



A typical organization includes many systems that will serve us as data producers.

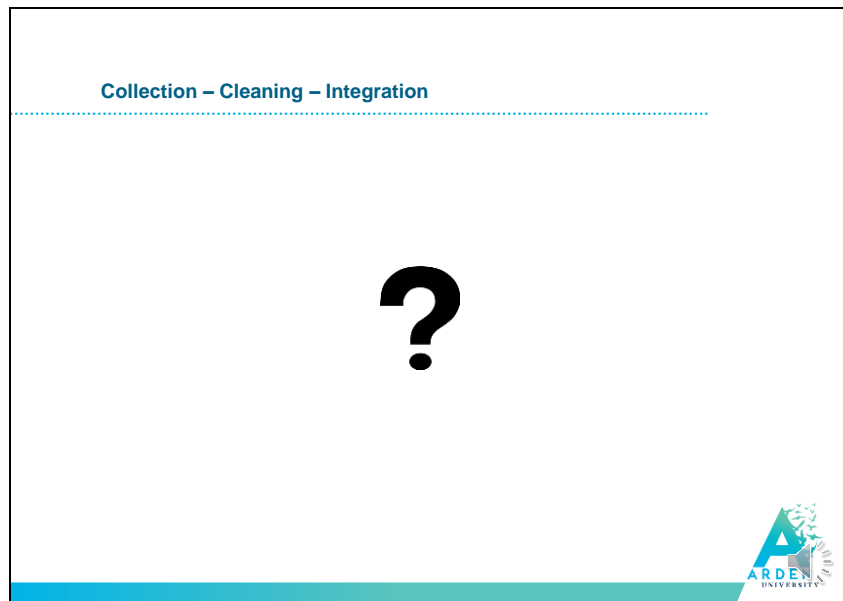
Whether these systems were designed with this task in mind or not, they are usually not coordinated to produce the same format of data.

Moreover, they are usually not designed to produce data in a format needed by a specific modelling technique. Therefore, a substantial amount of a data analytics project's time would typically be spent on collecting, cleaning and integrating data produced by different data sources.

We have no choice but to bring our data to a trustable and reliable level before we do anything with it, even just for simple reporting.

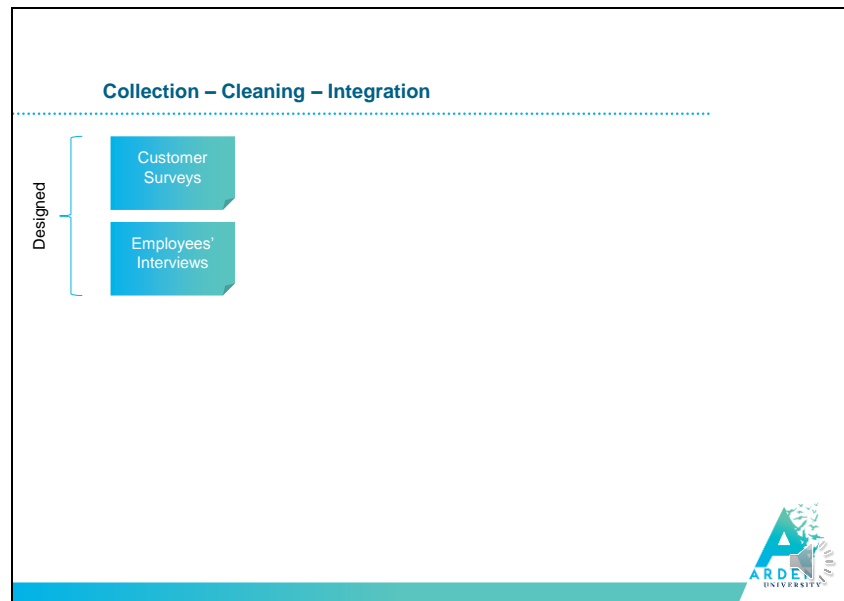
The CRISP-DM specification requires reports to be produced about how these processes were taken specifically.

The Data Handling module covers some of the issues connected to cleaning and integration. Therefore, in this lesson, as well as in the previous lesson, we will be focused more on collecting data. In this presentation we will introduce the high-level view of the process.



After formulating our business question, it is time for us to look around our organization and identify data sources which may be of relevance to it.

