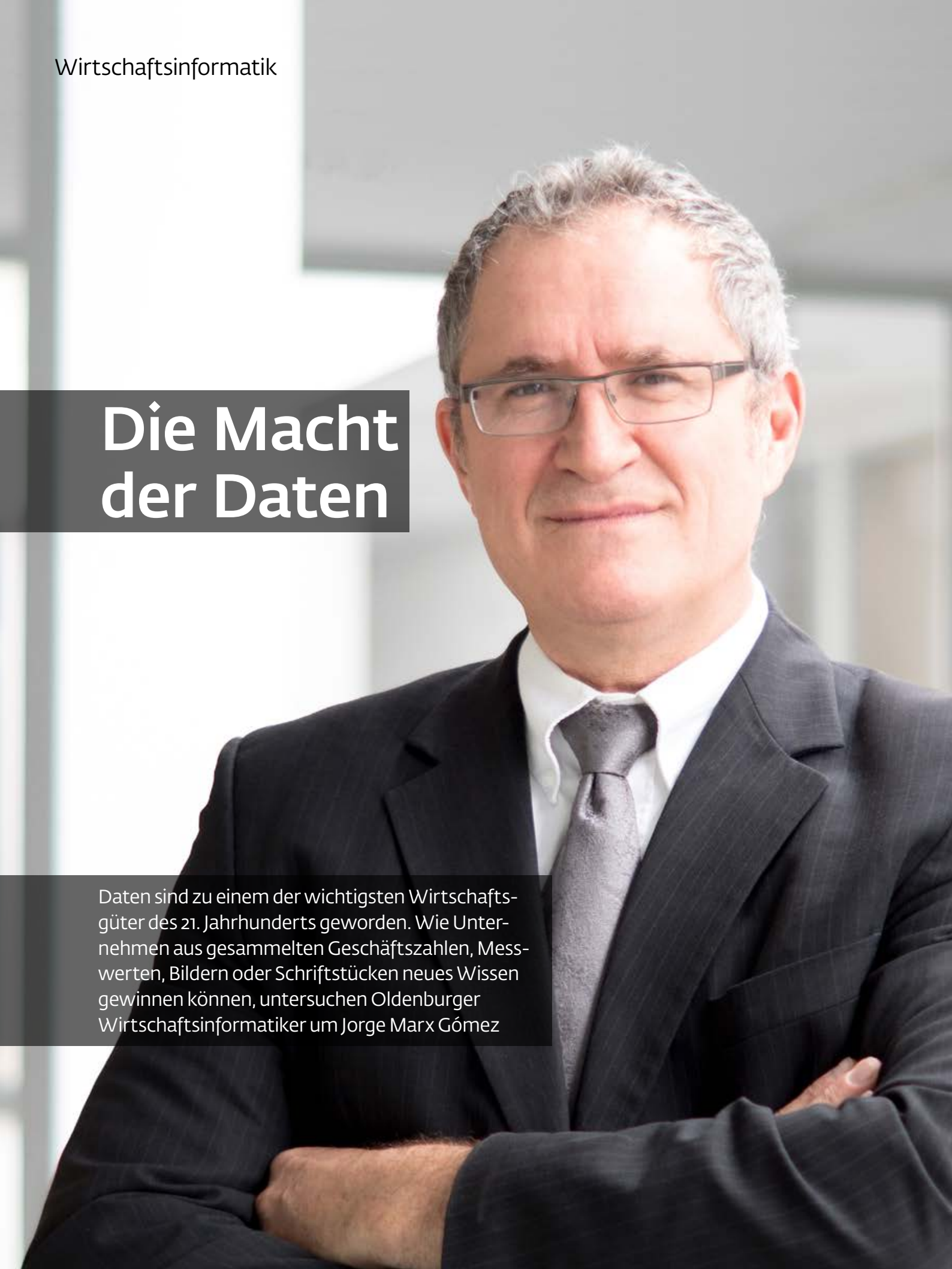
Die Macht der Daten



Daten sind zu einem der wichtigsten Wirtschaftsgüter des 21. Jahrhunderts geworden. Wie Unternehmen aus gesammelten Geschäftszahlen, Messwerten, Bildern oder Schriftstücken neues Wissen gewinnen können, untersuchen Oldenburger Wirtschaftsinformatiker um Jorge Marx Gómez

