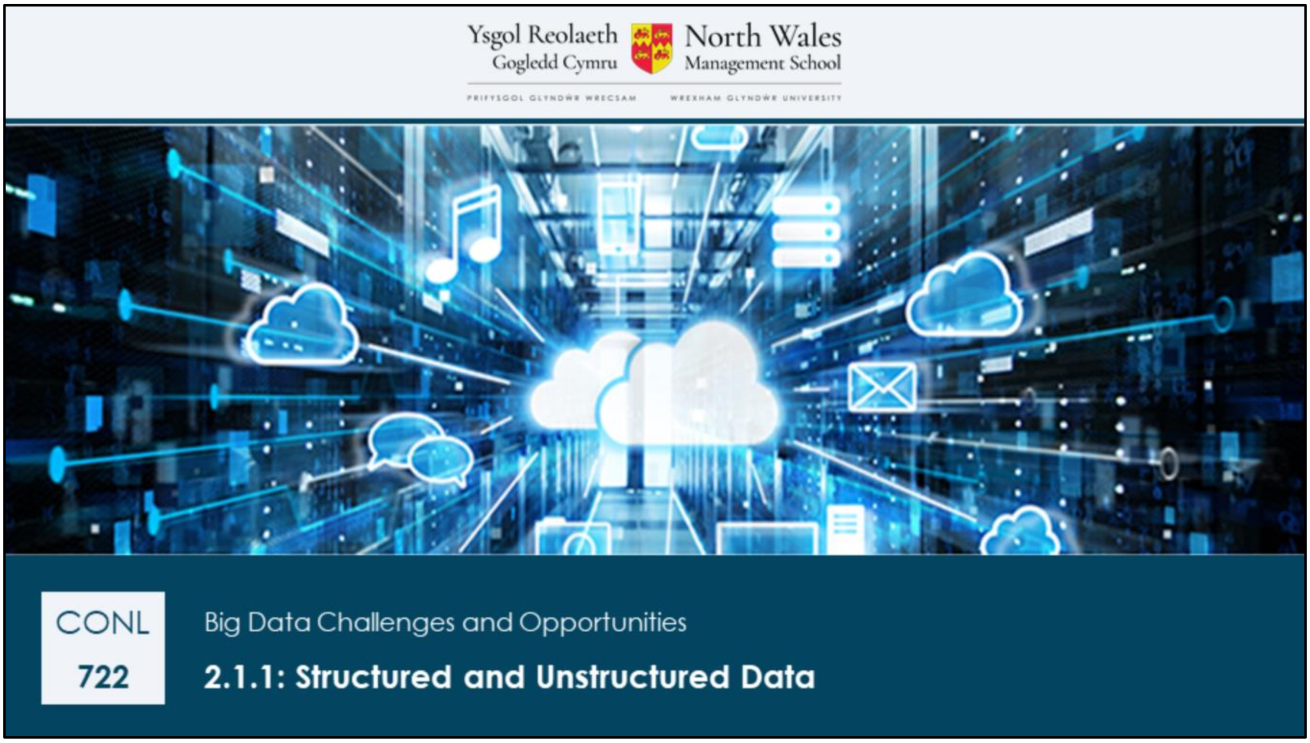


Ysgol Reolaeth
Gogledd Cymru

North Wales
Management School

PRIFYSGOL GLYNDŴR WRECSAM

WREXHAM GLYNDŴR UNIVERSITY

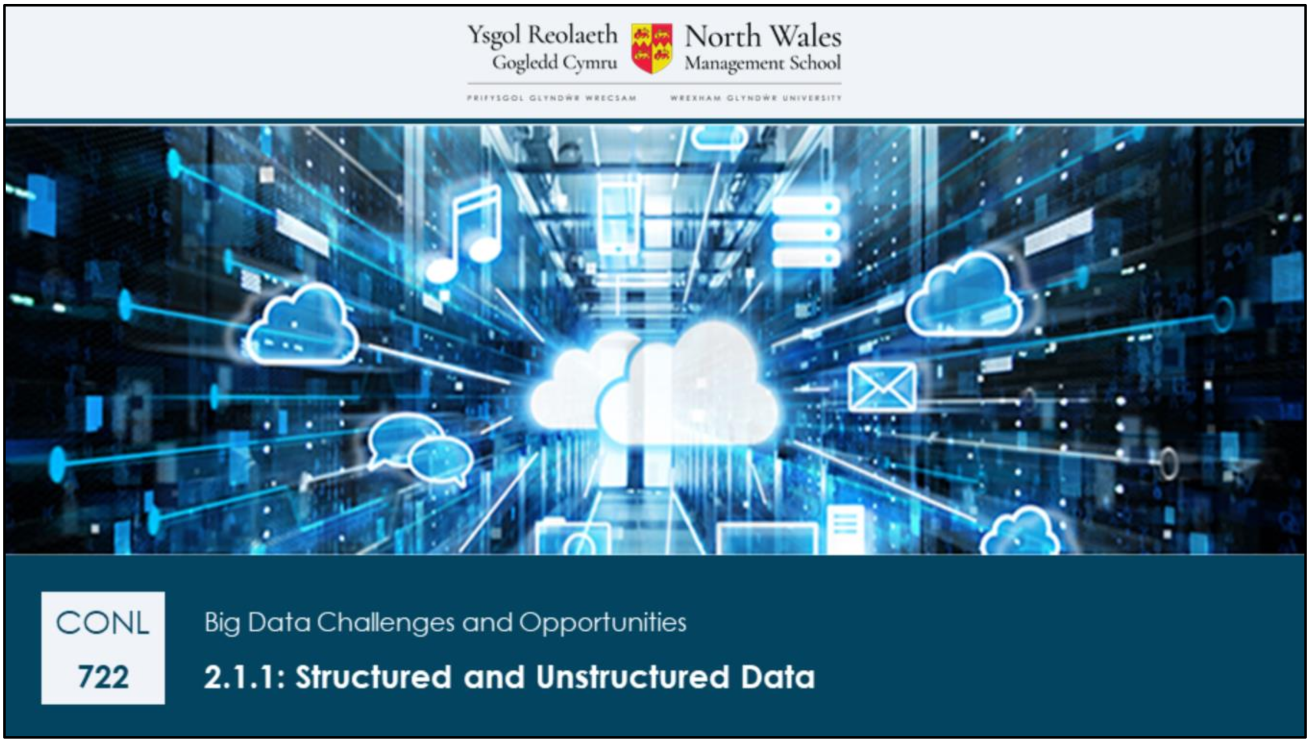
CONL
722

Big Data Challenges and Opportunities

2.1.1: Structured and Unstructured Data

During last week you have been introduced to Big data and its application. While discussing some of the examples you may have noticed that the Big Data Analytics uses complex types of data.

