Ysgol Reolaeth Gogledd Cymru North Wales Management School

PRIFYSGOL GLYNDŴR WRECSAM WREXHAM GLYNDŴR UNIVERSITY

CONL 722

Big Data Challenges and Opportunities

2.1.2: Identifiers and Identifier Systems

In the context of Information Systems each data objects must be identifiable and should be able to be described.
Dealing with structured data stored in a database each record is uniquely identified by its primary key and the attributes will describe the data objects.
The previous session explained that majority of the big data resources are unstructured and to process these data must be identified and described.

This video will discuss the concepts of identifying the data using unique identifiers.

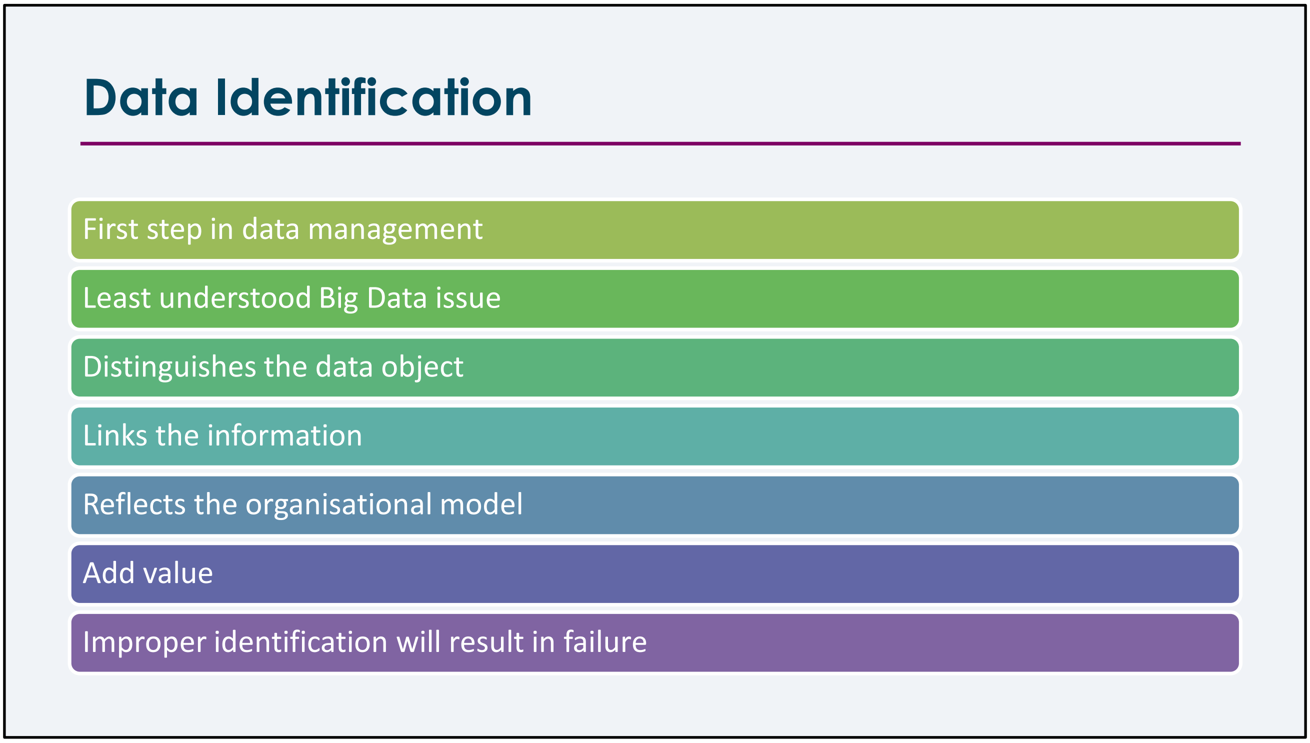
Ysgol Reolaeth Gogledd Cymru
North Wales Management School
PRIFYSGOL GLYNDŴR WRECSAM
WREXHAM GLYNDŴR UNIVERSITY

CONL 722

Big Data Challenges and Opportunities

2.1.2: Identifiers and Identifier Systems

Hello and Welcome to the first session for CONL722, Big Data challenges and opportunities.