Wer wissen möchte, wie „Big Data“ die Zukunft verändern könnte, muss nicht unbedingt das Silicon Valley in Kalifornien besuchen. Auch in Niedersachsen kann man einen Einblick in die faszinierende Welt der Datenwissenschaften bekommen – in die Möglichkeiten von Künstlicher Intelligenz (KI), maschinellem Lernen oder neuronalen Netzen. Etwa in Oldenburg, wo Lastenradfahrer eines Postdienstleisters in Zukunft von einer Computerbrille durch die Stadt gelotst werden. Oder in Wolfsburg, wo der Volkswagen-Konzern neue Methoden für datengetriebene Revisionsprüfungen entwickelt. Und auch in Wehnen, westlich von Oldenburg: In der Versuchsstation für Schweinehaltung der Landwirtschaftskammer Niedersachsen soll demnächst erprobt werden, wie moderne Data-Science-Verfahren Landwirte im Umgang mit ihren Tieren unterstützen können.

Über den großzügigen Buchten des Stalls sind Kameras installiert, Sensoren messen unter anderem die Luftfeuchtigkeit, den Ammoniakgehalt der Luft oder den Futterverbrauch. Die Digitalisierung des Schweinestalls, davon sind Veterinäre und Agrarforscherinnen überzeugt, kann zum Wohlergehen der Tiere erheblich beitragen. Im Projekt „DigiSchwein“, koordiniert von der Landwirtschaftskammer, arbeiten Wirtschaftsinformatiker der Universität und vom Informatikinstitut OFFIS um Prof. Dr. Jorge Marx Gómez daran, ein automatisches Farm-Managementsystem zu entwickeln, das die Landwirte zum Beispiel warnt, wenn es einem Tier nicht gut geht.

Daten intelligent nutzen

Das vom Bundeslandwirtschaftsministerium geförderte Projekt zählt zu einer ganzen Reihe von Vorhaben der Abteilung Wirtschaftsinformatik/Very Large Business Applications, in denen die intelligente Nutzung großer Datenströme im Vordergrund steht. Das Team ist unter anderem daran beteiligt, die Systeme für die Lastenradfahrer und den Volkswagen-Konzern zu entwickeln und arbeitet zudem an Anwendungen, die zeitlich hochaufgelöste Betriebsdaten von Windkraftanlagen auswerten, Kunden beim Online-Kauf von Brillen beraten, Routen planen und Mitfahrgelegenheiten organisieren. „Daten helfen Unternehmen dabei, neue Produkte zu finden, Märkte zu erschließen und innerbetriebliche Prozesse zu optimieren“, sagt Marx Gómez. Seine Mitarbeiter und er kooperieren deshalb mit Unternehmen vom Startup bis zum Großkonzern, mit Städten und Regionen in Deutschland, den Niederlanden oder Großbritannien und mit anderen Forschungseinrichtungen. Ihr eigenes Ziel: innovative Verfahren der Datenwissenschaften in Bereichen zur Anwendung zu bringen, wo sie bislang selten zu finden waren – in der Landwirtschaft, im Radverkehr oder auch in manchen Bereichen der Betriebswirtschaft.

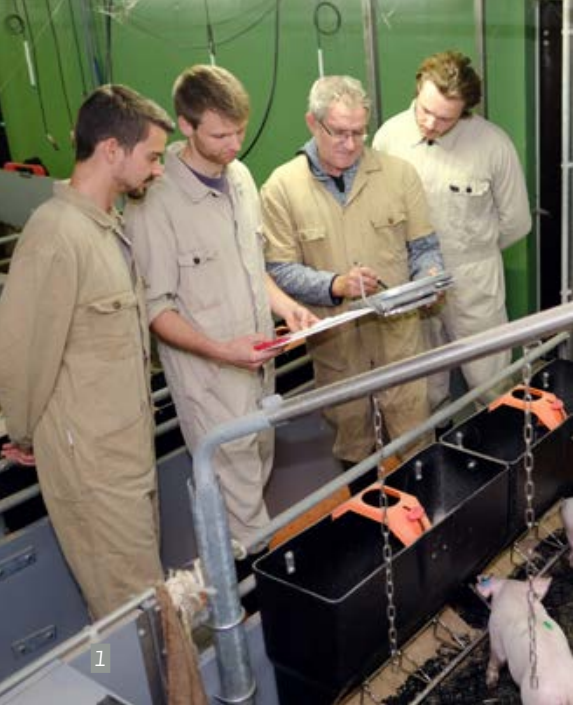
„Unsere Kompetenz besteht darin, vorhandene Verfahren wissenschaftlich gestützt auf neue Probleme anzuwenden. Dabei versuchen wir, das Optimum aus den Daten herauszuholen“, betont Marx Gómez. Er und sein Team beschäftigen sich sowohl mit sogenannten strukturierten Daten, etwa Zahlenwerten in Tabellen, die Hunderte von Spalten und Millionen von Zeilen lang sein können, als auch mit unstrukturierten Daten wie Texten, Bildern oder Videos. Die Mengen können dabei schnell in den Gigabyte-Bereich gehen – wie beim Projekt SmartHelm zur Unterstützung von Lastenradfahrern, die für Kurierfahrten in der Innenstadt unterwegs sind. Bilder einer Eye-Tracking-Kamera, die die Blickbewegungen des Fahrers aufnimmt, Audioaufnahmen und weitere Daten sollen in Echtzeit in ein Assistenzsystem für die Fahrer einfließen.