This week, we will be concentrating on identifying various types of data and while the majority of the data is unstructured how you will transform the data to add value and meaning.



Ysgol Reolaeth Gogledd Cymru

North Wales Management School

PRIFYSGOL GLYNDŴR WRECSAM

WREXHAM GLYNDŴR UNIVERSITY

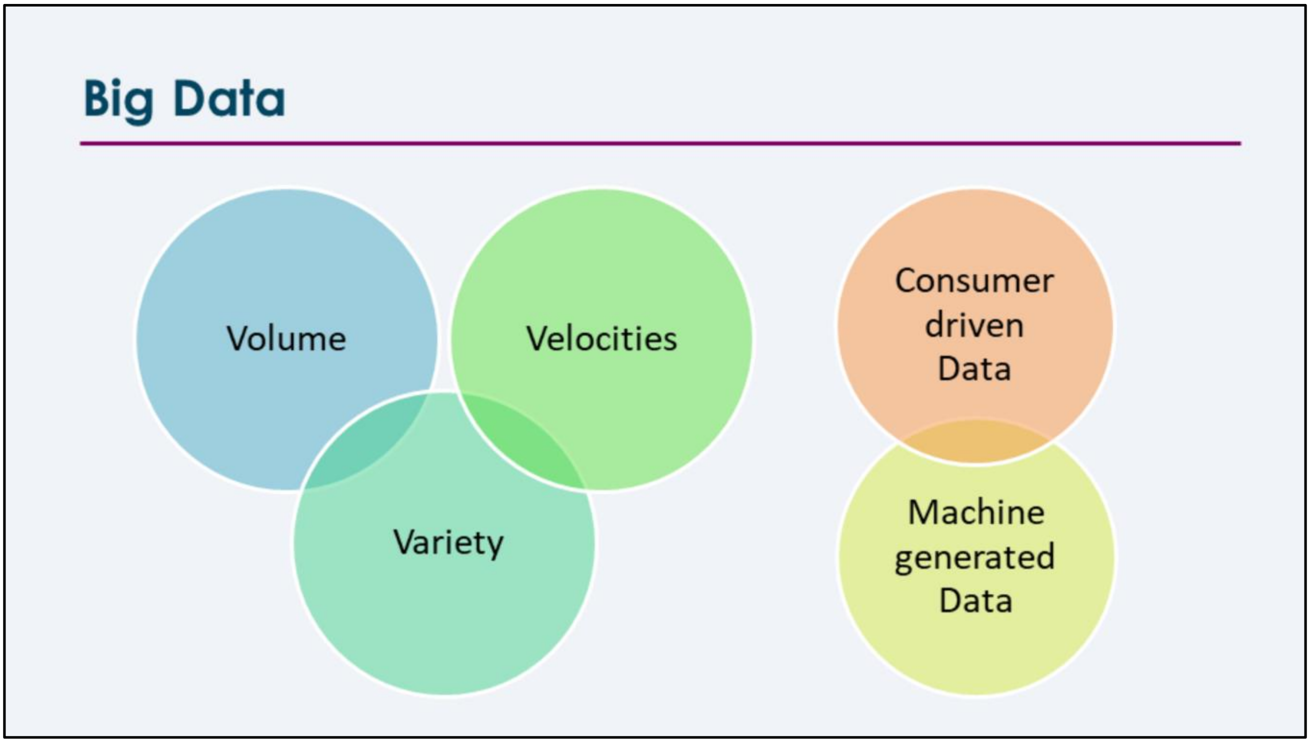
CONL 722

Big Data Challenges and Opportunities

2.1.1: Structured and Unstructured Data

During last week you have been introduced to Big data and its application. While discussing some of the examples you may have noticed that the Big Data Analytics uses complex types of data.

