

# Big Data: Structured and Unstructured Data

Thuan L Nguyen, PhD

## Slide 2: Big Data: Structured and Unstructured Data



*Structured and Unstructured Data (Source: Christine Taylor)*

## Slide 3: Big Data: Structured and Unstructured Data



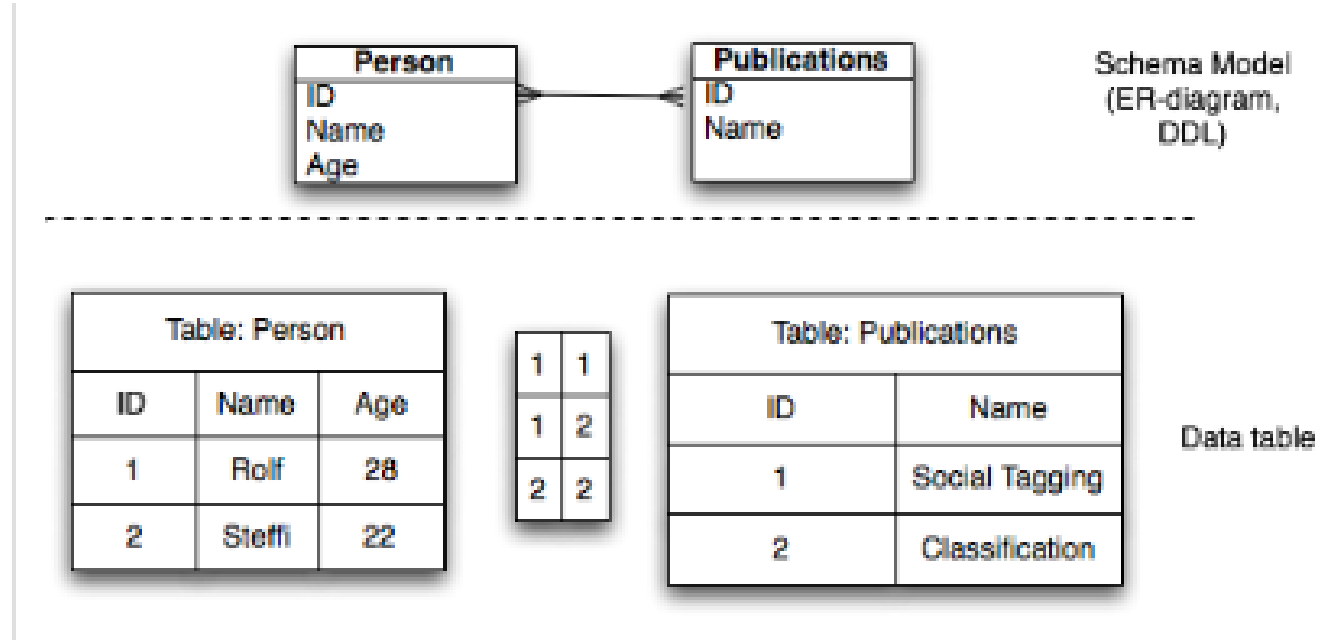
Structured Data

?

Unstructured Data

?

## Slide 4: Big Data: Structured and Unstructured Data



**Fig. 1:** Sample Table in a Relational Database System

Source: Sint et al.

## Slide 5: Big Data: Structured and Unstructured Data

### Structured Data:

- Is highly organized information found in a relational database
  - Is formatted in fixed fields, following a predefined schema
  - Is easily detectable via search operations or algorithms.
  - Is relatively simple to enter, store, query, and analyze
  - Is strictly defined in terms of field name and type (e.g. alpha, numeric, date, currency)
    - Is often restricted by character numbers or specific terminology.
- Language (SQL) is used to perform queries on structured data within relational databases.
- Structured data leaves out immense amounts of material that do not fit simply into a firm's organization of information.

## Slide 6: Big Data: Structured and Unstructured Data

### **Unstructured Data:**

- Is essentially everything else – that is not structured data
- Unstructured data:
  - Textual or non-textual
  - Human or machine-generated.

