INDUSTRY 4.0 for VET

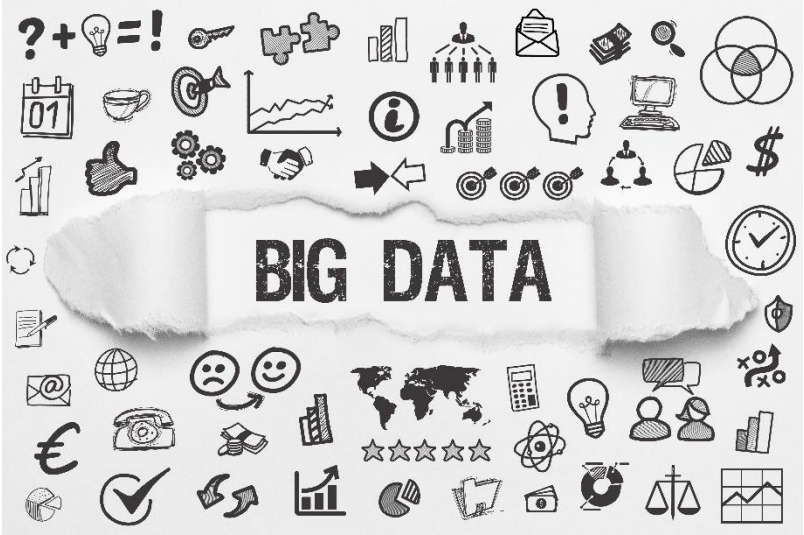
3. BIG DATA

3.1 The topic

The first introduction

Can you imagine that it would take a human being about 181 million years to download all the data from the Internet? These large amounts of data that are available today, and the way they are processed, are called **Big Data**.

As you will see in this unit, we are confronted with this almost daily - often without our knowledge. You will learn about the **advantages** but also the **dangers** of Big Data and why **correct handling** of these large amounts of data is often more important than the data itself.



The practical relevance - For this you will need the knowledge and skills

Not only IT specialists, almost everyone encounters the so-called large amounts of data in **everyday situations**, such as when visiting a doctor, surfing social media like Facebook and Instagram, searching on google or in a networked vehicle.

Knowing how large amounts of data are used and what **opportunities** but also **dangers** are associated with this can be relevant both for your **personal use** of the **Internet** and for your **professional life**, perhaps in a company that analyses large amounts of data.

Learning objectives and competences at a glance

This learning unit gives you a basic understanding of Big Data. You will get to know the **3-V model** and learn how large amounts of data are collected and analysed. You will then learn about the **purposes** for which the **knowledge** gained from the large amounts of data is used and the **risks** involved in handling it. You will see why **data protection** has become increasingly **important** in recent years and understand that **handling** Big Data poses major challenges for both companies and private individuals.

Learning Objectives
Understand and describe the term Big Data.
Know how to use Big Data.
Understand and explain how large amounts of data are collected and analysed.
Know what challenges and risks Big Data contains.

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

3.2 What is Big Data?

Did you know that around 90 percent of all the data available around the world today was generated in recent years? Due to the numerous new information and communication technologies, the **volume of data** worldwide has grown incredibly and offers previously unknown possibilities. **Big Data** stands for this **volume** of structured and unstructured **data**, which cannot be processed with conventional software or hardware due to its size.

These data volumes are created, among other things, with each of our **clicks on the Internet**. This can be, for example, a purchase on Amazon, a search query on Google, activity on social networks such as Instagram or Facebook etc.

However, large amounts of data alone do not make Big Data. Only the **analysis** and **processing** of these data volumes, e.g. by a company, distinguishes Big Data. In 2001, analyst **Doug Lane** created a definition of Big Data with his **3-V model** that is still recognised today. According to Lane, Big Data has the following three characteristics:

- **Volume:** Companies collect large volumes of data from various sources. These include intelligent devices (IoT) such as mobile phones, videos, social media, etc. In the past, it would not have been possible to store these large volumes of data; today, storage platforms exist for this purpose.

- **Velocity:** Companies are currently being flooded with data streams at unprecedented speeds that need to be processed quickly.

- **Variety:** The data collected is diverse and has a wide variety of formats: numerical data, which is available in structured form and stored in ordinary databases, can be part of Big Data, as well as unstructured text documents, data from financial transactions or e-mails.

