Machine Learning: Technische Grundlagen und einschlägige Aufgaben eines Data Scientists

Studienarbeit T3100

Abgabedatum: 07.05.2018

an der Dualen Hochschule Baden-Württemberg
Standort Stuttgart

Name:	Dennis Netzer
Matrikelnummer:	8242815
Kurs:	TWIW15PL
Studiengang:	Wirtschaftsingenieurwesen
Studienjahrgang:	2015
Betreuer:	Bert Miecznik

Eigenständigkeitserklärung

Name, Vorname:	Netzer, Dennis
Matrikelnummer:	8242815
Studiengang/Kurs:	TWIW15PL

Ich versichere hiermit, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Stuttgart, 07.05.2018

Ort, Datum	Unterschrift

Zusammenfassung

Wir leben in dem Zeitalter der Daten. Jeder produziert und konsumiert täglich Unmengen an Daten. Durch diese steigende Datenmenge und die steigende Rechenleistung der Computer, gewann das Feld des Machine Learnings zunehmend an Beliebtheit und Anwendbarkeit. Machine Learning hilft hierbei die Daten zu verstehen, Fragen mit Hilfe der Daten zu beantworten, Trends abzuleiten oder aufgrund der Daten Entscheidungen zu treffen.

Im zweiten Kapitel werden Lernverfahren des Machine Learning erklärt. Dabei werden zwei grundsätzliche Verfahren unterschieden, Supervised Learning und Unsupervised Learning. Das Supervised Learning bietet dabei dem Computern die Möglichkeit, anhand von richtigen Beispielen Zusammenhänge und Abhängigkeiten von Daten zu lernen. Beim Unsupervised Learning lernt der Computer selbstständig Muster und Zusammenhänge in Daten zu erkennen. Bei diesem Lernverfahren kommen unterschiedliche Algorithmen zum Einsatz, die auf mathematischen und statistischen Modellen beruhen. Diese Algorithmen können „Lernen“, indem sie ihre Parameter, aufgrund richtiger oder falscher Entscheidungen selbstständig anpassen. Eine Gruppe von diesen Algorithmen werden als künstliche neuronale Netze bezeichnet. Diese künstlichen neuronalen Netze bestehen aus mehren Ebenen, wobei sich in jeder Ebene mehrere Neuronen befinden. Nun werden die Daten über mehrere Ebenen hinweg analysiert. Dabei führt jedes Neuron eine Berechnung durch und sendet das Ergebnis der Berechnung zu einem oder mehreren Neuronen der nächsten Ebene. So werden die Daten über mehrere Ebenen hinweg abstrahiert, um schlussendlich zu einem Ergebnis zu kommen. Jede Verbindung zwischen zwei Neuronen ist über einen Zahlenwert gewichtet, welchen das künstliche neuronale Netz eigenständig anpasst, sodass beispielsweise eine möglichst genaue Vorhersagewahrscheinlichkeit erreicht wird.

Im dritten Kapitel geht es um den Data Scientist, als neu entstandene Berufsbezeichnung. Ein Data Scientist ist eine Person, welche versucht aus Daten monetäre Werte für Unternehmen zu schaffen. Dabei werden eine Vielzahl an Kompetenzen benötigt. Durch diese hohen Anforderungen an die Kompetenzen eines Data Scientists, herrscht ein starker Fachkräftemangel. Dieser wird, durch Automatisierung einiger Aufgaben des Data Scientists, versucht zu bekämpfen. Viele Firmen bieten dazu schon Tools zur automatisierten Datenverarbeitung, Datenaufbereitung und Modellbildung an. Allerdings ist die größte Herausforderung, die richtigen Fragen zu vorhandenen Daten zu stellen und aus Daten gewonnenes Wissen in monetäre Werte umzuwandeln. Daher werden weiterhin Data Scientists in Unternehmen benötigt. Die Tools unterstützen dabei, Ergebnisse schneller zu erzielen. Außerdem bieten diese Tools einen leichteren Einsteig in die Data Science.

