Big data – The methods behind the term

The umbrella term “big data” describes the various ways in which masses of data are acquired, processed, and analyzed. The term therefore also refers to a variety of technological applications. The following examples illustrate how data can be both acquired and analyzed.

Tracking

When a user calls up a website, a file called a cookie is stored in the browser. The cookie records usage data, also called tracking data, which are of interest to industry, for example. Many companies use big-data collections of this type to optimize their products and services or to increase their sales. For instance, if a news provider uses tracking services to determine which articles are only partially read or which articles are frequently read in succession, it can shorten texts or concentrate on particular topics in the future.

The myriad users of social networks such as Instagram also generate huge volumes of data that are saved and analyzed by the respective software. Social networks use these data to offer target group-specific advertising that they sell to companies. The advertising customer can select the target age of the people who will see its advertisement, where the people come from, whether they are male or female, and what they are interested in. Social network operators earn money with the acquired and processed data. To protect themselves from such applications, users can install browser extensions that at least limit tracking. Users can also restrict the use of cookies via the browser settings. However, complete suppression of cookies usually means that websites can no longer be used without restrictions.
Fingerprinting is a newer form of tracking that works without cookies. With this method, it is possible to uniquely identify people across various browsers. Information such as the operating system, the fonts used, or screen resolution is captured to create a unique profile. The first browser extensions to suppress tracking via fingerprinting to a large extent have been developed, such as CanvasBlocker for Firefox.

Data mining

With big-data methods, it is possible to determine correlations and patterns among multitudes of data. This results in models which can be used to make predictions. Weather forecasts are one example of such predictions. For this reason, statistical methods are used for big-data analyses. The approach of applying statistical methods to comprehensive databases to filter information from them is also referred to as data mining. Various types of analyses are used in the process, such as the following three:

Clustering

With this method, similarities between individual cases (for example, persons) within large masses of data are identified in order to form groups. These groups are called clusters. Many different characteristics can be used for cluster analysis. For example, parties in an election campaign develop various clusters of voter types based on age, gender, and level of education to obtain a more precise profile of the voters. Using this information, campaign aides can target voters more selectively.

Association analysis
With this method, dependencies of characteristics and rules can be established. Frequently, this has to do with decisions: If someone chooses A, he or she will also choose B next with a certain probability. For instance, if someone places a book in an online shopping cart and is then shown suggested book titles that may appeal to him or her, association analysis has taken place. Before such an automated recommendation is made, other customers' purchases are analyzed for dependencies that are transferred to this purchase.

Regression analysis
With this method, correlations are determined from available data. It is similar to association analysis and helps find out whether data are dependent or independent of each other. A simple regression considers only two values, such as a) the frequency of errors made in a two-hour in-class exam and b) the elapsed time. Predictions can be made based on a regression model, for example, that for every ten minutes that the in-class exam lasts, the error frequency increases by one error.