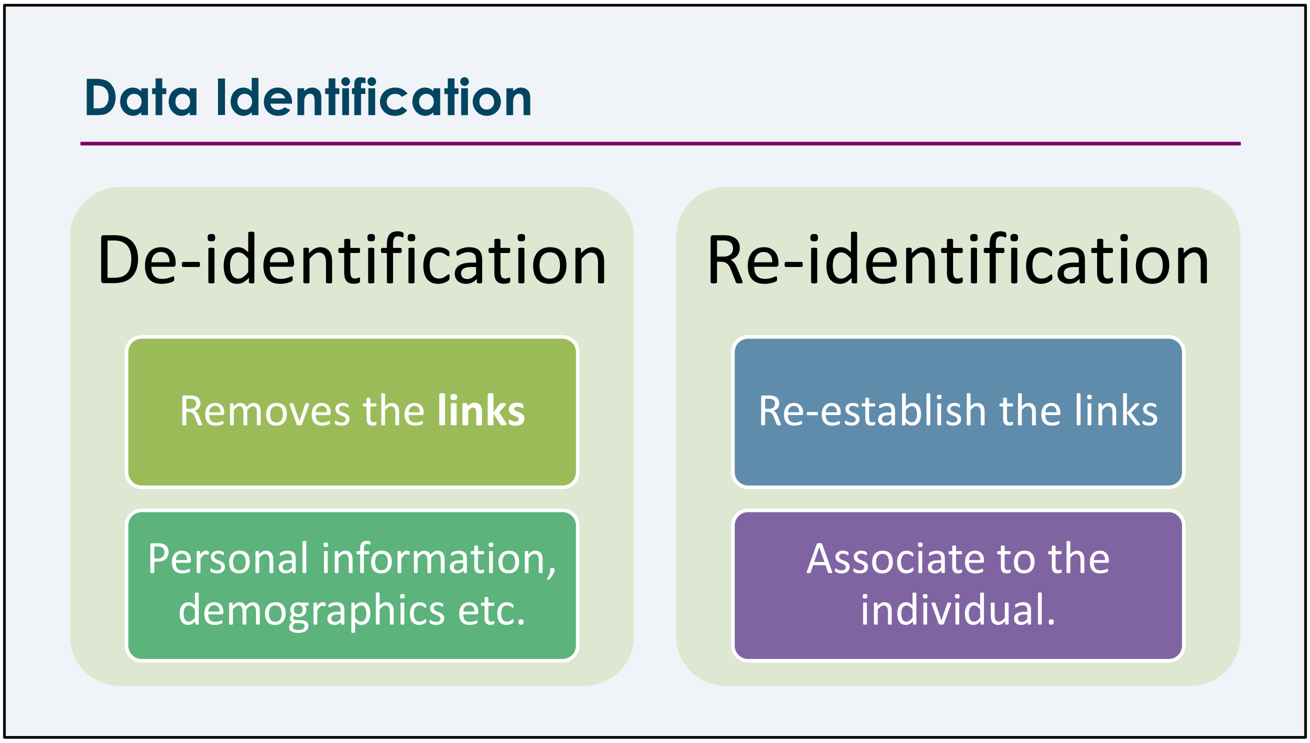


Data Identification

- First step in data management
- Least understood Big Data issue
- Distinguishes the data object
- Links the information
- Reflects the organisational model
- Add value
- Improper identification will result in failure

Data identification is the first step in data management, however, it is one of the least understood big data issue.
Only with the appropriate system of factors and data sources, the analysis can bring the needed insight. Data will have no meaning without a relevant identifier.
Data identifier will distinguish each data object and at the same time links the information associated with the identified data object.
The data objects and methods of identification will reflect the organisational model of big data resources.

Appropriate identification will add value to the big data resource. As I said earlier properly identified data sources with appropriate factors can be analysed to harness the power and this can then be properly analysed.
Hence ignoring the identification or improper identification will result in big data project failures.