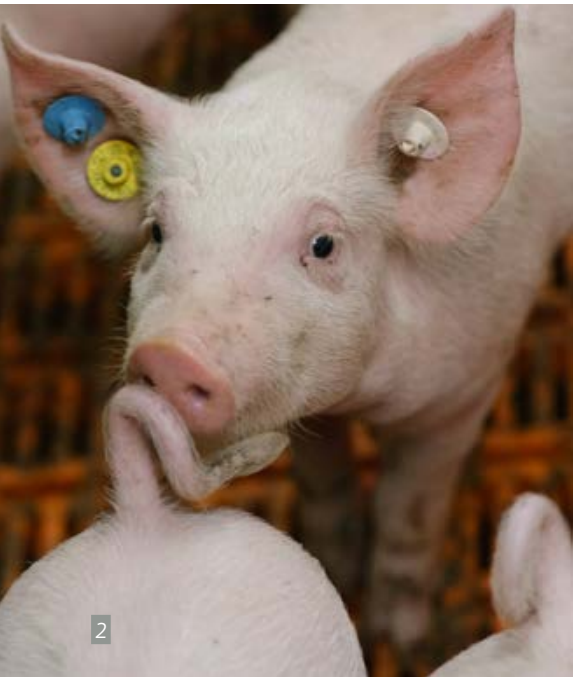
Die erste Aufgabe in einem Big-Data-Projekt besteht meist darin, zu überlegen, welche Informationen überhaupt nötig sind, um die jeweilige Aufgabe zu erfüllen. Im Fall von SmartHelm soll die fertige Plattform ermitteln, wie stark die Fahrer gerade vom Verkehr und anderen Umweltreizen abgelenkt sind. Abgestimmt darauf soll die Computerbrille die Kurierfahrer mit Informationen zu ihrer Route oder zu ihrem Auftrag versorgen, etwa, an welchem Hauseingang ein Paket abgeliefert werden muss. Das Team steht nun zum Beispiel vor der Frage, welche Informationen benötigt werden, um das Aufmerksamkeitslevel zu messen oder wie sehr sich die Auflösung der Kamera reduzieren lässt.

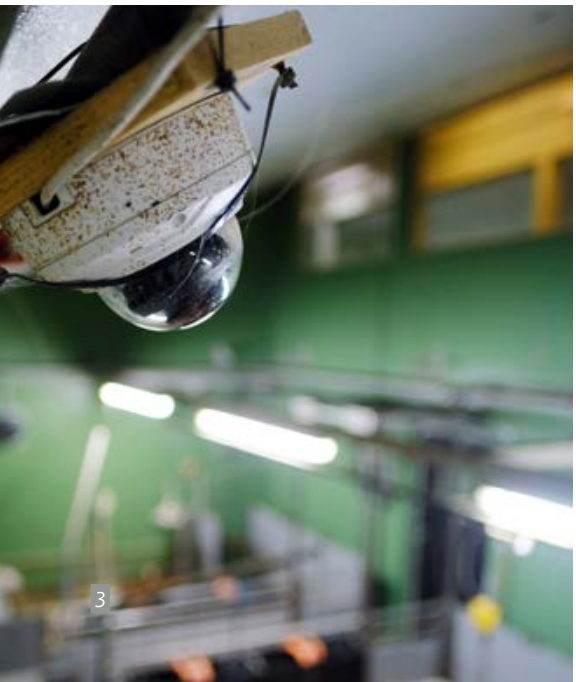
Aufwendige Vorbereitung

Im nächsten Schritt geht es in der Regel darum, Daten, die vielleicht aus unterschiedlichen Quellen stammen, von Fehlern zu befreien, überflüssige Informationen zu entfernen und alles in ein einheitliches Format zu überführen. Für diese vorbereitenden Arbeiten wird oft rund 70 Prozent der Zeit in einem Projekt benötigt. „Bei diesen Schritten muss man immer wieder überlegen: Wie kann ich möglichst viel aus meinen Daten herausholen?“, sagt Jan-Hendrik Witte, Mitarbeiter im Projekt Digi-Schwein. Daher sei dieser Prozess keine reine Fleißarbeit, auch „viel Kreativität und Gehirnschmalz“ seien nötig.

