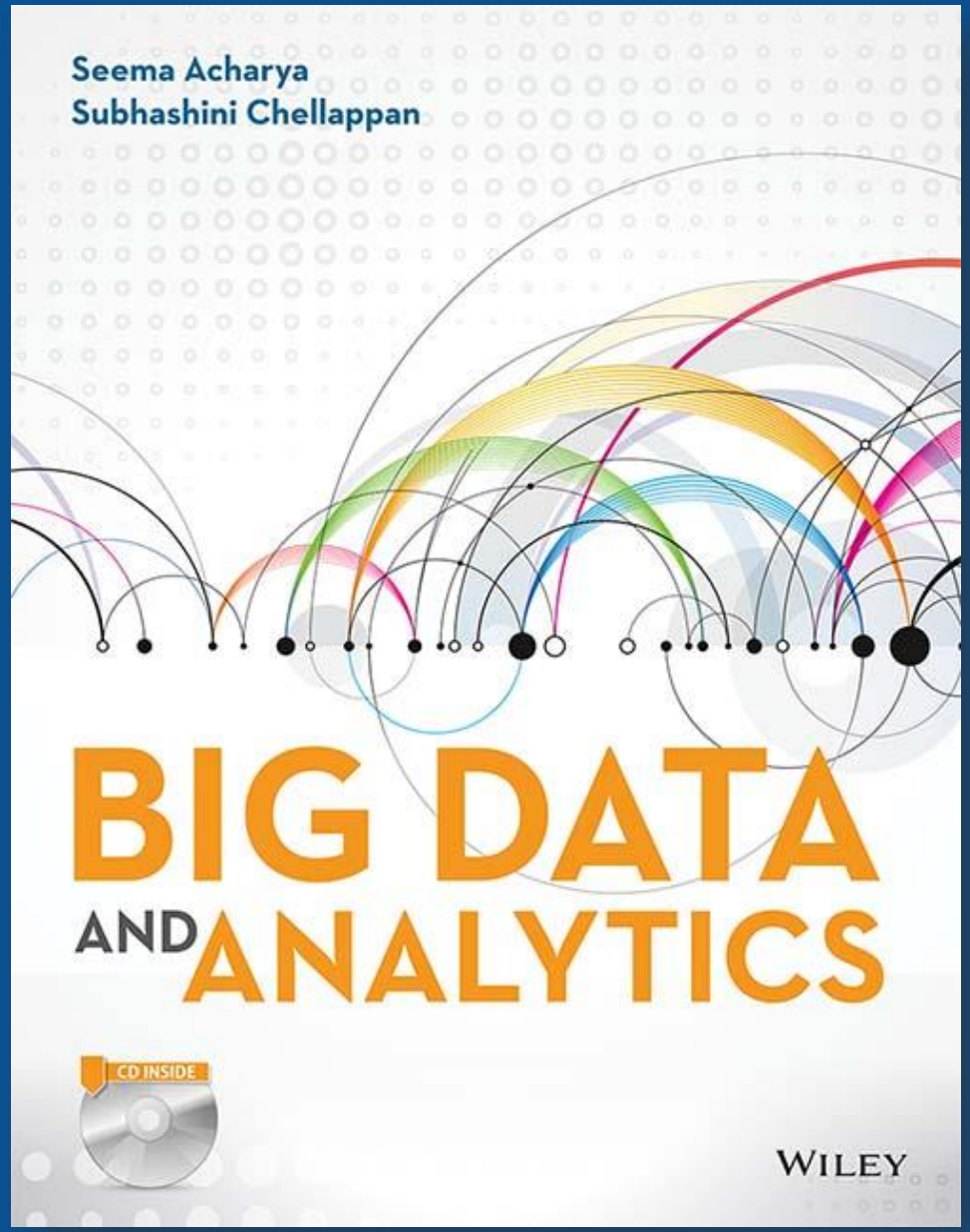


Big Data And Analytics

Seema Acharya
Subhashini Chellappan

Chapter 1

Types of Digital Data

Learning Objectives and Learning Outcomes

Learning Objectives	Learning Outcomes
Introduction to digital data and its types 1. Structured data: Sources of structured data, ease with structured data, etc. 2. Semi-Structured data: Sources of semi-structured data, characteristics of semi-structured data. 3. Unstructured data: Sources of unstructured data, issues with terminology, dealing with unstructured data.	a) To differentiate between structured, semi-structured and unstructured data. b) To understand the need to integrate structured, semi-structured and unstructured data.

