Konzept eines Modells zur ganzheitlichen Datenschutzbetrachtung unter Anwendung von Data Mining

Optimierung des Vorgehens zur Dokumentation personenbezogener Daten und Verarbeitungstätigkeiten durch den Einsatz innovativer Technologien

Can Gümüs[1], Wolfgang Köhler[2], Prof. Dr. Christian Schultz[3] und Prof. Dr. Christoph Rasche[4]

Abstract: Während die Digitalisierung weiter voranschreitet und immer größere Datenmengen verarbeitet werden, müssen zeitgleich steigende gesetzliche Anforderungen im Umgang mit Daten, insbesondere zum Schutz der Rechte und Freiheiten natürlicher Personen beachtet werden. Um die gesetzliche Konformität von Datenverarbeitungsprozessen sicherzustellen, sind Organisationen in der Pflicht, Transparenz über Verfahren zur Erfassung und Verarbeitung personenbezogener Daten herzustellen. Unternehmen greifen zunehmend auf innovative Datenanalytik-Technologien zurück, um Analysen großer Datenmengen durchführen zu können und Muster von oder Verbindungen zwischen Daten zu erkennen. Der Beitrag nimmt sich der Optimierung des Dokumentations- und Aktualisierungsprozesses von Verarbeitungstätigkeiten an und befasst sich mit der Entwicklung des Cross-Industry Standard Process for Data Mining Modells (CRISP-DM) zur Wahrung der Konformität durch den Einsatz von Data Mining.

Keywords: Datenschutz; Personenbezogene Daten; Verarbeitungstätigkeiten; Digitalisierung; Data Mining; Künstliche Intelligenz; Cross-Industry Standard Process for Data Mining

1 Einführung

Der fortschreitende Wandel hin zu digitalen Geschäftsmodellen und Arbeitsprozessen macht die Erhebung stetig wachsender Datenmengen notwendig. Gleichzeitig treten immer mehr Gesetze zum Schutz der Privatsphäre natürlicher Personen und deren personenbezogenen Daten in Kraft. Organisationen sehen sich zunehmend damit konfrontiert, detaillierte Informationen über Praktiken der Datenerfassung und -verarbeitung zu dokumentieren und Transparenz hinsichtlich interner Verfahren im Umgang mit Verarbeitungsprozessen sicherzustellen. So sind etwa Umfang und Zweck einer Verarbeitung vor Beginn der Datenverarbeitung gegenüber betroffenen Personen offenzulegen (Art. 13

[1] Burgstraße 29, 72213 Altensteig, can.guemues@gmx.de
[2] Universität Potsdam, Management, Professional Services & Sportökonomie, Karl-Liebknecht-Str. 24-25 14476 Potsdam, wkoehler@uni-potsdam.de
[3] VICTORIA Internationale Hochschule, Business Studies, Bernburger Straße 24/25, 10963 Berlin, christian.schultz@victoria-hochschule.de
[4] Universität Potsdam, Management, Professional Services & Sportökonomie, Karl-Liebknecht-Str. 24-25 14476 Potsdam, chrasche@uni-potsdam.de

DSGVO). Verarbeitungsprozesse sind in einem zentralen Verzeichnis der Verarbeitungstätigkeiten (VVT) festzuhalten (Art. 30 Abs. 1 DSGVO). Das Verzeichnis kann zu jeder Zeit von Datenschutzbehörden angefordert werden, wobei die Nachweispflicht dem Unternehmen obliegt (Art. 30 Abs. 4 DSGVO). Ausgenommen von den genannten Pflichten sind Organisationen mit weniger als 250 Beschäftigten – unter der Voraussetzung, dass durch die Verarbeitung kein Risiko für die Rechte und Freiheiten der Betroffenen besteht, die Verarbeitung nur gelegentlich erfolgt oder keine besonderen Datenkategorien gemäß Artikel 9 Absatz 1 oder Artikel 10 DSGVO verarbeitet werden (Art. 30 Abs. 5 DSGVO).

Die Praxis zeigt, dass insbesondere große, international operierende Organisationen die Umsetzung von Dokumentations- und Aktualisierungsanforderungen zu Verarbeitungstätigkeiten häufig auf Bereichs- oder Abteilungsebene herunterbrechen [KRS20]. Eine grundlegende Herausforderung betrifft die Sicherstellung von Vollständigkeit und Aktualität des zentralen VVT. Mangels fehlender Transparenz ist für viele Unternehmen die Gesamtheit der existierenden Verarbeitungsprozesse unbekannt. Somit sind weder die Ganzheitlichkeit noch der Erfüllungsgrad gesetzlicher Anforderungen in vollem Umfang überprüfbar. [KRS20]. Es fehlt eine Übersicht aller verarbeiteten personenbezogenen Daten, damit Verantwortliche der Datenverarbeitung die rechtskonforme Umsetzung datenschutzrechtlicher Vorgaben prüfen können. Überdies existieren meist keine einheitlichen Standards zur Dokumentation von Verarbeitungstätigkeiten. Daher sind Verarbeitungsprozesse oftmals heterogen organisiert und unvollständig oder fehlerhaft dokumentiert. Eine weitere Problemstellung resultiert aus der Dynamik und Vielfalt von Verarbeitungsprozessen. Während einige Verfahren beständig sind, befinden sich andere in einem ständigen Wandel. Ein regelmäßiger Turnus zur Aktualitäts- und Konformitätsprüfung ist daher nur bedingt geeignet [KRS20]. Sofern Datenkategorien verarbeitet werden, die ohne konkrete Kenntnis nicht als personenbezogene Daten identifizierbar sind, kann dies maßgeblich die Sicherstellung der Datenschutzkonformität beeinflussen [KRS20]. Werden personenbezogene Daten nicht als solche identifiziert, erfolgt auch keine Überführung und Zentralisierung der betroffenen Verarbeitungstätigkeiten im VVT. Im Falle einer behördlichen Prüfung drohen Unternehmen hohe Geldstrafen sowie Reputationsschäden.