This week, we will be concentrating on identifying various types of data and while the majority of the data is unstructured how you will transform the data to add value and meaning.



Big Data

Volume

Velocities

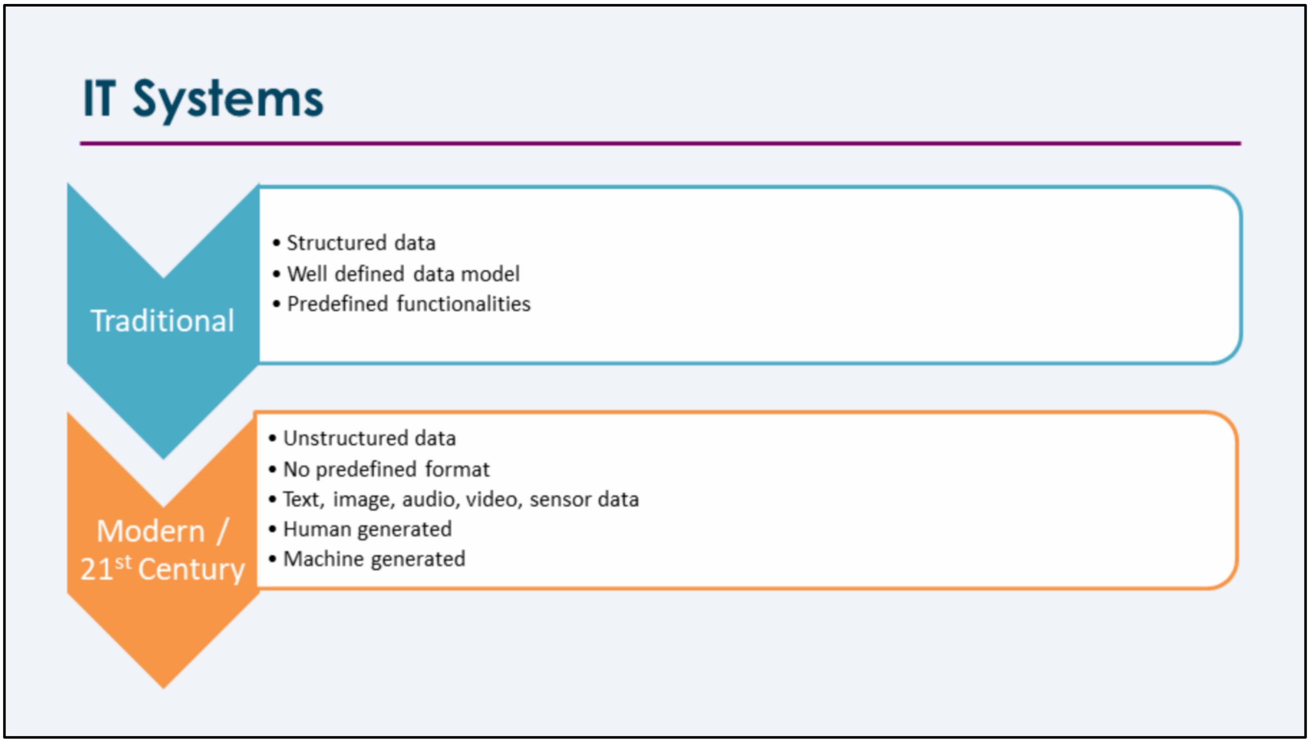
Variety

Consumer driven Data

Machine generated Data

Big data is Large Volume of Data with Varying degree of formats, complexities, and ambiguities (Variety), generated at different Velocities, which cannot be processed using traditional technologies, methods, algorithms or any off-the-shelf solutions.

The data source may include machine-generated data from sensor networks, clickstream data, scanning devices, machine log, aeroplane engines and consumer-driven data from social media, transactional data etc.