Definition
Big Data ...stands for a **large amount** of available **data** that is **analysed** and **processed** for a specific purpose. According to Doug Lane, Big Data is characterised by **volume**, **speed** and **diversity**.

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Big Data vs. Small Data

**"Let's shrink Big Data into Small Data ...
and hope it magically becomes Great Data."**

Unlike Big Data, Small Data refers to data in a volume and format accessible to humans. The following points show how Big Data can be distinguished from Small Data:

- **Targets:** Small Data is used for a defined goal, the use of Big Data often develops unexpectedly.
- **Location:** Small Data is generally stored in one place, usually in one file on the PC, while Big Data is usually spread across numerous files on different servers located in different countries.
- **Data structure:** Small Data is structured in a straight line, whereas Big Data can be unstructured and can have many file formats from different fields.
- **Data preparation:** only one end user is usually involved in the preparation of Small Data. In the case of Big Data, however, it is often the case that one group of people prepares the data, another group analyses the data and finally a third group uses the data. Each of these groups may have different objectives.
- **Durability:** Small Data is generally retained for a certain period of time after the completion of a project. In the case of Big Data, however, the data remains stored for an unlimited period of time.
- **Origin:** Small Data is stored within a short time and in specific units of measurement. Big Data, on the other hand, originates from different places, countries, companies, organisations, etc.
- **Reproducibility:** Small Data can generally be completely reproduced. Big Data, by contrast, originates from many different sources and is available in many forms that reproduction is impossible.
- **Quality:** the meanings of the data in a Small Data set are unambiguous, these data can therefore describe itself. Big Data, conversely, is much more complex and may also contain unidentifiable information that has no specific meaning. This can reduce the quality of the data.

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

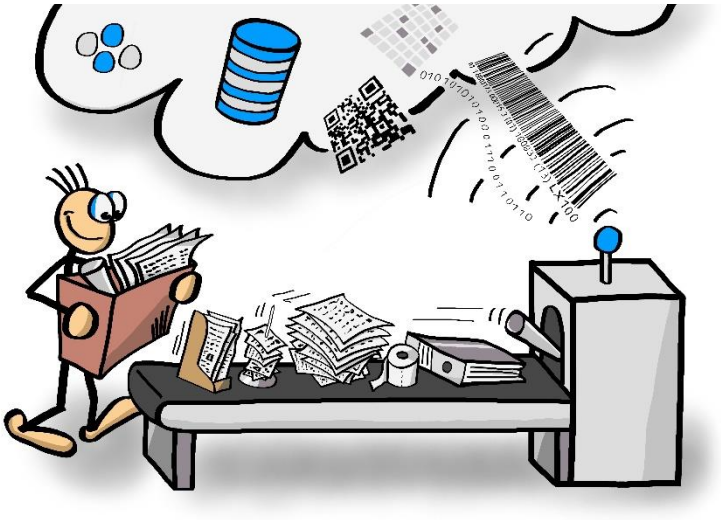
- **Analysis:** a single process is usually sufficient for the analysis of Small Data, since the data is analysed from only one computer file. In the case of Big Data, the data must be extracted, checked, reduced, etc. in a time-consuming process.

As you can see from the distinction between Big Data and Small Data, Big Data is literally often difficult to grasp.

3.3 Possible uses and opportunities of Big Data

The analysis of large amounts of data makes it possible to gain **insights**. These results can serve as a basis for decisions, for example, regarding the **strategic direction** of the company. Companies, for instance, want to learn more about the preferences of their customers in order to adapt their product range, advertising, and so on, to them.

Deep Learning also uses Big Data: this is a special method of **information processing** and a sub-area of **machine learning**. A machine is "fed" with large amounts of data, which is analysed and used to train the machine. The machine is able to link new information with each other and on this basis can make forecasts and make its own decisions. However, the result is only as good as the data, the machine has "learned" from



One example is a machine translation system that "learns" to correctly translate technical terms in a company by entering data (existing translations).
In addition, **authorities** and **secret services** use large amounts of data to detect discrepancies and anomalies that could indicate criminal or terrorist activities. In **science**, large amounts of data are used to investigate **complex natural phenomena** such as climate change or the occurrence of earthquakes and epidemics.

