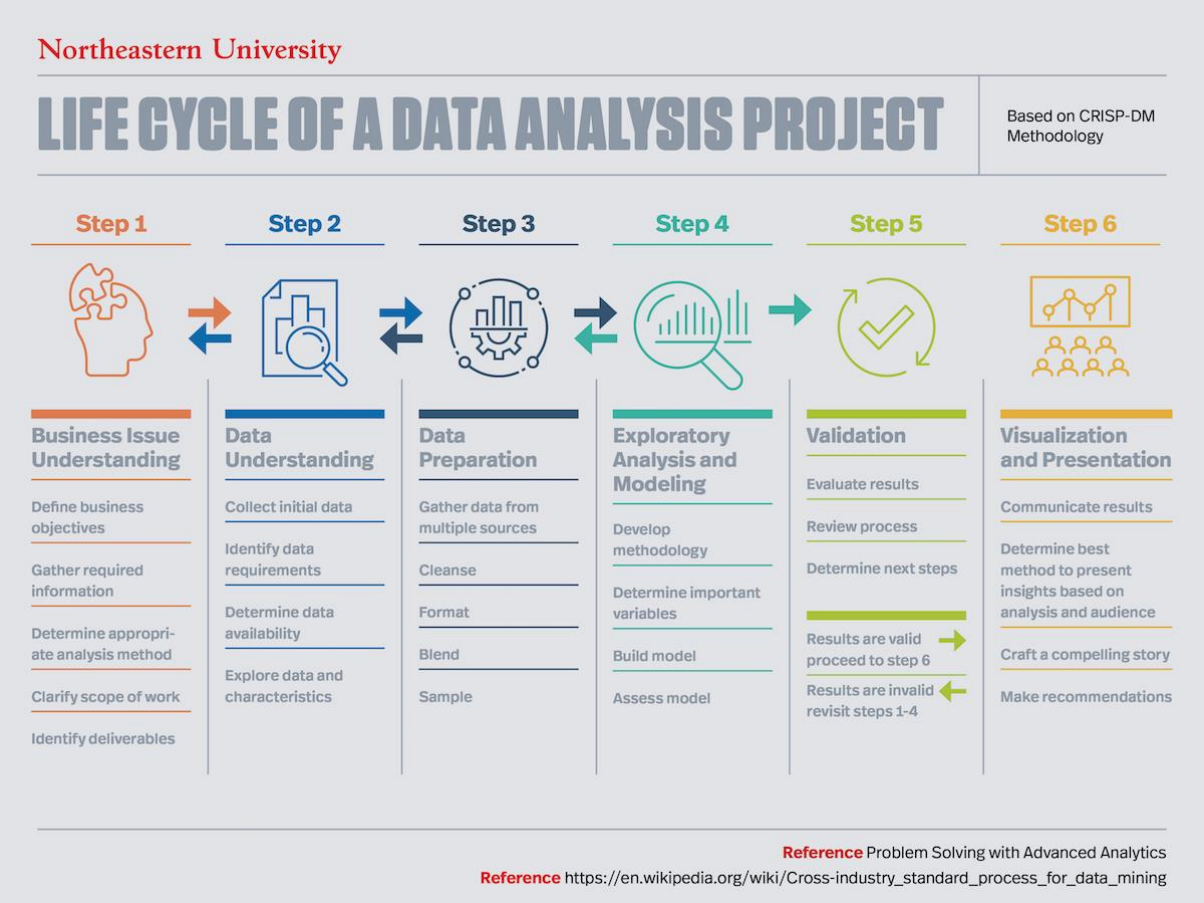
EDA
Elit Data Analytics
(data To Wisdom)

LIFECYCLE OF A DATA ANALYTICS PROJECT

What is the Data Analytics Lifecycle?

The data analytics lifecycle describes the process of conducting a data analytics project, which consists of six key steps based on the CRISP-DM methodology. According to Paula Muñoz, a Northeastern alumna, these steps include: understanding the business issue, understanding the data set, preparing the data, exploratory analysis, validation, and visualization and presentation.

