SESSION 2016-2017

B.TECH (CSE) YEAR: IV SEMESTER: VIII

BIG DATA AND ANALYTICS

(CSE448)

MODULE 1 (L1)

Presented By

Dilip Kumar Sharma , Rahul Pradhan, Vivek Kumar, Yogesh Gupta

Dept of Computer Engineering & Applications

GLA University India

What Happens in an **Internet Minute?**

And **Future Growth** is **Staggering**

Today, the number of networked devices = the global population

By **2015**, the number of networked devices = **2x** the global population

In **2015**, it would take you **5 years** to view all video crossing IP networks each **second**

intel

Do you know what happens in one minute on the Internet?

- In just one minute, more than 204 million emails are sent.
- Amazon rings up about $83,000 in sales.
- Around 20 million photos are viewed and
- 3,000 uploaded on Flickr.
- At least 6 million Facebook pages are viewed around the world.
- And more than 61,000 hours of music are played on Pandora while more than
- 1.3 million video clips are watched on YouTube.

Classification of Digital Data

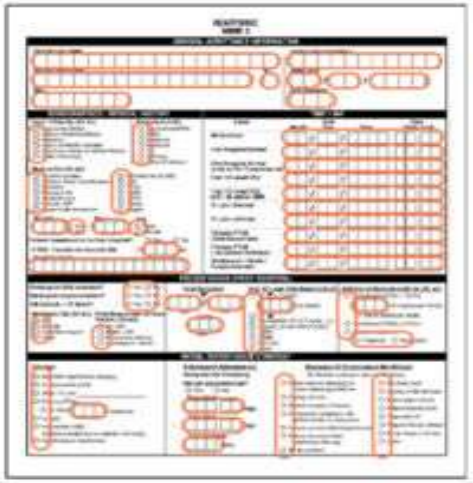
Digital data is classified into the following categories:

- Structured data
- Semi-structured data
- Unstructured data

STRUCTURED VS UNSTRUCTURED DOCUMENTS

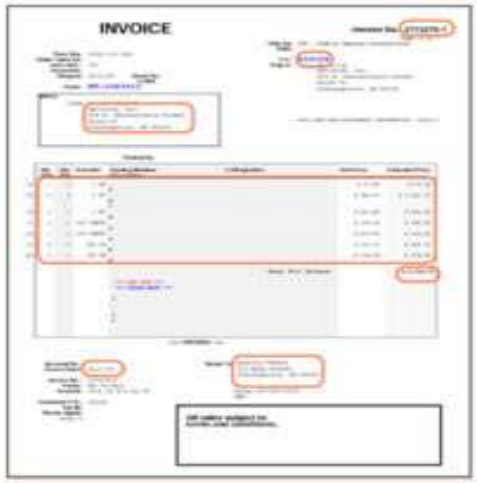
Software captures the image of a paper document allowing the information to be translated to electronic data without manual input. Recognition technologies have accelerating capabilities from optical character recognition (OCR) to intelligent character recognition (ICR). The technology differs for each type of document. Which style of documents do you need to automate?

Structured Document



- Surveys
- Questionnaires
- Tests
- Claim forms

Semi-structured Document



- Invoices
- Purchase orders
- Bills of lading
- Explanation of benefits

Unstructured Document

- Contracts
- Letters
- Articles
- Memos

Classification of Digital Data

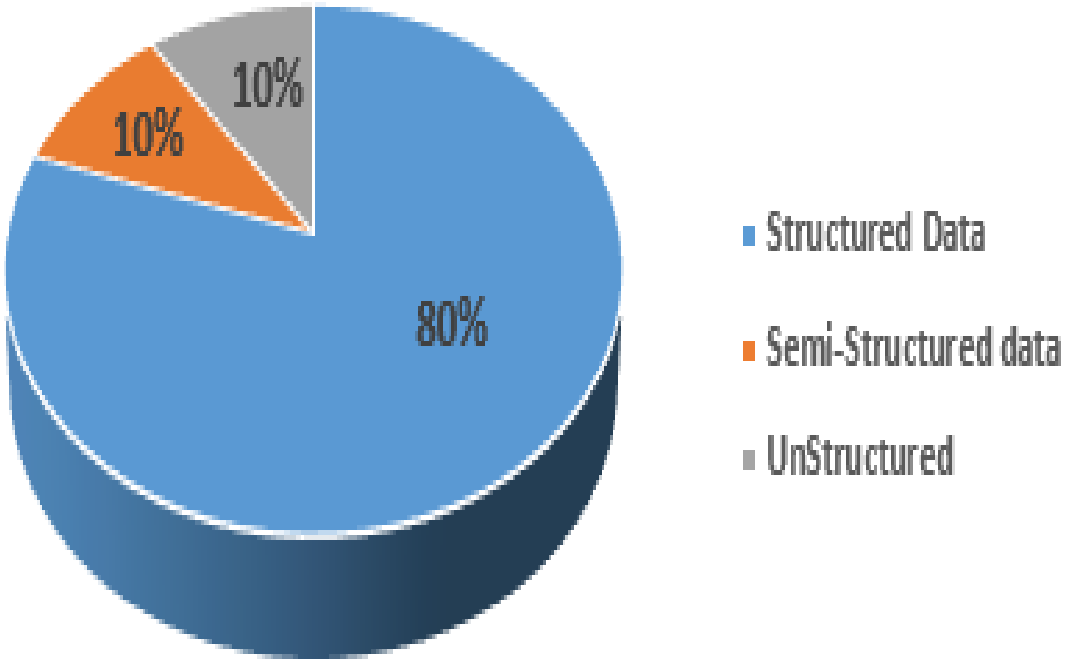
- **Unstructured data:**
  - This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program.
  - About 80-90% data of an organization is in this for example, memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white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

Classification of Digital 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;
- for example, en 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 their objects. Data stored in databases is an example of structured data.