However, the large amounts of data are not always handled **responsibly**. Some companies or institutions do not adhere to data protection regulations, which means that information is released to the public. This can be trivial, but in some cases it can also be dangerous and lead to **fraud** and **blackmail**.

Example
In 2015, the Ashley Madison fling portal, where people in search of an extramarital adventure can create a profile, became the victim of a hacker attack. As a result, information about the people registered on the portal became available on the Internet. Information on celebrity flings and personal information such as credit card numbers became public. In addition, those affected were asked by e-mail to pay a ransom so that their life partner would not find out about the profile on the fling portal.

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Remember
Large amounts of data can be used for the following **purposes**, among others: • strategic orientation of companies • Deep Learning • fighting against crime and terrorism • scientific investigation of natural phenomena (e.g. earthquakes and climate change) • unlawful evaluations which may lead to blackmail or fraud The decisive factor regarding Big Data is not so much the data itself as what happens to it.

Companies in particular benefit from analysing and evaluating Big Data. Both consciously and unconsciously, they generate and store vast amounts of data today. In the following, you will learn in detail what possibilities the correct analysis of large amounts of data offers companies.

Decision-making

By analysing the large amounts of data generated in the company, companies can identify patterns and filter out information. This enables companies to make better business decisions that increase the success of the company. By evaluating machine data, for example, it is possible to calculate at what intervals a machine breaks down. The company can use this knowledge to service the machine before it fails. Big Data is also used in the finance and insurance industry to better calculate risks.

Example
Ms Schmidt is 47 years old and would like to conclude a private health insurance. When visiting her insurance broker, she is surprised about the high costs and enquires. It turns out that her provider analyses large amounts of data in order to better calculate the individual insurance costs. The company finds out, for example, what particular health risks women of this age bear who, like Ms. Schmidt, are smokers, have no children and never do sports.

Increase in efficiency

Competitiveness is very important for companies. In order to keep up with the competition, companies need to design strategies to save costs without compromising performance. Analysing and connecting large amounts of data helps to do this.

Example
Have you ever heard that UPS drivers almost always turn right? That's because UPS has discovered, through big-data analysis, that this can save about $10 million a year. You're probably wondering how that's possible: the merging of various data sets, such as accident statistics, fuel consumption data, etc., has shown that UPS vehicles are much less likely to be involved in accidents if they don't turn left. This can save a lot of money, even if the routes become more complicated as a result.

Prediction in research and development

By making existing or potential customers or clients aware of their preference for certain products, research can identify and predict trends, design appropriate marketing strategies and develop tailor-made products. With the appropriate analytical methods, it is also possible, for example, to predict the rupture safety of a product while it is still being developed.

Co-funded by the Erasmus+ Programme of the European Union

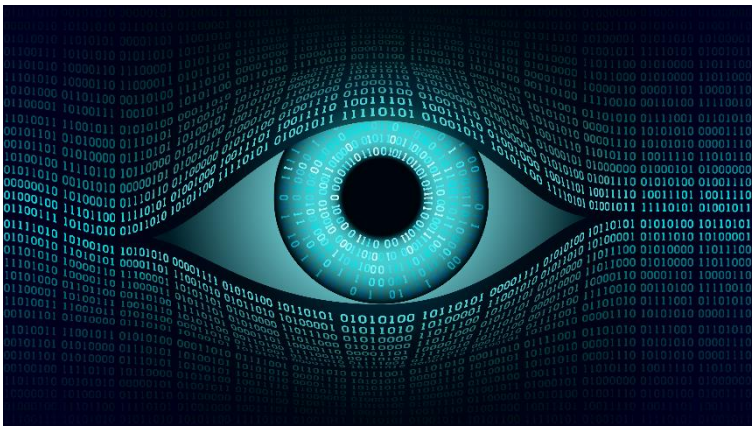
This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Example
An operator of an online web shop installs cookies and online tracking and tracks the movements of its visitors. He can determine where visitors come from, which products they click on, how often they visit the site, etc. With the help of this data, the operator can adapt the contents of the site and the products offered to the preferences of the visitors and thus increase his turnover.

