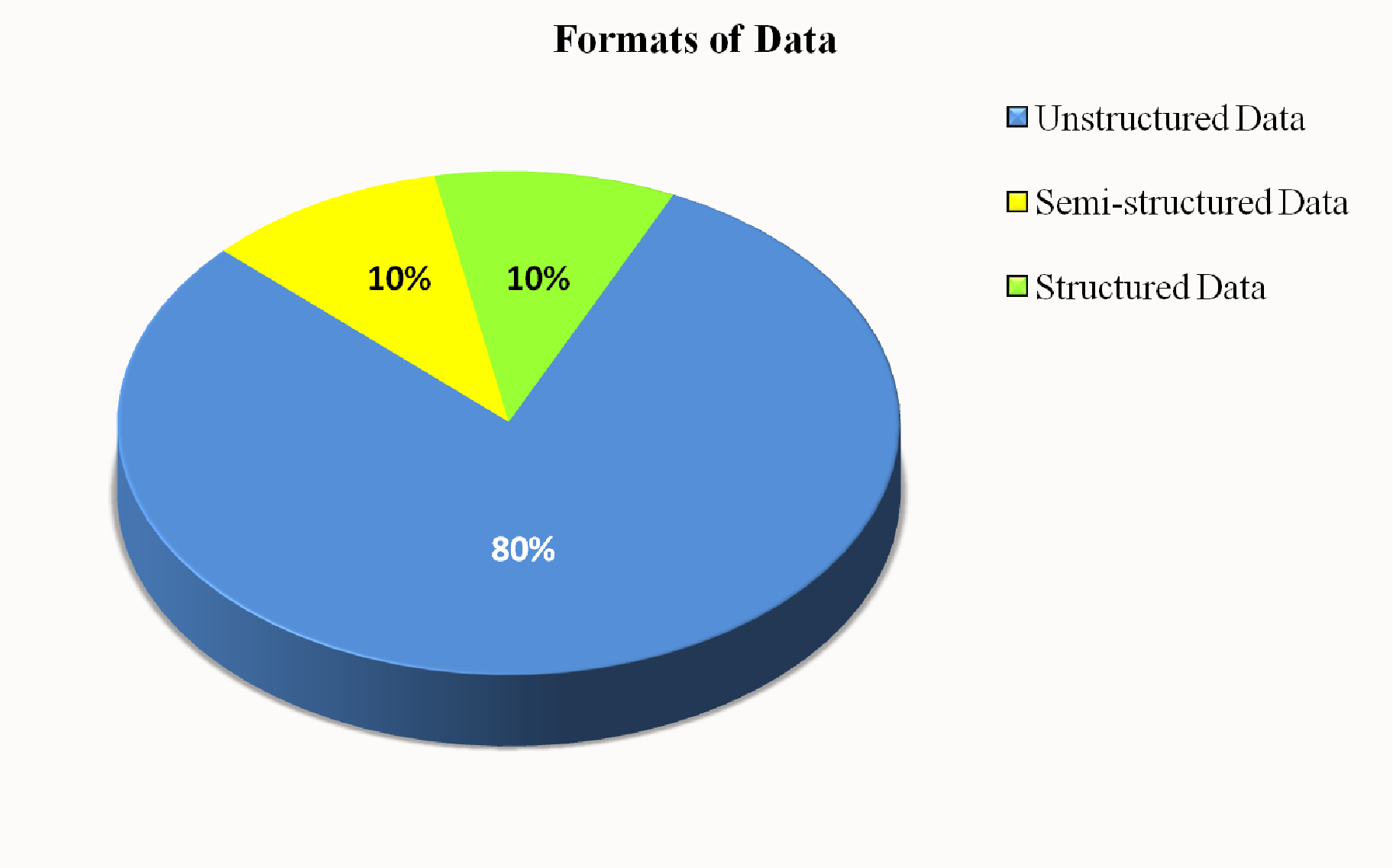
Digital data can be classified into three forms:

- Unstructured
- Semi-structured
- Structured

- Usually, data is in the unstructured format which makes extracting information from it difficult.
- According to Merrill Lynch, 80–90% of business data is either unstructured or semi-structured.
- Gartner also estimates that unstructured data constitutes 80% of the whole enterprise data.

Formats of Digital Data

Here is a percent distribution of the three forms of data -



Data Forms Defined-

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; for example, memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc.

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; for example, emails, XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.

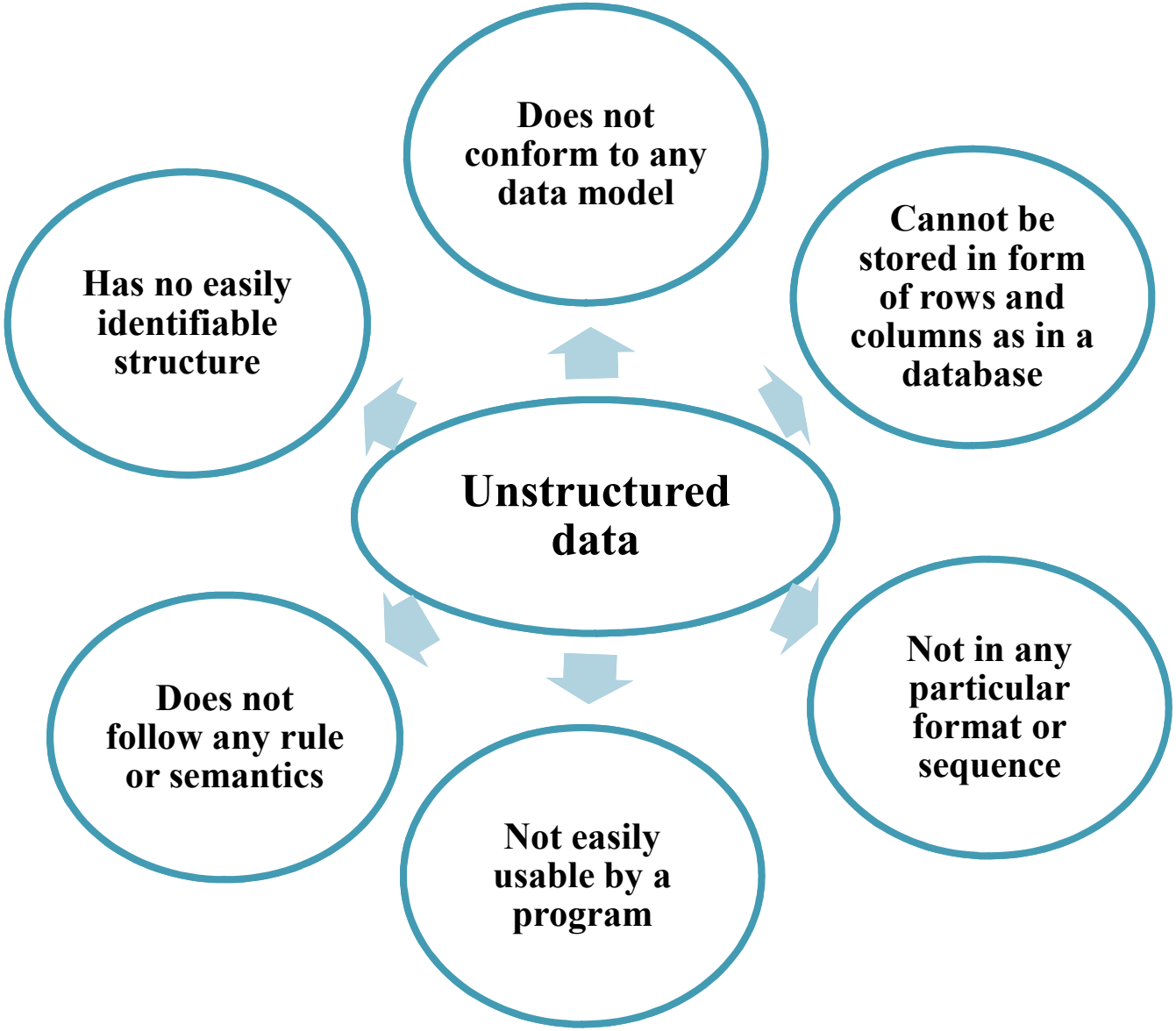
Structured data: This is the data which is in an organized form (e.g., in rows and columns) and can be easily used by a computer program. Relationships exist between entities of data, such as classes and their objects. Data stored in databases is an example of structured data.

