technische universität
dortmund

IT in Produktion
und Logistik

Prof. Dr. -Ing. Markus Rabe

Masterarbeit

aus dem Gebiet IT in Produktion und Logistik

Untersuchung des Einsatzes von Vorgehensmodellen des Knowledge Discovery in Databases für Bereiche der Logistik

Kandidat :	Nadine Beckmann
Matrikel-Nr.:	164193
Fakultät:	7 – Maschinenbau
Studienrichtung:	MA-Studiengang: Logistik
Prüfer:	Prof. Dr. -Ing. Markus Rabe
Betreuende Assistentin:	Dipl.- Inf. Anne Antonia Scheidler

Ausgegeben am:	26.05.2015
Eingereicht am:	10.11.2015

Inhaltsverzeichnis

1. Einleitung

Durch die Möglichkeit der Speicherung von großen Datenmengen ist im Laufe der Zeit auch die Menge gespeicherter und zu analysierender Daten angestiegen. Dies führt wiederum zu stetiger Verfügbarkeit und dem Verlangen, Wissen aus diesen Daten zu gewinnen und zu nutzen. Im Zuge dieser Entwicklung wird Wissen mit steigender Relevanz als eine wirtschaftliche Ressource gesehen und erfordert neue Methoden zur Wissensgewinnung und zum effizienten und effektiven Umgang mit Daten. Ein systematischer Umgang mit Daten, eine schnelle Beschaffung, Verwaltung und Bereitstellung sowie effiziente und effektive Analyse und Interpretation von Daten liefern Informationen, die einem Unternehmen einen wichtigen Wettbewerbsvorteil verschaffen können. Zunächst werden Daten in Archiven gesammelt. Das sind Datenbanken, die einen strukturierten Überblick über eine Ansammlung an Daten ermöglichen. Dort werden große Datenmengen gespeichert, jedoch nur selten wieder angefragt oder ausgewertet. Diese gespeicherten Daten können bedeutende Informationen enthalten, welche jedoch zum Teil nicht als explizite Einzelinformation gespeichert oder aufrufbar sind. Die Anforderung besteht darin, die Bedeutung der in diesen Daten verborgenen Informationen zu erkennen und Prozesse der intelligenten Informationsgewinnung aus Daten zu initiieren. Der Anstieg an Daten übersteigt jedoch zunehmend die Fähigkeit der menschlichen Analyse, weshalb das Data Mining, also das Schürfen nach Daten in Datenmengen und das Knowledge Discovery in Databases (KDD) als Gesamtprozess zur Wissensgewinnung aus Datenbanken zunehmend an Bedeutung gewinnen.

Anwendung findet der KDD-Prozess vor allem in den Bereichen Vertrieb und Marketing. Hier können anhand von Daten Informationen generiert werden, um spezialisierte Werbestrategien zu entwickeln, die schlussendlich den Umsatz fördern. Besonders durch die Automatisierung und die fortschreitende Globalisierung haben sich auch die Anforderungen an die Logistik verändert. Von zentraler Bedeutung sind mehr und mehr die Koordination von Beschaffungs-, Lieferungs- und Produktionsabläufen sowie deren Schnittstellen. Aus diesen vielfältigen Strukturen und verschiedenen Verantwortlichkeitsbereichen resultieren Datenmengen, welche die Notwendigkeit eines effektiven Informationsaustausches und Informationsgewinnung aus Daten mit sich bringen. Auf der Suche nach Optimierungspotenzial kommen hier innovative Informations- und Kommunikationstechnologien und besonders die Anwendung von KDD zum Einsatz.

Diese Arbeit soll effektive Einsatzmöglichkeiten von KDD in unterschiedlichen Logistikbereichen herausarbeiten. Ziel der Arbeit ist dabei, ein geeignetes KDD-Vorgehensmodell für die Logistik zu identifizieren. Anhand geeigneter Kriterien soll eine Auswahl an KDD-Verfahrensmodellen getroffen werden. Anschließend soll eine Gegenüberstellung der Vorgehensmodelle stattfinden und eine Bewertung erfolgen. Ein Teilziel ist die Betrachtungsweise der Logistik; ist diese als Ganzes zu sehen oder sollte eine Gliederung in verschiedene Bereiche stattfinden. In diesem Kontext soll zunächst eine Einteilung in die üblichen Bereiche der Logistik dargestellt werden. Demgegenüber wird Supply Chain Management als Aufbau, Organisation und Steuerung integrierter Logistikbereiche über den Gesamtwertschöpfungsprozess aufgezeigt. Anhand der Einsetzbarkeit der Vorgehensmodelle wird die Betrachtungsgrundlage des Logistikbereichs ausgewählt.

Entsprechend dieser Ziele soll zunächst eine Einordnung der Begriffe Data-Mining und KDD, danach ein Überblick über die Bereiche der Logistik erfolgen. Anschließend werden auf Grundlage wissenschaftlicher Literatur verschiedene Verfahrensmodelle des KDD-Prozesses

dargestellt. Eine Auswahl der gezeigten Modelle erfolgt anhand geeigneter Kriterien. Im Hauptteil der Arbeit sollen die Vorgehensmodelle des KDD in Bezug auf die Anwendbarkeit in einzelnen Logistikbereichen analysiert werden. Dazu werden Anforderungen an die Modelle abgeleitet und schließlich die Modelle mit ihren Vor- und Nachteilen gegenübergestellt. Mit dem Ziel, ein geeignetes Modell zu identifizieren, wird abschließend eine Bewertung vorgenommen und nach Möglichkeit eine Empfehlung gegeben. Zur Veranschaulichung der Anwendung des empfohlenen KDD-Vorgehensmodells soll eine konkrete Praxisanwendung beschrieben werden.

2. Bedeutung und Einteilung der Logistik

Nach Oeldorf & Olfert [2009, S. 42] ist Logistik die Organisation, die Planung und die Realisierung des gesamten Güter, Daten- und Steuerungsflusses entlang des Produktlebenszyklus innerhalb und zwischen Wirtschaftseinheiten. Diese Definition stellt jedoch nur eine der vielfältigen Varianten dar. Logistik befasst sich demnach mit der Summe der Prozesse des gesamten Flusses innerhalb eines Unternehmens. [Schulte-Zurhausen 2002, S. 293 f.]

Die Beschaffung, die Fertigung, die Distribution sowie die Entsorgung stellen damit Subsysteme der Logistik dar, wobei die Grundprobleme die zeitliche Synchronisation zwischen Gebrauch bzw. Verbrauch und der Herstellung liegen. Einzelproblemstellungen werden jedoch nie isoliert betrachtet, sondern unter Einbeziehung aller beteiligten Systeme. Aufgrund dieser übergreifenden Funktionen wird Logistik aus heutiger Sicht als Querschnittsfunktion verstanden. [Oeldorf & Olfert 2009, S. 42; Schulte-Zurhausen 2002, S. 293 f.]

2.1. Stellenwert der Logistik im Unternehmensumfeld

Funktionsorientiert wird die Betriebswirtschaft in die Teilbereiche Produktionswirtschaft, Beschaffung, Finanzwirtschaft und Personalwesen unterteilt. Dabei beschäftigt sich beispielsweise die Produktionswirtschaft mit der Modifizierung von Gütern wie Rohstoffen oder Halbzeugen unter Einsatz von Maschinen und Menschen zu wertigeren Gütern. Funktionsübergreifend wird dagegen die Logistik gesehen. Als betriebliche Querschnittsfunktion ist sie zuständig für die Bereiche Beschaffung, Produktion und Absatz; sie stellt ein Instrument zur Schaffung von Prozessketten dar. Ziel der Logistik ist es, die Verfügbarkeit des richtigen Gutes in der richtigen Menge, im richtigen Zustand, am richtigen Ort, zur richtigen Zeit, zu den richtigen Kosten zu sichern [Stiller 2015]. Die Logistik begleitet Materialien vom Rohzustand bis zur Fertigware. Neben dem Materialfluss steuert die Logistik auch den Informationsfluss mit dafür geeigneten Steuerungssystemen.

Die prozessorientierte Gestaltung, Lenkung und Entwicklung aller Aktivitäten über die gesamte logistische Kette wird als Supply Chain Management bezeichnet. Diese reicht von der Beschaffung der Rohmaterialien bis hin zum Endverbraucher, wobei nicht nur direkt vor- und nachgelagerte Prozesse untersucht werden. Waren- und Informationsflüsse werden firmenübergreifend und in beide Richtungen betrachtet. Der Aufbau und die Optimierung von Supply Chains werden von elektronischen Informationsflüssen unterstützt. Eine enge Zusammenarbeit mit allen Partnern ist die Voraussetzung für das funktionierende Supply Chain Management. *Nach Baumgarten* et al. [2004, S. 5 ff.] existiert darüber ein visionärer Supply Chain Management Zustand. Diese Steigerung stellt vollständig synchronisierte Abläufe dar und vermeidet unnötige Prozesse. Die Weiterentwicklung von Informations- und Kommunikationstechnologien ermöglicht es, Kundenwünsche und Nachfragevolumen elektronisch zu erfassen und allen Parteien durchgängig zur Verfügung zu stehen. Durch starke Integration des Endkunden resultiert eine Verbesserung der Marktkenntnisse, die bis zum Rohstofflieferanten einsichtig ist und die Basis aller Handlungen darstellt. Zu prüfen ist, ob die Anwendung von KDD und Data Mining, Werkzeuge darstellen, welche die Entwicklung in Richtung dieser Ziele begünstigen. [Oeldorf & Olfert 2009, S. 48 f.; Yakut 2015, S. 47 f.]

Ein funktionierender organisierter Datenfluss ist von immenser Bedeutung für eine effiziente Logistik. Eintreffende Aufträge lösen die Prozesse der Leistungserstellung aus und führen zu einem permanenten Datenfluss zwischen den Organisationsebenen. Dabei findet der Informationsaustausch statt, welcher sowohl als Anweisungen wie auch als Rückmeldungen fungieren kann. Ein funktionierender Materialfluss und Informationsaustausch erfordern möglichst synchrone Datenflüsse. [Gudehus 2012, S. 50]
Beim Durchlaufen der Logistikkette sind zudem Begleitdokumente erforderlich, die die Waren, Ladeeinheiten und Leistungen identifizieren und eine Kontrolle der Prozesssteuerung ermöglichen. Die Begleitinformation umfasst folgende Daten:

- eine Ident-Information zur Identifikation
- die Absenderinformation zur Kennzeichnung von Herkunftsort und Versender
- die Zielinformation zur Kennzeichnung des Bestimmungsortes und des Empfängers
- Steuerungsinformationen mit Angaben über Lieferwege, Zwischenstationen und Lagerorte. [Gudehus 2012, S. 51 f.]

Besonders mit der fortscheitenden Entwicklung zum volldigitalisierten Industrie 4.0-Unternehmen sowie dem „Internet der Dinge“ wird deutlich, dass der Bereich der Datenverarbeitung stetig an Bedeutung zunimmt. „In der Industrie 4.0 verzahnt sich die Produktion mit modernster Informations- und Kommunikationstechnik. Treibende Kraft dieser Entwicklung ist die rasant zunehmende Digitalisierung von Wirtschaft und Gesellschaft.“ [Kahlen 2015] Durch die Verknüpfung physischer Objekte mit einer internet-ähnlichen Struktur denken, lernen und organisieren intelligente Geräte ihren Weg zum Ziel selbst [Fischer 2015]. *Nach Prestifilippo* [2015] sollen künftig „Produktionsmittel, Maschinen und Anlagen die Fähigkeit erhalten, ihr Verhalten durch Selbstoptimierung und Rekonfiguration an gewandelte äußere Umstände, Auftrags- und Betriebsbedingungen anzupassen“.

Ziele der Logistik und ihrer Leistungssysteme lassen sich aus den übergeordneten Zielgrößen des Unternehmens ableiten. Neben humanitären und ökologischen Zielen, die als Rahmenbedingungen vorgegeben werden, leiten sich drei Hauptziele der Unternehmenslogistik ab. Während zu den humanitären Zielen die Sicherheit der Menschen, die Eliminierung unzumutbarer Arbeit, die Versorgung mit lebenswichtigen Gütern und die Entlastung des Menschen von körperlicher Arbeit und zu den ökologischen Zielen die Vermeidung von Abfällen, der Naturschutz und die Senkung des Energieverbrauchs gehören, sind die Hauptziele der Logistik folgende (**Abbildung 2-1**):

- Leistungserfüllung
- Qualitätssicherung
- Kostensenkung. [Gudehus 2012, S. 69 ff.]

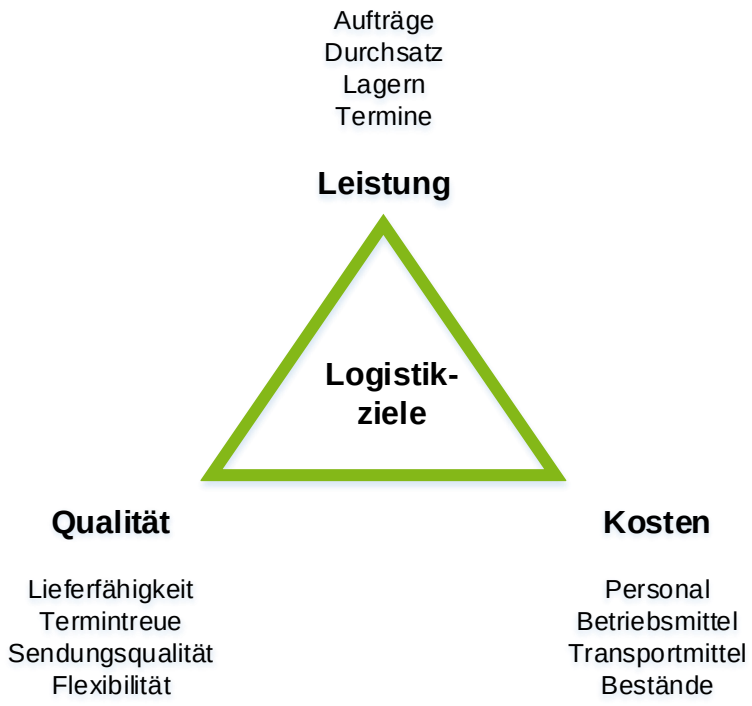

Abbildung 2-1: Ziele der Unternehmenslogistik [Gudehus 2012, S. 70]

„Anforderungsgemäße Leistungen und marktgerechte Qualität sind Voraussetzungen für gute Erlöse zu gewinnbringenden Preisen. In Verbindung mit geringen Kosten ergibt sich damit auch im Wettbewerb ein dauerhaft hoher Unternehmensgewinn." [Gudehus 2010, S. 70]

Maßstäblich für die **Leistungserfüllung** sind die kennzeichnenden Anforderungen. Vor der Planung und Realisierungen werden Einzelziele quantifiziert. Dazu gehören beispielsweise die Anzahl ausgeführter Aufträge, die Lagerung der Warenbestände und die Erfüllung von Serviceleistungen. Während des laufenden Betriebs werden diese Ziele anhand von Kennzahlen gemessen und bei Bedarf angepasst.

Zur **Qualitätssicherung** zählen Faktoren wie die Vollständigkeit von bearbeiteten Aufträgen, die Unversehrtheit, sowie die Lieferfähigkeit und die Einhaltung der vereinbarten Lieferzeiten. Die Qualitätsziele lassen sich in die Merkmale der Leistungsbereitschaft, der Sendungsqualität und Termintreue gliedern.

Ein zentrales Ziel der Logistik stellt die **Kostensenkung** dar, weil es sich dabei um ein übergeordnetes, grundsätzliches Ziel jeden Unternehmens handelt. Die Senkung der Kosten bei gleichbleibendem Erlös steigert die Wirtschaftlichkeit des Unternehmens. Zu den Logistikkosten zählen Lagerkosten, Bereitstellungskosten, Transportkosten, Bestandskosten und Steuerungskosten. [Gudehus 2012, S. 71 ff.; Seeck 2010, S. 5 ff.]

2.2. Bereiche und Modelle der Logistik

Logistik lässt sich nach verschiedenen Kriterien in unterschiedliche Bereiche aufgliedern. Zunächst soll eine Einteilung nach den Zielvorgaben dargestellt werden. Sie betreffen je einen „Aktionsbereich, der durch die Standorte und Funktionen der Quellen, Senken und Leistungsstellen sowie durch die vorgegebenen Material- und Datenströme definiert ist“ [Gudehus 2012, S. 4]. In Leistungsstellen werden mit Hilfe von Ressourcen wie Personen, Flächen, Maschinen und anderen Mitteln Aufträge und Anweisungen ausgeführt, um Produkte und Dienstleistungen zu erstellen. Ein Produkt stellt dabei das Ergebnis der Produktion in Form eines Sachziels dar, wohingegen die Dienstleistung ein immaterielles Gut darstellt. Für den weiteren Verlauf dieser Arbeit soll die Bezeichnung *Produkt* als übergeordneter Begriff definiert sein, der Sachleistungen und Dienstleistungen umfasst. Aufgabe der Leistungsstellen ist die Erbringung der geforderten Leistung zu geringen Kosten. Leistungsstellen verschiedener Räumlichkeiten lassen sich zu Leistungsbereichen zusammenfassen. Netzwerke aus Leistungsbereichen ergeben Leistungssysteme. [Gudehus 2012, S. 4, 9]

Mit dem Ziel der effizienten Versorgung von Haushalten, Unternehmen und dem Staat mit Gütern ist an erster Stelle die **Makrologistik** zu nennen. Ihre Aufgabe ist es, die Ströme zwischen Quellen und Senken zu ermöglichen, unabhängig vom Besitz der Quellen und Senken. Das beinhaltet sowohl die Verkehrs- und Güterströme, wie auch die Personenströme. Verkehrsnetze, Logistikzentren, geeignete Institutionen und Gesetzte ergänzen sich zu einer Infrastruktur, die rationelle Güterströme und Verkehrsflüsse ermöglichen.

Die **Mikrologistik** befasst sich mit der Versorgung einzelner Verbraucher und Unternehmen. Ihr Ziel ist es, den Güter- und Mobilitätsbedarf kostenoptimal zu decken. Zum Erreichen dieser Ziele sind Logistiksysteme, Beförderungsketten und Versorgungsnetzte zu organisieren.

Abbildung 2-2 zeigt die Bereiche der **Unternehmenslogistik**. Sie besteht aus der innerbetrieblichen wie auch der außerbetrieblichen Logistik. Für innerbetriebliche Logistik werden synonym auch Intralogistik, Betriebslogistik, Werkslogistik und Standortlogistik gebraucht. Gegenstand der innerbetrieblichen Logistik ist es, eine Verbindung zwischen Wareneingang, den internen Quellen und Senken sowie dem Warenausgang zu schaffen und zu organisieren. Die außerbetriebliche Logistik, auch Extralogistik, umfasst in Input-Richtung die Beschaffungslogistik, in Output-Richtung die Distributionslogistik und in Rücklaufrichtung die Entsorgungslogistik. Dadurch wird eine Verbindung zwischen Warenausgängen und Wareneingängen verschiedener Teilnehmer geschaffen.

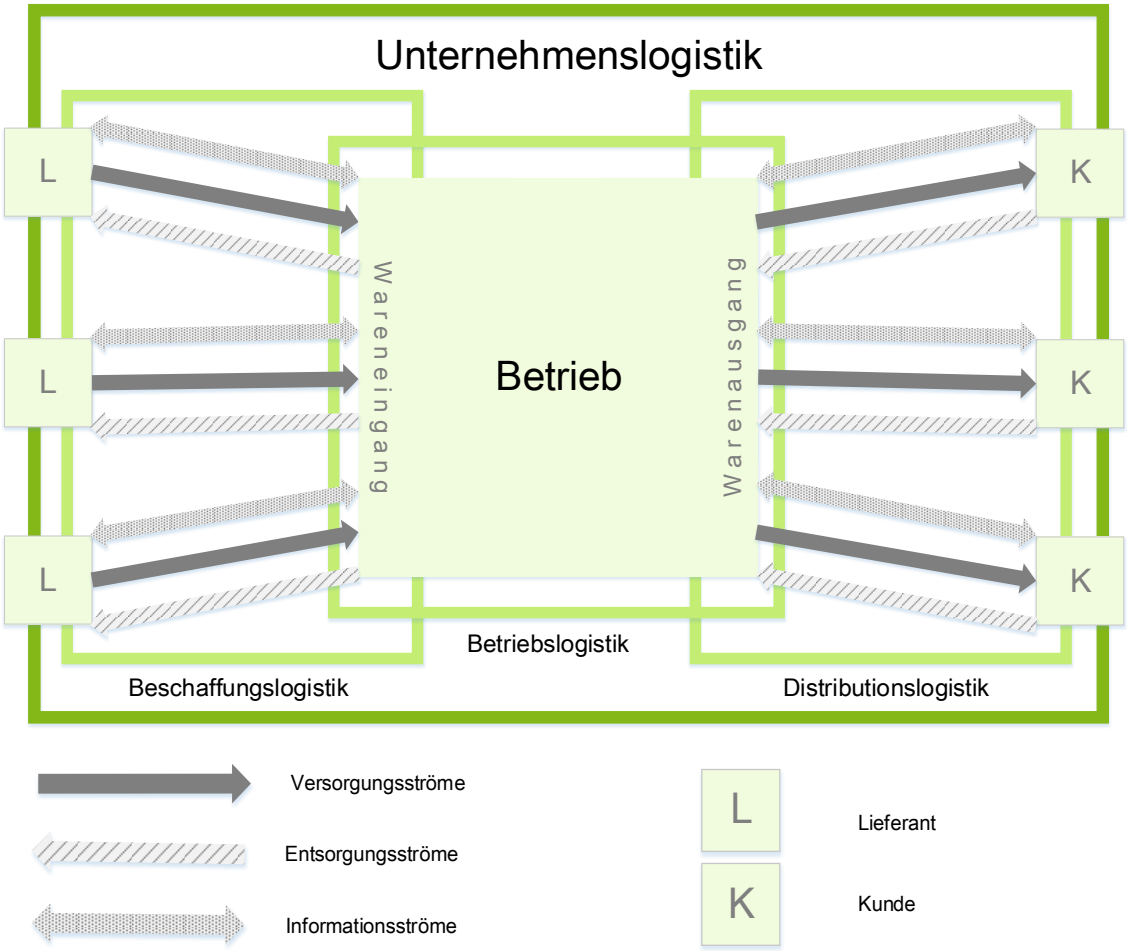

Abbildung 2-2: Bereiche der Unternehmenslogistik [Gudehus 2012, S. 5]

Die **Beschaffungslogistik** ist für den Zulauf von Waren zuständig. Sie stellt das Bindeglied zwischen Lieferanten und Empfängern her. Zu den Beschaffungsaufgaben zählen strategische sowie operative Entscheidungen. Erstere umfassen die Auswahl an Lieferanten, die Entscheidung über Make-or-Bye und Outsourcing, die Entwicklung von Einkaufsstrategien und die Qualitätssicherung, wobei diese Aufzählung nur einen Auszug an Aufgaben darstellen soll. Zu den operativen Aufgaben der Beschaffungslogistik zählen die Einkaufsplanung, die Verfügbarkeitsprüfung, sowie die Verfolgung und Prüfung der Bestellungen. [Hausladen 2014, S. 85 f.; Gudehus 2012, S. 5 f.]

Im Gegensatz zur Beschaffungslogistik befasst sich die **Distributionslogistik** mit der Verteilung der Fertigwaren zum Kunden. Je nach Betrachtungsweise sind Distribution und Beschaffung jedoch Aspekte der gleichen Logistikaufgabe. So ist die Distribution der Lieferanten aus Sicht des Empfängers Teil der Beschaffungslogistik. Die Beschaffungslogistik der Kunden wiederum stellt einen Teil der Distribution aus Sicht des Lieferanten dar.

Die **Produktionslogistik** umfasst die Aufgaben, die sich aus der Waren- bzw. Leistungserbringung ergeben. Die Produktionslogistik gliedert sich zwischen der Beschaffungslogistik und der Distributionslogistik ein und verbindet somit den Wareneingang eines Unternehmens mit dem Warenausgang. Zu ihren Aufgaben zählen zum Beispiel der Transport und die Lagerung von Rohmaterial, Hilfs- und Betriebsstoffen, Kauf- und Ersatzteilen

oder Halbzeugen und Fertigprodukten, sowie organisatorische Maßnahmen, die damit verbunden sind.

Die **Entsorgungslogistik** wird auch als inverse Logistik bezeichnet, da sie die Umkehr des Versorgens ist. Ihr Ziel besteht darin, Rückstände der Produktion, Abfälle, Verpackungsmaterial, Leergut und Reststoffe zu beseitigen. Das kann in Form von erneuter Verwendung, in der Weiterverarbeitung/Recycling oder Endlagerung geschehen. Zu den Aufgaben der Entsorgungslogistik zählen außerdem das Abtransportieren, das Lagern und das Aufbereiten.

Gegenstand der **Verkehrslogistik und Transportlogistik** ist die Beförderung von Gütern, Waren und Personen. Wichtige Teilaufgaben sind beispielsweise die Analyse der Warenströme, Standortplanung und die Tourenplanung. [Gudehus 2012, S. 4 ff.]

Der hierarchische Aufbau gliedert die Logistik in drei Organisationsebenen. Die drei Ebenen sind die **Administrative**, die **Dispositive** und die **Operative** Ebene. Die jeweiligen Aufgaben sind dabei folgende:

- **Administrative Ebene**
  - Unternehmensplanung
  - Strategieentwicklung
  - Programmplanung
  - Marketing
  - Verkauf
  - Einkauf
  - Finanz- und Rechnungswesen
  - Personalverwaltung
  - Controlling der Gesamtprozesse

- **Dispositive Ebene**
  - Auftragsdisposition
  - Auftragsverwaltung
  - Produktionsplanung
  - Arbeitsvorbereitung
  - Bestandsführung
  - Nachschubdisposition
  - Betriebsmitteldisposition
  - Auftragsverfolgung
  - Kontrolle der operativen Prozesse

- **Operative Ebene**
    - Auslösen der Prozesse
    - Steuern der Einzelvorgänge
    - Regeln der Prozesse
    - Überwachung der Prozesse
    - Sicherung der Durchführung

Anhand der jeweiligen Aufgaben lassen sich unterschiedliche Merkmale der Ebenen identifizieren. So definiert sich die **administrative Ebene** dadurch, dass ihre Aufgaben einen langen Entscheidungsraum von einigen Stunden bis zu Wochen haben. Ihre Entscheidungen basiert dabei auf unsicheren Informationen und unterliegen der Unternehmensleitung. Die **dispositive Ebene** unterliegt wiederum den Vorgaben der administrativen Ebene. Ihre Informationslage ist relativ gesichert und der Ausführungszeitraum beläuft sich auf mittlere Zeiten von Minuten bis einigen Stunden. An unterster Stelle der Hierarchie steht die **operative Ebene** unter den Vorgaben der dispositiven Ebene. Ihre Merkmale sind die kurze Reaktionszeit von Sekunden bis einige Minuten unter einer gesicherten Informationslage. [Gudehus 2012, S. 46 f.]

Im weiteren Verlauf sollen Modelle des Logistikmanagements vorgestellt werden. Zunächst wird dazu gezeigt, wie ein Logistiknetzwerk als ein System aus Knoten und Kanten modelliert werden kann.

„Moderne Logistiksysteme beruhen auf dem Konzept von Flüssen in einem Netzwerk, in dem Rechte, Güter, Finanzströme und Informationen von Quellen über Zwischenknoten zu Senken fließen und dabei Raum- und Zeitdifferenzen sowie Grenzen von Unternehmen überwinden." [Vahrenkamp & Mattfeld 2007, S. 5] Mit Netzwerken lassen sich Logistiksysteme anschaulich visualisieren. Zu beachten ist, dass mit der räumlichen und zeitlichen Veränderung eine Transformation der Güter stattfindet. Sie verändern sich hinsichtlich Menge, Sorte, Eigenschaft oder auch Information. Die Darstellung von Netzwerkmodellen beruht auf Knoten, Kanten und Gewichten, die an den Kanten haften. Die Bedeutung kann je nach Relation variieren. **Tabelle 2-1** zeigt dazu eine Auflistung der Bedeutungen der Objekte bei unterschiedlicher Relation.

Tabelle 2-1: Mögliche Bedeutung der Objekte [Vahrenkamp & Mattfeld 2007, S. 5]

Relation	Knoten	Kanten	Gewichte
Physisch	Orte	Verkehrswege	Distanzen
Logisch	Partner	Güterflüsse	Flüsse
Hierarchisch	Objekte	Ordnung	Mengen
Zeitlich	Zustände	Übergänge	Zeit

Die Abbildung von physischen Strukturen ist ein weit verbreiteter Anwendungsbereich. Knoten haben hier die Bedeutung von Orten, Kanten die der Verkehrswege, die zwei Knoten verbinden. Die Distanzen werden als Gewichte an den Kanten vermerkt (**Tabelle 2-1**). Diese Abbildung reduziert die Merkmale für den Betrachter auf die relevantesten Informationen. Als Beispiel dafür zeigt **Abbildung 2-3** einen Ausschnitt des deutschen Autobahnnetzes.