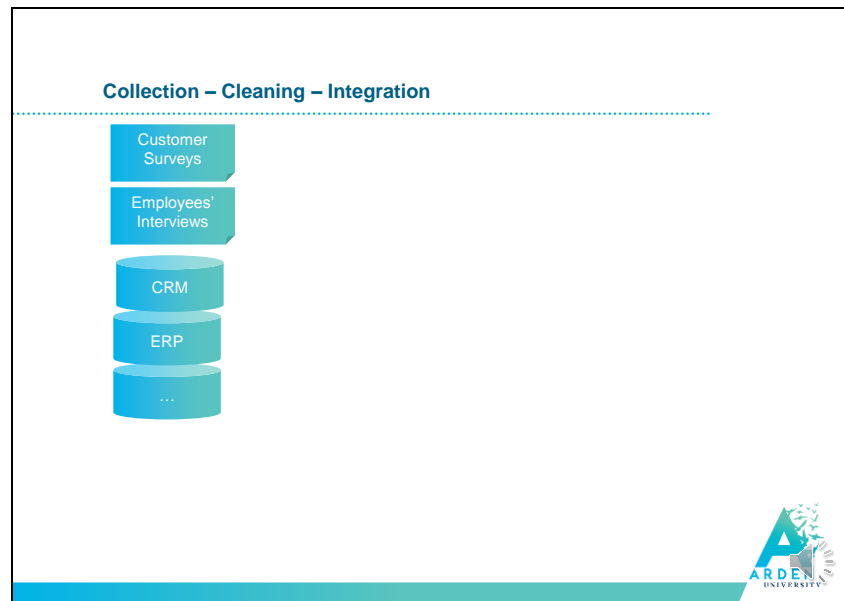
Slide 3



In the previous lesson we discussed designed methods for data collection. Thus, we may have collected data from a customer survey, or we may collected some data by interviewing our employees.

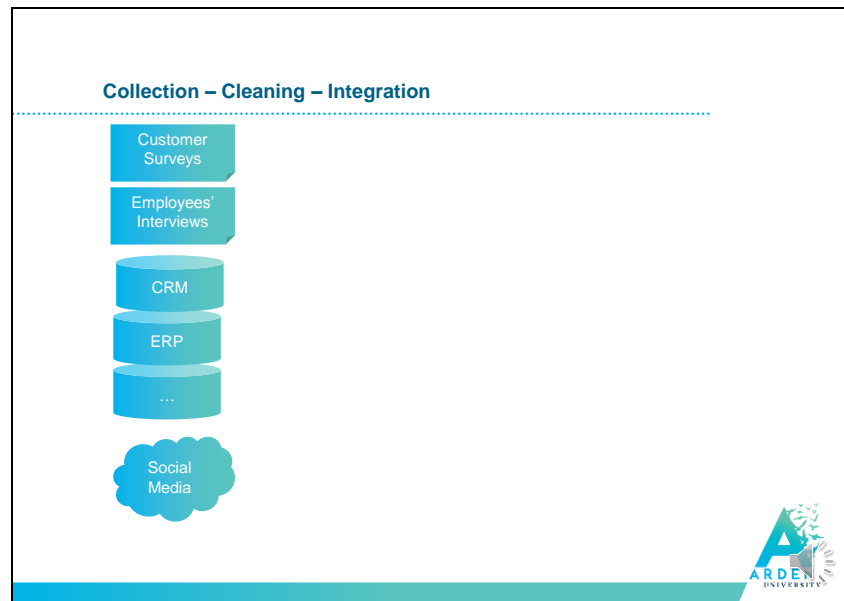
We also learned that designed methods bear their own costs and biases, and so we must strive to complement these with automatically collected data.

