Title: Data exploration

URL: https://en.wikipedia.org/wiki/Data_exploration

PageID: 43385931

Categories: Category:Data analysis, Category:Data management, Category:Data quality, Category:Machine learning

-----

Data exploration is an approach similar to initial data analysis , whereby a data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data, rather than through traditional data management systems. These characteristics can include size or amount of data, completeness of the data, correctness of the data, possible relationships amongst data elements or files/tables in the data.

Data exploration is typically conducted using a combination of automated and manual activities. Automated activities can include data profiling or data visualization or tabular reports to give the analyst an initial view into the data and an understanding of key characteristics.

This is often followed by manual drill-down or filtering of the data to identify anomalies or patterns identified through the automated actions. Data exploration can also require manual scripting and queries into the data (e.g. using languages such as SQL or R) or using spreadsheets or similar tools to view the raw data .

All of these activities are aimed at creating a mental model and understanding of the data in the mind of the analyst, and defining basic metadata (statistics, structure, relationships) for the data set that can be used in further analysis.

Once this initial understanding of the data is had, the data can be pruned or refined by removing unusable parts of the data (data cleansing), correcting poorly formatted elements and defining relevant relationships across datasets. This process is also known as determining data quality .

Data exploration can also refer to the ad hoc querying or visualization of data to identify potential relationships or insights that may be hidden in the data and does not require to formulate assumptions beforehand.

Traditionally, this had been a key area of focus for statisticians, with John Tukey being a key evangelist in the field. Today, data exploration is more widespread and is the focus of data analysts and data scientists ; the latter being a relatively new role within enterprises and larger organizations.

Interactive Data Exploration

This area of data exploration has become an area of interest in the field of machine learning . This is a relatively new field and is still evolving. As its most basic level, a machine-learning algorithm can be fed a data set and can be used to identify whether a hypothesis is true based on the dataset. Common machine learning algorithms can focus on identifying specific patterns in the data. Many common patterns include regression and classification or clustering , but there are many possible patterns and algorithms that can be applied to data via machine learning.

By employing machine learning, it is possible to find patterns or relationships in the data that would be difficult or impossible to find via manual inspection, trial and error or traditional exploration techniques.

Software

Trifacta – a data preparation and analysis platform

Paxata – self-service data preparation software

Alteryx – data blending and advanced data analytics software

Microsoft Power BI - interactive visualization and data analysis tool

OpenRefine - a standalone open source desktop application for data clean-up and data transformation

Tableau software – interactive data visualization software

See also

Exploratory data analysis

Machine learning

Data profiling

Data visualization

References