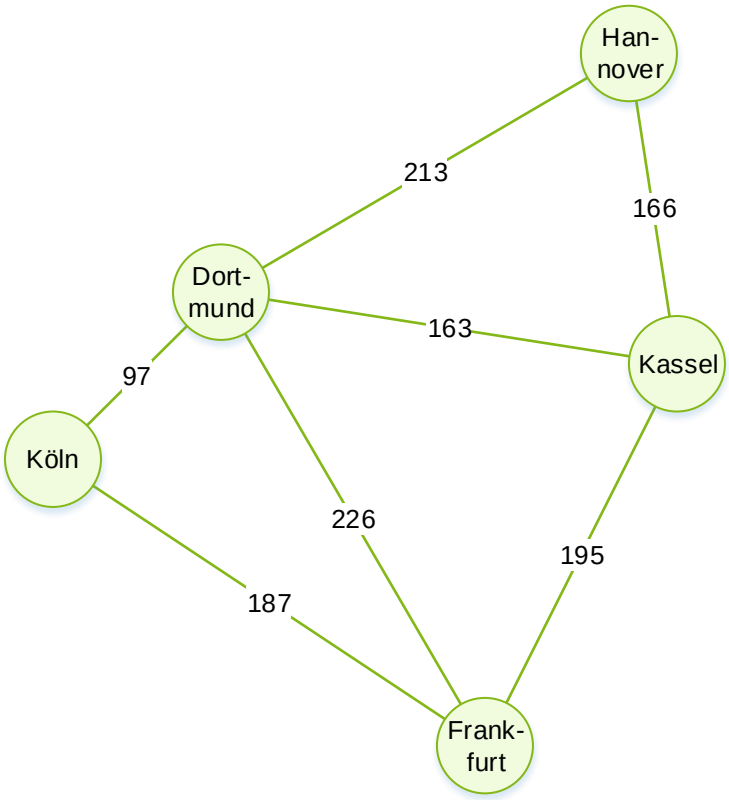

Abbildung 2-3: Ausschnitt aus dem Autobahnnetz Deutschland [Vahrenkamp & Mattfeld 2007, S. 6]

Bei der Modellierung ist die wesentliche Aufgabe des Modells zu beachten. Handelt es sich um ein Erklärungsmodell, stehen die Informationen des Modells im Vordergrund. Handelt es sich um ein Optimierungsmodell, dient dieses lediglich als Hilfsmittel zu Lösung eines übergeordneten Ziels. [Vahrenkamp & Mattfeld 2007, S. 7] Eine spezielle Netzwerkstruktur ist die Baumstruktur. Die Baumstruktur hat die Eigenschaft, Knoten mit einer minimalen Anzahl an Kanten zu verbinden. Bei der Auslegung von zum Beispiel Telefonnetzten oder Versorgungsleitungen findet die Baumstruktur ihre Anwendung. [Vahrenkamp & Mattfeld 2007, S. 26] Eine Abfolge von Knoten, beginnend bei einem Anfangsknoten, hinführend zu einem Endknoten, wird als Pfad bezeichnet. Hierbei kann ein Knoten mehrfach durchlaufen werden. Ohne diese Wiederholungsmöglichkeit von Knoten wird ein Pfad zu einem Weg. Sind Anfangs- und Endknoten gleich, entsteht ein Zyklus. [Vahrenkamp & Mattfeld 2007, S. 12 f.]

Zur Modellierung von Güterströmen in der Supply Chain wird mit dem **Transportmodell** gearbeitet. Güterströme können im Transportmodell von mehreren Anbietern zu verschiedenen Kunden führen. **Abbildung 2-4** zeigt ein Transportnetzwerk mit den Anbietern (1) bis (3), sowie den Kunden (A) bis (D). Die Nachfolgende **Tabelle 2-2** zeigt die dazu gehörende Transportmatrix mit den jeweiligen Mengen (x), die vom Anbieter (1, 2, 3) zum Kunden (A, B, C, D) transportiert wird. [Vahrenkamp & Mattfeld 2007, S. 101 f.]

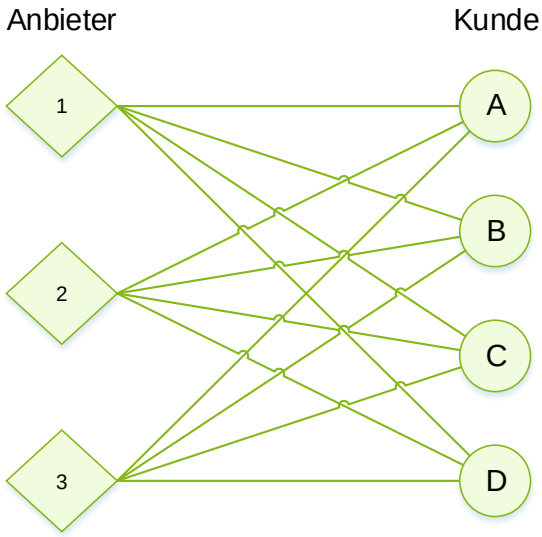

Abbildung 2-4: Ein Transportnetzwerk [Vahrenkamp & Mattfeld 2007, S. 102]

Tabelle 2-2: Transportmatrix

Anbieter/Kunde	A	B	C	D
1	x_{1A}	x_{1B}	x_{1C}	x_{1D}
2	x_{2A}	x_{2B}	x_{2C}	x_{2D}
3	x_{3A}	x_{3B}	x_{3C}	x_{3D}

Das zu lösende Transportproblem dieses Modells liegt darin einen optimalen Güterstrom zu bestimmen, der die Transportkosten minimiert und die Kapazitätsbeschränkungen der Anbieter berücksichtigt. Verfahren zur Bestimmung einer Lösung sollen an dieser Stelle nicht weiter erläutert werden.

Ein weiterer Modelltyp des Logistikmanagements sind **Modelle der Standortwahl**. In einem geographischen Raum sind die Warenströme zu koordinieren und die Kundennachfrage zu bedienen. Die Nähe zum Kunden und die Schnelligkeit der Kundenbelieferung sind dabei bedeutende Faktoren. Die Standortwahl „hängt von einer Vielzahl von Faktoren ab, so dass das Problem […] zu einem komplexen Entscheidungsproblem werden kann“ [Vahrenkamp & Mattfeld 2007, S. 141]. Beispiele dieser Faktoren sind:

- Politische Stabilität des betrachteten Landes
- Verfügbarkeit von Rohstoffen
- Klimatische Bedingungen
- Verfügbarkeit von Arbeitskräften
- Lohnniveau
- Verfügbarkeit von Grundstücken und Gebäuden
- Verkehrsinfrastruktur
- Lagegunst hinsichtlich Betriebskosten, Transportkosten und Serviceniveau

  [Vahrenkamp & Mattfeld 2007, S. 141]

Im Kontext des Supply Chain Managements spielt beispielsweise die Konfiguration von Netzwerken eine besondere Rolle. Gemeint ist damit die Festlegung der Standorte von Produktionswerken, Zulieferern und Lagerhäusern für Distributionssysteme. Die Modelle zur Standortwahl werden unterschieden in:

- **Diskrete Modelle:** Ein Netzwerk mit Informationen zu Distanzen. Standorte sind nur in den Knoten des Netzwerks möglich.
- **Kontinuierliche Modelle:** Basieren auf der euklidischen Ebene und lassen an jedem Punkt Standorte zu.
- **Semidiskrete Modelle:** Netzwerke mit euklidischen Daten. Standorte sind in den Knoten, aber auch auf beliebigen Punkten der Kanten möglich.

Weiter werden Standortmodelle nach den Begriffspaaren **statisch – dynamisch** und **deterministisch – stochastisch** unterschieden. Statisch – dynamisch bezieht sich auf die Dimension der Zeit. Dabei kennzeichnen sich statische Modelle dadurch, dass keine Zeitdimension berücksichtigt wird und stochastische Modelle dadurch, dass verschiedene Zeitperioden betrachtet werden. Bei der Einbeziehung stochastischer Elemente können Wahrscheinlichkeiten (z.B. über die Verfügbarkeit eines Produkts) oder auch Warteschlangen-Theorien berücksichtigt werden. [Vahrenkamp & Mattfeld 2007, S. 143]

Ein letzter Modelltyp, welcher vorgestellt werden soll, behandelt Modelle zur **Rundreise- und Tourenplanung**. Als Rundreiseproblem werden Fragestellungen behandelt, die eine Rundreise durch ein Netzwerk beschreiben. Unterschieden werden kanten- von knotenorientierten Rundreisen, je nachdem ob jede Kante oder jeder Knoten mindestens einmal besucht werden soll. Beispielsweise handelt es sich bei der Zeitungsverteilung um kantenorientierte, bei der Auslieferung von Paketen eines Paketdienstes um knotenorientierte Rundreise. [Vahrenkamp & Mattfeld 2007, S. 226, 231]

Die Tourenplanung erweitert die Problemstellung um Restriktionen wie zum Beispiel die Anzahl der Fahrzeuge, Kapazität der Fahrzeuge und Zeitrestriktionen. Ziel ist es, die Gesamttransportdistanz zu minimieren. Zunächst müssen dabei Kunden Touren zugeordnet werden, anschließend liegt jeweils ein Rundreiseproblem vor. Diese beiden Entscheidungen hängen jedoch voneinander ab. [Vahrenkamp & Mattfeld 2007, S. 275]

3. Knowledge Discovery zur Wissensgewinnung

In diesem Kapitel wird in die Thematik des Wissensmanagements eingeführt. Dabei soll zunächst der Zusammenhang der Begriffe Daten, Informationen und Wissen erläutert werden. Das Wachstum an gesammelten Daten, die in immer größer werdenden Datenbanken gespeichert werden, führt zu steigenden Anforderungen an die Analyseverfahren. Die gespeicherten Daten können Informationen von hohem Interesse enthalten. Sind herkömmliche Verfahren der Datenanalyse nicht mehr praktikabel oder ausreichend, werden KDD-Prozesse und das Data Mining eingesetzt. Im zweiten Teil dieses Kapitels folgt daher eine Einführung in das Thema des Data Mining und darüber hinaus in das KDD, welches den kompletten Prozess um das Data Mining beschreibt.

3.1. Daten, Informationen und Wissen

Der Wert des Wissens als Produktionsfaktor gewinnt zunehmend an Bedeutung. Neben der bereits verbreiteten Disziplin des Informationsmanagement wird daher auch vermehrt das Wissensmanagement zur festen Managementdisziplin vieler Unternehmen.

Für den Begriff „Wissen“ liegt zwar ein konventionelles Verständnis vor, jedoch existiert keine einheitliche, präzise Definition. Je nach verwendeter Definition ergeben sich unterschiedliche Eigenschaften für Wissen. So beziehen einige Definitionen auch Fertigkeiten, Fähigkeiten und Informationen mit ein. Als Beispiel sei hier die unterschiedliche Auffassung von Wissen in den Bereichen der Soziologie und der Informatik genannt. Was in der einen Fachrichtung unter Information verstanden wird, definiert die andere bereits als Wissen. [Doberstein 2011]

Um ein möglichst großes Spektrum abzubilden, wird für diese Arbeit eine offene Definition herangezogen.

Nach Segler [1985, S. 138] wird unter Wissen alles verstanden, „was der Akteur zur Generierung von Aktionen, Verhalten, Lösungen etc. verwendet, unabhängig von der Rationalität oder Intentionalität der Wissenselemente, also sowohl wissenschaftliche Erkenntnisse und Theorien, praktische Regeln und Techniken, als auch Patentrezepte, Eselsbrücken, Weltbilder, Bräuche, Aberglauben und religiöse und mystische Vorstellungen aller Art.“

Bei der Vielzahl an Definitionen wurde während der Recherche deutlich, dass trotz der Unterschiede auch bedeutende Gemeinsamkeiten auftreten. Wissen ist prinzipiell allgegenwärtig, endlos existierend und unbegrenzt kopierbar. Damit ist das Wissen eine Ressource, die durch Teilung nicht verbraucht werden kann. Des Weiteren ist Wissen immateriell; es ist an einen Wissensträger gebunden um existent zu sein. [Doberstein 2011]

Wie in **Abbildung 3-1** dargestellt, unterscheidet *Werner* [2004, S. 17ff.] natürliche und unnatürliche Wissensträger. Als unnatürliche Wissensträger werden hier alle speichernden Medien zusammengefasst, die nicht an Personen gebunden sind. Dahingegen sind natürliche Wissensträger Gruppen oder Organisationen, die über das Speichern des Wissens hinaus auch in der Lage sind mit diesem Wissen rational zu handeln. Die kleinste Einheit eines natürlichen Wissensträgers stellt ein Individuum dar.

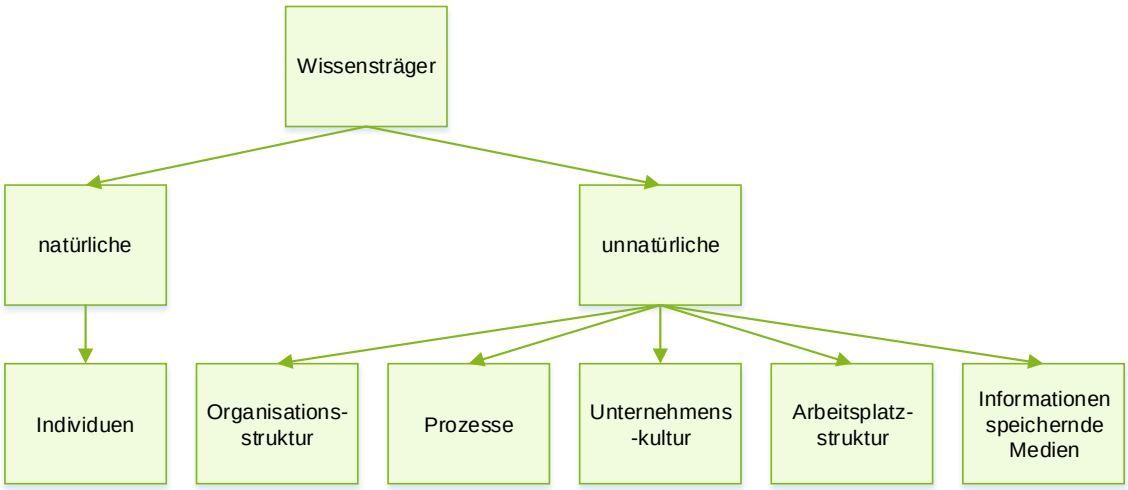

Abbildung 3-1:Wissensträger im Unternehmen [Werner 2004, S. 18]

Mit Augenmerk auf die Wirtschaft und Unternehmen soll Wissen einer Organisation möglichst einfach erfasst, erweitert, genutzt, gespeichert und verteilt werden. Neben dem humanorientierten Ansatz, Wissen von Person zu Person weiter zu geben, rückt hier der technologische Ansatz immer mehr in den Vordergrund. Mit Hilfe technischer Komponenten soll eine organisatorische Wissensbasis geschaffen werden und u.a. zur Sicherung der Wettbewerbsfähigkeit dienen. Daten, Informationen und vor allem Wissen sollen die Gesamtleistung des Unternehmens erhöhen. Zur Erreichung dieser Ziele muss die Entwicklung und Verbreitung von Wissen angetrieben werden. [Lehner 2014, S. 40; Kratzer & Van Veen 2014]

"Entwicklungen zeigen, dass Wissensmanagement vielleicht die wichtigste Herausforderung für Manager in der Zukunft wird, weil Fragen über Wissens- und Informationserzeugung und -verbreitung immer mehr in den Vordergrund gedrängt werden, die Selbständigkeit der Mitarbeiter stetig wächst, und auch die Möglichkeiten der Informationstechnologie sprunghaft zunehmen. " [Kratzer & van Veen 2014]

Mit dem Ziel der Wettbewerbsfähigkeit ergibt sich die Notwendigkeit, Wissen zu generieren um dieses nachhaltig umzusetzen. Die Wissenstreppe *nach North* [2011, S. 35 f.] zeigt Wissen und Informationen als Elemente in der Entwicklung zur Wettbewerbsfähigkeit. [Schmid 2013, S. 10]

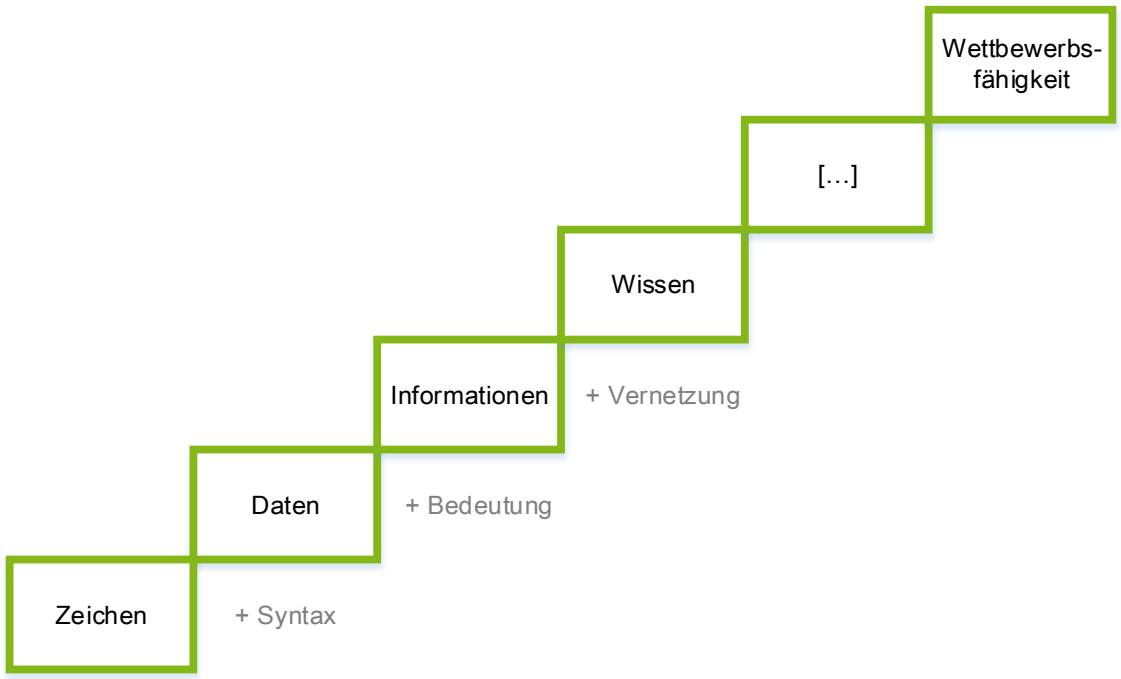

Abbildung 3-2: Die Wissenstreppe [North 2011, S. 36]

Wie **Abbildung 3-2** zeigt, entwickelt sich Wissen durch die Vernetzung von Informationen, welche wiederum aus Daten gewonnen werden. Zeichen werden durch eine präzise Regelung, beispielsweise einem Code oder Syntax zu Daten. *Nach North* [2011, S. 36f.] sind Daten Symbole, „die noch nicht interpretiert sind […]. Zu Informationen werden diese Daten erst, wenn ein Bezug hergestellt ist." In der Entwicklung zur Wettbewerbsfähigkeit führt Wissen in der richtigen Anwendung zu Kompetenzen, die wiederum bei Einzigartigkeit einen Wettbewerbsvorteil erzielen können.

Um verfügbares Wissen richtig anwenden zu können, bedarf es einer Strukturierung. Einfache Lösungen hierfür stellen Wissensdatenbanken dar. Inhalte lassen sich hier gliedern und klassifizieren. Für die Filterung der Informationen ergeben sich folgende Anforderungen:

- **Auswahl hochwertiger Informationsquellen** – Die Informationsquelle sollte ein gewisses Vertrauensniveau aufweisen. Die Richtigkeit der Daten ist die Basis für das resultierende Wissen.
- **Erkennen von duplizierten Informationen** – Identisches Wissen kann in unterschiedlichen Daten enthalten sein. Redundante Publikationen sollten vermieden werden.
- **Identifikation des Originals** – Beim Vorhandensein duplizierter Informationen ist die Originalquelle zu bevorzugen, um Verfälschungen zu vermeiden.
- **Identifikation relevanter Information** – Nur die für den Leistungsprozess relevanten Informationen werden benötigt.
- **Vermittlung der Sprachen von Autor und Nutzer** – Autoren und Nutzer verwenden unterschiedliche Begriffe um Inhalte zu präsentieren. Es ist eine Kommunikationsform zu wählen, die zwischen den Parteien vermittelt.
- **Klassifikation der Informationen** - Eine abschließende Klassifizierung sorgt für eine Übersichtliche Darstellung der Informationen. [Hoppe 2013]

Um größtmögliche Optimierungspotenziale aus Geschäftsprozessen zu generieren, ist ein weiteres Werkzeug des Wissensmanagements die Integration von Wissensprozessen. Genau wie Geschäftsprozesse, sollen die Wissensprozesse modellierbar und sichtbar gemacht werden. Neben dem Einsatz von implizitem Wissen liegt ein Hauptaugenmerk auf der im Anschluss folgenden Dokumentation. Neues resultierendes Wissen bedeutet einen Vorteil für den nächsten Durchlauf der Prozesse. [Allweyer 2005, S 297 f.]

Eine weitere Methode im Wissensmanagement beschreibt das „Lessons Learned“. Auch hier hat die Dokumentation von Wissen eine hohe Bedeutung. Um Wissen effizient in Organisationen einzusetzen gilt es, das Wissen der Mitarbeiter zugänglich zu machen. Ziel dabei ist, Wiederholungfehler zu vermeiden und Optimierungspotenzial zu steigern. Aus praktischen Erfahrungen vorangegangener Prozesse und Projekte werden Erkenntnisse gewonnen, die wenn sie in geeigneter Weise dokumentiert werden, zum Erfahrungswert bzw. zur „Lessons Learned“ werden. Die erfolgreiche Anwendung ist mit einer Selbstreflexion nach Durchführung eines Projekts oder Prozesses verbunden. Diese Aufbewahrung von Wissen hat den Vorteil, dass die bereits gewonnen „Lessons Learned“ auch nach Ausscheiden der betroffenen Mitarbeiter im Unternehmen bestehen bleiben und dadurch die Einarbeitung neuer Mitarbeiter vereinfacht wird. [Angermeier 2008; Probst et al. 2012, S. 136; Reinmann-Rothmeier et al. 2001, S 115 f.]

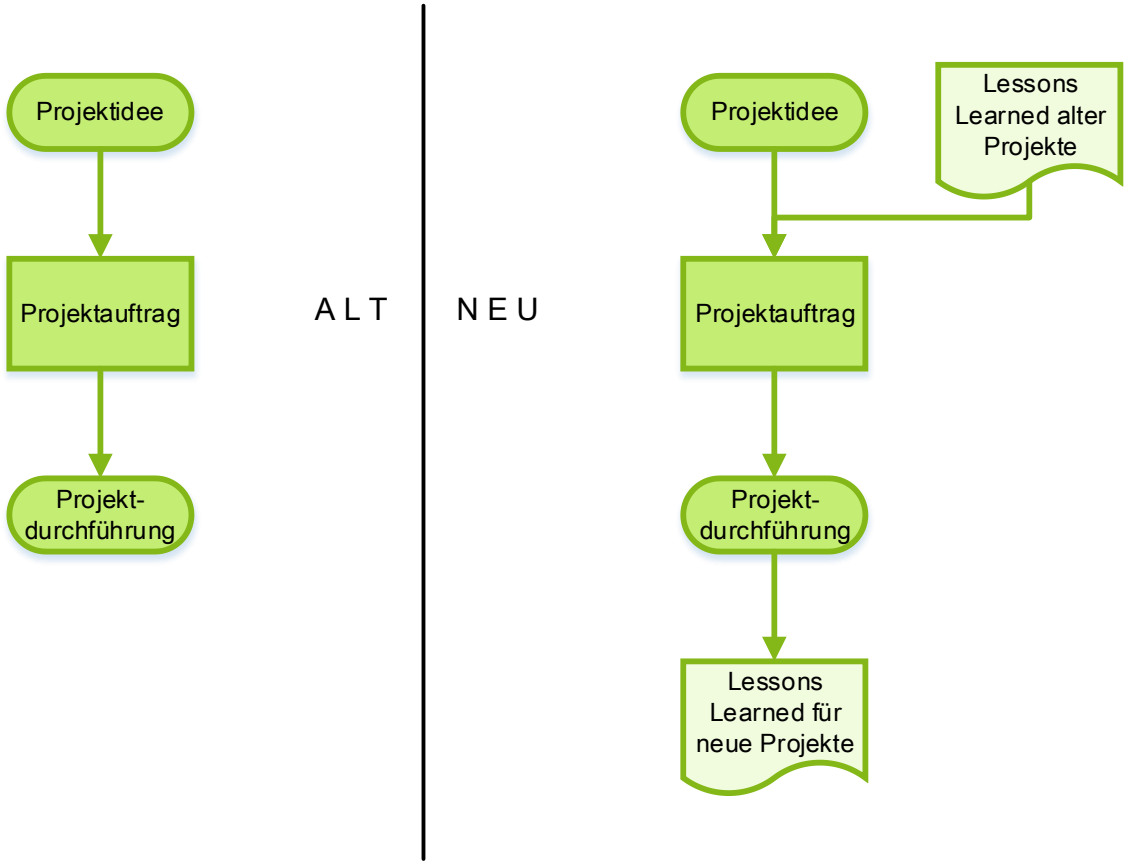

Abbildung 3-3: Integration von Lessons Learned im Projektprozess [Probst et al. 2012, S. 136]

Wikis dienen als weiteres Wissensmanagement-Werkzeug im Bereich Planung und Dokumentation. „Wikiwiki“ bedeutet übersetzt „schnell“ oder „sich beeilen“. Eine Wiki ist eine Software. Sie stellt Inhalte schnell und möglichst unkompliziert zur Verfügung, ist webbasierend und ermöglicht allen Nutzern die Inhalte sowohl zu betrachten als auch zu ändern. Die Wiki ist als eine benutzerfreundliche Plattform für kooperatives Arbeiten gedacht. Zu unterscheiden sind Wikis in zwei Anwendungsmöglichkeiten. Im World Wide Web (WWW) eingesetzte Wikis sind uneingeschränkt. Das bekannteste Beispiel dafür ist „Wikipedia“. Daneben können Wikis auch als Werkzeuge in geschlossenen Arbeitsgruppen eingesetzt werden. Dort unterstützen sie die aktive Kommunikation zwischen Mitarbeitern und helfen bei der Analyse, Strukturierung, der Erweiterung und dem Transfer von Wissen. Als Nachteil der Wiki-Software ist der Umgang hinsichtlich des Urheberrechts zu nennen. Im Wiki können Anwender anonym Inhalte sichten und bearbeiten. Ein Unternehmen will zumeist Rollen und Rechte individuell verwenden und möglichst Beiträge dem Wissensträger zuordnen. Weitere Anforderungen an Wikis für den Einsatz im Wissensmanagement in Unternehmen sind:

- **Sicherheit** – In Unternehmen eingesetzte Wikis sind in der Regel nur für Mitarbeiter verfügbar. Zusätzlich ist es möglicherweise notwendig, bestimmte Bereiche zu schützen. Nutzer müssen für den Zugang zu verschiedenen Bereichen über bestimmte Berechtigungen verfügen. Des Weiteren muss wie auch andere Unternehmens-Software die Wiki die spezifischen Sicherheitsanforderungen der Organisation erfüllen.
- **Suche** – Um vollständigen Zugriff auf alle relevanten Informationen zu gewährleisten, ist bei der Suchfunktion entscheidend, neben der Textsuche auch die Möglichkeit der Suche nach angehängten Dateien zu gewährleisten.
- **Benutzerfreundlichkeit** – Um Schulungs- und Einführungsaufwand zu reduzieren, ist die Einfachheit der Software wichtig. Auch die Akzeptanz bei den Mitarbeiten und die aktive Nutzung verbessern sich durch die Einfachheit der Bedienung. Dazu zählen zum Beispiel die Benutzeroberfläche der Software, die Vorstrukturierung oder das Vorhandensein von Vorlagen.

Diese Auflistung stellt nur einen beispielhaften Auszug aus weiteren Anforderungen dar. [Figura & Gross 2013; Ebersbach et al. 2008, S. 14 f.]

3.2. Data Mining und Knowledge Discovery in Databases

Beim Umgang mit Datenanalysen lassen sich zwei grundlegende Problemtypen unterscheiden. Ist das Ziel, bestehende Annahmen oder Theorien mittels Daten zu verifizieren oder zu widerlegen, handelt es sich um hypothesengetriebene Fragestellungen, auch Top-Down-Ansatz genannt. Hier kommen klassische Analyseverfahren wie die Clusteranalyse oder die Varianzanalyse zum Einsatz. Beispielsweise ist das Ziel der Varianzanalyse die Ursachen-Wirkungszusammenhänge von Variablen festzustellen, um kausale Beziehungen zu bestätigen oder zurückzuweisen. Im Vergleich dazu ist das Ziel der hypothesenfreien Analyse neue Erkenntnisse zu gewinnen. Hier wird auch von datengetriebener Analyse oder Bottom-Up-Verfahren gesprochen. Ein Verfahren ist als Data-Mining-Verfahren geeignet, wenn es das Ziel der Mustererkennung in Datenbeständen verfolgt. [Knobloch & Weidner 2000, S. 4f.]

In der Literatur ist eine Vielzahl an Definitionen für den Data-Mining-Begriff vertreten. Dabei begrenzen eine Reihe von Autoren die Data-Mining-Verfahren der Datenanalyse auf Verfahren aus den Bereichen der künstlichen Intelligenz, der Statistik und des maschinellen Lernens.