Unstructured Data

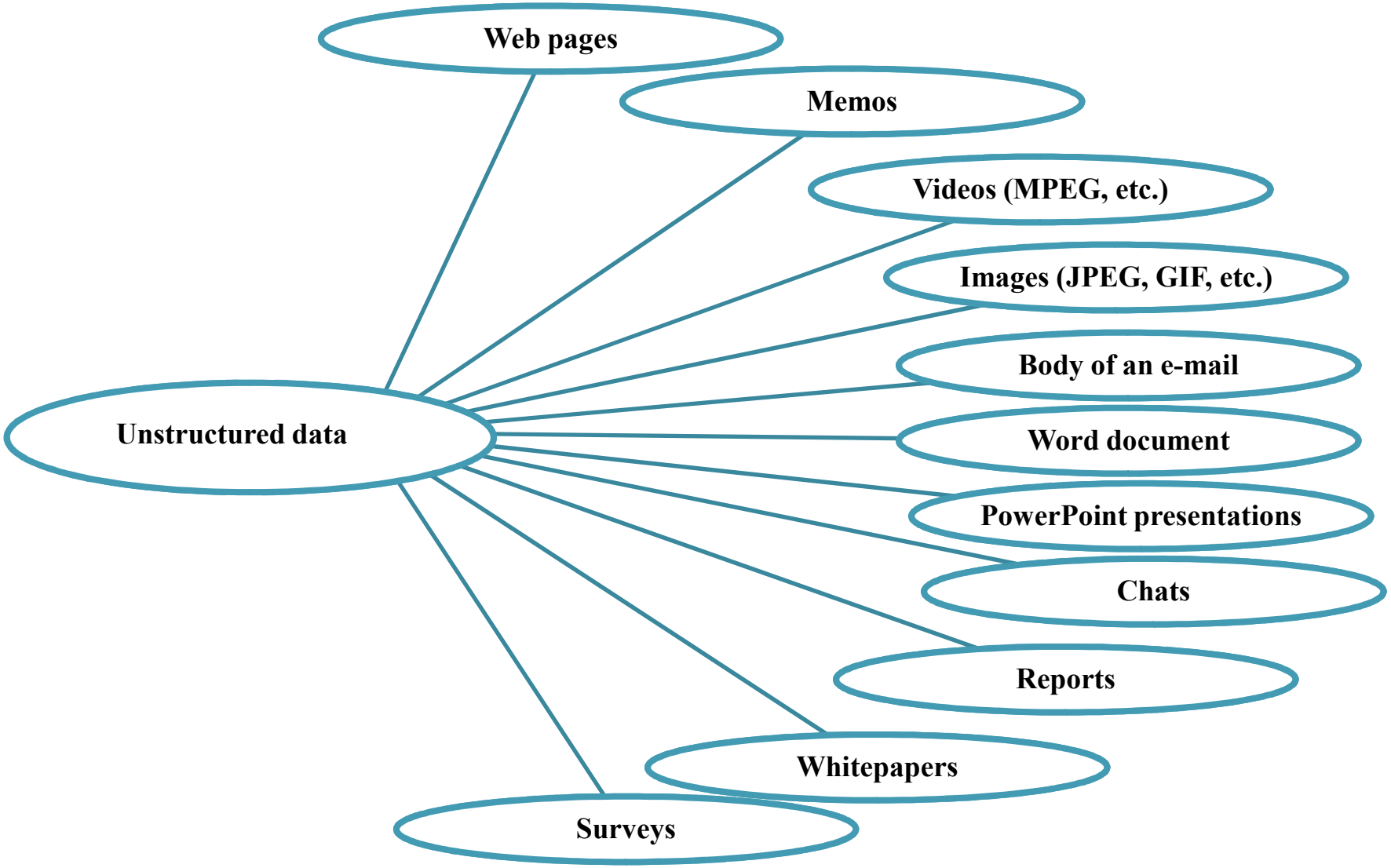
Unstructured Data – Getting to Know

- Dr. Ben, Dr. Stanley, and Dr. Mark work at the medical facility of “GoodLife”. Over the past few days, Dr. Ben and Dr. Stanley had been exchanging long emails about a particular case of testinal problem. Dr. Stanley has chanced upon a particular combination of drugs that has cured gastro-intestinal disorders in his patients. He has written an email about this combination of drugs to Dr. Ben.
- Dr. Mark has a patient in the “GoodLife” emergency unit with quite a similar case of gastro-intestinal disorder whose cure Dr. Stanley has chanced upon. Dr. Mark has already tried regular drugs but with no positive results so far. He quickly searches the organization's database for answers, but with no luck. The information he wants is tucked away in the email conversation between two other “GoodLife” doctors, Dr. Ben and Dr. Stanley. Dr. Mark would have accessed the solution with few mouse clicks had the storage and analysis of unstructured data been undertaken by “GoodLife”.
- As is the case at “GoodLife”, 80-85% of data in any organization is unstructured and is an alarming rate. An enormous amount of knowledge is buried in this data. In the above Stanley's email to Dr. Ben had not been successfully updated into the medical system in the unstructured format.
- Unstructured data, thus, is the one which cannot be stored in the form of rows and as in a database and does not conform to any data model, i.e. it is difficult to determine the meaning of the data. It does not follow any rules or semantics. It can be of any type and is hence unpredictable.

Characteristics of Unstructured Data



Where does Unstructured Data Come from?



Where does Unstructured Data Come from?

- Broadly speaking, anything in a non-database form is unstructured data.

- It can be classified into two broad categories:
  - Bitmap objects : For example, image, video, or audio files.
  - Textual objects : For example, Microsoft Word documents, emails, or Microsoft Excel spread-sheets.

- Refer to figure in the previous slide - Let us take the above example of the email communication between Dr. Ben and Dr. Stanley. Even though email messages like the ones exchanged by Dr. Ben and Dr. Stanley are organized in databases such as Microsoft Exchange or Lotus Notes, the body of the email is essentially raw data, i.e. free form text without any structure.
- A lot of unstructured data is also noisy text such as chats, emails and SMS texts.
- The language of noisy text differs significantly from the standard form of language.

A Myth Demystified

- Web pages are said to be unstructured data even though they are defined by HTML, a markup language which has a rich structure.
- HTML is solely used for rendering and presentations.
- The tagged elements do not capture the meaning of the data that the HTML page contains. This makes it difficult to automatically process the information in the HTML page.
- Another characteristic that makes web pages unstructured data is that they usually carry links and references to external unstructured content such as images, XML files, etc.

How to Manage Unstructured Data?

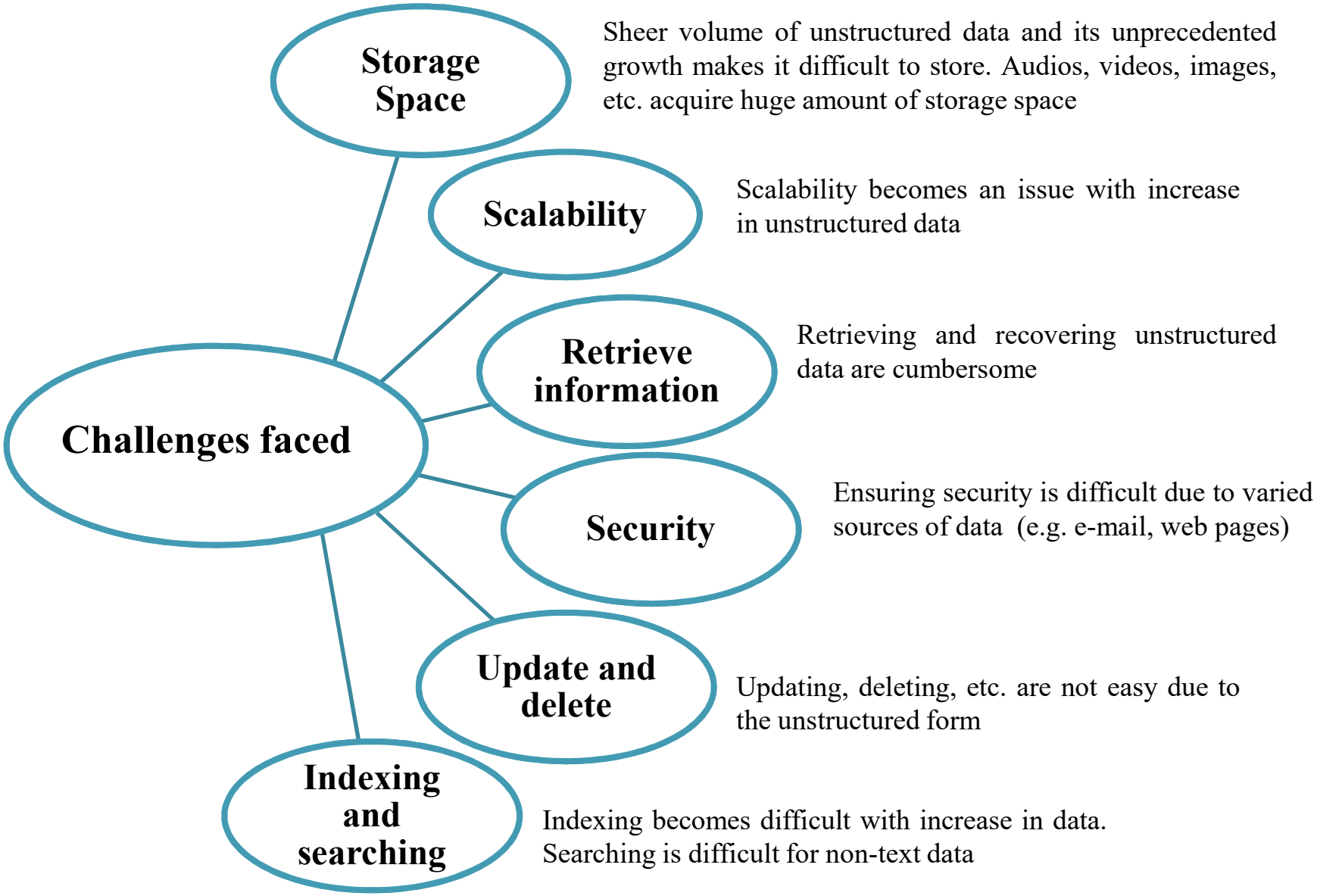
Let us look at a few generic tasks to be performed to enable storage and search of unstructured data:

Indexing: Let us go back to our understanding of the Relational Database Management System(RDBMS). In this system, data is indexed to enable faster search and retrieval. On the basis of some value in the data, index is defined which is nothing but an identifier and represents the large record in the data set. In the absence of an index, the whole data set/ document will be scanned for retrieving the desired information. In the case of unstructured data too, indexing helps in searching and retrieval. Based on text or some other attributes, e.g. file name, the unstructured data is indexed. Indexing in unstructured data is difficult because neither does this data have any predefined attributes nor does it follow any pattern or naming conventions. Text can be indexed based on a text string but in case of non-text based files, e.g. audio/video, etc., indexing depends on file names. This becomes a hindrance when naming conventions are not being followed.