Session Plan

Lecture time	45 to 60 minutes
Q/A	15 minutes

Agenda

Types of Digital Data

- **Structured**
  - Sources of structured data
  - Ease with structured data
- **Semi-Structured**
  - Sources of semi-structured data
- **Unstructured**
  - Sources of unstructured data
  - Issues with terminology
  - Dealing with unstructured data

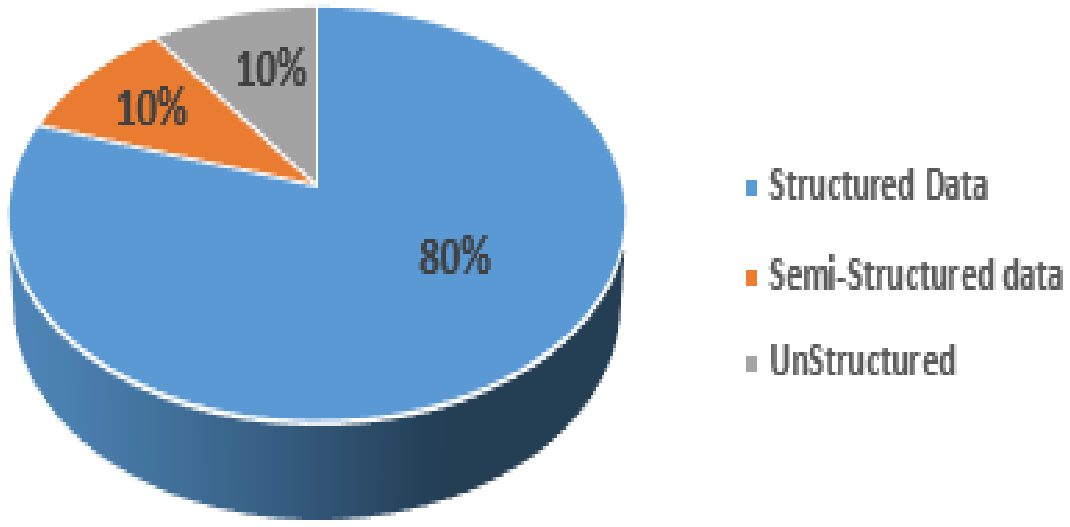
Classification of Digital Data

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

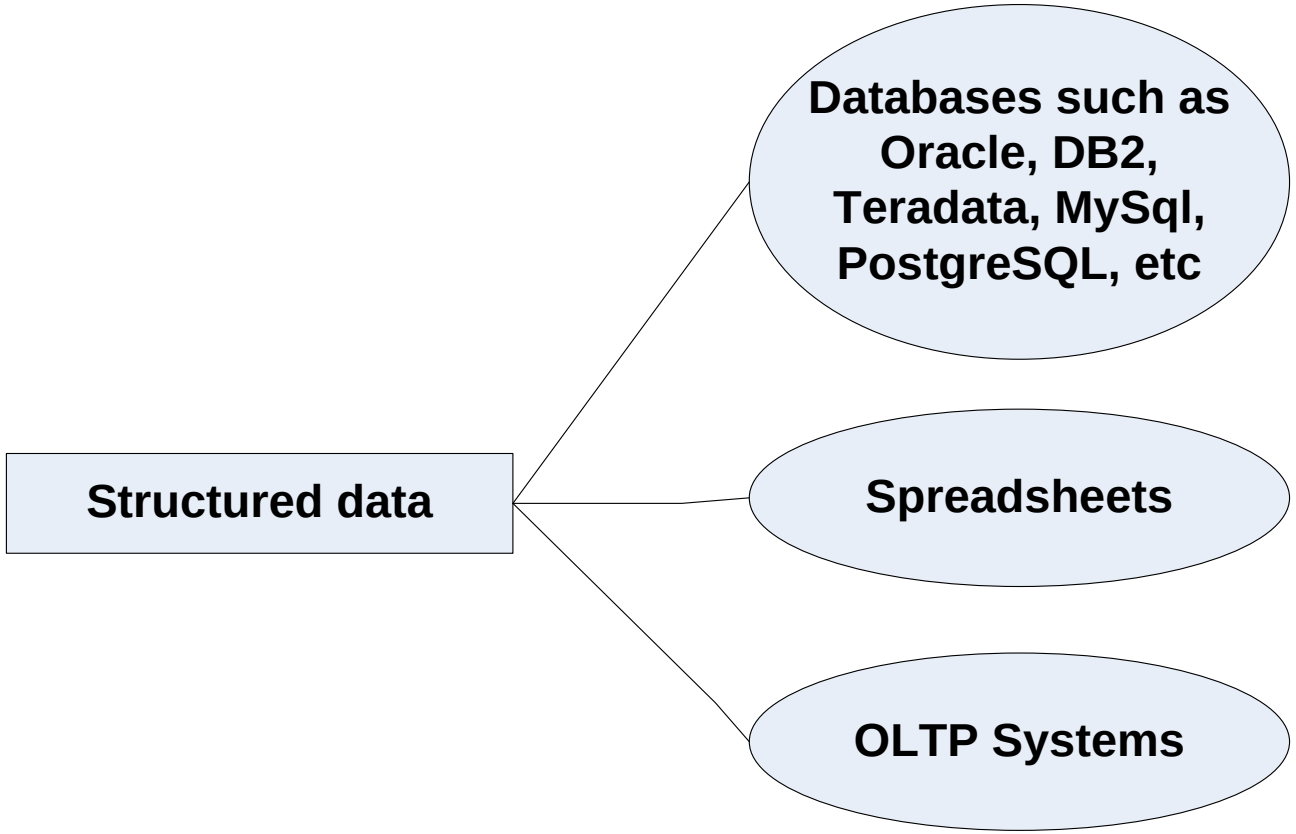


Structured Data

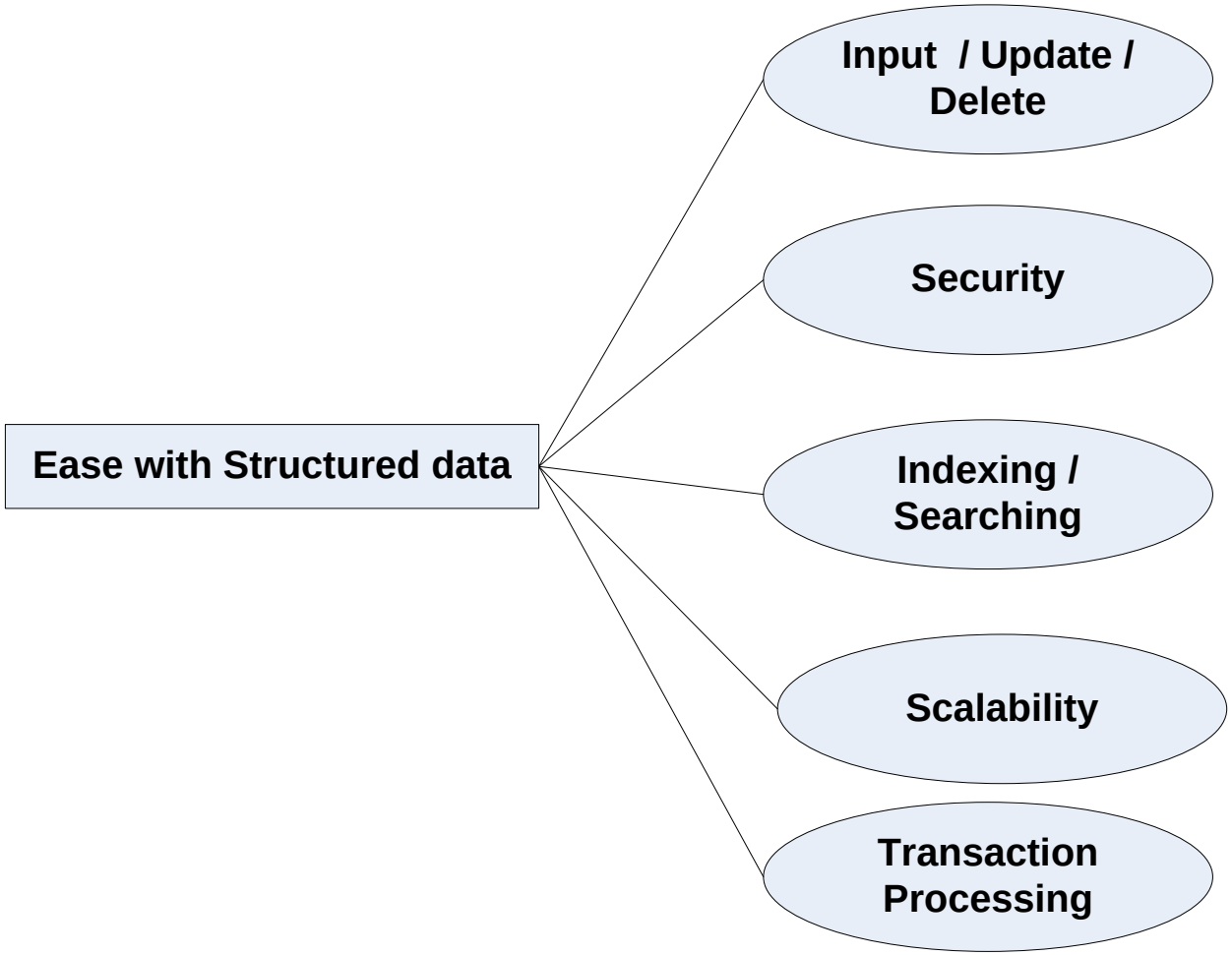
Structured Data

- 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.

