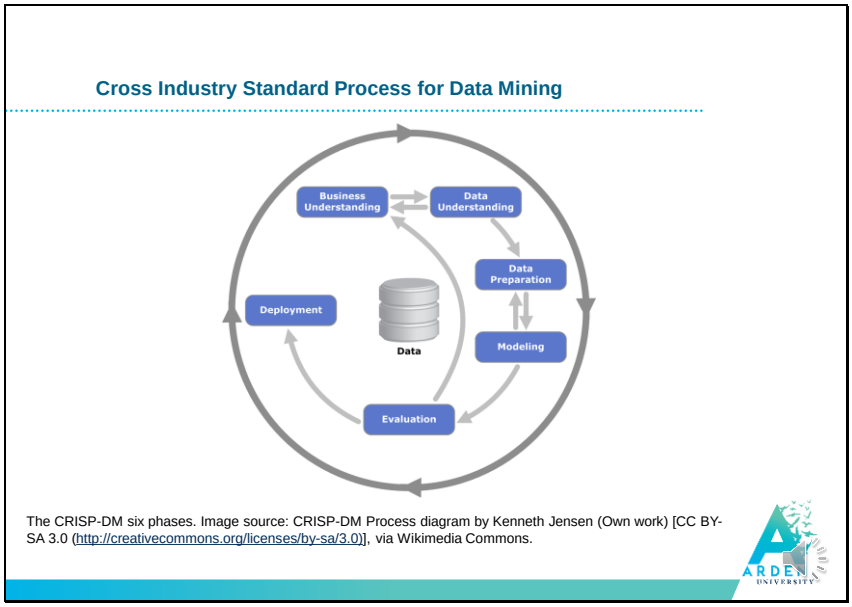
Slide 1

Slide 2



Various models have been developed with the aim of modelling a structured approach to a data analytics study or project.

In this module we will focus on CRISP-DM, which is the most commonly used model today. CRISP-DM stands for The Cross Industry Process for Data Mining. It shows the data mining (or analytics) process in a way which is totally independent on either the industry, tool or application.

CRISP-DM was initiated in late 1996 by three partners: Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR.

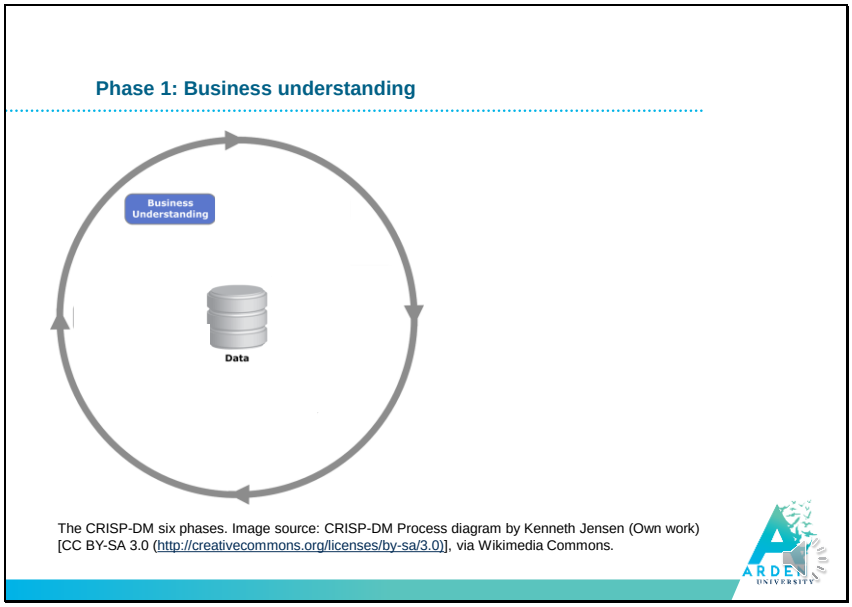
The step-by-step guide was released in 2000 (please see its references in the wider reading list of this lesson).

CRISP-DM shows a typical life cycle of a data analytics project, which is composed of six main phases.