Beim folgenden Schritt, der Datenanalyse, erfolgt der eigentliche Wissenszugewinn. „Weil die Rechner in den letzten Jahren immer schneller geworden sind, können sie natürlich immer mehr Informationen verarbeiten“, sagt Marx Gómez. „Gleichzeitig gibt es auch neuartige Ansätze, um mehr Wissen aus Daten zu extrahieren, vor allem KI-Methoden.“ Unter das Stichwort Künstliche Intelligenz (KI) fallen verschiedene Verfahren, bei denen Computer Probleme eigenständig lösen und schließlich aus Erfahrung lernen sollen. Eingesetzt wird KI im Alltag bereits für zahlreiche Aufgaben, etwa bei der Gesichtserkennung oder bei der automatischen Übersetzung von Texten.







Im Fall von DigiSchwein, dem Vorhaben in Wehnen, will das Team unter anderem herausfinden, ob es möglich ist, das sogenannte Schwanzbeißen frühzeitig zu erkennen – eine Verhaltensstörung, die vielfältige Ursachen haben kann und nicht vollkommen verstanden ist. „Wir sind natürlich keine Experten für Schweinehaltung, daher ist es wichtig, dass wir eng mit unseren Partnern kooperieren, unter anderem mit Veterinären der Tierärztlichen Hochschule Hannover und der Universität Göttingen“, berichtet Witte. Aufgabe von ihm und seinem Kollegen Johann Gerberding ist es, die Daten aus dem Schweinestall so aufzubereiten, dass mögliche bislang unbekannte Zusammenhänge zum Verhalten der Tiere ans Licht kommen.

Dafür speisen die beiden zunächst Trainingsdaten in ein Bilderkennungsprogramm ein: Bilder, die Buchten in Schweineställen von oben zeigen. Das Programm soll lernen, die Umrisse von Schweinen zu erkennen – eine Aufgabe, für die es bislang keine passende Bilderkennungssoftware gibt. Die beiden Informatiker teilen dem Programm daher mit, welche der identifizierten Silhouetten tatsächlich zu einem Schwein gehören und welche nicht. So wird das Programm mit der Zeit besser und erkennt die Tiere auf Bildern immer zuverlässiger – ein Vorgehen, das als maschinelles Lernen bezeichnet wird und zu den wichtigsten Disziplinen der KI gehört. Eine zentrale Herausforderung besteht darin, das System so robust wie möglich zu gestalten, damit es stallspezifische Probleme – etwa Spinnweben über dem Kameraobjektiv oder wechselnde Lichtverhältnisse – bewältigen kann.
Im nächsten Schritt kombinieren die beiden Informatiker die unstrukturierten Daten der Videokameras mit den strukturierten Zahlenkolonnen der Sensoren. Mithilfe der Projektpartner wollen sie Anhaltspunkte dafür finden, unter welchen Umständen Problemsituationen wie das Schwanzbeißen beginnen. Indizien könnten beispielsweise ein steigender Ammoniakgehalt der Luft oder ein veränderter Wasserverbrauch sein, während sich die Schweine gleichzeitig häufiger anrempeln. Um verdächtige Muster in den Daten zu finden, wenden die Forscher Methoden des Deep Learnings an – einer speziellen Form des maschinellen Lernens, bei dem sogenannte künstliche neuronale Netze zum Einsatz kommen. Diese mathematischen Modelle simulieren Netzwerke von Nervenzellen im Gehirn und lernen anhand von eingespeisten Daten selbstständig Regeln.

Dabei kommt es auch auf die Datenmenge an: „Man kann ein Ereignis wie das Schwanzbeißen, dessen Ursachen bislang letztlich noch unklar sind, am besten vorhersagen, wenn man so viele Informationen wie möglich hat“, sagt Witte. Denn bei allen KI-Verfahren gelte: Jedes Modell ist letztlich nur so gut wie die Daten, die man hineinsteckt. Das fertige Farm-Managementsystem soll den Landwirt dann über das Smartphone benachrichtigen können, wenn im Stall irgendetwas passiert, das von der Norm abweicht.