Personalised customer service

By storing customers' decisions, companies are able to provide them with personalised customer service. For example, if a user watches a particular movie or series on Netflix, the system will save it and the next time the user logs in, recommendations will be based on the movies or series the user has already watched. But this personalised offer does not always meet approval:

Example
When Mr. Maier realises that his old mountain boots are no longer usable, he searches on Google for "mountain boots new for men". He is overwhelmed by the many different offers and Mr. Maier also discovers that many products cannot be delivered to his home country, Austria. Mr. Maier decides to get personal advice in a specialist shop and also buys a pair of mountain boots. Nevertheless, he sees more and more advertising for mountain boots on the Internet in the coming days and weeks, as his search query has been saved and analysed on Google. Mr Maier is irritated and feels observed. He decides not to place any more search queries on Google in the future.



Let's recap once again:

Remember
Companies have numerous opportunities to use Big Data to be more successful. These include: • **Decision making:** Big Data analysis enables companies to make better business decisions and better assess risks. • **Increased efficiency:** Analysing and linking data (such as weather and congestion data with fuel prices) helps companies to make processes more efficient. • **Forecasting in the field of research and development** With the help of Big Data, predictions can be made regarding trends, characteristics of a product, etc. • **Personalised customer service** By storing the decisions made by customers, companies can offer them personalised customer service on their next visit.

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

3.4 How is Big Data analysed?

You have learned how Big Data is defined and what options there are for using the large amounts of data. In this chapter, we will go into more detail and deal with the **analysis** of Big Data. This specialist field is known as **Big Data Analytics**.

Big Data Analytics – Theory

The first step is to collect **large amounts of data** from different sources, which have different formats. This is often done using search queries. Then the data is **prepared** for further processing. One problem is often that large amounts of data are available in an unstructured form and in completely different formats and therefore cannot be captured by conventional database software.

Big Data Analytics therefore uses **complicated processes** to extract and capture the data. The data is then **analysed** using special Big Data software. Finally, the results are **processed** and **presented**.

It is important that the software used is capable of quickly implementing many search requests and quickly importing and processing the various data records. In order to be even more powerful, many systems do not use the **hard drive space** (like conventional database applications) for data processing, but rather the usually much faster **main memory**. This way, the access speed can be increased, and analyses can be performed almost in real time.

Remember
The **analysis** of Big Data can be roughly divided into three different areas: • Procurement of data from many and various sources by using search queries • Evaluation and optimisation of the collected data • Data analysis and the summary and presentation of results A **powerful** and **suitable software** is very important for that.

Big Data Analytics – In practice

It is interesting to note that Big Data Analytics is still in its infancy in most companies and the **opportunities** offered are **far from being exhausted**. On average, companies analyse only a little more than **a third** of the data generated by digital contact with their customers (e.g. via online shops or websites).

The reason for this is often the strict **data protection** regulations which make Big Data Analytics more difficult. The laws and regulations that govern data protection are discussed in more detail in the following chapter. In reality, however, in many respects companies are not yet ready to effectively use the large amounts of data for themselves. The following areas play an important role:

First of all, it is advisable to distribute the results correctly: the **data sources** should come from different areas, the results should be used in several areas of the company. A suitable **strategy** is also required: a company should know for what purpose the large amounts of data are being analysed. A suitable **corporate**

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

culture is also very important, new technologies, for example, should not be rejected in principle, but rather considered realistically.

Most companies do not have their own department for data analysis. Nevertheless, some employees should bring the **necessary expertise** with them or acquire it in training courses. New employees may need to be hired. Responsibilities and authorisations must also be defined within the company.

Efficient technology, in the form of suitable **Big Data analysis tools**, is required for the analysis. However, which tools are suitable depends on the previously defined strategy or the defined purpose of the analysis. Last but not least, a suitable data protection strategy is also essential to ensure that personal data of individuals is not disclosed to the public. A dedicated data protection expert within the company ensures that the analysis of the data complies with the applicable laws and regulations.

Remember
In summary, the following points are important for **Big Data Analytics** to **succeed** in a company: • a Big Data strategy - defining the purpose of the analysis • a suitable corporate culture - openness to new technologies • personnel with the necessary know-how - training or recruiting • a powerful technology - appropriate big-data analysis tools • an appropriate data protection policy - compliance with applicable laws and regulations

3.5 Challenges and risks of Big Data

In the previous chapters you have witnessed how complex it is to analyse and use Big Data. At least as complex are the challenges and risks associated with large amounts of data. Probably the biggest **challenge** for companies in connection with Big Data is **data protection**:

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Although companies have been paying more attention to data protection in recent years, there are still problems. For example, personal data of Internet users will be used without their consent and the persons concerned can be identified, controlled and in the worst case blackmailed.