Data Identification

De-identification

Removes the **links**

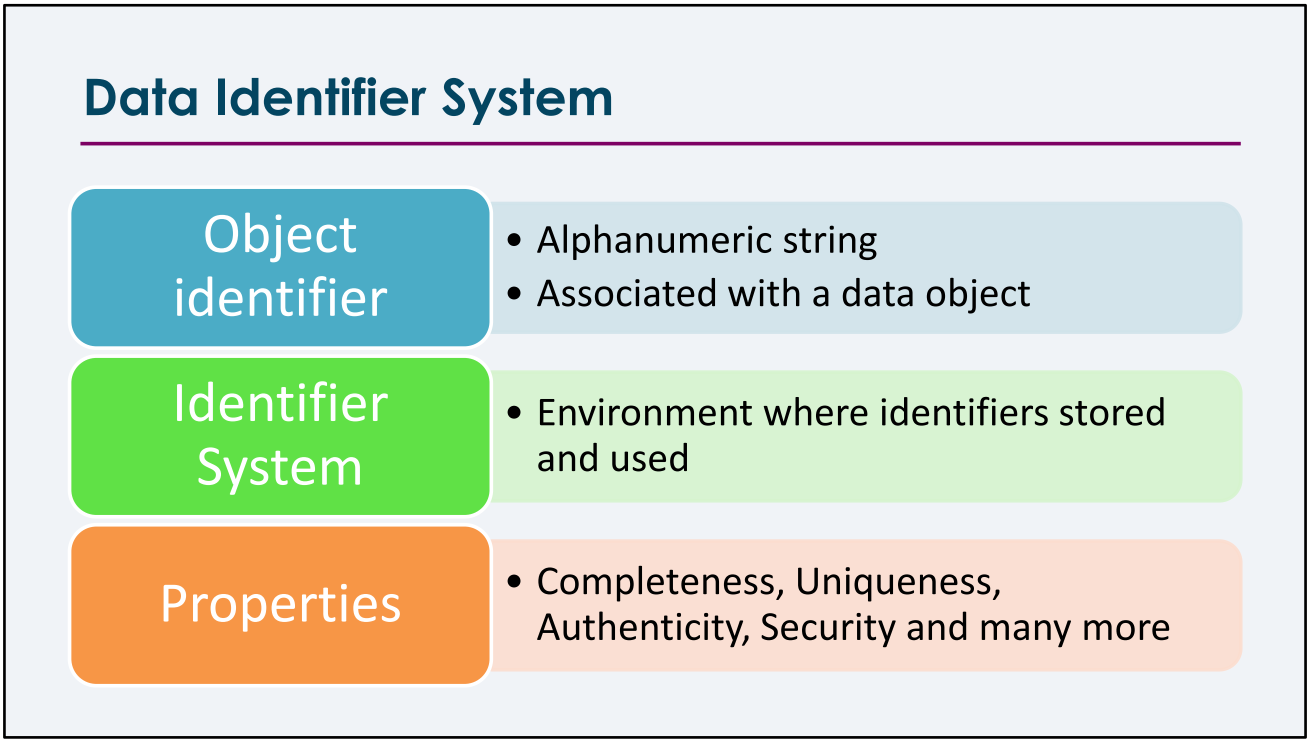
Personal information, demographics etc.

Re-identification

Re-establish the links

Associate to the individual.

Once the data objects are properly identified, it can be de-identified and re-identified.
De-identification removes the **links** that associate any data object to an individual.
Removes the link to the personal information, demographics etc.
Re-identification establishes the links that are removed by de-identification.
Which means it reassigns the link between the information and the person associated with that information.