Andere Autoren verstehen sämtliche rechnergestützte Datenanalyseformen unter Data-Mining. *Fayyad et al.* [1996, S. 6] liefern die Definition, “Data Mining is a step in the KDD process consisting of particular data mining algorithms that, under some acceptable computational efficiency limitations, produces a particular enumeration of patterns“ Es soll bisher unbekannte Zusammenhänge aus Datenmengen ermitteln und Muster bzw. Trends erkennen. [Knobloch 2000, S. 3; Farkisch 2011, S. 97]

Unterscheiden lassen sich Data-Mining-Probleme je nach Autonomiegrad. Der Eingriff des Anwenders kann sich sowohl auf die Lenkung als auch auf die vorgegebenen Hypothesen beziehen. Weiterhin können Analysen hinsichtlich ihres Zieles der Mustererkennung oder Musterbeschreibung unterschieden werden. Mustererkennungsverfahren sind zumeist nicht überwacht. Überwachte Methoden hingegen werden zur Beschreibung vorgegebener Muster genutzt. Dabei ist anzumerken, dass die Vorgabe von Mustern als hypothesengetriebener Ansatz verstanden werden kann, was einen fließenden Übergang zwischen Top-Down- und Bottom-Up-Anwendungen erkennen lässt. **Abbildung 3-4** zeigt zusammenfassend die Einordung der Data-Mining-Ansätze. [Yakut 2015, S. 15; Knobloch 2000, S. 27; Knobloch & Weidner 2000, S. 5]

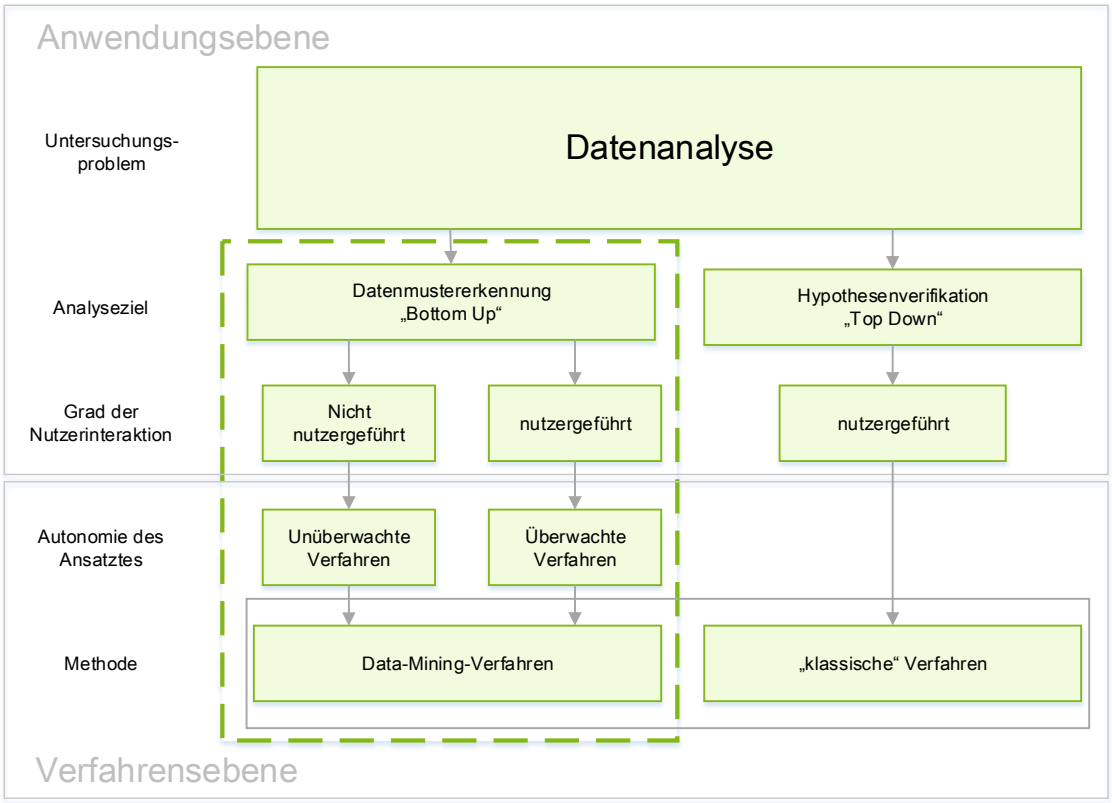

Abbildung 3-4: Einordnung von Data-Mining-Problemen [Knobloch 2000, S. 14]

Die spezifischen Verfahren und Techniken des Data Mining sind zurückzuführen auf die Statistik, maschinelles Lernen, künstliche neuronale Netzte sowie Datenbanktechnologien. Die Zusammenführung dieser Verfahren und Techniken wie auch fortschreitende Datenbanktechnologien und verbesserte Algorithmen zur Analyse machen ein erfolgreiches Data Mining möglich. An dieser Stelle sollen Beispiele aus den einzelnen Bereichen aufgeführt werden, die jedoch nur einen Auszug der Vielzahl der Verfahren und Techniken darstellen.

- **Statistik** mit dem Beispiel: Clusteranalyse

  Dabei werden Datensätze in Gruppen unterteilt, die Ähnlichkeiten aufweisen. Anstatt jeden einzelnen Datensatz zu betrachten, können bei der Clusteranalyse die verschiedenen Klassen verglichen werden, die ähnliche Daten zusammenfassen. Betrachtet werden schließlich die Charakteristika der Gruppen, nicht mehr die einzelnen Daten. Zur Einteilung der Daten werden beispielsweise deterministische Verfahren angewandt.

- **Neuronale Netze** mit dem Beispiel: Backpropagation Netzmodell

  Vorbild der neuronalen Netzte ist das Gehirn. Durch Interaktion mit Reizen lernt und passt sich das Netzwerk an. Dies führt i. d. R. zu nicht linearen Verzweigungen und komplexen Funktionen. Ein großer Nachteil findet sich darin, dass nur numerische Daten verarbeitet werden können. Die Struktur der neuronalen Netze zeigt Neuronenschichten, die hintereinandergeschaltet sind. Jedes Neuron ist mit jedem Neuron der nachfolgenden Schicht gekoppelt. Die erste Schicht enthält die Eingabemuster von außen, die letzte Schicht die Ausgabe. Sind Eingabe- und Ausgabeschicht nicht direkt verbunden, entstehen versteckte Schichten. Diese sind bei Netzmodellen wie beispielsweise dem Backpropagation Netzmodell vorhanden und ermöglichen die parallele Verarbeitung der Ausgabe der vorangegangenen Schicht. Verbindungtopologien bestimmen die Fließeigenschaften der Daten. So besteht die Möglichkeit durch Rückkoppelungen gespeicherte Daten mit aktuellen Werten zu verarbeiten.

- **Maschinelles Lernen** mit dem Beispiel: Entscheidungsbäume

  Hier findet eine weitgehend automatisierte Klassifizierung statt. Auf Grundlage von induktivem und überwachtem Lernen werden wesentliche Charakteristika anhand der Daten selbstständig ermittelt. Die visuelle Darstellung ist eine Baumstruktur, die zu einem einfachen Verständnis verhilft.

- **Datenbanktechnologien** mit dem Beispiel: Data Warehouse

  Ein Data Warehouse ist eine Sammlung von Daten. Aufgebaut ist es in Datenbanken, die den individuellen Bedarf der Entscheidungsträger nach benötigten Informationen bereitstellen. [Schinzer et al. 1999, S. 14, 97 ff.]

Zusammenfassend beschreibt Data-Mining das Aufdecken bedeutender Informationen in großen Datenmengen zur Generierung neuer Hypothesen. In der Literatur wird Data Mining und KDD vermehrt synonym gebraucht, *nach Fayyad et al.* ist Data Mining ein Teilprozess des KDD. Durch die Einbettung des Data Mining in den KDD-Prozess wird deutlich, dass Data Mining und KDD nicht isoliert betrachtet werden können. Die durch Data Mining entdeckten Muster werden erst durch geeignete Interpretation oder Bewertung zu brauchbarem Wissen. *Nach Fayyad et al.* wird „KDD als nicht-triviale Entdeckung gültiger, neuer, potenziell nützlicher und verständlicher Muster in Datenbeständen“ verstanden. Die Definition zeigt, dass die datengetriebene Entdeckung von Informationen, Anforderungen an die Ergebnisse stellt. Dies sind Gütekriterien, denen die Muster entsprechen müssen. Diese Anforderungen werden nun ausführlicher aufgeführt: [Knobloch & Weidner 2000, S. 4]

- **Entdeckung von Mustern**: Ein Muster ist eine allgemeingültige Aussage, die freie Variablen enthält. Konkrete Aussagen entstehen durch einsetzten konkreter Werte. In einer Datenmenge stellt ein Muster schließlich eine Hypothese über dieses Gebiet dar. [Fayyad et al. 1996, S. 585]
- **Nicht-Trivial**: Der Data-Mining-Prozess arbeitet mit Algorithmen, die einen gewissen Grad an Suchautonomie aufweisen. Die statistische Berechnung von Fakten ist im Data-Mining-Prozess uninteressant.
- **Gültigkeit:** Die Gültigkeit ist das Ergebnis der Überprüfung einer Hypothese. Dieses Ergebnis legt die Entscheidung fest, ob eine Hypothese als wahr behandelt werden kann. Beziehen sich Ergebnisse auf zu wenig Objekte der Gesamtdatenmenge stellen sie kein statistisch signifikantes Ergebnis dar. Auch fehlende oder fehlerhafte Daten führen zu unsicheren Ergebnissen. Ebenso spielt der Faktor Zeit bezüglich der Gültigkeit eine große Rolle. [Fayyad et al. 1996, S. 585; Knobloch 2000, S. 39 ff.]
- **Neuartigkeit**: Es handelt sich um bislang unbekannte neue Muster. Bereits bekanntes Wissen stellt im KDD-Projekt keinen Wert da. Ebenso sind Redundanzen herauszufiltern, bei der mehrere Muster denselben Sachverhalt beschreiben. Weiter sind auch triviale Resultate uninteressant. [Knobloch 2000, S. 39 ff.]
- **Potentielle Nützlichkeit**: Die Ergebnisse des Data-Mining-Prozessen sollen den Nutzen mit sich bringen, die Erreichung der Ziele des Anwenders zu unterstützen. Mangelnde Nützlichkeit liegt vor, wenn Resultate zwar die erforderliche Gültigkeit und Neuartigkeit erfüllen, jedoch nicht in Handlungen umgesetzt werden können. Nicht umsetzbare Ergebnisse sind daher weil irrelevant zu filtern. [Knobloch 2000, S. 39 ff.]
- **Verständlichkeit**: Anzustreben ist die Präsentation des Musters in einer universellen Form. Dies soll die Weiterverarbeitung erleichtern und zu einem besseren Verständnis der Information führen. [Knobloch & Weidner 2000, S. 4; Knobloch 2000, S. 14f.]

KDD als übergeordneter Prozess bezieht über das Data Mining hinaus Aspekte der Vorverarbeitung und der Interpretation mit ein und vereint diese zu einem Lösungsverfahren der Wissensentdeckung. Die drei Phasen Vorverarbeitung, Durchführung und Interpretation lassen sich in weitere Teilaufgaben zerlegen. In der Literatur lassen sich hierzu eine Vielzahl an Vorgehensmodellen finden. Eine als zweckmäßig erachtete Einteilung zur Orientierung soll hier vorgestellt werden. Es werden folgende Teilschritte unterschieden. [Yakut 2015, S. 15; Knobloch 2000, S. 27; Knobloch & Weidner 2000, S. 5]

- **Selektion der Daten**: Im ersten Schritt erfolgt die Definition der zu verwendenden Daten. Nach bestimmten Kriterien werden relevante Datensätze aus Quellenbeständen extrahiert. Diese Auswahl erachtet sich als Herausforderung, da im Voraus oft nicht bekannt ist, welche Größen welchen Einfluss auf das Ergebnis haben. Zur Selektion gehört weiter die Elimination von unbrauchbaren Daten. Auch die zeitliche Beständigkeit muss gewährleistet sein. Verändern sich Daten über den Zeitraum der Analyse, können Verfälschungen des Ergebnisses auftreten.
- **Exploration der Daten**: Um bestmögliche Ergebnisse im KDD-Prozess zu erzielen, ist qualitativ hochwertiges Datenmaterial Grundvoraussetzung. In der Explorationsphase sollen Fehler und Mängel aufgedeckt werden. Ein gutes Analyseergebnis hängt von der Korrektheit und Zuverlässigkeit der zu analysierenden Daten ab. Fehlerhafte Daten können Ergebnisse verfälschen ohne dass der Anwender Kenntnis darüber hat. Ein

gewisses Zeitbudget, das die Korrektheit und Qualität der Daten den weiteren Verlauf des KDD-Prozesses maßgeblich beeinflusst, sollte für diese Phase eingeplant werden.

- **Manipulation der Daten**: Rohdaten sind für den anschließenden Data-Mining-Prozess meist ungeeignet und führen zu Fehlern. Mögliche Probleme lassen sich in die Klassen Verfügbarkeit, Inhalt und Qualität und Repräsentation aufgliedern. Verfügbarkeit meint dabei das Fehlen wichtiger Datensätze. Diese können unter Umständen aus anderen Quellen entnommen oder berechnet werden. Zur Verbesserung der Qualität der Daten werden Redundanzen beseitigt und nicht korrekt gefüllte Datenfelder aufgefüllt. Die Repräsentation bezieht sich auf die Darstellungsform. Werden Daten aus mehreren Quellen integriert, herrscht in den meisten Fällen keine einheitliche Syntax. Ein gemeinsames Format wird in der Phase der Manipulation gefunden und umgesetzt.
- **Analyse der Daten**: Sind die Daten von geeigneter Qualität, kann die eigentliche Analyse durchgeführt werden. Dieser Schritt stellt das Data-Mining-Verfahren dar. Ein für die konkrete Problemstellung geeignetes Verfahren wird ausgewählt und auf die vorbereiteten Daten durchgeführt. Dieser Prozess erweist sich vermehrt als iterativer. Das Berechnungsmodell muss nach dem Durchlaufen angepasst werden und ein erneuter Analyseprozess wird durchgeführt. Je nach Ergebnis kann auch eine andere Verfahrensalternative gewählt werden. Folgende Auflistung zeigt die Anwendungsbereiche, die jeweilige Aufgabenstellung und eine Auswahl von Methoden.
    - **Segmentierung**

      Aufgabenstellung: Bildung von Klassen aufgrund von Ähnlichkeiten der Objekte

      Methoden: Neuronale Netze, Clusteranalysen
    - **Klassifikation**

      Aufgabenstellung: Identifikation der Klassenzugehörigkeit von Objekten auf der Basis gegebener Merkmale

      Methoden: Diskriminanzanalyse, Neuronale Netze, Entscheidungsbaum-verfahren
    - **Prognose**

      Aufgabenstellung: Prognosen der Werte einer abhängigen, kontinuierlichen Variablen auf Basis einer funktionalen Beziehung

      Methoden: Regressionsanalysen, Neuronale Netze, Entscheidungsbaum-verfahren
    - **Assoziation**

      Aufgabenstellung: Aufdeckung von strukturellen Zusammenhängen in Datenbanken mit Hilfe von Regeln

      Methoden: Assoziationsanalysen [Yakut 2015, S. 23]
- **Interpretation der Daten**: Die in der vorangegangenen Analyse entdeckten Muster erfordern eine anschließende Interpretation, um Wissen entstehen zu lassen und Handlungsmaßnahmen einzuleiten. Vor der eigentlichen Interpretation wird zunächst die Brauchbarkeit der Analyseresultate geprüft. Triviale oder im Sinne des Data Mining uninteressante Aussagen werden bei dem weiteren Vorgehen nicht beachtet. Die

Beurteilung erfolgt nach den beschriebenen Kriterien Gültigkeit, Neuartigkeit, Nützlichkeit und Verständlichkeit. Zur konkreten Interpretation bedarf es neben analytisch-methodischen Kenntnissen auch umfassende Kenntnisse des betroffenen Fachgebietes. Um möglichst gute Ergebnisse zu erzielen ist es ratsam, ein Team aus Experten zu bilden. Auch die Erkenntnisse der Explorationsphase liefern bereits wertvolle Hinweise, die die Bedeutung der Analyseergebnisse klarer erscheinen lassen.

Die erzielten Ergebnisse sind sinnvollerweise anhand der Datenbasis zu prüfen. Ergebnisse liefern oft neue Fragestellungen. An dieser Stelle wird deutlich, dass der KDD-Prozess als Zyklus gesehen werden kann oder zumindest Rücksprünge in vorangegangene Phasen erlaubt. [Knobloch 2000, S 28 ff.]

- **Bereitstellung und Sicherung**: Die Ergebnisse sind in kundenorientierter Art und Weise zu dokumentieren. Die Bereitstellung des entdeckten Wissens erfolgt in schriftlicher Form. Berichte über die Durchführung, Bereitstellung, Überwachung und Wartung sind anzulegen. [Sharafi 2012, S. 59]

4. Von der Methode zum Vorgehensmodell des Knowledge Discovery in Databases

In diesem Kapitel geht es um die verschiedenen Vorgehensmodelle des KDD. Zunächst werden die Begriffe Methode und Vorgehensmodell definiert und voneinander abgegrenzt, sodann die Entwicklung und Zusammenhänge von Methoden und Vorgehensmodellen vorgestellt. Kapitel 4.2 zeigt spezifische Vorgehensmodelle für den KDD-Prozess.

4.1. Abgrenzung der Begriffe Methode und Vorgehensmodell

Mit dem Begriff „Methode“ wird im Allgemeinen ein planmäßiges Verhalten und Handeln zur Erreichung von Zielen bezeichnet. Eine Methode ist ein Verfahren bestehend aus Regeln und Annahmen. Sie ist Grundlage zum Erreichen von Erkenntnissen oder Herstellung von Ergebnissen, die der Entwicklung von Hypothesen dient. Abzugrenzen ist der Begriff „Methode“ von dem Begriff des „Werkzeuges“. Eine Methode beschreibt in dem Zusammenhang lediglich die Art und Weise des Einsatzes von Werkzeugen. Die Methode entscheidet, welche Werkzeuge in welcher Reihenfolge zum Erreichen des Ziels angewandt werden. [Alby et al. 2009]

Ein Vorgehensmodell organisiert einen Prozess der Durchführung eines Vorhabens in verschiedene, strukturierte Abschnitte, denen wiederum entsprechende Methoden der Organisation zugeordnet sind. Das Vorgehensmodell beschreibt die einzelnen Phasen des Prozesses und strukturiert die Maßnahmen. [Sharafi 2012, S. 56] **Tabelle 4-1** zeigt Ziele auf, die methodisches Vorgehen verfolgen.

Tabelle 4-1: Ziele methodischen Vorgehens [Schmidt 2003, S. 39 f.]

Ziele	Erläuterung
Zielorientiertes Vorgehen	Es soll sichergestellt werden, dass die Ziele der Verantwortlichen (Entscheider)erkannt und verfolgt werden
Projektbegleitende Steuerung sicherstellen	Der oder die verantwortlichen Entscheider sollen kontinuierlich den Projektfortschritt steuern - die wichtigen Weichen stellen - da die Auftragnehmer in der Regel als Stäbe über keine eigenen Befugnisse verfügen • dadurch sollen kostspielige Fehlentscheidungen frühzeitig erkannt werden • die Entscheidungsträger sollen dadurch den Projektfortschritt besser nachvollziehen können, um deren Entscheidungsfähigkeit und -bereitschaft zu fördern
Planungshilfen durch einen Vorgehensleitfaden	Die Organisationsarbeit soll sich an einem Ablaufmodell orientieren, so dass • ein standardisiertes Vorgehen möglich ist, das die Koordination aller Beteiligten erleichtert • die Grundstruktur eines Projektablaufes nicht jedes Mal wieder neu geplant werden muss

Begrenzungen erkennen	Es sollen nur für Bereiche Vorschläge erarbeitet werden, die auch verändert werden dürfen. Den Handlungsspielraum einengende Vorschriften - was ist zu beachten, welche Restriktionen sind einzuhalten, was darf nicht herauskommen, was muss unbedingt herauskommen - sollen so früh wie möglich bekannt sein
Beherrschen komplexer Probleme	Es soll gewährleistet werden, dass • die gedankliche Auseinandersetzung mit einem Problem systematisiert (geordnet) und vereinfacht wird • bei der Arbeit im Detail der Überblick erhalten bleibt • Einzellösungen miteinander verträglich sind • Insellösungen vermieden werden
Rationalisierungspotenziale nutzen	Mehrfach benötigte Faktoren (Informationen, Sachmittel, Programme etc.) sollen • möglichst nur einmal entwickelt oder bereitgestellt werden • möglichst standardisiert werden

Besonders in der Bearbeitung von Projekten entsteht die Frage nach zweckmäßigem Vorgehen. Jedes Projekt bietet andere Herausforderungen, warum ein standardisiertes Vorgehen zunächst nicht möglich erscheint. Jedoch zeigt sich, dass „beim Organisieren Problemlösungsprozesse ablaufen, die bestimmte Strukturen aufweisen“ [Schmidt 2003, S. 36]. Ein Vorgehensmodell gilt als zeitlicher Leitfaden und organisiert den Ablauf von Projekten. Das Standardvorgehensmodell *nach Schmidt* [2003, S. 78 ff.] kann für beliebige Projekte angewendet werden. Es besteht aus folgenden Planungsschritten:

- **Auftrag:** Aufträge steuern den Prozess. Die Inhalte eines Projekts sind mit dem Auftraggeber zu Beginn des Projektes abzustimmen und im Auftrag festzuhalten. [Schmidt 2003, S. 78 f.]
- **Erhebung / Analyse:** Als Erhebung wird die Sammlung relevanter Informationen verstanden. Verschiedenen Techniken und Werkzeuge wie Befragungen, Beobachtungen und Multimomentaufnahmen können die Erhebung unterstützen. Von welcher Art diese Daten sind, hängt von dem Projektauftrag ab. Die anschließende Analyse meint die wertfreie Ordnung der vorangegangen Erhebung. Auch hier stehen Techniken zur Unterstützung zur Verfügung. [Schmidt 2003, S. 79 ff.]
- **Würdigung:** „Die Würdigung setzt sich wertend mit dem Ist-Zustand auseinander. Sie fragt nach Stärken und Schwächen, Chancen und Risiken der gegenwärtig vorhandenen Lösung.“ [Schmidt 2003, S. 83]
- **Lösungsentwurf:** Auf Grundlage der Ergebnisse der Erhebung und Analyse werden Lösungsvarianten entworfen. Zunächst werden diese ohne Wertung gesammelt. Mit steigender Variantenvielfalt erhöht sich die Chance auf die bestmögliche Lösung. Eine große Auswahl ermöglicht das Erkennen von ausschlaggebenden Vor- und Nachteilen. [Schmidt 2003, S. 84 f.]
- **Bewertung:** In der Bewertung erfolgt die Einschätzung, wie weit die einzelnen Varianten die formulierten Ziele erreichen. [Schmidt 2003, S. 85]
- **Auswahl:** An die Bewertung schließt die Auswahl einer Lösungsvariante an. Diese eröffnet oftmals den Auftrag für einen neuen Auftrag.

Dieser Ablauf kann folglich als Zyklus betrachtet werden, in dem der Output sogleich den Anstoß für einen weiteren Durchlauf dieser Planungsschritte gibt. Je nach Projekt sind einige Besonderheiten zu berücksichtigen, weshalb sich Varianten für verschiedene Anwendungsbereiche entwickeln. [Schmidt 2003, S. 40 ff.] Auf Analyseprojekte im industriellen Umfeld bezogen werden *nach Knobloch* [2000, S. 45 ff.] folgende 4 Schritte als zweckmäßig erachtet. Im Verlauf dieser Arbeit sollen diese vier Schritte als Basis dienen.

- **Spezifikation des Untersuchungsproblems**

  Ausgangspunkt einer Analyseaktivität im industriellen Umfeld ist eine Problemstellung. Eine Fragestellung formuliert die konkrete Handlungsabsicht und definiert erwünschte Ziele. Alle Mitarbeiter einer Organisation haben die Möglichkeit potenzielle Analyseziele zu identifizieren. Diese können aus Beobachtungen, Befragungen oder auf andere Weise ermittelt werden. Auch verfügbares Datenmaterial kann Analyseprojekte anstoßen. Weiterhin muss dieses Material identifiziert und auf seine Eignung getestet werden. Vor Start der Analyse sollte geprüft werden, ob es sich um eine einmalige oder sich wiederholende Analyse handelt, da im repetitiven Fall Maßnahmen zur Einbettung in die Geschäftsprozesse eingeleitet werden sollten. Weiter müssen rechtliche Fragen wie beispielsweise zum Datenschutz geklärt und die Realisierbarkeit und Wirtschaftlichkeit bewertet werden. [Knobloch 2000, S. 45 f.]

- **Durchführung der Untersuchung**

  Die Durchführung bezieht sich auf den eigentlichen Analysevorgang. Zur erfolgreichen Analyse ist ein Verfahren und geeignetes Werkzeug zu wählen, welche das verfolgte Ziel auf Basis der verfügbaren Daten am besten erreicht. Die Auswahl ist dabei vom jeweiligen Untersuchungsproblem und den gegebenen Ressourcen wie Zeit, Budget, Personal abhängig. Einige Verfahren erfordern spezifische Fachkenntnisse und sind mit hohem Aufwand verbunden; dennoch ist eine große Auswahl an Verfahren grundsätzlich vorteilhaft. Zur Durchführung von Analyseprojekten empfiehlt es sich eine jeweilige Projektgruppe aus Analyse- und Fachspezialisten zu bilden. Wiederkehrende gleichartige Analysesituationen sind bevorzugt von Fachspezialisten durchzuführen. Bei der Durchführung verschiedenartiger Problemstellungen ist es sinnvoll, Analyseexperten mit dem erforderlichen Know-how heranzuziehen. [Knobloch 2000, S. 46]

- **Umsetzung der Untersuchungsergebnisse**

  Besondere Erwartungen werden an Data-Mining-Projekte gestellt. Es werden Ergebnisse erwartet, die den hohen Aufwand und die investierten Kosten rechtfertigen. Im besten Fall ergeben sich neue Erkenntnisse, die nutzenbringende Handlungsmöglichkeiten eröffnen. Die Erwartungen können nicht immer erfüllt werden.

  Um den ökonomische Vorteil aus der Datenanalyse zu gewinnen, müssen die Ergebnisse in Handlungskonsequenzen abgeleitet werden. Es sind zwei Formen für die Verwendung der Analyseergebnisse möglich. Neben der einmaligen Umsetzung von Aktionen kann neu entdecktes Wissen auf operative Vorgänge Einfluss nehmen. Die Einteilung von Lieferanten in Risikogruppen wird beispielsweise in die Stammdaten aufgenommen und beeinflusst den weiteren Umgang mit diesen Lieferanten. [Knobloch 2000, S. 46 f.]

- **Evaluierung der Untersuchungssituation**

  Die Basis zur kontinuierlichen Verbesserung bilden die Bewertung der Untersuchungssituation selbst und die Maßnahmen, die zur Umsetzung der gewonnenen

Erkenntnisse eingeleitet werden. Die Bewertung des Analyseprojekts erfolgt anhand einer Gegenüberstellung des Zielzustands mit dem tatsächlich erzielten Erfolg. Für folgende Analysen können daran bereits neue Verbesserungspotenziale fest gemacht werden.

Ziel der Projektanalyse ist es, den betrieblichen Nutzen zu messen um die Untersuchungssituation und die Methode beurteilen zu können. Der betriebliche Nutzen bezieht sich auf Handlungsobjekte wie Kunden, Produkte oder Lieferanten. Der Ertrag ist dem Aufwand gegenüberzustellen. Die Untersuchungssituation beinhaltet Elemente der Zielsetzung, Datenbasis und das eingesetzte Verfahren. Zu beurteilen ist, ob die vordefinierten Ziele zufriedenstellend erreicht wurden.

Die Methodenbeurteilung muss ebenso Bezug auf das Untersuchungsproblem nehmen. Bewertet werden die Genauigkeit der Ergebnisse, die Genauigkeit der Beschreibung der Daten sowie die Zuverlässigkeit und Verständlichkeit des Modells und der Ergebnisse.

Der Ergebniswert wird anhand einer Gegenüberstellung der erwarteten Ergebnisse mit den erzielten gemessen. Die während der Phase erkannten Verbesserungspotenziale können neue Untersuchungsanstöße initiieren. Es wird empfohlen, Erfahrungen aus diesen Analyseprojekten zu kommunizieren, um andere Abteilungen auf die Potenziale hinzuweisen und an diese Erfahrungen anknüpfen zu können. [Knobloch 2000, S. 47 f.]