Inhaltsverzeichnis

Abkürzungen

Abkürzung	Bedeutung
KNN	Künstliche Neuronale Netze
DNN	Deep Neural Network
CRISP-DM Modell	Cross Industry Standard Process for Data Mining Modell
Central Processing Unit	CPU
Graphics Processing Unit	GPU
Gerneral Purpose Graphics Processing Unit	GP-GPU

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Was bedeutet Machine Learning

Machine Learning ist ein Teilgebiet der künstlichen Intelligenz, bei dem Algorithmen systematisch angewandt werden, um Wissen aus Beziehungen der Daten und der Informationen zu gewinnen. Diese Algorithmen ermöglichen dem Computer zu lernen. Eine Definition dieses Lernens wurde von Tom. M. Mitchell aufgestellt: *„A computer program is said to learn from experience E with respect to some class of tasks T and performance P, if its performance at tasks in T, as measured by P, improves with experience E.“* [36] Eine Aufgabe besteht aus einem realen Problem, welche aus einfachen Ja/Nein Entscheidungen, wie das Kategorisieren einer E-Mail nach Spam oder kein Spam, oder komplexeren Entscheidungen, wie die Bestimmung des genauen Verkaufswertes einer Immobilie. Die Performance kann in diesem Zusammenhang die Anzahl richtiger Vorhersagen sein. Die Erfahrung besteht aus Daten, welche genutzt werden, um daraus die Parameter des Algorithmus zu adjustieren. [41] Werden dem Algorithmus zum Beispiel neue Datensätze zur Verfügung gestellt, die er kategorisieren soll und die Fehlerquote dieser Kategorisierung ist bei dem aktuellen Algorithmus hoch, kann der Algorithmus seine Parameter selbstständig ändern, sodass die Fehlerquote für die bereits bekannten Daten minimiert wird. Die angewandten Algorithmen können so Störungen des existierenden Modells erkennen und sich selber umstrukturieren und anpassen. [3] Um das Lernproblem richtig zu verstehen, müssen diese drei Eigenschaften, Aufgabe, Performance und Erfahrung, genau identifiziert werden. Mit steigender Komplexität des Problems, steigt auch die Komplexität der zu nutzenden Algorithmen zur Erstellung eines geeigneten Modells. Ein Modell ist hier die Anwendung des richtigen Algorithmus auf die Daten, durch welche das Problem gelöst werden soll und das daraus resultierende Ergebnis. Eine grobe Einteilung der vorhandenen Algorithmen bietet Kapitel 2.

1.2 Wo wird Machine Learning eingesetzt

Durch das steigende Volumen, die steigende Vielfalt und die Geschwindigkeit des Datenstroms, sowie die zunehmende Komplexität der Daten, können Menschen die Daten nicht mehr analysieren und verstehen. Durch steigende Rechenleistung, wird es immer leichter Computern diese Aufgabe zu übergeben und damit große Datenmenge zu analysieren. [3] Die Maschinen nutzen die Daten um daraus die verwendeten Algorithmen anzupassen,

sodass diese auf neue Probleme angewandt werden können, die der Maschine nicht explizit 'beigebracht' wurden. Durch die steigende Menge an Daten aus denen die Maschine lernen kann, steigt auch der Lernerfolg, sodass die Algorithmen verlässlicher auf neue Probleme angewandt werden können. Durch diese parallele Entwicklung der steigenden Datenmenge und der steigenden Rechenleistung, wachsen die Anwendungsgebiete solcher maschinellen Lernalgorithmen ständig an. [31] Bekannte Anwendungsgebiete sind hierbei autonomes Fahren, Börsenhandel, Gesundheitswesen, Suchmaschinen oder Verkaufsempfehlungen. [33] Und nicht nur in der Sprachverarbeitung werden Machine Learning Algorithmen eingesetzt, sondern auch in der Übersetzung von Sprachen. [16] Doch auch Industrieunternehmen können immer mehr durch die Fortschritte in Machine Learning und der künstlichen Intelligenz profitieren. Beispielsweise helfen Machine Learning Algorithmen beim Einstellen neuer Arbeitskräfte, indem die Bewerbungen auf das Vorhandensein bestimmter Qualifikationen geprüft werden. Im Marketingbereich helfen Algorithmen neben dem Erstellen personalisierter Werbung auch bei dem Messen des Erfolges bestimmter Werbemaßnahmen. So können Algorithmen beispielsweise die Dauer und die Position an der ein Markenzeichen bei bestimmten Events oder Werbespots auftaucht messen. Ein weiterer wichtiger Anwendungsfall in der Industrie ist die sogenannte Predictive Maintenance. Dabei werden Daten, welche Einfluss auf den Verschleiß beispielsweise eines Werkzeugs haben, mit Hilfe verschiedene Sensoren gesammelt. Machine Learning Algorithmen können diese Daten auswerten und schon bevor es zum Ausfall einer Maschine oder eines Werkzeugs kommt die zuständigen Arbeitskräfte darauf hinweisen. So können Ausfallzeiten minimiert werden. [60] [59]