Sources of Structured Data



Ease with Structured Data



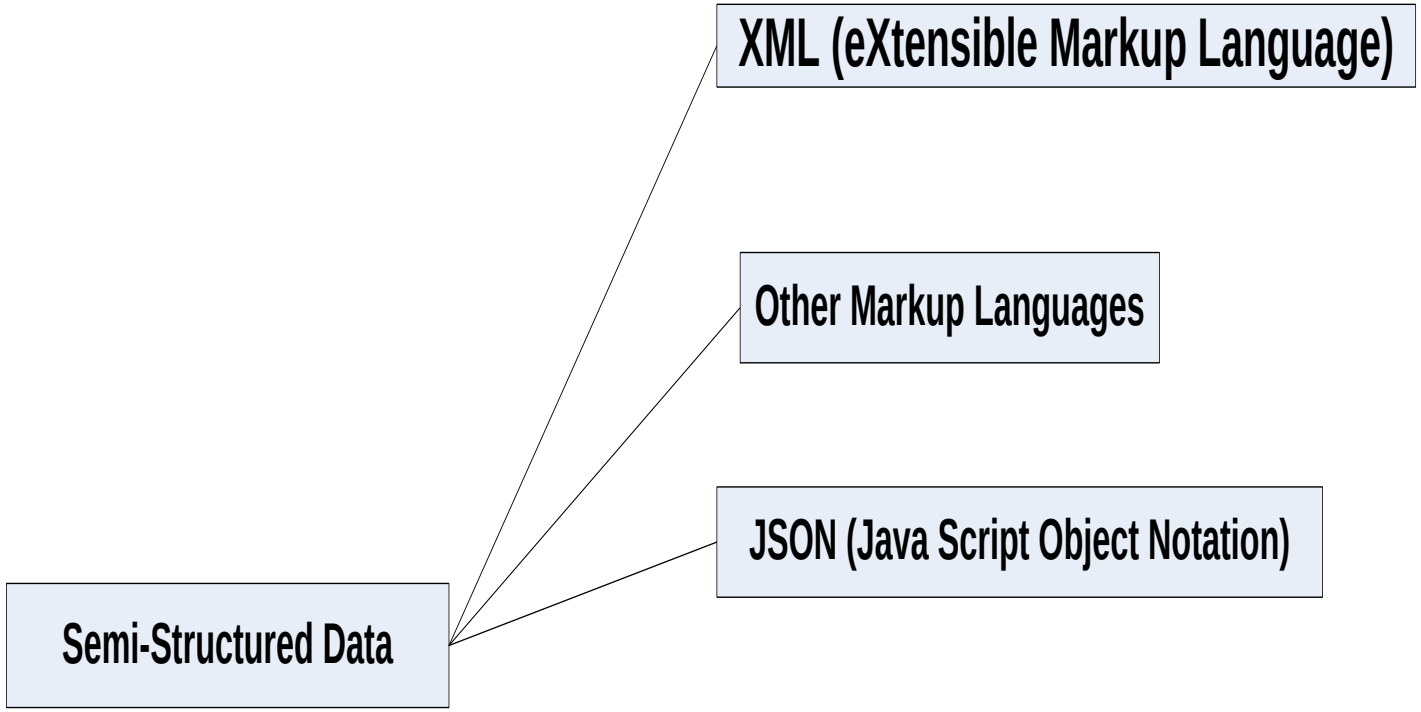
Semi-structured Data