4.2. Vorgehensmodelle für das Knowledge Discovery in Databases

Für das KDD sind, wie auch für eine Vielzahl anderer Themengebiete, verschiedene Modelle entwickelt worden. Sie stellen das Vorgehen dar und geben somit einen Leitfaden für den Prozess der Durchführung. Ein KDD Vorgehensmodell besteht aus einer Abfolge von Vorgehensschritten. [Kurgan & Musilek 2006, S. 4 f.] Die Notwendigkeit eines Vorgehensmodells wurde erstmals 1989 *von Fayyad et al.* [1996, S. xiii] diskutiert. Als Grund für die Entwicklung und den Bedarf der Verwendung von Modellen ist die bessere Planung und Steuerung von KDD-Prozessen zu sehen. Die Durchführung anhand eines vorgegebenen Plans erspart Zeit und Kosten, erleichtert das Verständnis und erhöht die Akzeptanz solcher Projekte. [Sharafi 2012, S. 57]

Eines der ersten Vorgehensmodelle des KDD wurde von *Fayyad et al.* [1996] veröffentlicht. Seitdem wurden weitere KDD-Modelle für Wissenschaft und Industrie entwickelt. Dabei bestehen alle Modelle aus mehreren aufeinander folgenden Schritten, die zudem oft Schleifen und Iterationen enthalten. Jeder Schritt liefert nach erfolgreichem Abschluss ein Ergebnis, das den nachfolgenden Schritt anstößt und dieses als seine Eingabe empfängt. Die Aufgaben der Vorgehensmodelle sind es, Verständnis über das Projekt und die Daten zu erlangen, Datenaufbereitung und Analysen durchzuführen, Auswertungen zu betreiben und die Anwendung der Ergebnisse. [Kurgan & Musilek 2006, S. 4 f.]

Im Folgenden sollen zunächst fünf Modelle beschrieben werden, die in der Entwicklung der KDD-Vorgehensmodelle den größten Einfluss haben. Die Auswahl basiert auf einer detaillierten Literatursuche mit mehreren Indexdiensten, durchgeführt von *Kurgan & Musilek* [2006]. Verwendet wurden dabei SCOPUS, eine Zitations- und Abstractdatenbank für wissenschaftliche Journalbeiträge, CiteSeer (Scientific Literature Digital Library, deutsch: Digitale Bibliothek

wissenschaftlicher Literatur), eine Suchmaschine und Zitationsdatenbank für wissenschaftliche Literatur im Internet und andere. [Kurgan & Musilek 2006, S. 14]

Das erste veröffentliche Modell ist somit das aus neun Schritten bestehende Modell *von Fayyad et al.* [1996, S. 10 f.]. Das nächste Modell, bestehend aus fünf Schritten *von Cabena et al.* wurde 1998 entwickelt. *Anand & Bucher* entwickelten ebenso im Jahr 1998 ein Modell, bestehend aus acht Schritten. Das vierte Modell umfasst sechs Schritte und hat seine Wurzeln im Jahr 1996. Der **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (DRISP-DM) wurde in Kooperation von vier Unternehmen entwickelt und im Jahr 2000 offiziell. Das fünfte Modell, ebenfalls aus dem Jahr 2000, umfasst sechs Schritte und wurde *von Cios et al.* vorgestellt. **Tabelle 4-2** zeigt eine Übersicht der hier eingeleiteten Modelle. In der letzten Spalte der **Tabelle 4-2** wird ein generisches Modell vorgeschlagen, welches auf den fünf vorgestellten Modellen basiert. Das generische Modell kumuliert die Informationen der anderen Modelle und stellt eine verdichtete Ansicht dar. [Kurgan & Musilek 2006, S. 5]

Tabelle 4-2:Vergleichende Gegenüberstellung der wichtigsten KDD-Modelle [Kurgan & Musilek 2006, S. 6]

Modell	**Fayyad et al.**	**Cabena et al.**	**Anand & Bucher**	**Crisp-DM**	**Cios et al.**	**generisches Modell**
Jahr	1996	1998	1998	2000	2000	
Einsatz	Wissenschaft	Praxis	Wissenschaft	Praxis	Wissenschaft	
Anzahl an Schritten	9	5	8	6	6	6
Vorgehen	1 Developing an d Understanding of the Application Domain	1 Business Objectives Determination	1 Human Resource Identifination	1 Business Understanding	1 Understanding the Problem Domain	1 Application Domain Understanding
			2 Problem Specifikation			
	2 Creating a Target Data Set	2 Data Preparation	3 Data Prospecting	2 Data Understanding	2 Understanding the Data	2 Data Understanding
			4 Domain Knowledge Elicitation			
	3 Data Cleaning and Preprocessing		5 Methodology Identification	3 Data Preparation	3 Preparation of the Data	3 Data Preparation and Identification of DM Technology
	4 Data Reduction and Projection		6 Data Preprocessing			
	5 Choosing the DM Task					
	6 Choosing the DM Algorithm					
	7 DM	3 DM	7 Pattern Discovery	4 Modeling	4 DM	4 DM
	8 Interpretating Mined Patterns	4 Domain Knowledge Elicitation	8 Knowledge Post-processing	5 Evaluation	5 Evaluation of the Diskovered Knowledge	5 Evaluation
	9 Consolidating Discovered Knowledge	5 Assimilation of Knowledge		6 Deployment	6 Using the Discovered Knowledge	6 Knowledge Consolidation and Deploymant

Die oben gezeigte **Tabelle 4-2** enthält Informationen über das Jahr der Veröffentlichung, die Anzahl der Vorgehensschritte sowie das Einsatzgebiet der Modelle. Zudem führt es gegenüberstellend die einzelnen Teilschritte der jeweiligen Modelle auf. Der Vergleich zeigt, dass die KDD-Vorgehensmodelle ähnliche Schritte aufweisen und die gleiche Abfolge der Schritte angeben. Auf diesen Modellen basierend ist das generische Modell in der letzten Spalte der Tabelle aufgebaut. Es beschreibt ein zusammengefasstes Modell, basierend auf den fünf in **Tabelle 4-2** dargestellten Modellen. Deutlich wird die große Ähnlichkeit zu dem Modell CRISP-DM und dem Modell *nach Cios et al*., welche ebenso durch sechs Schritte beschrieben sind. Ein Grund für die Parallelen ist, dass diese beiden Modelle später entwickelt wurden, basierend auf Erfahrungen der älteren Modelle. Des Weiteren lassen sich auch die älteren Modelle auf diese generischen sechs Schritte reduzieren, indem mehrere originale Schritte zusammengefasst werden. So beinhaltet Schritt 3 „Data Preparation and Identification of DM Technology“ des generischen Modells die Teilschritte 3 „Data Cleaning and Preprocessing“, 4 „Data Reduction and Projection“, 5 „Choosing the DM Task“ und 6 „Choosing the DM Algorithm“ *nach Fayyad et al*. [Kurgan & Musilek 2006, S. 6] Für den weiteren Verlauf dieser Arbeit soll dem generischen Modell keine weitere Bedeutung zukommen, da es auf einer zu geringen Basis aufgebaut ist.

Weitere Vorgehensmodelle zeigen **Tabelle 4-3** und **Tabelle 4-4.** Es wird kein Anspruch auf Vollständigkeit erhoben. Die Tabellen enthalten Informationen über das Jahr der Veröffentlichung der Modelle, die Anzahl der Schritte, das Vorgehen der Modelle und die jeweiligen Quellenangaben. Aus den Tabellen wird ersichtlich, dass stetig neue Modelle entwickelt werden. Die dargestellten Modelle reichen von der Veröffentlichung aus dem Jahr 1996 bis 2010. Später veröffentlichte Modelle werden nicht weiter aufgeführt, da sie über keine ausreichende Quellenbasis verfügen. Weiterhin zeigt sich auch eine Vielfalt an Varianten bezüglich der Anzahl der Schritte. Vertreten sind Modelle, die aus drei bis sieben Schritten bestehen. Inhaltlich lassen sich alle gängigen Modelle mit gewisser Variation und teils unterschiedlichem Fokus auf vier vereinfachte Prozessschritte reduzieren. Diese sind:

- Vorbereitung
- Datenvorverarbeitung
- Methodenanwendung
- Ergebnisinterpretation

Die Vorverarbeitung kann dabei weiter in Aufgabendefinition und Datenauswahl gegliedert werden. [Sharafi 2012, S. 57]

Tabelle 4-3: Weitere KDD-Vorgehensmodelle (Tabelle 1 von 2)

Modell	**Adriaans & Zantinge**	**Brachmann & Ananad**	**Reinartz & Wirth**	**Berry & Linoff**	**SEMMA von SAS**	**John**	**Cooley et al.**
Jahr	1996	1996	1996	1997	1997	1997	1999
Anzahl der Schritte	5	6	6	4	5	6	3
Vorgehen	Data Selektion	Task Discovery	Requirement Analysis	Identifying the Problem	Sample	Define a Problem	Preprocessing
	Cleaning Enrichment	Data Discovery	Knowledge Aquisition	Analysing the Problem	Explore	Extract Data	Mining Algorithms
	Coding	Data Cleaning	Preprocessing	Taking Action	Modify	Data Engineering	Pattern Analysis
	DM	Model Development	Postom Extraction	Measuring the Outcome	Model	Algorithm Engineering	-
	Reporting	Data Analysis	Post Processing	-	Assess	Run Mining Algorithm	-
	-	Output Generation	Deployment	-	-	Analyse Results	-
Quellen	Adriaans & Zantinge 1996 Kurgan & Musilek 2005	Arndt 2008 Säuberlich 2000	Arndt 2008 Säuberlich 2000	Berry & Linoff 2011 Kurgan & Musilek 2006	Cleve & Lämmel 2014 Kurgan & Musilek 2006 Talia & Trunfio 2012	Arndt 2008 Säuberlich 2000	Arndt 2008

Tabelle 4-4: Weitere KDD-Vorgehensmodelle (Tabelle 2 von 2)

Modell	**Edelstein**	**5 As Martines de Pisón**	**Haglin et al.**	**Petersohn**	**Runkler**	**Hippner & Wilde**	**Wrobel et al.**
Jahr	2001	2003	2005	2005	2010		
Anzahl der Schritte	5	5	7	7	4	7	6
Vorgehen	Identifying the Problem	Assess	Goal Identification	Aufgaben-definition	Vorbereitung	Aufgaben-definition	Anwendung verstehen
	Preparing the Data	Access	Target Data Creation	Datenselektion	Vorverarbeitung	Auswahl der Daten	Extraktion / Integration
	Building the Model	Analyze	Data Preprocessing	Daten-aufbereitung	Muster-erkennung	Daten-aufbereitung	DM Verfahren wählen/ Analysedaten erzeugen
	Using the Model	Act	Data Transformation	Datananalyse	Nachbearbeitung	Auswahl der DM Verfahren	Verfahrens-anwendung
	Monitoring the Model	Automate	DM	Modell-evaluierung	-	Anwendung der DM Verfahren	Ergebnis-verarbeitung
	-	-	Evaluation and Interpretation	Anwendung des Analysemodells	-	Interpretation/ Evaluation	Umsetzung
	-	-	Take Action steps	Ergebnis-interpretation	-	Anwendung der Ergebnisse	-
Quellen	Kurgan & Musilek 2005	Talia & Trunfio 2012	Kurgan & Musilek 2005	Petersohn 2005 Sharafi 2012	Runkler 2010 Sharafi 2012	Hippner & Wilde 2001	Wrobel 1998 Wrobel et al. 2013

4.3. Knowledge Discovery in Industrial Databases

Speziell für die Anwendung in industriellem Umfeld liegen jedoch spezifische Herausforderungen vor. Die Datenanalyse setzte große zusammengeführte Databestände voraus, die in der Industrie so oft nicht vorliegen. Hier ist eine Vorbereitung und Integration nützlicher Daten aus verschiedenen Quellen nötig. Weiter fordern Analyseprojekte der Industrie möglichst Ergebnisgenerierung auf Basis von Ist-Daten in Echtzeit und ausreichende Integration von Experten. Die Akzeptanz von KDD im industriellen Umfeld und die Frage nach Datenschutz sind weitere Faktoren, die eine besondere Betrachtung von KDD-Projekten in der Industrie erforderlich machen. Unter dem Begriff Knowledge Discovery in Industrial Databases (KDID) versteht sich ein weiterentwickeltes, standardisiertes Vorgehensmodell, welches die Herausforderungen der Analyse in industriellen Datenbeständen berücksichtigt.

Der KDID-Prozess basiert auf den ursprünglichen KDD-Vorgehensmodellen, nimmt jedoch zusätzliche bisher wenig beachtete Faktoren in den Fokus. Besonderer Bedeutung kommen dabei der Einbeziehung von Experten aus den unterschiedlichen Anwendungsbereichen und der Selbstkontrolle durch Setzung von Meilensteinen zu. Im Folgenden werden die KDID-Schritte kurz erläutert und anschließend graphisch (**Abbildung 4-1**) dargestellt. [Deuse et al. 2014, S. 378 ff.]

Ähnlich den KDD-Vorgehensmodellen sind im ersten Schritt (Projektziele und Data Mining Aufgabenstellung definieren) die Projektziele zu bestimmen und daraus die Data-Mining-Aufgaben zu definieren. Industriell bedingt ist im zweiten Schritt (Ist-Zustand der IT-Struktur und des Expertenwissens aufnehmen) eine enge Einbindung der Anwender vorgesehen. Um einen Einblick in die Datenquellen zu schaffen, ist weiter der Ist-Zustand der IT-Infrastruktur zu untersuchen. Der dritte Schritt (Vorstudie Durchführen) ist wiederum angelehnt an bekannte Strukturen der KDD-Vorgehensmodelle. Anhand einer Datenprobe ist die Datenqualität zu prüfen, um Auffälligkeiten aufzudecken. Mit dem positiven Abschluss dieses Schrittes ist der erste im KDID-Vorgehensmodell ebenfalls neu integrierte Meilenstein erreicht. Der vierte Schritt (Datenbeschaffung) unterteilt sich in „Integration der Daten aus IT-Systemen durchführen“ und in „Daten erfassen und speichern“. Es erfolgt eine Integration der Daten aus verschiedenen Quellen und je nach Bedarf werden zusätzliche Daten erfasst. Im Hinblick auf industrielle Daten stellt dieser Schritt eine besondere Herausforderung dar, weil diese oftmals stark heterogenen Charakter aufweisen und unvollständig und inkonsistent sind. Die folgenden Schritte (Datenvorverarbeitung durchführen, Data Mining Modell erstellen und anwenden, Ergebnisse hinsichtlich der Zielerreichung interpretieren, gewonnenes Wissen in Planungs- und Entscheidungsprozesse integrieren) können als KDD-typische Schritte beschrieben werden, wobei nach der Ergebnisinterpretation ein weiterer Meilenstein angesetzt ist. Der KDID-Prozess schließt im letzten Schritt (IT-Prototyp zur Wissensentdeckung und –nutzung erstellen) mit der Entwicklung eines IT-Werkzeuges ab. Dieses soll sowohl die Ausführung des Data-Mining-Modelles als auch die Rückführung der Resultate realisieren. Die Umsetzung kann dabei ein einfacher Bericht bis hin zur Implementierung einer automatischen Datenauswertung sein. [Deuse et al. 2014, S. 383 ff.]

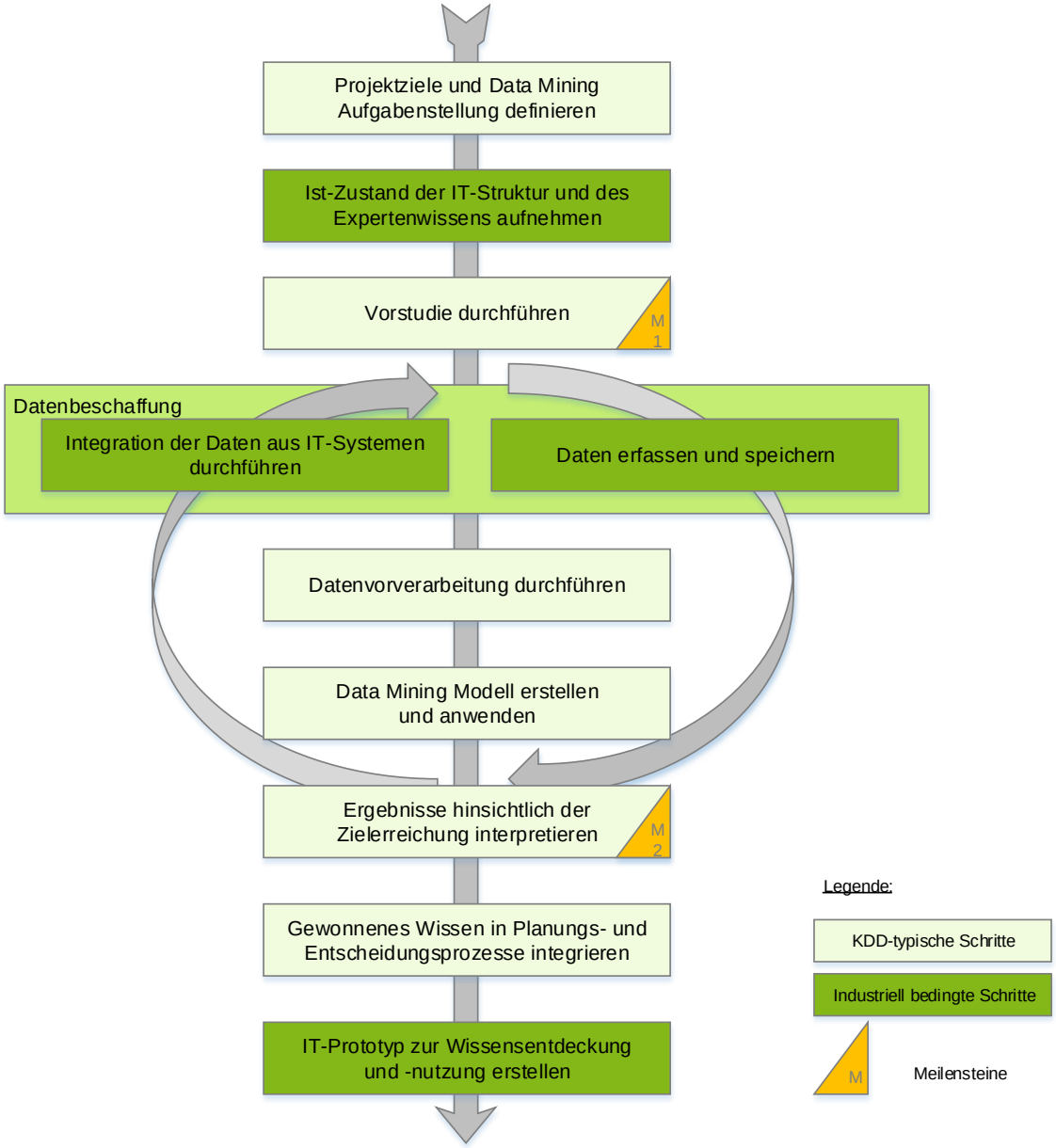

Abbildung 4-1: Knowledge Discovery in Industrial Databases [Deuse et al. 2014, S. 383]

Neben diesem KDID-Ansatz *nach Deuse* werden auch andere Modelle zur Einbettung in ein betriebliches Umfeld in der Literatur diskutiert. Der KDID-Prozess stellt eine interessante und vielversprechende Weiterentwicklung dar und soll deshalb nicht unerwähnt bleiben. Eine nähere Betrachtung dieser Ansätze ist im Verlauf dieser Arbeit jedoch nicht vorgesehen, da nur eine mangelhafte Quellenbasis vorhanden ist.

5. Anforderungen an die Vorgehensmodelle des Knowledge Discovery in Databases

In diesem Kapitel werden die Möglichkeiten des Einsatzes von Data-Mining-Methoden in Logistikbereichen dargestellt. Daraus werden Anforderungen an konzeptgebende Vorgehensmodelle des KDD für die praktische Anwendung abgeleitet. Anschließend werden die in **Kapitel 4.2** vorgestellten Vorgehensmodelle hinsichtlich dieser Anforderungen untersucht.

5.1. Data Mining im Anwendungskontext der Logistik

Neben dem Materialfluss ist der Umgang mit Daten ein wesentlicher Faktor der Logistik. Daten werden kontinuierlich erzeugt, müssen aufgenommen, verarbeitet und übertragen werden (**Kapitel 2.1**).

Daten der Logistik fallen an verschiedenen Stellen an. Neben den Auftragsdaten, die die jeweiligen Prozesse anstoßen und steuern, fallen kontinuierlich Maschinendaten und Sensordaten an. In der Fülle an Daten liegen Potenziale, daraus nützliche Informationen als Grundlage für neue Dienste zu ziehen. Die Einsicht, dass aus Datenbeständen besonders in der Wissenschaft wertvolle Informationen und daraus Wissen generiert werden kann, ist keineswegs neu; mit der Entwicklung von Data Mining und KDD wächst jedoch auch der potentielle Nutzen in wirtschaftlichen Zweigen zunehmend an.

Mögliche Data-Mining-Anwendungsbereiche der Logistik werden nun exemplarisch aufgeführt:

- **Risikogruppen:** Viele Unternehmen bieten bevorzugt Standardartikel und Standardleistungen an. Die Nachfrage wächst jedoch nach Sonderartikeln und kundenspezifischen Leistungen. Darüber hinaus können auch Aufträge, Produktionsprozesse, Lieferanten etc. vom Standard abweichen. Ein reibungsloser Ablauf hängt von vielen voneinander abhängigen Faktoren ab. Der Ausfall eines einzelnen „Zahnrades“ kann einen Prozess und das ganze Unternehmen massiv stören. Welchen Einfluss welche Störung hat und welche Präventivmaßnahmen zur Vermeidung in Betracht kommen, sind ohne genaue Analyse nicht zu bestimmen. Data-Mining-Methoden erkennen diese Risikofaktoren, teilen sie in Risikogruppen ein und geben so eine Basis, um Prozesse entsprechend zu gestalten, Maßnahmen einzuleiten oder Entscheidungen zu fällen. Eintrittswahrscheinlichkeiten und mögliche Schadenshöhen können bewertet werden und daraus resultierende Maßnahmen zur Prävention und Minimierung von Risiken eingeleitet werden.
- **Automatisierung:** Der Einsatz von Automaten, welche die Übernahme von Funktionen der Produktion und der Steuerung beinhalten, hat eine Produktivitätssteigerung, eine Flexibilisierung und damit die Erhöhung der Wirtschaftlichkeit zum Ziel. Je nach Umfang der automatischen Prozesse wird in Teil- und Vollautomatisierung unterschieden. Steigende Lieferanforderungen und das Streben nach immer kürzer werdenden Bearbeitungszeiten für jeden Prozess erfordern Optimierung. In Produktions- oder Logistiknetzwerken werden mittels Data-Mining-Verfahren Cluster identifiziert, innerhalb derer ein höherer Grad an autonomer Steuerung zugelassen werden kann. Ein

vollautomatisches Lager ermöglicht in diesem Zuge beispielsweise eine effizientere Kommissionierung und Bereitstellung der Versand-Paletten.

- **Maschinenwartung:** Wartung bezeichnet die vorbeugende Instandhaltung. Dazu gehören Pflegemaßnahmen von Anlagen und Maschinen wie das Reinigen, Justieren, Schmieren und andere Maßnahmen zur Verminderung und Vermeidung von Verschleiß. Die Aufzeichnung und Weiterverarbeitung durch Data-Mining-Methoden von Maschinendaten ermöglicht eine Analyse der Maschine in Echtzeit. Sie kann über den gesamten Lebenszyklus gesteuert und optimiert werden. Das ermöglicht zum Beispiel vorrausschauende Wartung und einen höheren Grad an Automatisierung. Über bekannte Fehlermeldungen hinaus können bisher unbekannte Fehlermuster entdeckt werden und entsprechendes Handeln frühzeitig eingeleitet werden.

- **Lagerhaltung:** Bei der Lagerung verschiedenster Produkte fallen große Mengen Daten an. Dazu zählen Mengen, Abmessungen, Lagerplätze, Adressen, Mindesthaltbarkeitsdaten, Umschläge, Bedarfe und Weiteres. Aufgabe der Lagerhaltung ist die Zwischenspeicherung von Rohstoffen und Produkten, um einen stabilen Materialfluss sicherzustellen. Diskrepanzen, die im Verkauf oder Einkauf auftreten, werden so überbrückt. Der Zielkonflikt besteht darin, Bestände möglichst niedrig zu halten, jedoch eine hohe Verfügbarkeit zu gewährleisten. Data Mining kann dazu dienen, Lagerbestände bei gleichbleibender Kundenbedarfsdeckung zu regulieren. Außerdem können hinsichtlich der Lagerverwaltung Lagerplätze anhand der zu kommissionierenden Aufträge so vergeben werden, dass Materialien, die vermehrt in Aufträgen zusammen erscheinen, zusammen gelagert werden.

- **Lieferantenauswahl:** Bei der Auswahl eines geeigneten Lieferanten ist nicht nur der Preis entscheidend. Neben dem Preis, der in die eigene Preiskalkulation mit einfließt, ist es sinnvoll weitere Auswahlkriterien zu definieren und verschiedene Angebote zu vergleichen. Weitere Kriterien sind eine geringe Fehlerquote, die Qualität der Waren, Liefertermintreue, Flexibilität, Preisgarantie und Angebotstransparenz, um nur einige zu nennen. Zur Entscheidungsunterstützung können Data-Mining-Methoden der Klassifikation behilflich sein. Lieferanten werden nach verschiedenen Indikatoren gefiltert und bewertet. Auch die Bewertung bereits vorhandener Lieferanten kann mit Data Mining Methoden beobachtet werden, um Veränderungen und Risikos frühzeitig zu erkennen.

- **Transportwege:** Die moderne Logistik ist ein Geflecht von Netzwerken, das Quellen (Lager und Umschlagszentren) mit Senken (Unternehmen, Haushalten und Konsumenten) verbindet. Die Distributionslogistik versucht Kundenzufriedenheit durch kurze Lieferzeiten mit der optimalen Kapazitätsnutzung und verschiedenen Liefermöglichkeiten zur Kosteneinsparung zu kombinieren. Data-Mining-Methoden bestimmen geeignete Lieferstrecken unter Einbeziehung verschiedener Faktoren wie zu versendende Mengen, Sendungsaufkommen, Regionen des Netzwerkes und anderes.

- **Liefertreue:** Je höher die Liefertreue, desto höher die Kundenzufriedenheit. Die Einhaltung von Terminen gehört zu den wichtigsten Logistikzielen eines Unternehmens. Dabei können sowohl die Verspätung als auch die verfrühte Auslieferung erhebliche Probleme auslösen. Die fehlende Verfügbarkeit von Produkten kann zu schwerwiegenden Lücken im laufenden Produktionsprozess führen. Eine verfrühte Auslieferung kann Engpässe im Lagerwesen auslösen. Data-Mining-Methoden können

Faktoren und sogar Ursachen identifizieren, die Terminabweichungen auslösen und Kriterien liefern, die diese verhindern können.

- **Mängel und Ausschuss:** Ausschuss meint den Anteil an Ausbringungen, der wegen Mängeln erst nach einer Nachbesserung oder gar nicht verwendet werden kann. Ursachen für Ausschuss sind Materialfehler, Bearbeitungsfehler, Transportschäden und andere. Ausschuss ist entweder durch Nacharbeit zu verbessern oder wird als Abfall deklariert und entsorgt wodurch in beiden Fällen Kosten entstehen, die vermieden werden sollten. Die Analyse und Auswertung von Maschinendaten mittels Data Mining kann zu einer wesentlichen Verbesserung der Prozesse führen. Bislang unentdeckte Fehlerquellen werden identifiziert und können beseitigt werden. Dadurch können neben der Ausschussrate auch Durchlaufzeiten und Wartezeiten optimiert werden.

- **Retouren:** Retouren sind Rücklieferungen an den Lieferanten. Gründe für Retouren sind beispielsweise Reparaturen, Falschlieferungen, Missfallen, mehrere Varianten zur Auswahl bestellt und Qualitätsabweichungen. Da einigen Gründe von den subjektiven Einschätzungen des Kunden abhängen, lassen sich nicht alle Retourengründe beseitigen. Im Zusammenhang mit Beschädigungen und Falschlieferungen sind Maßnahmen zur Vermeidung von Mängeln und Ausschuss heranzuziehen. Auch die Automatisierung von Prozessen, wie beispielsweise die scannergestützte Kommissionierung und die RFID-Technologie sind Maßnahmen zur Vermeidung von Retouren. RFID ist die Abkürzung für Radio Frequency Identification und zählt zu den Identifikationstechniken. „Das Ziel von RFID-Systemen ist die Identifikation beliebiger Objekte in logistischen Prozessketten, sowie die Verknüpfung von Informationen mit diesen Objekten“ [Krieger 2015]. Um aus den Retouren den größtmöglichen Nutzen ziehen zu können bedarf es einem analytischen Beschwerdemanagements. Segmentierungsmethoden des Data Mining gruppieren Aufträge und analysieren Rücksendegründe. Basierend auf diesen Untersuchen können Rückschlüsse auf die Ursachen der Retouren gezogen werden und Gegenmaßnahmen, wie eine angepasste Produktbeschreibung, Reduzierung von fehlerhaften Lieferungen und andere veranlasst werden.

Kapitel 5.1 zeigt, dass sich bezüglich Data Mining im industriellen Umfeld großes Potenzial ergibt. Der mögliche Einsatz von Data Mining ist anhand der Beispiele vorgestellt worden. Die Analyse der Prozessdaten deckt verborgene Informationen auf und führt zu neuem Wissen. Eine Umfrage des Fraunhofer Instituts für Produktionstechnik zeigt jedoch, dass Data Mining im industriellen Umfeld fremd ist. Die Umfrage verdeutlicht, dass es bei der Anwendung von Data Mining und damit von KDD besondere Faktoren zu beachten gibt. **Abbildung 5-1** zeigt das Ergebnis der Umfrage.

