Big und Smart Data

Herausforderungen in der Prozessindustrie

In der Prozessindustrie fallen eine Vielzahl unterschiedlicher, heterogener Daten an, und das Gesamtsystem kann aufgrund seiner Komplexität und Dynamik nicht komplett formal beschrieben werden. Daher untersuchen die Projekte Sidap und FEE die Eignung von Big- Data- und Smart-Data-Ansätzen in dieser Domäne. Obwohl beide Projekte unterschiedliche Ansätze verfolgen, ergeben sich gemeinsame Herausforderungen. Dieser Beitrag fasst diese zusammen und zeigt Lösungsansätze auf, beispielsweise durch die Schaffung eines gemeinsamen Datenverständnisses oder die Anreicherung der Daten mit Zusatzinformation.

SCHLAGWÖRTER Big Data / Smart Data / Prozessindustrie / Datenvorverarbeitung / Datenanalyse / Datenmanagementplattform / Mensch-Maschine

Big and smart data – Challenges in the process industries

In process industries, large amounts of heterogeneous information are generated. However, the overall system is too complex and dynamic to be fully described in a formal way. The projects SIDAP and FEE investigate the use of big data and smart data in this domain. Though the projects have different focuses, they face a number of similar challenges. These challenges are summarized and ways are suggested to overcome them, like establishing a basis for the common understanding of data or the enrichment of data with additional meta-information.

KEYWORDS big data / smart data / process industries / data pre-treatment / data analysis / data management platform / human-machine interface

JENS FOLMER, IRIS KIRCHEN, EMANUEL TRUNZER, BIRGIT VOGEL-HEUSER, TU MÜNCHEN
THORSTEN PÖTTER, BAYER AG
MARKUS GRAUBE, SEBASTIAN HEINZE, LEON URBAS, TU DRESDEN
MARTIN ATZMÜLLER, UNIVERSITÄT KASSEL
DAVID ARNU, RAPIDMINER

Prozesstechnische Anlagen generieren im täglichen Betrieb einen kontinuierlichen Strom an Messdaten. Hinzu kommen beispielsweise Qualitätswerte, Auftrags- und Wartungsdaten. Big Data mit seinen Charakteristiken Volume, Variety und Velocity [4] hat das Versprechen abgegeben, aus den Unmengen an Daten sinnvolle Schlussfolgerungen zu ziehen. Im Gegensatz zu wissensbasierten Systemen [15], die formalisiertes Wissen zur Analyse nutzen, ist hierbei kein Vorwissen über den Inhalt der Daten notwendig. Aber auch Big-Data-Ansätze versuchen zunehmend, mehr Vorwissen in die Analyse zu integrieren, um die komplexen Sachverhalte besser erklären zu können und die Ergebnisse der Analysen stark zu verbessern. Der Begriff Smart Data drückt diese Anpassung von Big Data hin zur wissensunterstützten Analyse von großen Datenmengen und die Nutzung der Analyseergebnisse aus. Die zugrundeliegenden Methoden und Ansätze können hierbei an unterschiedlichsten Stellen in der Prozessindustrie eingesetzt werden. So befassen sich die aktuellen Forschungsprojekte *Frühzeitige Erkennung und Entscheidungsunterstützung für kritische Situationen im Produktionsumfeld* (FEE) beziehungsweise *Skalierbares Integrationskonzept zur Datenaggregation, -analyse, -aufbereitung von großen Datenmengen in der Prozessindustrie* (Sidap) mit der Vorhersage von Geräteausfällen in chemischen Anlagen, der Analyse von Alarmschauern [1] und der Vorhersage der Prozessqualität, um Assistenzsysteme zur besseren Führung der Anlagen anbieten zu können.

In großen Anlagen kommt es immer wieder zu anormalen Situationen. Diese werden sichtbar, wenn sie zu einer großen Anzahl an Alarmen führen, die allerdings die Sicht auf die Problemursachen verdecken können. In kritischen Situationen, zum Beispiel Pumpenausfall oder Überschäumen, bleiben oft nur wenige Minuten für eine angemessene Reaktion, bevor sicherheitsgerichtete Steuerungen die Anlage in einen sicheren Zustand, oft die Abschaltung, überführen. Die Kürze der verfügbaren Zeit erlaubt Anlagenfahrern in der Regel nur eine Reaktion basierend auf dem eigenen Erfahrungsschatz zum Abwenden von Produktionsausfällen – nicht aber eine tiefergehende Situationsanalyse und Ableitung von Handlungen zur wirtschaftlich-technisch optimalen Problembewältigung. Aus diesem Grund ist das Ziel von FEE, sich abzeichnende kritische Situationen frühzeitig zu erkennen und den Operator mit Hilfe von Big-Data-Technologien rechtzeitig prädiktiv zu alarmieren [2]. Damit kann dieser proaktiv mit dem Problem umgehen und wird geringer belastet. Andererseits soll das System die Entscheidungsfindung der Anlagenfahrer in kritischen Situationen unterstützen, indem es vergangene ähnliche Situationen mit Potenzial zur Übertragung der Auswirkungen ausgibt.

Demgegenüber stellt Sidap die Vorhersage von Geräteausfällen in den Vordergrund. Die Geräte verschiedener Anbieter erzeugen in einer heterogenen IT-Landschaft eine Flut von Daten, darunter Nutzungs-, Wartungs- und Qualitätsdaten. Bisher werden diese Daten von den Unternehmen häufig in unterschiedlichen IT-Systemen gesammelt und nur als lokales Ereignis (in einer Anlage) betrachtet. Aggregierte Daten aus mehreren Anlagen werden von den Unternehmen selten weiterverwendet. Der Gerätehersteller sammelt unabhängig davon eigene Daten. Das Potenzial einer übergreifenden Analyse möglichst aller Daten wird derzeit nicht genutzt. Ziel von Sidap ist die Entwicklung und Erprobung von Big-Data-Technologien für diese innovativen und wettbewerbsrelevanten Nutzungsszenarien. Es werden unternehmensübergreifende, sichere und skalierbare Datenintegrationsarchitekturen, sowie Analysemethoden zur Datenaggregation und zur Unterstützung der Entscheidungsfindung im Betrieb entworfen. Dies erfolgt in enger Zusammenarbeit zwischen Betreibern und Geräteanbietern der Prozessindustrie, IT-Anbietern und Forschung. Sidap entwickelt hierzu eine datengetriebene sowie serviceorientierte Integrationsarchitektur. Diese Integrationsarchitektur macht vorhandene Strukturinformation und Daten aus dem Engineering und den Prozessleitsystemen unter Berücksichtigung ihrer unterschiedlichen Semantik in abstrahierter, integrierter und zugriffsgeschützter Form für interaktive Analysen zugänglich. So können Gerätehersteller anhand von