Slide 4



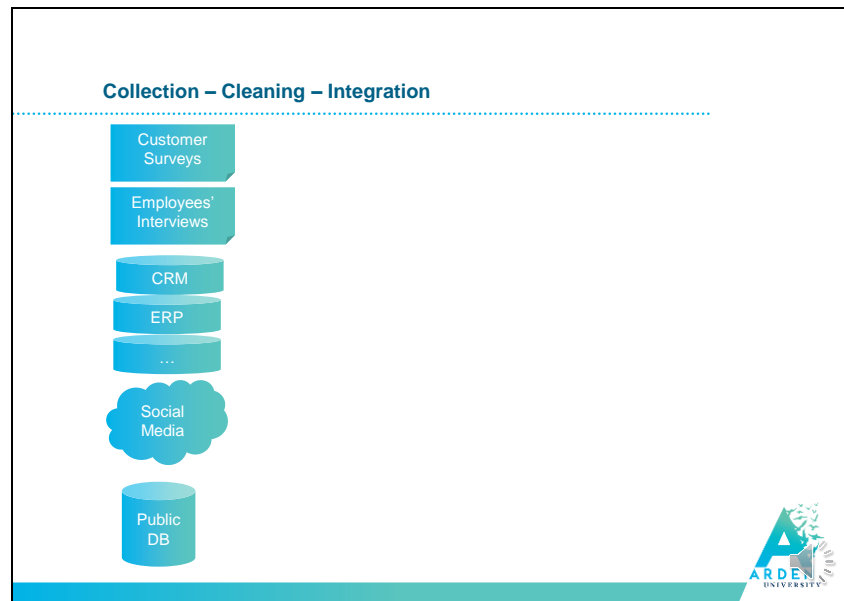
Some organizational databases such as CRM or ERP can be used to extract information about customers or other resources. Some of this data might have been manually entered into these systems, and thus might have many missing and erroneous values.

Slide 5



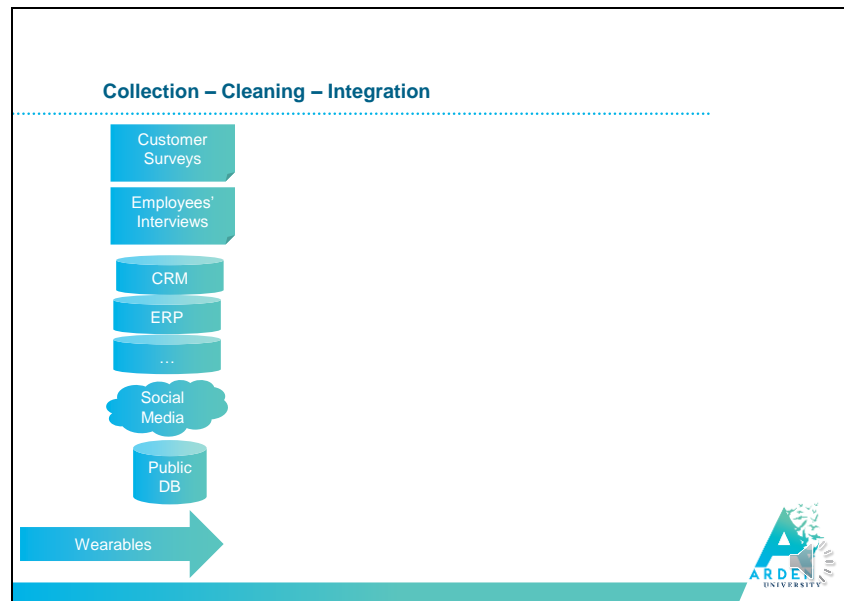
We might also want to add some data about our own brand or about our competitors from social media sites.

Slide 6



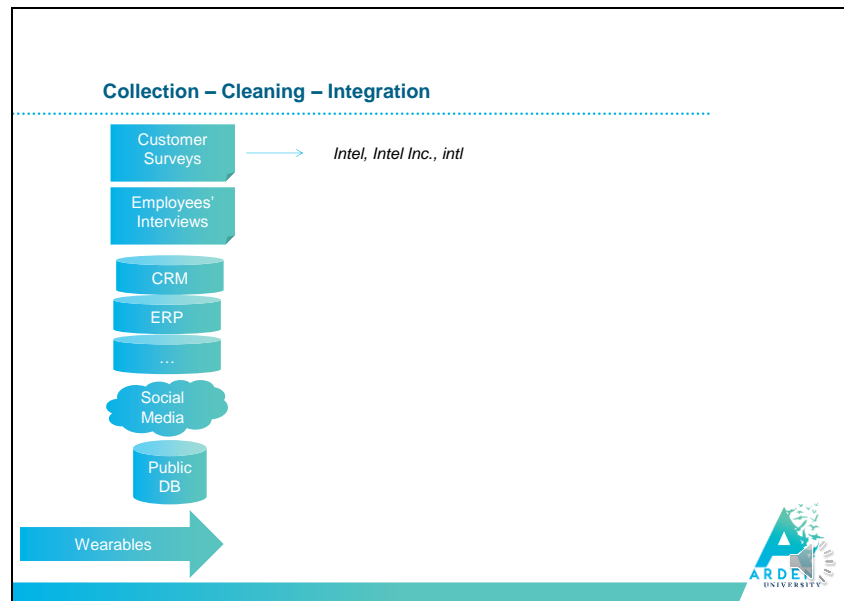
Some financial/environmental data that is publicly available