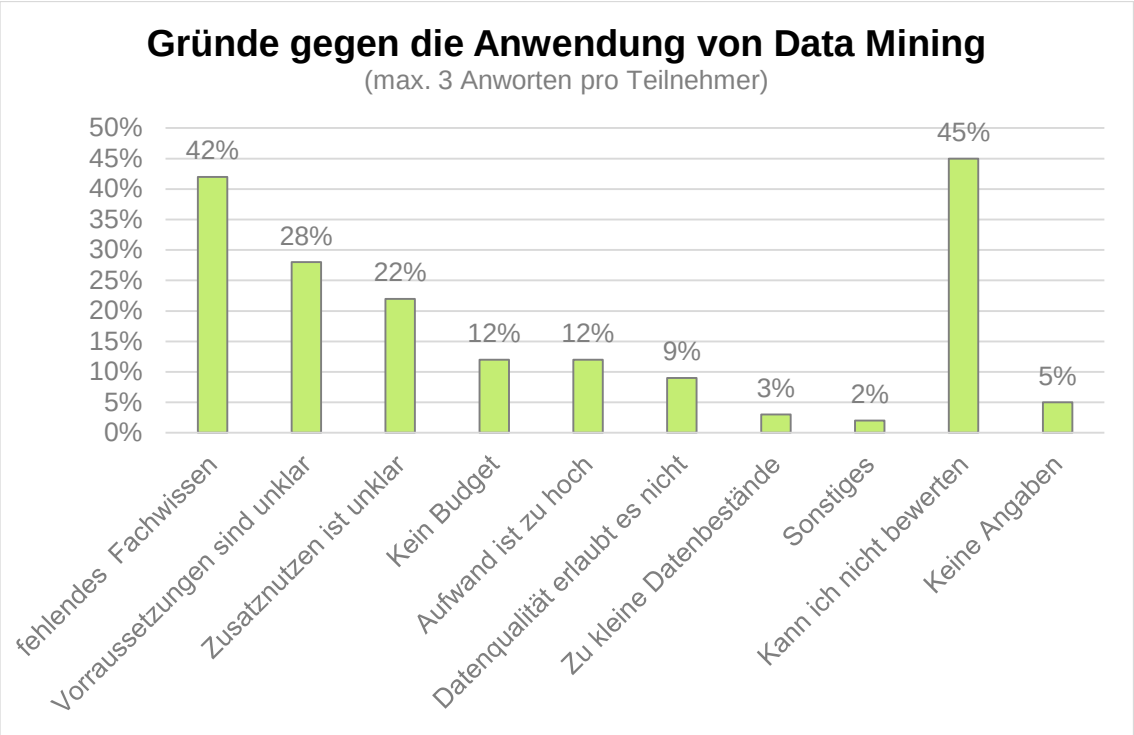

Abbildung 5-1: Gründe gegen Data Mining, Umfrage des Fraunhofer Instituts für Produktionstechnik [Westkamp et al. 2014, S. 21]

5.2. Anforderungen für den praktischen Einsatz von Knowledge Discovery Vorgehensmodellen

Bei der Anwendung verschiedener Vorgehensmodelle des KDD ist allen Modellen das gleiche übergeordnete Ziel zuzuschreiben: Nutzen aus Daten zur Verbesserung der Produkte im Unternehmen zu ziehen und somit die Generierung von Wettbewerbsvorteilen.

Dieses Ziel setzt jedoch voraus, dass Unternehmen das in diesen Daten verborgene Potenzial erkennen und Prozesse der intelligenten Informationsgewinnung initiieren und geeignet strukturieren. Potenziale bei der Anwendung des KDD liegen in den bislang unentdeckten Mustern von Daten. KDD ermöglicht intelligente Weiterentwicklung von Prozessen, Steuerung und Optimierung in Echtzeit und zeigt Potenziale der Automatisierung auf.

Wie in **Kapitel 5.1** exemplarisch aufgeführt, gibt es verschiedenste Einsatzgebiete für Data Mining in der Logistik. Die einzelnen Methoden des Data Mining sollen hier nicht weiter diskutiert werden. Gegenstand dieser Arbeit ist der Einsatz der KDD-Modelle, in die das Data Mining eingebettet ist. Für die erfolgreiche Anwendung der KDD-Modelle ergeben sich anhand der Einsatzmöglichkeiten bereits einige Anforderungen, die im Folgenden aufgeführt und erläutert werden.

Bei der Anwendung und Durchführung von KDD-Vorgehensmodellen stellt sich zu Anfang eine wichtige Frage: Auf Basis welcher im Unternehmen bereits vorhandenen Daten können Analysen zu den Prozessen durchgeführt werden? Wie schon der Aufbau der in **Kapitel 4.1** gezeigten Vorgehensmodelle des KDD verdeutlicht, ist die Basis zur Durchführung eines KDD-Projektes sowie der Data-Mining-Schritt jeden Modelles abhängig von einer geeigneten Datenbasis. Zu entscheiden ist demnach, welche Daten zielführend und ob diese im Unternehmen vorhanden sind. Beispielsweise können Daten aus Systemen für Enterprise Ressource Planning, Customer Relationship Management oder Sensor- und Maschinendaten herangezogen werden. Ein

Enterprise-Ressource-Planning-System oder kurz ERP-System dient der funktionsbereichsübergreifenden Unterstützung sämtlicher in einem Unternehmen ablaufenden Geschäftsprozesse [Vahrenkamp 2015]. Customer Relationship Management oder kurz CRM ist zu verstehen als ein strategischer Ansatz, der zur vollständigen Planung, Steuerung und Durchführung aller interaktiven Prozesse mit den Kunden genutzt wird. [Holland 2015]

Trotz technischen Fortschritts sind vollständig autonome KDD-Systeme nicht möglich. Erfolgreiches KDD ist vielmehr eine komplexe Interaktion zwischen Menschen und Daten. Der Mensch bedient sich dabei geeigneter Methoden und Werkzeuge. Diese Umstände verdeutlichen, dass den an dem Prozess beteiligten Personen oder Personengruppen eine bedeutende Rolle zukommt. Auch die Auswahl geeigneter Daten hängt von dem Fachwissen des eingesetzten Experten ab. Über die Anwendungsdomäne sollte folglich genügend **Expertenwissen** zur Verfügung stehen und genügend Fachpersonal zur Ausführung der Analyse vorhanden sein.

Die richtige Auswahl hängt weiter von der gegebenen Zielsetzung ab. Von der präzisen Zieldefinition hängt maßgeblich der Projekterfolg ab. Dem essentiellen Schritt der Zieldefinition muss in der Initiierungsphase von Projekten und im weiteren Verlauf höchste Aufmerksamkeit gewidmet werden. Auch hier liegt große Verantwortung im Fachwissen beteiligter Personen. Zudem setzt die Zieldefinition ein **Verständnis von Anforderungen** aus Sicht des Auftraggebers voraus. Die Anforderung an die Durchführung eines KDD-Projektes ist eine Auseinandersetzung mit dem gegebenen Problem, den verfügbaren Ressourcen und den Vorstellungen des Auftraggebers. Das Ergebnis ist ein gemeinsamer Vertrag mit den vereinbarten Konditionen. Die Zieldefinition stellt die Basis der Datenauswahl da.

Moderne Datenbanken bestehen in der Regel aus verschiedenen Datentypen. Auch der Umgang mit fehlerhaften Daten oder fehlenden Daten muss geregelt sein. Insbesondere die Auswirkung von Datenmängeln und Fehlern wird selten sofort erkannt. Vielmehr tritt dies erst beim Durchlaufen der Analysen in Augenschein. Der KDD-Prozess kann daher in den wenigsten Fällen geradlinig durchlaufen werden; er muss in jedem Schritt die **Möglichkeit für Rücksprünge** haben.

Um ein Analysevorhaben auf große Datensätze anwenden zu können, ist rechnergestützte Verarbeitung unerlässlich. Da ein großer Anteil des Aufwands für KDD-Projekte der Datenvorverarbeitung zukommt, sind Datenquellen wie das Data-Warehouse-System von Vorteil. Diese Daten haben gewisse Verarbeitungsschritte bereits bei der Eingabe zum Data Warehouse durchlaufen. Zudem dürfen Data-Mining-Verfahren nicht den Datenschutz gefährden. Für die Verarbeitung der anfallenden Daten besteht die Forderung nach **geeigneter Software**, die die KDD- Prozessschritte unterstützt.

Im Schritt der eigentlichen Datenanalyse liegt die Herausforderung in der Entscheidung für die richtigen Untersuchungsverfahren. Hierzu findet sich in der Literatur eine Vielzahl an Methoden. Die meisten beruhen auf Konzepten des maschinellen Lernens, der Statistik und der künstlichen Intelligenz. KDD-Systeme sollten eine entsprechende **Auswahl an alternativen Data Mining Werkzeugen** liefern.

Ein hohes Maß an Bedeutung kommt der Interpretationsphase der KDD-Vorgehensmodelle zu. Ziel ist es, aus den gewonnenen Erkenntnissen, Handlungsmaßnahmen abzuleiten. Hierbei handelt es sich um einen vorwiegend kreativen Prozess, der nur eingeschränkt unterstützt werden kann. Auch Experten mit umfassenden Domänenkenntnissen erreichen hier ihre Grenzen. Durchführungen von weiteren Untersuchungen und Analysen zur Abwägung der Auswirkungen verschiedener Alternativen müssen unter Umständen eingeleitet werden. Folglich

macht das die Notwendigkeit der **zyklischen Eigenschaft der KDD-Vorgehensmodelle** deutlich.

Für den Gebrauch der KDD-Vorgehensmodelle im praktischen Einsatz liegt die Anforderung in der **Einbettung des Modells in das betriebliche Umfeld**. Die Erkenntnis über interessante Informationen aus Daten genügt im praktischen Einsatz der Modelle nicht aus. Weitaus wichtiger ist das Ableiten geeignetere Maßnahmen, die dem Unternehmen zielbringenden Nutzen erweisen. Die aus dem extrahierten Wissen gewonnenen Maßnahmen müssen dabei bewertbar sein. Idealerweise handelt es sich um einen zyklischen Prozess. Anhand der Ergebnisse lassen sich schrittweise Verbesserungen einleiten.

5.3. Untersuchung der Vorgehensmodelle hinsichtlich der gestellten Anforderungen

Die in **Tabelle 4-2**, **Tabelle 4-3** und **Tabelle 4-4** aufgeführten Vorgehensmodelle sollen nun mit den Anforderungen aus **Kapitel 5.2** abgeglichen werden. Dabei geht es darum, ob die jeweiligen Modelle die Anforderungen erfüllen können. Entscheidungsgrundlagen dafür werden die in den Tabellen angeführten Prozessschritte sowie die ebenso in den Tabellen angegebene Literatur zu den einzelnen Modellen sein.

Einbeziehen von Experten

Hinsichtlich der Forderung nach Einbeziehung von geeigneten Experten und dem Bewusstsein, dass eine erfolgreiche KDD-Anwendung eine komplexe Interaktion zwischen einem Menschen und einer Datenbank ist, wird besonders bei dem Modell *nach Anand & Bucher* der Auswahl geeigneten Personals ein eigener Prozessschritt, „Human Resource Identification“ zugeordnet. Die Modelle *nach Adriaans & Zantinge*, *Cooley et al.* und *Runkler* vernachlässigen die Einbeziehung von Experten gänzlich und beginnen direkt mit Schritten wie der Datenauswahl und der Datenvorverarbeitung. Die übrigen Vorgehensmodelle erwähnen nicht explizit die Auswahl von geeignetem Fachpersonal, aber aufgrund der angegebenen Prozessschritte ist davon auszugehen, dass in gewissem Maße Experten in den KDD-Prozess einbezogen werden.

Verständnis von Anforderungen

Eine Problemdefinition liefern bis auf wenige Ausnahmen alle hier angegebenen Vorgehensmodelle. Dabei werden Ziele festgelegt, woraus sich die jeweiligen Anforderungen an das Projekt ergeben und ein Verständnis über die Gegebenheiten entsteht. Die Bezeichnung dieses Prozessschrittes variiert. So werden Schritte wie „Identifying the Problem“, „Anwendung verstehen“, „Assess“ und anderen für dieser Aufgabe verwendet. Die Vorgehensmodelle *nach Adriaans & Zantinge* und *Cooley et al.* sowie das Modell *SEMMA* verzichten auf einen Prozessschritt zur Aufgabendefinition bzw. zum Verstehen der Anwendung. Das Modell *nach Runkler* erfüllt diesen Punkt nur bedingt. Der Prozessschritt „Vorbereitung“ beinhaltet sowohl die Datenauswahl und -sammlung als auch Planung und Merkmalsgenerierung, wodurch eine Vorbereitung auf das Projekt, wenn auch in geringem Maß, gegeben ist.

Möglichkeit für Rücksprünge

Rücksprünge ermöglichen es Fehler, die erst im Verlauf des Prozesses erkannt werden, auszugleichen. Diese iterative Eigenschaft realisiert eine ständige Verbesserung des Prozesses selbst. Da diese Eigenschaft nicht an den Prozessschritten selbst erkennbar ist, ist die Einordnung in einigen Fällen nicht möglich, da auch die Literatur nur wenige Informationen dazu anbietet. Ausdrücklich vorhanden ist die Möglichkeit für Rücksprünge in den Modellen *nach Fayyad et al.*, *Runkler* und dem *CRISP-DM*. Die Modelle nach *Brachmann & Anand* und dem

Modell *SEMMA* erfüllen diese Punkt nur bedingt. Rücksprünge sind zwar im Modell *SEMMA* möglich, aber nur in einem definierten Prozessschritt. Im Modell *nach Brachmann & Anand* wird beschrieben, dass die Anwendung der Analysemethoden nicht nur einmal erfolgt, sondern solange wiederholt wird, bis ein zufriedenstellendes Ergebnis erreicht wird, was darauf schließen lässt, dass auch in diesem Modell Rücksprünge erfolgen. Es wird jedoch nicht deutlich, ob diese Rücksprünge in beliebigen Prozessschritten möglich oder auf bestimmt beschränkt sind. Die graphische Darstellung des Modells *nach Wrobel et al.* lässt ebenso erkennen, dass Rücksprünge möglich sind, um aus bereits entdeckten Wissen neue Potenziale zu ziehen und die Datenauswahl zu erweitern oder anzupassen. Rücksprünge aus jeder beliebigen Phase sind jedoch nicht verzeichnet. Im Modell *nach Hippner & Wilde* wird eine Toolbox von Methoden beschrieben, in welcher die einzelnen Prozessschritte zwar in der standartmäßigen Reihenfolge dargestellt werden, aber kein Anspruch auf die Einhaltung dieser Reihenfolge erhoben wird. Eine Anpassung des Vorgehens sowie die Wiederholung einzelner Schritte sind daher auch in diesem Modell möglich.

Geeignete Software

Auch über den Einsatz von geeigneter Software ist zu den meisten Modellen nicht viel bekannt. Das Modell *SEMMA* wurde speziell für das Data-Mining-Tool der Firma SAS entwickelt. Im eigentlichen Sinne ist *SEMMA* demnach kein richtiges KDD-Vorgehensmodell, sondern nur ein Abbild der Software. Im Fall der Entwicklung von *CRISP-DM* wurde eine Software der Firma SPSS entworfen, die zu den Entwicklern des Modells *CRISP-DM* zählen. Spezielle Software der übrigen Modelle bleibt in der Literatur unerwähnt. Jedoch gibt es neben kommerziellen Produkten (SPSS Clementine, SAS Enterprise Miner und weitere) auch Open Source Software. Zu den bekanntesten gehören RapidMiner, Konstanz Information Miner (KNIME) und Waikato Environment for Knowledge Analysis (WEKA). Der Einsatz ist unabhängig vom Vorgehensmodell und wird anhand spezifischer Faktoren gewählt, die projektabhängig variieren. Softwareunterstützung ist demnach weitestgehend in allen Modellen möglich, wobei vorwiegend der Data-Mining-Schritt unterstützt wird. In wieweit die jeweiligen anderen Prozessschritte in der Software Berücksichtigung finden, bleibt unklar.

Auswahl alternativer Data Mining Werkzeuge

Der eigentliche Data-Mining-Prozess beinhaltet eine Reihe verschiedener Algorithmen und Verfahren zur Mustererkennung in den gegebenen Daten. Je nach Aufgabenstellung und Zielsetzung eignen sich verschiedene Maßnahmen zur Bewältigung des Data-Mining-Schritts. Anhand der Beschreibungen der einzelnen Vorgehensmodelle ist davon auszugehen, dass sie alle über eine Auswahl an Methoden verfügen. In einigen Fällen wird der Auswahl geeigneter Data-Mining-Werkzeuge ein eigener Prozessschritt zugeschrieben. Im Modell *nach B*rachmann & *Anand* ist es der Schritt „Model Development“, im Modell *nach Fayyad et al.* sind es sogar zwei auf einander folgende Schritte „Choosing the Data Mining Task“ und „Choosing the Data Mining Algorithm“. Weiter haben auch die Modelle *nach John* mit „ Algorithm Engineering“, nach *Edelstein* mit „Building the Model“ und *nach Hippner & Wilde* mit „Auswahl der Data Mining Verfahren“ dieser Forderung einen separaten Schritt zugeteilt.

Zyklische Eigenschaft

Die zyklische Eigenschaft der Modelle gewährleistet einen kontinuierlichen Prozess, bei der eine ständige Überprüfung und Optimierung des Prozesses selbst erfolgt. Anhand der Vorgehensmodelle ist in den meisten Fällen nicht ersichtlich, ob diese eine zyklische Vorgehensweise berücksichtigt. Im Fall *CRISP-DM* wird der zyklischen Eigenschaft eine hohe Bedeutung zugesprochen, was durch die graphische Darstellung des Modells verdeutlicht wird.

Das Modell *nach Wrobel et al*. zeigt ebenso anhand der graphischen Darstellung einen zyklischen Verlauf. Im Modell *SEMMA* wird ebenfalls ausdrücklich von einem Kreislaufmodell gesprochen. Ebenso wird im Modell *nach Hippner & Wilde* darauf hingewiesen, dass die gewonnenen Erkenntnisse des Prozesses als Aufgabendefinition für neue KDD-Projekte dienen können. Über die anderen hier betrachteten Modelle liegen keine Informationen bezüglich dieser Eigenschaft vor.

Einbettung in das betriebliche Umfeld

Bezüglich der Einsetzbarkeit der verschiedenen Vorgehensmodelle in der Praxis ist die Einbettung in das betriebliche Umfeld von besondere Bedeutung. Dabei geht es um das Verständnis, die Ergebnisse in der Praxis anzuwenden und Handlungskonsequenzen abzuleiten. Die Berücksichtigung dieses Schrittes findet unterschiedliche Ausprägung in den aufgeführten Modellen. Im 5 A's Modell *nach Martine de Pisón* schließt das Modell mit dem Prozessschritt „Automate" ab. Da keine weiteren Informationen darüber vorliegen, kann damit sowohl die Anpassung der betriebsinternen Prozesse, als auch die Weiterverarbeitung der Ergebnisse gemeint sein. Die Modelle *nach Fayyad et al., Cabena et al., Cios et al., Reinartz & Wirth, Haglion et al, Hippner & Wilde* und *Wrobel et al.* und der *CRISP-DM* haben alle gemeinsam, dass nach der eigentlichen Data-Mining-Durchführung noch zwei abschließende Prozessschritte folgen. Dabei handelt es sich um den Schritt der Nachverarbeitung, in der die gefundenen Muster interpretiert und bewertet werden und anschließend um den Schritt der Anwendung und Umsetzung. Die Prozessschritte sind mit „Unsing the Discovered Knowledge", Assimilation of Knowledge", "Deployment", "Anwendung der Egebnisse" oder "Umsetzung" bezeichnet. Dadurch ist sichergestellt, dass die gefundenen Ergebnisse den Zugang in das betriebliche Umfeld erhalten. Die Modelle *nach Anand & Bucher, Adriaans & Zantinge, Brachmand & Anand, Berry & Linoff, John, Cooley et al., Edelstein, Petersohn* und *Runkler* und das Modell *SEMMA* schließen mit einer Interpretationsphase ab. In dieser werden die erkannten Muster bewertet, jedoch fehlen die aktive Umsetzung und die Ableitung von Handlungsmöglichkeiten für das betriebliche Umfeld.

Tabelle 5-1 und **Tabelle 5-2** fassen die Erkenntnisse über die Erfüllung der Anforderungen der einzelnen Vorgehensweisen übersichtlich zusammen.

Tabelle 5-1: Anforderungen an die Vorgehensmodelle (Tabelle 1 von 2)

	Fayyad et al.	Cabena et al.	Anand & Bucher	Crisp-DM	Cios et al.	Adriaans & Zantinge	Brachmann & Ananad	Reinartz & Wirth	Berry & Linoff	SEMMA von SAS
Einbeziehen von Experten	[✗]	[✗]	✓ ✓	[✗]	[✗]	✗	[✗]	[✗]	[✗]	[✗]
Verständis über Anforderungen	✓ ✓	✓ ✓	✓ ✓	✓ ✓	✓ ✓	✗	✓ ✓	✓ ✓	✓ ✓	✗
Möglichkeit für Rücksprünge	✓ ✓	?	?	✓ ✓	?	?	✓	?	?	[✓]
Geeignete Software	✓	✓	✓	✓ ✓	✓	✓	✓	✓	✓	✓ ✓
Auswahl an alternativen Data Mining Werkzeugen	✓ ✓	✓	✓ ✓	✓	✓	✓	✓ ✓	✓	✓	✓
Zyklische Eigenschaft	?	?	?	✓ ✓	?	?	?	?	?	✓ ✓
Einbettung des Modelles in betriebliches Umfeld	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗

voll erfüllt : ✓ ✓ erfüllt : ✓ nicht bekannt: ? nicht erfüllt: ✗

Tabelle 5-2: Anforderungen an die Vorgehensmodelle (Tabelle 2 von 2)

	John	Cooley et al.	Edelstein	5 As Martines de Pisón	Haglin et al.	Petersohn	Runkler	Hippner & Wilde	Wrobel et al.
Einbeziehen von Experten	[✗]	✗	[✗]	[✗]	[✗]	[✗]	✗	[✗]	[✗]
Verständis über Anforderungen	››	✗	››	››	››	››	›	››	››
Möglichkeit für Rücksprünge	?	?	?	?	?	?	››	›	›
Geeignete Software	›	›	›	›	›	›	›	›	›
Auswahl an alternativen Data Mining Werkzeugen	››	›	››	›	›	›	›	››	››
Zyklische Eigenschaft	?	?	?	?	?	?	?	›	››
Einbettung des Modelles in betriebliches Umfeld	✗	✗	✗	[›]	›	✗	✗	›	›

voll erfüllt : ›› erfüllt : › nicht bekannt: ? nicht erfüllt: ✗

6. Gegenüberstellung der Vorgehensmodelle des Knowledge Discovery in Databases

Zunächst erfolgt eine Auswahl repräsentativer Vorgehensmodelle, da eine Betrachtung aller in **Kapitel 4.1** gezeigten Modelle den Umfang vorliegender Arbeit überschreiten würde. Die Kriterien der Auswahl orientieren sich an den aufgestellten Anforderungen aus **Tabelle 5-1** und **Tabelle 5-2** und aktuellen Umfragen zum Einsatz der Modelle. Die Auswahl wird in **Kapitel 6.1** eingehend erläutert. Eine detaillierte Beschreibung der ausgewählten Vorgehensmodelle erfolgt in **Kapitel 6.1.1** und folgende.

6.1. Auswahl geeigneter Vorgehensmodelle des Knowledge Discovery in Databases

Mit dem Ziel geeignete Modelle für den Einsatz im logistischen Umfeld zu untersuchen, werden Modelle ausgewählt, die bereits in der Praxis angewendet werden. **Abbildung 6-1** zeigt eine Umfrage aus den Jahren 2007 und 2014 über den Einsatz von Vorgehensmodellen.

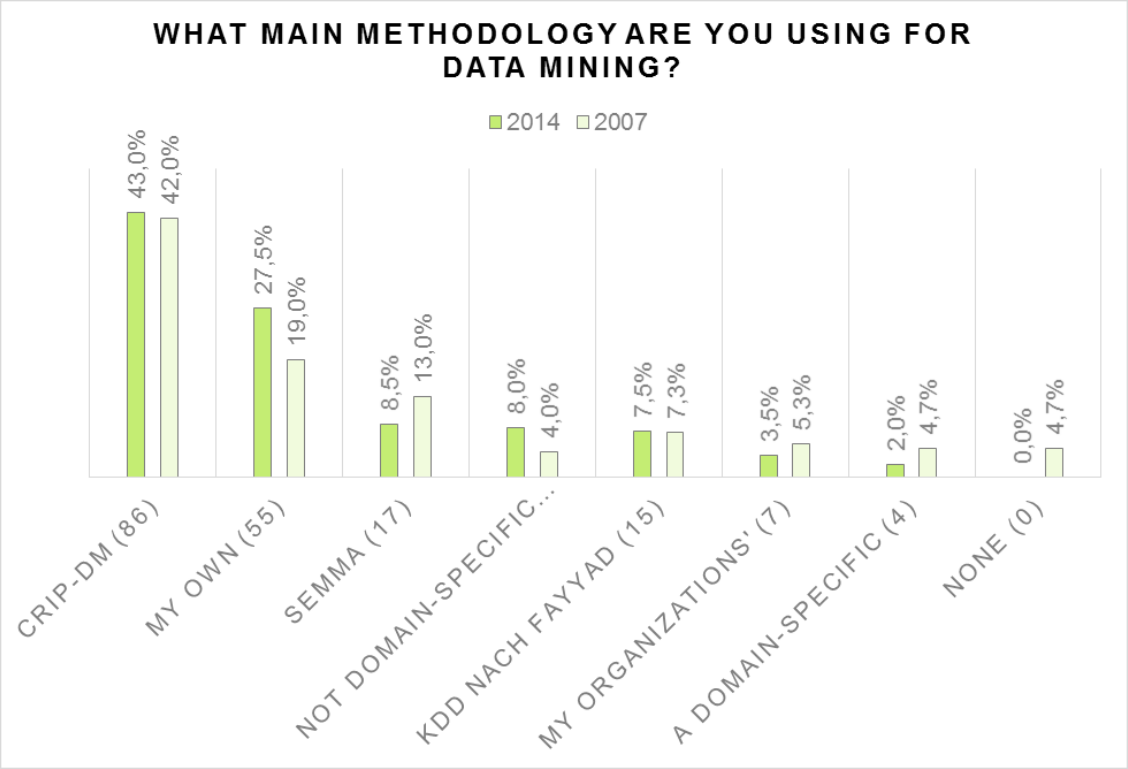

Abbildung 6-1: Umfrage aus den Jahren 2007 und 2014 über die eingesetzten Data Mining Vorgehensmodelle [Piatetsky-Shapiro 2014]

Die Umfrage verdeutlicht, dass neben selbstentwickelten Methoden besonders die standardisierten Verfahren *CRISP-DM* und *SEMMA* eingesetzt werden, auf die daher im weiteren Verlauf näher eingegangen wird.

Des Weiteren wird das Modell *nach Fayyad et al.* vertieft, als eines der bekanntesten und ersten Modelle des KDD.

Die weitere Auswahl erfolgt anhand **Tabelle 5-1** und **Tabelle 5-2**. Auch hier wird deutlich, dass die Modelle *nach Fayyad* et al. und der *CRISP-DM* die meisten Anforderungen erfüllen. Die Modelle *nach Adriaans & Zantinge* und *nach Cooley et al.* hingegen erfüllen nur zwei Anforderungen und scheiden daher aus der Auswahl aus.

Im Hinblick auf den Einsatz im Logistikumfeld stellt der Faktor der Einbettung in das betriebliche Umfeld eine wichtige Komponente dar. Die Modelle, die diese Anforderung nicht erfüllen, werden daher für die weitere Untersuchung nicht weiter betrachtet. Dieses sind die Modelle *nach Anand & Bucher*, *Brachmann & Anhand, Berry & Linoff, John, Edelstein, Petersohn* und *Runkler*.

Im Vergleich der übrigen Modelle stellen sich große Parallelen heraus. So sind die Modelle *nach Caberna et al.* und *Cios et al.* dem *CRISP-DM* zum einen in kurzem zeitlichen Abstand zueinander veröffentlicht und in der Abfolge und Bezeichnung der Prozessschritte sehr ähnlich. Die Auswahl des *CRISP-DM* soll daher genügen.

Ebenso wird das Modell *nach Martines de Pisón* nicht weiter betrachtet, da sich hier Parallelen zum Modell *SEMMA* von SAS ergeben.

Die Modelle *nach Reinhartz & Wirth, Haglin et al, Hippner & Wilde* und *Wrobel et al.* ähneln sich in ihrem Aufbau und der Prozessschrittbezeichnung ebenfalls in großem Maße. Für die weitere Betrachtung wird das Modell *nach Hippner & Wilde* ausgewählt, da es in der direkten Bewertung die meisten Anforderungen erfüllt und für eine detailliertere Darstellung mehr Informationen in der Literatur vorhanden sind als für die vergleichbaren Modelle.