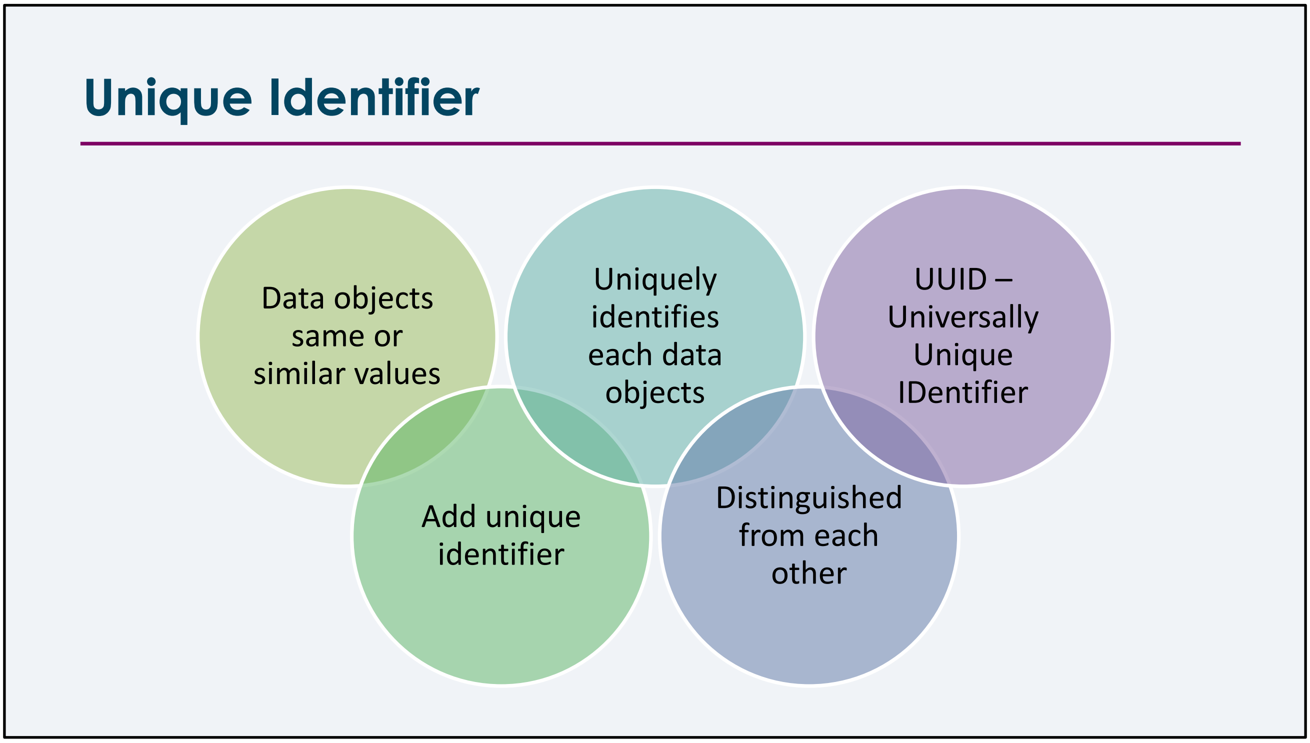
Data Identifier System

Object identifier
- Alphanumeric string
- Associated with a data object

Identifier System
- Environment where identifiers stored and used

Properties
- Completeness, Uniqueness, Authenticity, Security and many more

Data or the object identifier can be an alphanumeric string associated with a data object.
The role is to uniquely identify or distinguish the object from other objects.
The identifier systems provide a permanent environment for storing and using identifiers.
Good identifier systems have many properties like completeness, uniqueness, authenticity, security, autonomy etc.
A detailed listing of these properties and explanation can be found in chapter 3 of the core text. Please spare some time to read the chapter.



