CS4650 Topic 7:
Data Analytics Lifecycle

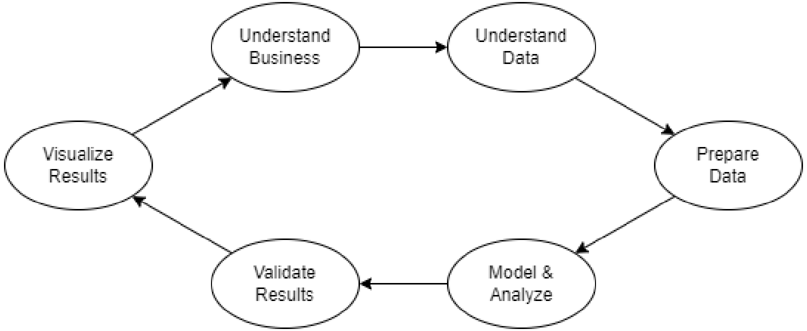
Data Analytics Lifecycle

- A pile of random data is basically useless.
- However, by carefully gathering, cleaning, and managing the data, then performing analytics and modeling the data, useful information can be gleaned.
- Over the years, the *process* of collecting, managing, and mining the data has been studied, resulting in the *Data Analytics Lifecycle*.
- This lifecycle is a series of steps that are performed, however:
    - There is not perfect agreement on the number, names, or descriptions of the steps.
    - The steps are not necessarily followed in order, and there are many iterations that can occur.
- The lifecycle gives us a framework for understanding the methodology.

Data Analytics Lifecycle

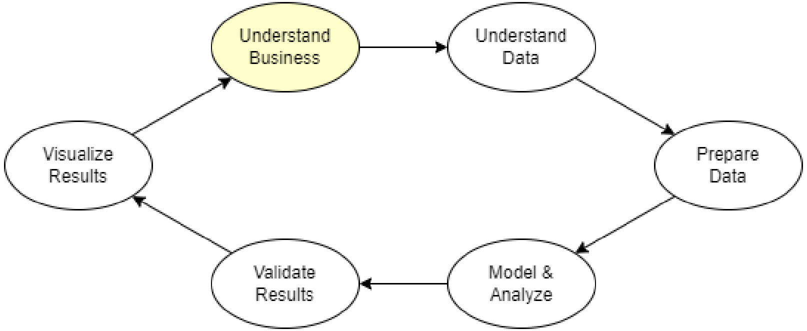
- The Data Analytics Lifecycle helps you to manage a data analytics project.
- This will assist in efficiency and help to get the best results for the client.

Data Analytics Lifecycle



- The Lifecycle is represented in a circular form, but in practice the flow is not necessarily so regular:
    - As more is learned in the process, it might make sense to return to an earlier step and refine the operations there.
    - In some cases, a step in this cycle might be skipped.
    - Sometimes additional steps are performed.
- Consequently, this is a *guideline*.

1. Understand the Business Issues

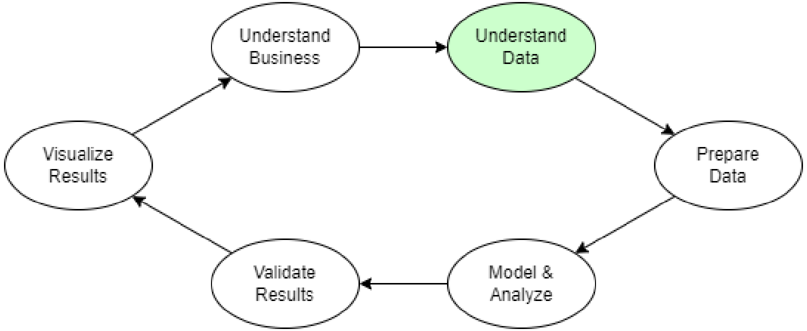


- This step identifies what you are trying to uncover from the data.
  - What are the expectations?
  - What are the objectives?
  - What type of analysis is being sought?
  - What are the deliverables?
  - What is the purpose?

1. Understand the Business Issues

- In this stage, you might even formulate an initial hypothesis to test.
- Data learning might also begin in this phase.
- Without a clear understanding of these, it may be difficult to proceed: Where do you begin, and where to you end, and have you succeeded?
- This phase is focused on the business requirements.

2. Understand the Data Set

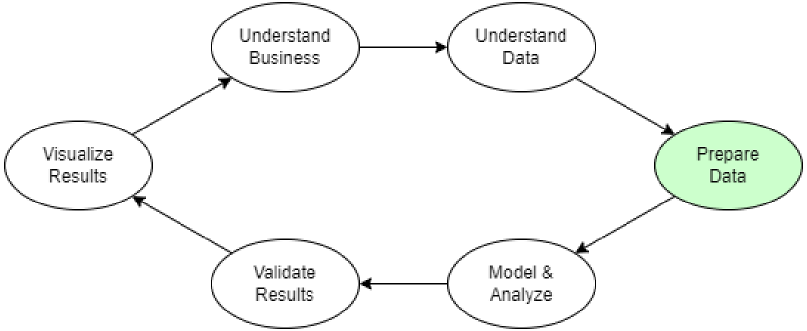


- Gather the data set:
  - Data Entry, either through manual techniques or automated systems within the organization.
  - Data Acquisition, from external sources
  - Signal Reception, capturing data from digital devices.
- Identify the values within the data records:
  - What is necessary information
  - What is useful optional information
  - What is irrelevant information