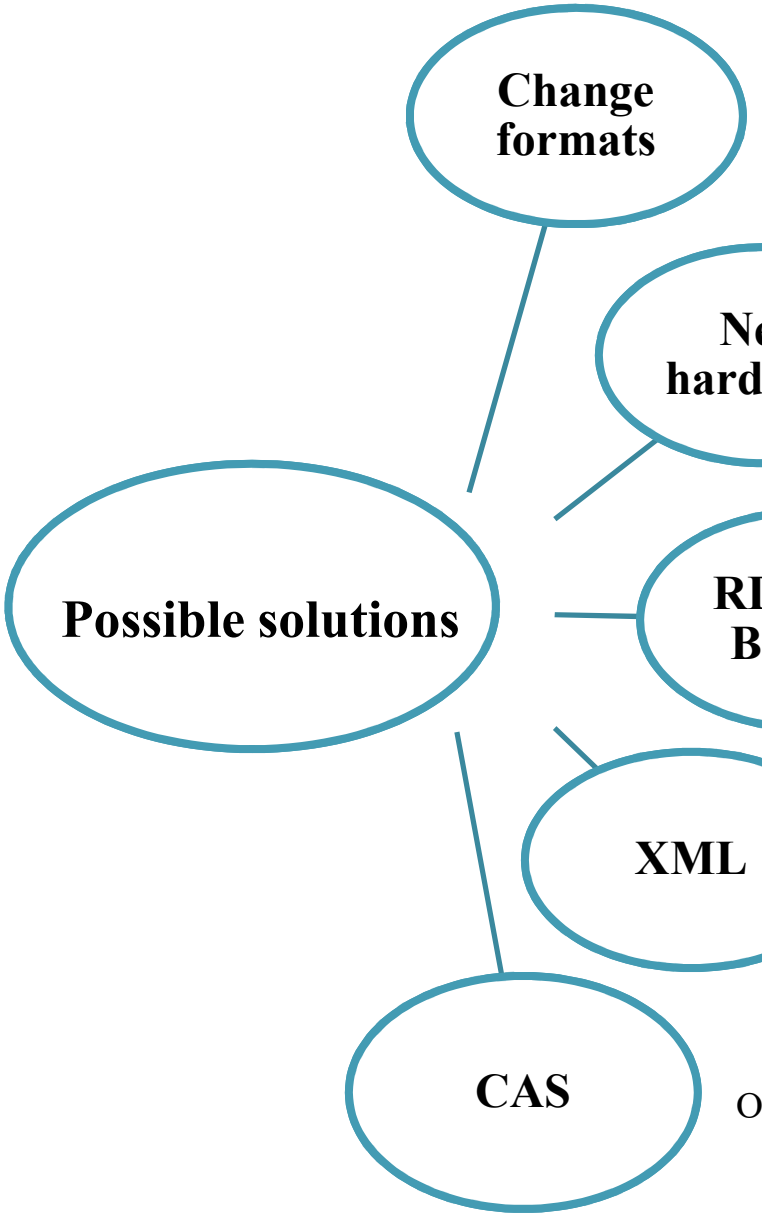
Tags/Metadata: Using metadata, data in a document, etc. can be tagged. This enables search and retrieval. But in unstructured data, this is difficult as little or no metadata is available. Structure of data has to be determined which is very difficult as the data itself has no particular format and is coming from more than one source.

Classification/Taxonomy: Taxonomy is classifying data on the basis of the relationships that exist between data. Data can be arranged in groups and placed in hierarchies based on the taxonomy prevalent in an organization. However, classifying unstructured data is difficult asidentifying relationships between data is not an easy task. In the absence of any structure ormetadata or schema, identifying accurate relationships and classifying is not easy. Since the datais unstructured, naming conventions or standards are not consistent across an organization, thusmaking it difficult to classify data.CAS (Content Addressable Storage): It stores data based on their metadata. It assigns 2 uniquename to every object stored in it. The object is retrieved based on its content and not its location.It is used extensively to store emails, etc.

How to Store Unstructured Data?



How to Store Unstructured Data?



Unstructured data may be be converted to formats which are easily managed, stored and searched. For example, IBM is working on providing a solution which converts audio , video, etc. to text

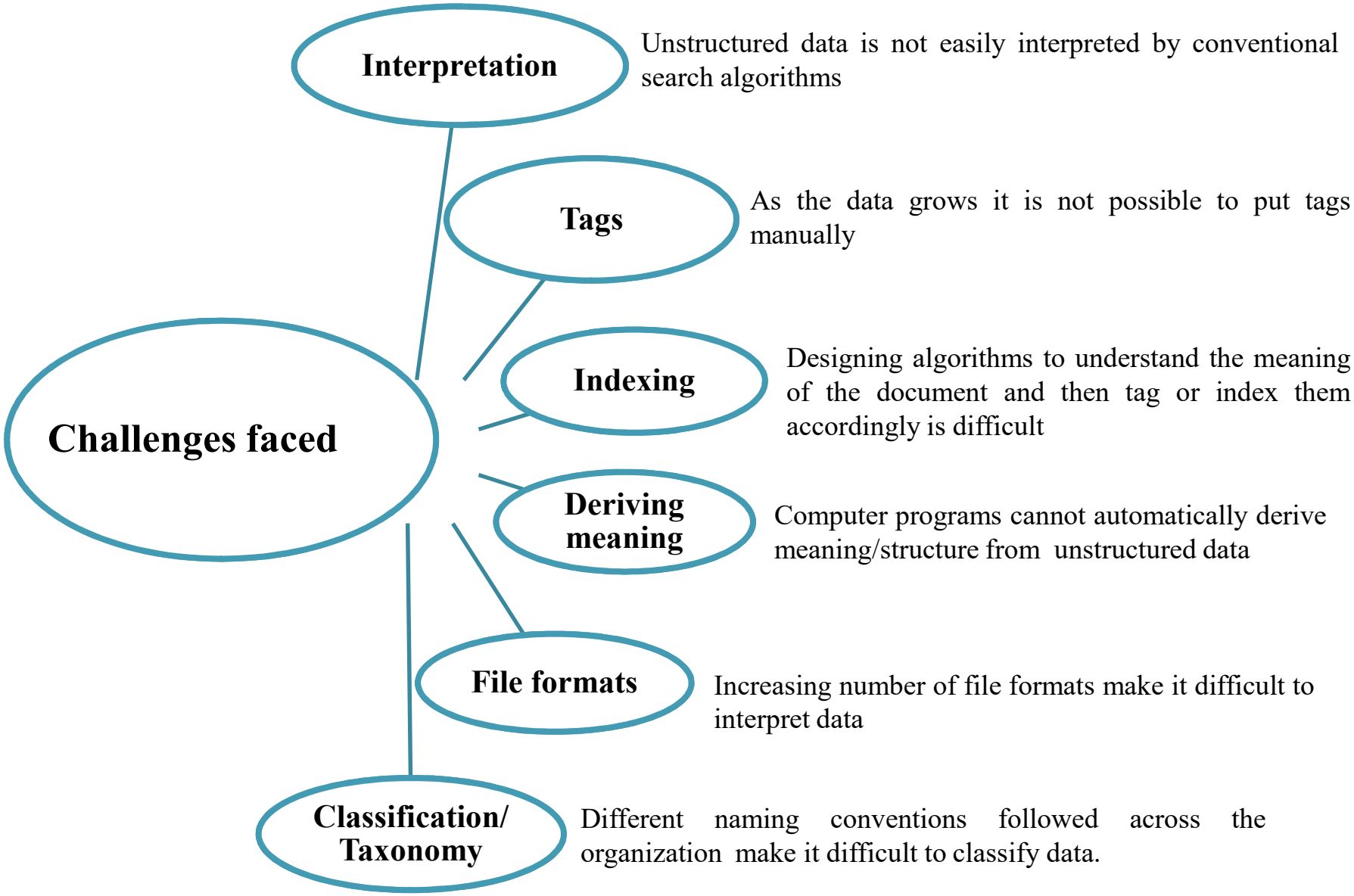
Create hardware which support unstructured data either compliment the existing storage devices or be a stand alone for unstructured data