Unique Identifier

Data objects same or similar values

Add unique identifier

Uniquely identifies each data objects

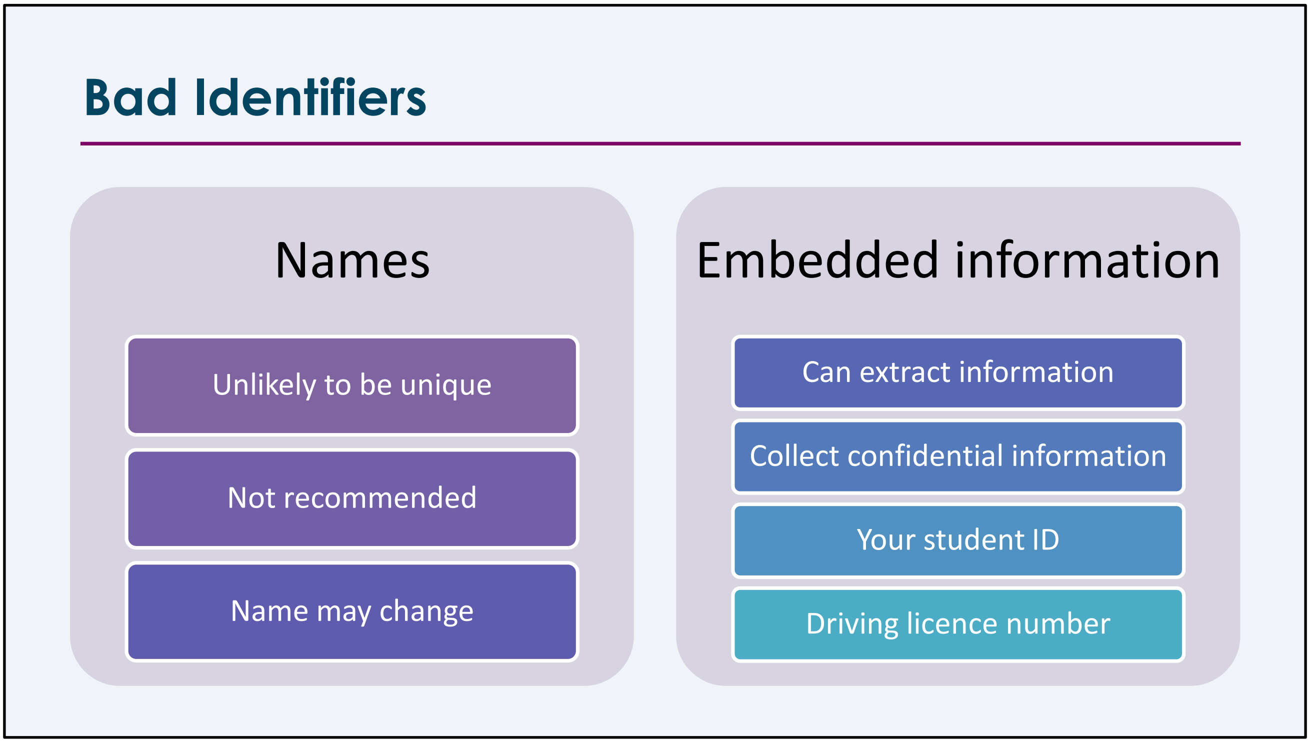
Distinguished from each other

UUID – Universally Unique IDentifier

Data objects are place holders for data values and descriptions. Value could be the same in more than one object.
Associating a unique identifier to a data object guarantees the uniqueness of the object. Very similar to assigning unique student id to each one of you.
Individual data object will become distinguishable from each other, but not form itself.
Your student ID is automatically generated as you enrol. It is generated by combining a few different set of details and is only valid to identify you here in the University.

Whereas big data deals with data objects generated and managed across the globe, hence need an identifier that is universally unique.

Universally Unique Identifier is an example for one type of an algorithm which creates 128-bit long identifier.
60 bit is computed using the computer timestamp from where the data object is generated.
Collisions are possible where the algorithm generates the same ID for different objects, however, it is extremely low as less than even one every quintillion (10^18, 10 to the power 18). , it is almost 1 every 2.7 quintillion.