Slide 7



Or perhaps some data streaming from sensors in real time

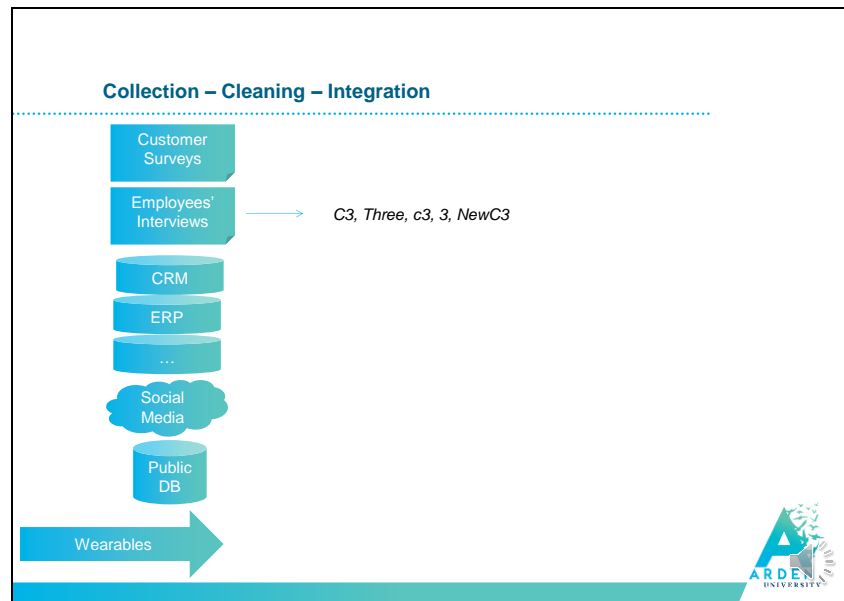
Slide 8



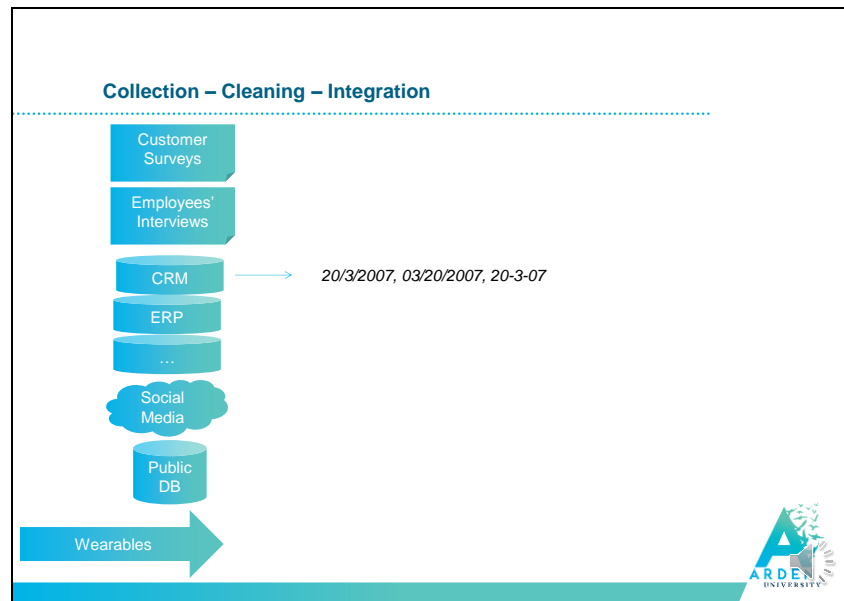
All this data arrives through data streams, spreadsheets, log files, relational databases, etc. Each has its own analysis unit and its own cleaning issues.