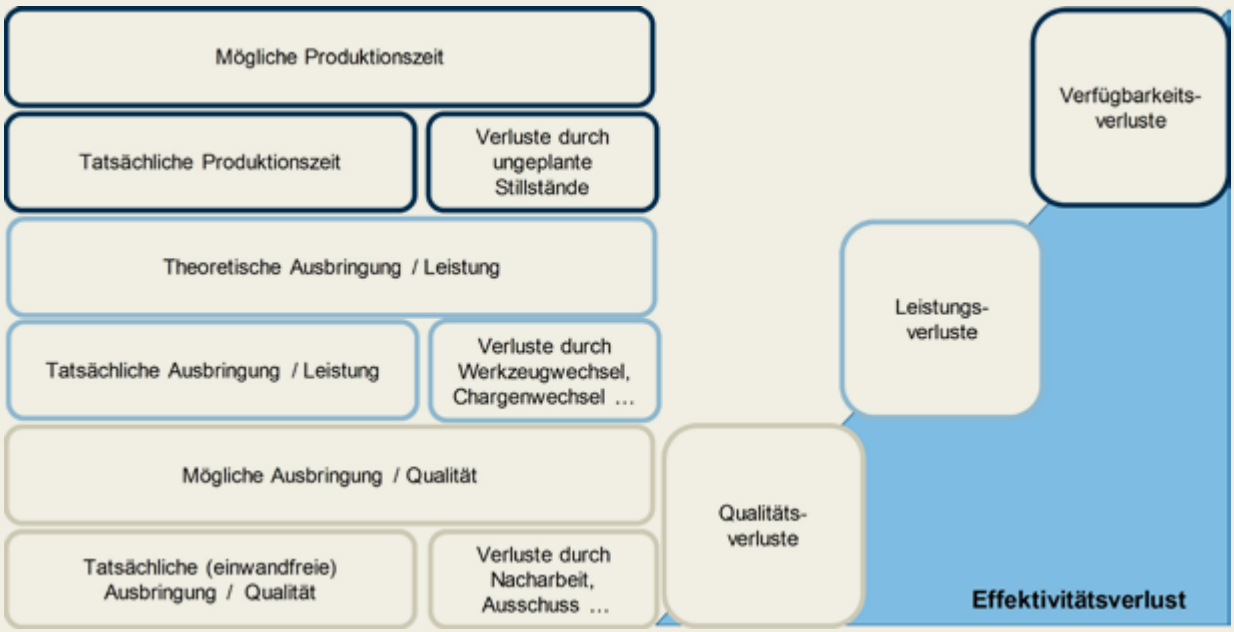

BILD 1: Faktoren für die Beeinflussung der Overall Equipment Effectiveness (OEE)

ausgewählten Nutzungsdaten ihrer Geräte in den Produktionsanlagen und der Wartungs- und Reparaturdaten Gerätestörungen analysieren. Basierend auf der Datenauswertung können Zusammenhänge identifiziert und somit präventiv mögliche Fehler identifiziert und vor dem Eintritt Abhilfemaßnahmen getroffen werden. Für den Anlagenbetreiber wird eine optimale Nutzung der Geräte und damit einhergehend ein möglichst störungsfreier Betrieb sichergestellt.

Beide Forschungsprojekte zielen auf eine Erhöhung der Overall Equipment Effectiveness (OEE) ab. Bei Sidap steht die integrale Betrachtung der Lebenszykluskosten der Feldgeräte in verfahrenstechnischen Anlagen im Fokus. Die OEE kann durch unterschiedliche Faktoren, wie Verbesserung der Produktqualität, Erhöhung der Anlagenleistung oder Steigerung der Anlagenverfügbarkeit, erreicht werden, vergleiche Bild 1. In Sidap soll die Anlagenverfügbarkeit durch die Vorhersage von Geräte- und Equipmentausfällen und die entsprechende Umplanung der Wartung erreicht werden. Die Nutzung der verfügbaren Daten verbessert die Kenntnisse der Geräte- und Equipmentzustände über Unternehmen und Unternehmensstandorte hinweg. In FEE wird ebenfalls die Verfügbarkeit durch das Vermeiden kritischer Situationen mit möglicher Abschaltung adressiert. Hinzu kommt eine Steigerung der Qualität durch eine bessere Prozessführung.

Im Fokus steht ferner die Erhöhung der OEE durch die integrale Betrachtung der Lebenszykluskosten über Unternehmensgrenzen hinweg, zwischen den Herstellern der Feldgeräte und des Equipments, dem Betreiber der Anlagen beziehungsweise der Instandhaltung und, bezogen auf das Reengineering von Altanlagen oder das Engineering von neuen Anlagen, auch dem Engineering. In den letzten Jahren überwog demgegenüber eine stark vom Einkauf und den Kosten des Einkaufs dominierte Betrachtung, die allerdings nicht dazu führte, langlebige Geräte einzusetzen. Der Grund lag auch darin, dass belastbare Ausfallzahlen und Betriebskosten für die Equipments nicht vorlagen.

1. ANWENDUNGSFÄLLE FÜR BIG DATA IN DER PROZESSINDUSTRIE

Big-Data-Ansätze sind für die Prozessindustrie und die beim Betrieb der Anlagen anfallenden Daten höchst relevant. Spezifische Anwendungsfälle sind hierbei notwendig, um den Nutzen durch solche Ansätze zu verdeutlichen.

1.1 Identifikation von kritischen Zuständen

Nach dem Eintritt eines kritischen Ereignisses in einer Prozessanlage werden üblicherweise Maßnahmen ergriffen, um dieses Ereignis in Zukunft komplett zu verhindern, etwa durch zusätzliche Sensorik, kürzere Wartungsintervalle, bessere Prozessführung oder geänderte Vorschriften. In einigen Fällen ist es jedoch nicht oder nur mit erheblichen Aufwand möglich, die Ursache des Problems zu lösen, während es

bewährte Maßnahmen gibt, um mit den Symptomen umzugehen. Ein Beispiel für diese Art von wiederkehrenden kritischen Ereignissen ist das Überschäumen in Kolonnen. Eine Warnung, dass ein solches Ereignis in naher Zukunft bevorsteht, kann es ermöglichen, präventive Maßnahmen, zum Beispiel die Hinzugabe eines Anti-Schaummittels zielgerichteter einzusetzen und somit die negativen Auswirkungen der präventiven Tätigkeit auf die Verfügbarkeit und Qualität des Prozess zu minimieren. Dabei weisen Prozesswerte vor einem solchen Ereignis häufig charakteristische Muster auf, die durch eine Analyse erkannt werden können. Kritische Situationen können jedoch ferner unabhängig von bekannten Ereignissen auftreten. Hier ist ein Vergleich mit einem vorher angelernten Modell des zu detektierenden Ereignisses nicht möglich. Dennoch ist es für die Führung des Prozesses eine wichtige Information, falls sich die Anlage anders verhält, als sie sich bei ähnlichen Bedingungen früher verhalten hat. Dies ermöglicht es dem Anlagenfahrer, aktiv eine Diagnose zu starten, um sich anzeichnende kritische Situationen zu vermeiden. Die Eingangsinformation für eine Analyse bildet die Gesamtheit der verfügbaren Daten über den aktuellen Zustand der Anlage. Dazu gehören Prozesswerte, Alarme, Meldungen, Labormessungen und Planungsinformation. Aus diesen wird ein Modell des normalen Zustands der Anlage gebildet, gegen welches die aktuelle Situation verglichen wird. Daraus kann eine Anomaliebewertung der gesamten Anlage, von Teilanlagen, einzelnen Modulen und Signalen erfolgen. Diese Vorgehensweise erlaubt es dem Operator, sich auf die sich ungewöhnlich verhaltenden Teile der Anlage zu konzentrieren.

1.2 Spezialfall: Regelarmaturen

Für die Vorhersage von Geräteausfällen werden unter anderem Prozess-, Wartungs- und Auslegungsdaten benötigt. Diese werden derzeit in verschiedenen Datenbanken bei den jeweiligen Unternehmen gespeichert. So hat der Betreiber die Prozessdaten zur Verfügung, während der Hersteller des Equipments genauere Information zur Auslegung der Regelarmatur besitzt. Bisher müssen diese Daten für eine Analyse meist händisch zusammengefügt werden. Aus diesem Grund wird eine verbesserte unternehmensübergreifende Systemvernetzung angestrebt, um schneller und effektiver auf diese Daten zugreifen zu können und Datenintegration und -aggregation so weit wie möglich zu automatisieren. Ausgangsbasis für die Analyse ist die erarbeitete Klassifikation von Armaturfehlern, die die Erkennungs- und Prädikationsmerkmale auflistet. Die Klassifikation beinhaltet die unterschiedlichen Sichtweisen auf eine Regelarmatur aus den verschiedenen Phasen des Lebenszyklus und aus Sicht der verschiedenen Stakeholder. Typische Fehlerbilder können zum Beispiel Kegelverschleiß, Fouling oder das Haken des Ventils sein. Je nach Zielsetzung der Analyse, zum Beispiel Erkennung von Nullpunktverschiebungen, Prädikation von Ventilverschleiß, werden aus dieser Klassifikation die Anforderungen an das Analysemodell und die auszuwertenden Daten abgeleitet. Das Ziel der Analyse ist, ungeplante Anlagenabschaltungen zu vermeiden und das Equipment in der Anlage kontinuierlich zu verbessern, zum Beispiel durch die Weiterentwicklung der Regelarmaturen für die nächste Produktgeneration oder durch die Anpassung der Ventilauslegung an die tatsächlichen Bedingungen in der Anlage.