Um den Problemstellungen entgegenzuwirken und Konformität zu gewährleisten, greifen Organisationen vermehrt auf Technologien wie etwa Data Mining zurück. Sie ermöglichen die nahtlose Analyse großer Datenmengen, transparente Visualisierungen und Überwachung von Datenströmen und lassen zusammenhängende Muster und Abhängigkeiten zwischen Daten erkennen [Ha16]. Im Fokus des vorliegenden Beitrags steht die Frage, wie Data Mining zur Einhaltung des Datenschutzes in Unternehmen entwickelt werden kann.

Ein besonderes Augenmerk liegt auf der Pflege des VVT, da dieses das zentrale Element der europäischen Datenschutzgrundverordnung (DSGVO) darstellt. Es wird geprüft, welche technologischen Anpassungen zur Lösung der Problemstellungen essentiell sind und wie der Einsatz von Data Mining effizient umgesetzt, die Komplexität der Arbeitsvorgänge verringert und die Flexibilität von Geschäftsprozessen erhöht werden kann.

2 Verwandte Arbeiten

Der folgende Abschnitt befasst sich mit einer Vorstellung bereits existierender Vorgehensmodelle zur Sicherstellung datenschutzrechtlicher Vorgaben, die auf dem Einsatz innovativer Technologien der Datenanalytik beruhen.

Im Beitrag von Becker und Buchkremer wird die Entwicklung eines agilen Vorgehensmodells erläutert, mit dessen Hilfe aufsichtsrechtliche Anforderungen durch Einsatz einer sogenannten Regulatory Technology Lösung implementierbar seien [BB18]. Die Autoren heben die Relevanz eines harmonischen Zusammenspiels zwischen Technologie und menschlichen Experten für agile Implementierungsprozesse hervor und kommen zu dem Schluss, dass iterative Vorgehen für die Analyse regulatorischer Anforderungen im Kontext des Datenschutzes erfolgsentscheidend sind.

Kittel beschreibt in einem Artikel, wie Agilität bei Geschäftsprozessen mit Datenschutzbezug sichergestellt werden kann [Ki13]. Es zeigt sich, dass Ad-hoc-Änderungen von Geschäftsprozessen dieser Art eine vorausgehende Kontrolle regulatorischer Datenschutzanforderungen unbrauchbar machen. Kittel stellt einen modellbasierten Ansatz zur Ad-hoc-Integration von Datenschutzkontrollen in Arbeitsabläufen vor, durch den die Abhängigkeiten zwischen Agilität und Compliance verringert werden sollen.

Ein weiteres Vorgehensmodell zur Vorbereitung auf datenschutzrechtliche Anforderungen wird von Wirnsperger, Buchholz und Wolff erarbeitet [BWW16]. Das Modell berücksichtigt rechtliche, technische, organisatorische und prozessuale Aspekte. Beginnend mit einer Vorprüfung und einer Umfeldanalyse zur Erfassung aller personenbezogenen Daten in Geschäftsprozessen solle der Status Quo auf Basis einer Fit-/Gap-Analyse erfasst sowie ein Risiko- und Maßnahmenplan erarbeitet werden.

Das von Chapman et. al. entwickelte CRISP-DM-Modell stellt die Entwicklung und Umsetzung spezifischer Data-Science-Projekte durch den Einsatz von Data Mining und künstlicher Intelligenz in den Mittelpunkt [Ca00]. Das Modell gilt als Standardvorgehen für die Ausführung von Data-Mining-Projekten und ist für diverse Projekte der künstlichen Intelligenz zur Sicherstellung des Datenschutzes anwendbar. Da das CRISP-DM-Modell in den Kontext des aktuellen Technologiestands eingeordnet ist, wird es als Rahmen für die vorliegende Untersuchung verwendet.

Während ein Großteil aktueller Untersuchungen den Einfluss regulatorischer Datenschutzvorgaben auf die Entwicklung intelligenter Technologien diskutieren, widmen sich einige wenige Quellen dem Unterstützungsgrad innovativer Technologien und deren Anwendungspotentialen zur Wahrung des Datenschutzes. Inwiefern Data Mining jedoch speziell bei der Verarbeitungsdokumentation und -aktualisierung in einem VVT unterstützt, wird in der Wissenschaft nicht vertieft betrachtet. Nach aktuellem Stand existiert kein Vorgehensmodell für diesen spezifischen Anwendungsfall.

3 Vorgehensmodell und methodische Unterstützung

3.1 Anforderungsanalyse

Die Anwendung einer Anforderungsanalyse hat unmittelbaren Einfluss auf die zielgerichtete Entwicklung des CRISP-DM-Modells. Zur vollständigen Ermittlung aller Anforderungen an CRISP-DM wird zunächst ein umfassender Anforderungskatalog entwickelt. Der Katalog differenziert zwischen technologischen Anforderungen, die primär die zu erbringenden Funktionalitäten, Mechanismen und Leistungen des Data Mining zur Gewährleistung der Konformität betreffen und Anforderungen von Seiten des Datenschutzrechts zur Pflege eines zentralen VVT. Letzteres orientiert sich an den Regularien der DSGVO aus Artikel 30.

