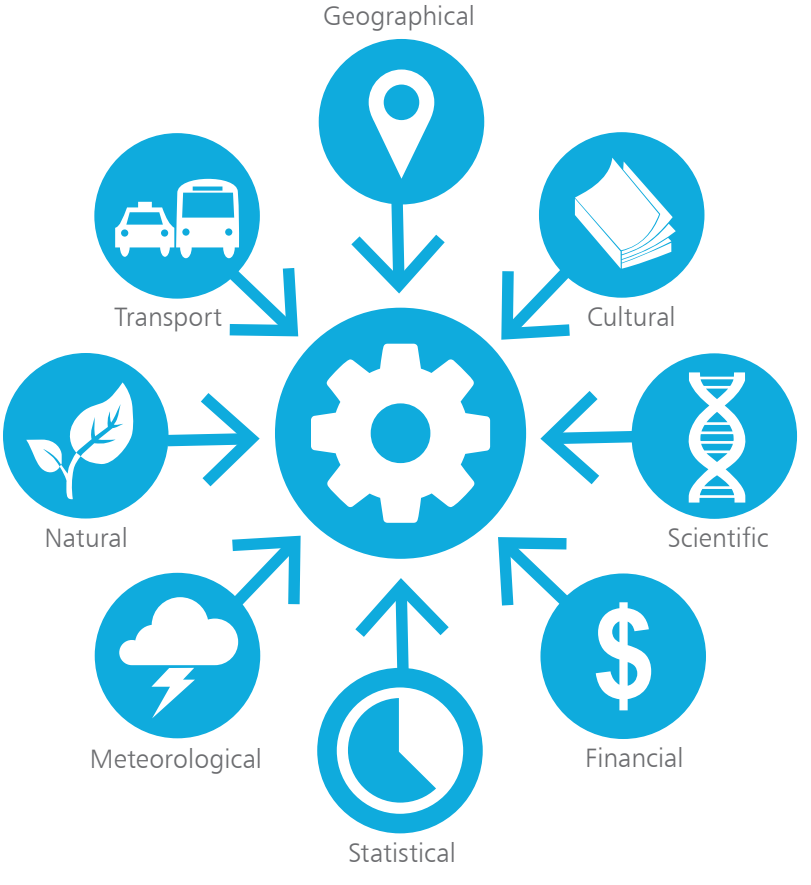
Understanding data

The term "data" refers to items of information that describe a (qualitative) status or a (quantitative) measure of magnitude. Various types of data is collected from a huge range of sources and reported for analysis to reveal pattern and trend insights:

This illustration depicts only some of the many data types that can be reported for analysis.



Around 13 billion devices are connected to the internet today. This is predicted to grow to 50 billion by 2020.

Data is increasingly being collected by devices that are able to report measurements for analysis via the internet ("The Cloud"). For example, devices that have temperature and humidity sensors can report measurements for instant analysis of climate conditions. The recent rapid decline in the cost of device sensors has given rise to the "Internet of Things" (IoT) that can easily and cheaply report vast amounts of data – this is often referred to as "big data". Big data consists of extremely large data sets that can best be analyzed by computer to reveal pattern and trend insights.

...cont'd

Data analysis (a.k.a. "data analytics") is the practice of converting collected data into information that is useful for decision-making. The collected "raw" data will, however, typically undergo two initial procedures before it can be explored for insights:

- **Data processing** – the raw data must be organized into a structured format. For example, it may be arranged into rows and columns in a table format for use in a spreadsheet.
- **Data cleaning** – the organized data must be stripped of incomplete, duplicated, and erroneous items. For, example, by the removal of duplicated rows in a spreadsheet.

After the data has been processed and cleaned it can be explored to discover its main characteristics. This may require further data cleaning to refine the data to specific areas of interest, or may require additional data to better understand its messages. Descriptive statistics, such as average values, might be calculated to understand the data. Algorithms might be used to identify associations within the data. Data visualization might also be used to produce a graphical representation of the data for examination.

After the data has been analyzed, the results can be communicated using data visualization to present tables, plots, or charts that clearly and efficiently convey the key messages within the data. Tables provide information in which the user can look up a specific number, whereas plots and charts provide information in a way that encourages the eye to make comparisons.

"R" is an interpreted programming language and software environment that is widely used for data analysis and visualization. The "RStudio" Integrated Development Environment (IDE) is often used with R, as RStudio provides a code editor, debugging features, and visualization tools that make R easier to use. The popularity of R has grown rapidly in recent years as the increase in big data has made data analysis more important than ever.

The R programming software and RStudio IDE are both available for Windows, Linux, and macOS operating systems, and both are used throughout this book to demonstrate R for data analysis.

"Data Science" is the study of how data can be turned into a valuable resource.

"Data Mining" is the process of searching large data sets to identify patterns.

"Data Product" is digital information that can be purchased.