2. SYSTEMVERNETZUNG

Viele der aufgezeigten Anwendungsfälle können nur durch die Vernetzung von bereits bestehenden Systemen adressiert werden. Die Autoren unterscheiden zwischen der unternehmensinternen und der unternehmensübergreifenden Vernetzung.

2.1 Unternehmensinterne Integration unterschiedlicher Datenarten

Zur Lösung anlagenspezifischer Problemstellungen steht hier zu allererst die Integration der unterschiedlichen Datenquellen an, die in einer Anlage vorhanden sind und üblicherweise seit Jahren aufgezeichnet werden. Dabei sind die Daten von Sensoren, aus Engineering- und anderen Datenbanken, aus Prozess-Informations-Management-Systemen (PIMS) sowie aus Schichtbüchern und Betriebsvorschriften sehr heterogen und müssen für die Verknüpfung in einem zentralen Data Warehouse auf eine gemeinsame semantische Basis übertragen werden. In FEE wurde aufgrund der Heterogenität der Daten entschieden, die Verknüpfung der Datenpunkte mit Hilfe von Semantic Web/ Web Ontology Language (OWL) durchzuführen [14]. Dabei werden unter anderem die einzelnen Alarmmeldungen und Datentags semantisch als RDF aufbereitet und unter Nutzung einer selbst erstellten Ontologie mit dem dazugehörigen Datensatz sowie durch Nutzung der Planungsdaten (R&I-Daten) miteinander verknüpft.

2.2 Unternehmensübergreifende Nutzung von Engineering- und Prozessinformation

Für das Beispiel der Ventildiagnose sind in der Datenanalyse Betriebs- und Auslegungsdaten aus unterschiedlichen Quellen zu kombinieren, um suboptimale oder kritische Zustände erkennen und bewerten zu können. Dies erfordert oftmals die Vernetzung über Unternehmensgrenzen hinweg, da die notwendigen Daten nicht gesammelt im Unternehmen selbst vorliegen. Durch eine solche unternehmensübergreifende Nutzung der Daten kann Wissen extrahiert werden, das sonst aufgrund lückenhafter Datenlage im Verborgenen bleiben würde. Im Fokus steht hierbei die

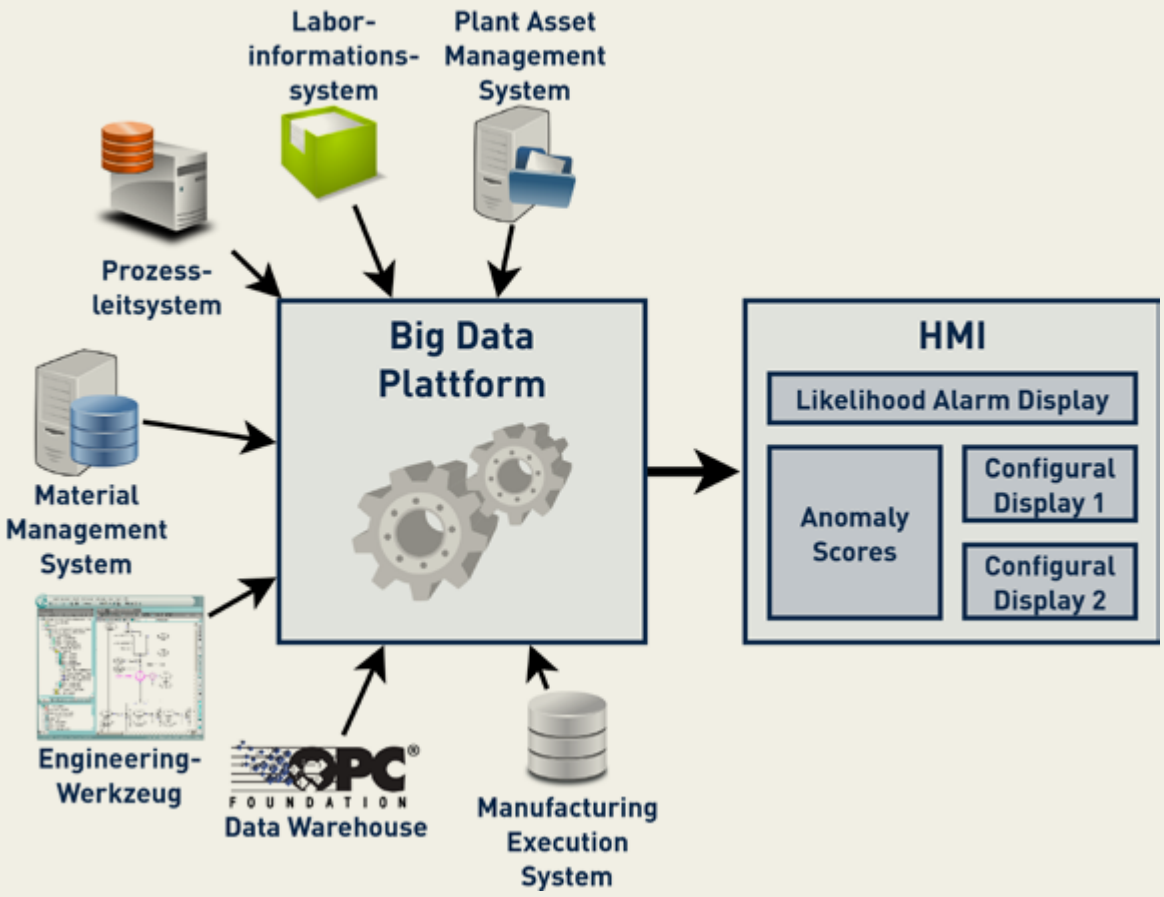

BILD 2: Integration verschiedener Daten aus einer Anlage zur Überführung in Big-Data-Analyseplattformen

Datensicherheit und -integrität: zunächst müssen die Daten, die zwischen den Unternehmen geteilt werden sollen, gemeinsam ausgewählt werden, sodass einerseits die vertrauliche Information über den technologischen Prozess geschützt und andererseits die wesentliche Information für die Beurteilung des Geräts bereitgestellt werden können. Die Rohdaten werden gemäß den Vorgaben des Datenbesitzers, zum Beispiel des Anlagenbetreibers, automatisch anonymisiert. Dies beinhaltet zum Beispiel das Entfernen unnötiger Metainformation oder die Normalisierung von Datenreihen. Des Weiteren muss neben einer manipulationssicheren und verschlüsselten Übertragung der Daten eine sichere Speicherung der Daten am Ort der Verwendung gewährleistet werden.