Store in relational databases which support BLOBs which is Binary Large Objects

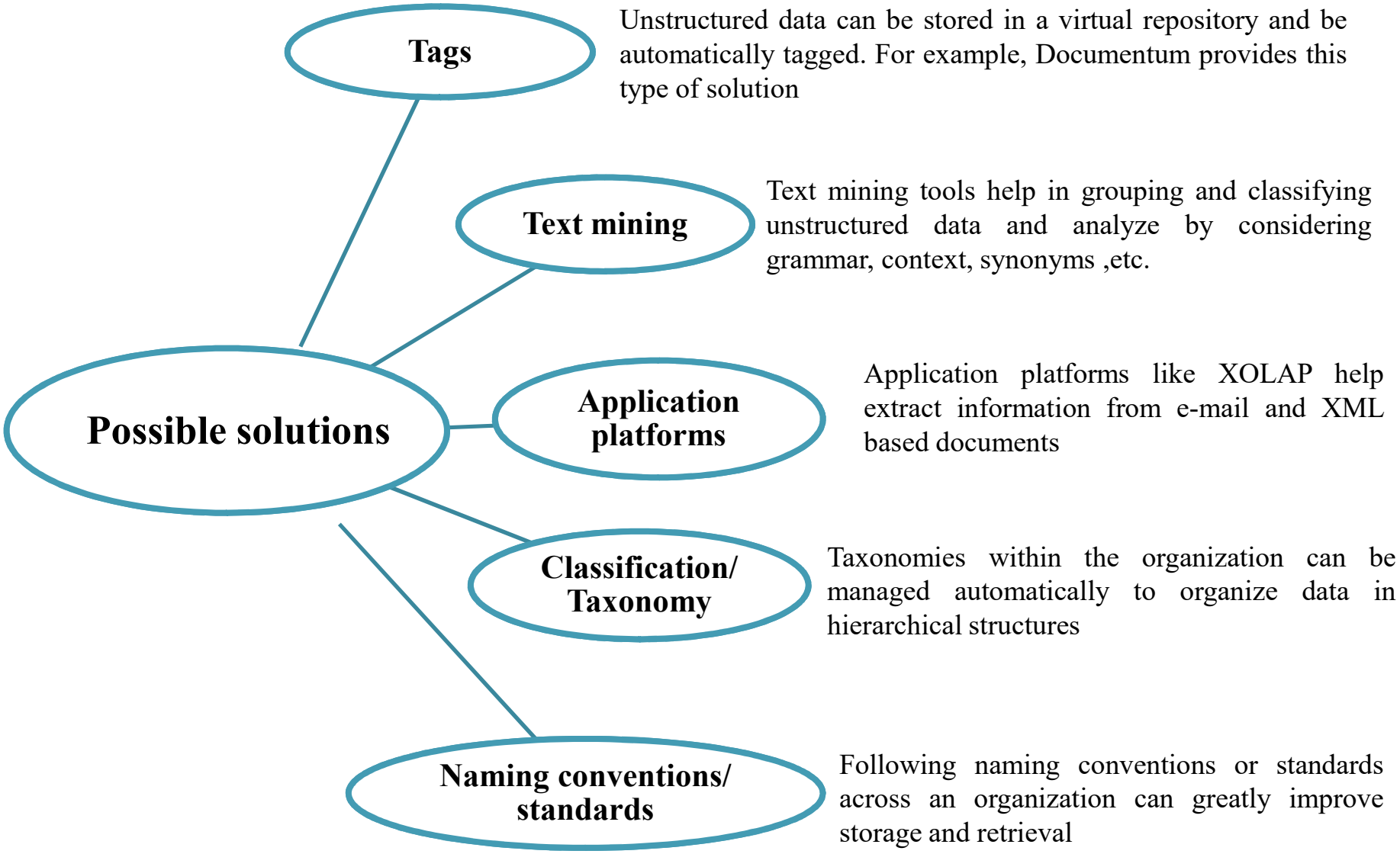
Store in XML which tries to give some structure to unstructured data by using tags and elements

Organize files based on their metadata

How to Extract Information from Unstructured Data?



How to Extract Information from Unstructured Data?



UIMA

- ✓ UIMA (Unstructured Information Management Architecture) is an open source platform from IBM which integrates different kinds of analysis engines to provide a complete solution for edge discovery from unstructured data.
- ✓ In UIMA, the analysis engines integration and analysis of unstructured information and bridge the gap between structured and unstructured data.
- ✓ UIMA stores information in a structured format. The structured resources can be mined, searched, and put to other uses. The information obtained from structured sources is also for sub-sequent analysis of unstructured data.
- ✓ Various analysis engines analyze unstructured data in different ways such as:
  - Breaking up of documents into separate words.
  - Grouping and classifying according to taxonomy.
  - Detecting parts of speech, grammar, and synonyms.
  - Detecting events and times.¢ Detecting relationships between various elements.

Further Reading

- *http://www.information-management.com/issues/20030201/6287-1.html*
- *http://www.enterpriseitplanet.com/storage/features/article.php/11318_3407161_2*
- *http://domino.research.ibm.com/comm/research_projects.nsf/pages/uima.index.html*
- *http://www.research.ibm.com/UIMA/UIMA%20Architecture%20Highlights.html*

Answer a Quick Question

Ask the participants of the learning program to state some more examples of **Unstructured data**

Do it Exercise

Search, think and write about two best practices for managing the growth of unstructured data

Semi-structured Data

Semi-structured Data

- Semi-structured data does not conform to any data model i.e. it is difficult to determine the meaning of data neither can data be stored in rows and columns as in a database but semi-structured data has tags and markers which help to group data and describe how data is stored, giving some metadata but it is not sufficient for management and automation of data.

- Similar entities in the data are grouped and organized in a hierarchy. The attributes or the properties within a group may or may not be the same. For example two addresses may or may not contain the same number of properties as in

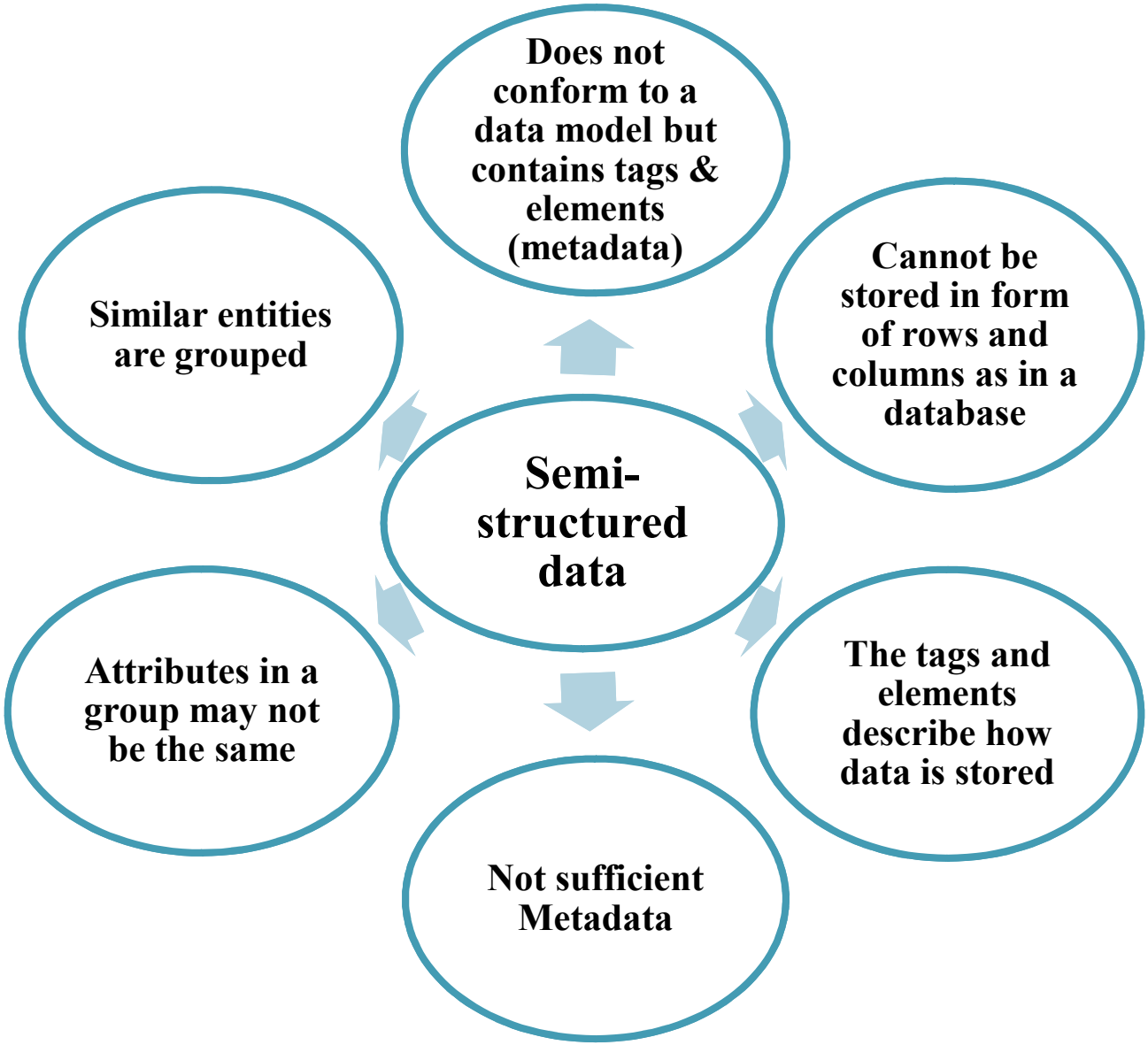
  Address 1
  <house number><street name><area name><city>
  Address 2
  <house number><street name><city>