Approximate Percentage Distribution of Digital Data

- Approximate percentage distribution of digital data



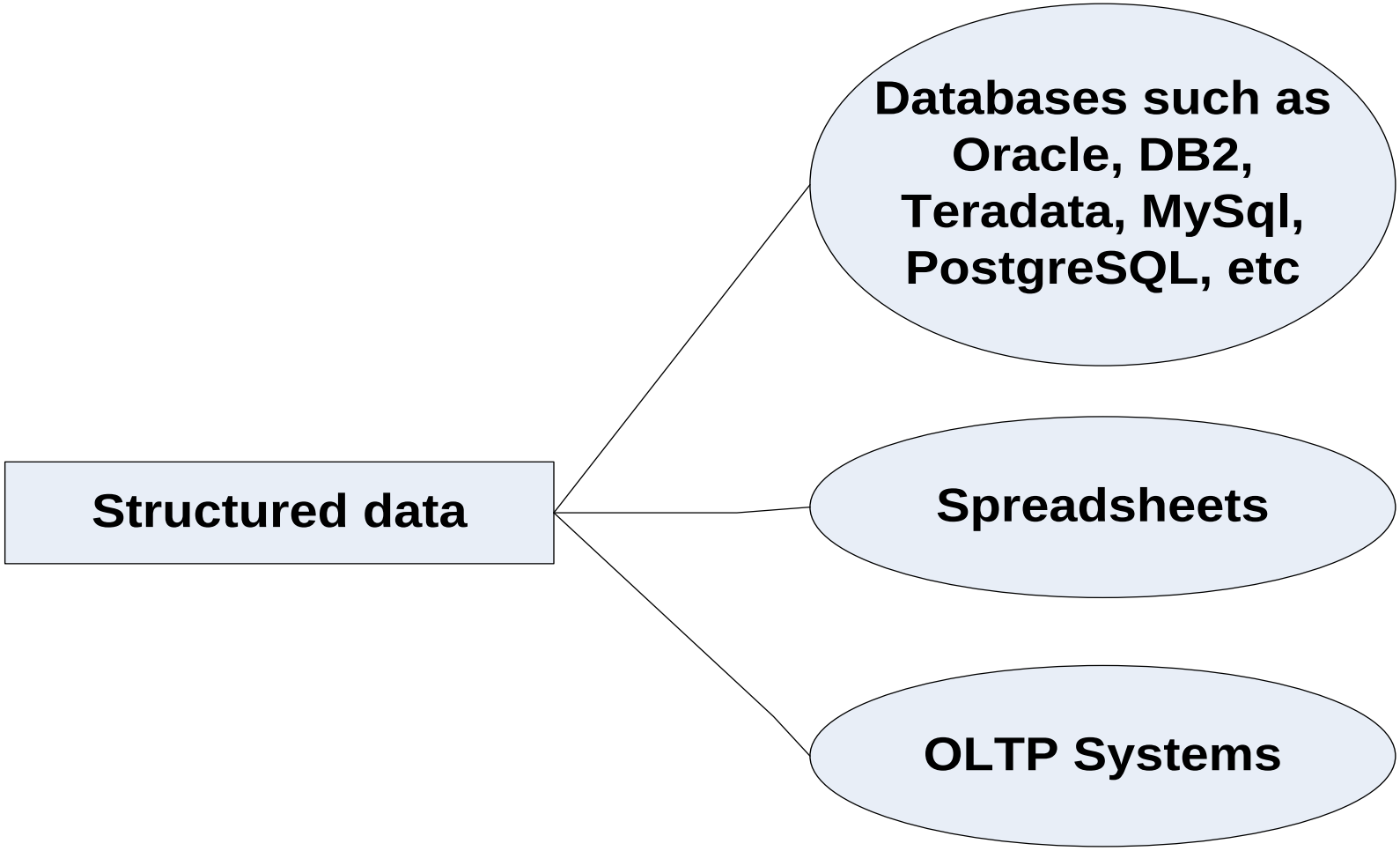
Structured Data

- 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.
- 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

- If your data is highly structured, one can look at leveraging any of the available RDBMS
- [Oracle Corp. — Oracle, IBM — DB2, Microsoft — Microsoft SQL Server, EMC — Greenplum, Teradata — Teradata, MySQL (open source), PostgreSQL (advanced open source) etc.] to house it.
- These databases are typically used to hold transaction/operational data generated and collected by day-to-day business activities. In other words, the data of the On-Line Transaction Processing (OLTP) systems are generally quite structured.

Sources of Structured Data



Ease of Working with Structured Data

The ease is with respect to the following:

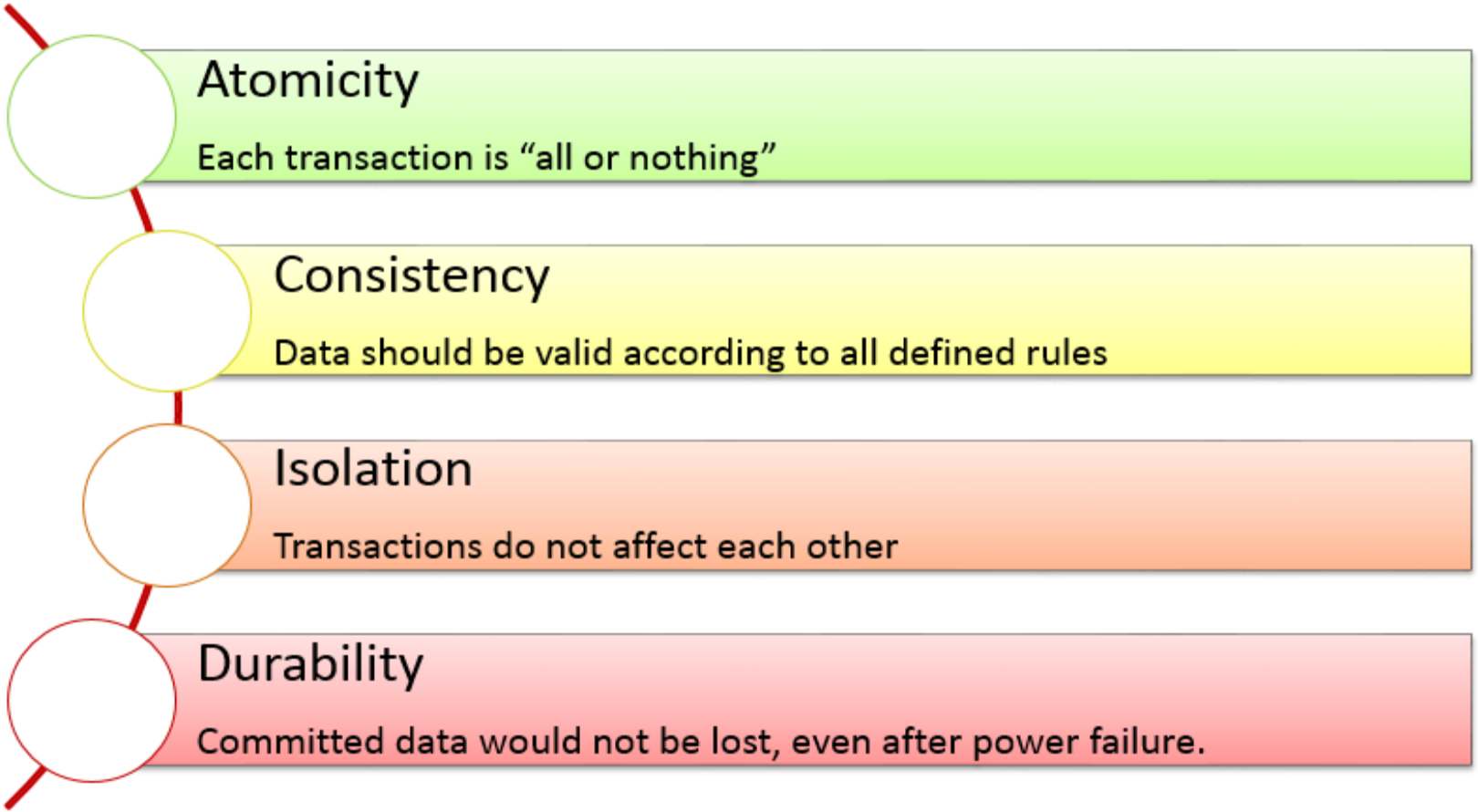
- **Insert/update/delete:** The Data Manipulation Language (DML) operations provide the required ease with data input, storage, access, process, analysis, etc.
- **Security:** How does one ensure the security of information? There are available check encryption and tokenization solutions to warrant the security of information throughout its lifecycle.
- Organizations are able to retain control and maintain compliance adherence by ensuring that only authorized individuals are able to decrypt and view sensitive information.