## Slide 7: Big Data: Structured and Unstructured Data



*Sources of Big Data (Source: ADEC Group)*

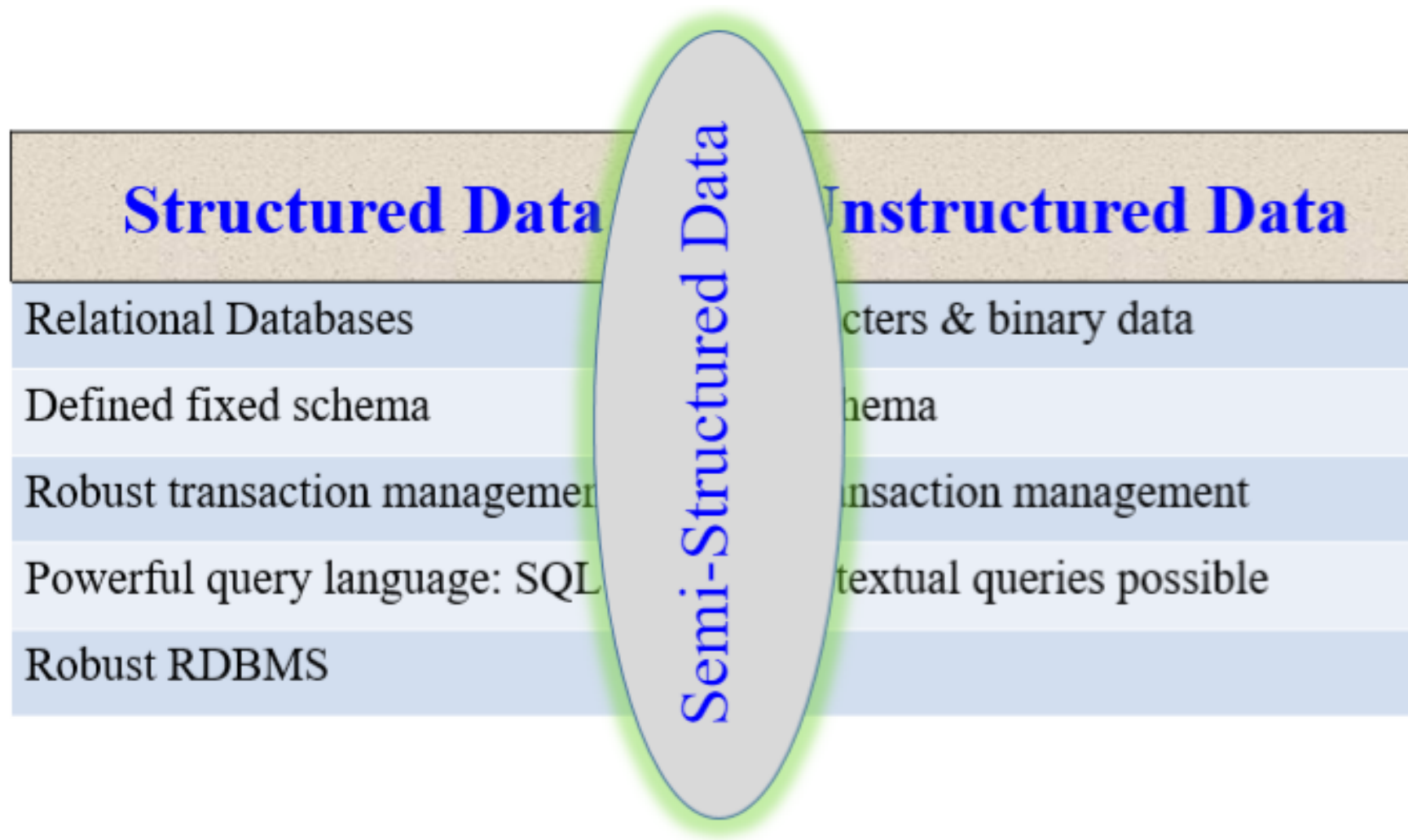
# Slide 8: Big Data: Structured and Unstructured Data