3. VORGEHENSMODELL FÜR DATENGETRIEBENE ANALYSEN

Intensive Datenanalytik wird nicht um ihrer selbst willen betrieben, sondern muss zu allererst durch geeignete Szenarien, wie in Abschnitt 2 beschrieben, einen Mehrwert für die Betreiber aufzeigen. Aus diesen lassen sich die notwendigen und verfügbaren Daten (online und offline) für den betrachteten Kontext analysieren. Danach kann das System entweder in einem nutzerzentrierten FEE- oder gerätezentrierten Sidap-Prozess entwickelt werden. Die Nutzerzentrierung, siehe Bild 3, bietet sich an, wenn der Mehrwert für den Endnutzer noch nicht vollkommen klar definiert ist. Mit Hilfe von User Stories und einfachen Prototypen werden in der Diskussion mit den Anwendern iterativ funktionale und nicht-funktionale Anforderungen abgeleitet. Hier hat sich bewährt, diese in einen Workflow für die Analytik zu überführen und die Big-Data-Potenziale und Herausforderungen abzuleiten. Hierauf aufbauend werden unter Einsatz spezifischer Techniken zur Analyse umfangreicher und heterogener Daten neuartige Echtzeitmethoden entwickelt, wobei historische Daten auf den aktuellen Kontext bezogen werden. Im Rahmen der vorgesehenen Softwarelösung zur Analyse der Daten werden zunächst vorhandene Daten analysiert und statistische Modelle entwickelt, die im

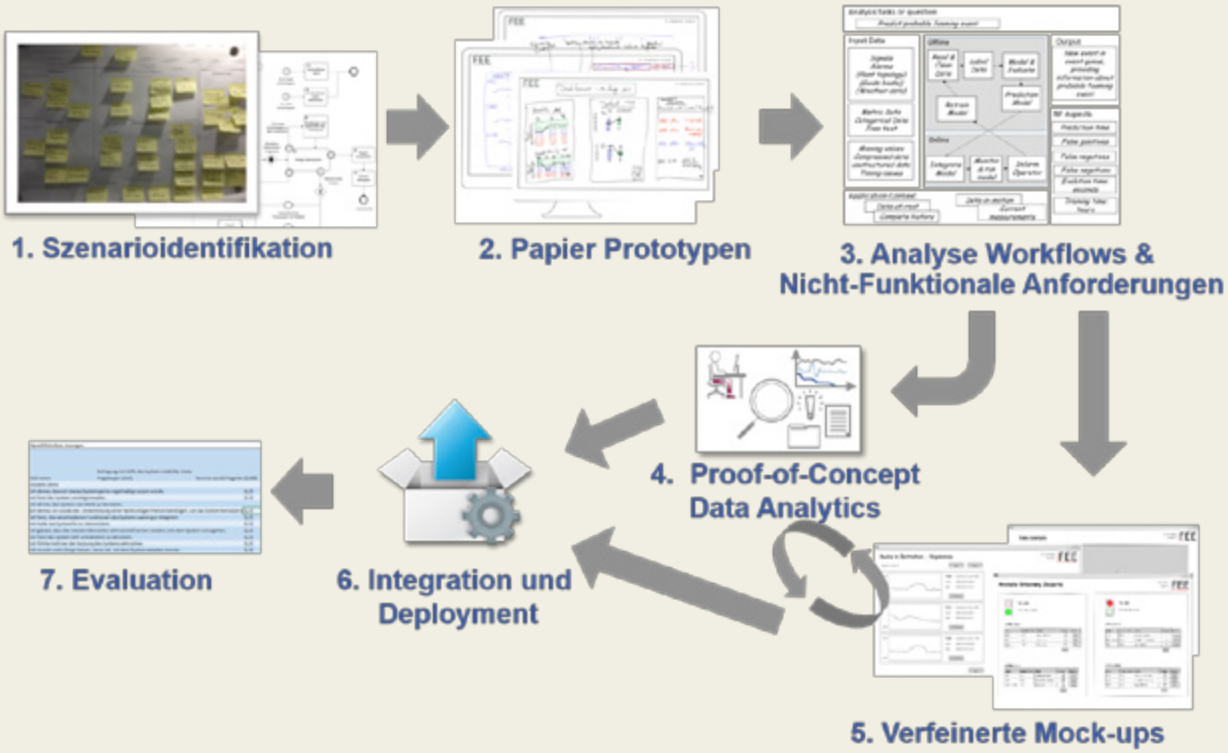

BILD 3: Allgemeines Vorgehensmodell für die Entwicklung datengetriebener Analysemodelle

Sinn eines lernenden Systems im Laufe der Zeit an neue Daten und Ereignisse angepasst werden. Die so erzeugten Modelle werden anschließend permanent zur Entscheidungsunterstützung angewendet. Parallel hierzu wird ein User Interface zur Visualisierung der Analyseergebnisse mit spezieller Betrachtung von Unsicherheiten der Ergebnisse [13] erstellt. Als letzter Schritt steht die Integration in den Betrieb an. Hier ist eine Kopplung an Live-Daten notwendig und die Integration der Analytikergebnisse in die relevanten Dashboards der Entscheider.

Datengetriebene Analyseprojekte zeichnen sich meist durch einen vergleichbaren Datenfluss durch die verschiedenen Systeme und Ebenen aus, vergleiche Bild 3 [7]. Zunächst müssen, sofern mehrere Datenquellen mit unterschiedlichen Formaten mit einbezogen werden sollen, die Rohdaten in ein generisches Datenformat gebracht werden (Datenintegration). Hierfür werden Datenadapter genutzt, die die heterogenen Datenformate und Schnittstellen, zum Beispiel Daten in csv-Tabellenformat, Daten aus dem Manufacturing Execution System (MES) oder über eine OPC-Schnittstelle, an die Analyse anbinden. Dieses dient zum gemeinsamen Verständnis der Daten und zur Zugänglichmachung dieser Daten für die Analyse. Durch das modulare Adapterkonzept können somit einfach neue Datenquellen zugänglich gemacht werden. Anschließend werden diese Daten aggregiert und mit zusätzlicher Metainformation aus anderen Quellen angereichert, sowie, falls notwendig, anonymisiert. Bevor die Daten für eine Analyse benutzt werden können, müssen diese aufbereitet werden (data curation). In diesem Schritt werden zum Beispiel Ausreißer aus dem Datensatz entfernt oder die Abtastraten der einzelnen Signale synchronisiert. Die Integration, Aggregation und Aufbereitung der Daten wird durch eine Datenmanagementplattform, eine Middleware zwischen Datenquellen und Analyse, übernommen. Durch Interaktion des Datenanalysten mit dem zu erstellenden Modell, zum Beispiel Zeitreihenmodell oder physikalische Modelle, wird mit Hilfe der Daten ein Modell trainiert. Der Datenanalyst beeinflusst hierbei sowohl die Modellparameter, als auch iterativ den vorhergegangenen Schritt der Datenaufbereitung. Ein historischer Speicher mit zuvor gelernten Modellen, Parametersätzen und zusätzlichen Metadaten kann dazu genutzt werden, vorhandenes Wissen in das zu entwickelnde Modell mit einfließen zu lassen. Ist der Datenanalyst mit den Ergebnissen des Modells zufrieden, wird dieses in die Produktivumgebung eingebracht und kann zur Diagnose und Prädiktion verwendet werden. Neben den optimierten Parametern