Ease of Working with Structured Data

- **Indexing:** An index is a data structure that speeds up the data retrieval operations (primarily the SELECT DML statement) at the cost of additional writes and storage space, but the benefits that ensue in search operation are worth the additional writes and storage space.
- **Scalability:** The storage and processing capabilities of the traditional RDBMS can be easily scaled up by increasing the horsepower of the database server *(increasing the primary and secondary or peripheral storage capacity, processing capacity of the processor, etc.)*.

Ease of Working with Structured Data

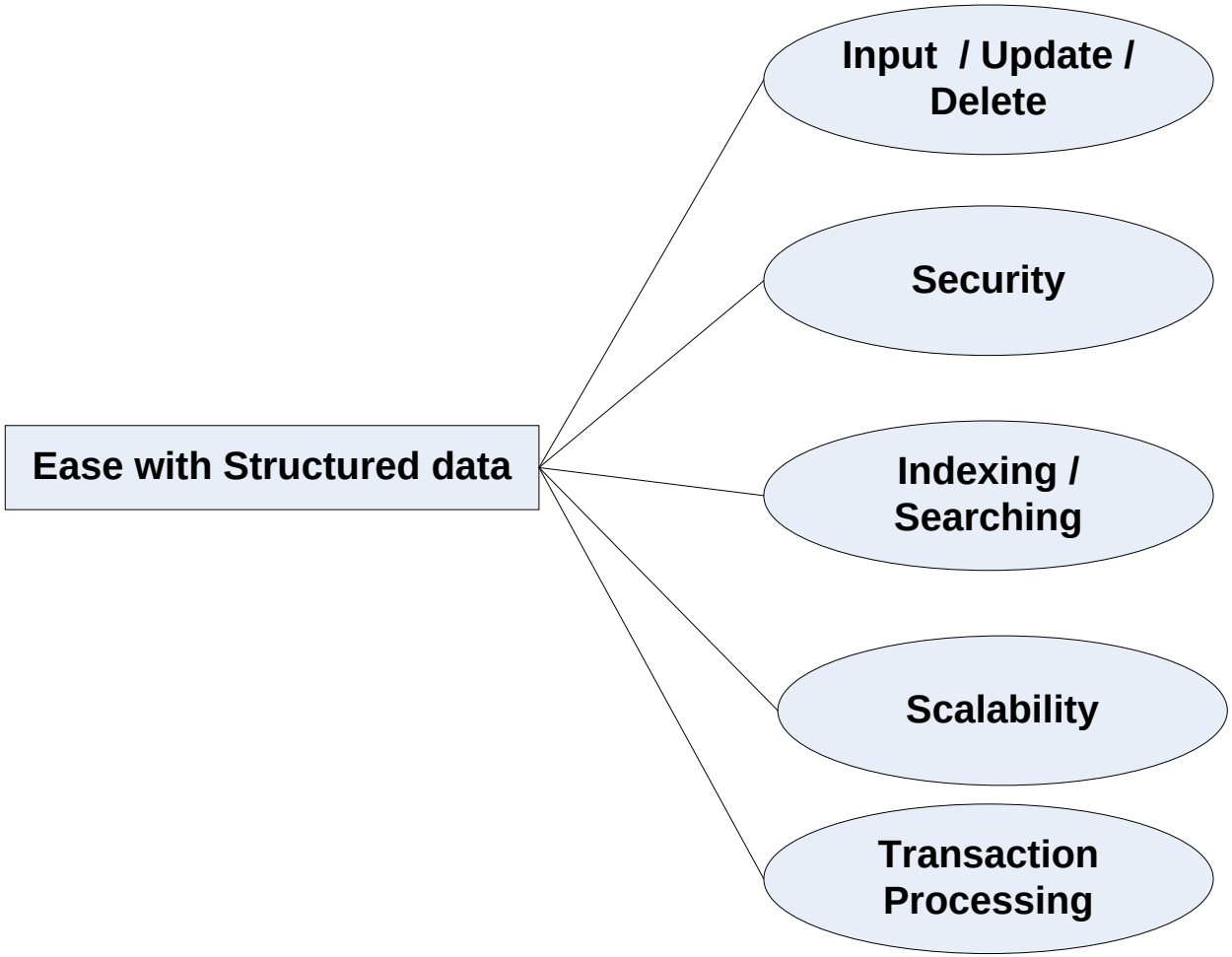
- **Transaction processing:** RDBMS has support for Atomicity, Consistency, Isolation, and Durability (ACID) properties of transaction.



Ease of Working with Structured Data

- **Transaction processing:** RDBMS has support for Atomicity, Consistency, Isolation, and Durability (ACID) properties of transaction.
  - ***Atomicity:*** A transaction is atomic, means that either it happens in its entirety or none of it at all.
  - ***Consistency:*** The database moves from one consistent state to another consistent state. In other words, if the same piece of information is stored at two or more places, they are in complete agreement.
  - ***Isolation:*** The resource allocation to the transaction happens such that the transaction gets the impression that it is the only transaction happening in isolation.
  - ***Durability:*** All changes made to the database during a transaction are permanent and that accounts for the durability of the transaction.

Ease with Structured Data



Semi-structured Data

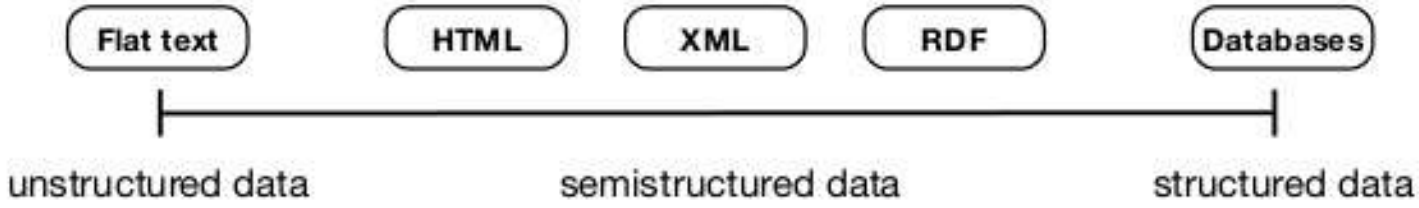
- This is the data which does not conform to a data model but has some structure.
- However, it is not in a form which can be used easily by a computer program.
- Example, emails, XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.