Definition

Personal data

... refers to **data** that relates to a **person** and allows conclusions to be drawn about their **personality**. This includes, for example, Werner Kogler's license plate number, your neighbour's date of birth or Bill Gates' account balance.

An example of a **data protection violation** in connection with Big Data is the case of the Ashley Madison fling portal, which was already mentioned as an example in chapter 2. In this case, the **personal data** became **public** and was used to **blackmail** the owners of the data.

Data protection regulations and laws help to protect consumers from abuse. The **basis** of the **general data protection law** in the European Union and in Austria is the **General Data Protection Regulation**, which became effective on 25 May 2018.

Excursus

The General Data Protection Regulation

The General Data Protection Regulation, or GDPR for short, is called in its entirety "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of individuals with regard to the processing of personal data and on the free movement of such data and repealing

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Directive 95/46/EC". It is directly applicable in Austria and is supplemented by the Data Protection Act (DPA) and the Directive on Data Protection.

This regulation enables EU citizens to better control the collection and use of their personal data. This should strengthen consumer confidence in the individual companies. Existing rights of EU citizens are consolidated in the GDPR, and new rights are also established. The rights established in the GDPR include:

- **simplified access to personal data** - this includes providing comprehensive, clear and comprehensible information on the processing of the data
- a new **right** to **data transferability** - personal data will be transferred in a simplified way
- a clearer **right to** erasure **("right to be forgotten")** - data are deleted if a citizen does not agree to his or her data being processed and there is no legitimate reason to keep them
- a right to be **informed about hacked personal data** - companies and organisations shall immediately inform the persons concerned about serious violations of the protection of personal data. The responsible data protection supervisory authority must also be notified

For companies, the GDPR is intended to create more business opportunities and to promote innovation with numerous measures. These include:

- the creation of **uniform EU-wide rules**, which will lead to major savings
- the **appointment of a data protection officer** within authorities and companies dealing with large data sets
- the **designation of a single point of contact** in their own country to which businesses must turn
- the creation of **EU rules for third country companies** to which third country companies must adhere when offering goods or services or monitor how people behave
- the creation of **rules that promote innovation** and ensure that data protection rules are taken into account at an early stage in the development of services or products
- the use of **data protection-compatible techniques** such as **pseudonymisation** (replacement of passages in a data record that make it possible to identify the associated person) and **encryption** (data is encrypted so that it can only be read by authorised persons)
- removing **reporting** obligations for companies in order to promote the free movement of personal data within the European Union
- carrying out **impact assessments** when the processing of the data is likely to threaten the rights and freedoms of individuals

The complete General Data Protection Regulation can be accessed at https://eur-lex.europa.eu/legal-content/DE/ALL/?uri=CELEX%3A32016R0679.

A further challenge is that the existing employees in companies do not always have the necessary **expertise** and are not open to the possibilities that the analysis of large amounts of data offers the company.

Time and resources are often wasted because those involved are not clear about the goal of a Big Data-project or what infrastructure is required for it. Finding and retaining **competent employees** is usually difficult for companies, as they are in great demand.

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Moreover, **Big Data Technology** is diverse and **confusing** for beginners. Have you ever heard of Spark, Hadoop MapReduce, Cassandra or Hbase? These are Big Data technologies with different features and benefits.

In addition, technologies are evolving at a rapid pace, so companies often simply can't keep up with the pace of adoption. Therefore, for companies that are considering using a Big Data analysis, **expert advice** is useful.

Another point is that Big Data projects are very **expensive**. This applies both to companies that choose an on-premise model and to those that prefer a cloud model. The difference is that with an **on-premise model**, the company uses the big data software in its own data centre and is responsible for its operation and maintenance. In a **cloud- model**, on the other hand, the software is only rented by the company and the data remains with the provider.

Definition
On-Premise-Model ...refers to a solution where the company **buys** or **leases** Big Data software and deploys it in its **own data centre**. The company has to take care of the hardware itself, and it also takes responsibility for the use of the software and the data is stored at the company.