Semi-structured Data

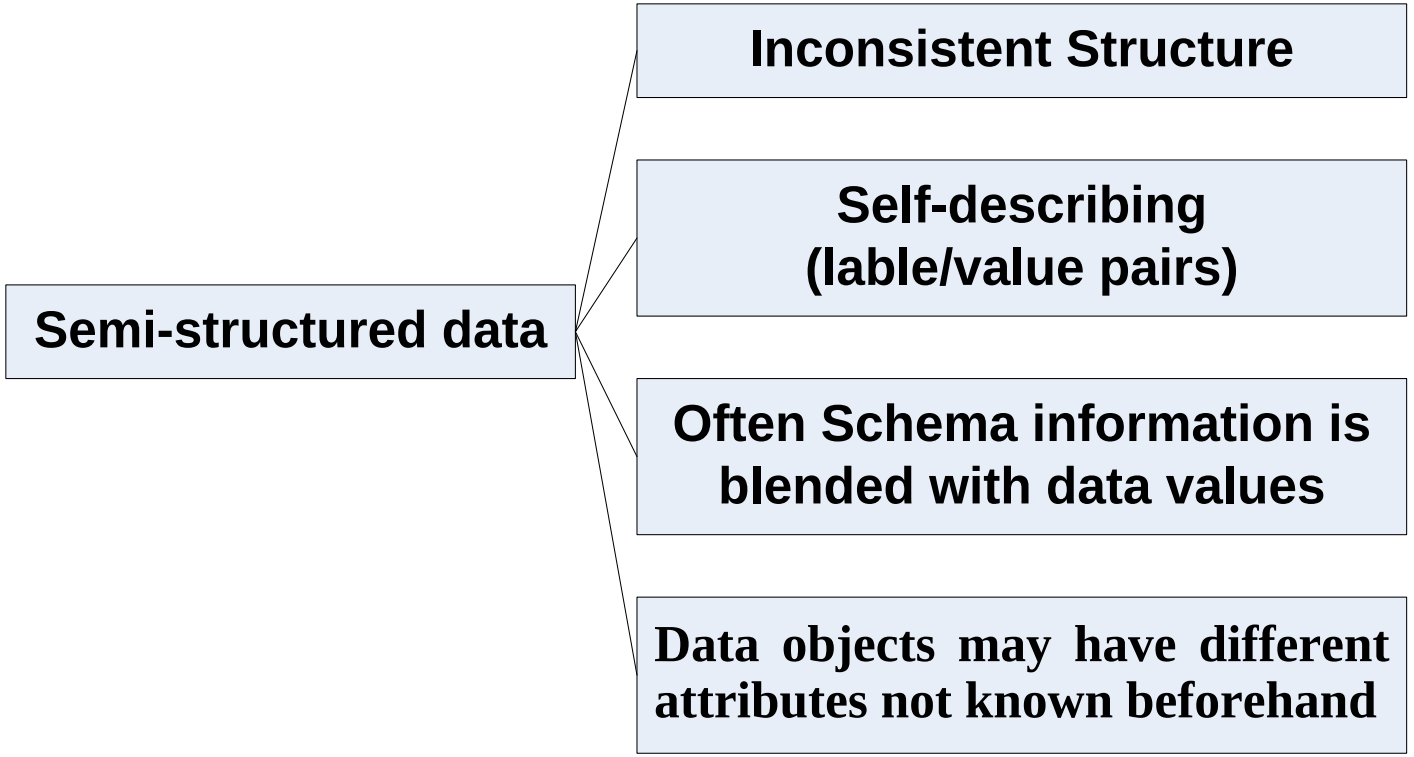
- This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program.