Semi-structured Data

It has the following features:

- It does not conform to the data models that one typically associates with relational databases or any other form of data tables.
- It uses tags to segregate semantic elements.
- Tags are also used to enforce hierarchies of records and fields within data.
- There is no separation between the data and the schema.
- The amount of structure used is dictated by the purpose at hand.
- In semi-structured data, entities belonging to the same class and also grouped together need not necessarily have the same set of attributes.
- And if at all, they have the same set of attributes, the order of attributes may not be similar and for all practical purposes it is not important as well.

The data landscape

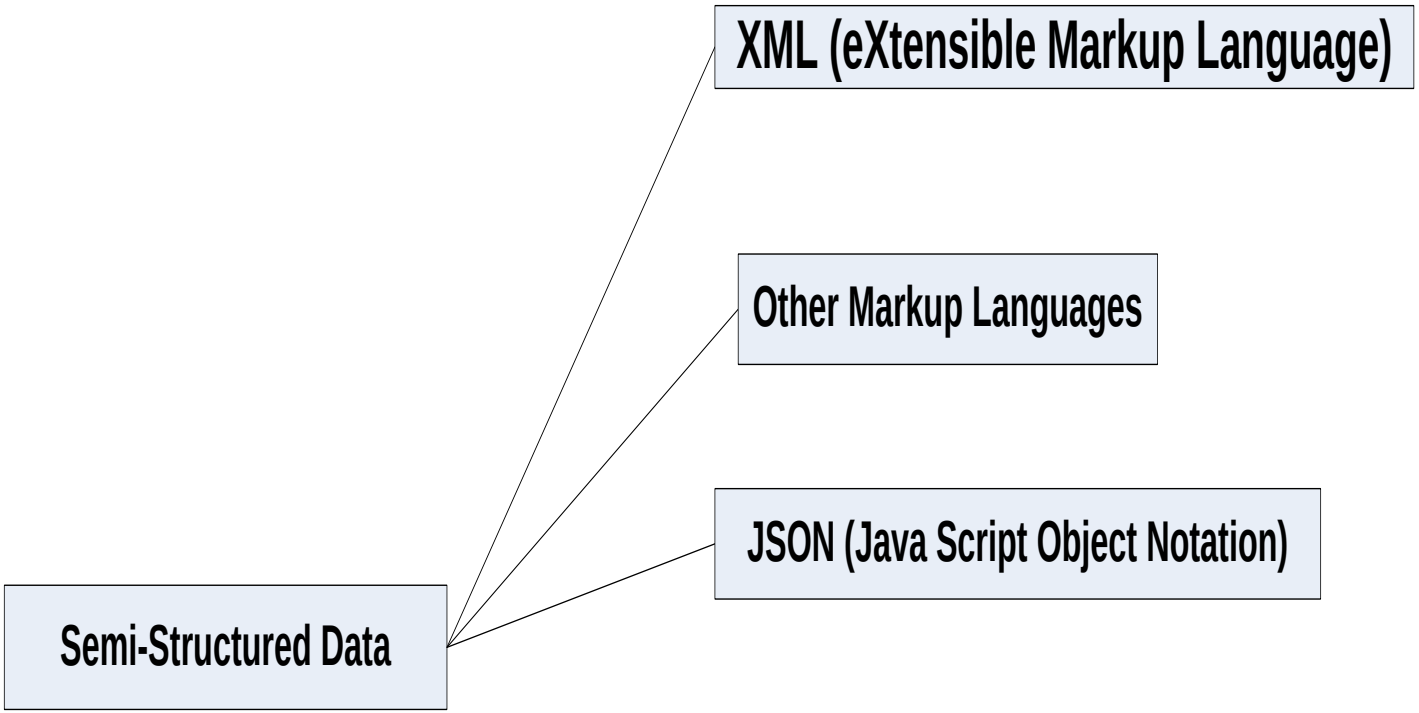


- **Semistructured data**
  - Lack of fixed, rigid schema
  - No separation between the data and the schema, self-describing structure (tags or other markers)