Bad Identifiers

Names

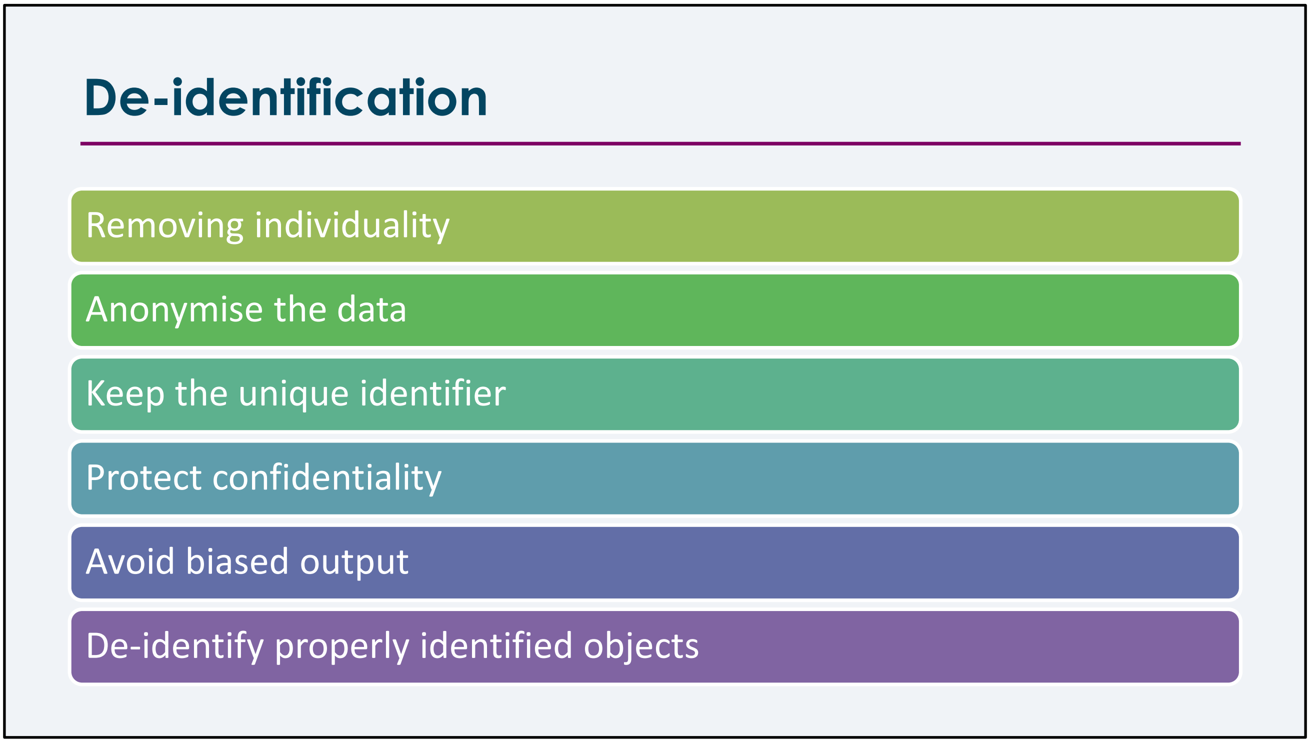
- Unlikely to be unique
- Not recommended
- Name may change

Embedded information

- Can extract information
- Collect confidential information
- Your student ID
- Driving licence number

You should always think about and consider what makes a good unique identifier.
Could it be just a random number? Don’t forget that not all possible identifiers are good.
Make sure you are familiar with what can be categorised as bad identifiers and why and avoid them.

For example, consider identifiers using names. It is very unlikely to be unique – is this is the only reason? No.
Even if it is unique, it is not recommended as an identifier because they can be volatile.
for example, the name may change, especially female surname, department name, course name they all can change.
Any identifiers with embedded information should also be avoided. Familiarity with the system will help to extract information.
Various embedded details can be interpreted to collect confidential information.
For example, check your student ID, can you extract any information from that? What about your Driving licence number?



De-identification

- Removing individuality
- Anonymise the data
- Keep the unique identifier
- Protect confidentiality
- Avoid biased output
- De-identify properly identified objects