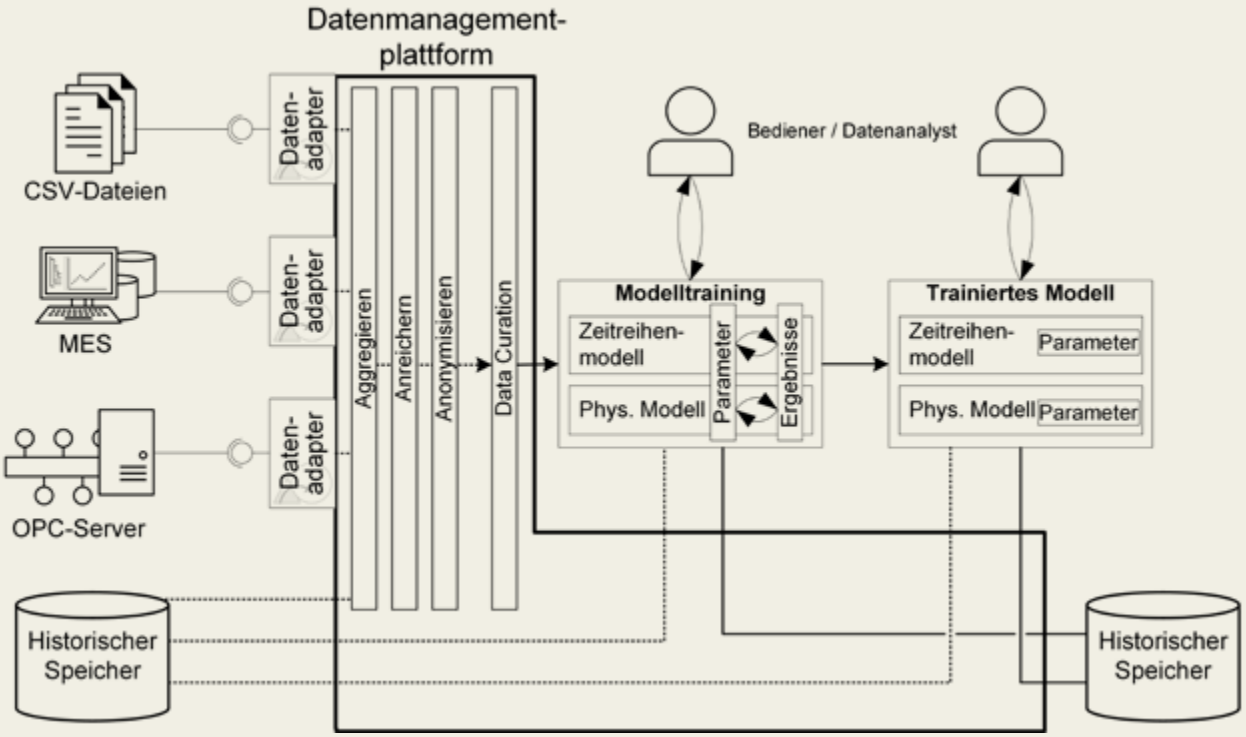

BILD 4: Datenmanagementplattform und -flussdiagramm (in Anlehnung an [5])

können ferner die Ergebnisse der Analysen wieder in den historischen Datenspeicher zurückgesichert werden. Die Datenmanagementplattform abstrahiert somit die Schnittstelle zwischen Rohdaten und der Analyse selbst.

4. EINGESETZTE METHODEN UND ERGEBNISSE DER DATENANALYSE

Für die Analyse der Daten müssen zunächst geeignete Methoden ausgewählt und angepasst werden. Nachfolgend werden anhand der verschiedenen Anwendungsfälle die verwendeten Methoden und Ergebnisse diskutiert.

4.1 Schadensfallidentifikation von Geräten am Beispiel von Ventilen

Zur Identifikation von Schadensfällen bei Ventilen wurden parallel eine modellbasierte und eine signalbasierte Methode entwickelt. Unerwünschte Anhaftungen am Ventilkegel und Kegelverschließ als häufige Schadensfälle rücken in den Fokus der zu betrachtenden Analyse. Die Modelle, sowie die Ergebnisse, innerhalb des Projektes Sidap, werden in den folgenden zwei Abschnitten erläutert. Insgesamt wurden für die Beispiele 237.072.054 Datenpunkte in die Analyse miteinbezogen. Ziel war es, durch verschiedene Modelle Anomalien im Ventilverhalten zu erkennen. Diese lassen bei längerfristigem Auftreten auf einen Ventildefekt schließen.

4.1.1 Modellbasierte / signalbasierte Methode

Die modelbasierte Methode [6] betrachtet den Hub des Ventils (h) als Eingangsgröße und den Durchfluss (F) als Ausgangsgröße des Systems. Beide Größen werden durch ein physikalisches Modell, abgeleitet von der DIN EN 60534 [8], in den relativen Durchflusskoeffizient überführt. Zur Berechnung der Soll-Ist-Abweichung wird der Soll-Durchflusskoeffizient durch das Ventilmodell und der Ist-Durchflusskoeffizient nach dem Ventil verglichen. Diese Abweichung lässt sich dazu verwenden, einen Schadensfall zu erkennen. Im Falle eines im Vergleich zum theoretischen Wert zu großen Durchflusses kann von einem Kegelverschleiß ausgegangen werden. Es vergrößert sich die freie Durchflussquerschnittsfläche im Ventil mit der Folge eines erhöhten Volumenstroms. Andererseits kann durch Fouling innerhalb des Ventils der gemessene Durchfluss unter den berechneten fallen. Durch diese Charakteristika ist es somit möglich, Ventilverschleiß und Fouling zu detektieren. Bei der signalbasierten Methode werden

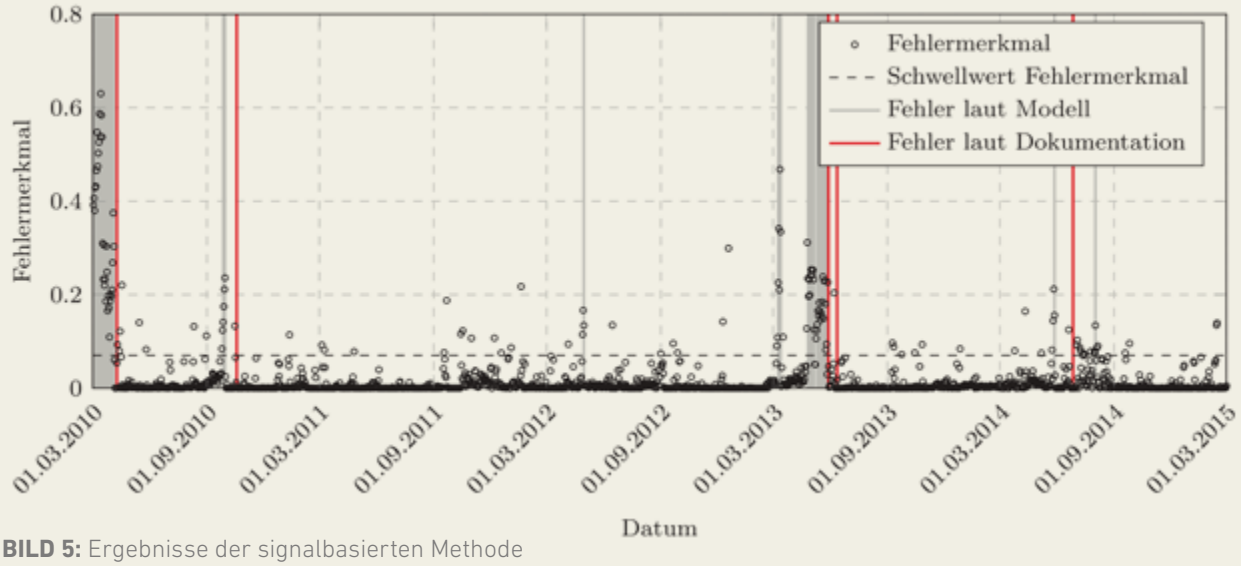

BILD 5: Ergebnisse der signalbasierten Methode

der Ventilhub und der Druck am Ausgang des Ventils in die Analyse mit einbezogen. Aus dem vorliegenden Datensatz werden Zeitreihenmodelle generiert, um das zeitliche Verhalten der Messgrößen zu beschreiben. Die zeitabhängigen Veränderungen der Modellkoeffizienten, und damit des Ventilverhaltens, werden herangezogen, um Schadensfälle zu erkennen. Zur Zeitreihenanalyse werden Autoregressive Moving Average (Arma) Modelle sowie Vektor Autoregressive (VAR) Modelle eingesetzt,

4.1.2 Ergebnisse der Datenanalyse (inklusive Datenbeschreibung)

Für die Validierung der Methoden steht ein Datensatz mit historischen Prozessdaten einer automatisierten prozesstechnischen Anlage über fünf Jahre zur Verfügung, der aus den Daten des laufenden Produktionsbetriebs des Industriepartners generiert wurde. Dieser beinhaltet Werte zu den Messgrößen Ventilhub, Druck und Durchfluss. Zusätzlich wurden aufgetretene Fehler des Ventils innerhalb des Aufzeichnungszeitraumes mit Zeitstempel und Beschreibung dokumentiert.