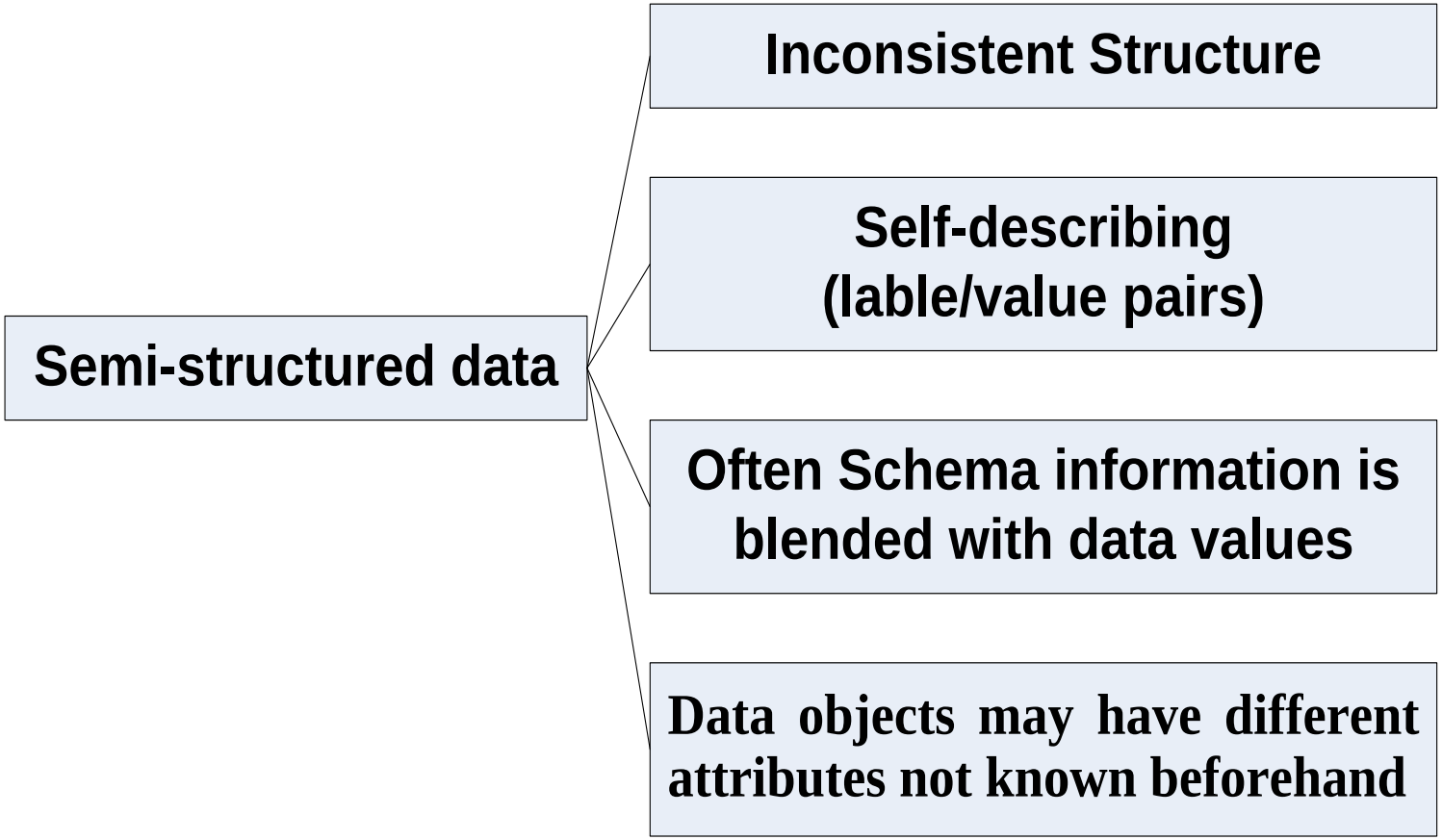
Sources of Semi-structured Data

- Amongst the sources for semi-structured data, the front runners are “XML” and “JSON”.
- **XML:** eXtensible Markup Language (XML) is hugely popularized by web services developed utilizing the Simple Object Access Protocol (SOAP) principles.

Sources of Semi-structured Data



Characteristics of Semi-structured Data



Sources of Semi-structured Data

- **JSON:** Java Script Object Notation (JSON) is used to transmit data between a server and a web application.
- JSON is popularized by web services developed utilizing the Representational State Transfer (REST) - an architecture style for creating scalable web services.
- MongoDB (open-source, distributed, NoSQL, documented-oriented database) and Couchbase (originally known as Membase, open-source, distributed, NoSQL, document-oriented database) store data natively in JSON format.

Sources of Semi-structured Data

An example of HTML is as follows:

```
<HTML>
    <HEAD>
    <TITLE>Place your title here</TITLE>
    </HEAD>
<BODY BGCOLOR="FFFFFF">
     <CENTER><IMG SRC="clouds.jpg" ALIGN="BOTTOM"x/CENTER>
     <HR>          <a href="http://bigdatauniversity.com">Link Name</a>
     <Hl>this is a Header</Hl>
     <H2>this is a sub Header</H2>
     Send me mail at <a href="mailto:support@yourcompany.com"> support@yourcompany.com</a>.
     <P>a new paragraph!
     <PxB>a new paragraph!</B>
     <BRxBxI>this is a new sentence without a paragraph break, in bold italics.</Ix/B>
     <HR>
</BODY>
</HTML>
```

Sources of Semi-structured Data

Sample JSON document

{

_id:9,

BookTitle: “Fundamentals of Business Analytics”,

AuthorName: “Seema Acharya”,

Publisher: “Wiley India”,

YearofPublication: “2011”

}

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.

INSANELY FAST CLIP

WHERE DOES IT COME FROM?

INTERNAL

Transactions 88%

Log Data 73%

Emails 57%

Brand social media properties

Customer service centers

Mobile market research data

Employee performance reviews

Consumer survey data

Candidate interviews

EXTERNAL

Social Media 43%

Audio Data 38%

Photo & Videos 34%

Merchandising photos

Crowd sourcing

Web scraping

Social media

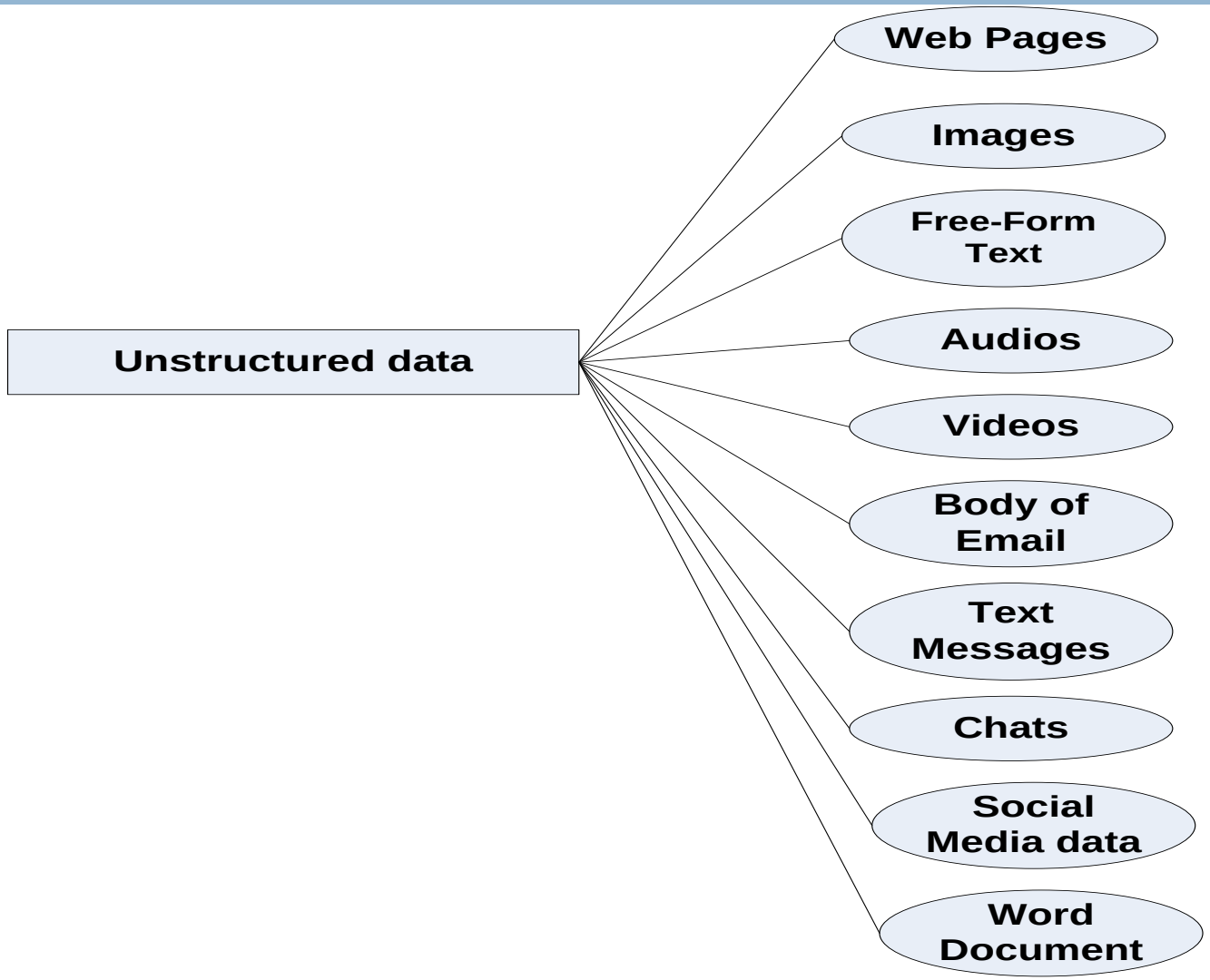
Blogs and chat rooms

Consumer product reviews

In-store video monitoring

2 Minutes

Sources of Unstructured Data



HOW VOLUMINOUS IS IT?