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

Sources of Semi-structured Data



Characteristics of Semi-structured Data

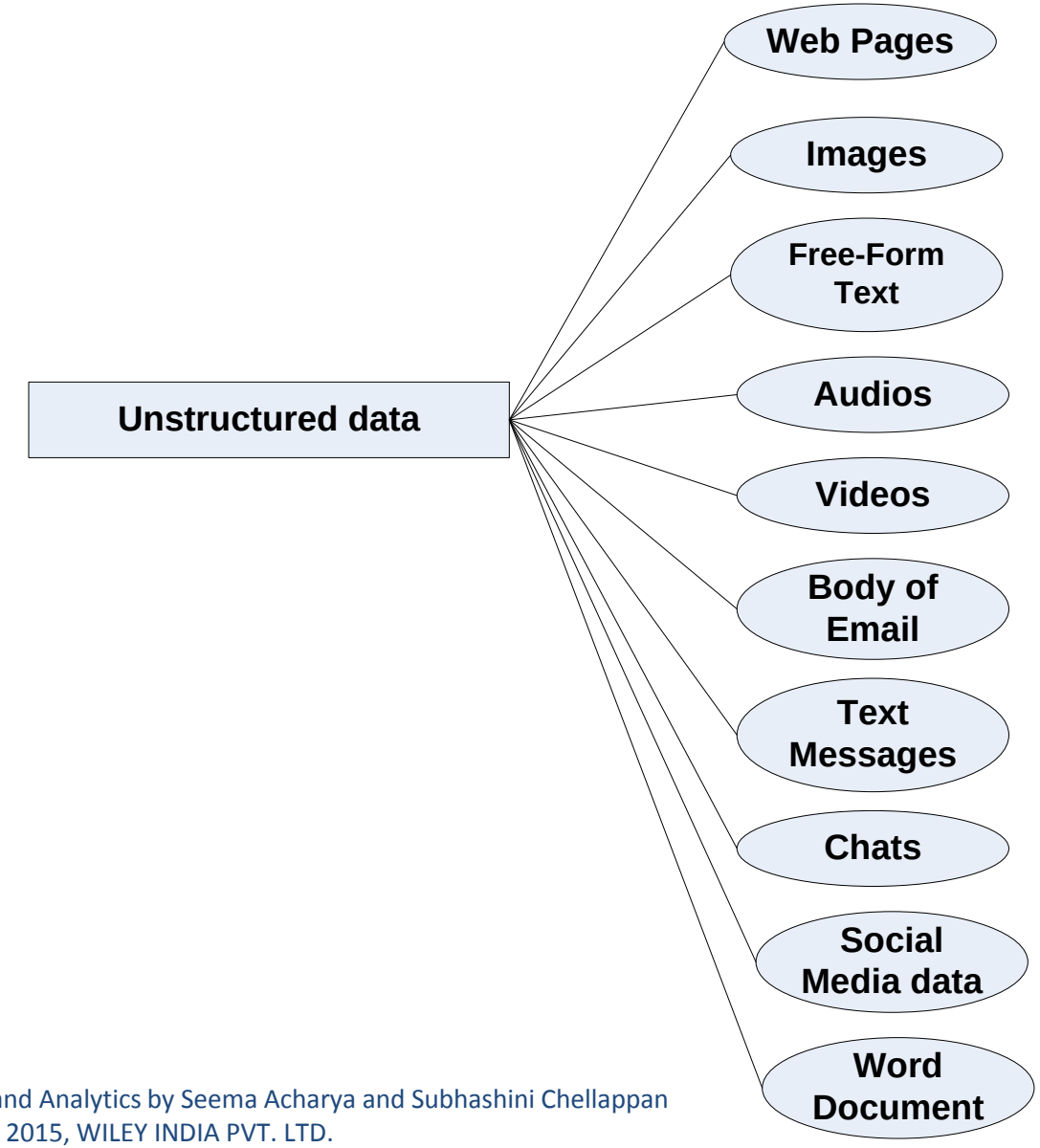


Unstructured Data

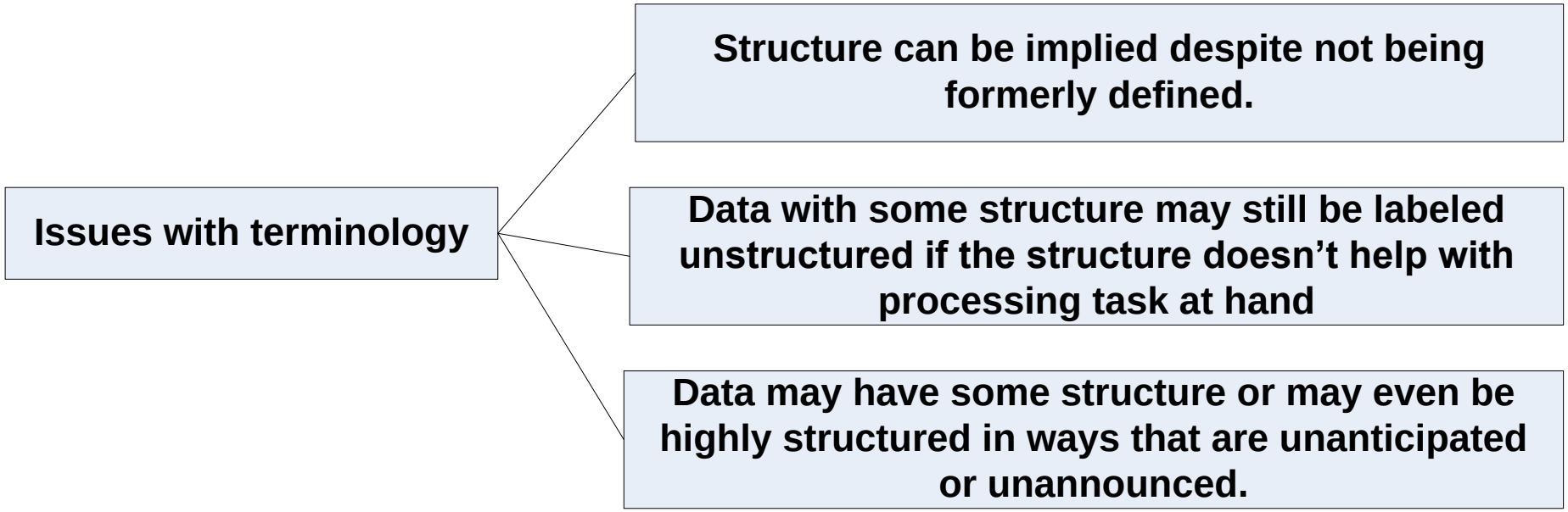
Unstructured Data

- 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.

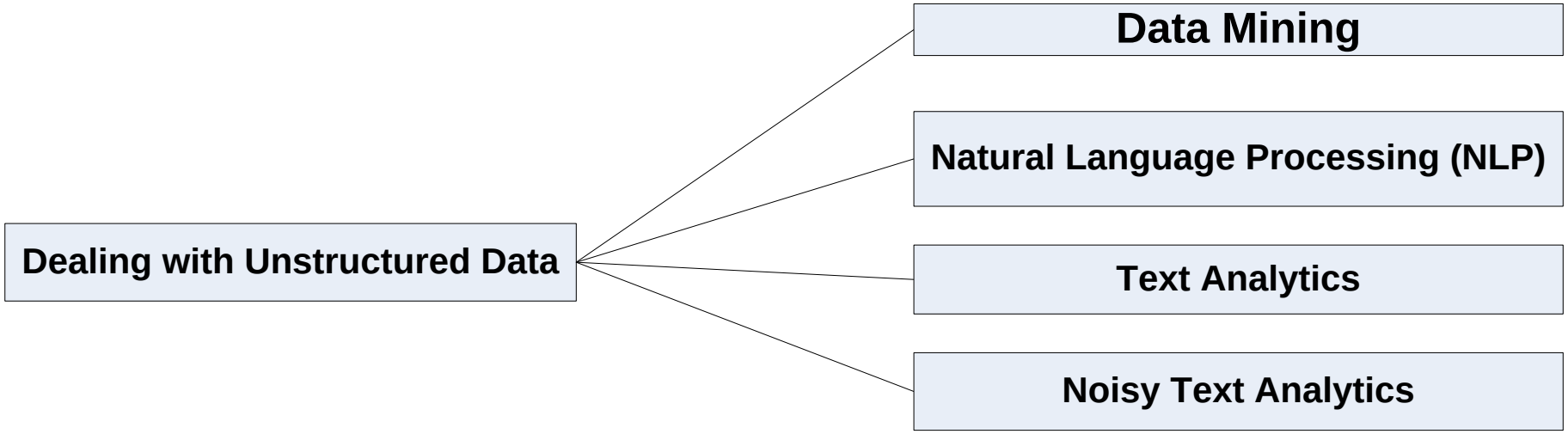
Sources of Unstructured Data



Issues with terminology – Unstructured Data



Dealing with Unstructured Data



Answer a few quick questions …

Match the following

Column A	Column B
NLP	Content analytics
Text analytics	Text messages
UIMA	Chats
Noisy unstructured data	Text mining
Data mining	Comprehend human or natural language input
Noisy unstructured data	Uses methods at the intersection of statistics, Artificial Intelligence, machine learning & DBs
IBM	UIMA

Answer Me

- Which category (structured, semi-structured, or unstructured) will you place a Web Page in?
- Which category (structured, semi-structured, or unstructured) will you place Word Document in?
- State a few examples of human generated and machine-generated data.

Summary please…

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

References ...

Further Readings

- http://data-magnum.com/the-big-deal-about-big-data-whats-inside-structured-unstructured-and-semi-structured-data/
- http://www.webopedia.com/TERM/S/structured_data.html
- http://en.wikipedia.org/wiki/UIMA

Thank you