The power of data

Data has become one of the most important economic assets of the 21st century. Jorge Marx Gómez and a team of Oldenburg business informatics researchers are investigating how companies can glean new knowledge from accumulated business figures, measured values, images and text documents

You don't need to travel all the way to Silicon Valley to find out how big data could change the future. Right here in Germany, in Lower Saxony, you can get a glimpse of the fascinating world of data science and the vast possibilities of artificial intelligence (AI), machine learning and neural networks. In Oldenburg, for example, where cargo bike couriers working for a postal company will soon be guided through the city by augmented reality glasses. Or in Wolfsburg, where the Volkswagen Group is developing new methods for data-driven auditing. And also at the Chamber of Agriculture in Lower Saxony's experimental station for pig farming in Wehnen, where tests are about to find out how modern data science methods can optimize livestock management.

Cameras have been installed above the spacious pigpens while sensors measure humidity, ammonia levels in the air, feed consumption and other variables. Veterinarians and agricultural researchers are convinced that pigpen digitalisation can make a significant contribution to animal welfare. In the "DigiSchwein" project coordinated by the Chamber of Agriculture, a team of business informatics researchers from the university and the affiliated OFFIS computer science institute headed by Prof. Dr. Jorge Marx Gómez is developing an automated farm management system that would warn farmers if problems arise in their livestock.

Intelligent use of data

Digischwein, which is funded by the Federal Ministry of Agriculture, is one of a series of projects at the Department of Business Informatics/ Very Large Business Applications, all of which focus on the intelligent use of large volumes of data to optimize decision making. In addition to the above-mentioned systems for cargo bike couriers and for the Volkswagen Group, the team is also working on applications that analyse high-resolution operating data from wind power plants, provide advice and guidance for customers buying glasses online, plan routes, and organize car-sharing. "Data helps companies to find new products, access new markets and optimize internal processes," says Marx Gómez. He and his colleagues cooperate with various companies, from start-ups to large corporations, as well as with several cities and regions in Germany, the Netherlands and the UK, and also with other research institutions. The team's goal is to apply innovative data science procedures in areas where they have rarely been used until now – agriculture, cycling and certain areas of business administration.

"Our area of expertise is applying existing, scientifically-based techniques to new problems and putting the data to optimal use," stresses Marx Gómez. He and his team deal with both "structured data", such as numerical values in tables with hundreds of columns and millions of lines, and "unstructured data" such as texts, images and videos. The quantities are often in the gigabyte range – as in the SmartHelm project, in which the researchers are developing an assistance system for cargo bike couriers delivering packages in congested city centres. Images from an eye-tracking camera that records the courier's eye movements as well as audio recordings and other data will all flow into the system.

The first step in a big data project is usually to determine what information is needed to accomplish the task at hand. In the case of SmartHelm, the data management platform should be able to store data from various internal sources as well as from external portals. All this information is analysed to measure how much of the courier's attention is absorbed by traffic and other environmental stimuli. Then in line with this assessment the smart glasses provide the couriers with information about the route or their current assignment, such as which entrance a parcel should be delivered to. This means that the courier doesn't need to search for the information on a smartphone, which can be dangerous in traffic. The team is currently focusing on how the different data sources may be combined and analysed to reliably measure attention levels. Furthermore, an important aim is to reduce the number of individual sensors to improve comfort of the drivers and wearability of the helmet.

Laborious preparations

The next step is usually to eliminate errors from data that may have been gathered from various sources, remove superfluous information and convert all the data into a standardized format. This preliminary work often takes up about 70 percent of the time in a project. "During these steps you have to keep asking: How can I get the most out of my data?" explains Jan-Hendrik Witte, a researcher involved in the DigiSchwein project. "So the process is not just laborious, it also requires a lot of creativity and brain power."

In the next step the data is analysed. This is where new knowledge is gleaned. "Computers have become faster and faster in recent years, so they can also process more and more data," says Marx Gómez. "At the same time, new ways to extract more knowledge from data are emerging, most of them involving AI." The term Artificial Intelligence (AI) refers to various methods by which computers solve problems independently and ultimately learn from experience. AI is already in use in numerous everyday applications such as facial recognition and automated text translation.

In the case of DigiSchwein, the project based in Wehnen, one of the team's objectives is to find out whether it is possible to detect tail-biting at an early stage. This is a behavioural disorder that can have many causes and is not fully understood. "Obviously, we're not pig farming experts, so we need to cooperate closely with our partners, in-

cluding veterinarians from the University of Veterinary Medicine Hannover and the University of Göttingen," Witte explains. He and his colleague Johann Gerberding have the task of processing the data from the pigpens in such a way that it brings to light previously unknown correlations in the animals' behaviour.