2 Minutes

2000

Every 2 minutes, we generate an equal amount of data to what was created from the start of time to 2000.

1600 Exabytes

2015

By the end of 2015, enterprise unstructured data will cross 1600 Exabytes.

In 2018, out of all online traffic,

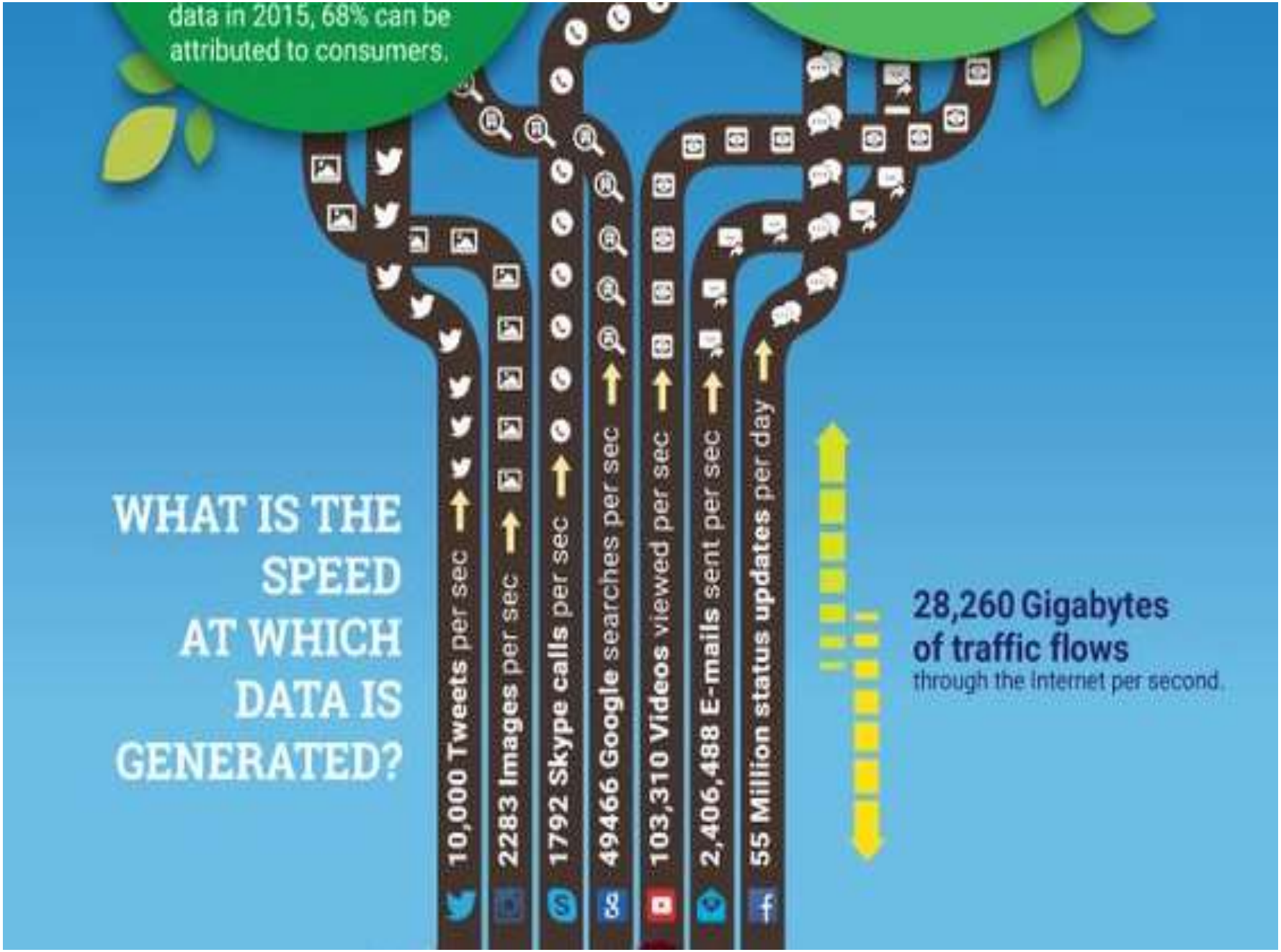
mobile video will account for **69%**.

$600 BILLION

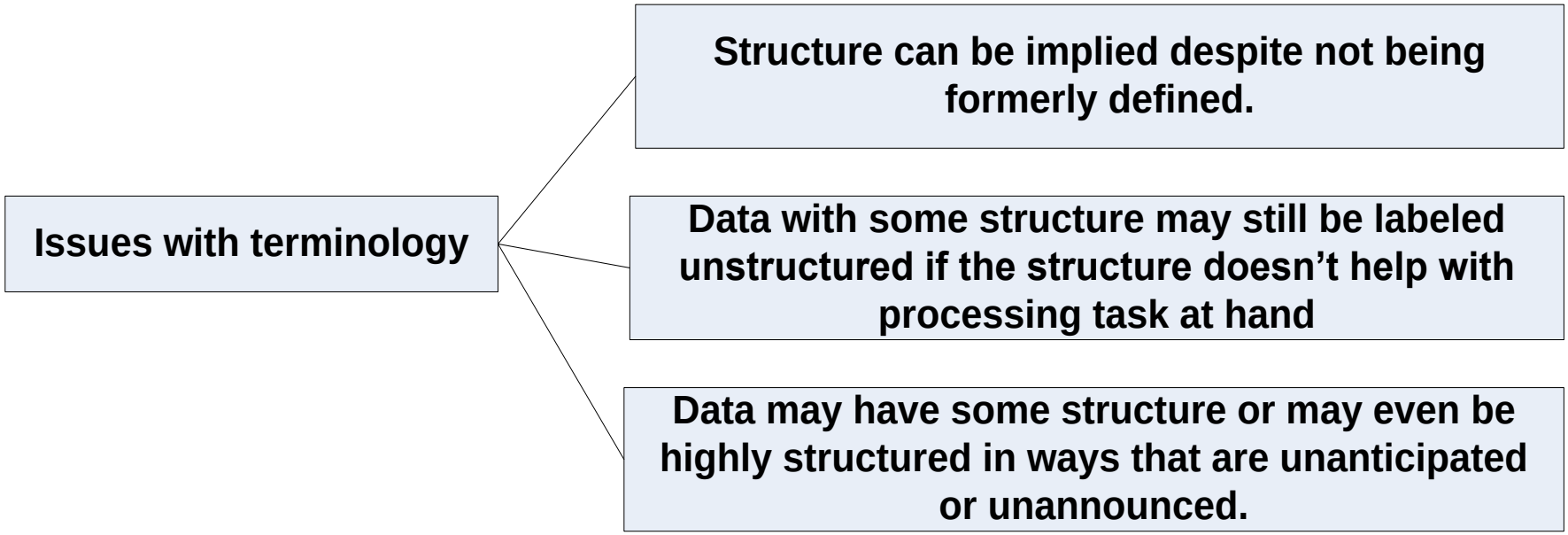
Every year, companies in the US shell out $600 billion to manage bad quality data