6.1.1. Knowledge Discovery in Databases nach Fayyad

Das bekannteste Vorgehensmodell wurde von Fayyad et al. [1996, S. 10 ff.] entwickelt und im Jahr 1996 veröffentlicht. Durch die Anwendung ihres Modells streben die Autoren das Ziel an, hochwertiges Wissen aus Datenansammlungen zu gewinnen. [Sharafi 2012, S. 60] **Abbildung 6-2** veranschaulicht das Vorgehensmodell *nach Fayyad et al.*

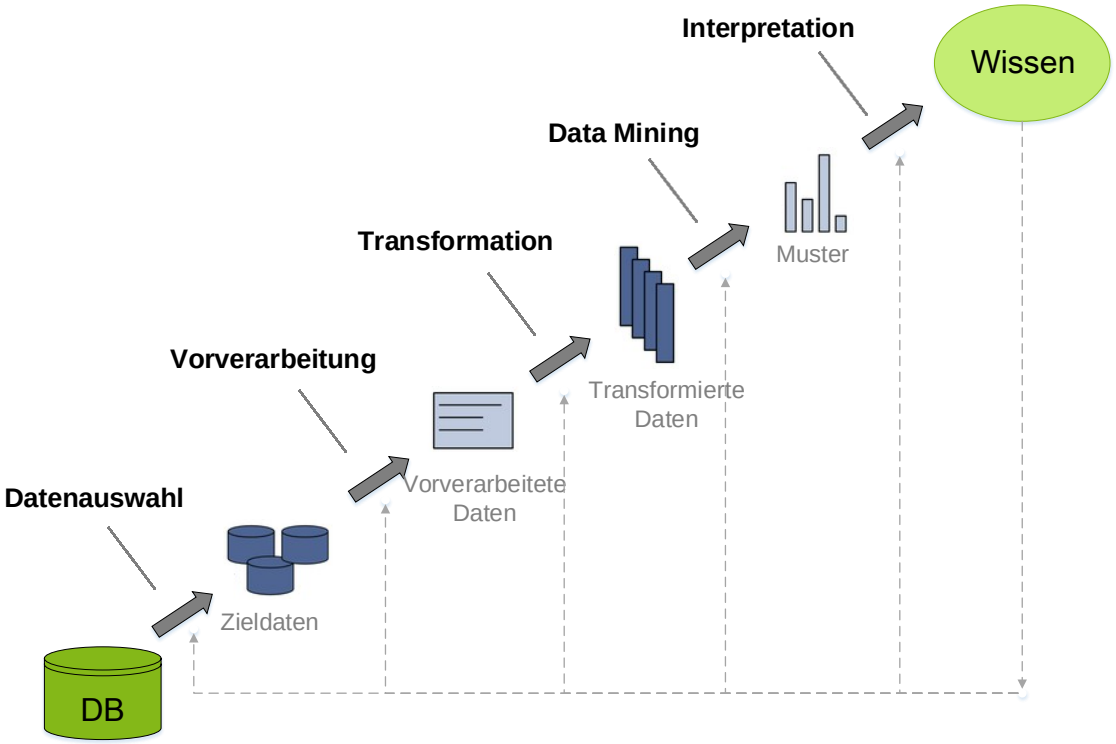

Abbildung 6-2:Der KDD-Prozess *nach Fayyad* et al. [1996, S. 10]

Abbildung 6-2 gibt einen Überblick über den KDD-Prozess *nach Fayyad et al.*, stellt aber nicht jeden Schritt des Vorgehensmodelles im Detail dar. Der Prozess beginnt mit dem Verständnis für die Anwendungsdomäne (Developing and Understanding oft the Application Domain – graphisch nicht dargestellt). Es folgen die Auswahl geeigneter Daten (Crating a Target Data set) und die Vorverarbeitung dieser Daten (Data Cleaning and Preprocessing) zur Schaffung einer Datenbasis. Im vierten Schritt werden die Daten transformiert (Data Reduction and Projection). Der in **Abbildung 6-2** dargestellte nächste Schritt des Data Mining wird in drei Unterpunkte gegliedert. Der Data-Mining-Schritt beinhaltet die Data-Mining-Methodenauswahl (Choosing the Data Mining Task), gefolgt von der Algorithmen- und Hypothesenauswahl (Choosing the Data Mining Algorithm) und schließlich der Mustersuche (Data Mining). Die gewonnenen Muster werden interpretiert (Interpretating Mined Patterns) und gewonnene Erkenntnisse der Domäne zurückgeführt (Consolidating Disvovered Knowledge – graphisch nicht dargestellt). Die neun Teilschritte sind detaillierter beschrieben: [Sharafi 2012, S. 61; Fayyad et al. 1996, S. 10 ff.]

- **Domänenverständnis und Zieldefinition:** Der erste Schritt eines Data-Mining-Projektes ist die Zieldefinition. Darauf aufbauend wird der weitere Prozess geplant. Zur Zieldefinition gehört, dass die durchführende Person ein Verständnis über den Fachbereich und das Wissensgebiet aufbaut und unter Einbeziehung wirtschaftlicher Faktoren Ziele erarbeitet. [Sharafi 2012, S. 61]

- **Datenselektion:** In diesem Schritt werden Daten ausgewählt. Diese Datenmenge wird als Basis für das folgende durchzuführende Knowledge Discovery definiert. Die Selektion beinhaltet die Auswahl der Daten, die Zusammenführung von Daten aus verschiedenen Quellen sowie die Behebung von Integrationsproblemen. Probleme sind beispielsweise die Redundanz in den Daten und das Format der Daten.
- **Datenvorverarbeitung und -bereinigung:** Um die Verfälschung von Ergebnissen zu verhindern, müssen Fehler in den Daten erkannt und eliminiert werden. Auftretende Fehler können Ausreißer, Messfehler, Übertragungsfehler, Inkonsistenzen oder auch fehlende Daten sein.
- **Datentransformation:** Im Zuge der Transformation werden die Daten für die anschließende Data Mining Anwendung vorbereitet. Dazu werden die Datendimensionen reduziert und die Anzahl der zu betrachtenden Variablen angepasst. Zu den Methoden der Transformation gehört zum Beispiel die Normierung, die die Daten verschiedener Wertebereiche vergleichbar macht oder die Generalisierung, welche die einzelnen Sachverhalte zusammenfasst und somit vereinfacht.
- **Data Mining:** Dieser Schritt umfasst die Teilschritte **Data-Mining-Methodenwahl, Algorithmen- und Hypothesenwahl** und **Mustersuche.** Das Ziel dieser Schritte ist es Muster in den vorbereiteten Daten zu finden, d.h. in diesem Schritt des KDD-Prozesses findet die eigentliche Datenanalyse statt. Unterschieden werden zwei Arten von Zielen: Die *Verifikation*, bei der zuvor bekannte Hypothesen geprüft werden und die *Entdeckung* neuer Muster. Zum Erreichen der Ziele stehen verschiedene Methoden zur Auswahl.
- **Musterinterpretation:** Damit aus den gefundenen Mustern gewünschtes Wissen entstehen kann, müssen sie in einer Form präsentiert werden, die eine Interpretation ermöglicht. Die gefundenen Muster werden visualisiert und Wissen generiert. Eine wirksame Methode der Visualisierung ist die graphische Darstellung. Je nach Ergebnis sind unterschiedliche Visualisierungstechniken einzusetzen.
- **Wissensnutzung und –verarbeitung:** Das abgeleitete Wissen muss anschließend verwertet werden. Dazu wird es dokumentiert und entweder zur Nutzung an die Domänen zurückgegeben oder genutzt, um den Prozess erneut zu starten und zu sensibilisieren um noch weiteres implizites Wissen zu entdecken.

6.1.2. Cross Industry Standard Process for Data Mining

Ein sehr verbreitetes Modell stellt der *CRISP-DM* dar. Dieses Modell wurde auf Basis von Erfahrungen und aus dem Bedarf der Praxis entwickelt. Seine Entstehungsgeschichte beginnt bereits 1996. Ein Konsortium aus verschiedenen Unternehmen wie Daimler und SPSS entwickelte basierend auf Erfahrungen ein standardisiertes Konzept. Offiziell wurde das Vorgehensmodell *CRISP-DM* im Jahr 2000. Der Prozess besteht aus sechs Phasen, die in bestimmten Beziehungen zu einander stehen.

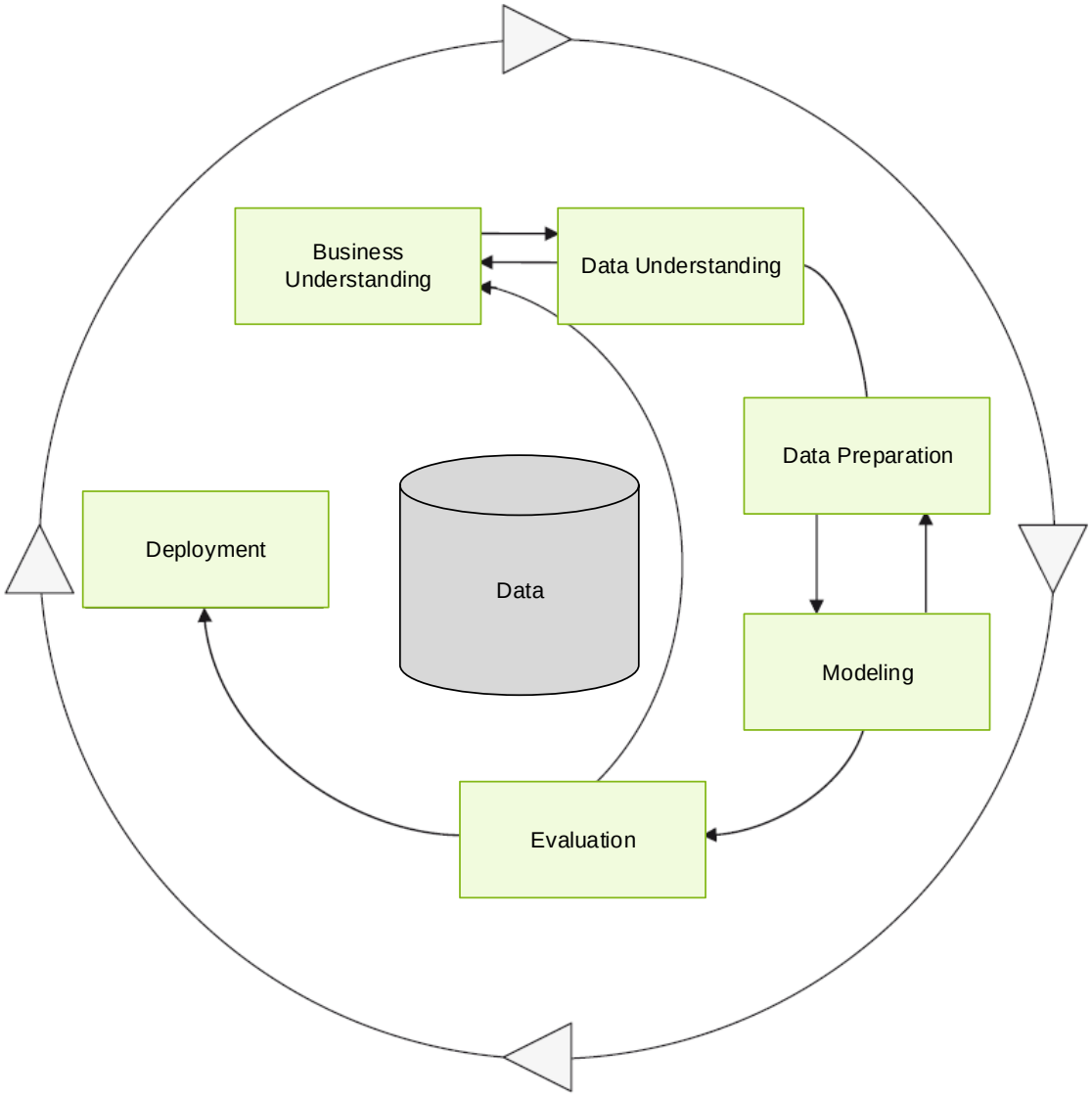

Abbildung 6-3: Das CRISP-DM Vorgehensmodell [Yakut 2015, S. 19]

Die sechs Phasen des Modells sind „Business understanding“, „data understanding“, data preparation“, „modelling“, „evaluation“ und deployment“. Ausgangspunkt für dieses Modell ist der Datenbestand. Diese Daten durchlaufen die genannten Phasen, wobei die Reihenfolge nicht starr vorgegeben ist. **Abbildung 6-3** zeigt eine idealisierte Darstellung des Prozesses. In der praktischen Umsetzung sind nicht alle Phasen klar getrennt. Zudem sind Rückkoppelungen zwischen den Phasen jederzeit möglich und erwünscht. Der äußere Kreis veranschaulicht dazu den zyklischen Charakter des Modells. Gewonnene Erkenntnisse ergeben in einem konkreten Projekt oft weiterführende Fragestellungen. Auch ermöglicht die mehrfache Anwendung des Modells eine ständige Anpassung und regelmäßige Verbesserung. Die praktische Anwendung

dieses Vorgehensmodells wird durch ein begleitendes Handbuch unterstützt, den IBM SPSS Modeler CRISP-DM Guide. Im Folgenden schließt sich eine detailliertere Betrachtung der einzelnen Phasen an [Sharafi 2012, S. 65 f.]:

- **Business Understanding:** Datenanalysen haben das Ziel, aus Unternehmenssicht Vorteile zu schaffen oder Probleme zu beseitigen. Aus Sicht des Auftraggebers liegt der Fokus daher zunächst darauf, das Verständnis des Problems und die Anforderungen zu klären. Aus der Problemstellung und den zu erreichenden Zielen wird ein erster Plan erstellt, wie diese Ziele zu erreichen sind. Ergebnis dieser Phase ist ein erster Vorgehensentwurf.
- **Data Understanding:** In der Phase Data Understanding erfolgt die Datensammlung. Von großer Bedeutung ist besonders die Qualität dieser Daten, um Probleme frühzeitig erkennen zu können und zu verringern. Außerdem sollen erste Hypothesen über verborgene Informationen in der Datensammlung gebildet werden.
- **Data Preparation:** Die Datensammlung wird in diesem Schritt vorverarbeitet. Die Rohdaten werden dabei solange bearbeitet bis die Datensätze in der gewünschten Form vorliegen. Zu den durchzuführenden Aufgaben gehören die Datenauswahl, die Reinigung, die Konstruktion, die Integration und die Formatierung der Daten. Auch eine Strategie zum Umgang mit Ausreißern und fehlenden Werten sollte entwickelt werden.
- **Modelling:** Die Phase Modelling umfasst die Auswahl und den Einsatz verschiedener Techniken zur Datenanalyse. Ausgehend von der Problemstellung werden Verfahren und Algorithmen bestimmt. Da verschiedene Techniken mit verschiedenen Datenformaten arbeiten können, sind Rücksprünge in vorangegangene Phasen möglich.
- **Evaluation:** Die Ergebnisse der Phase Modelling sind auf die anfangs festgelegten Anforderungen hin zu evaluieren. Zum Ende dieser Phase soll die Entscheidung getroffen werden, ob die Ergebnisse den wirtschaftlichen Randbedingungen und Erwartungen in hinreichendem Maße gerecht werden.
- **Deployment:** Die Ergebnisse werden dem Auftraggeber in einer nutzbaren Weise präsentiert. Der Output der Analyse gelangt somit in die Organisation zurück. Die Realisierung und Umsetzung kann nun auf Basis des gewonnenen Wissens umgesetzt werden.

  [Sharafi 2012, S. 66 f.; Yakut 2015, S. 19 f.]

6.1.3. Sample, Explore, Modify, Model, Assess

Sample, Explore, Modify, Model, Assess oder kurz *SEMMA* ist eine Entwicklung des Softwareunternehmens SAS. Anhand der unternehmenseigenen Software SAS Enterprise Miner hat das Unternehmen SAS ein Vorgehensmodell zur Umsetzung von KDD-Projekten entwickelt. Das Modell ist dabei stark an die Software gebunden.

Abbildung 6-4: SEMMA [Yakut 2015, S. 21]

- **Sample:** Im ersten Schritt des Modells geht es darum aussagekräftige Stichproben zu ziehen, die im weiteren Verlauf analysiert werden. Diese Stichproben sollen möglichst die Eigenschaften der gesamten Datenmenge wiederspiegeln. Alle Ausprägungen der Grunddaten sollen in der Stichprobe abgebildet sein. Die Stichprobenbildung soll eine schnelle Bearbeitung ermöglichen.
- **Explore:** Die explorative Datenanalyse hat das Ziel, unbekannte Anomalien und Beziehungen in dem Datenbestand aufzudecken.
- **Modify:** In der dritten Phase werden die Daten für die anschließende Analyse vorbereitet. Dazu gehören beispielsweise die Transformation und Formatierung der Daten, um eine Einheitlichkeit zu erzeugen.
- **Model:** Im vorletzten Schritt wird das eigentlich Data Mining durchgeführt. Ausgewählte Werkzeuge für das Data Mining suchen nach Regeln und Mustern in den bereitgestellten Daten.
- **Assess:** Abschließend erfolgt die Bewertung der Ergebnisse. Der Erfüllungsgrad der Anforderungen wird ermittelt und die Ergebnisse daraufhin bewertet.

6.1.4. Knowledge Discovery in Databases nach Hippner & Wilde

Das Vorgehensmodell *nach Hippner & Wilde* beschäftigt sich mit Schritten, die von der Beschäftigung mit der Anwendung bis zur Umsetzung der Ergebnisse reichen. Der KDD-Prozess wird als dynamisch und iterativ beschrieben. Eine grafische Darstellung zeigt **Abbildung 6-5**.

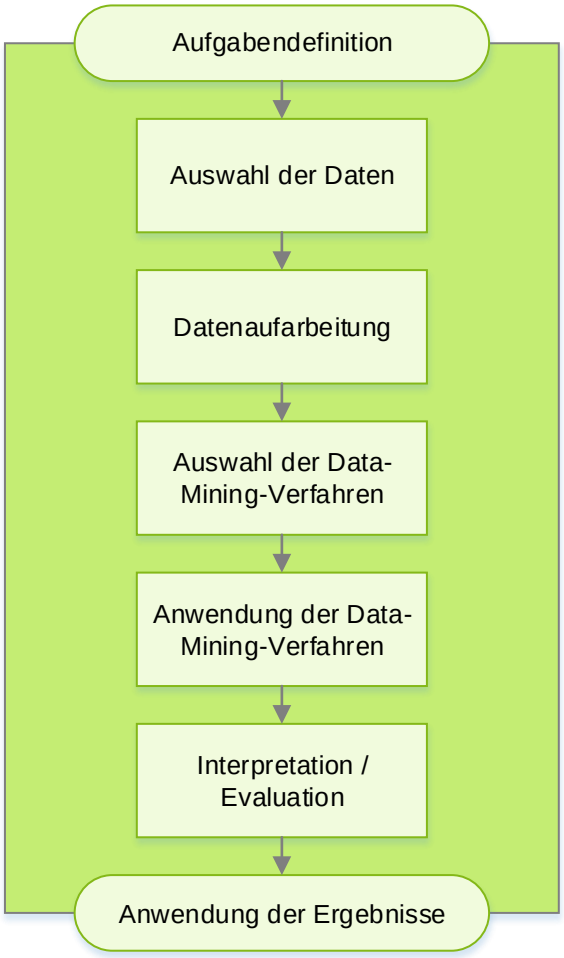

Abbildung 6-5: Der KDD-Prozess von Hippner & Wilde

Die Graphik zeigt einen Vorgang von 7 Schritten. Die Reihenfolge soll dabei nicht als starres Schema betrachtet werden, sondern vielmehr als eine nach der standardmäßigen Reihenfolge angeordnete Methodenauswahl, die zur Entscheidungsfindung helfen soll. Das Modell soll als Toolbox verstanden werden, welche verschiedene Methoden bereithält. Die einzelnen Schritte des Modells sind:

- **Aufgabendefinition:** Am Anfang steht die wirtschaftliche Problemstellung. Aus dieser werden Aufgabendefinition abgeleitet und schließlich die Data-Mining-Ziele erzeugt.
- **Auswahl der Daten:** Aufbauend auf der Zielformulierung werden anschließend Daten gesammelt und auf ihre Qualität geprüft.
- **Datenaufbereitung:** Unter der Datenaufbereitung verstehen *Hippner & Wilde* das Vorbereiten der Daten für die anschließende Analysephase. Darunter fallen die Transformation, die Anreicherung, die Reduktion und die Behandlung von Fehlern in der Datenmenge.

- **Auswahl der Data-Mining-Verfahren:** Die eigentliche Analysephase wird nach *Hippner & Wilde* in zwei Aufgabenbereiche gegliedert. Zunächst findet nur die Auswahl eines geeigneten Werkzeugs statt.
- **Anwendung der Data-Mining- Verfahren:** In der Anwendungsphase werden die gebildeten Datenmengen analysiert.
- **Interpretation/ Evaluation:** Auf jeder Data-Mining-Phase folgt eine Interpretation und Evaluation. Die Ergebnisse werden aufgabenbezogen gefiltert und die gewonnenen Resultate aus betriebswirtschaftlicher Sicht bewertet. Neben den Ergebnissen wird auch der Data-Mining-Prozess selbst geprüft.
- **Anwendung der Ergebnisse:** Der Prozess schließt mit der Integration und Anwendung des neu entstandenen Wissens in die Geschäftsprozesse ab. Die Ergebnisse sollen als Entscheidungsgrundlage oder als Aufgabenstellung für neue KDD-Projekte dienen.

6.2. Knowledge Discovery in Databases im Kontext der Logistik

Nachdem im vorangegangenen **Kapitel 6.1** einige KDD-Vorgehensmodelle vorgestellt wurden, werden diese nun auf logistische Bereiche angewendet. Die Anwendung von Knowledge Discovery in Databases fordert die Umsetzung von Daten in Informationen und von Informationen in Handlungen. Diese Handlungen sollen den Ertrag erzielen, der den Aufwand für die Analyse rechtfertigt. Die Nutzung von Daten stellt einen wichtigen Faktor für die Wettbewerbsfähigkeit dar. Sie können helfen, neue Erträge in den unterschiedlichsten Geschäftsfeldern zu schaffen. In der praktischen Umsetzung stellt sich jedoch heraus, dass eine häufig geforderte Automatisierung des Data-Mining-Prozesses nur partiell möglich ist und dass Analyseprojekte eine feste Einbettung in das betriebliche Umfeld benötigen, um Mehrwert für die Organisation zu erzielen. Um ein möglichst vollkommenes Prozessverständnis hinsichtlich der Ziel- und Einflussgrößen sowie der Wirkungszusammenhänge zu schaffen, werden Erklärungsmodelle herangezogen, die einen Leitfaden darstellen. Data-Mining-Verfahren ermöglichen hypothesenfreie Analysen und die Zusammenführung von menschlichem Wissen mit dem aus Daten explizierten Wissen. Zur erfolgreichen Durchführung bedarf es an interdisziplinärer Zusammenarbeit verschiedener Fachspezialisten um Herausforderungen der Modellkonstruktion und der Interpretation zu bewältigen.

Im weiteren Verlauf dieses Kapitels wird der Einsatz der KDD-Vorgehensmodelle in den Bereichen der Logistik geprüft. Das allgemeine Handlungsschema zur Durchführung von Analyseprojekten für die Einbettung in das betriebliche Umfeld umfasst vier Schritte (**Kapitel 4.1**):

- Spezifikation des Untersuchungsproblems
- Durchführung der Untersuchung
- Umsetzung der Untersuchungsergebnisse
- Evaluierung der Untersuchungssituation

Anhand dieses Handlungsschemas wird nachfolgend der Einsatz von KDD-Projekten mit Bezug auf Bereiche der Logistik vorgestellt.

Spezifikation des Untersuchungsproblems

Bezogen auf das logistische Umfeld ergeben sich die obergeordneten Ziele Leistungserfüllung, Qualitätssicherung und Kostensenkung gemäß der Hauptziele der Logistik, die in **Kapitel 2.1** bereits vorgestellt wurden. In den verschiedenen Teilbereichen der Logistik entstehen daraus abgeleitet unterschiedliche Teilziele. Fragestellungen und Problemstellungen der einzelnen Bereiche sind beispielsweise:

- **Beschaffungslogistik**
    - Welche Lieferanten werden ausgewählt?
    - Entscheidungen über Make-or-Bye und Outsourcing
    - Wie kann ich die Einkaufsplanung verändern um eine bessere Verfügbarkeit zu gewährleisten?
- **Distributionslogistik**
    - Welches ist die effektivste Verteilroute?
    - Können Kunden kategorisiert werden?
    - Wie wird mit Kunden kommuniziert und welchen Nutzen bringt das?
- **Produktionslogistik**
    - Welche Lagerhaltungsstrategie ist auszuwählen?
    - Wie kann die Kommissionierung verbessert werden?
    - Wie kann Lagerkapazität eingespart werden?
    - Wie kann die Maschinenverfügbarkeit optimiert werden?
- **Entsorgungslogistik**
    - Wie wird Abfall schon während der Produktion vermieden / Können Abfälle vermieden werden?
    - Entdecken effizienter Wege der Abfallentsorgung
- **Verkehrslogistik**
    - Wo liegen die Schwachstellen eines Stroms, und wie können sie behoben werden?
    - Welches sind die schnellsten, günstigsten oder risikoärmsten Verbindungen zwischen Standorten?

Die Forderung nach geeignetem Datenmaterial ist mittels der vorgestellten Logistikmodelle theoretisch erfüllt. Wie in **Kapitel 2.2** dargestellt, werden Logistiknetzwerke anhand von Knoten und Kanten modelliert, die auf einer geeigneten Datenbasis beruhen. Transportmodelle beinhalten beispielsweise Datensätze und Informationen über Mengen, Standorte der Quellen und Senken, je nach Auslegung des Modells auch Kosten und Zeiten. Faktoren, die zum Beispiel in Modelle der Standortwahl mit einfließen, müssen gegebenenfalls in vergleichbare und für Analysen verwertbare Datensätze transformiert werden. In der Theorie sind Daten in der Logistik vielfach vorhanden. Da eine lückenlose Datensammlung jedoch oft nicht gegeben ist, muss die Speicherung der relevanten Daten, die Vollständigkeit und Richtigkeit je nach Projekt geprüft werden.

Mit Bezug auf das Modell *nach Fayyad et al.* wird die Spezifikation des Untersuchungsproblems in den Phasen „Domänenverständnis und Zieldefinition“, „Datenselektion“, „Datenvorverarbeitung und -bereinigung“ und „Datentransformation“ behandelt. Am Anfang steht die Identifizierung der Ziele. Allgemein soll ein Verständnis über den zu untersuchenden Bereich entwickelt werden, bevor die Auswahl der Daten und deren Vorbereitung und Transformation folgt. Betont wird im Vorgehensmodell *nach Fayyad et al.* auch der Umstand, dass die definierten Ziele in messbare Data-Mining-Ziele übersetzt werden müssen.

Auch bei dem *CRISP-DM* umfasst die Spezifikation des Untersuchungsproblems mehrere Modellphasen. Betroffene Phasen sind „Business Understanding“, „Data Understanding“ und „Data Preparation“. Im ersten Schritt wird das Untersuchungsumfeld betrachtet, auf dessen Basis Ziele formuliert werden. Betonung liegt in diesem Schritt auf der Zusammenarbeit von Datenanalysten und Kunden, um Kundenwünsche und realisierbare Möglichkeiten zu Anwendungszielen zu integrieren. Im Anschluss erfolgt die Situationsbewertung, bei der eine Analyse und Beurteilung der vorhandenen Ressourcen, der äußeren Bedingungen, der zu treffenden Annahmen und der Erfordernisse für die anstehende Wissensentdeckung vorgenommen wird. Bei den Ressourcen wird sowohl die vorhandene Hardware geprüft als auch die Software, d.h. das Vorhandensein unterstützender Data-Mining-Werkzeuge. Die zu analysierenden Daten sind nach Art und Quelle zu bewerten und bereits vorhandenes Wissen und Hintergrundwissen ist zu prüfen. Im Hinblick auf das relevante Personal sind die Qualifikationen und die zeitliche Verfügbarkeit der betroffenen Personen zu prüfen. Des Weiteren sollte das Projekt auf mögliche Risiken geprüft werden; darauf aufbauend ist für denkbare Risikofälle ein Plan zu erstellen, um im Eintrittsfall schnellstmöglich zu reagieren. Die formulierten Anwendungsziele werden schließlich in messbare Data-Mining-Ziele übersetzt. Diese Ziele müssen anhand definierter Erfolgsfaktoren qualitativ und quantitativ bewertbar sein. Bei der Auslegung sollte bereits die Deploymentaufgabe berücksichtigt werden. Ergebnis dieser Phase ist der Projektplan, der Faktoren wie die Dauer des Projekts, die benötigten Mittel, die einzelnen Phasen mit Herausarbeitung der kritischen Schritte, wichtige Entscheidungspunkte und Iterationen enthält. Im besten Fall findet die Auswahl der Data-Mining-Werkzeuge bereits in dieser Phase statt, um im folgenden Schritt bereits zielgerichtet Daten auszuwählen. Die Datensammlung besteht aus verfügbar stehenden Daten und der Bestimmung von notwenigen Daten, die generiert werden müssen. Die ausgewählten Daten werden anhand ihrer Attribute wie dem Format und Beziehungen betrachtet und verfeinert. In diesem Schritt können gegebenenfalls bereits neue Hypothesen erzeugt werden, die zu einer Korrektur der bisherigen Data-Mining-Ziele führt. Abschließend wird die Datenmenge auf ihre Qualität geprüft, indem die Vollständigkeit und die Korrektheit untersucht werden. Ziel der Datenpräparationsphase ist die Bereitstellung eines geeigneten Datensatzes für die anschließende Analyse. Diese Phase der Vorbereitung ist häufig die arbeitsintensivste. Aus den gesammelten Daten werden unter technischen Bedingungen wie der Qualität und dem maximalen Datenvolumen Daten ausgewählt. Fehlende Werte werden durch Standartwerte erzeugt und Redundanzen entfernt. Daten werden transformiert und aus verschiedenen Quellen zu einer homogenen Menge zusammengefasst. Abschließend muss die Syntax der Daten vereinheitlicht werden. Alle durchgeführten Veränderungen sind zu dokumentieren, um diese Aktivitäten in späteren Phasen rekonstruieren zu können.