3.2 CRISP-DM

CRISP-DM folgt einem iterativen Kreislauf mit insgesamt sechs Phasen, ohne dabei einen konkreten Endpunkt festzulegen. Stattdessen kann jede Phase und deren Iterationen, je nach Problemstellung mehrfach durchlaufen und ausdifferenziert werden. Jede Wiederholung des Gesamtprozesses bringt neue Fragestellungen hervor und kann zu einer Prozessoptimierung beitragen. Das Modell schreibt keine starre Sequenzierung der einzelnen Phasen vor. Rückkopplungen, die sich etwa aus unvorhergesehen Problemfaktoren oder unzureichender Qualität eines Zwischenergebnisses ergeben, sind durchaus möglich und gewünscht. [Ca00]

Phase 1. Zur Erlangung vollständiger Kenntnis über die Geschäftsanforderungen und konkrete Aufgabenstellung hat eine präzise Erörterung der betriebswirtschaftlichen Problemstellung zu erfolgen [Ca00]. Dabei sind die zu erreichenden Zielkriterien festzulegen. Diese werden in Anforderungen an die Datenanalyse überführt, woraufhin ein konkreter Umsetzungsplan unter Berücksichtigung zeitlicher, personeller und sachlicher Ressourcen aufzusetzen ist [CL16].

Phase 2. Im nächsten Schritt werden relevante Datenbestände selektiert, deren Verarbeitung zur Erfüllung der zuvor bestimmten Ziele notwendig ist [CL16]. Es wird eine Datensammlung mit Beschreibung der typischen Eigenschaften der relevanten Daten angelegt, um ein generelles Verständnis über die selektierten Daten aufzubauen. Die Phase mündet letztlich in einer Bewertung der Datenqualität und -quantität [CL16].

Phase 3. Die Datenvorbereitung zielt auf die Auswahl der finalen Datenmenge ab, die in das Data-Mining-System integriert und entlang vordefinierter, anwendungsspezifischer Algorithmen analysiert werden soll [Ca00]. Es bedarf einer klaren Differenzierung zwischen irrelevanten und relevanten Daten. Das Ergebnis der Datenauswahl hängt von der jeweiligen Zielsetzung des Data-Science-Projektes ab. Ferner sind die Daten zu bereinigen, um eine Data-Mining-Verarbeitung zu ermöglichen. Diese Phase entscheidet darüber,

welche speziellen Merkmale und Charakteristiken die nachfolgende Modellbildung berücksichtigen soll [CL16].

Phase 4. Die Modellbildung nimmt sich der eigentlichen Datenanalyse an, indem ein Modell zum Umgang mit den selektierten Daten entwickelt wird [CL16]. Nach Auswahl und Parametrisierung einer passenden Modellierungstechnik wird ein Testmodell entwickelt, mit dessen Hilfe die Präzision und Qualität des Entwicklungsergebnisses geprüft und bewertet wird. Die Algorithmen der Modellbildung unterscheiden zwischen einem *Trainieren* und *Anwenden*, wobei das Modell entweder auf Basis des gewonnenen Wissens aus historischen Daten trainiert oder auf neue, bisher unbekannte Datensätze angewendet wird.

Phase 5. Zur Evaluation des Entwicklungsergebnisses wird die eingangs festgelegte Zielsetzung mit dem erarbeiteten Data-Mining-Verfahren abgeglichen. Für den Fall, dass die gewünschte Qualität des Modells zur Erfüllung der Zielkriteren nicht vollständig oder nur in Teilen erreicht wurde, muss CRISP-DM erneut durchlaufen werden [CL16].

Phase 6. Den Abschluss bildet die Planung und Umsetzung der Implementierung des Data Mining im Unternehmen. Das Modell kann je nach Anwendungsfall auf existierende oder auf neue, bislang unbekannte Datenbestände angewendet werden [Ca00].

4 Konzipierung des CRISP-DM-Modells

4.1 Anforderungsspezifizierung

Technologische Anforderungen. Tabelle 1 zeigt einen Überblick der technologischen Anforderungen an das Data-Mining-System.

Technologische Anforderungen
Zugriff auf den gesamten Datenpool des Unternehmens
Erschließen aller im Unternehmen verfügbaren, (un-)strukturierten Daten
Identifikation und Strukturierung aller verarbeiteten personenbezogenen Daten
Erfassung aller existierenden Verarbeitungsprozesse
Zentrale Steuerung der Pflege eines VVT
Vollständige Dokumentation aller Verarbeitungsprozesse im VVT
Gewährleistung kontinuierlicher Aktualität des VVT
Automatische Anpassung und Aktualisierung von Verarbeitungsprozessen
Erkennen von Trends, Veränderungen und datenschutzrechtlichen Anforderungen

Tab. 1: Technologische Anforderungen an das Data-Mining-System

Datenschutzrechtliche Anforderungen. Tabelle 2 zeigt einen Überblick der datenschutzrechtlichen Anforderungen an die Dokumentation von Verarbeitungstätigkeiten in

einem VVT gemäß Artikel 30 Absatz 1 DSGVO. Gleiches gilt für Auftragsverarbeiter unter Ausschluss der Beschreibung und Kategorisierung der Verarbeitungszwecke, der Beschreibung und Kategorisierung aller Datenempfänger im In- und Ausland sowie der Löschfristen der verschiedenen Datenkategorien (Art. 30 Abs. 2 DSGVO).

Datenschutzrechtliche Anforderungen
Name und Kontaktdaten des verantwortlichen Datenverarbeiters
Beschreibung und Kategorisierung der Verarbeitungszwecken
Beschreibung des Betroffenen und Kategorisierung der betroffenen Personen
Beschreibung und Kategorisierung der personenbezogenen Daten des Betroffenen
Beschreibung und Kategorisierung aller Datenempfänger im In- und Ausland
Beschreibung der Übermittlung in Drittländer oder internationale Organisationen und deren Benennung
Löschfristen der verschiedenen Datenkategorien
Dokumentation der technischen und organisatorischen Maßnahmen

Tab. 2: Datenschutzrechtliche Anforderungen

4.2 Grundlagen

Die erforderlichen Inhalte eines VVT ergeben sich aus der Analyse manuell gepflegter Verzeichnisse in der Praxis. Dies dient der nachgelagerten Lösungssuche, indem ermittelt wird, welche Strukturierungen und Klassifizierungen der relevanten Daten das Data-Mining-System zur vollumfänglichen Dokumentation zu berücksichtigen hat. Im Kontext der CRISP-DM-Entwicklung werden schließlich Regeln, Korrelationen und Muster zwischen Daten und deren Verarbeitungstätigkeiten abgeleitet. Eine beispielhafte Übersicht des Aufbaus eines VVT und der zu dokumentierenden Inhalte und Informationen ist in Tabelle 3 gegeben.