Slide 3

Each one of the six phases is composed of a set of tasks and outputs, such as reports, presentations and processed datasets.

Slide 4



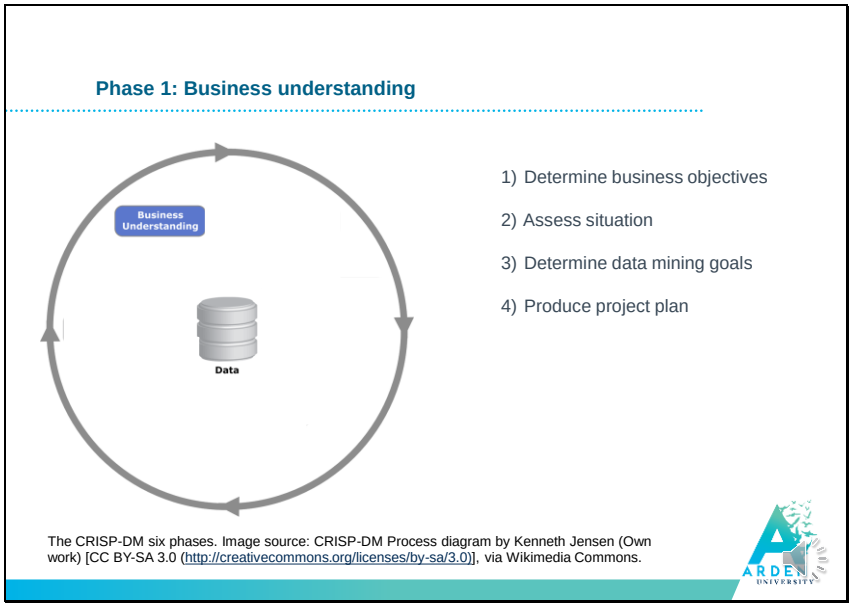
The first phase is the business understanding phase.

It is about figuring out what the business perspectives of the project are and deriving from these the project objectives and requirements.

The aim of this phase is to eventually translate the business questions into a data analytics problem definition, and to design a preliminary plan out of it.

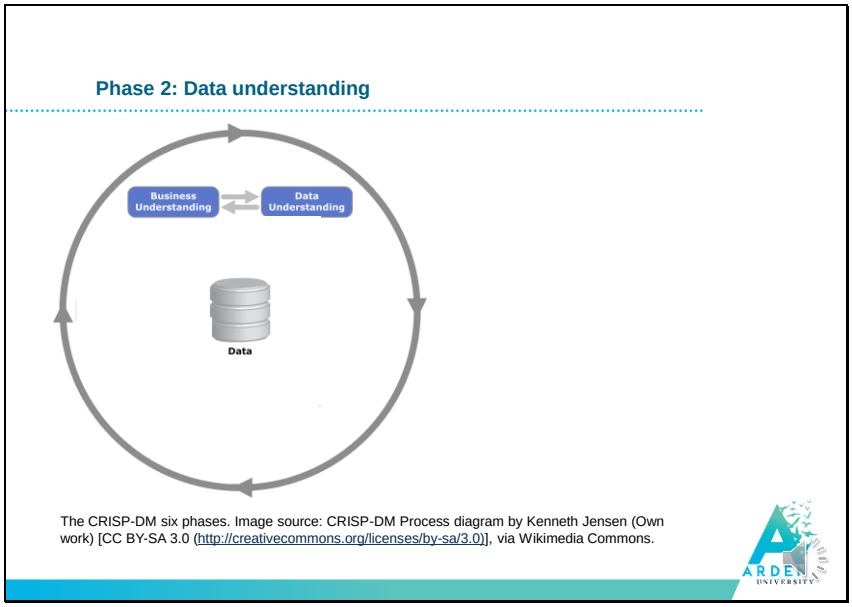
The Business Understanding phase includes four main tasks:

Slide 5



- Determine business objectives
- Assess situation
- Determine data mining goals
- Produce project plan