IT Systems

Traditional

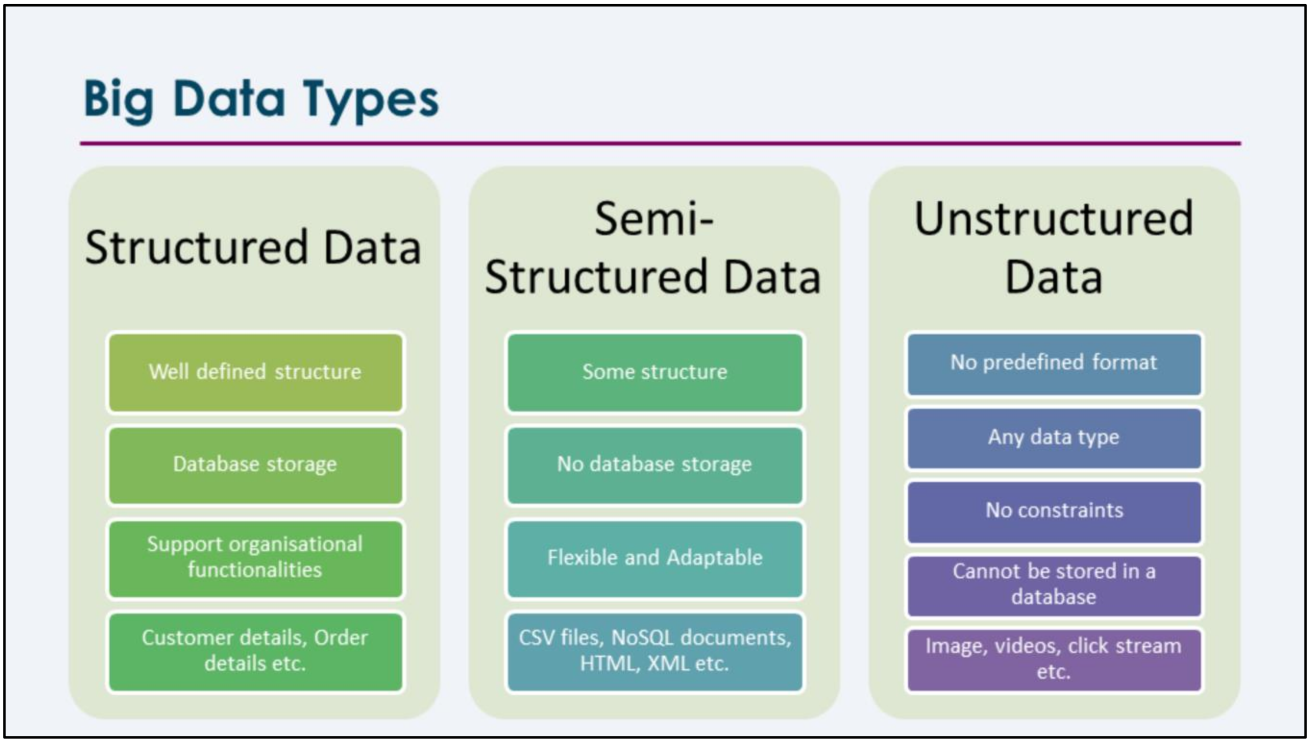
- Structured data
- Well defined data model
- Predefined functionalities

Modern / 21st Century

- Unstructured data
- No predefined format
- Text, image, audio, video, sensor data
- Human generated
- Machine generated

Traditional IT Systems were designed to generate and manage structured data. Created with a well defined data model and every data handled by the system is expected to fit within the predefined structure.
The system will support Predefined functionalities, closely related to the day to day running of the organisation, also known as transaction processing system. Most of the systems were created to move the organisational activities from a paper based system to a computerised systems.

Modern / 21st Century systems are aimed to handle unstructured data. They don’t have any predefined format. It could be text, image, audio, video, or even sensor data.
The systems must deal with human generated data like documents, email, eBook, logs, blogs, social n/w etc. as well as store and interpret machine generated data like sensor data, satellite images, radar data, surveillance data etc.