Themenbereiche eines VVT	**Potentielle Inhalte**
Dokumentation der Kontaktdaten des verantwortlichen Daten- oder Auftragsverarbeiters	Name, Funktion, E-Mail-Adresse, Telefonnummer und Anschrift des Verantwortlichen oder Auftragsverarbeiters
Dokumentation der Verarbeitungsprozesse	Bezeichnung, Beschreibung, Datenherkunft, Verwendetes IT-System
Dokumentation des Zwecks der Datenverarbeitung	Zweckkategorie, Zweckänderung, Zweck (Mit-)Bestimmung durch Dritte
Dokumentation des datenverarbeitenden Systems	Name des datenverarbeitenden Systems
weitere	

Tab. 3: Themenbereiche und potentielle Inhalte eines VVT

4.3 Konzeption

Phase 1. Zur Generierung eines exakten Verständnisses der Aufgaben- und Problemstellung werden im ersten Schritt die erwarteten Projektziele sowie -ergebnisse festgelegt. Im vorliegenden Kontext leiten sich diese aus den technologischen und datenschutzrechtlichen Anforderungen ab. Tabelle 4 zeigt die potentielle Zielsetzung des Data-Mining-Vorhabens.

Zielformulierung
Steigerung der Transparenz
Steigerung der Effizienz bei Anpassungen an Verarbeitungsprozesse
Verfolgbarkeit bei Prozessaktivitäten
Verfügbarkeit relevanter Informationen
Reduzierung der Arbeitsauslastung von Fachbereichen eines Unternehmens
Reduzierung des Abstimmungsaufwands zwischen den Fachbereichen
Standardisierung des Vorgehens
Vollständige Identifikation und Strukturierung personenbezogener Daten
Vollständige Erfassung aller Verarbeitungsprozesse
Unterstützung der Dokumentation der Verarbeitungsprozesse im VVT
Sicherstellung kontinuierlicher Aktualität des VVT

Tab. 4: Zielformulierung

Weiterhin ist eine ausführliche Risikoanalyse durchzuführen [CL16]. Ermittelte Risiken sind gemäß projektspezifischer Kriterien zu bewerten und individuell zu analysieren. Dies kann mittels einer Risiko-Matrix erfolgen, wobei die Eintrittswahrscheinlichkeit und Schadenhöhe für jedes Risiko geschätzt und in der Risiko-Matrix visualisiert werden. Zur Reduktion besonders schwerwiegender Risiken sind Gegenmaßnahmen festzulegen. Das Risikomanagement hat über die gesamte Projektdauer hinweg zu erfolgen.

Anhand der Zielsetzung werden Erfolgskriterien zur finalen Bewertung des Entwicklungsergebnisses spezifiziert [Ca00]. Das Vorhaben ist dann erfolgreich, wenn die Gesamtheit aller personenbezogenen Daten bekannt ist und diese entlang charakteristischer Merkmale strukturiert werden. Entsprechend sind Verarbeitungstätigkeiten automatisch zu erfassen, sodass Mitarbeitende bei Zentralisierung und Aktualisierung des VVT unterstützt werden. Weitere Erfolgskriterien betreffen die Reduzierung der Fehleranfälligkeit, des manuellen Arbeitsaufwands, der Komplexität sowie der Arbeitsauslastung innerhalb der Fachbereiche und -abteilungen.

Die Festlegung von Unternehmenszielen orientiert sich vor allem an der Firmenkultur und Vision eines Unternehmens, weshalb eine allgemein gültige Aussage nur bedingt möglich ist. Nichtsdestominder spiegeln sich zumeist einige Unternehmensziele wie Integrität, Verlässlichkeit und Vertrauenswürdigkeit in verschiedenen Unternehmen wider und ge-

hen demnach mit den Grundsätzen und Schutzzielen der Datensicherheit und des Datenschutzes (Compliance) einher. Die strikte Einhaltung dieser drei Faktoren sind in der heutigen Zeit zur Wahrung der Wettbewerbsfähigkeit [He18] und Rechtmäßigkeit unabdingbar.

Die wichtigsten Fragestellungen zur Aufbereitung der ersten Phase sind nachfolgend zusammengefasst.

- Wie ist die Ausgangssituation und Problemstellung?
- Welche Ziele und Ergebnisse sollen durch Data Mining erreicht werden?
- Welche Risiken (finanziell, rechtlich, organisatorisch) können auftreten?
- Was sind die Erfolgskriterien und Unternehmensziele?
- Wie ist die aktuelle Unternehmenssituation?
- Welche Ressourcen sind zur Umsetzung des Vorhabens verfügbar?
- Welche Kosten sind für welchen Nutzen aufzubringen?
- Wurde ein Projektmanagementsystem etabliert und ein Projektplan aufgesetzt?

Phase 2. Zur Selektion der relevanten Datenbestände muss geklärt werden, was unter personenbezogenen Daten verstanden wird und wodurch sich diese kennzeichnen. Die DSGVO definiert personenbezogene Daten als „Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen“ (Art. 4 Abs. 1 DSGVO). Personen gelten dann als „identifizierbar“, wenn sie sich (in-)direkt eindeutig identifizieren lassen (Art. 4 Abs. 1 DSGVO). Es wird dann von personenbezogenen Daten gesprochen, wenn die erhobenen Daten einen direkten Bezug zu einer betroffenen Person hervorbringen.

