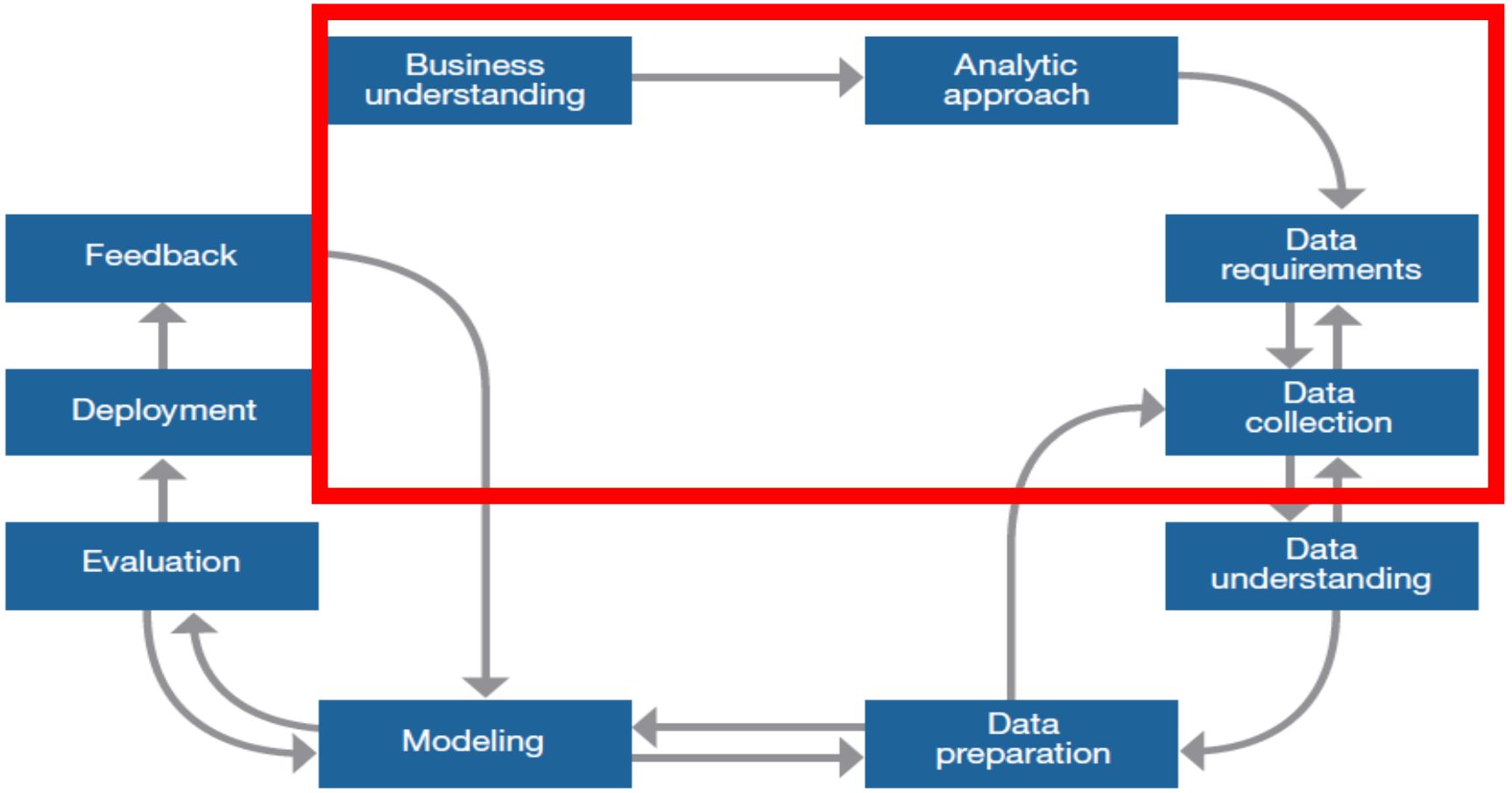
Business Understanding and Data Collection

PERTEMUAN XII

Foundational Methodology for Data Science



Business Understanding

Every project starts with business understanding. We must keep in mind that a Data Science project is **definitely a Business project**, so it must always be oriented on achieving results focused on the business and have a global vision aligned with the business strategy.

The business sponsors who need the analytic solution play the most critical role in this stage

Define the problem, project objectives, and solution requirements from business perspective

What is the goal? How do you define “success” and how can you measure it?

Business Understanding

Business Understanding example :

Traffic Problem: Traffic congestion wastes time and money

Clear question: How can we optimize traffic light duration using data on traffic patterns, weather, and pedestrian traffic?

Measurable outcomes:

- % decrease in commute time
- % decrease in length/duration of traffic jams

Analytic Approach

Once the business problem has been clearly stated, the data scientist can define the analytic approach to solving the problem

This stage entails expressing the problem in the context of statistical and machine-learning techniques, so the organization can identify the most suitable ones for the desired outcome

In brief, analytic approach is how to express problem in context of statistical and machine learning techniques

Analytic Approach

"Predicting revenue in the next quarter?" → **Regression**

"Does this patient have cancer A, cancer B, or are they healthy?" → **Classification**

"Are there groups of users that seem to be have similarly to each other?" → **Clustering**

"How can I target discounts to specific customers?" → **Recommendation/Personalization**

Data Requirements

The chosen analytic approach determines the data requirements.