## Unstructured Data:

- Can be human-generated:
  - Text files: Word processing, spreadsheets, presentations, email, logs.
  - Email contents
  - Social Media: Data from Facebook, Twitter, LinkedIn.
  - Website: YouTube, Instagram, photo sharing sites.
  - Mobile data: Text messages, locations.
  - Communications: Chat, IM, phone recordings, collaboration software.
  - Media: MP3, digital photos, audio and video files.
  - Business applications: MS Office documents, productivity applications.
- Can be machine-generated:
  - Satellite imagery: Weather data, land forms, military movements.
  - Scientific data: Oil and gas exploration, space exploration, seismic imagery, atmospheric data.
  - Digital surveillance: Surveillance photos and video.
  - Sensor data: Traffic, weather, oceanographic sensors.



## Slide 9: Big Data: Structured and Unstructured Data



## Slide 10: Big Data: Structured and Unstructured Data

Structured Data

### Semi-Structured Data

- Not so rigidly defined as structured data
- Not unmanageable like unstructured data
- Maintain internal tags and markings
- Tags & marks can be used to identify data elements and enable information grouping and hierarchies

Unstructured Data

# Slide 11: Big Data: Structured and Unstructured Data

Structured Data

## Semi-Structured Data

Examples: Emails (as a whole)

Examples: CSV files

CSV (Comma-Separated Values) files may contain multiple lines. Each line is a CSV string in which each piece of information is separated from others by a comma.

*“1234567890, A Good Book, John Smith, May 1<sup>st</sup> 2018”*

Unstructured Data

## Slide 12: Big Data: Structured and Unstructured Data

Structured Data

### Semi-Structured Data

- Data in XML (Markup Languages) format
- Data in JSON format  
(JSON: Javascript Object Notion - Name: Value)
- Data in NoSQL Databases

Unstructured Data



## Slide 13: Big Data: Structured and Unstructured Data

Structured Data

### Semi-Structured Data

- Data in XML (Markup Languages) format
  - ) A semi-structured document language
  - ) Is a set of document encoding rules that human-beings and machine can read
  - ) Very flexible, can be used to transfer data between entities in the Internet

Unstructured Data

# Slide 14: Big Data: Structured and Unstructured Data

Structured Data

## Semi-Structured Data

An example of a simple XML document:

```
<note>  
<to>Tom</to>  
<from>John</from>  
<heading>Reminder</heading>  
<body>We will have a lunch out today</body>  
</note>
```

Unstructured Data

## Slide 15: Big Data: Structured and Unstructured Data

Structured Data

### Semi-Structured Data

- Data in XML (Markup Languages) format
- Data in JSON format  
(JSON: Javascript Object Notion - Name: Value)

Unstructured Data

# Slide 16: Big Data: Structured and Unstructured Data

Structured Data

## Semi-Structured Data

An example of a simple piece of data in JSON:

```
{ "name": "John", "age": 30, "car": "GM" }
```

- JSON objects are surrounded by curly braces {}.
- JSON objects are written in key/value pairs.
- Keys must be strings, and values must be a valid JSON data type (string, number, object, array, boolean or null).
- Keys and values are separated by a colon.
- Each key/value pair is separated by a comma.
- Very flexible
- The structure is interchangeable among languages, JSON excels at transmitting data between web applications and servers

Unstructured Data



## Slide 17: Big Data: Structured and Unstructured Data

Structured Data

### Semi-Structured Data

- Data in XML (Markup Languages) format
- Data in JSON format  
(JSON: Javascript Object Notion - Name: Value)
- Data in NoSQL Databases

Unstructured Data

## Slide 18: Big Data: Structured and Unstructured Data

Structured Data

### Semi-Structured Data

- Data in XML (Markup Languages) format
  - Data in JSON format  
(JSON: Javascript Object Notion - Name: Value)
  - Data in NoSQL Databases
- > Does not follow the rules and principles of relational databases and SQL
- > Is flexible enough to store data/information that does not fit into the records/tables formats

Unstructured Data