„Dark Data“ – ungenutzte Daten

Um Abweichungen von der Norm geht es auch im Projekt DIfA (Data Intelligence for Audit), an dem die Universität

1 Im Projekt DigiSchwein arbeiten Jan-Hendrik Witte (l.), Jorge Marx Gómez (2.v.r.) und Johann Gerberding (r.) mit Marc-Alexander Lieboldt von der Versuchsstation (2.v.l.) zusammen.

2 Um Indikatoren für Verhaltensauffälligkeiten wie das Schwanzbeißen frühzeitig zu entdecken, entwickelt das Team KI-Verfahren.

3 Kameras über den Buchten liefern Daten für die automatische Bilderkennung.

Oldenburg beteiligt ist – und zwar in Geschäftsprozessen. Während bei DigiSchwein Messwerte, also Zahlen, und unstrukturierte Bilddaten zusammen ausgewertet werden, bestehen viele der in Unternehmen gesammelten Daten aus Texten. „Grob gesagt sind etwa 80 Prozent der Geschäftsdaten unstrukturiert, vieles davon wird nie genutzt“, sagt der Oldenburger Wirtschaftsinformatiker Gerrit Schumann, der ebenfalls zum Team von Marx Gómez gehört. Fachleute sprechen auch von „Dark Data“, ungenutzten Daten, die in jedem Unternehmen irgendwo im System schlummern.

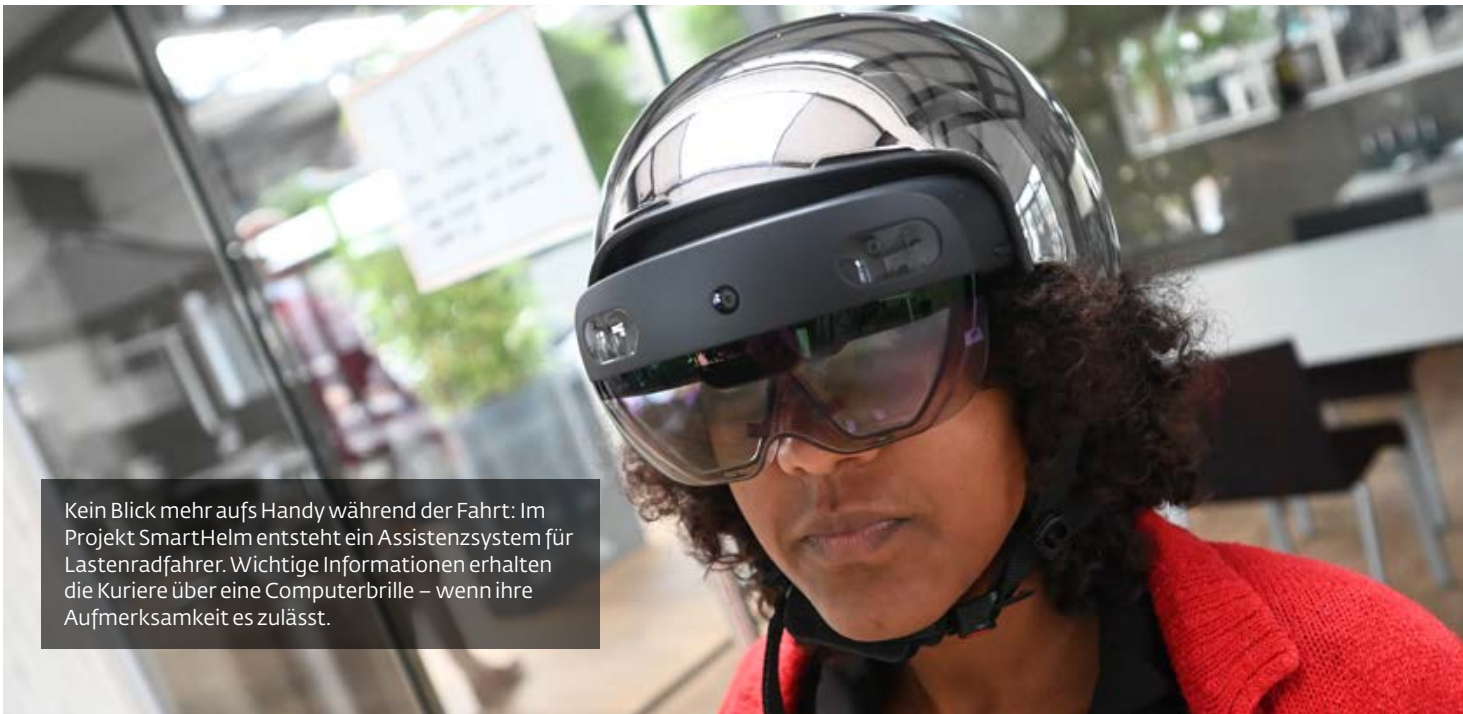
In Zusammenarbeit mit dem Volkswagen-Konzern entwickeln Schumann und sein Projektkollege Jakob Nonnenmacher derzeit ein System, das sowohl in strukturierten Daten wie Excel-Tabellen oder Datenbankexporten als auch in unstrukturierten Daten wie Verträgen, Protokollen oder Richtlinien und anderen Texten nach Anomalien suchen soll. „Das können Bestellvorgänge sein, die ungewöhnlich lange dauern, oder auch Hinweise auf Korruption“, erläutert Schumann. Mitarbeiterinnen und Mitarbeiter der Konzernrevision sollen das System für ihre Arbeit nutzen können. Diese übergreifende Abteilung überprüft betriebswirtschaftliche Prozesse in Unter-Einheiten wie dem Einkauf oder Vertrieb der einzelnen Marken von Volkswagen und sucht nach allem, was dem Unternehmen Schaden zufügen könnte. Der Konzern verwendet bereits sogenannte Massendatenanalysen, um die Berge von Geschäftsdaten zu verarbeiten, die in jeder der zwölf Marken anfallen. „Dabei suchen die Revisoren allerdings typischerweise nach bereits bekannten Anomalien, und sie fokussieren sich bislang ausschließlich auf strukturierte – also tabellarisch darstellbare – Daten“, berichtet Schumann.