Big Data Types

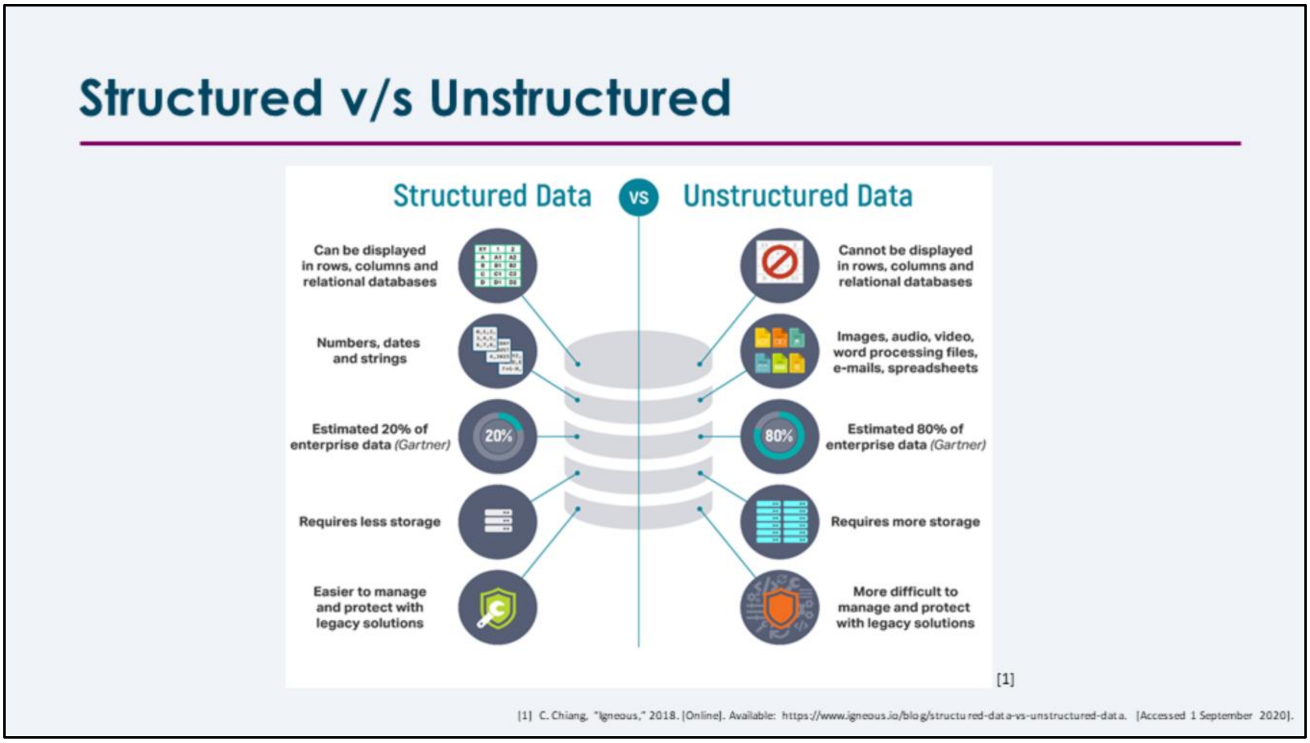
Structured Data	Semi-Structured Data	Unstructured Data
Well defined structure	Some structure	No predefined format
Database storage	No database storage	Any data type
Support organisational functionalities	Flexible and Adaptable	No constraints
Customer details, Order details etc.	CSV files, NoSQL documents, HTML, XML etc.	Cannot be stored in a database
		Image, videos, click stream etc.

Data, in general, can be classified as Structured data, semi-structured data and unstructured data.
As you have seen in the previous slide the traditional IT systems support structured Data, where the data has a well-defined structure and most often stored in a database.
The data is designed to support various organisational functionalities and often managed by Structured Query Language.
Customer details, order details etc. can be examples for structured data. Every customer will have a Customer ID, Name, Address, Phone number etc.
Order Details will include Order ID, Customer ID, Order date, Total etc.

Semi-Structured Data has some structure and share common characteristics, but do not have a rigid structure like a tabular form in a relational database, hence it is not feasible to store semi-structured data in a database.
The advantage of the semi-structured data is, to a certain extent it is flexible and adaptable and can accommodate variations in the structure.
CSV files, NoSQL documents, HTML, XML etc. are some examples. Unstructured Data, most of the Big Data resource falls into this category.

As the name suggests un-structured data has no predefined format. It can take any type of data and has no constraints imposed by the data architecture.
So it is quite obvious that it cannot be stored in a database.
Any data without the structure can be an example, hence the list will be very long and

will grow all the time.
Just get a feel of what can be considered here is a brief list of unstructured data, image, videos, clickstream data, sensor data, log files etc.