- For example an e-mail follows a standard format

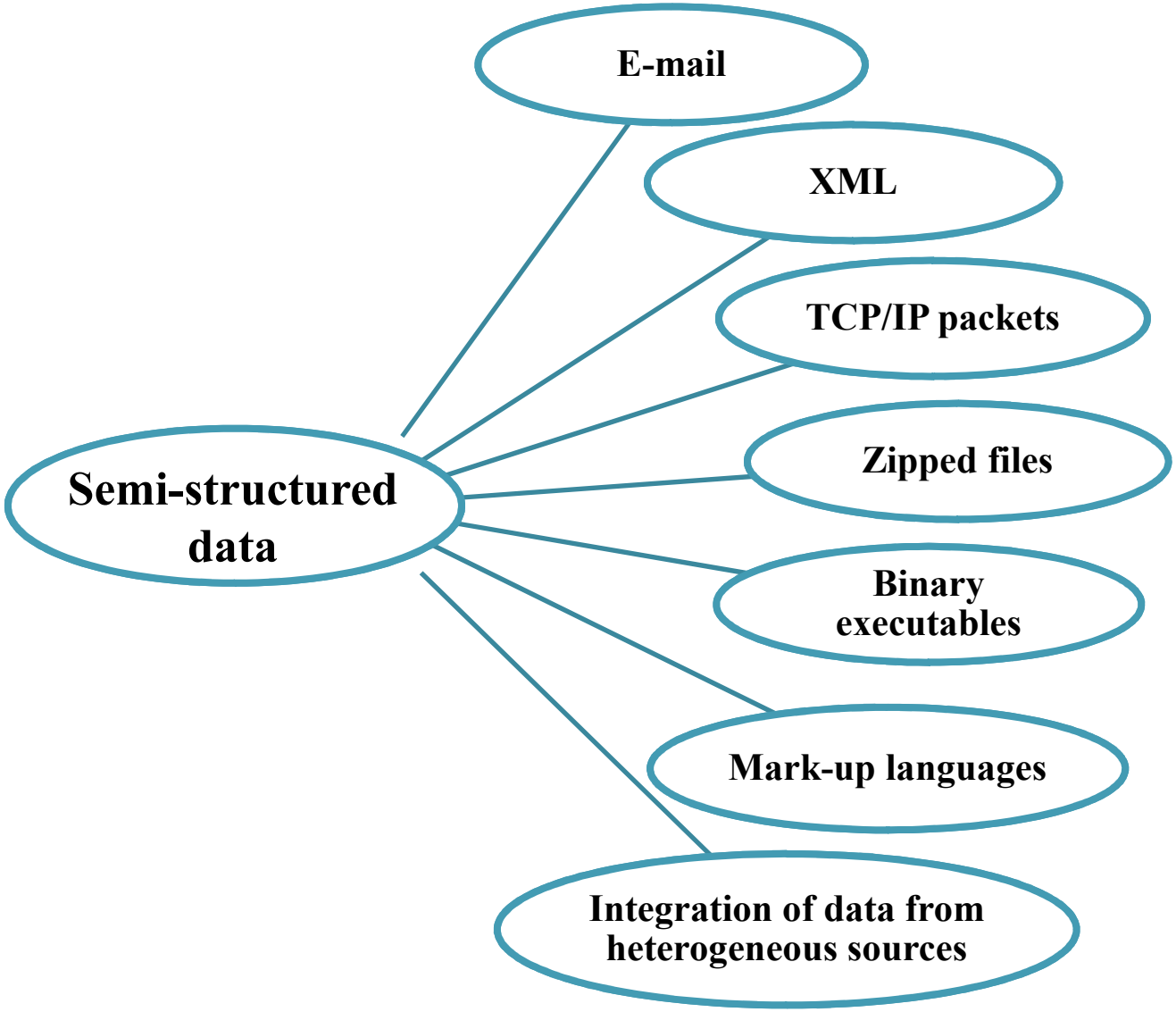
  To: <Name>
  From: <Name>
  Subject: <Text>
  CC: <Name>
  Body: <Text, Graphics, Images etc. >

- The tags give us some metadata but the body of the e-mail contains no format neither is such which conveys meaning of the data it contains.
- There is very fine line between unstructured and semi-structured data.

What is Semi-structured Data?



Where does Semi-structured Data Come from?

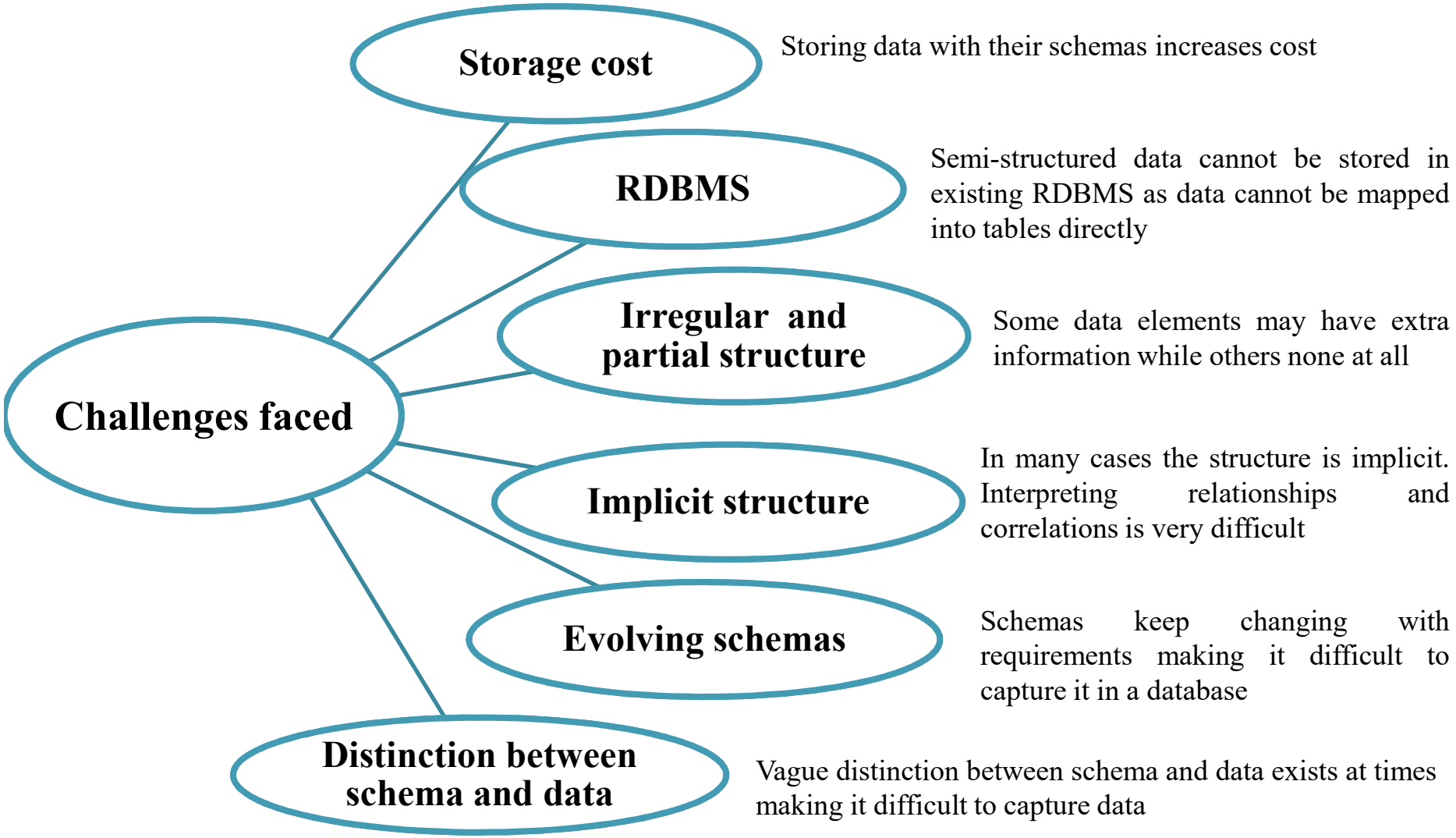


How to Manage Semi-structured Data?

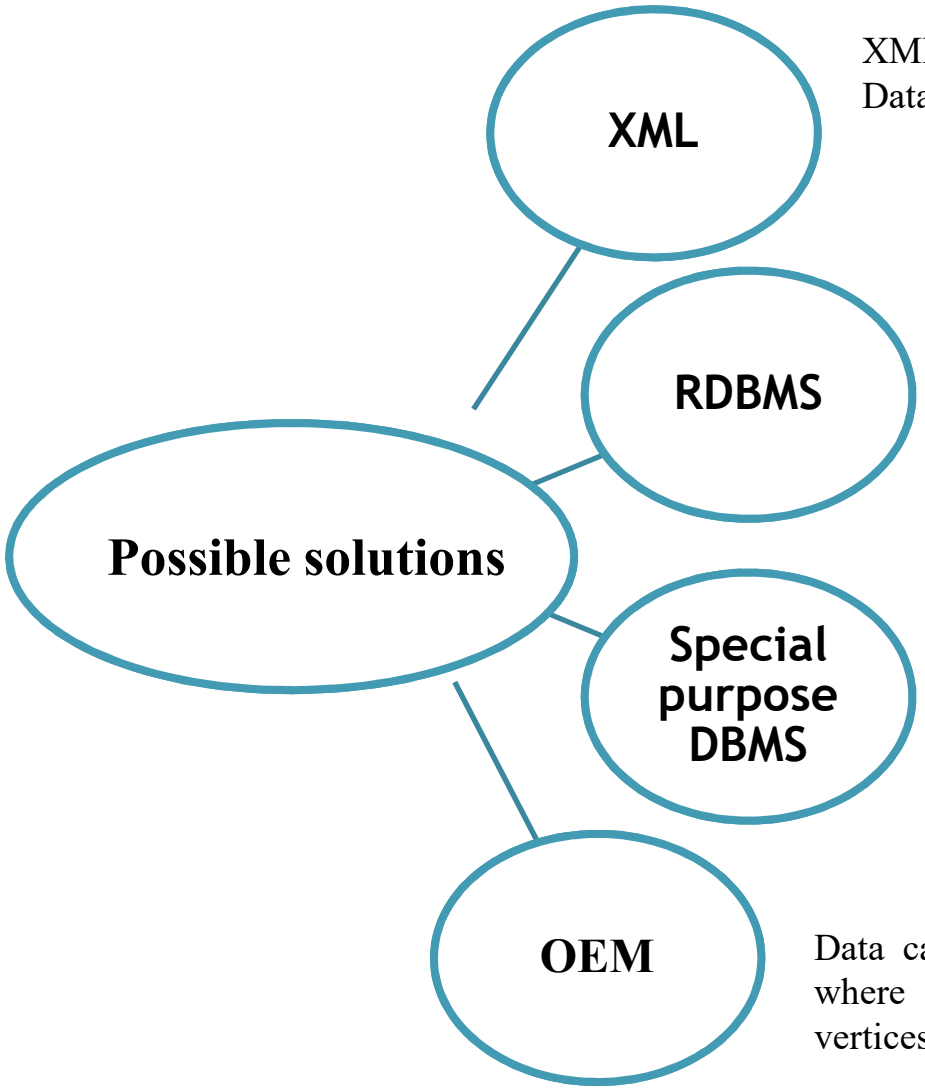
Some ways in which semi-structured data is managed and stored

Schemas	Graph-based data models	XML
• Describe the structure and content of data to some extent • Assign meaning to data hence allowing automatic search and indexing	• Contain data on the leaves of the graph. Also known as 'schema less' • Used for data exchange among heterogeneous sources	• Models the data using tags and elements • Schemas are not tightly coupled to data

How to Store Semi-structured Data?