Specifically, the analytic methods to be used require certain **data content**, **formats** and **representations**, guided by domain knowledge.

Data Collection

In the initial data collection stage, data scientists identify and gather the available data resources—structured, unstructured and semi-structured—relevant to the problem domain.

Available data? Obtain data? Revise data requirements or collect more data?

Gathering Data

Data could be gathered through several sources, such as:

1. Internal company data (excel, internal databases, etc)
2. Web API’s, Web scraping
3. Dataset via public data
4. Dataset via open data

Open Data

Open Data is defined as structured data that is machine-readable, freely shared, used and built on without restrictions - Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness.

Open Data

The Open Definition provides a more detailed definition of Open Data. To summarize the most important points:

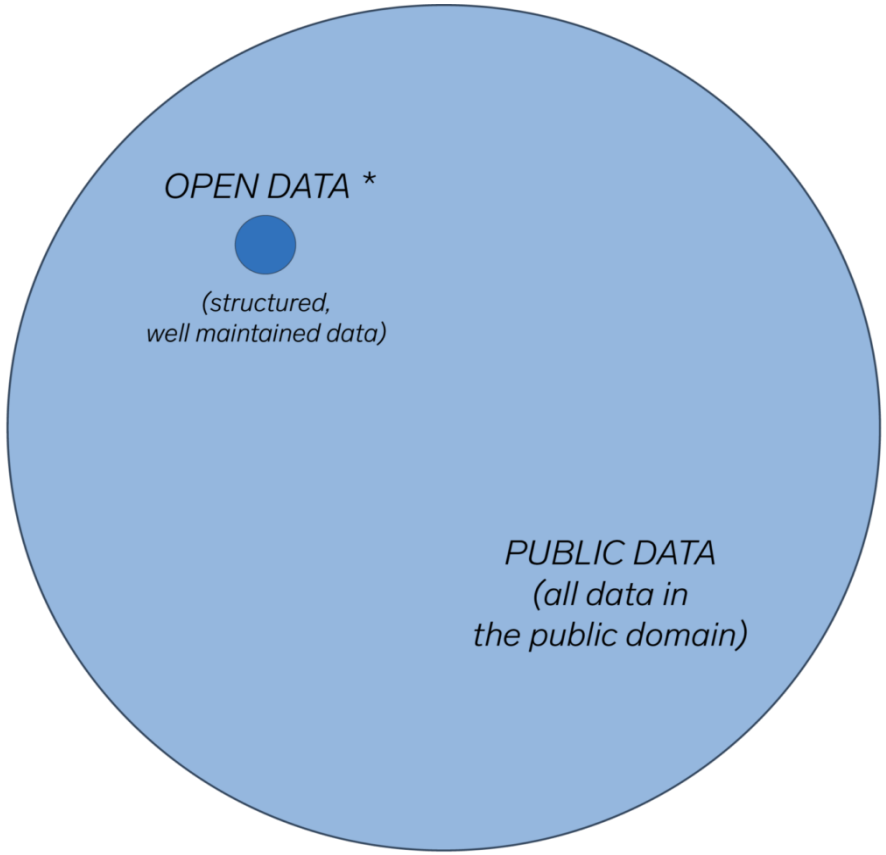
Availability and Access: the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.

Re-use and Redistribution: the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.

Universal Participation: everyone must be able to use, re-use and redistribute. There should be no discrimination against fields of endeavor or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

Public Data vs Open Data

Open data and content can be freely used, modified, and shared by anyone and for any purpose. Meanwhile, Public data can be defined as all information in the public domain that are only accessible via requests (less accessible).



** According to the Open Data Barometer's Global Report 2017, only **7%** of key datasets across 115 countries were considered open. The open data circle size is **7%** of data otherwise considered public.*

Open Data Sources

There are several free Open Data sources anyone can use, such as:

1. World Bank Open Data https://data.worldbank.org/
2. Kaggle https://www.kaggle.com/datasets
3. UNICEF Dataset https://data.unicef.org/
4. WHO Open Data https://www.who.int/gho/database/en/
5. IBM Data Asset eXchange (DAX) https://developer.ibm.com/exchanges/data/

References

Foundational Methodology for Data Science – IBM Analytics White Paper

The Data Science Process by Polong Lin - https://www-01.ibm.com/events/wwe/grp/grp304.nsf/vLookupPDFs/Polong%20Lin%20Presentation/$file/Polong%20Lin%20Presentation.pdf

https://cognitiveclass.ai/courses/data-science-with-open-data

https://medium.com/enigma/what-is-public-data-938e086f363f

https://developer.ibm.com/blogs/ibm-data-asset-exchange-dax-free-open-data-ai/

https://www.techedgegroup.com/blog/data-science-process-problem-statement-definition

https://www.kaggle.com/c/prudential-life-insurance-assessment/overview

https://www.kaggle.com/c/restaurant-revenue-prediction/overview

https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/overview