Founded in 1898, **Northeastern** is a global, experiential, research university built on a tradition of engagement with the world



The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) is a process model that serves as the base for a data science process. It has six sequential phases:

1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

Published in 1999 to standardize data mining processes across industries, it has since become the most common methodology for data mining, analytics, and data science projects.

Exploring the 6 CRISP-DM Phases

I. Business Understanding

Any good project starts with a deep understanding of the customer's needs. Data mining projects are no exception and CRISP-DM recognizes this.

The *Business Understanding* phase focuses on understanding the objectives and requirements of the project. Aside from the third task, the three other tasks in this phase are foundational project management activities that are universal to most projects:

1. **Determine business objectives:** You should first "thoroughly understand, from a business perspective, what the customer really wants to accomplish." (CRISP-DM Guide) and then define business success criteria.
2. **Assess situation:** Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.
3. **Determine data mining goals:** In addition to defining the business objectives, you should also define what success looks like from a technical data mining perspective.
4. **Produce project plan:** Select technologies and tools and define detailed plans for each project phase.

While many teams hurry through this phase, establishing a strong business understanding is like building the foundation of a house – absolutely essential.

II. Data Understanding

Next is the *Data Understanding* phase. Adding to the foundation of *Business Understanding*, it drives the focus to identify, collect, and analyze the data sets that can help you accomplish the project goals. This phase also has four tasks:

1. **Collect initial data:** Acquire the necessary data and (if necessary) load it into your analysis tool.
2. **Describe data:** Examine the data and document its surface properties like data format, number of records, or field identities.
3. **Explore data:** Dig deeper into the data. Query it, visualize it, and identify relationships among the data.
4. **Verify data quality:** How clean/dirty is the data? Document any quality issues.

III. Data Preparation

A common rule of thumb is that 80% of the project is data preparation.

This phase, which is often referred to as “data munging”, prepares the final data set(s) for modeling. It has five tasks:

1. Select data: Determine which data sets will be used and document reasons for inclusion/exclusion.
2. Clean data: Often this is the lengthiest task. Without it, you’ll likely fall victim to garbage-in, garbage-out. A common practice during this task is to correct, impute, or remove erroneous values.
3. Construct data: Derive new attributes that will be helpful. For example, derive someone’s body mass index from height and weight fields.
4. Integrate data: Create new data sets by combining data from multiple sources.
5. Format data: Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations.

IV. Modeling

What is widely regarded as data science’s most exciting work is also often the shortest phase of the project.

Here you’ll likely build and assess various models based on several different modeling techniques. This phase has four tasks:

1. Select modeling techniques: Determine which algorithms to try (e.g. regression, neural net).
2. Generate test design: Pending your modeling approach, you might need to split the data into training, test, and validation sets.
3. Build model: As glamorous as this might sound, this might just be executing a few lines of code like “reg = LinearRegression().fit(X, y)”.
4. Assess model: Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design.

Although the CRISP-DM Guide suggests to “iterate model building and assessment until you strongly believe that you have found the best model(s)”, in practice teams should continue iterating until they find a “good enough” model, proceed through the CRISP-DM lifecycle, then further improve the model in future iterations.

V. Evaluation

Whereas the A*ssess Model* task of the *Modeling* phase focuses on technical model assessment, the *Evaluation* phase looks more broadly at which model best meets the business and what to do next. This phase has three tasks:

1. Evaluate results: Do the models meet the business success criteria? Which one(s) should we approve for the business?
2. Review process: Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed.
3. Determine next steps: Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

VI. Deployment

"Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise."

–CRISP-DM Guide

A model is not particularly useful unless the customer can access its results. The complexity of this phase varies widely. This final phase has four tasks:

1. **Plan deployment:** Develop and document a plan for deploying the model.
2. **Plan monitoring and maintenance:** Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase (or post-project phase) of a model.
3. **Produce final report:** The project team documents a summary of the project which might include a final presentation of data mining results.
4. **Review project:** Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.

Your organization's work might not end there. As a project framework, CRISP-DM does not outline what to do after the project (also known as "operations"). But if the model is going to production, be sure you maintain the model in production. Constant monitoring and occasional model tuning is often required.