For example, in data coming from surveys, people may have used ambiguous naming

Slide 9

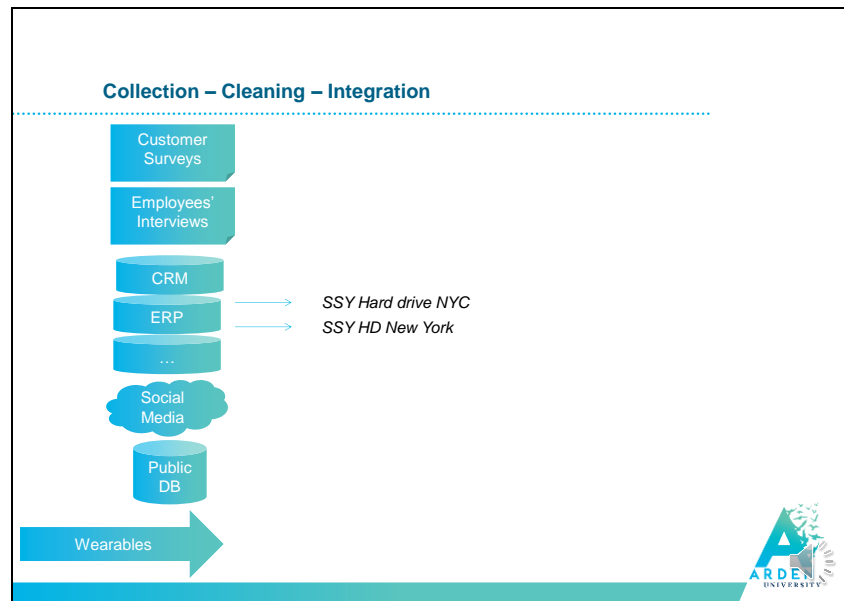


Coded data from interviews might have some coding inconsistencies, maybe because some coding systems have been updated.

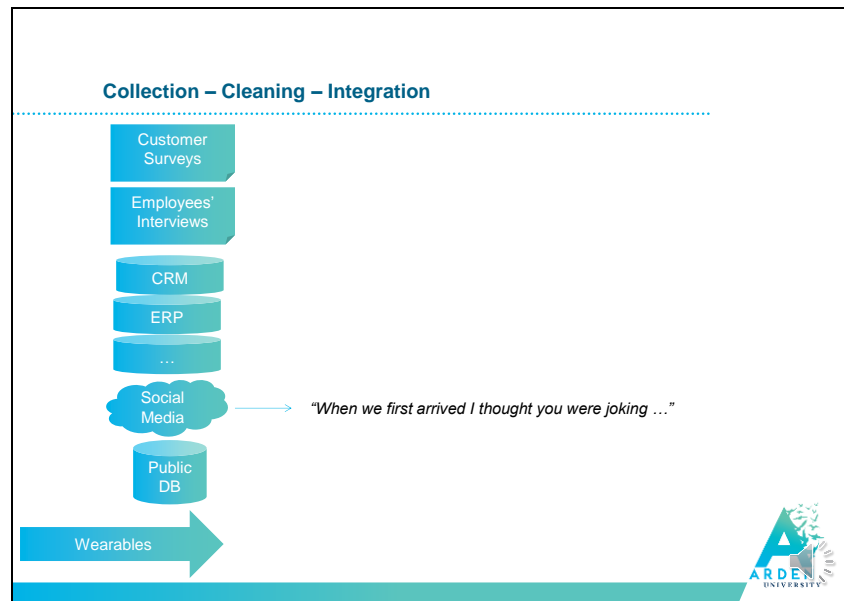


Date formats, or other measurement units, might be formatted differently.