Die Elemente und Kategorien der personenbezogenen Daten sind zu ermitteln. Dies erlaubt die Auswahl und Entwicklung eines passenden Data-Mining-Vorgehens. Datenelemente und -kategorien leiten sich aus den in Tabelle 3 dargelegten Informationen zur Dokumentation von Verarbeitungstätigkeiten ab. Die Erkenntnisse dienen der Entwicklung von Mustern und Regeln, die Data Mining zur vollumfänglichen Identifikation der personenbezogenen Daten, zur Ableitung und Dokumentation der resultierenden Verarbeitungstätigkeiten sowie dem Segmentieren der Elemente und Kategorien anzuwenden hat.

Verantwortliche mit weniger als 250 Beschäftigten sind gemäß Artikel 30 Absatz 5 DSGVO dazu verpflichtet, zu erheben, ob mit der Verarbeitung der personenbezogenen Daten besondere Risiken für die Rechte und Freiheiten für die Betroffenen einhergehen und inwieweit besondere Kategorien personenbezogener Daten gemäß Artikel 9 DSGVO verarbeitet werden. Weiterhin ist der Turnus der Datenverarbeitung zu bestimmen.

Phase 3. Die Datenvorbereitung gliedert sich in die Schritte Selektion und Integration, Säuberung, Reduktion und Transformation von Daten [Ca00]. Die Relevanz einer Datenselektion und -integration resultiert aus den verschiedenen Datenbanken und Quellen, aus denen Daten potentiell entstammen. Nach erfolgter Datenselektion sind die Daten in einer konsistenten Datenbasis mit schlüssigen Datensätzen zu vereinheitlichen. Probleme, die

bei der Integration auftreten können, sind unter anderem Entitäten-Identifikationsprobleme, Redundanzen, Widersprüche oder Datenwertkonflikte [CL16].

Danach ist der Datenbestand manuell zu bereinigen. Es ist darauf zu achten, dass eingefügte Werte durch Bereinigung informationsneutral sind, ohne eine Verfälschung der vorhandenen Dateninformationen herbeizuführen. Neben fehlenden Daten stellen ebenso verrauschte Daten und Ausreißer oder inkonsistente und falsche Daten mögliche Problemstellung dar, die während des Säuberungsprozesses zu unterbinden sind [CL16].

Eine Reduktion der Daten ist dann notwendig, wenn ein Datensatz zur Ausführung des Data Mining zu groß ist. Einerseits kann die Komplexität eines Datensatzes mit Hilfe einer zeilen- oder spaltenweise Aggregation [CL16] verringert werden. Mehrere Daten werden also auf Basis von charakteristischen Attributen zusammengefasst. Ein konkreter Anwendungsfall ist etwa das Clustern von Datenelementen und -kategorien durch Anwendung der spaltenweisen Aggregation. Personenbezogene Daten können so von dem Rest des Datenbestandes abgespalten und kategorisiert werden. Eine zweite Lösung bietet die Dimensionsreduktion als Vorwärtsauswahl oder Rückwärtseliminierung [CL16], indem Stichproben einer repräsentativen Teilmenge der selektierten Daten durchgeführt werden. Eine Erfassung aller personenbezogener Daten kann etwa mittels der Vorauswahl erfolgen, indem alle Daten, die keinen direkten Personenbezug aufweisen, durch sukzessive Aufnahme neuer Anforderungen gelöscht werden.

Ziel der Datentransformation ist die Überführung der Daten in eine brauchbare Form, um in das Data-Mining-System integriert werden zu können. Verfahren zur Datentransformation sind etwa Codierungen, Zeichenketten (z.B. Umlaute), Maßeinheiten und Skalierungen, Kombinationen oder Separierungen von Attributen, Berechnungen abgeleiteter Werte, Aggregationen oder Datenglättungen (z.B. Regression) [CL16].

Phase 4. Die Modellbildung des Data Mining unterscheidet zwischen Potential- und Beschreibungsaufgaben. Potentialaufgaben umfassen die Datenklassifikation und das Ableiten von Prognosen, wohingegen Beschreibungsaufgaben der Segmentierung oder dem Aufstellen von Assoziationen zwischen Datensätzen dienen [CL16]. Bei der Klassifikation erfolgt eine Zuordnung eines Datenobjekts zu einer vordefinierten Klasse entlang charakteristischer Merkmale. Die Prognose hingegen zielt auf die Entwicklung eines Bewertungsmodells zur fortlaufenden Ermittlung stetiger Werte ab. Bisher unbekannte, numerische Merkmale werden auf Basis anderer Merkmale oder erlangter Erkenntnisse vorausgesagt und Abhängigkeiten zwischen diversen Variablen hergestellt. Im Rahmen der Segmentierung wird die Gesamtheit aller Daten in Teilmengen unterteilt und mehrere Datenobjekte mit gemeinsamen Merkmalen zu einer homogenen Gruppe zusammengeführt. Im Fokus der Assoziation steht die Ermittlung und Beschreibung von Mustern zwischen Datenobjekten, die in einer bestimmten Relation zueinanderstehen. Beispiele für Data-Mining-Verfahren sind Entscheidungsbäume, Cluster-Algorithmen oder Regressionen.

Für den vorliegenden Anwendungsfall muss primär eine ganzheitliche Erfassung und Strukturierung aller personenbezogenen Daten vorgenommen werden. Verschiedene Datenkategorien sind zur einheitlichen Dokumentation im VVT zu einem einzigen Datenelement zu reduzieren. Dafür eignen sich die Klassifikation und Segmentierung.