Das Modell *SEMMA* ordnet der Phase Spezifikation des Untersuchungsproblems ebenfalls drei Prozessschritte zu, jedoch fehlt dabei eine modellunterstütze Zieldefinition. In der ersten Phase, dem „Sampling“ werden bereits Daten gesammelt und teilweise transformiert. Weiter werden die Daten in Trainingsdaten und Testdaten unterteilt. Die Trainingsdaten werden für die Durchführung der Analyse verwendet, die Testdaten für die Validierung der Ergebnisse der Analyse. Die

Datenmenge wird außerdem mit Hilfe statistischer Methoden bearbeitet, um das Volumen zu reduzieren. In der Phase „Explore“ soll ein Verständnis über die Daten geschaffen werden, um Trends oder beispielsweise Untergruppen zu identifizieren und den Prozess zu verfeinern. In der Modify-Phase erfolgt die Vorverarbeitung der Daten. Hier werden die Daten vergleichbar mit dem *CRISP-DM-Modell* auf das eigentliche Analyseverfahren vorbereitet.

Die Phasen *nach Hippner & Wilde*, die sich auf die Spezifikation des Untersuchungsproblems beziehen, sind „Aufgabendefinition“, „Auswahl der Daten“ und „Datenaufbereitung“. Die Problemstellung wird bestimmt und daraus Ziele abgeleitet. Diese werden in Data-Mining-Ziele übersetzt und in einem Projektplan zusammengestellt. Der Projektplan dient als Basis für die anschließende Datenauswahl. Diese werden katalogisiert und einer qualitativen Bewertung unterzogen. Aufbauend darauf erfolgt die Datenvorbereitung. Darunter verstehen *Hippner & Wilde* die Transformation, die Anreicherung, die Reduktion und die Behandlung von Fehlern.

Durchführung der Untersuchung

Die Durchführung der Untersuchung bezieht sich auch im logistischen Umfeld auf den Data-Mining-Vorgang. Dabei gibt es eine große Bandbreite verschiedener Werkzeuge, die hier zur Anwendung kommen. Eine Beschreibung der verschiedenen Anwendungsbereiche und der dazu gehörenden Aufgabenstellung sowie eine Auswahl an Methoden werden in **Kapitel 3.2** aufgezeigt. Folgende Auflistung zeigt beispielhafte Anwendungsmöglichkeiten aus den Bereichen der Logistik. Die jeweilige Zuordnung ist dabei nicht immer trennscharf, da Überschneidungen bei Analysen und Einsatzbereichen vorkommen.

- **Segmentierung**

  Bespiel: Bestandsmanagement – In der Lagerverwaltung bieten Data-Mining-Verfahren großes Unterstützungspotenzial. Um einen störungsfreien Materialfluss zu gewährleisten, wird Material nach Merkmalen gruppiert. Material, welches zeitnah an gleichen Standorten benötigt wird, kann so auch im Lager an gleichen Lagerstandorten aufbewahrt oder so gelagert werden, dass bei der Kommissionierung effiziente Routenplanung möglich wird.

- **Klassifikation**

  Beispiel: Lieferantenmanagement – Im Rahmen von Supply-Chain-Management nimmt das Lieferantenmanagement einen hohen Stellenwert ein. Die Auswahl eines Lieferanten kann im Wettbewerb entscheidend sein. Bei der Vergabe ist mittels Data Mining eine Klassifizierung anhand Logistik- und Qualitätskennzahlen behilflich.

- **Prognose**

  Beispiel: Maschinenwartung – Unerwartete Ausfälle von Anlagen einer Fertigungskette können erhebliche Folgen haben, wie etwa den Stillstand der Produktion. Die vorbeugende Wartung spielt daher eine wichtige Rolle. Die Bildung von Algorithmen durch Data-Mining-Analysen ermöglicht die Erstellung von Prognosen, wann Wartungen notwendig sind. Dadurch kann auch die Lebensdauer der Anlagen und die Ersatzteilversorgung optimiert werden.

- **Assoziation**

  Beispiel: Qualitätsüberwachung - Mittels Data Mining wird automatisch nach kritischen Mustern in Produktionsdaten gesucht, die Mängel verursachen. Ursachen-Wirkungs-

Zusammenhänge können in die bisherigen Prozesse eingebracht werden und verbessern diese langfristig.

Das Modell *nach Fayyad et al.* beschreibt die Durchführung der Untersuchung anhand von drei Prozessschritten, der „Data-Mining-Methodenwahl“, der „Algorithmen- und Hypothesenwahl“ und der „Mustersuche“. Auf diese Weise wird deutlich, dass der Data-Mining-Schritt aus einzelnen Teilschritten und Entscheidungen besteht. Entsprechen der Problemstellung und Zielsetzung sind geeignete Methoden zu wählen, denen wiederum verschiedene Algorithmen und Werkzeuge untergeordnet sind. Unerwähnt bleibt der Einsatz spezialisierten Fachpersonals.

Der *CRISP-DM* umschreibt die Durchführung der Untersuchung im Prozessschritt „Modeling“. Dieser umfasst ebenfalls die Auswahl geeigneter Data-Mining-Werkzeuge. Betont wird, dass die jeweiligen Tools anhand der gegebenen Parameter von einem geeigneten Anwender eingestellt werden müssen und dadurch eine Anpassung der Daten erforderlich sein kann. Ein Rücksprung in vorhergegangene Phasen kann an dieser Stelle notwendig sein. Vor Analysestart ist ein Testlauf durchzuführen, um die Qualität und Gültigkeit des Modelles zu bestimmen, weshalb an dieser Stelle die Trennung der Daten in Trainings- und Testdaten stattfindet. Die Anwendung der Analyse ist nutzergesteuert, der Bediener legt die Parameterbelegung fest. Die Ergebnisse und die zugehörigen Parameter sind zu dokumentieren, um Unterschiede der Modelle in Beziehung mit der dazugehörigen Einstellung bringen zu können. Die Modelle sind mittels der Testdaten zu validieren, um sie mit den zuvor festgesetzten Data-Mining-Zielen zu vergleichen. Bei nicht zufriedenstellenden Ergebnissen wird die Modeling-Phase wiederholt.

Die Phase „Model“ stellt im Rahmen des *SEMMA*-Modells die Anwendung der Data-Mining-Verfahren dar. Zunächst werden die verschiedenen Verfahren in der Modeling-Phase ausgewählt. Zur Verfügung stehen neuronale Netzte (Neural networks), baumbasierte Modelle (Tree based models), Logistikmodelle (logistic models) und andere statistische Modelle (other statistical models). Unerwähnt bleibt der Einsatz spezialisierten Fachpersonals.

Der Ansatz von Hippner & Wilde gliedert die Durchführung der Untersuchung in zwei Phasen. In der ersten dieser Phasen findet die Auswahl der Data-Mining-Methode statt, womit auch die Auswahl der Werkzeuge gemeint ist. Im anschließenden Schritt erfolgt die Anwendung des Verfahrens.

Umsetzung der Untersuchungsergebnisse

Bei der praktischen Umsetzung von KDD-Projekten erfüllen bei weitem nicht alle der entdeckten Muster die Kriterien, die zu umsetzbaren Handlungen führen. Vor der Interpretation der Ergebnisse ist es daher sinnvoll diese zu filtern. Ursachen mangelhafter Interessantheit sind in **Kapitel 3.2** aufgeführt. Beispiele dafür sind:

- **Mangelnde Gültigkeit**

  Beispiel: Dynamische Veränderungen der Datensätze oder veränderbares Verhalten von Kunden führen zu Wertverlusten der Analyseergebnisse.

- **Mangelnde Neuartigkeit**

  Beispiel: Die Analyse ergibt, dass die Lagerung von Holzkohle und Grillanzündern am selben Ort vorteilhaft ist.

- **Mangelnde Nützlichkeit**

  Beispiel: Die Analyse ergibt, dass der Einsatz eines vollautomatischen Präzisionswerkzeuges die Bearbeitung vereinfacht, effizienter und schneller macht. Aufgrund der Kosten ist die Anschaffung des Werkzeuges jedoch irrelevant.

- **Mangelnde Verständlichkeit**

  Beispiel: Eine Ansammlung von Werten ohne Bezug oder Ordnung, ist für den Menschen kaum zu bewerten. Es sind graphische Darstellungen und natürlich sprachliche Beschreibungen wie Wenn-Dann-Regeln geeignet.

Die Ergebnisse der Analyse sollten einen definierten Filter durchlaufen, der in das Data-Mining-System integriert ist, um dem Anwender eine eingeschränkte Auswahl darzubieten. Die Interpretation der Ergebnisse erfordert im nächsten Schritt analytisch-methodisches Know-how und umfassende Fachkenntnisse. Zusätzliches Wissen über den Fachbereich sowie geeignete Präsentationswerkzeuge unterstützen des Schritt der Interpretation.

Nach Fayyad et al. werden diese Kriterien in den Phasen „Musterinterpretation“ und „Wissensnutzung und –verarbeitung“ bearbeitet. Nach welchen Kriterien die Interpretation und Bewertung abläuft, wird im Modell nicht näher erläutert. Abgeschlossen wird das Modell von *Fayyad et al.* mit der Anwendung des resultierten Wissens und der Generierung von Berichten.

Der *CRISP-DM* verwendet für die Umsetzung der Untersuchungsergebnisse die Phasen „Evaluation“ und „Deployment“. An die Data-Mining-Phase anschließend werden die Ergebnisse einer kritischen Betrachtung unterzogen. Sie werden hinsichtlich der gesetzten Ziele untersucht, im Anschluss in eine anwenderfreundliche Sprache übersetzt und ihre Neuartigkeit beurteilt. Anhand der gewonnenen Erkenntnisse über die entdeckten Muster wird über das weitere Vorgehen entschieden. Danach erfolgt die Übergabe der Ergebnisse in die Deployment-Phase; weitere Iterationen der Modeling-Phase können gestartet oder neue Projekte angestoßen werden. Auch mehrere Alternativen können parallel verfolgt werden. Die Verwendung der Ergebnisse erfolgt in der „Deployment-Phase“. Diese Phase ist die Schnittstelle zwischen dem Analysen und dem Kunden. Für die erfolgreiche Anwendung ist ein Ausführungsplan zu erstellen, der die Strategie zur Umsetzung in einzelnen Schritten organisiert. Im *CRISP-DM* wird außerdem darauf hingewiesen, dass eine strategische Überwachung der Ergebnisse nach Abschluss des Projekts zu planen ist.

Im *SEMMA*-Modell findet die Umsetzung der Untersuchungsergebnisse in der Phase „Assess“ statt. Die Ergebnisse werden hinsichtlich ihrer Brauchbarkeit und dem Maß der Zielerfüllung gemessen. Das Resultat dieser Phase können neue Miningziele für nachfolgende Projekte sein oder falls die Ergebnisse nicht von ausreichender Qualität sind, einen Rücksprung in die Explore-Phase veranlassen. Die Anwendung von Handlungskonsequenzen findet in diesem Ansatz keine Berücksichtigung.

Der KDD-Prozess von *Hippner & Wilde* verwendet für die Umsetzung der Untersuchungsergebnisse die Phasen "Interpretation/ Evaluation" und „Anwendung der Ergebnisse“. Nach der Data-Mining-Phase werden die gefundenen Muster nach Relevanz gefiltert und bewertet. Das KDD-Projekt schließt mit der Verwendung des neuen Wissens ab. Handlungskonsequenzen werden in die operativen Geschäftsprozesse eingebettet, dienen als Entscheidungshilfen oder geben den Anstoß für weitere KDD-Projekte. Die beiden Phasen "Interpretation/ Evaluation" und „Anwendung der Ergebnisse“ *nach Hippner & Wilde* sind dabei nicht streng getrennt.

Evaluierung der Untersuchungssituation

In der Logistik beginnt ein KDD-Projekt mit der Frage nach den mit der Analyse verfolgten Zielen und Zielvorgaben (Soll-Werte). Ziel des KDD ist es, durch die Anwendung des gewonnenen Wissens eine Verbesserung gegenüber der Ausgangssituation herzustellen. Um den erzielten Nutzen messen zu können, werden Kennzahlen benötigt, die einen Ist-Zustand mit den Soll-Werten vergleichbar machen. Kennzahlen der Logistik sind zum Beispiel:

- **Administrative Ebene der Logistik**
    - Logistikkosten: Ergebnisse der Logistikkostenabrechnung
- **Dispositive Ebene der Logistik**
    - Lieferzuverlässigkeit: Verhältnis von termingerecht gelieferten Bedarfen zur Gesamtzahl der Aufträge
    - Lieferbereitschaft: Relation von ab Lager erfüllten Anforderungen zu der Gesamtzahl
    - Leistungsauslastung eines Lagers: Anzahl belegter Plätze und Anzahl freier Plätze
- **Operative Ebene der Logistik**
    - Mengen und Strukturdaten: Anzahl täglicher Anlieferungen

Eine ganzheitliche Betrachtung ermöglicht Performance Measurement, ein Prozess zur Identifizierung von Leistungsindikatoren (Kennzahlen), die eine Aussage über das Maß der Zielerreichung im Hinblick auf Qualität, Zeit und Kosten ermöglichen (Performance). Ziel dabei ist die Ermittlung, ob Ergebnisse den Intensionen entsprechen. Es findet eine mehrdimensionale Leistungsmessung statt, die herkömmliche Kennzahlensysteme (Messung von Umsatz, Gewinn) um Einflussgrößen wie Kundenzufriedenheit, Anzahl an Neukunden und Leitung von Mitarbeitern erweitert. Beispielsweise die Balanced Scorecard, ein Konzept zur Messung, Dokumentation und Steuerung von Strategiefindung und -umsetzung eines Unternehmens oder einer Organisation im Hinblick auf seine Vision gehört zu den bekanntesten Performance Measurement Systemen die in der Praxis eingesetzt werden

Der KDD-Prozess im Modell *CRISP-DM* behandelt die Evaluation der Untersuchungssituation in zwei Phasen. So werden neben den beschriebenen Vorgängen zur Umsetzung der Untersuchungsergebnisse in den Phasen „Evaluation“ und „Deployment“ auch Forderungen der Evaluation der Untersuchungssituation nachgegangen. Die Analysemethode und die Gesamtheit der Schritte des Prozesses werden überprüft um Verbesserungspotenziale zu erkennen. Der Prozess wird rückblickend betrachtet um Bereiche zu identifizieren, die bis zum aktuellen Zeitpunkt übersehen wurden oder denen nicht ausreichend Relevanz zugeordnet wurde. Die einzelnen Phasen werden dafür hinsichtlich ihrer Durchführung betrachtet. Die Deployment-Phase schließt mit einer Untersuchung des gesamten Projektes ab. Dazu wird ein Bericht erstellt, der Angaben zu der Qualität, der Zeit, den Kosten und der Gründe für aufgetretene Abweichungen von den vordefinierten Zielen beinhaltet. Weiter wird ein Ausblick in Form von Implementierungsplänen gegeben. Anhand von Interviews aller Projektbeteiligten werden auch Arbeitsbedingungen während der Projektdurchführung und Erfahrungen für spätere Projekte festgehalten.

Im Modell von *Hippner &Wilde* wird in der Phase "Interpretation/ Evaluation“ die Evaluierung der Untersuchungssituation während der Interpretation mit einbezogen. An dieser Stelle soll der

Prozess des Data Mining geprüft werden. Neben der Untersuchung der Methode bleibt die Untersuchung der Durchführung jedoch unerwähnt.

Das Modell von *Fayyad et al.* und das Modell *SEMMA* gehen auf die Evaluierung der Untersuchungssituation in keiner ihrer Phasen ein.

Zusammenfassend zeigt dieses Kapitel eine Übersicht über die verschiedenen Herangehensweisen der einzelnen KDD-Vorgehensmodelle im Einsatz in Bereichen der Logistik. Dazu wurden die Modelle anhand spezifischer Herausforderungen für den praktischen Einsatz von Analyseprojekten im industriellen Umfeld untersucht und erörtert, in welchem Maße die Modelle die Anforderungen erfüllen.

6.3. Vergleich der ausgewählten Vorgehensmodelle mit abschließender Beurteilung

Die in der Literatur unterschiedenen Vorgehensmodelle des KDD unterteilen den Prozess grundlegend nach dem gleichen Schema. Auch die vier in **Kapitel 6.2** vorgestellten Modelle wählen diese Vorgehensweise, wobei sich die Aufteilung und Benennung der Phasen unterscheiden. Es lässt sich jedoch feststellen, dass die inhaltlichen Unterschiede eher gering sind. Die ersten Phasen der Modelle von *Fayyad et al.* (Domänenverständnis und Zieldefinition) und *Hippner & Wilde* (Aufgabendefinition) und dem *CRISP-DM* (Business Understanding) verfolgen vom Grundgedanken das Ziel, einer Einarbeitung in das Projekt und ein Verständnis über die Ziele des Projekts zu erlangen. Das Modell *SEMMA* verzichtet im Vergleich zu den anderen auf eine Definitions- oder Einarbeitungsphase. Die Phase der Datensammlung und die Phase der Interpretation der Data-Mining-Ergebnisse sind in allen Modellen zwar in unterschiedlicher Bezeichnung, (Datenselektion, Sample, Data Understanding, Auswahl der Daten – Musterinterpretation, Assess, Evaluation, Interpretation/Evaluation) aber mit der gleichen Intension vertreten. Größere Unterschiede sind in der Vorverarbeitung zu erkennen. Die Phase im *CRISP-DM* (Data Preparation) und dem Modell von *Hippner & Wilde* (Datenaufbereitung) sind dabei sehr ähnlich strukturiert. *Fayyad et al.* unterteilt diese Phase in die Schritte „Datenvorverarbeitung/Bereinigung“ und „Datentransformation“. Das Ziel dieser Schritte ist gleichzusetzten mit denen der Phasen „Datenaufbereitung“ und „Data Preparation“. Die beiden Phasen „Explore“ und „Modify“ im SEMMA-Prozess haben auch die Vorbereitung der Daten auf die Data-Mining-Durchführung zum Ziel, jedoch bilden diese teilweise schon in der Explore-Phase durchgeführten Analysen den Hauptunterschied zu den anderen Modellen. Der Schritt des eigentlichen Data Mining wird in den Vorgehensmodellen von *Fayyad et al.* in drei Teilschritte, in dem Modell von *Hippner & Wilde* in zwei Teilschritte zerlegt. Vorbereitend auf die Data-Mining-Analyse wird *nach Hippner & Wilde* zunächst in einem eigenen Prozessschritt das Data-Mining-Verfahren ausgewählt. *Nach Fayyad et al.* werden zunächst die Methode und anschließend auch Algorithmen und Hypothesen in einer eigenen Phase ausgewählt. Der abschließenden Umsetzung der Ergebnisse wird im *SEMMA* Modell kein Prozessschritt zugeordnet. Das Modell endet mit der Interpretationsphase. Die übrigen drei Modelle gehen über die Interpretation mit einer Anwendungsphase hinaus.

Weiter unterscheidet sich auch das Einsatzgebiet der Modelle. Beim KDD-Vorgehensmodell von *Fayyad et al.* wird von einem umfassenden Datenbestand ausgegangen. Die Integration von Experten des Anwendungsgebietes ist nur bei der Interpretation vorgesehen. Das Modelle *SEMMA,* der *CRISP-DM* und das Modell von *Hippner & Wilde* werden vorwiegend in der

Wirtschaft eingesetzt, wobei speziell das Modell von *Hippner & Wilde* für den Einsatz im Marketing gedacht ist.

In der Anwendung der Vorgehensmodelle weist der Detaillierungsgrad deutliche Unterschiede auf. Der *CRISP-DM* ist insgesamt ein sehr umfassender Ansatz. Als Hilfsmittel steht ein eigenes Handbuch zur Verfügung (IBM SPSS Modeler CRISP-DM Guide). Das Modell gibt eine strukturierte Vorgehensweise in unterschiedlichen Detaillierungsebenen vor. In den Modellen von *Fayyad et al.* und *Hippner & Wilde* sind sehr große Ähnlichkeiten auch zum *CRISP-DM* zu finden, jedoch werden die Phasen nicht so detailliert in ihre einzelnen Aufgaben gegliedert. Das Modell von *Hippner & Wilde* legt die Betonung auf fließende Übergänge in der Umsetzung und anwendungsabhängige Anpassung der Reihenfolge der Schritte. In der praktischen Umsetzung zeigt dieses Modell hohe Flexibilität. Im Vergleich der Anwendungsfreundlichkeit zeigt das *SEMMA*-Modell die größten Abweichungen. Rücksprünge in andere Phasen des Modells sind nur von der Assess zur Explore-Phase möglich. Das Fehlen der ersten und letzten Phase sowie die teilweise in der Explore-Phase stattfindenden Voranalysen gehören außerdem zu den Hauptunterschieden.

Auf die Kosten für den Einsatz der Vorgehensmodelle bezogen, ist ein Vergleich nicht sinnvoll. Die anfallenden Gesamtkosten richten sich nach Faktoren, die je nach Projekt deutlichen Schwankungen unterliegen.

Ein Vergleich der vorhandenen Data-Mining-Software zeigt jedoch große Unterschiede. Es gibt eine Reihe an kostenfreier wie auch kostenpflichtiger Software. Zu den Kostenpflichtigen gehören beispielsweise der IBM SPSS Modeler und der SAS® Enterprise Miner. Der IBM SPSS Modeler unterstützt über das Data Mining hinaus den gesamten KDD-Prozess und ist speziell für den *CRISP-DM* entwickelt worden. Dem gegenüber ist der *SEMMA*-Prozess speziell auf die Data-Mining-Software des Unternehmens SAS (SAS® Enterprise Miner, SAS® Factory Miner) zugeschnitten, weshalb *SEMMA* stark an diese gebunden und wenig flexible einsetzbar ist.

Tabelle 6-1 fasst die Gegenüberstellung der Vorgehensmodelle vergleichend zusammen:

Tabelle 6-1: Gegenüberstellung der ausgewählten Modelle

Model	Fayyad	SEMMA	CRISP-DM	Hippner & Wilde
Jahr	1996	1997	2000	-
Anzahl an Schritten	9	5	6	7
Einsatzgebiet	Wissenschaft	Wirtschaft	Wirtschaft	Wirtschaft/ Marketing
Vorgehen	Domänenverständnis und Zieldefinition	-	Business Understanding	Aufgabendefinition
	Datenselektion	Sample	Data Understanding	Auswahl der Daten
	Datenvorverarbeitung Bereinigung	Explore	Data Preparation	Datenaufbereitung
	Datentransformation	Modify		
	DM Methodenwahl		Modelling	Auswahl der Data-Mining-Verfahren
	Algorithmen und Hypothesenwahl	Model		Anwendung der Data-Mining-Verfahren
	Mustersuche			
	Musterinterpretation	Assess	Evaluation	Interpretation / Evaluation
	Wissensnutzung und Verarbeitung	-	Deployment	Anwendung der Ergebnisse
Praxiseinsatz	hoher Bekanntheitsgrad	häufiger praktischer Einsatz	häufiger praktischer Einsatz	-
Software	-	SAS® Enterprise Miner	IBM SPSS Modeler	-
Flexibilität	mittel	gering	hoch	sehr hoch
Detaillierung	hoch	mittel	sehr hoch (Handbuch)	hoch

Auf Basis der angeführten Unterschiede erfolgt an dieser Stelle eine abschließende Beurteilung der Ergebnisse. Ziel der Beurteilung ist es, nach Möglichkeit Empfehlungen bezüglich der KDD-Vorgehensmodelle im Einsatzgebiet der Logistik zu geben.

Für die Anwendung von KDD-Projekten im industriellen Umfeld sind folgende Aspekte von besonderer Bedeutung:

- Die Analyse mittels Data Mining dient als Mittel zum Zweck, der Fokus liegt auf den resultierenden Maßnahmen
- Die Möglichkeit zur Messung der erzielten Ergebnisse
- Für die Umsetzung von KDD-Projekten ist Fachwissen, Datenwissen und Data-Mining-Know-how erforderlich
- Es handelt sich um einen Kreislauf, bei dem schrittweise Verbesserung angestrebt wird
- Das Projekt kann eine regelmäßig oder eine einmalige Aktivität sein

Mit Blick auf die vier betrachteten Modelle wird deutlich, dass alle Modelle einem groben gleichen Leitfaden folgen, jedoch auch erhebliche Unterschiede in den Modellen vorhanden sind. Aufgrund der gegebenen Eigenschaften ist *SEMMA* nur bedingt für den Einsatz in der Logistik geeignet. *SEMMA* ist an die betriebseigene Software von SAS gebunden und umfasst wenige Prozessschritte, die das Modell in das betriebliche Umfeld eingliedern. Die Einbeziehung von Expertenwissen fehlt gänzlich.

Dagegen ist das Vorgehensmodell von *Fayyad et al.* mit neun Prozessschritten, weitaus differenzierter. Dieser Detaillierungsgrad, der sich auf eine einzige Ebene bezieht, ist jedoch für die Anwendung im industriellen Umfeld nicht optimal. Weiter geht das Modell von einem großen vorhandenen Datenbestand aus. Die Integration von Experten erfolgt lediglich in der Umsetzung der Ergebnisse. Die Anwendung des Vorgehensmodells von *Fayyad et al.* ist daher in der Wissenschaft angesiedelt.

Das Vorgehensmodell von *Hippner & Wilde* und der *CRISP-DM* ähneln sich in großem Maße. Die Modelle beinhalten die ausschlaggebenden Prozessschritte in zyklischer Anordnung. Besonders der *CRISP-DM* hat einen hohen Detaillierungsgrad bezüglich der Aufgaben in den einzelnen Prozessschritten. Der *CRISP-DM* gehört, wie in **Kapitel 6.1** deutlich wird zu den bekanntesten Modellen und wird bereits praktisch angewendet. Das Vorgehensmodell von *Hippner & Wilde* ist für die Anwendung im Marketing vorgesehen, lässt sich aufgrund seiner Beschaffenheit jedoch auch für den Einsatz in der Logistik heranziehen.

Der Einsatz der Modelle im Logistikbereich ist mit Einschränkungen folglich durchaus möglich, jedoch sind einige Aspekte zu bedenken, die eine Anpassung der Modelle erforderlich macht. Der Einsatz im industriellen Umfeld bringt spezifische Faktoren mit sich, die in den gezeigten KDD-Vorgehensmodellen zum Teil ganz vernachlässigt werden oder nur in mangelndem Maße Berücksichtigung finden. Der Ansatz eines KDID-Vorgehensmodells, wie in **Kapitel 4.3** angedeutet, kann ein zielführender Ansatz dahingehend sein.

7. Prototypische Anwendung eines ausgewählten Vorgehensmodells

Nachdem in den vorigen **Kapiteln 6.2 und 6.3** die Vorgehensmodelle auf die Anwendbarkeit in der Logistik untersucht wurden, erfolgt an dieser Stelle der Einsatz an einem Anwendungsbeispiel. Bei der Vorgehensweise dient das KDD-Modell von *Hippner & Wilde*, welches in **Kapitel 6.1.4** vorgestellt, wird als Leitfaden. Es umfasst alle wichtigen Teilschritte und ist mit seiner hohen Flexibilität für die Anwendung im industriellen Umfeld gut geeignet. Das angeführte Beispiel orientiert sich am Forschungsvorhaben „Diagnose und Optimierung von Materialflusssteuerungen“ [Wustmann et al. 2010].

Das vorliegende Beispiel soll die Anwendbarkeit von KDD in der Logistik das Potential vorzeigen, das mit dem Einsatz von KDD realisierbar ist. Für dieses Vorhaben wird das KDD-Vorgehensmodell von *Hippner & Wilde* am Beispiel einer Gepäckförderanlage im Flughafen angewendet. Eine manuelle Analyse der Ereignisdaten ist aufgrund der immer komplexer und flexibler werdenden Anlagen nur für stark eingeschränkte Bereiche der Anlage möglich. Eine strukturierte umfassende Analyse ermöglicht das Data Mining und damit die Anwendung von KDD-Modellen.

Die Anwendung orientiert sich zunächst an der standartmäßigen Reihenfolge der Prozessschritte **(Kapitel 6.1.4)**:

Im ersten Schritt erfolgt die **Aufgabendefinition.** Grundlage der Betrachtung ist eine Gepäckförderanlage im Flughafen. Es handelt sich demnach um eine lange, spurgebundene Förderstrecke von ca. 40 km mit zahlreichen Abzweigungen und Zusammenführungen. Die Anforderung besteht darin Koffer, in vorgegebener Zeit von A nach B, zu transportieren. Herausforderungen ergeben sich durch erforderliche Pufferfunktionen für Frühgepäck oder unterschiedliche Lastphasen, das sind zum Beispiel Zeiträume mit viel Gepäck oder mit erhöhtem Aufkommen von Transfer-Gepäck.

