

One cannot think about Artificial Intelligence without thinking about data, as data is an essential part of AI. In order for an AI algorithm to output any prediction, it has to be fed with large volumes of data. Apart from its use in predictive analytics, data has become a key input driving growth, enabling businesses to extract valuable insights and improve the decision-making process.

Data as a general concept, refers to the fact that some existing knowledge of information is represented or coded in some form suitable for valuable usage or processing. In this article, we explain the different types of data and data sources companies can leverage to implement Artificial Intelligence and improve the decision-making process.

PRIMARY AND SECONDARY SOURCES OF DATA

To analyze, present, and interpret information from the data, there has to be a process of gathering and sorting the data. There are different methods to gather data, all of which fall into two categories: **primary data source** and a **secondary data source**.

The term primary data refers to the data originated by a researcher himself, while secondary data is the already existing data collected by agencies and organizations for the purpose of conducting an analysis. Primary data sources can include surveys, observations, questionnaires, experiments, personal interviews, and more. The data from ERP (Enterprise Resource Planning) and CRM (Customer Relationship Management) systems

can also be used as a primary source of data. On the contrary, secondary data sources can be government publications, staging websites, publications from independent research labs, journal articles, etc. The transformed "raw" data set into another format, in the process of data wrangling, can also be seen as a secondary data source. Secondary data can be a key concept in terms of data enrichment when the primary source data is not solid enough with information, and it can improve the precision of the analysis by adding more attributes and variables to the sampling.

QUANTITATIVE AND QUALITATIVE DATA

Data can be defined by a set of variables with qualitative or quantitative nature.

Qualitative Data

Qualitative data refers to the data that can provide insights and understanding about a particular problem.

Quantitative Data

Quantitative data, as the name suggests is one that deals with quantity or numbers. This numerical data can be determined by categories or so-called classes.

Although both types of data can be considered as separate entities providing different outcomes and information about a sample, it is important to understand that both types are often needed to perform quality

analysis. Without knowing why are we seeing a certain pattern in behavioral data, we may try to solve the wrong problem, or the right problem incorrectly. A real-life example would be collecting qualitative data about customer preferences, and quantitative data about the number and the age of customers in order to analyze the level of customer satisfaction and find a pattern or correlation of changing preferences with different customer age groups.

TYPES OF DATA SOURCES

Data can be captured in many different shapes, some may be easier to extract than others. Having data in different shapes requires different storage solutions and should therefore be approached in different ways. At Kantify, we distinguish between three shapes of data: **structured data**, **unstructured data**, and **semi-structured data**.

Structured Data

Structured data is tabular data, containing columns and rows which are very well defined. The main advantage of this type of data is being easily stored, entered, queried, modified, and analyzed. Structured data is often managed by Structured Query Language, or SQL - a programming language created for managing and querying data in relational management systems.

Unstructured Data

Unstructured data is the rawest form of any data, and it can be in any type or file: pictures and graphic images, webpages, PDF files, videos, emails, word processing documents, etc. This data is often stored in repositories of files. Extracting valuable information out of this type of data can be somewhat challenging. For example, a text can be analyzed by extracting the topics it covers and whether the text is positive or negative about them.

Semi-structured Data

As the name implies, semi-structured data is a cross between structured and unstructured data. A semi-structured data may have a consistent defined format, however, the structure may not be very strict. The structure may not be necessarily tabular and parts of the data may be incomplete, or contain differing types. An example can be photos of other graphics tagged with keywords, making it easy to organize and locate graphics.

HISTORICAL AND REAL-TIME DATA

Historical datasets can help in answering exactly the types of questions that decision-makers would like to benchmark against real-time data. Historical data sources can be best suited for building or modifying predictive or prescriptive models, and offering insights that can improve long-term and strategic decision making. The basic definition of real-time data explains it as a data that is passed along the end-user as quickly as it is gathered. Real-time data can be enormously valuable in things like traffic GPS systems, in bench-marking different kinds of

analytics projects and for keeping people informed through instant data delivery.

In predictive analytics, both types of data sources should be given equal consideration, as both can help in predicting and identifying future trends.

INTERNAL AND EXTERNAL DATA

Internal data

Internal data is information gathered within an organization and can cover areas such as personnel, operations, finance, maintenance, procurement, and many more. Internal data can provide information on employee turnover, sales success, profit margins, structure and dynamics of an organization, etc.

External data

External data is the information gathered from outside, including customers, staging websites, agencies, and more. For example, external data gathered from social media can provide insights about the behavior, preferences, and motivations of customers. At this stage, you may wonder if internal data is the same as primary data, and external data the same as secondary data. This is close but slightly different. The categorization of internal and external data sources is mostly in terms of where the data comes from - whether it was collected from your organization or from a source outside your organization. The notion of primary/secondary data rather refers to the purpose and time-frame for which the data was collected -

whether it was collected by the researcher for a precise project, or in the form of another source, even within the same organization.

These different types of data sets can be found within an organization, but they can also be found in external data sources like the internet. We help companies make smarter and data-driven decisions with the help of Artificial Intelligence and Machine Learning. If you are curious to find out how your organization can leverage data to boost growth, don't hesitate to contact us, and one of our team members will be back at you shortly with more insights!