Definition
Cloud-Model ...refers to a solution in which a company purchases the Big Data software as a **service**; the provider takes responsibility for maintenance and operation. The company pays a rental price which includes the hardware, operation and maintenance costs. With this solution, the data is stored at the provider.

If a company decides to use an on-premises solution, it must invest in new hardware and hire new employees to operate and maintain the system. In the case of a cloud solution, the company only needs to hire employees to operate and maintain the system, and the company must pay for the cloud services.

After all, the **quality** of data is often poor, and companies are faced with the challenge of harmonising data from different sources of varying quality. For example, an online merchant analyses data from social media, website logs, call centres and websites that have different formats.

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

But even when all the problems mentioned have been solved, it is often not that simple for companies to gain useful **insights** from the large amounts of data. If information is **linked** together and wrong conclusions are drawn, for instance, this can be dangerous.

For example, a person may be considered uncreditworthy by a bank that performs a Big Data analysis because he or she lives in the same neighbourhood as many uncreditworthy people and drives the same car as many people who are considered uncreditworthy. The following example also shows why the correct use of the large amounts of data is crucial:

Example
An online retailer relies on Big Data Analytics, which is based on historical data about customer behaviour. It turns out that people who buy black sneakers often add a pair of black sneaker socks. The retailer adjusts his range for the spring accordingly. However, just before the beginning of spring, a well-known rapper posts a photo of himself with black sneakers and yellow socks on Instagram. Many young people are therefore looking for yellow socks to go with their black sneakers, but unfortunately the online retailer soon runs out of them because he was not prepared for the rush. The retailer simply used the wrong Big Data strategy, relying only on historical results and ignoring other important data sources such as social media, shops of competitors, etc.

Remember
In summary, these are the main **challenges** that companies face when using Big Data: • ensuring **data security** - compliance with the General Data Protection Regulation (GDPR • **professional competence** of the employees - **proficient use** of the diverse and rapidly developing Big Data technology • **high costs** of Big Data projects (hardware and software or rental costs, staff, maintenance etc.) • **poor quality** of data, standardisation of data in different formats and with different quality • **correct interpretation** of the results

As you have noticed, Big Data offers enormous possibilities and opportunities that companies have not even come close to exploiting. However, the large volumes of data are also associated with challenges and risks that should not be underestimated and are unsettling for many people. The decisive factor in ensuring that Big Data is used successfully without causing harm to other people is therefore **responsible** and **proficient** handling of the large volumes of data.

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

3.6 Summary

Big Data refers to large amounts of data that can no longer be processed with conventional software or hardware, and for which processing and analysis is performed for a specific **purpose**. In contrast to **Big Data**, **Small Data** refers to data that are accessible to humans due to their volume and format.

We encounter these large amounts of data in everyday situations, for example when surfing social media or searching on Google. To better define Big Data, analyst Doug Lane designed the **3-V model**, which states that Big Data is characterised by **volume**, **speed** and **diversity**.

Large amounts of data can be used, among other things, to improve the **strategic orientation** of companies, for **Deep-Learning-Systems**, to **fight crime and terrorism**, for the **scientific investigation of natural phenomena** (e.g. earthquakes and climate change), but also for **illegal evaluations** that can lead to blackmail or fraud. The decisive factor is not so much the large volumes of data itself, but what happens to it.

Companies can use Big Data to **increase their business success**. Among other things, Big Data Analytics enables companies to **make** better **business decisions** and **assess risks** with greater accuracy. In addition, the **efficiency** of business processes can be **increased** when data is analysed, evaluated and linked together. Big Data helps companies in research and development to make **predictions** about trends, product characteristics, etc. Finally, the knowledge gained from Big Data can also be used to offer **personalised customer service**.

For a successful Big Data analysis, an appropriate **Big Data strategy**, a suitable **corporate culture**, **personnel** with the necessary **know-how**, **efficient technology** and, last but not least, a **suitable data protection strategy** are required. But analysing and processing big data not only offers opportunities and chances, but also poses challenges and risks.

A major challenge for companies is to **ensure data security** and to comply with the **General Data Protection Regulation**. In addition, it is often difficult for companies to find and retain **suitable professionals** who can handle the **complex Big Data Technology**. Big data projects are also associated with **high costs** and **data quality** is often **poor**. Finally, the right **conclusions** must be drawn from the results of data analysis and the right **decisions** must be made.

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.