The two scientists first feed training data, in this case images showing the pens from above, into an image recognition algorithm, which can also be referred to as a "model." The goal is that the model learns to recognize the shapes of pigs – a task which none of the currently available image recognition software can perform. The researchers essentially tell the model which of the shapes that it identifies belong to a pig and which do not. In this way the programme gradually gets better and better at recognising the animals on the images – a procedure known as machine learning and one of the most important subdisciplines of AI. A key challenge in this project is to make the system as robust as possible so that it can cope with problems specific to the pigpen, such as cobwebs on the camera lens or changing light conditions.

In the next step, the researchers combine the unstructured data from the video cameras with the structured columns of numbers from the sensors. With the help of other partners in the project they will then look for clues pointing to situations in which problems such as tail-biting arise, for example elevated ammonia levels in the air, or changes in water consumption accompanied by increased jostling among the pigs. To detect suspicious patterns in the data, the researchers use deep learning methods – a special form of machine learning based on artificial neural networks. These mathematical models simulate networks of nerve cells in the human brain and learn rules autonomously on the basis of the data fed into them.

The amount of data also plays a critical role here: "The best way to predict an event such as tail-biting, the causes of which are as yet unknown, is to have as much information as possible about this specific problem," says Witte. After all, a model is only as good as the data you put into it, regardless of which AI method you use, he explains. Once completed, the farm management system should be able to alert the farmer via smartphone whenever something happens in the shed that deviates from the norm.

Dark data

Deviations from the norm are also the subject of the DIfA (Data Intelligence for Audit) project, another undertaking in which the University of Oldenburg is involved. In this case the focus is on deviations in business processes. Whereas DigiSchwein evaluates measured values, i.e. numbers, and unstructured image data together, most of the data collected in companies consists of text. "Roughly speaking, about 80 percent of business data is unstructured, and much of it is never used," says Gerrit Schumann, a business informatics researcher from Oldenburg who is also a member of Marx Gómez's team. Experts refer to this as "dark data", unused data that lies dormant within the systems of every company.

In cooperation with the Volkswagen Group, Schumann and his colleague

1 Jan-Hendrik Witte (left), Jorge Marx Gómez (second from right) and Johann Gerberding (right) are working with Marc-Alexander Lieboldt from the experimental station (second from left) on the DigiSchwein project.

2 The team is developing AI methods to identify indicators of abnormal behavior such as tail biting in pigs.

3 Cameras positioned above the pig pens provide data for automated image recognition.

Jakob Nonnenmacher are currently developing a system that looks for anomalies in structured data such as Excel tables or database exports, as well as in unstructured data such as contracts, protocols, guidelines, and other text documents. "Such anomalies might be ordering processes that are taking longer than usual or indications of fraud," explains Schumann. The system is to be used by employees in the company's internal auditing department. This independent department checks business processes in sub-divisions such as the purchasing or sales departments of the individual Volkswagen brands, looking for anything that could pose a risk to the company. The Volkswagen Group already uses so-called mass data analyses to process the mountains of business data generated by each of its twelve brands. "However, the auditors typically look for anomalies that are already known to them, and so far they have focused exclusively on structured – i.e. tabular – data," Schumann explains.

Nonnenmacher, an external PhD student at the University of Oldenburg working for the Volkswagen Group, is developing a new approach for analysing structured data that doesn't rely on already suspected anomalies. Schumann, on the other hand, focuses on methods for analysing unstructured corporate data. He is currently testing various methods from the field of Natural Language Processing, which deals with the automated processing of text or speech by machines. Words, sentences and entire paragraphs in text documents are converted into machine-readable, mathematical expressions. By analysing the results in a procedure known as "text mining", complex information can be extracted from documents, for example sentiments, contradictions and attempts to conceal information.

To find out which information is actually relevant for their project partners at Volkswagen, Schumann and Nonnenmacher conducted a series of interviews with auditors at the company. "Expert knowledge plays a key role in all our projects when it comes to interpreting results and validating the respective model," Marx Gómez stresses.

The programmes that emerge from these projects can often do amazing things. In the DIfA project, the final system would work as follows: The auditors upload the data to be checked, either as structured tables or unstructured texts; the system can then choose from around two dozen different algorithms for analysing the data, and decides – depending on the type of data – which of these it will use for each analysis. After analysing the data, the algorithm extracts the data sets in which it has identified anomalies, classifies them using a points system, and provides an explanation of what exactly the anomaly is.

This would take a weight off the auditors' shoulders, as it would substantially reduce the risks posed by dark data and previously undetected irregularities. (uk)

No more looking at a mobile phone while driving: the SmartHelm project is developing an assistance system for bike couriers. They receive relevant information via smart glasses – but only when opportune.