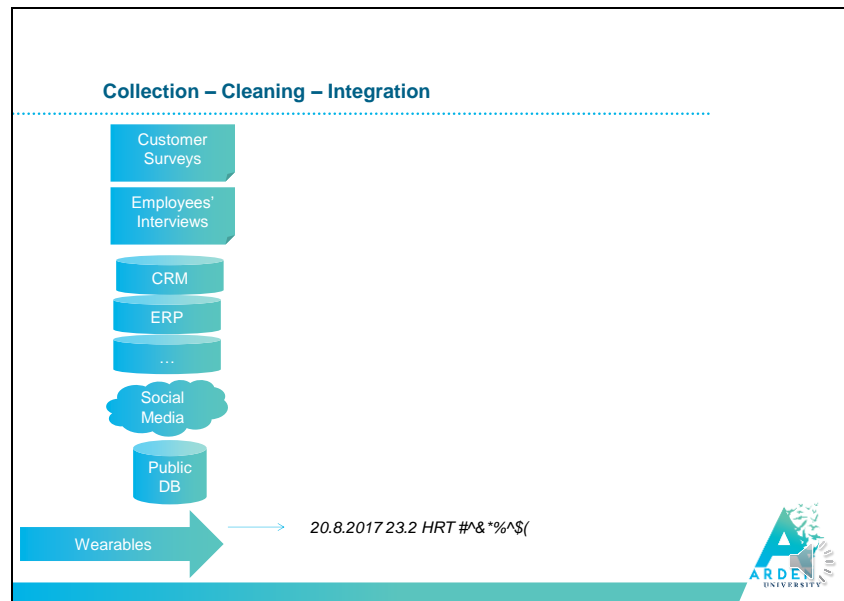
Slide 11



Records might be duplicated.



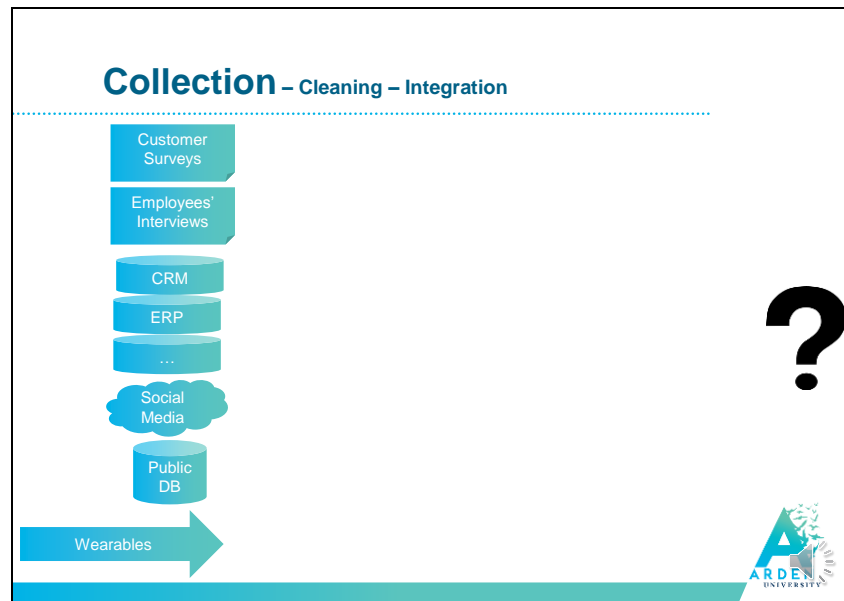
We might need to handle unstructured data, and to think how to extract structured data out of it. Here we should consider, for example, abbreviations, slang or domain specific terms.



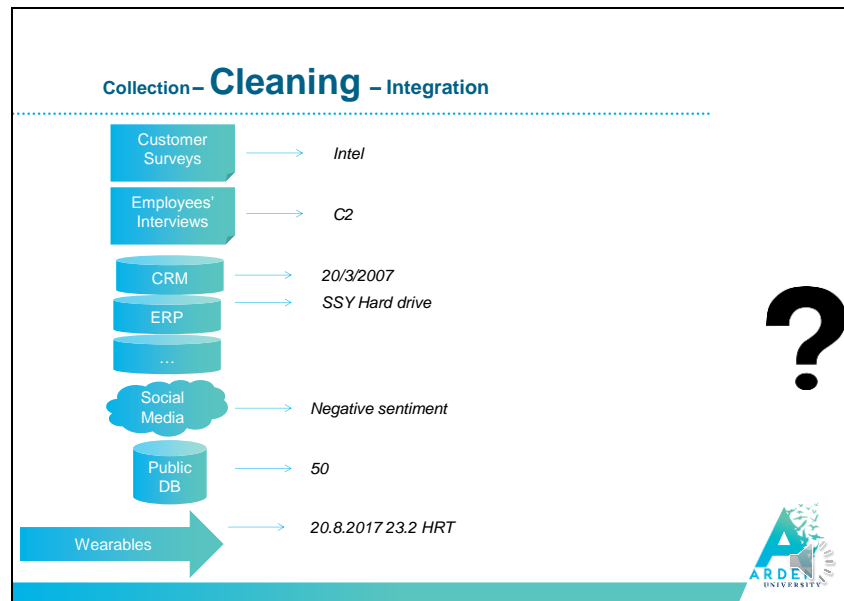
Also, data might be missing, incomplete or corrupted.