68%

Out of the total unstructured data in 2015, 68% can be attributed to consumers.


data in 2015, 68% can be attributed to consumers.
WHAT IS THE SPEED AT WHICH DATA IS GENERATED?
10,000 Tweets per sec
2283 Images per sec
1792 Skype calls per sec
49466 Google searches per sec
103,310 Videos viewed per sec
2,406,488 E-mails sent per sec
55 Million status updates per day
28,260 Gigabytes of traffic flows
through the Internet per second.

Issues with terminology – Unstructured Data



BIG DATA INFORMATION VARIES, AND COMES FROM MANY SOURCES

Nearly all of the information today is unstructured data.

Unstructured data accounts for more than

90%

of data in today's **organizations**.

Enterprises have liability or responsibility for nearly

80%

of the information in the **digital universe**.

How to Deal with Unstructured Data?

- Today, unstructured data constitutes approximately 80% of the data that is being generated in any enterprise.

MANAGE THE SURGE IN

UNSTRUCTURED DATA

WITH

INTELLIGENT NETWORKING

Massive amounts of new digital information are flooding enterprise networks everyday. A powerful transformation is happening. People and machines are interacting with each other differently over networks. All information today is unstructured data.

ARE YOU PREPARED TO MEET THESE CHALLENGES?

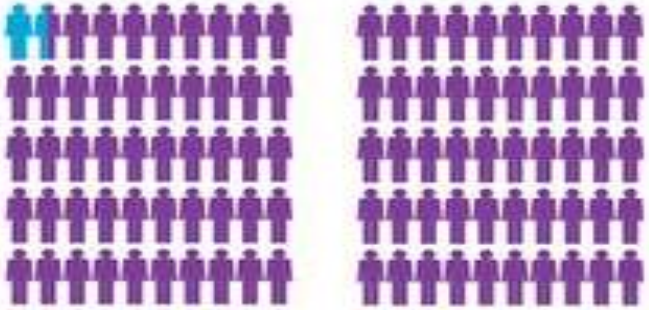
The average enterprise will need to **manage**

50 times

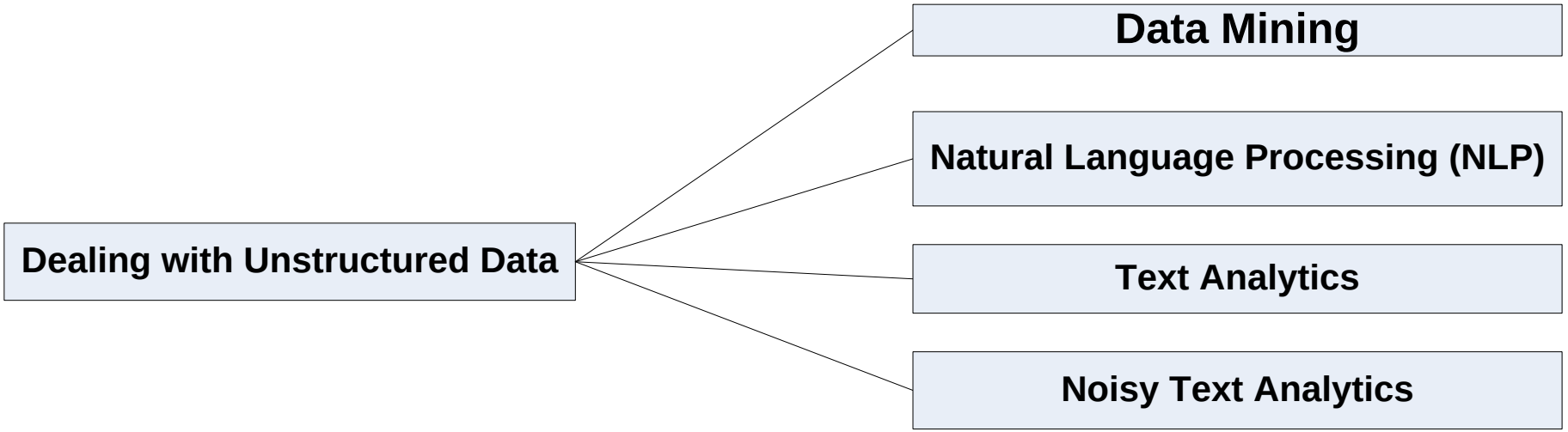
more information by **2020.**

While **increasing** IT staff by only

1.5%



Dealing with Unstructured Data



Issues with "Unstructured" Data

- **Data Mining:**
  - First, we deal with large data sets.
  - Second, we use methods at the intersection of artificial intelligence, machine learning, statistics, and database systems to unearth consistent patterns in large data sets and/or systematic relationships between variables.
  - It is the analysis step of the "knowledge discovery in databases" process.

Issues with "Unstructured" Data

Few popular data mining algorithms are as follows:

- ***Association rule mining:***
  - It is also called "market basket analysis" or "affinity analysis".
  - It is used to determine "What goes with what?"
  - It is about when you buy a product, what is the other product that you are likely to purchase with it.
  - For example, if you pick up bread from the grocery, are you likely to pick eggs or cheese to go with it.

Issues with "Unstructured" Data

- ***Regression analysis:***
  - It helps to predict the relationship between two variables.
  - The variable whose value needs to be predicted is called the dependent variable and the variables which are used to predict the value are referred to as the independent variables.

Issues with "Unstructured" Data

- ***Collaborative filtering:***
  - It is about predicting a user's preference or preferences based on the preferences of a group of users.
- For example, take a look at Table next slide.
  - We are looking at predicting whether User 4 will prefer to learn using videos or is a textual leaner depending on one or a couple of his or her known preferences.
  - We analyze the preferences of similar user profiles and on the basis of it, predict that User 4 will also like to learn using videos and is not a textual learner.

Issues with "Unstructured" Data

Table . Sample Record depicting learner's preferences for model of learning

Issues with "Unstructured" Data

- **Text Analytics or Text Mining:** Compared to the structured data stored in relational databases, text largely unstructured, amorphous, and difficult to deal with algorithmically.
- Text mining is the process of gleaning high quality and meaningful information (through devising of patterns and trends by means of statistical pattern learning) from text.
- It includes tasks such as text categorization, text clusterirg, sentiment analysis, concept/entity extraction, etc.