Slide 6



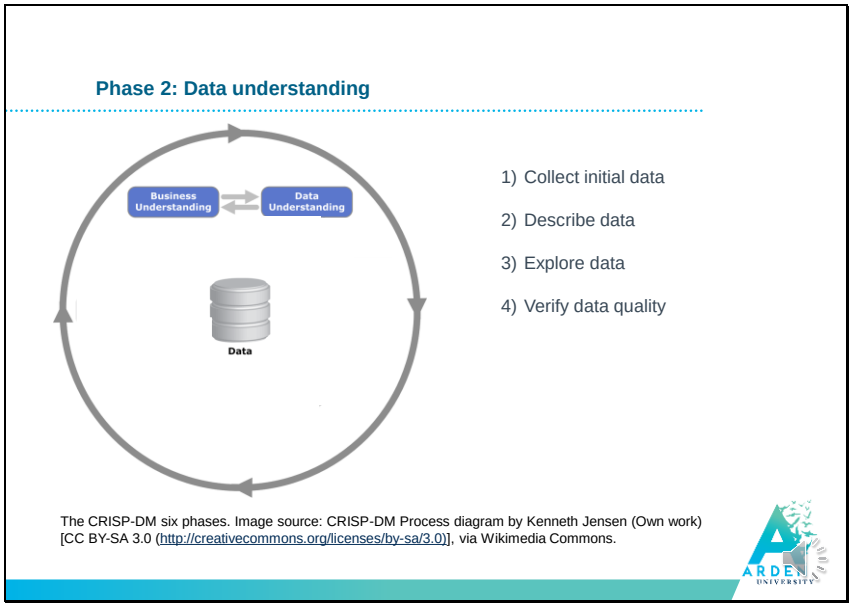
Introducing the inductive approach, as we discussed in the previous lesson, necessitates that we deeply understand the data we have.

The data understanding phase will typically happen almost in parallel with the business understanding phase, as they are highly dependent in one another.

This phase will include an initial data collection and a process of familiarization with the data. This is done in order to identify data quality issues and to assess the strengths and weaknesses of the data as early as possible in the process.
Often, we discover during this phase that the data needed to answer the business questions may not be available or easily accessible. In this case, we may need to go back to the business understanding phase and tweak and adapt it to the reality.

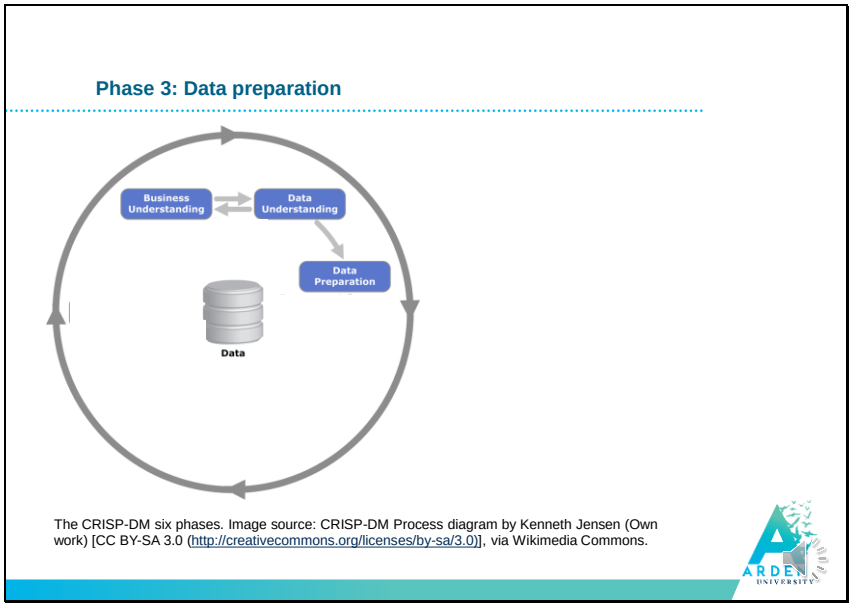
The data understanding phase includes four main tasks :

Slide 7



1) Collect initial data
2) Describe data
3) Explore data
4) Verify the data quality

Slide 8



Once we are familiar with the data, we can go on and prepare it for the modelling phase.