Structured v/s Unstructured

Structured Data VS Unstructured Data

Can be displayed in rows, columns and relational databases

Numbers, dates and strings

Estimated 20% of enterprise data *(Gartner)*

Requires less storage

Easier to manage and protect with legacy solutions

Cannot be displayed in rows, columns and relational databases

Images, audio, video, word processing files, e-mails, spreadsheets

Estimated 80% of enterprise data *(Gartner)*

Requires more storage

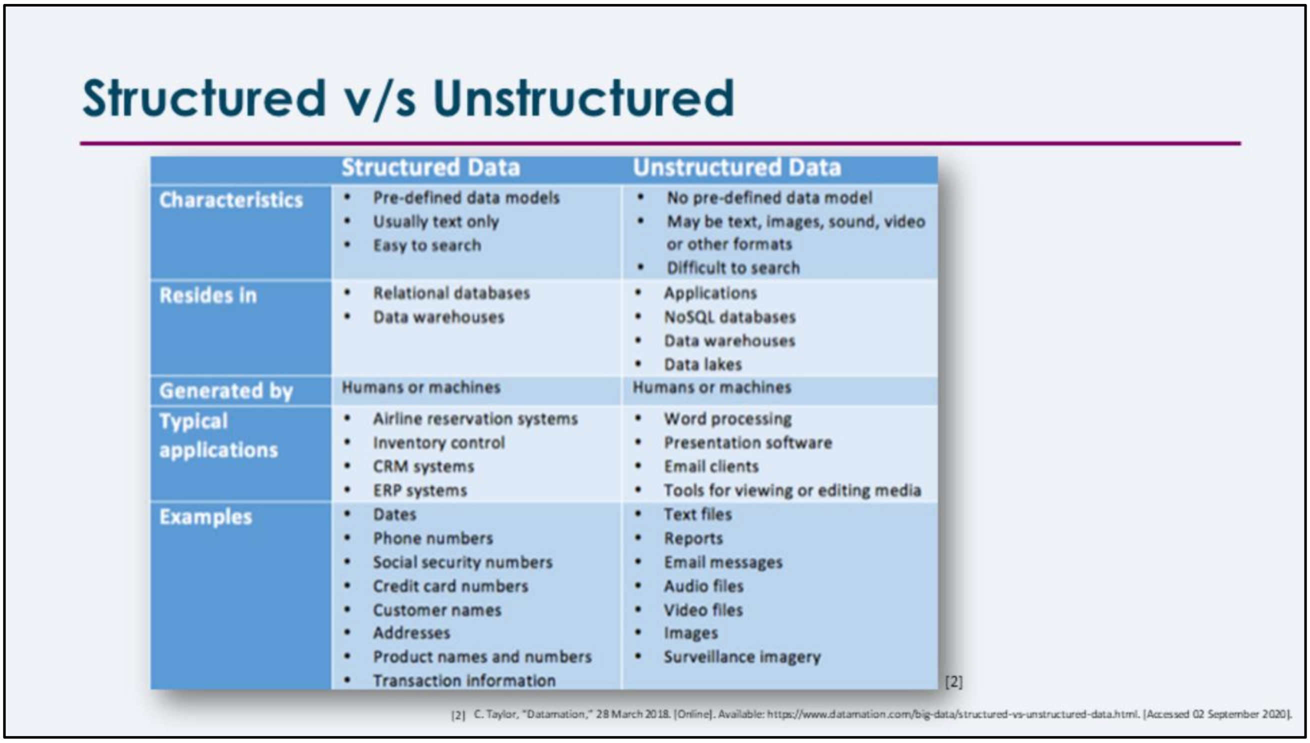
More difficult to manage and protect with legacy solutions

[1]

[1] C. Chiang, "Igneous," 2018. [Online]. Available: https://www.igneous.io/blog/structured-data-vs-unstructured-data. [Accessed 1 September 2020].

The diagram shown on the slide provides some basics comparison between structured and unstructured data. Please pause the video and have a read.

Slide 6:The diagram shown on the slide provides some basics comparison between structured and unstructured data. Structured data v/s Unstructured data in terms of appearance, general data types, the quantity of data generated, storage requirements and management issues etc.



Structured v/s Unstructured

	Structured Data	Unstructured Data
Characteristics	• Pre-defined data models • Usually text only • Easy to search	• No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	• Relational databases • Data warehouses	• Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	• Airline reservation systems • Inventory control • CRM systems • ERP systems	• Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	• Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

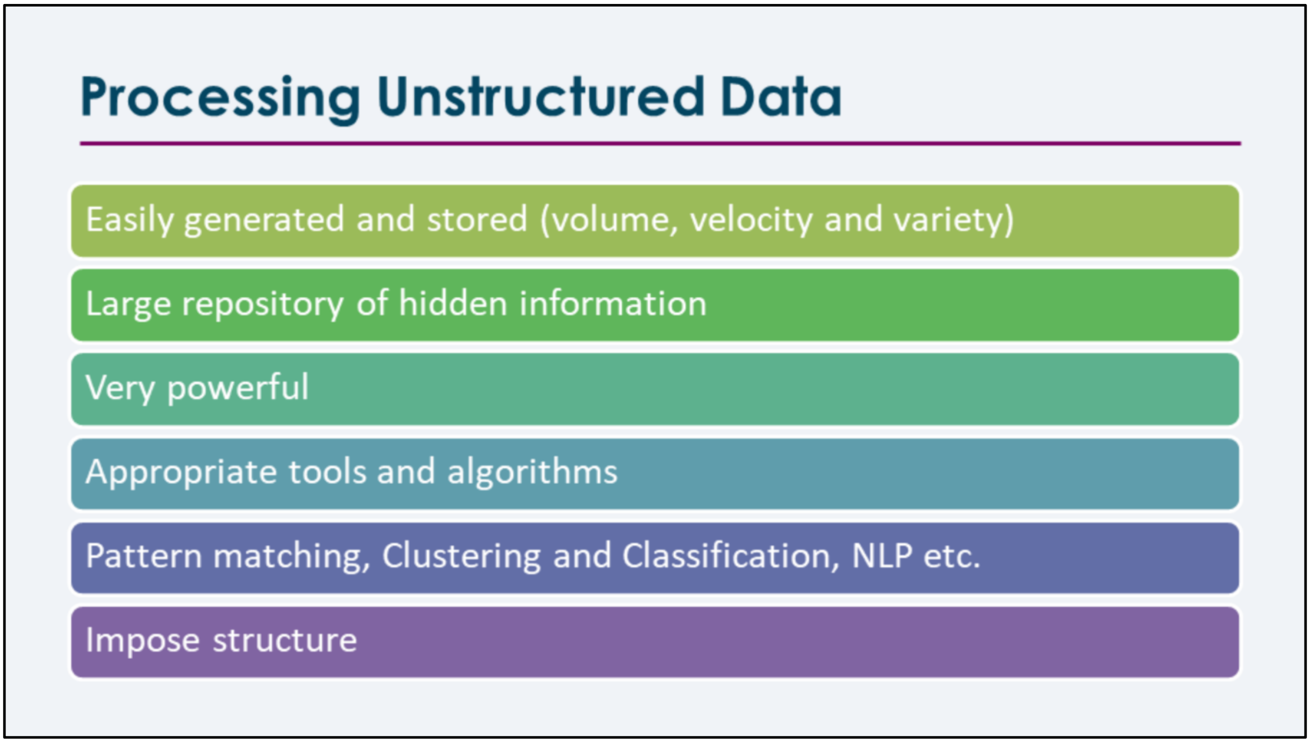
[2]