Issues with "Unstructured" Data

- **Natural language processing (NLP):** It is related to the area of human computer interaction. It about enabling computers to understand human or natural language input.
- **Noisy text analytics:** It is the process of extracting structured or semi-structured information from noisy unstructured data such as chats, blogs, wikis, emails, message-boards, text messages, etc.
- The noisy unstructured data usually comprises one or more of the following: Spelling mistakes, abbreviations, acronyms, non-standard words, missing punctuation, missing letter case, filler words such as “uh”, “urn”, etc.

Issues with "Unstructured" Data

- **Manual tagging with metadata:** This is about tagging manually with adequate metadata to provide the requisite semantics to understand unstructured data.
- **Part-of-speech tagging:** It is also called POS or POST or grammatical tagging. It is the process reading text and tagging each word in the sentence as belonging to a particular part of speech such aj “noun”, “verb”, “adjective”, etc.

Issues with "Unstructured" Data

- **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 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/

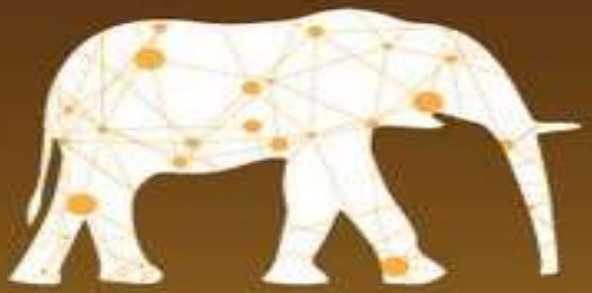
WHAT DOES THE FUTURE LOOK LIKE?

MARKET

$50.1 Billion

The potential of the Big Data market in 2017.

SIZE



44 Zettabytes

The size of the digital universe in 2020.

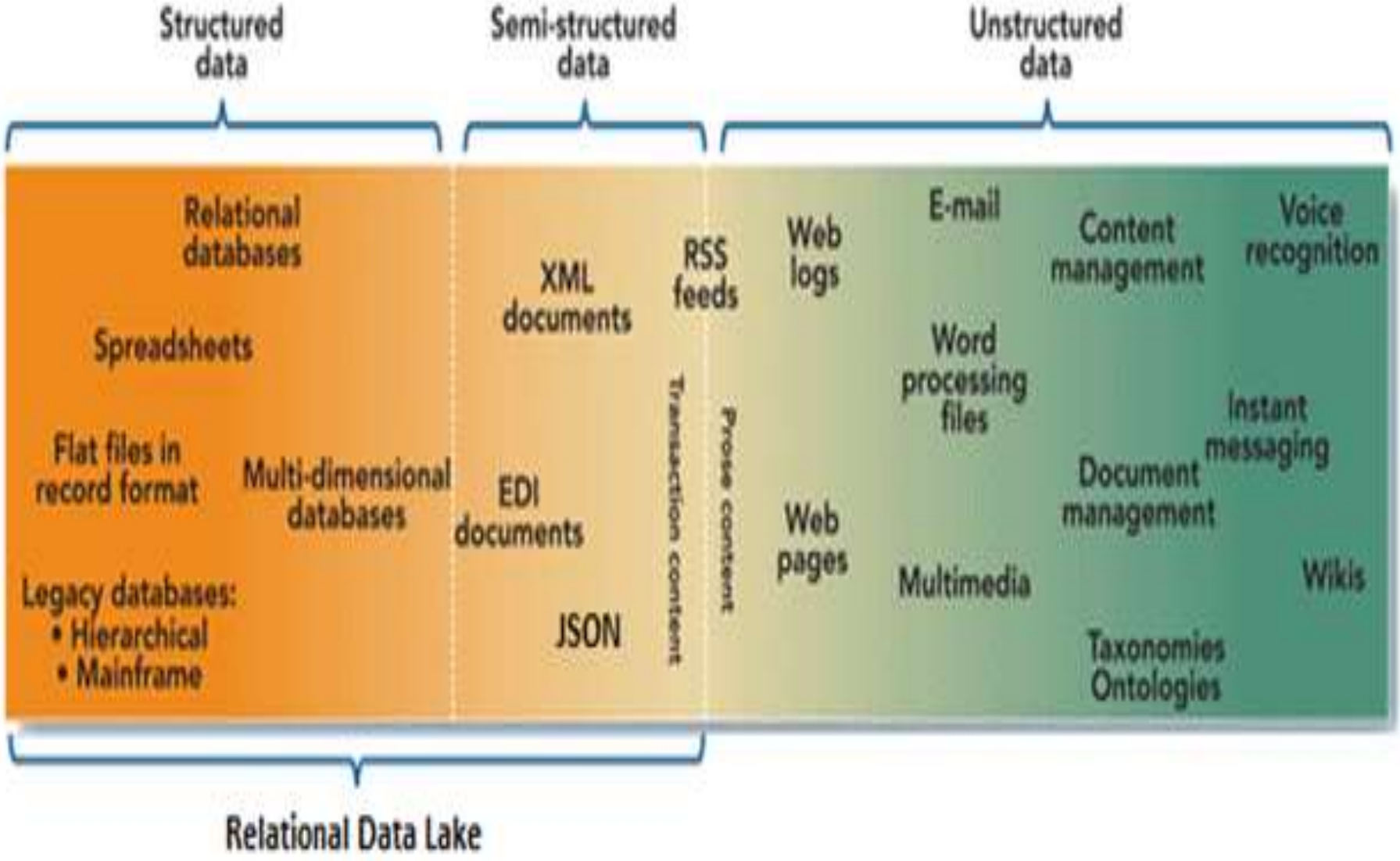
GROWTH

197,000 Petabytes

The volume of global mobile data traffic in 2019.

Summary

- *Structured data:* It conforms to a data model. For example, RDBMS conforms to relational daci model. It has a pre-defined schema.
- *Semi-structured data:* For this format of data, little metadata is available, but is insufficient. Semi-structured data have a self-describing structure. There is little or no separation between data and schema.
- *Unstructured data:* This data is growing by the day and growing by leaps and bounds. It has innumerable sources such as human generated (social media data, emails, word documents, pre-sentations, audio and video files that we create and share every day, etc.) and machine generated data (sensors, web server logs, call data records, etc.).


Structured data
Semi-structured data
Unstructured data
Relational databases
Spreadsheets
Flat files in record format
Multi-dimensional databases
Legacy databases:
• Hierarchical
• Mainframe
XML documents
RSS feeds
EDI documents
JSON
Transaction content
Prose content
Web logs
Web pages
E-mail
Word processing files
Multimedia
Content management
Document management
Taxonomies
Ontologies
Voice recognition
Instant messaging
Wikis
Relational Data Lake

TEST ME

A. Place Me in the Basket

Structured	Unstructured	Semi-Structured

Following words are to be placed in the relevant basket:

Email MS Access

Images Database

Chat conversations

Relations/Tables

Facebook Videos

MS Excel XML

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

Question‘s Answer ??

- 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.