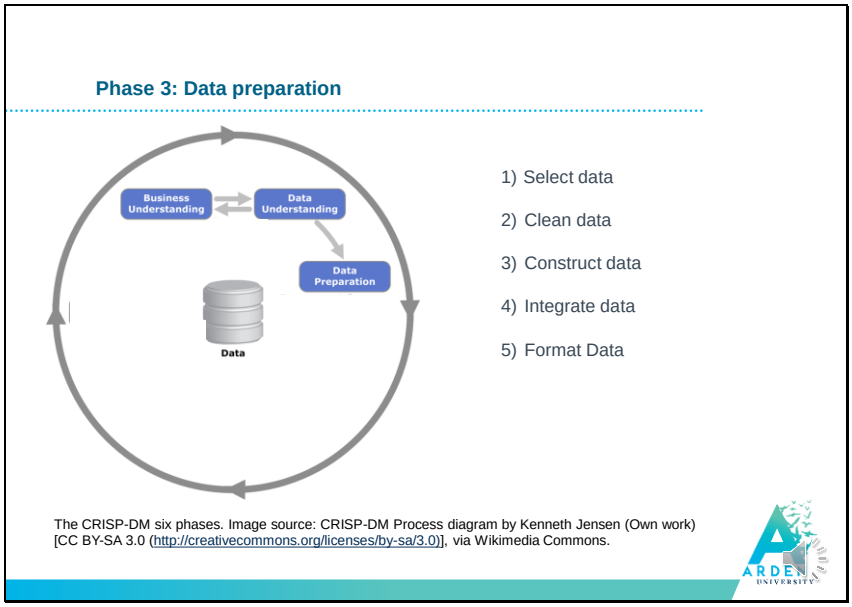
Usually, the format of the originally collected data does not fit the tabular format that most analysis tools can work with (as we discussed in the previous lesson).

In addition, real-world data is usually noisy and includes many missing values. This makes it hard or even impossible sometimes for the modelling techniques to work with.

This is the most time-consuming phase, and thus, the Data Handling module concentrates on it.

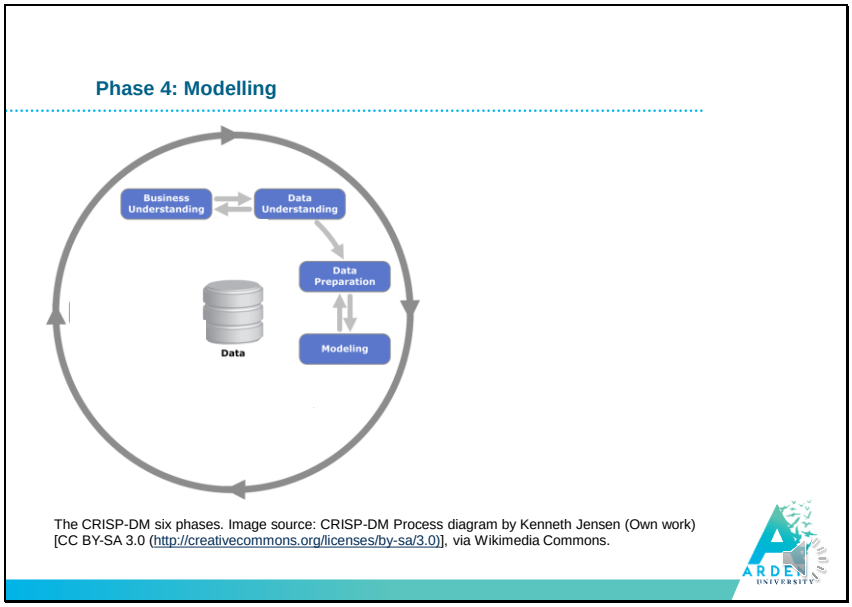
The data preparation phase includes five main tasks:

Slide 9



- Select data
- Clean data
- Construct data
- Integrate data
- Format Data

Slide 10

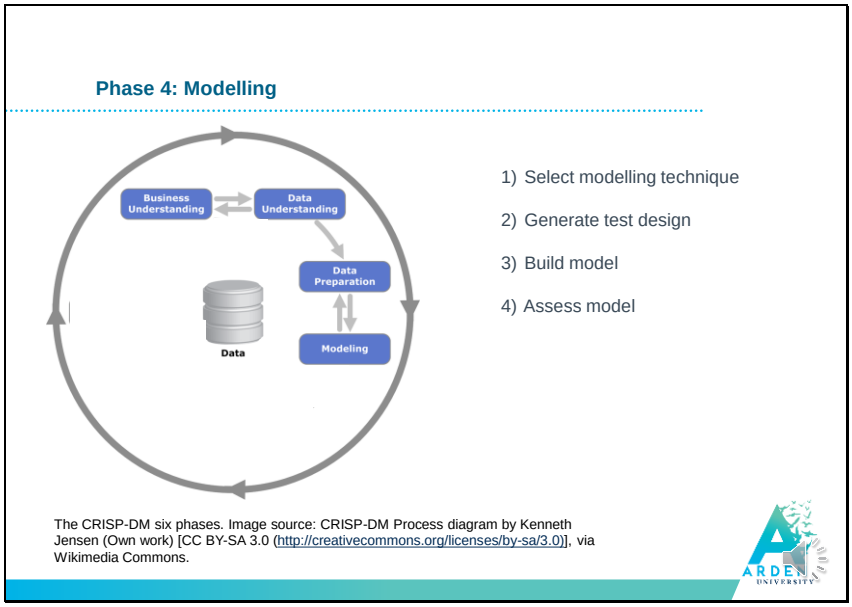


At this stage we feed our cleaned, formatted and integrated dataset into our pallet of modelling tools.

Various modelling techniques will be selected and applied to the data at this phase. Each aims at discovering some meaningful patterns or models in it.
Not all techniques work with all formats and types of data. This means that at this stage we sometimes need to go back to the data preparation phase and then back again to the modelling, until we are settled with the chosen modelling technique.

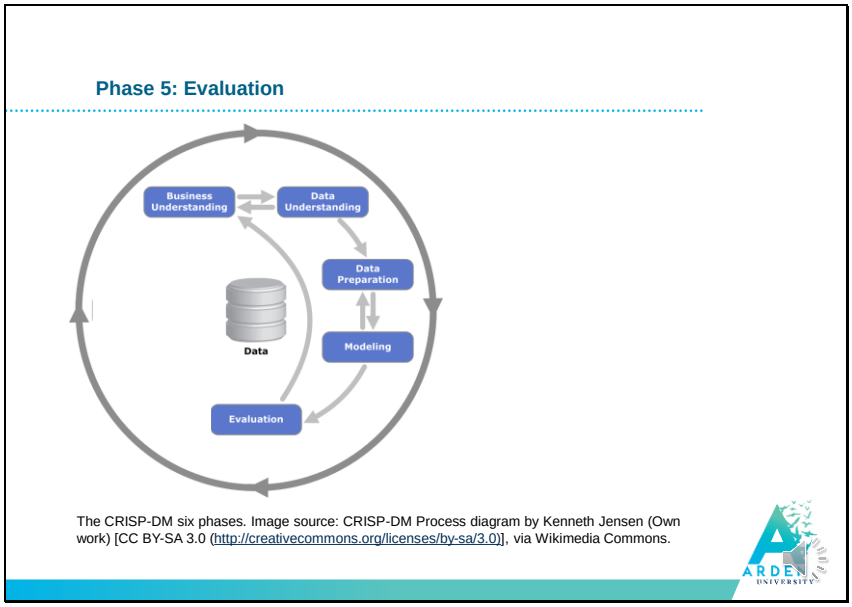
The Modelling phase includes four main tasks, which, like the other phases, will probably be applied more than once:

Slide 11



- Select the modelling technique
- Generate test design
- Build model
- Assess model