Zu Beginn erlaubt die Klassifikation eine Kategorisierung personenbezogener Daten gemäß charakteristischer Merkmale, durch die natürliche Personen eindeutig identifizierbar sind. Um eine solche Separierung zu erreichen, müssen dem Data-Mining-System die Merkmalanforderungen bekannt sein. Die Anforderungen ergeben sich vorrangig aus der Definition personenbezogener Daten des Artikel 4 DSGVO. Beispiele für charakteristische Merkmalanforderungen zur Datenklassifikation sind bspw. Name, Anschrift, E-Mail-Adresse oder Telefonnummer.

Überdies sind die als relevant klassifizierten Datenobjekte zu segmentieren und bestehende Datenkategorien bestimmten Datenelementen zuzuweisen. Eine Möglichkeit stellt das Rule-based Reasoning [Ch20] dar, indem Regeln entlang des Wenn-Dann-Sonst-Prinzips [FPZ95] erarbeitet werden. Ein Beispiel zur Segmentierung personenbezogener Daten kann etwa über die Regel *„Wenn die Datenkategorie die Angabe Name oder Anschrift enthält, dann sind diese personenbezogenen Daten dem Element persönliche Kontaktinformationen zuzuweisen“* erfolgen. Dieses Schema muss für alle identifizierten Datenkategorien und -elemente umgesetzt werden – unter der Prämisse, dass die Möglichkeit einer Segmentierung besteht. Zur besseren Veranschaulichung ist die Vorgehensweise der Regelbildung in Abbildung 1 dargestellt.

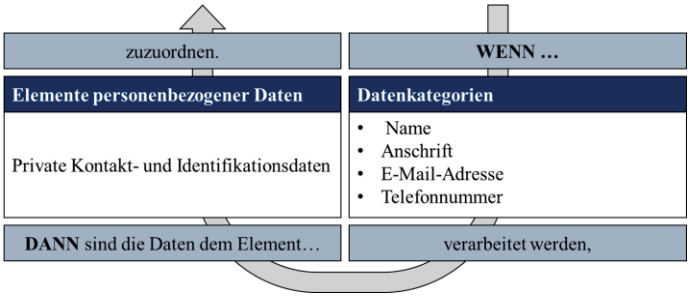

Abb. 1: Wenn-Dann-Regel zur Segmentierung personenbezogener Daten

Hinzu kommt die Notwendigkeit, auf Grundlage der identifizierten und segmentierten Daten resultierende Verarbeitungsprozesse zu erfassen, zu dokumentieren und zu aktualisieren. Hierfür kann auf die Assoziation und Prognose zurückgegriffen werden.

Mit Hilfe des Assoziationsverfahrens lassen sich Abhängigkeiten und Muster zwischen Daten und Verarbeitungstätigkeiten feststellen. Das stellt sicher, dass personenbezogene Daten, die nicht als solche kenntlich sind, identifiziert werden können. Eine Möglichkeit stellt dabei die semantische Interoperabilität [Gö10] zur Kollaboration zwischen diversen IT-Systemen mittels Klassifikationssystemen, Taxonomien oder Nomenklaturen dar. Die

semantische Interoperabilität eignet sich im Speziellen zur Erfassung und Dokumentation von Verarbeitungstätigkeiten auf Grundlage der verarbeiteten personenbezogenen Daten. Das Data-Mining-System wird dazu befähigt, Informationen mit den IT-Systemen der Organisation auszutauschen und so bspw. den Zweck einer Verarbeitung zu ermitteln. Mittels eines intelligenten und vernetzten Zusammenspiels zwischen IT-Systemen lassen sich die zentral im VVT festzuhaltenden Inhalte und Informationen detektieren und einheitlich dokumentieren. Dadurch wird ebenso eine Standardisierung der Dokumentation erreicht.

Als letzter Schritt unterstützt die Prognose dabei, Zusammenhänge zwischen bekannten und bisher unbekannten Merkmalattributen herzustellen und Trendentwicklungen zu prognostizieren. Treten etwa Änderungen in den Angaben personenbezogener Daten auf, können diese analysiert und bei dokumentierten Verarbeitungstätigkeiten aktualisiert werden. Weiterhin können Risiken aufgrund von Datenlecks oder bei Nicht-Einhalten der geltenden Datenschutzvorgaben präventiv gemeldet werden.

Es lässt sich festhalten, dass eine Kombination der vier Data-Mining-Vorgehen essentiell ist, um den geschilderten Herausforderungen und Problemstellungen zu begegnen und die identifizierten Anforderungen zu erreichen.

Phase 5. Im Kontext der Evaluation werden die Analyseergebnisse geprüft. Ob die Umsetzung erfolgreich ist, ergibt sich aus einer Ermittlung des Erfüllungsgrads der initial spezifizierten Erfolgs- und Zielkriterien. Zentrale Fragestellung ist, ob der erwünschte betriebswirtschaftliche Nutzen durch das Entwicklungsergebnis erzielt wird [CL16].

Im weiteren Verlauf ist eine Analyse der auftretenden Fehler durchzuführen, woraus sich unter Umständen weitere Optimierungspotentiale ergeben [Ca00]. Tritt dieser Umstand auf, kann in eine der vorangegangenen Phasen zurückgekehrt und so das Data-Mining-Vorhaben sukzessive verbessert werden.

Phase 6. Den Abschluss bildet die praktische Implementierung des Data-Mining-Systems. Zur optimalen Einsatzvorbereitung wird ein im Detail ausgearbeitetes und strukturiertes Vorgehen zum künftigen Monitoring des Data Mining und der resultierenden Analyseergebnisse vorgesetzt. Des Weiteren muss eine ausreichende Motivation der Mitarbeitenden der Organisation, in der das System Anwendung finden soll sowie eine umfassende Unterstützung der durch das Data Mining betroffenen Mitarbeitenden (z.B. IT-Abteilung, Datenschutzbeauftragter, Fachabteilung etc.) gegeben sein, um das Scheitern des Projekts zu verhindern. Das System ist in den Regelbetrieb der Organisation zu überführen und in laufende Prozesse einzubetten.