When data is missing, we might consider three main options:

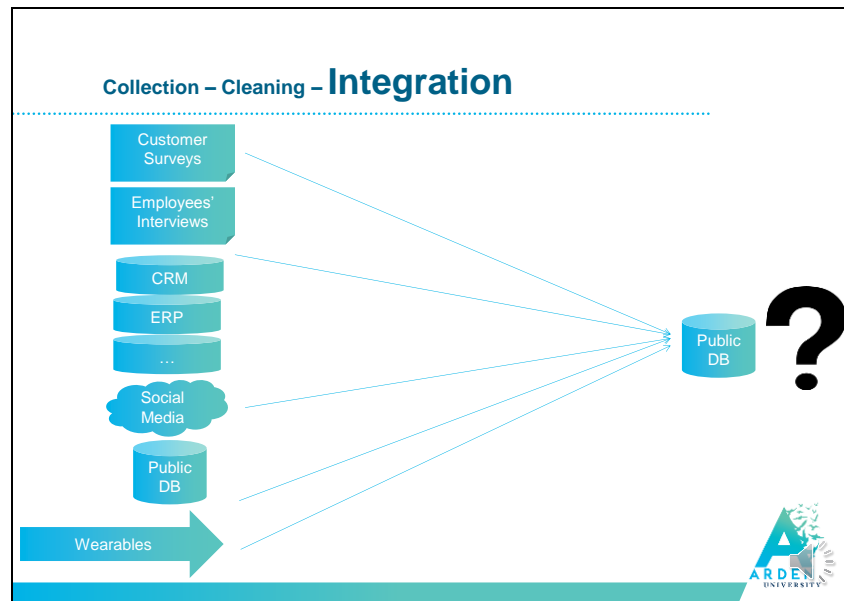
- to conduct the analysis without the affected variable
- to estimate its value from average, median. etc.
- to simply delete the specific records.



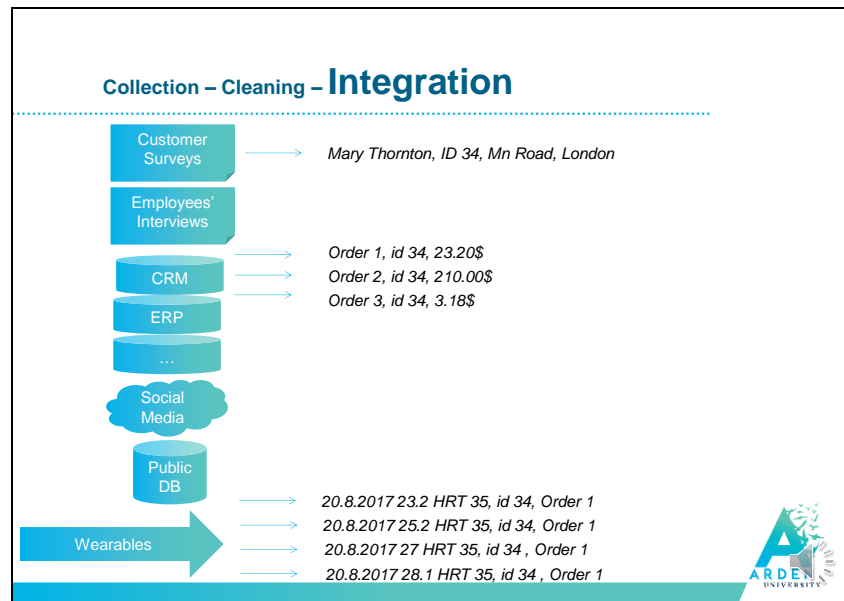
We not only have to understand the format and unit of analysis of our business question, but we also have to be very familiar with the data arriving from the data sources we wish to use for collection.



We will need to clean our data: to remove duplicates, to normalize units, to merge ambiguous values, to fix errors where possible, to extract structured features from unstructured text and to deal with missing values.