Wiederholte Leistungsmängel eines Systems können für den Betreiber zu Beeinträchtigungen führen. Beispiele für Mängel einer Gepäckförderanlage sind:

- lange Durchlaufzeiten
- Warteschlagen
- Das Gepäckstück kommt am falschen Flughafen an
- Das Gepäckstück fällt vom Förderband
- Stau am Zoll

Ein grundlegendes wirtschaftliches Interesse liegt folglich darin, die Anlagen zu optimieren, die maximale Leistungsfähigkeit gezielt zu nutzen und lokale Reserven zu identifizieren. Standartmäßig benutzte Simulationsmodelle beantworten die Frage, ob ein System die Forderungen nach Durchsatz und Durchlaufzeit erfüllt, jedoch nicht wie gut es das tut oder welche Potenziale verborgen bleiben. Um diese Fragen zu klären, besteht zu Beginn die Aufgabe darin, Kriterien der Qualität der Anlage und Kenngrößen zu bestimmen. Bewertungskriterien für Materialflusssysteme lassen sich aus den übergeordneten Unternehmenszielen ableiten. Für die Materialflussplanung ergeben sich folgende Ziele:

- Min. Durchlaufzeit
    - Minimierung von Wartezeiten
    - kleinere Pufferbereiche
    - Reduzierung von Warteschlangen
- Max. Termintreue
- Max. Ausbringung
- Max. Auslastung
- Max. Flexibilität
- Min. Kosten

Die Gewichtung der Ziele hängt vom Transportgut und dem Unternehmen ab. Um die Funktionsgüte des Materialflusssystems zu messen, müssen quantifizierbare, messbare Merkmale identifiziert werden. Entsprechend der Gewichtung der Ziele lassen sich verschiedene Kenngrößen ableiten.

- Warteschlangenlänge
- Auslastung
- Blockadewahrscheinlichkeit
- Durchlaufzeit
- Kapazität

Ziel der Analyse der Gepäckförderanlage ist die Charakterisierung von Unregelmäßigkeiten in den Ereignisdaten zur Identifizierung von systeminternen Schwachstellen und Leistungsreserven. Eine Unregelmäßigkeit wird hier als numerische Abweichung in positive oder negative Richtung vom definierten Normal- oder Optimalwert verstanden.

Bereits ab dem nächsten Schritt (**Auswahl der Daten)** wird bereits deutlich, dass die die Phasen nicht trennscharf betrachtet werden können und wiederholte Rückschritte vorgenommen werden.

Um die hohe Komplexität des Systems auf eine Menge an relevanten Daten zu reduzieren und zu veranschaulichen wird zunächst ein Netzwerkmodell (**Kapitel 2.2**) entwickelt. Da der Transportweg zwischen einzelnen Punkten analysiert werden soll, wird ein Knoten-Kanten-Modell gewählt. Das topologische Modell besteht aus folgenden Objekten:

- **Statische Objekte**
    - *Knoten* = Ereignisort

      Beispiel: 114 Check-in-Schalter und 17.000 einzelne Komponenten wie Scanner, Weichen, Gepäckausgabe und Sicherheitskontrollen
    - *Kanten* = Verbindungen zwischen Ereignisorten

      Beispiel: Übergänge zwischen den Knoten, ca. 40 km Förderstrecke
    - *Pfade* = Verbindungen zwischen nicht benachbarten Ereignispositionen

- **Bewegliche Objekte**
  - *Fördergut* = Transporteinheit

    Beispiel: 15.000 Gepäckstücke pro Stunde

Um Kernpunkte des Systems zu bilden, bietet sich im nächsten Schritt (**Auswahl der Data-Mining-Verfahren** und **Anwendung der Data-Mining- Verfahren**) eine Clusteranalyse an. Dadurch wird die Anzahl der Knoten vermindert und es entsteht ein auf die wichtigen Knoten reduzierter Graph. Für die gebildeten Cluster können anschließend relevante Kenngrößen gemessen, berechnet, verglichen und erneut interpretiert werden. Dieser iterative Vorgang dient einer sich anpassenden Analyse, um möglichst sinnvolle Bereiche zu identifizieren und ausschlaggebende Abweichungen zu erfassen. Betont werden soll an dieser Stelle auch, dass durch die Clusterbildung zwar ein vereinfachtes Modell entsteht, jedoch keine Ergebnisse verworfen werden. Einzelergebnisse werden vielmehr zusammengefasst und können bei Bedarf wieder vereinzelt werden. **Abbildung 7-1** zeigt die prinzipielle Zusammenfassung von Knoten und Kanten.

Abbildung 7-1: Zusammenfassen von Knoten und Kanten

Das Clustern gibt für den weiteren Verlauf darüber Auskunft, an welchen Orten im System Ereignisse zu erfassen sind, um effizient Schwachstellen zu lokalisieren. Für die Datenbasis **(Auswahl der Daten)** werden nun Ereignisse vor und nach Verzweigungen, Zusammenführungen und Kreuzungen erfasst. Daneben fließen auch Daten aus den Quellen und Senken mit in die Analyse ein. Für Kanten, die keine Elemente wie beispielsweise Kreuzungen aufweisen ist eine Datenerhebung nicht notwendig. Ergeben Zwischenergebnisse jedoch, dass diese Kanten fehlerhaft funktionieren, kann ein nachträgliches Einbeziehen von Messdaten dieser Kanten notwendig sein.

Die **Auswahl der Daten** bezieht sich im Beispiel des Gepäckförderers auf die statischen Informationen wie die Förderstrecken des Netzwerkmodells, einzelne Pfade sowie die Speicherung folgender Zustandsänderungen (Ereignisse) im System:

- **Ort des Ereignisses**: Der Punkt im System, an dem das Ereignis gemessen wird (Gepäckannahme und diverse Lichtschranken)
- **ID des Fördergutes**: Die Kennzeichnung des Fördergutes (Zielflughafen des Gepäckstückes, Kennung des Besitzers)
- **Zeitstempel**: Der Zeitpunkt des Ereignisses

Bei der Bewertung der Datenqualität und in der Phase der **Datenaufbereitung** werden die Ereignisdaten aufbereitet. Dieser Schritt ist der aufwendigste Teil des Analyseprojekts. Fehlerhafte Daten werden bearbeitet, Ausreißer identifiziert, beurteilt und behandelt, neue Faktoren aus gemessenen Daten abgeleitet wie beispielsweise Kantenlasten und Zeiten, die

Gepäckstücke benötigen, um von A nach B zu gelangen oder Zwischenankunftszeiten. Die Software RapidMinier kann hierzu hilfreich eingesetzt werde und geht mit verschiedenen Techniken automatisch mit z.B. fehlenden Werten um. RapidMiner ist eine Software für Data-Mining-Anwendungen und kann auf allen gängigen Betriebssystemen verwendet werden. Außerdem deckt RapidMiner sowohl Forschungs- als auch industrielle und wirtschaftliche Anwendungsbereiche ab.

Die **Auswahl der Data-Mining-Verfahren** erfolgt anhand der Zielsetzung der Identifikation und Charakterisierung von Schwachstellen des Systems. Zur Aufdeckung von strukturellen Zusammenhängen eignet sich eine Assoziationsanalyse. Hierbei werden kritische Muster gesucht, die Mängel verursachen. Außerdem sind Prognosen denkbar, die die Knoten hinsichtlich verschiedener Lastphasen analysieren. Auf die eigentliche Data-Mining-Phase **(Anwendung der Data-Mining- Verfahren)** soll auch im dargestellten Beispiel der Gepäckförderanlage nicht detaillierter eingegangen werde. Eine Darstellung würde den Rahmen dieser Arbeit überschreiten und steht zudem nicht im Fokus vorliegender Untersuchung.

Bezüglich der **Evaluation** gibt es in RapidMiner verschiedenen Möglichkeiten um die Güte eines Modells zu bestimmen. Der Data-Mining-Prozess wird geprüft, indem aus dem vorhandenen Datensatz Trainings- und Testdaten generiert werden. Die Testdaten dienen dazu, das Modell auf seine Genauigkeit zu prüfen. Weiter findet in der Phase **Interpretation/ Evaluation** eine Charakterisierung der Ergebnisse statt. Kriterien dafür sind:

- Differenzierung nach Ausprägung: größere Abweichungen vom Optimalwert weisen auf kritischere Ergebnisse
- Unterscheidung nach Ort der Messung: Haupt- und Nebenpfade haben unterschiedliche Prioritäten
- Häufigkeit einer Ausprägung: ist es die erste Auffälligkeit oder ereignet sie sich vermehrt
- Unterscheidung zwischen bewegtem und stationärem Objekt: ergeben sich Schwachstellen bezüglich bestimmten Objekten, oder an bestimmten Knoten und Kanten
- Schwachstellengruppierung: Ergebnisse von bewegten Objekten, die sich im gleichen Stau befinden

An dieser Auflistung wird deutlich, dass die Bewertung objektbasiert ist. Sie bezieht sich auf das Fördergut oder auf einen Ort. Zudem bedarf es für die Bewertung an ausreichend Hintergrundwissen bezüglich des untersuchten Systems. Die Interpretation ist folglich in Zusammenarbeit von Analyseexperten und Fachpersonal durchzuführen um ausreichend Domänenverständnis zu gewährleisten. Schwachstellen die im Beispiel der Gepäckförderanlage entdeckt werden, beschreiben:

- Über- oder Unterlast
- Stau, vorübergehende Blockierungen
- Nicht vorgesehene Transportvorgänge wie Schleifenfahrten
- Leistungsreserven

Weiter sollen Aussagen darüber getroffen werden, ob nicht optimales Verhalten durch spontane Blockaden entsteht, die sich auch schnell wieder lösen, oder ob es ein Aufschaukeln ist, welches eine weitergehende Analyse bedarf um Aussagen über Ursachen treffen zu können. Weitere Anomalie-Erscheinungen können in Kreuzungsbereichen durch zwei aus verschiedenen

Richtungen eintreffenden Fördergütern auftreten. Positionsbezogene Unregelmäßigkeiten zeigen zeitbereichsbasierte Auswertungen von Kanten. Diese sollen unterschiedliche Inanspruchnahme der Kanten zu verschiedenen Zeiten zeigen. Der Schritt der **Interpretation** ist demzufolge ein weiterer iterativer Prozess. Ziel dieses Analyseschrittes ist die Identifizierung von Primär-Ursachen. Demnach werden mehrere Anomalien zur Bündelung der Schwachstellen zu Ursachen zugeordnet.

Das neu entstandene Wissen wird in der Phase **Anwendung der Ergebnisse** in den Prozess integriert. Dazu werden die Ergebnisse in Form von Handlungsanweisungen in den Prozess und an die Verantwortlichen zurückgegeben. Die vorangegangene Ursachen-Wirkungs-Diagnose zeigt betroffene Ursprünge der Schwachstellen auf. Zum Beispiel kann zielgenau bestimmt werden, welcher Pfad warum zu Verzögerungen führt. Mögliche Anpassungen können Vorfahrtsregelungen an Kreuzungen sein oder die Umleitung einer definierten Menge Fördergut über andere Pfade zu bestimmten Lastphasen. Falls die Beseitigung einer Ursache aus technischen Gründen nicht umsetzbar ist, fließen die Erkenntnisse dennoch als Grundlage für Planungssicherheit in den Prozess ein.

Anhand dieses Anwendungsbeispiels zeigt sich die Komplexität eines KDD-Projektes. Die Anwendung des Vorgehensmodells von Hippner & Wilde eignet sich aufgrund der Flexibilität sehr gut. Die Übergänge der Phasen sind fließend und variabel gestaltbar. Die Wissensentdeckung gestaltet sich als iterativer Prozess, bei dem schrittweise Parameter angepasst werden müssen. Die größte Herausforderung besteht im Vorhandensein benötigter Daten, in der Datenqualität und dem entsprechenden Fachwissen von Fachexperten und Analysten.

8. Zusammenfassung und Fazit

In der vorliegenden Arbeit wird ein Überblick über die KDD-Vorgehensmodelle mit dem Schwerpunkt der Anwendung in Bereichen der Logistik gegeben.

Zunächst erfolgt die Darstellung relevanter Grundlagen der Logistik und der Wissensentdeckung durch KDD und Data Mining. Nach einer Auflistung von KDD-Vorgehensmodellen wird eine Auswahl geeigneter Modelle getroffen, die auf einen praktischen Einsatz geprüft werden. Anwendungsbezogen werden dazu relevante Kriterien abgeleitet. Zur Visualisierung wird eine tabellarische Übersicht erarbeitet. Basierend auf einer detaillierteren Darstellung der ausgewählten Modelle wird die Anwendungsmöglichkeit dieser Modelle auf den Einsatz im logistischen Umfeld untersucht. Danach anschließend erfolgen ein Vergleich der Ergebnisse vorangegangener Untersuchungen und eine tabellarische Gegenüberstellung. Abschließend wird die Bewertung und eine prototypische Anwendung anhand eines geeigneten Vorgehensmodells vorgestellt.

Die eigentliche Datenanalyse ist selten problematisch, Schwierigkeiten ergeben sich aufgrund der polystrukturierten Daten. Daten der Logistik erhalten durch weitreichende Vernetzung hohe Komplexität. Von bedeutender Wichtigkeit ist die Auswahl qualitativer Daten und geeigneter Prognosemodelle. Vorteil der Anwendung von Vorgehensmodellen ist eine hohe Erfolgschance durch die bereits erprobte Herangehensweise. Ein flexibles Modell zur individuellen Anpassung je nach Projekt sollte dennoch gegeben sein. Die erfolgreiche Anwendung erfordert eine optimale Zusammenarbeit der verschiedenen Experten sowie die Integration menschlichen Fachwissens mit Wissen aus Prozessdaten. Mit der Entwicklung der Informationstechnik sowie der zunehmenden Möglichkeiten, heterogene Daten aus verschiedenen Quellen automatisch zu vereinen und zu verwalten wird Data Mining und KDD weiter an Bedeutung gewinnen.

Basierend auf den Ergebnissen vorliegender Arbeit wird deutlich, dass KDD-Vorgehensmodelle auch in der Logistik gewinnbringend eingesetzt werden können. Dabei ist zu bedenken, dass die Möglichkeit zur maschinellen Auswertung großer Datensätze keine Garantie für eine gesteigerte Wettbewerbsfähigkeit darstellt. In der praktischen Umsetzung von KDD-Projekten ist die Abstimmung der Daten mit geeigneten Modellen und Expertenwissen von herausragender Wichtigkeit. Daher ergibt sich in vorliegender Arbeit eine Abstufung für die Eignung der verschiedenen Modelle. Eine Vielzahl der in der Literatur vorgestellten Vorgehensmodelle vernachlässigt den Einsatz menschlicher Anwender und die Wichtigkeit fachlicher Kompetenz. Weiter wird deutlich, dass die Einbettung in ein betriebliches Umfeld und die Analyse industrieller Daten weitere Schwierigkeiten darstellen. Eine entsprechende Umsetzung in den hier vorgestellten Vorgehensmodellen ist nach vorliegender Untersuchung nur bedingt erfüllt. Besonders in der Untersuchung der vier ausgewählten Modelle werden Unterschiede dahingehend verdeutlicht. Das Modell *SEMMA* gehört zu den bekanntesten KDD-Vorgehensmodellen und wird bereits häufig eingesetzt, jedoch zeigen sich deutliche Defizite für den Einsatz im logistischen Umfeld. Dazu ist es an die von SAS entwickelte Software gebunden und bietet damit wenig Spielraum für flexible Anpassungen. Das Modell von *Fayyad et al.* ist für den Einsatz in Bereichen der Logistik ebenfalls nicht optimal geeignet. Bei der Untersuchung zeigt sich, dass dieses Modell aufgrund seiner Eigenschaften vorwiegend für den Einsatz in der Wissenschaft (z.B.: DNA-Sequenzierung) vorgesehen ist. Das Modell von *Hippner & Wilde* und der *CRISP-DM* eigenen sich mit Einschränkungen am besten für den logistischen Einsatz. Der

CRISP-DM überzeugt durch einen hohen Detaillierungsgrad. Die einzelnen Phasen sind in weitreichende Unteraufgaben gegliedert, die dem Anwender einen genauen Leitfaden vorgeben. Außerdem wird der Überprüfung des Modells und den äußeren Bedingungen des Projekts eine große Bedeutung zugeteilt. Der Prozess von *Hippner & Wilde* betont höchste Flexibilität. Die Reihenfolge der Schritte unterliegt keiner Vorgabe und Rücksprünge sind jederzeit möglich. Trotzdem gibt es genügend Vorgaben zur Orientierung.

Das abschließende Fazit dieser Arbeit lautet folglich, dass großes Potenzial für den Einsatz von KDD im logistischen Umfeld vorhanden ist. Unter den hier vorgestellten KDD-Vorgehensmodellen eignen sich besonders der *CRISP-DM* und das Modell von *Hippner & Wilde*. Jedoch ergeben sich für die Analyse von industriellen Daten charakteristische Herausforderungen. Außerdem kommt besondere Wichtigkeit der Integration von Expertenwissen zu. Für den optimalen Einsatz von KDD im logistischen Umfeld bedarf es daher einer Anpassung der bisher bekannten Vorgehensmodelle.

Literaturverzeichnis

Adriaans, P. & Zantinge, D. (1996): Data Mining. Amsterdam: Addison-Wesley Longman.

Alby, T.; Braun, D.; Pfleger, S. (2009): Projektmanagement: Definitionen, Einführungen und Vorlagen. http://projektmanagement-definitionen.de/glossar/methode/. 06.08.2015.

Allweyer, T. (2005): Geschäftsprozessmanagement. Strategie, Entwurf, Implementierung, Controlling. Herdecke, Witten: W3L.

Angermeier, G. (2008): Lessons Learned. https://www.projektmagazin.de/glossarterm/lessons-learned. 01.07.2015.

Arndt, D. (2008): Customer Information: Ein Referenzmodell für die Informationsversorgung im Customer Relationship Management. Göttingen: Cuvillier.

Baumgarten, H.; Darkow, I.; Zadek, H. (2004): Supply Chain Steuerung und Services. Berlin, Heidelberg: Springer.

Berry, M. J. & Linoff, G. S. (2011): Data Mining Techniques. For Marketing, Sales, and Customer Relationship Management. (3. Aufl.), Indianapolis: Wiley.

Cleve, J. & Lämmel, U. (2014): Data Mining. Wismar: De Gruyter Oldenbourg.

Deuse, J.; Erohin, O.; Lieber, D. (2014): Wissensentdeckung in vernetzten, industriellen Datenbeständen. In: (H. Lödding (Hrsg.)) Industrie 4.0. Wie intelligente Vernetzung und kognitive Systeme unsere Arbeit verändern., Berlin: Gito, S. 373-395.

Doberstein, S. (2011): Was ist Wissensmanagement?. http://www.community-of-knowledge.de/wissensmanagement/. 09.06.2015

Ebersbach, A.; Glaser, M.; Heigl, R.; Warta, A. (2008): Wiki. Kooperation im Web. (2. Aufl.), Berlin et al.: Springer.

Farkisch, K. (2011): Data-Warehouse-Systeme kompakt. Aufbau, Architektur, Grundfunktionen. Berlin, Heidelberg: Springer.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (1996): Advances in Knowledge Discovery and Data Mining, Cambridge et al.: MIT Press

Figura, M. & Gross, D. (2013): Die Qual der Wiki-Wahl. Wikis für Wissensmanagement in Organisationen. https://www.pumacy.de/publikationen/studien/wikis-fuer-wissensmanagement/. 02.07.2015.

Fischer, J. (2015): Internet der Dinge. http://www.internet-der-dinge.de/. 18.09.2015.

Gudehus, T. (2012): Logistik 1. Grundlagen, Verfahren und Strategien. (4. Aufl.), Berlin, Heidelberg: Springer.

Hausladen, I. (2014): IT-gestützte Logistik. Systeme - Prozesse- Anwendungen. (2. Aufl.), Wiesbaden: Springer Gabler.

Hippner, H.; Grieser, L.; Wilde, K. D. (2011): Data Mining - Grundlagen und Einsatzpotenziale in analytischen CRM-Prozessen. In: B. Hubrich, K.D. Wilde, H. Hippner (Hrsg.) Grundlagen des CRM: Strategie, Geschäftsprozesse und IT-Unterstützung, Wiesbaden: Gabler.

Hippner, H. & Wilde, K.D. (2001): Der Prozess des Data Mining im Marketing, in: (Hippner, H.; Küsters, U.; Meyer, M.; Wilde, K.D. (Hrsg.)): Handbuch Data Mining im Marketing, Wiesbaden, S. 21-91.

Holland, H. (2015): Gabler Wirtschaftslexikon, Stichwort: Customer Relationship Management. http://wirtschaftslexikon.gabler.de/Archiv/5072/customer-relationship-management-crm-v10.html. 29.09.1015.

Hoppe, T. (2013): Wissensmanagement in Theorie und Praxis. http://www.community-of-knowledge.de/beitrag/semantische-filterung-ein-werkzeug-zur-steigerung-der-effizienz-im-wissensmanagement/. 23.06.2015.

Jarke, M. (2014): Interview mit Stefan Wrobel zum Thema "Angewandte Big-Data-Forschung". In: Wirtschaftsinformatik. 56 (2014) 5, S. 333-334.

Kahlen, C. (2015): Industrie 4.0. http://www.plattform-i40.de/I40/Navigation/DE/Industrie40/WasIndustrie40/was-ist-industrie-40.html;jsessionid=B07FFE85FA85864F56AA5A1B6497EBA1. 18.09.2015.

Kamber, M.; Han, J.; Pei, J. (2012): Data mining. concepts and techniques. (3. Aufl.), Amsterdam et al.: Elsevier.

Knobloch, B. (2000): Der Data-Mining-Ansatz zur Analyse betriebswirtschaftlicher Daten. In: Bamberger Beiträge zur Wirtschaftsinformatik, (2000) 58.

Knobloch, B. & Weidner, J. (2000): Eine kritische Betrachtung von Data-Mining-Prozessen. Ablauf, Effizienz und Unterstützungspotenziale. In: Data Warehousing 2000. Methoden, Anwendungen, Strategien, (2000), S.345-365.

Kratzer, J.& van Veen, K. (2005). Über die Bedeutung der Analyse sozialer Netzwerke für das moderne Wissensmanagement. Von Community of knowledge: http://www.community-of-knowledge.de/beitrag/ueber-die-bedeutung-der-analyse-sozialer-netzwerke-fuer-das-moderne-wissensmanagement/. 09.06.2015.

Krieger, W. (2015): Gabler Wirtschaftslexikon, Stichwort: RFID. http://wirtschaftslexikon.gabler.de/Archiv/83828/rfid-v7.html. 28.09.2015

Kurgan, L. A. & Musilek, P. (2006): A survey of Knowledge Discovery and Data Mining process models. In: The Knowledge Engineering Review, (2006) 1, S. 1-24

Lehner, F. (2014): Wissensmanagement. Grundlagen, Methoden und technische Unterstützung. (5. Aufl.), München: Carl Hanser.

North, K. (2011): Wissensorientierte Unternehmensführung. Wertschöpfung durch Wissen. (5. Aufl.), Wiesbaden: Gabler.

Oeldorf, G. & Olfert, K. (2009): Kompakt-Training Materialwirtschaft. (3. Aufl.), Ludwigshafen: Kiehl.

Petersohn, H. (2005): Data Mining. Verfahren, Prozesse, Anwendungsarchitektur. München: Oldenbourg Wissenschaftsverlag GmbH.

Pfitzner, M. & Wieland, U. (2014): Interdisziplinäre Datenanalyse für Industrie 4.0. In: Controlling & Management Review. 58 (2014) 7, S. 80-85.

Piatetsky-Shapiro, P. (2014): Kdnuggets. http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html. 24.08.2015.

Prestifilippo, G. (2015): Wandelbare IT-Systeme. Logistiksoftware. In Logistik für Unternehmen, Jg. (2015) 9, S. 47-49.

Probst, G.; Raub, S.; Romhardt, K. (2012): Wissen managen. Wie Unternehmen ihre wertvollste Ressource optimal nutzen. (7. Aufl.), Wiesbaden: Springer Gabler.

Reiner, R. (2013): Anwendung von Data Mining zur Ableitung von Planungsregeln in der flexibilitätsorientierten Prozessindustrie. In: (H. Zsifkovits und S. Altendorfer-Kaier (Hrsg.)):

Logistische Modellierung: 2. Wissenschaftlicher Industrielogistik-Dialog in Leoben (WilD), München: Hampp, S. 110-119.

Runkler, T. A. (2010): Data Mining. Methoden und Algorithmen intelligenter Datenanalyse. (W. Bible, R. Kruse, B. Nebel (Hrsg.)), Wiesbaden: Vieweg + Teubner.

Säuberlich, F. (2000): KDD und Data Mining als Hilfmittel zur Entscheidungsunterstützung. Frankfurt am Main et al.: Lang.

Schinzer, H.; Bange, C.; Mertens, H. (1999): Data Warehouse und Data Mining. Marktführende Produkte im Vergleich. (R. Thome (Hrsg.)), 2. Aufl.), München: Vahlen.

Schmid, H. (2013): Barrieren im Wissenstransfer. Ursachen und deren Überwindung. (H. Krcmar (Hrsg.)),Wiesbaden: Springer Gabler.

Schulte-Zurhausen, M. (2002): Organisation. (3. Aufl.), München: Vahlen

Schwarzer, B. & Krcmar, H. (2004): Wirtschaftsinformatik. Grundzüge der betrieblichen Datenverarbeitung. (B. Pietschmann, D. Vahs (Hrsg.)),(3. Aufl.), Stuttgart.:Schäffer-Poeschel.

Seeck, S. (2010): Erfolgsfaktor Logistik. Klassische Fehler erkennen und vermeiden. Wiesbaden: Gabler.

Segler, T. (1985): Evolution von Organisationen. Ein evolutionstheoretischer Ansatz zur Erklärung der Entstehung und des Wandels von Organisationsformen. Frankfurt am Main: Lang.

Sharafi, A. (2012): Knowledge Discovery in Databases. Eine Analyse des Änderungsmanagements in der Produktentwicklung. Dissertation, Technische Universität München, Wiesbaden: Springer Gabler.

Steinlein, U. (2004): Data Mining als Instrument der Responseoptimierung im Direktmarketing: Methoden zur Bewältigung niedriger Responsequoten. Göttingen: Cuvillier.

Stiller, G. (2015): wirtschaftslexikon24.com. http://www.wirtschaftslexikon24.com/d/logistik/logistik.htm. 23.08.2015.

Talia, D. & Trunfio, P. (2012): Service-Oriented Distributed Knowledge Discovery. London: CRC Press.

Vahrenkamp, R. (2015): Gabler Wirtschaftslexikon, Sichwort: Enterprise-Resource-Planning-System. http://wirtschaftslexikon.gabler.de/Archiv/17984/enterprise-resource-planning-system-v12.html. 29.09.2015

Vahrenkamp, R. & Mattfeld, D. (2007): Logistiknetzwerke. Modelle für Standortwahl und Tourenplanung. Wiesbaden: Gabler.

Werner, M. (2004): Einflussfaktoren des Wissenstransfers in wissensintensiven Dienstleistungsunternehmen. Eine explorativ-empirische Untersuchung bei Unternehmensberatungen. Universität Duisburg-Essen, Wiesbaden: Deutscher Universitäts Verlag.

Weskamp, M.; Tama, A.; Schatz, A. (2014): Einsatz und Nutzenpotenziale von Data Mining in Produktionsunternehmen. Fraunhofer IPA

Wrobel, S. (1998): Data Mining und Wissensentdeckung in Datenbanken. In: Künstliche Intelligenz, Heft 1/1998.

Wrobel, S.; Joachims, T.; Morik, K. (2013): Maschinelles Lernen und Data Mining. In: (G. Günther, J. Schneeberger, U. Schmid (Hrsg.)): Handbuch der Künstlichen Intelligenz, München: De Gruyter, S. 405-472.

Wustmann, D.; Vasyutynskyy, V.; Schmidt, T.; Kabitzsch, K. (2010): Diagnose und Optimierung von Materialflusssteuerung. Schlussberiecht des Forschungsvorhaben AiF-Nr. 15770 BR. Technische Universität Dresden, Professur für Technische Logistik, Professur für Technische Informationssysteme. https://www.bvl.de/files/441/481/522/578/15770BR_MaterialflussDiagnose_TUD.pdf. 13.10.2015

Yakut, Y. (2015): Erzielen von Wettbewerbsvorteilen durch Data Mining in Produktion und Logistik. Dissertation, Hamburg: disserta Verlag.

Abbildungsverzeichnis

Tabellenverzeichnis

Eidesstattliche Versicherung

Beckmann, Nadine 164193

Name, Vorname Matr.-Nr.

Ich versichere hiermit an Eides statt, dass ich die vorliegende ~~Bachelorarbeit~~/Masterarbeit* mit dem Titel

Untersuchung des Einsatzes von Vorgehensmodellen des Knowledge Discovery in Databases für Bereiche der Logistik

selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Dortmund, den 10.11.2015 _____

Ort, Datum Unterschrift

*Nichtzutreffendes bitte streichen

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz - HG -)

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird gfls. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin“) zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Dortmund, den 10.11.2015 _____

Ort, Datum Unterschrift