Slide 12



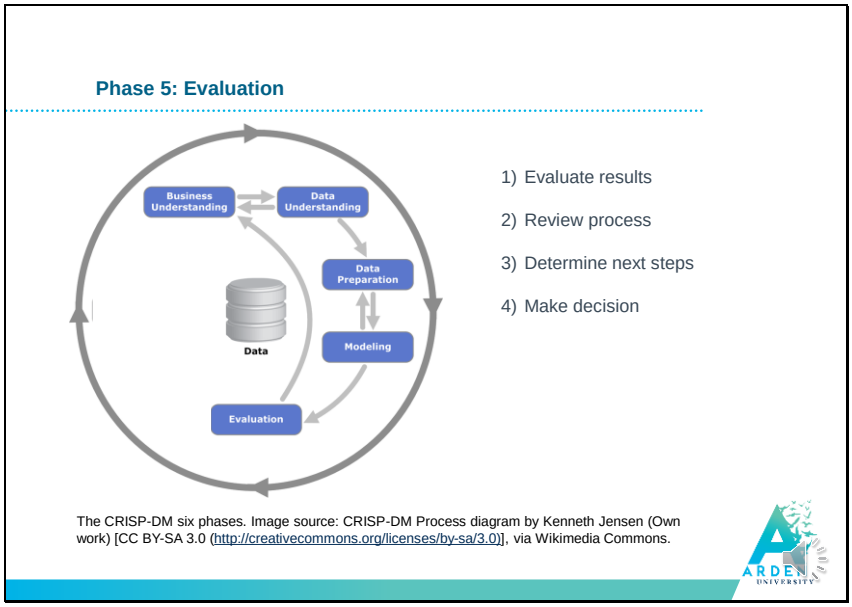
The fifth phase is about evaluating the model resulting from the modelling phase's outcomes.

Here we evaluate how accurate the results are, whether they can be generalized to new data, and whether they can be applied to the real world in the coming deployment phase.

One of the most important goals of this phase is to ask ourselves whether the model satisfies the business goals. This is really our last reasonable opportunity to go back to the first phase, and face our initial business understanding.

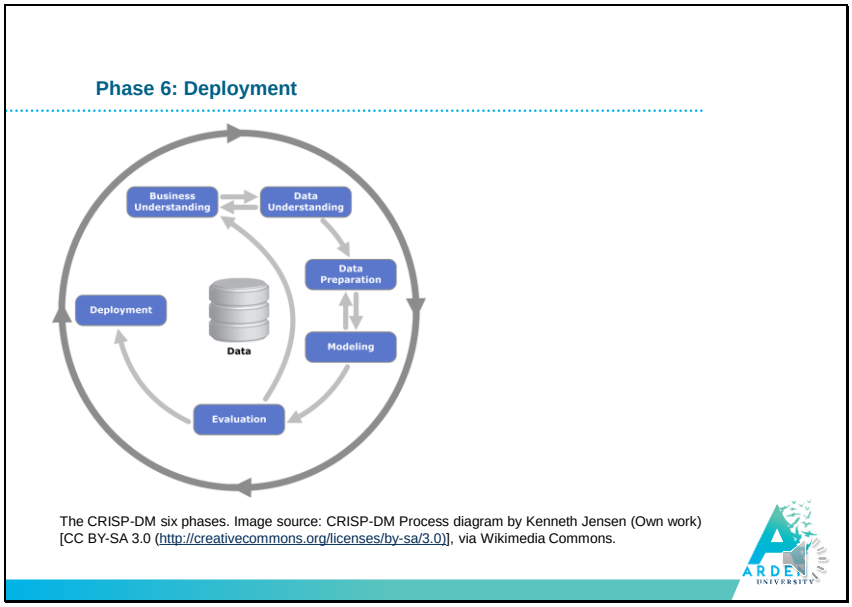
The evaluation phase includes four main tasks:

Slide 13



- Evaluate results
- Review process
- Determine next steps
- Make decision

Slide 14

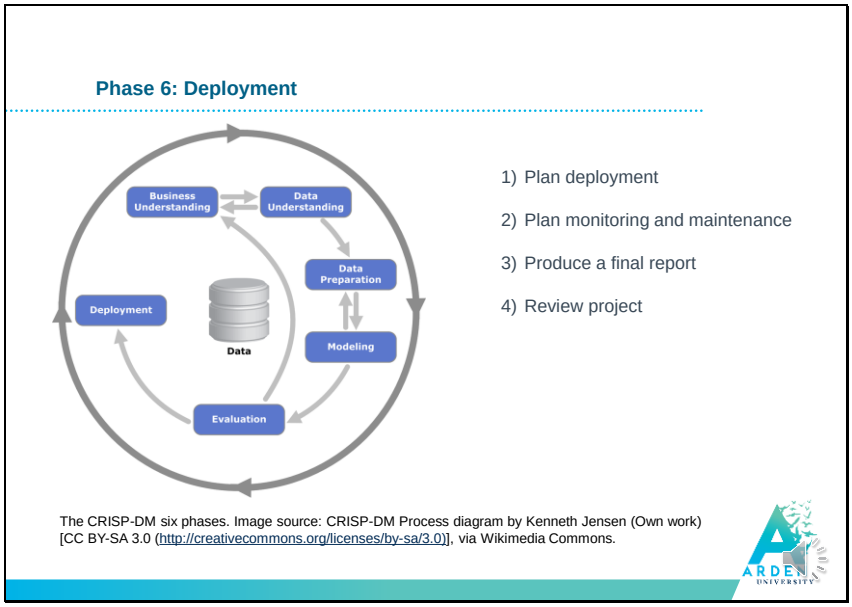


Deployment is the last phase of the data analytics project, although it actually marks the beginning of transitioning the project into a production environment, a transitioning which may well require many more changes and returns to initial phases.

At this phase, the model's results are put to real use for the first time. After having been trained and validated by existing datasets, and following an existing set of business assumptions, in this phase new data and new, many times, unexpected business scenarios will test it.

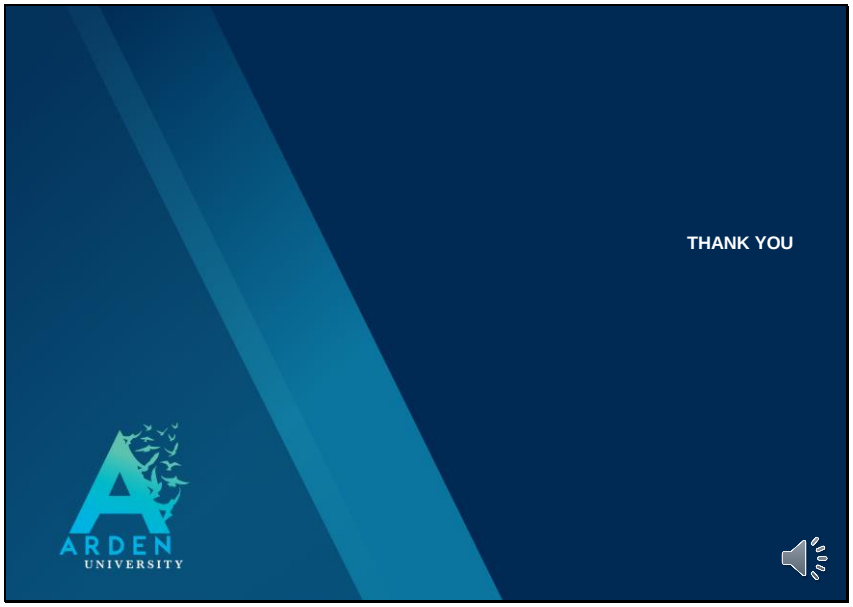
The deployment phase includes four main tasks:

Slide 15



- Plan deployment
- Plan monitoring and maintenance
- Produce a final report
- Review project

Slide 16



In this lesson we will briefly discuss each of these phases in order to be able to have a sufficient overlook of data analytics projects.

Thank you.