How to Store Semi-structured Data?



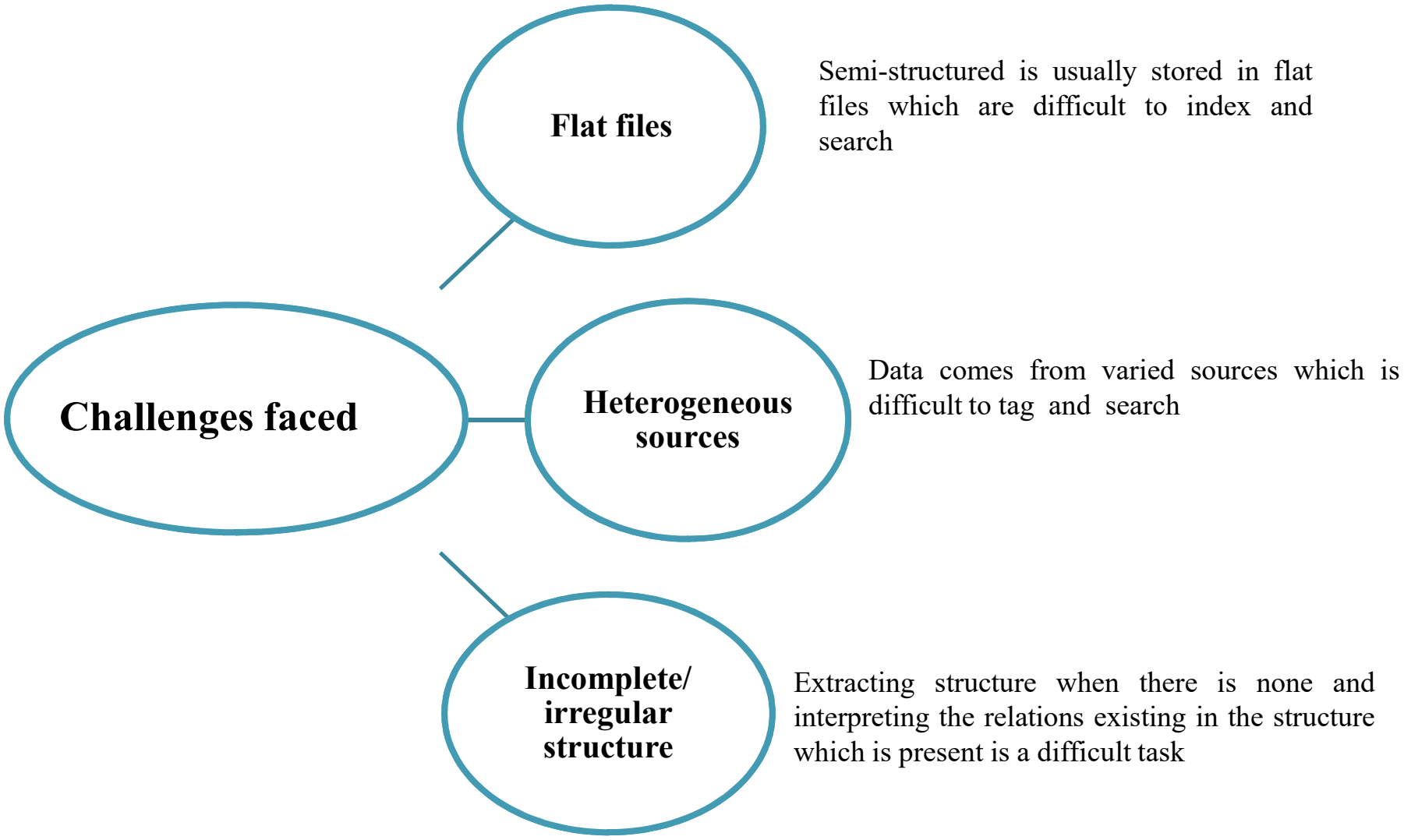
XML allows to define tags and attributes to store data. Data can be stored in a hierarchical/nested structure

Semi-structured data can be stored in a relational database by mapping the data to a relational schema which is then mapped to a table

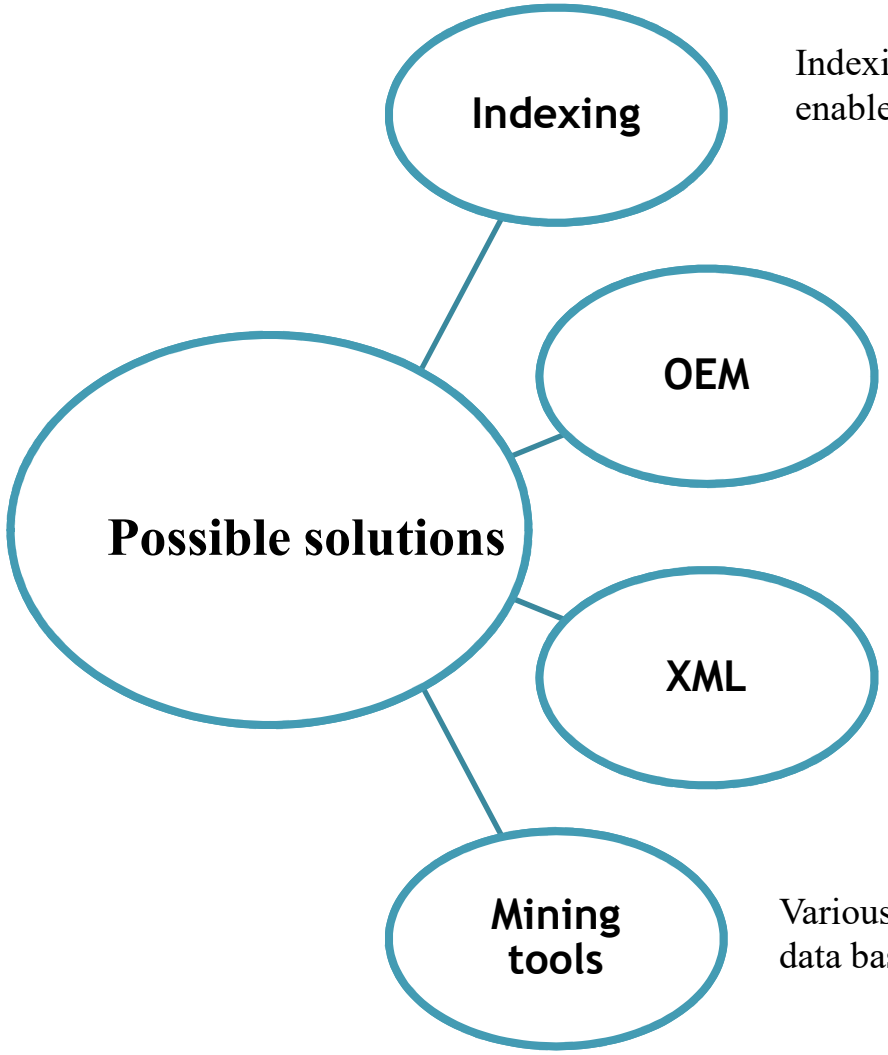
Databases which are specifically designed to store semi-structured data

Data can be stored and exchanged in the form of graph where entities are represented as objects which are the vertices in a graph

How to Extract Information from Semi-structured Data?



How to Extract Information from Semi-structured Data?



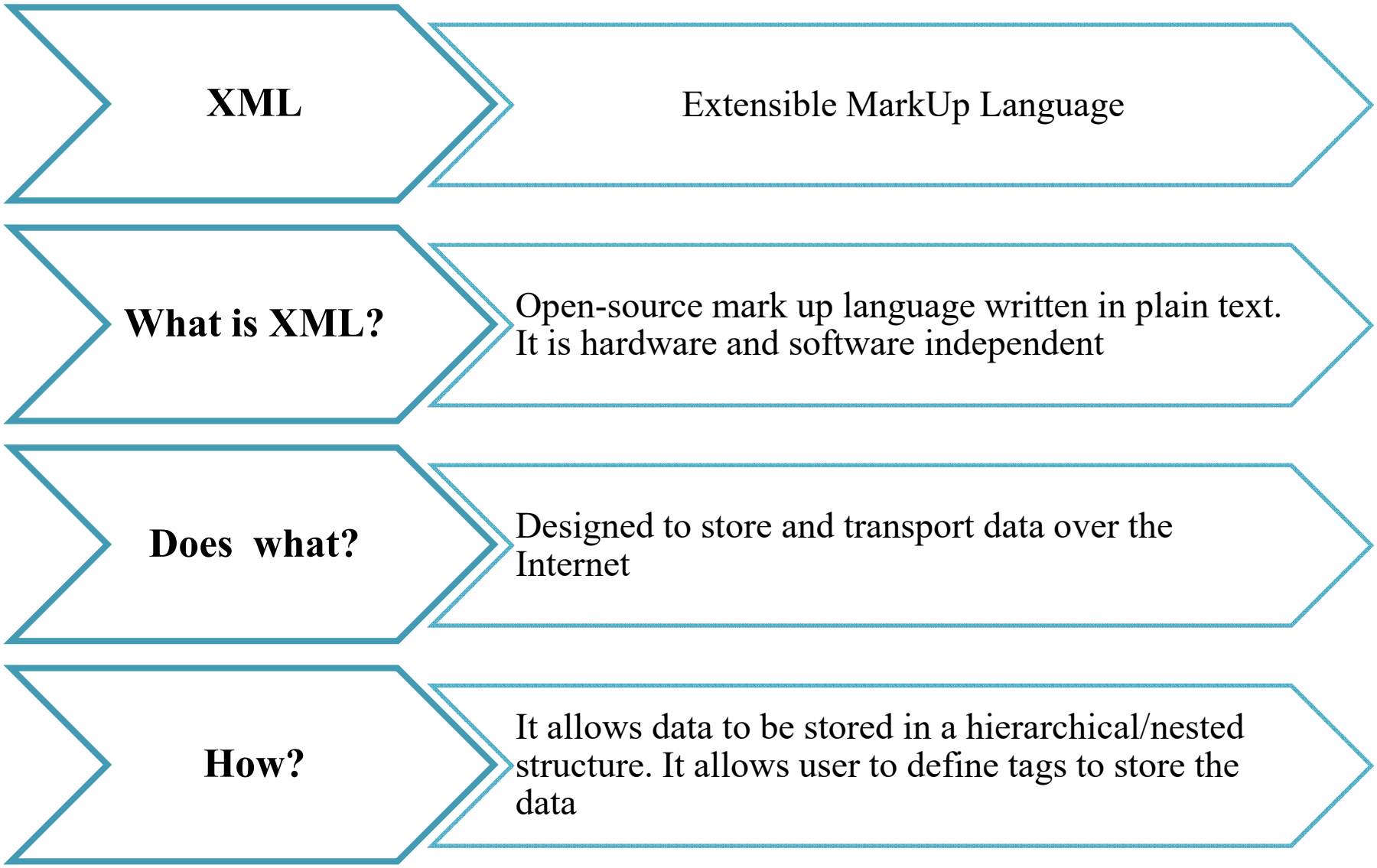
Indexing data in a graph-based model enables quick search