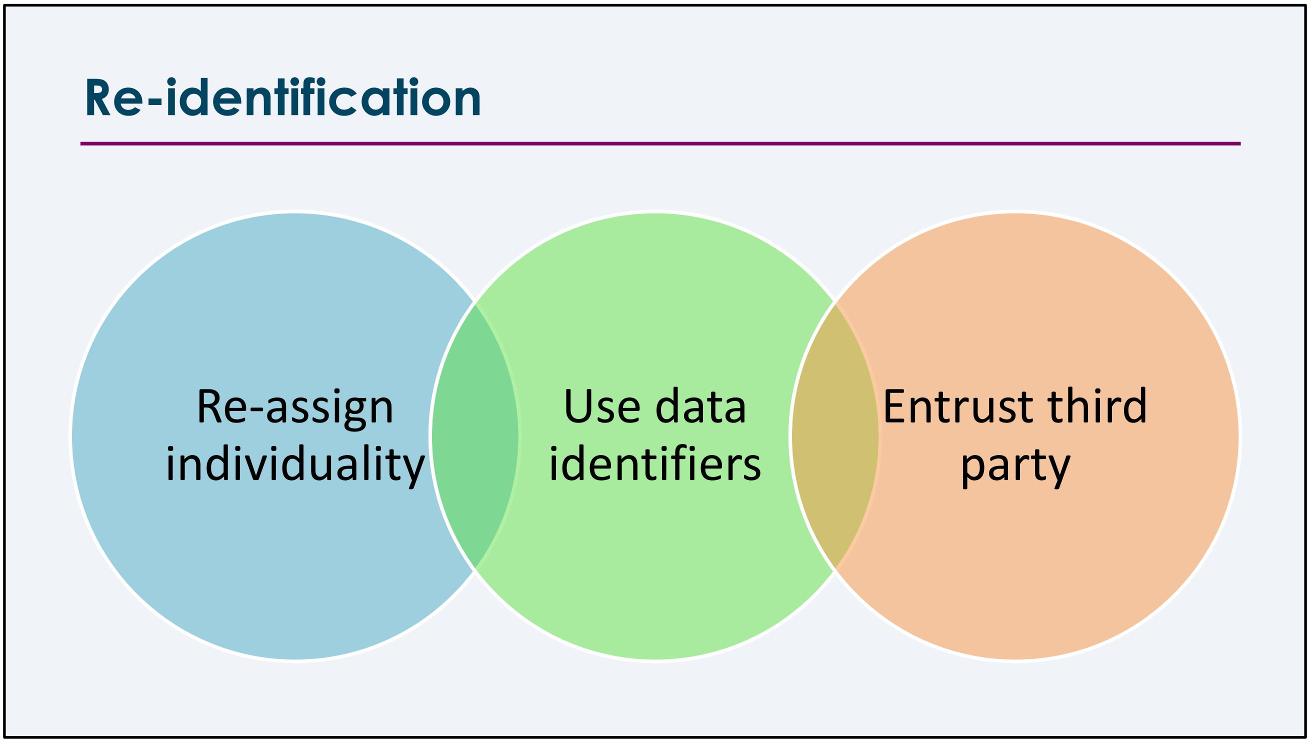
Data analysis should ensure anonymity. You should remove every possible detail that may personalise the data object, in other words, remove the link between the data object and the individual whom that data belongs to.
The process of removing the association between the data object and the individual is known as de-identification.

De-identification is not removing the identifier, in fact, is about using the identifier and the relevant information while avoiding every personal detail.

For example, analysing the success of online masters, use student details and the results. However the name of the students could be avoided, and depending on the analysis you may be able to avoid DOB, address etc. We will revisit this next week while discussing data reduction.

To a certain extent, removing the personalisation will ensure privacy and will protect confidentiality.
Also helps in avoiding biased opinions or output.

Don’t forget the fact that de-identification is only successful if the data objects are properly identified.