[2] C. Taylor, "Datamation," 28 March 2018. [Online]. Available: https://www.datamation.com/big-data/structured-vs-unstructured-data.html. [Accessed 02 September 2020].

The details on this slide are also very similar to the previous slide, comparing the two on slightly different factors as well as examples for application and data itself.
Pause the video and read the details.

This diagram provides a comparison between the structured data and unstructured data in terms of its characteristics, where is it stored, how it is generated, applications of them along with some examples.

Here you can see why the unstructured is fall under the big data umbrella.



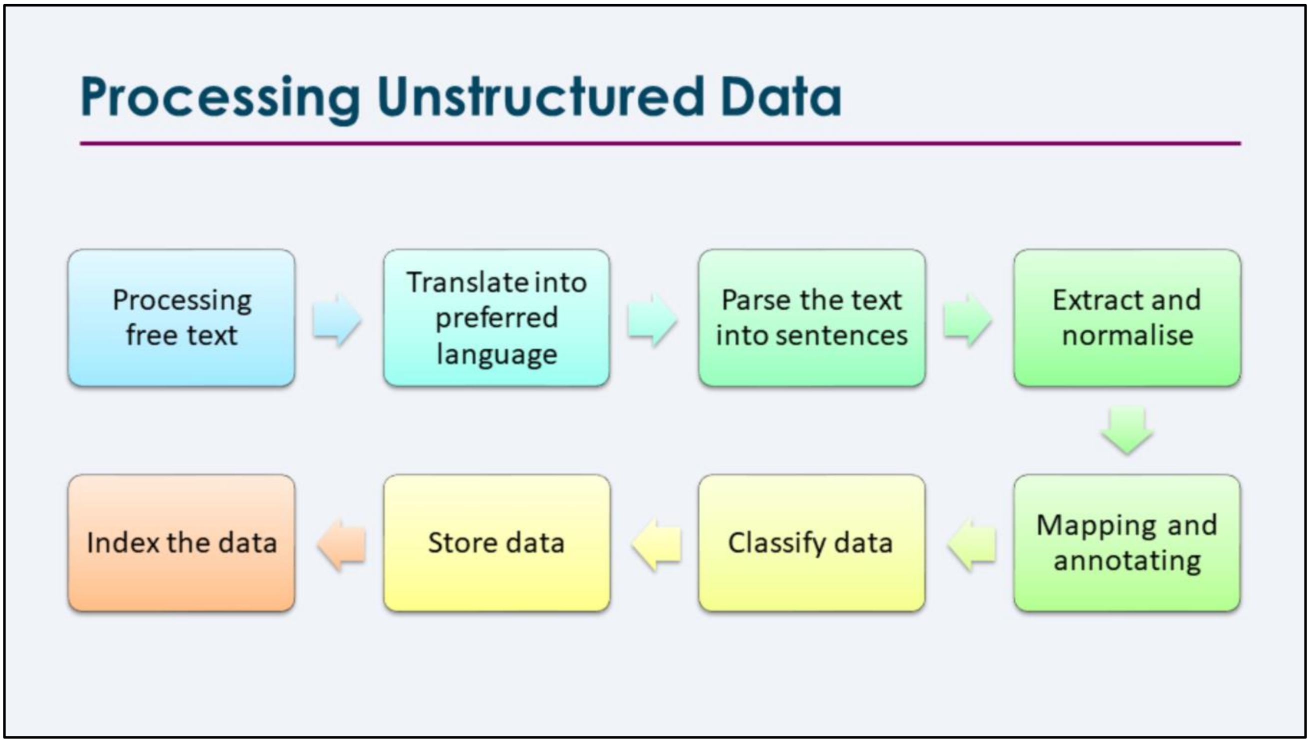
Processing Unstructured Data

- Easily generated and stored (volume, velocity and variety)
- Large repository of hidden information
- Very powerful
- Appropriate tools and algorithms
- Pattern matching, Clustering and Classification, NLP etc.
- Impose structure

Unstructured data is easily generated and can also be easily stored (volume, velocity and variety).
It is certainly a large repository of hidden information.
However, processing the unstructured data is much more complex in comparison to structured or even semi-structured data.