Der vorliegende Datensatz wurde in einen Trainings- und einen Testdatensatz unterteilt. Durch die Optimierung der Modellparameter unter Verwendung des Trainingsdatensatzes wurden die Modelle an das zu untersuchende Ventilverhalten angepasst. Mittels Testdatensatz wurden die Modelle anschließend validiert und das Ergebnis anhand der Receiver Operating Characteristics evaluiert. Auf der einen Seite konnte bei beiden Methoden ein hoher Anteil der Fehler richtig erkannt werden. Auf der anderen Seite wurden jedoch auch einige Fehlalarme (False Positives) durch die Modelle ausgegeben, vergleiche Bild 5. In weiteren Arbeiten werden die bestehenden Ansätze durch den Einbezug zusätzlich generierter Metainformation verfeinert und darüber hinaus zusätzliche Analysemethoden angewandt werden.

4.2 Identifikation kritischer Situationen

Kritische Zustände in prozesstechnischen Anlagen lassen sich durch verschiedene Verfahren feststellen. Im Folgenden wird der Ansatz im Rahmen des Projekts FEE beschrieben, bei dem Anomalien durch die Analyse der Prozessdaten und Alarmmuster beschrieben werden.

4.2.1 Anomalieerkennung anhand von numerischen Prozessdaten

Ziel der Anomalieerkennung ist es, vom normalen Verhalten abweichende Situationen zu erkennen. Nach der Definition von Hawkings [9] liegt ein solches Verhalten vor, wenn eine Beobachtung so sehr von anderen Beobachtungen abweicht, dass die Vermutung entsteht, dass sie von einem anderen Prozess erzeugt wurde. Im Kontext von Produktionsanlagen ist die Herausforderung, trotz sich ändernder Anlagenzustände, ein verlässliches Detektionsverfahren zu erstellen. Für die Bewertung, ob eine Anomalie vorliegt, kann eine Beobachtung entweder mit einem zuvor als normal klassifizierten Soll-Zustand verglichen werden, oder es wird das Verhalten mit der lokalen Nachbarschaft (k-nearest neighbour) verglichen [12]. Beide Verfahren liefern einen numerischen Indikator (anomaly score), der angibt, wie abweichend die jeweilige Situation ist.

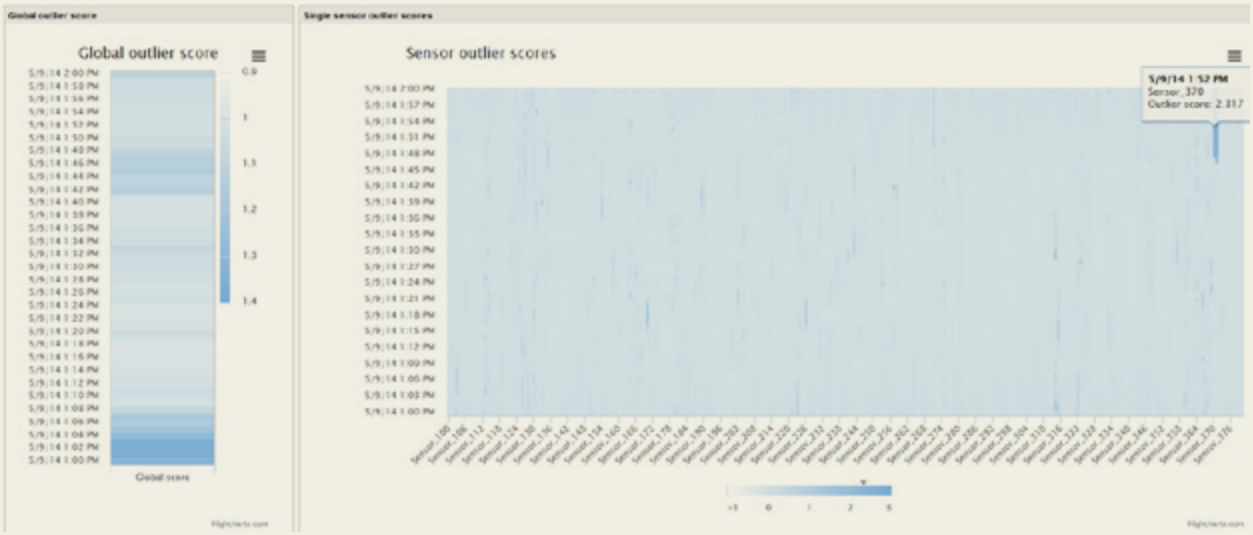

BILD 6: Anomaly-Scores über die Zeit für Gesamtanlage (links) und auffällige Sensoren (rechts)

Bei der Vielzahl der eingehenden Signale ist eine manuelle Überwachung aller Indikatorvariablen nicht mehr möglich.

4.2.2 Alarmmuster

Als zusätzliches Verfahren zur Analyse von Anomalien wurde das HypGraphs-Verfahren [10] entwickelt, das Sequenzen von Alarmen zwischen Anlagenteilen als Markov-Kette erster Ordnung modelliert. Mit Hilfe eines datengetriebenen Modells des Normalzustands werden Anomaliehypothesen überprüft und mit Hilfe eines Bayes'schen Verfahrens gewichtet. Die Evidenzwerte erlauben einen einfachen Vergleich der Hypothesen untereinander. Starke Abweichungen der Markov-Kette gegenüber dem Normalverhalten liefern Hinweise auf kritische Alarmsequenzen und können auch als Anomalie-Score dargestellt werden.

4.2.3 Ergebnisse der Datenanalyse

Die Algorithmen wurden für jeweils eine Anlage der Anwendungspartner über einen Zeitraum von mehreren Monaten erstellt. Die Anomalie-Scores werden dabei als Heatmaps über einen wählbaren Zeitraum und eine Auswahl an Sensoren (Teilmenge der Anlage) aufgetragen. Insbesondere die Darstellung großer Mengen an Sensoren über einen langen Zeitraum ermöglicht so einen Überblick über den Zustand der Anlage (siehe Bild 6). Neben dem Wechsel der Lastzustände werden ferner andere Situationen dargestellt, die größere Teile der Anlage betroffen haben. Mittels geeigneter Sortierung kann im Anschluss auf Wirkzusammenhänge geschlossen werden, wobei sich insbesondere die Sortierung von allen Elementen mit Anomaly-Scores über einer bestimmten Grenze als hilfreich erwiesen hat. Eine quantitative Bewertung der Ergebnisse ist schwierig, da der Begriff einer Anomalie nicht eindeutig definiert ist und im Zweifel auch bislang unbekannte Ereignisse erkannt werden sollen. In den beobachteten Zeiträumen zeigte sich aber, dass auffällige Ereignisse, wie zum Beispiel Lastwechsel oder ein Aufschwingen der Anlage, einen hohen Anomaly-Score aufzeigen. Insbesondere die Verknüpfung von hohen Anomaly-Scores mit den korrespondierenden Messwerten, erwies sich für die Beurteilung des Anlagenverhaltens als sehr nützlich.

5. HERAUSFORDERUNGEN FÜR DIE UMSETZUNG

Eine Vielzahl an gemeinsamen Herausforderungen wurde im Rahmen der Projekte identifiziert. Diese betreffen unter anderem Probleme bei der Datenintegration, der Definition von Systemschnittstellen und Unvollständigkeit der Daten und Dokumentation, sowie den Umgang mit Fehlalarmen und Anlagenevolution. Für viele dieser Herausforderungen können derzeit im Rahmen der Projekte nur Lösungsansätze, aber keine konkreten Lösungen gegeben werden.

5.1 Integrationsprobleme / semantische Lücken

Obwohl die Forderung nach einer Integration von Anlagendaten nicht neu ist, zeichnet sich diese Aufgabe in der Realität durch erhebliche Herausforderungen aus.