5 Diskussion

Die ersten drei CRISP-DM-Phasen beanspruchen etwa 50 bis 70 Prozent des Arbeitsaufwands zur Entwicklung des Data Mining, wobei die einzelnen Phasen manuell vorzubereiten und umzusetzen sind [Wu20]. Ein direkter Vergleich des Status Quo zur Pflege

eines VVT und den Phasen des CRISP-DM impliziert, dass die initialen drei CRISP-Phasen gleichermaßen im manuellen Pflegeprozess eines VVT stattfinden. Unternehmen, die bereits ein VVT pflegen, haben die Schritte im Optimalfall durchlaufen. Auch wenn zur Pflege eines VVT kein Data-Mining-System etabliert werden soll, ist es sinnvoll, die Phasen gewissenhaft umzusetzen. Organisationen sollten nach Ausführung der initialen Phasen in Erwägung ziehen, ihre Ergebnisse in KI-Algorithmen und Regeln zu überführen und die manuellen Arbeitsaufwände auf ein Data-Mining-System zu verlagern. Unternehmen können auf ihrem bisherigen Arbeitsstand aufbauen, die Inhalte entsprechend dem dargestellten Vorgehen anpassen und letztlich in ein Data-Mining-Modell überführen. Jedoch muss das CRISP-DM-Modell nicht zwangsläufig in der Implementierung eines Data-Mining-Systems münden, auch wenn dies zu einem deutlichen Anstieg der Produktivität beiträgt. Stattdessen sehen sich Organisationen aufgrund der Rechenschaftspflicht ohnehin damit konfrontiert, Transparenz hinsichtlich der Datenverarbeitungen und -flüsse sicherzustellen. Um ebendiese Transparenz zu erreichen, haben Organisationen die initialen Phasen Business Understanding, Data Understanding und Data Preparation zur Erfassung, Dokumentation und Aktualisierung aller existierenden Verarbeitungsprozesse aufzubereiten. Folglich zieht das dargestellte Vorgehen kein Mehraufwand nach sich, sondern bietet Organisationen im Gegenteil die Möglichkeit, Synergien zu nutzen und in Zukunft bedarfsorientiert auf ihrem bisherigen Arbeitsstand aufzubauen, um eine technologische Unterstützung und Optimierung des Vorgehens zur Pflege eines VVT herbeizuführen.

Zwar kann Data Mining bei der vollständigen Erfassung und Kategorisierung personenbezogener Daten, dem Ableiten von Verarbeitungstätigkeiten sowie der Dokumentation und Aktualisierung im VVT unterstützten. Jedoch sind nachgelagert weiterhin manuelle Aufwände notwendig. Bspw. verantwortet das Rechtswesen einer Organisation die Zentralisierung von Verarbeitungstätigkeiten im VVT, die Überführung neuer datenschutzrechtlicher Vorgaben in konkrete Anforderungen an Data Mining oder die konstante Überwachung der Qualität der Analyseergebnisse.

Eine weitere Handlungsempfehlung betrifft die Erweiterung des CRISP-DM-Modells um die Monitoring-Phase. Neben einer fortwährenden Wartung des Systems sind auch die Ergebnisse der Datenanalysen durch Verantwortliche der Datenschutzorganisation (z.B. Datenschutzbeauftragter etc.) zu überwachen, da diese bestens mit den rechtlichen Grundlagen vertraut sind. Es ist sicherzustellen, dass alle Mitarbeitenden, die in Zukunft Berührungspunkte mit dem System haben, umfassend geschult werden.

Darüber hinaus ist ein Datensicherheitskonzept gemäß Artikel 32 DSGVO zu entwickeln, um softwareseitige Störungen und Systemausfällen vorzubeugen. Im Falle eines Absturzes wird etwa der Zugriff auf das zentrale Verzeichnis verweigert. Daher müssen in regelmäßigen Abständen automatische Backups des VVT durchgeführt werden und Mitarbeitende der Organisation dafür Sorge tragen, das System durch Schutzmaßnahmen (z.B. technisch-organisatorische-Maßnahmen) abzusichern.

Data Mining erlaubt eine ganzheitliche, transparente und zentrale Steuerung des VVT-Prozesses, sodass ein technologischer Einsatz zur Wahrung der Konformität geeignet ist.

Durch Einhalten der Vorgaben des Artikel 30 DSGVO wird das Haftungsrisiko wesentlich reduziert und der aktuell gelebte Prozess in der Praxis flexibilisiert und vereinfacht. Die ersten vier Phasen des Modells können als Basis für vergleichbare Aufgabenstellungen genutzt und spezifiziert werden.

Nichtsdestominder zeigt sich der Unterstützungsgrad des Data Mining erst durch Entwicklung entlang einer realen Aufgabenstellung und tatsächlichen Implementierung des Systems. Darüber hinaus ist die Modellentwicklung auf Annahmen und Theorien gestützt. Die CRISP-DM-Phasen wurden zwar detailliert ausgearbeitet, jedoch ohne die technologische Ebene vertieft zu betrachten. Dies ist unter anderem der Tatsache geschuldet, dass kein Testmodell unter realen Umständen entwickelt und zu Testzwecken implementiert wurden. Eine Aussage über die tatsächliche Um- und Einsetzbarkeit des Data-Mining-Systems kann somit nicht getroffen werden. Grundsätzlich ist die technische Sicht bei der Entwicklung des Vorgehensmodells unterrepräsentiert. Es kann vorkommen, dass Probleme, die in der Anwendung und Programmierung des Data-Mining-Systems auftreten, nicht vollständig erkannt und berücksichtigt wurden. Daher wird empfohlen, das System in der Praxis zunächst umfassend zu testen und mit erfolgreichem Abschluss der Testphase auf weitere Bereiche des Unternehmens auszuweiten. Ferner können in Zukunft vor- und nachgelagerte Teilprozesse oder weitere Anforderungen der DSGVO berücksichtigt und auf ähnliche Weise intelligent gesteuert und optimiert werden.