Re-identification

Re-assign individuality

Use data identifiers

Entrust third party

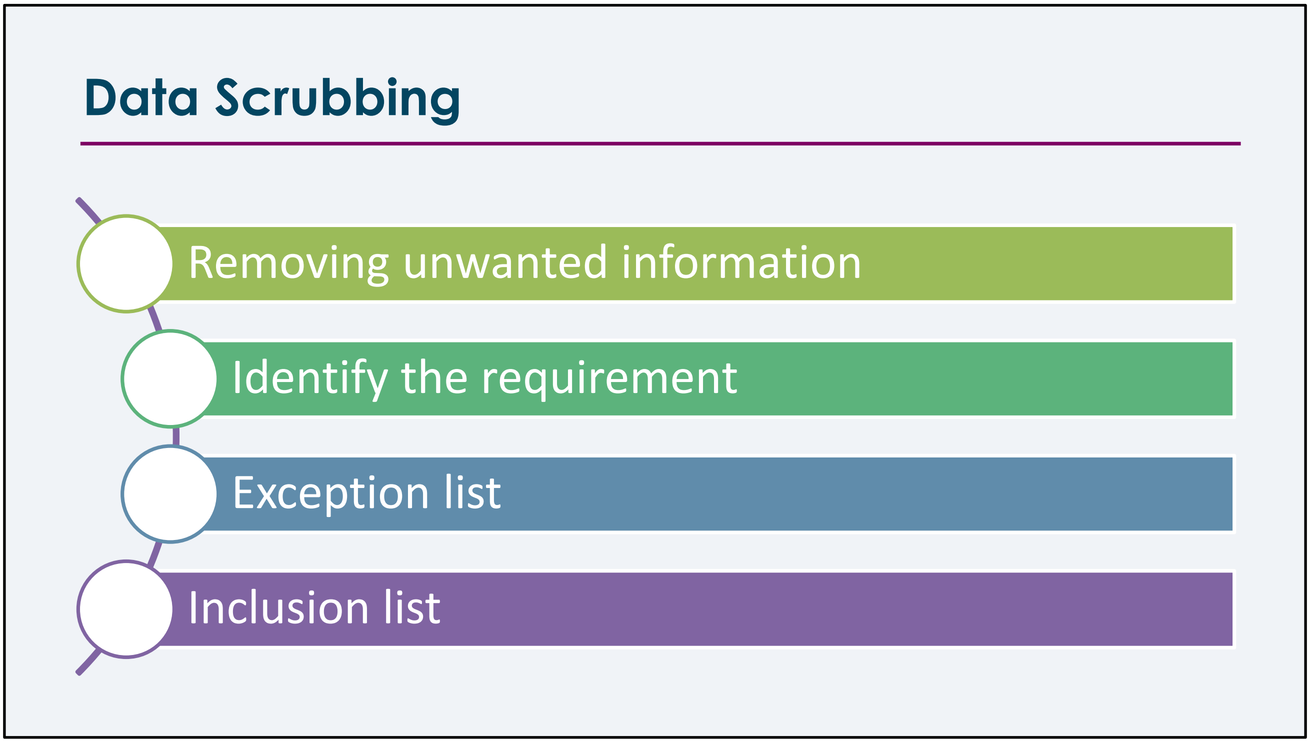
Occasionally the analysis may have to bring back the individuality of the data object. Especially in some scientific research and analysis

Re-identification is all about reconnecting the data objects to the individuals.
This should be carried out using the data identifiers. Emphasising the fact that data identification is the first and the most important step.

De-identifying removes the association to the personal data, but as we discussed in the previous slide it will not remove the association to the identifier.

Re-identification uses the data identifier from the de-identified records to find the original data objects which have the same data identifier.

This may mean trusting the third-party with records that connect to the individuals.



Data Scrubbing

- Removing unwanted information
- Identify the requirement
- Exception list
- Inclusion list

Data scrubbing is about removing unwanted information. Some times this will be misinterpreted and de-identification.
Scrubbing normally applied on de-identified data objects, in other words, they are not the same.
Scrubbing starts after de-identification.

There are different methods for data scrubbing.
You can create an exception list, listing the information that needs to be avoided.
But this can’t always guarantee a clean data.

Another method is to create a list of acceptable data that can be included in the final data set, that is after the de-identification and scrubbing.
Data objects which are not in the list will be automatically deleted.

Next

Metadata

Data about the data

This lesson has discussed the data types and explained the 80% of data resources in the Big data systems are of unstructured data.
We have also discussed uniquely identifying data objects and the importance of de-identifying to protect privacy and confidentiality.

Data identifiers do not describe what information is available in the data object.

Metadata is the data about the data, in other words in the descriptive details of the data objects.
The next lesson will be concentrating on Metadata and Classification Ontologies.