Beispielsweise weisen die Daten, die einem Equipment zugeordnet werden müssen, in unterschiedlichen Datenbanken verschiedene Bezeichnungen oder Einheiten auf. Zudem unterscheiden sich die Abtastraten der einzelnen Signale und müssen vor der Analyse auf eine gemeinsame Referenzzeit übertragen werden (Zeitsynchronisation). Änderungskomprimierte Signale, vor allem Sollwerte, die nur bei Änderung übertragen werden, erfordern erweitertes Wissen beim Auffüllen der Datenreihen. Zudem ist häufig spezifische Information für die Auswertung von Daten in Kommentarfeldern versteckt (zum Beispiel Soll- versus Istwerte), die eine automatische Integration erschweren. Lösungsansätze bieten die Verknüpfung der Rohdaten mit zusätzlicher Metainformation und ein durchgängiges Datenmodell zur Datenrepräsentation. Da diese aber je nach Anwendungsfall zu konzipieren sind, ist ihre Realisierung komplex.

5.2 Systemschnittstellen

Die Heterogenität der einzelnen Daten drückt sich ebenso in der Gestaltung der Schnittstellen der Systeme aus. Diese gehen üblicherweise von unterschiedlichen Datenmodellen aus. Einige der Probleme sind die Modellierung in unterschiedlichem Detaillierungsgrad und die Nutzung von verschiedenen Größeneinheiten. Darüber hinaus ist die Erfassung von relevanter Information, etwa wichtige Daten aus Stellungsreglern, nicht immer ohne Mehraufwand möglich, weil die dafür notwendige Kommunikation entweder nicht konfiguriert wurde oder von der Datenübertragungsrate zu langsam ist (zum Beispiel bei Hart). Des Weiteren stellt sich die Frage nach der optimalen Datenhaltung, die je nach konkretem Anwendungsfall unterschiedlich beantwortet werden muss. Einerseits kann es sinnvoll sein, eine zentrale Datenhaltung zu etablieren, um den Zeitbedarf für eine Abfrage großer Datenmengen möglichst klein zu halten. Andererseits weist diese Form der zentralisierten Datenhaltung den Nachteil auf, dass der eigentliche Eigentümer der Daten nicht mehr im Besitz dieser ist. Zur Überwindung dieser Problemstellung müssen die Daten entweder ausreichend gesichert und anonymisiert auf dem zentralen Speicher abgelegt, oder aber dezentral bei den jeweiligen Eigentümern gespeichert werden. Dies kann weiterhin die zu übertragenden Datenmengen je nach Anwendungsfall verringern, da nicht ständig ein kontinuierlicher Datenstrom zwischen den einzelnen Datenbanken übertragen werden muss, sondern nur gezielt für die Analyse benötigte Daten abgefragt werden. Auch hier bildet ein gemeinsames Datenverständnis die Grundlage zur Überwindung der Hindernisse.

5.3 Datenvollständigkeit

Die Instrumentierung prozesstechnischer Anlagen soll oftmals nur den einwandfreien Betrieb sicherstellen und wurde nicht für die Datenanalyse ausgelegt, es fehlt beispielsweise die Temperatur des Mediums. Die Messdatenarchivierung wird meist durch die Dokumentationspflicht des Betreibers bestimmt, sodass Datenpunkte aggregiert und nicht separat gespeichert werden, die für eine nachgelagerte Datenanalyse aber hilfreich wären. Die gespeicherten Datensätze sind oftmals nicht mit dem Betriebszustand der Anlage zum Zeitpunkt der Messung verknüpft. Mittels einer vorgelagerten Datenvorverarbeitung sind deshalb Abtastraten zu synchronisieren und instationäre Betriebsphasen sowie Zeiten mit fehlerhafter Messdatenerfassung aus den Rohdaten zu entfernen. Weiterhin können durch den Einbezug von Prozesswissen zusätzlich Datenreihen vervollständigt oder ergänzt werden.

5.4 Dokumentationslücken

Für die Anwendung von Big-Data-Ansätzen ist digital vorliegende Information eine Voraussetzung. Selbst wenn Formulare bereits digitalisiert sind, häufig bei Schichtbüchern, und nicht auf Papier (Aufnahme von Schadensfällen in Werkstätten) vorliegen, führen Freitexteingaben zu unterschiedlichen Einträgen für denselben Fehler (vergleiche „Ventil klemmt“ und „Ventil fährt nicht auf“). Die Informationsextraktion aus solchen unstrukturierten Daten ist eine große Herausforderung. Neben der Entitätsauflösung ist vor allem das Matching der Entitäten mit Domänenbegriffen und deren Synonymen schwierig. Bei nicht digital vorgehaltenen Planungsdaten ist zudem ein Re-engineering mit Wissensextraktion aus Bildern (P&IDs in PDF) notwendig. Hier besteht Bedarf für weitere Forschung und eine zunehmende Digitalisierung der Prozesse. Neben der in Abschnitt 5.1 diskutierten semantischen Lücke ist der Dokumentationsgrad unternehmens- und personenabhängig.

5.5 Anlagenevolution (Änderungen, Umbauten)

Eine weitere Problematik liegt in der stetigen Änderung von prozesstechnischen Anlagen über ihre Lebensdauer. So werden bei planmäßigen Stillständen Komponenten ausgewechselt und modernisiert sowie Prozessführungsstrategien im Zuge einer kontinuierlichen Optimierung angepasst. Damit verändert sich das Verhalten einer Anlage, und die angelernten Modelle für die Datenanalytik verlieren über die Zeit an Wert. Nichtsdestotrotz lassen sich zu einem gewissen Teil veraltete Modelle weiter nutzen, da sich einerseits nicht das komplette Verhalten einer Anlage ändert und andererseits, weil direkt nach einem Umbau oder einer Umstellung noch keine Daten für das Anlernen neuer Modelle zur Verfügung stehen. So ist das ständige Anlernen und Weiterlernen von Big-Data-Modellen ein fester Teil im Lebenszyklus von Anlagen und Geräten und muss als Prozess gelebt werden.

5.6 Fehlalarme

Die Güte einer Analyse kann nicht nur durch die korrekt erkannten Fehler- oder Ausnahmezustände charakterisiert werden. Ebenso bedeutend ist die Rate an Fehlalarmen bei der das Modell eine Meldung ausgibt, obwohl keine kritische Situation oder ein Fehler in der Anlage vorliegen. Fehlalarme werden beispielsweise durch instationäre Betriebsbedingungen hervorgerufen, bei denen das Modell eine Anomalie zum gelernten, normalen Anlagenverhalten feststellt. Die Berücksichtigung des aktuellen Betriebszustands kann daher die Rate an Fehlalarmen senken. Fehlalarme schädigen zum einen das Vertrauen in die Ergebnisse der Datenanalytik und können andererseits konkret zu unnötigen Maßnahmen veranlassen, zum Beispiel das Herunterfahren einer Anlage zur Fehlerbeseitigung, was einen vermeidbaren Verfügbarkeitsausfall der Produktionsanlage zur Folge hat. Je nach Anwendungsfall und Zielsetzung ist deshalb abzuwägen, wie das Modell angepasst werden muss.