Es bleibt zu erwähnen, dass das CRISP-DM-Modell einem Standardmodell entspricht, welches der Standardisierung diverser Anwendungsfälle dient und damit nicht nur auf die Pflege eines VVT begrenzt ist. Stattdessen kann das CRISP-DM-Modell beliebig erweitert und ebenso auf andere Ausgangssituationen übertragen werden. Das Modell kann demnach genau wie VVT in der Praxis vielfältig ausfallen. Aus diesem Grund findet im Rahmen des vorliegenden Beitrags eine generische Darstellung und Entwicklung des Modells Anwendung. Auf die Beschreibung einzelner, konkreter Anwendungsfälle wird in diesem Zusammenhang bewusst verzichtet.

Als weiterer Forschungsbedarf kann das Process Mining und dessen Kombination mit Data Mining betrachtet werden. Process Mining vereint die Vorteile des Data Mining mit denen der Prozessmodellierung, sodass eine effiziente Überwachung und Erstellung von komplexen Echtzeitprozessen möglich ist [Re20]. Durch Standardisierung von Prozessen kann einerseits die Transparenz erhöht werden und damit Schwachstellen der aktuellen Prozessumsetzung effizient geprüft und bei Bedarf verbessert werden [NP19]. Andererseits werden durch Prozessautomatisierungen Redundanzen reduziert, Engpässe vermieden und damit einhergehend Kosten reduziert werden [NP19].

In einem nächsten Schritt kann das System um maschinelles Lernen erweitert werden. Diese Form der künstlichen Intelligenz ermöglicht die Entwicklung von Handlungsempfehlungen und Generierung von Maßnahmen anhand großer Datenbestände eines Prozesses [Re20]. Gemeinhin wird dieses Vorgehen als Predictive Process Mining bezeichnet. Anhand prozessbezogener Daten erkennt das System relevante Kausalitäten und erklärt diese. Während das System automatisch Trends und Muster

ableitet, werden die entwickelten Maßnahmen durch Mitarbeitende des Unternehmens bewertet und schließlich umgesetzt.

In einer vertieften Betrachtung sind die Potentiale einer Kombination des Process- und Data Mining zur Erreichung einer ganzheitlichen Standardisierung und Sicherstellung datenschutzrechtlicher Vorgaben zu untersuchen.

Literaturverzeichnis

[BB18] Becker, M.; Buchkremer, R.: Implementierung einer Regulatory Technology Lösung bei Finanzinstituten unter Berücksichtigung agiler Vorgehensmodelle. In (Mikusz, M.; Volland, A.; Engstler, M., Hrsg.): Projektmanagement und Vorgehensmodelle 2018 - Der Einfluss der Digitalisierung auf Projektmanagementmethoden und Entwicklungsprozesse. Köllen Druck+Verlag GmbH, Bonn, S. 125-134, 2018.

[BWW16] Buchholz, S.; Wirnsperger, P. J.; Wolff, D.: Zeitgemäßer Datenschutz in der datengetriebenen Wirtschaft - Effektive Umsetzung der EU-Datenschutz-Grundverordnung (DSGVO). Deloitte GmbH, 2016.

[Ca00] Chapman, P. et.al.: CRISP-DM 1.0: Step-by-step data mining guide. The CRISP-DM consortium, 2000.

[Ch20] Chowdhary, K.R.: Fundamentals of Artificial Intelligence. Springer India, 2020.

[CL16] Cleve, J.; Lämmel, U.: Data Mining. De Gruyter Oldenbourg, Berlin, Boston, 2016.

[FPZ95] Frye, D.; Palfai, T.; Zelazo, D.P.: Theory of mind and rule-based reasoning. In (Cognitive Development, Hrsg.): Cognitive Development, Volume 10, Issue 4. S. 483-527, 1995.

[Gö10] Gödert, W.: Semantische Wissensrepräsentation und Interoperabilität. In (Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e. V., Hrsg.): Information - Wissenschaft & Praxis, 61. Jahrgang, Nr. 1. S. 5-28, 2010.

[Ha16] Hackett, D.: Big Data in Life Insurance, https://www.mlc.com.au/content/dam/mlc/documents/pdf/media-centre/big-data-report.pdf, Stand: 14.05.2021.

[He18] Hellmann, R.: IT-Sicherheit: Eine Einführung. De Gruyter Oldenbourg, Berlin, Boston, 2018.

[Ki13] Kittel, K.: Agilität von Geschäftsprozessen trotz Compliance. In (Wirtschaftsinformatik Proceedings, Hrsg.): Wirtschaftsinformatiok Proceedings 2013. S. 967-981, 2013.

[KRS20] Köhler, W.; Schultz, C.; Rasche, C.: Das 100% Problem im Datenschutz. G-Forum Konferenz, Karlsruhe, 2020.

[NP19] Peters, R.; Nauroth, M.: Process-Mining Geschäftsprozesse: smart, schnell und einfach. Springer Gabler, Wiesbaden, 2019.

[Re20] Reinkemeyer, L.: Process Mining in Action Principles, Use Cases and Outlook. Springer Nature Switzerland AG, 2020.

[Wu20] Wuttke, L.: Datasolut. Von CRISP-DM: Grundlagen, Ziele und die 6 Phasen des Data Mining Prozess, https://datasolut.com/crisp-dm-standard/, Stand: 14.05.2021.