Allows data to be stored in a graph-based data model which is easier to index and search

Allows data to be arranged in a hierarchical or tree-like structure which enables indexing and searching

Various mining tools are available which search data based on graphs, schemas, structure, etc.

XML – A Solution for Semi-structured Data Management



XML – A Solution for Semi-structured Data Management

XML has no predefined tags

```
<message>
<to> XYZ </to>
<from> ABC </from>
<subject> Greetings </subject>
<body> Hello! How are you? </body>
</message>
```

The words in the <> (angular brackets) are user-defined tags

XML is known as self-describing as data can exist without a schema and schema can be added later

Schema can be described in XSLT or XML schema

Further Reading

- *http://queue.acm.org/detail.cfm?id=1103832*
- *http://www.computerworld.com/s/article/93968/Taming_Text*
- *http://searchstorage.techtarget.com/generic/0,295582,sid5_gci1334684,00.html*
- *http://searchdatamanagement.techtarget.com/generic/0,295582,sid91_gci1264550,00.html*
- *http://searchdatamanagement.techtarget.com/news/article/0,289142,sid91_gci1252122,00.html*

Answer a Quick Question

What is your take on this….

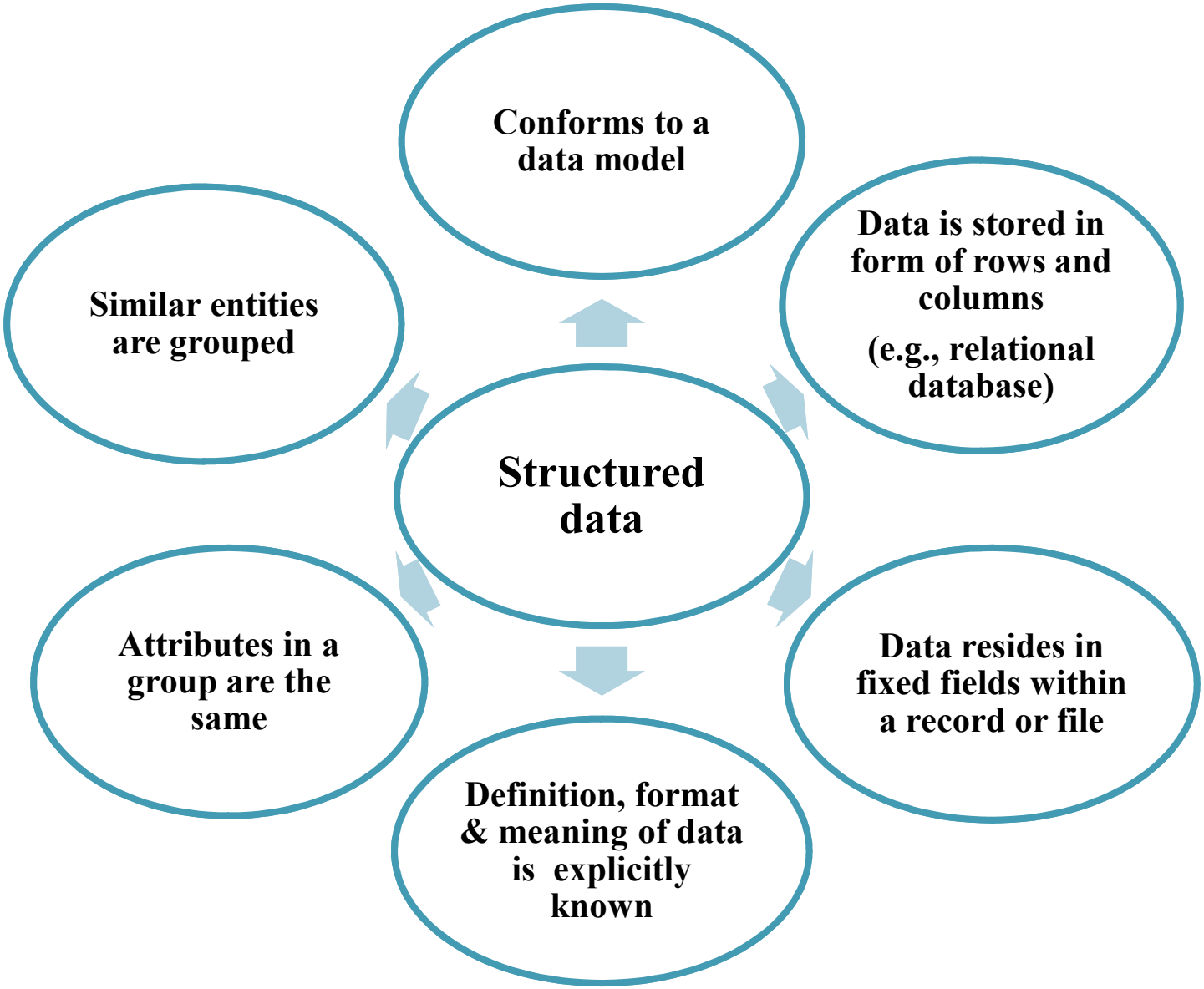
A Web Page is unstructured. If yes, why?

Structured Data

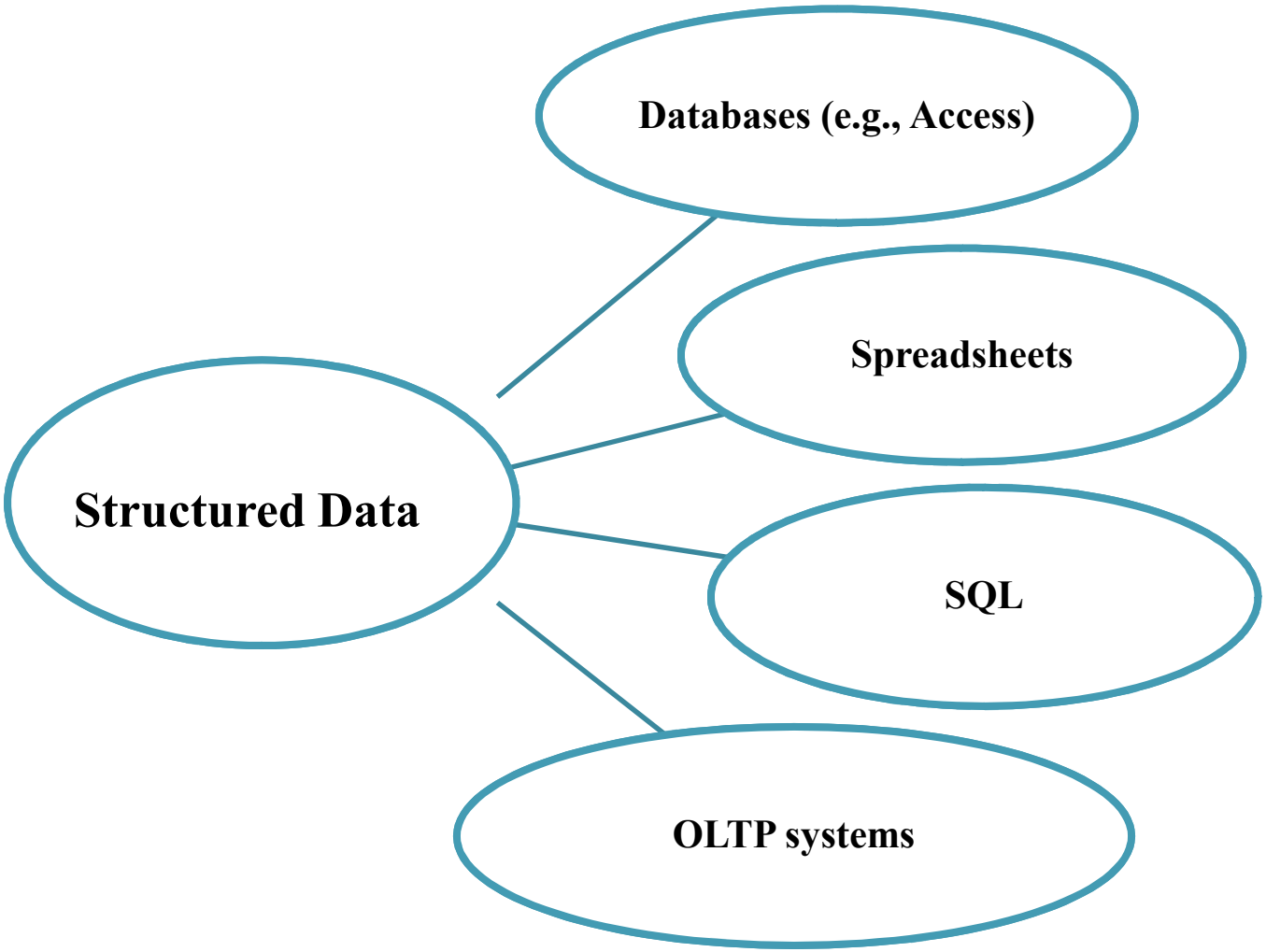
Structured Data

- Structured data is organized in semantic chunks (entities)
- Similar entities are grouped together (relations or classes)
- Entities in the same group have the same descriptions (attributes)
- Descriptions for all entities in a group (schema)
  - have the same defined format
  - have a predefined length
  - are all present
  - and follow the same order

What Is Structured Data?