ZUSAMMENFASSUNG UND AUSBLICK

Ansätze aus dem Bereich Big Data und Smart Data in der Prozessindustrie entwickeln sich zu einem nützlichen Instrument für die Unterstützung während des Betriebs. Es gibt eine große Bandbreite an Anwendungsmöglichkeiten, die in Projekten wie Sidap und FEE evaluiert werden. Die mathematischen Algorithmen existieren bereits für eine Vielzahl der identifizierten Problemstellungen. Jedoch sind Out-Of-The-Box-Versprechungen für Big-Data-Lösungen in der Prozessindustrie selten möglich. Als problematisch erweist sich die Integration und Vorverarbeitung von Daten sowie

REFERENZEN

[1] Vogel-Heuser, B., Schütz, D. und Folmer, J. (2015). Criteria-based Alarm Flood Pattern Recognition using Historical Data from Automated Production Systems (aPS). *Mechatronics,* 31, S. 89-100

[2] Atzmueller, M., Klöpper, B., Mawla, H. A., Jäschke, B., Hollender, M., Graube, M., Arnu, D., Schmidt, A., Heinze, S., Schorer, L., Kroll, A., Stumme, G. und Urbas, L. (2016). Big data analytics for proactive industrial decision support. *atp edition,* 58(9), S. 62-74

[3] Wiczorek, R. und Manzey, D. (2010). Is Operators' Compliance with Alarm Systems a Product of Rational Consideration? In: Proc. HFES 54, Santa Monica: Human Factors and Ergonomics Society, S. 1722-1726. Verfügbar unter: http://journals.sagepub.com/doi/abs/10.1177/1071181311551061

[4] McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J. und Barton, D. (2012). Big Data. The management Revolution. *Harvard Bus Review, 90*(10), S. 61-67

[5] Trunzer, E., Kirchen, I., Folmer, J. und Vogel-Heuser, B. (2017). A Flexible Architecture for Data Mining from Heterogeneous Data Sources in Automated Production Systems. In: Proceedings 18th IEEE International Conference on Industrial Technology (ICIT), IEEE

[6] Folmer, J., Schrüfer, C., Fuchs, J., Vermum, C. und Vogel-Heuser, B. (2016). Data-Driven Valve Diagnosis to Increase the Overall Equipment Effectiveness in Process Industry. In: Proceedings 14th IEEE International Conference on Industrial Informatics (INDIN), IEEE. Verfügbar unter: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7800953

[7] Nemati, H. R.; Steiger, D. M.; Iyer, L. S.; Herschel, R. T. (2002). Knowledge warehouse. An architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. *Decision Support Systems,* 33(2), S. 143–161.

[8] DIN EN 60534-2-1 (2012).Industrial-process control valves - Part 2-1: Flow capacity - Sizing equations for fluid flow under installed conditions. http://www.beuth.de

[9] Hawkins, D. (1980). Identification of Outliers. London: Chapman and Hall

[10] Atzmueller, M., Schmidt, A., Kloepper, B. und Arnu, D. (2016). HypGraphs: An Approach for Modeling and Comparing Graph-Based and Sequential Hypotheses. In: Proc. ECML-PKDD Workshop on New Frontiers in Mining Complex Patterns (NFMCP), Springer

[11] Runde, S., Fay, A., Schmitz, S. und Epple, U. (2011). Wissensbasierte Systeme im Engineering der Automatisierungstechnik. *at – Automatisierungstechnik*, 59(1)

[12] Jäschke, B. und Kroll, A. (2016). Ein Nächste-Nachbarn-Ansatz zur Anomaliedetektion bei Massendaten aus kontinuierlich betriebenen Chemieanlagen. In: Proceedings 26. Workshop Computational Intelligence, KIT Scientific Publishing

[13] Pfeffer, J., Graube, M. und Urbas, L. (2014). Interaktion mit Big Data in Industriellen Wertschöpfungsnetzwerken. In: VDI-Berichte. Bd. 2222, VDI

[14] Graube, M., Pfeffer, J., Ziegler, J., und Urbas, L. (2012). Linked Data as Integrating Technology for Industrial Data. *International Journal of Distributed Systems and Technologies,* 3(3), S. 40–52

DANKSAGUNGEN

Wir bedanken uns sowohl beim BMWi für die Förderung des Projekts SIDAP (Förderkennzeichen 01MD15009F) und beim Bundesministerium für Bildung und Forschung für die Förderung des Projekts FEE (Förderkennzeichen 01IS14006), als auch bei den Kooperationspartnern Bayer, Evonik, IBM, Samson, Uni Kassel, ABB, BASF, Ineos und PCK für die Unterstützung. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

die Notwendigkeit projektspezifischer Aufwände. Der kontinuierliche Aufbau von Expertise in Bezug auf Datenintegration, -analyse und -auswertung bei den Betreibern ist deshalb zwingend erforderlich, um Big-Data-Ansätze in der Prozessindustrie sinnvoll zu etablieren. Weiterhin müssen Lösungsanbieter die abweichenden Rahmenbedingungen und Problemstellungen berücksichtigen, um in anderen Bereichen etablierte Methoden auf die Prozessindustrie zu übertragen.

MANUSKRIPTEINGANG
23.12.2016
Im Peer-Review-Verfahren begutachtet

AUTOREN

JENS FOLMER (geb. 1981), M.Sc. angewandte/technische Informatik, ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Automatisierung und Informationssysteme der TU München. Er ist Mitglied des Leitungskreises des Lehrstuhls und leitet die Forschungsgruppe „Big Data in automatisierten Produktionssystemen“. Die Themenschwerpunkte orientieren sich an den Technologien „Industrie 4.0“ und „Cyberphysikalische-Produktionssysteme (CPPS)“.

Technische Universität München,
Fakultät für Maschinenwesen,
Lehrstuhl für Automatisierung und Informationssysteme,
Boltzmannstraße 15, 85748 Garching bei München,
TEL. +49 (0) 89 289 164 27,
folmer@ais.mw.tum.de

IRIS KIRCHEN (geb. 1989), M.Sc. Wirtschaft mit Technologie, ist wissenschaftliche Mitarbeiterin am Lehrstuhl für Automatisierung und Informationssysteme der TU München. Ihr Tätigkeitsfeld umfasst die datengetriebene Analyse von Prozessdaten aus automatisierten Produktionssystemen zur Unterstützung des Anlagenbetreibers.

EMANUEL TRUNZER (geb. 1990), M.Sc. Chemieingenieurwesen, ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Automatisierung und Informationssysteme der TU München. Sein Forschungsinteresse gilt der Integration von Expertenwissen in den Datenanalyseprozess und dem Datenmanagement.

Prof. Dr.-Ing. **BIRGIT VOGEL-HEUSER** (geb. 1961) leitet als Ordinaria den Lehrstuhl für Automatisierung und Informationssysteme (ehemals: Lehrstuhl für Informationstechnik) der TU München. Sie forscht an der Entwicklung und Systemevolution verteilter intelligenter eingebetteter Systeme in mechatronischen Produkten und Produktionsanlagen. Gleichzeitig ist sie Sprecherin des Sonderforschungsbereiches (SFB) 768 „Zyklenmanagement von Innovationsprozessen“ und Mitinitiatorin sowie Mitglied des Coordination Boards vom Schwerpunktprogramm (SPP) 1593 „Design For Future – Managed Software Evolution“.

MARKUS GRAUBE (geb. 1985), Dipl.-Ing. Mechatronik, arbeitet am Lehrstuhl für Prozessleittechnik der Technischen Universität Dresden. Seine Forschungsinteressen umfassen die semantische Informationsmodellierung und die Mensch-Maschine-Interaktion.

SEBASTIAN HEINZE (geb. 1990), Dipl.-Ing. Informationssystemtechnik, arbeitet seit 2015 als wissenschaftlicher Mitarbeiter am Lehrstuhl für Prozessleittechnik der Technischen Universität Dresden. Sein Forschungsgebiet umfasst die Förderung von Kollaboration im industriellen Umfeld mittels neuen Interaktionstechniken.

Prof. Dr.-Ing. **LEON URBAS** (geb. 1965) ist Inhaber der Professur für Prozessleittechnik und Leiter der Arbeitsgruppe Systemverfahrenstechnik an der Technischen Universität Dresden. Er beschäftigt sich mit der digitalen Transformation in der Prozessindustrie.

MARTIN ATZMUELLER (geb. 1976) ist Assistant Professor an der Universität Tilburg, Visiting Professor an der Universität Paris und Privatdozent an der Universität Kassel. Seine Forschungsgebiete umfassen Themen wie Data Science, Data Mining, Netzwerkanalyse, ubiquitäre soziale Medien sowie Big Data.

DAVID ARNU (geb. 1983), M. Sc. Statistik und maschinelles Lernen, arbeitet als Data Scientist in der Forschungsabteilung der RapidMiner GmbH. Seine Forschungsinteressen umfassen die prädiktive Analytik und Big Data.