Using appropriate tools and algorithms will help to harvest the power of unstructured data.
Some of the techniques that are used in the data analysis are Natural Language Processing, Pattern Sensing, Clustering and Classification, Text-mining, Sentiment Analysis etc.

Processing may also require imposing some structure on the unstructured data.
The following slide will explain a simple example.



Processing Unstructured Data

Processing free text

Translate into preferred language

Parse the text into sentences

Extract and normalise

Mapping and annotating

Classify data

Store data

Index the data

For example, processing free text to get retrieve some information may involve imposing a structure to the free text.
To accomplish this, the first step is translating the text into the preferred language and parse the text into sentences which is the fundamental unit of a natural language.

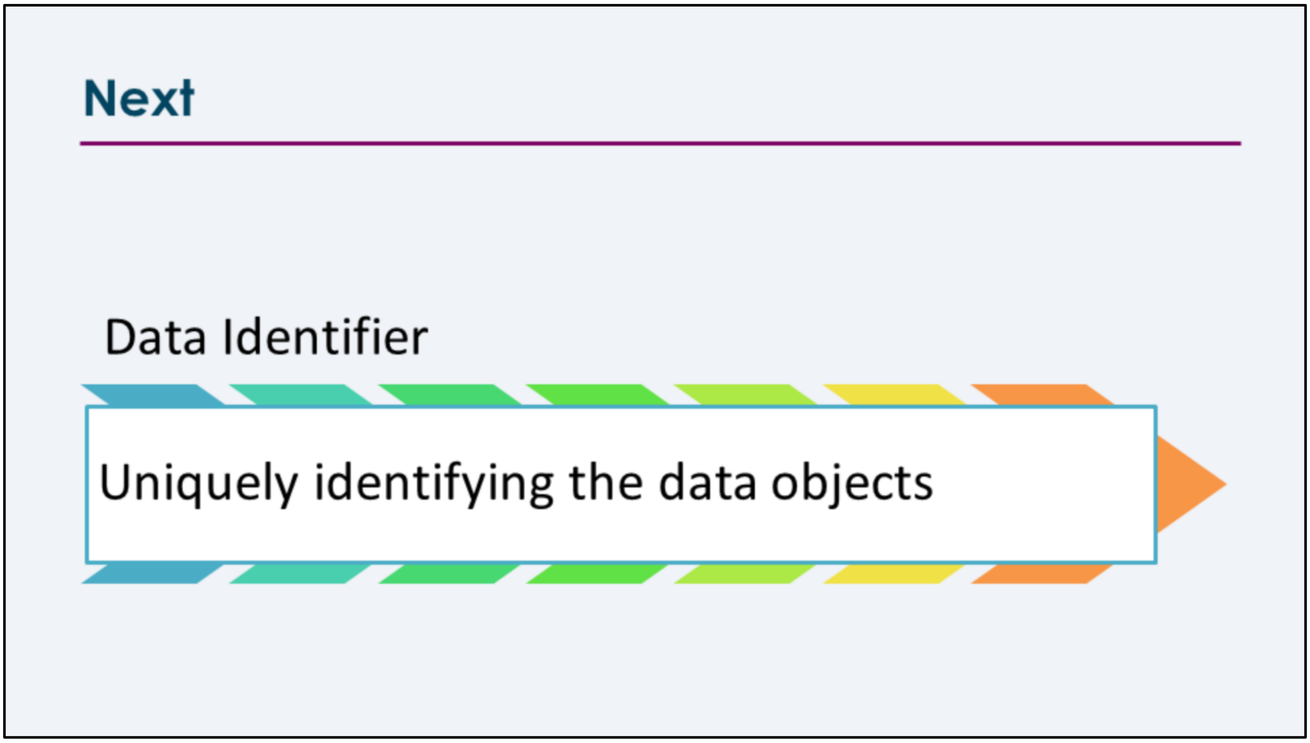
Then extract and normalise the conceptual terms from the text, normalise could also be interpreted as minimise.

Map the identified terms into standard nomenclature and then annotate or tag them.
Now the data will have some descriptions and values.

The next step is associating the data into an appropriate classification.
The classified data can then be assigned to a data storage and retrieval system.
Finally indexing the data item in the system.

Think about this in larger scale, in relation to Big data resources and big data analytics.

These concepts are explained in detail with various examples and case studies in Chapter 2, Providing Structure to Unstructured data of the core text. Please spend some time reading the book chapter.



Next

Data Identifier

Uniquely identifying the data objects

Big data resources are mainly unstructured. These data objects must be identified and described.
Using data identifiers each data objects can be uniquely identified.
Once identified the relationships and associations could be derived and analysed.

The next video will discuss the importance of data identifiers in the Big Data domain and the challenges in identifying and assigning unique data identifiers.