Where does Structured Data Come from?



Structured Data: Everything in its Place

- **Fully described datasets**
- **Clearly defined categories and sub-categories**
- **Data neatly placed in rows and columns**
- **Data that goes into the records is regulated by a well-defined structure**
- **Indexing can be easily done either by the DBMS itself or manually**

Structured Data

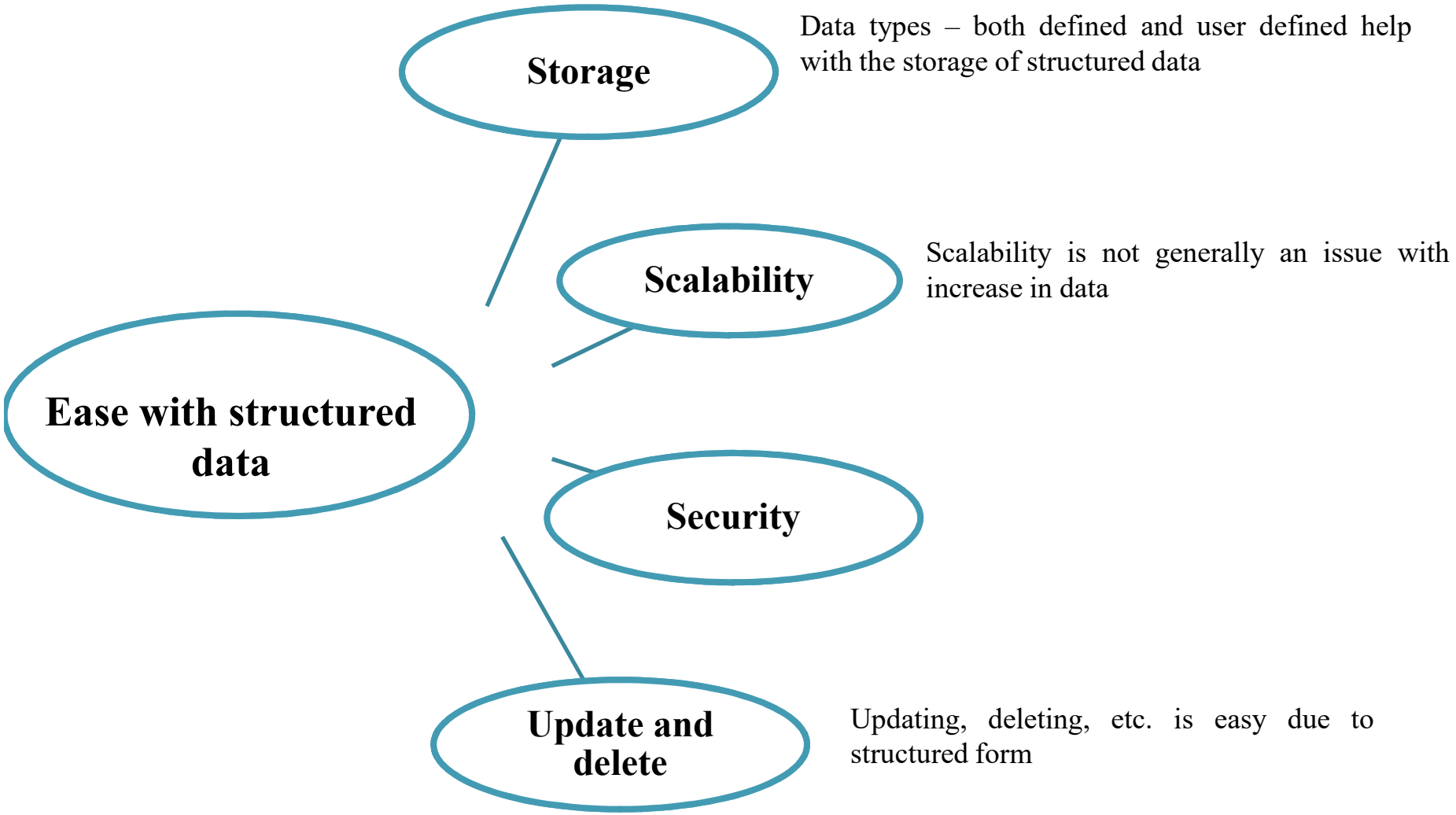
Semi-structured

Name	E-mail
Patrick Wood	ptw@dcs.abc.ac.uk, p.wood@ymail.uk
First name: Mark Last name: Taylor	MarkT@dcs.ymail.ac.uk
Alex Bourdoo	AlexBourdoo@dcs.ymail.ac.uk

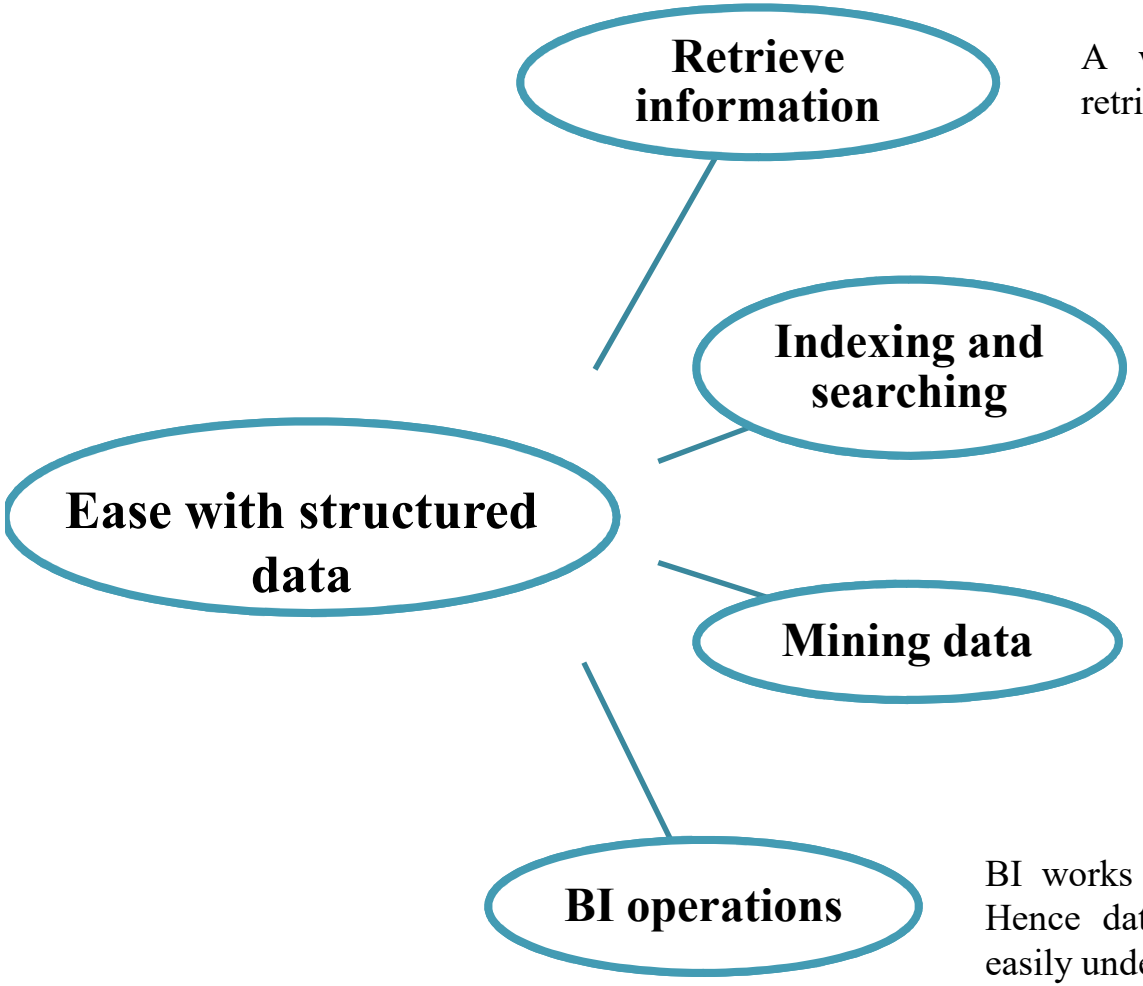
Structured

First Name	Last Name	E-mail Id	Alternate E-mail Id
Patrick	Wood	ptw@dcs.abc.ac.uk	p.wood@ymail.uk
Mark	Taylor	MarkT@dcs.ymail.ac.uk	
Alex	Bourdoo	AlexBourdoo@dcs.ymail.ac.uk	

Ease with Structured Data-Storage



Ease with Structured Data-Retrieval



A well-defined structure helps in easy retrieval of data

Data can be indexed based not only on a text string but other attributes as well. This enables streamlined search

Structured data can be easily mined and knowledge can be extracted from it

BI works extremely well with structured data. Hence data mining, warehousing, etc. can be easily undertaken

Further Readings

- *http://www.govtrack.us/articles/20061209data.xpd*
- *http://www.sapdesignguild.org/editions/edition2/sui_content.asp*

Do it Exercise

Think and write about an instance where data was presented to you in **Unstructured, semi-structured and structured data format**

Summary please…

Ask a few participants of the learning program to summarize the lecture.