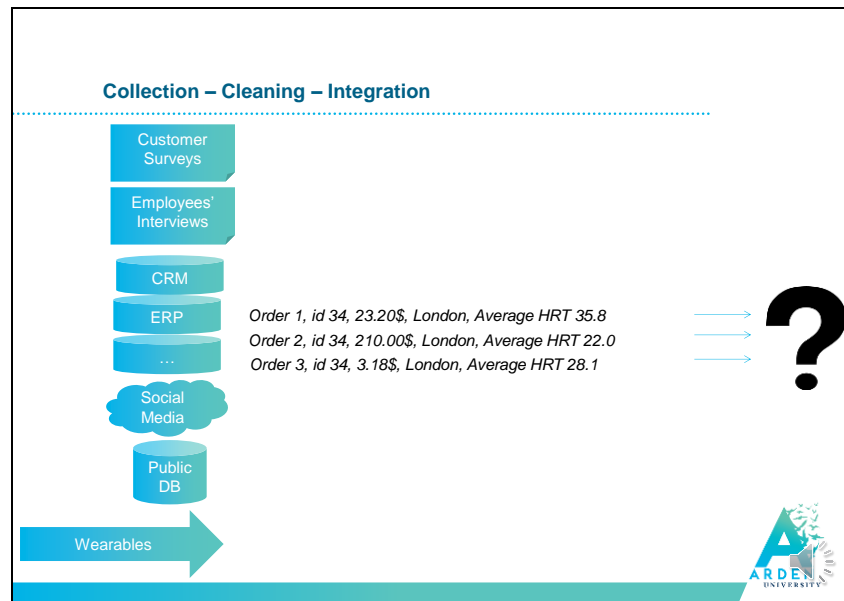


Then we will need to merge all of these into one dataset, with a single format, a single storage system, which speaks the language of our business question and is already prepared for analysis.



Different systems speak in different units of analysis.

For example, in the survey, we might have a record for each customer, while an organizational system might store a record for each of the customer's orders, while a sensor might send a timestamped record of the order each millisecond.




If we are asking about orders, we will need to add the customer-anonymized details to each order record, and aggregate higher grained data about these orders.

Collection – Cleaning – Integration

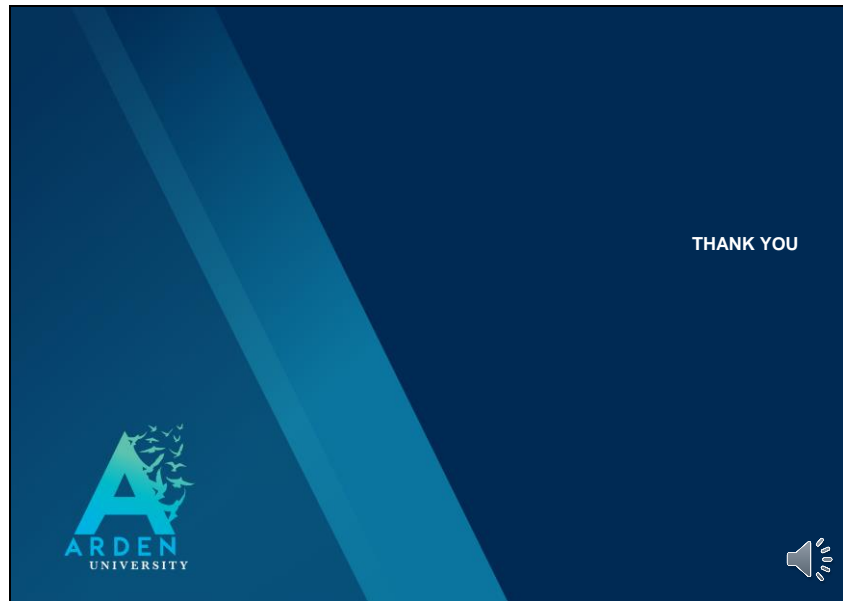
Order 1, id 34, 23.20\$, London, Average HRT 35.8
Order 2, id 34, 210.00\$, London, Average HRT 22.0
Order 3, id 34, 3.18\$, London, Average HRT 28.1

→ ?



At this point, we are no more worried about our data sources. Our data is collected, cleaned and integrated in the format and reliability level suitable for our analysis.

Slide 20



Thank you for listening.