2. Understand the Data Set

- See what values might need cleaning:
  - What might be missing, and what to do if it is
  - What data doesn't make sense
  - What data might be duplicated
  - What might have spelling errors
  - What information might be necessary or optional, but is in an inconvenient format
- This phase focuses on the information requirements

3. Preparing the Data

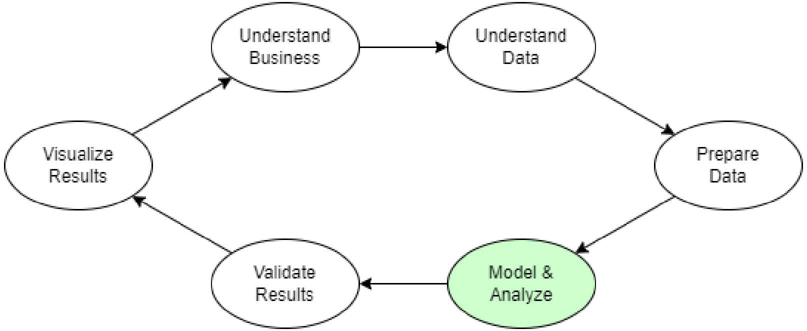


- In this step, the data is cleaned.
- If values are missing, enter a valid substitute
  - Maybe discard the data
  - Maybe enter an average data score for this value.
    - Care must be taken not to skew the data or influence the output

3. Preparing the Data

- Convert values to a more convenient representation
- Perhaps for each value give an acceptable range, then scan the data (by program) to find records with out-of-bounds values.
- This phase results in developing a model.

4. Perform Exploratory Analysis and Modeling

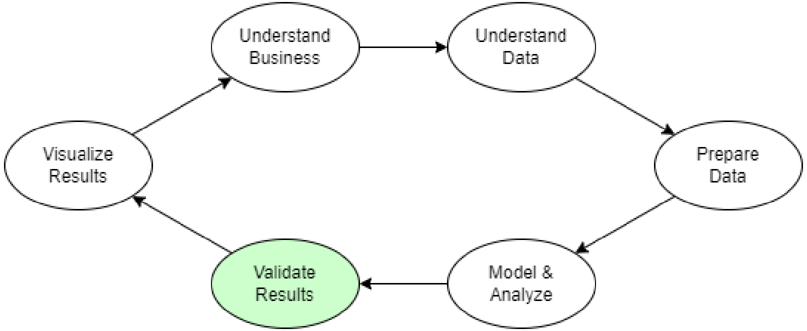


- The data may be divided into test, training and production datasets.
- The models planned in the previous step are built and refined, then used to test the data

4. Perform Exploratory Analysis and Modeling

- Seek the answers to the objectives
- Determine the best statistical modeling method for the data/objective.
    - This may be done by trying various approaches to see which produces in the most useful results.

5. Validate Your Data

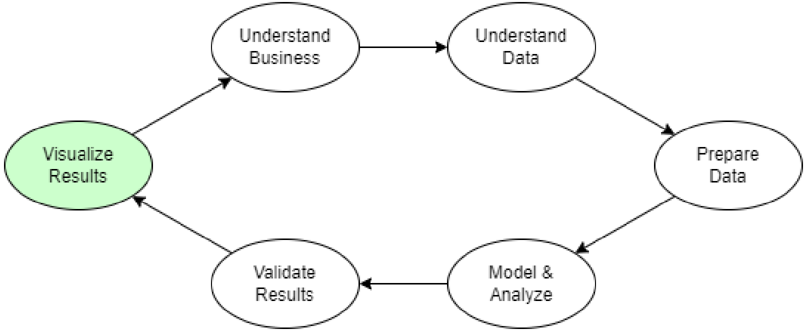


- Assess the data and the results: do we have the correct information for the deliverable?
- Did the models work properly?
- Does the data need more cleaning?

5. Validate Your Data

- Did you find the answers to the questions being asked?
- Iterate to earlier stages if necessary
- Develop a summarized narrative of the results.

6. Visualize and Present Findings



- Once you have the answers to the questions being sought, begin the data visualization.
- Find a visualization format that highlights the answer to the question, while reducing any irrelevant information that would obscure the answer.

6. Visualize and Present Findings

- Remember not to *spin* the results, to fudge the results to highlight your desired or preconceived ideas and to minimize indications that the results were not what you expected.

Usefulness of the Data Analytics Lifecycle

- The Data Analytics Lifecycle is a roadmap to help guide the process of performing an analysis.
- Being able to organize the project will increase the efficiency of the process, and will help to minimize errors or distractions.
- We will be using the Data Analytics Lifecycle throughout this class.