Nonnenmacher, externer Doktorand der Universität Oldenburg im Volkswagen-Konzern, entwickelt einen neuen Analyseansatz für strukturierte Daten, der ohne Vorannahmen zu möglichen Auffälligkeiten auskommt. Schumann fokussiert sich hingegen auf Verfahren zur Analyse von unstrukturierten Unternehmensdaten. Der Wirtschaftsinformatiker testet derzeit unterschiedliche Verfahren aus der Computerlinguistik, einem Fachgebiet zur maschinellen Verarbeitung natürlicher Sprache. Wörter, Sätze oder ganze Absätze in Textdokumenten werden dabei in mathematische Größen umgewandelt. Durch deren Analyse, das sogenannte „Text Mining“, lassen sich Dokumenten komplexe Informationen entnehmen, etwa Stimmungen, Widersprüche oder Tendenzen zur Verschleierung.

Um herauszufinden, was für ihre Projektpartner bei Volkswagen tatsächlich relevant ist, haben Schumann und Nonnenmacher eine Reihe von Interviews mit Revisoren geführt. „Das Expertenwissen spielt in allen unseren Projekten eine tragende Rolle, um die Ergebnisse zu interpretieren und das jeweilige Modell zu validieren“, betont Marx Gómez.

Die Programme, die am Ende entstehen, können dann oft erstaunliche Dinge leisten. Bei DIfA soll das fertige System so funktionieren: Die Revisoren laden die zu prüfenden Daten hoch, egal ob strukturierte Tabellen oder unstrukturierte Texte. Das System hat rund zwei Dutzend verschiedene Verfahren zur Prüfung der Daten zur Auswahl und entscheidet selbstständig, welches davon es für die jeweilige Analyse einsetzt. Nach der Datenprüfung wirft das Programm diejenigen Datensätze aus, in denen es Anomalien gefunden hat, klassifiziert diese anhand eines Punktesystems – und liefert eine Erklärung, worin genau die Anomalie besteht.

Für die Revisoren wäre das eine erhebliche Erleichterung: Die Risiken, die sich durch „Dark Data“ und bislang unentdeckte Unregelmäßigkeiten ergeben, wären dann erheblich kleiner. (uk)



Kein Blick mehr aufs Handy während der Fahrt: Im Projekt SmartHelm entsteht ein Assistenzsystem für Lastenradfahrer. Wichtige Informationen erhalten die Kuriere über eine Computerbrille – wenn ihre Aufmerksamkeit es zulässt.