2 Lernverfahren des Machine Learning

2.1 Lerndaten

Damit eine Maschine lernen kann, werden Daten benötigt, aus denen sie lernt. Dazu müssen in einem ersten Schritt Daten mit entsprechendem Problembezug gesammelt werden. Dazu werden alle historischen Daten, die bereits im Unternehmen vorliegen aus den vorhandenen Datenquellen verwendet. Wenn zusätzlich zu den vorhandenen Daten weitere Daten benötigt werden, können diese über andere Quellen wie Internet, Befragungen, Simulationen oder Experimente bezogen werden. Es ist auch möglich eine lernende Maschine zu entwickeln, die aktiv weiteres Wissen erwirbt, indem sie automatisiert Datenquellen, wie beispielsweise das Internet durchsucht. [19] Diese Daten werden anschließend im Hinblick auf die Menge, die Attribute und deren Beziehungen analysiert und beschrieben. Ziel ist es die Daten zu verstehen. Das heißt, es muss verstanden werden, welchen Einfluss die Attribute auf das zu lösende Problem haben und welche Beziehungen unter den Attributen herrschen. Aufgrund der steigenden Datenmengen und der steigenden Komplexität der Daten ist das Verstehen der Einflüsse und der Beziehungen oft nicht möglich. Hier können Alorithmen verwendet werden, welche die vorhandenen Daten auf auf Beziehungen und Abhängigkeiten untersuchen und dessen Ziel es ist auch nur diese Beziehungen darzustellen. Des Weiteren sollen mögliche Fehler, die bei dem späteren Algorithmus auftreten können identifiziert werden. [48] Nach dieser ersten Analyse und Beschreibung der Daten, werden diese für die bevorstehende Analyse durch den maschinellen Lernalgorithmus aufbereitet. Dazu müssen unterschiedliche Datensätze zusammengefasst werden und in ein einheitliches Format gebracht werden, sodass die Daten von dem maschinellen Lernalgorithmus verarbeitet werden können. [3] Die Daten werden meist in Tabellenform gespeichert. Jede Reihe ist dabei ein Datensatz, der die Attribute und das Resultat einer Beobachtung enthält. Jede Spalte enthält dabei ein anderes Attribut auch Feature genannt. Diese Features beeinflussen die abhängige Variable, also das Resultat. Bei den angesprochenen Algorithmen, welche die Daten auf Muster untersuchen, entfällt der Schritt der Datenaufbereitung. In Tabelle 2.1 sind beispielsweise vier Attribute dargestellt, die einen Einfluss auf den Verkaufspreis eines Hauses haben.
Die Datenaufbereitung gehört zu den Aufgaben, die am meisten Zeit in Anspruch nehmen. Zu dieser Datenaufbereitung gehört zum einen der Prozess der Generierung von Attributen. Dabei werden aus den vorhandenen Attributen weitere nützliche Attribute abgeleitet, ohne Redundanzen entstehen zu lassen. Bei einem weiteren Prozess der Attributselektion,

Tabelle 2.1: Datentabelle Verkaufspreis

Attribut 1	**Attribut 2**	**Attribut 3**	**Attribut 4**	**Resultat**
Anzahl Zimmer	**Wohnfläche in (m^2)**	**Anzahl Autostellplätze**	**Gartenfläche (m^2)**	**Hauspreis (€)**
4	80	1	20	152.000
5	130	3	80	320.000
2	50	0	0	110.000
3	120	2	35	240.000
2	40	2	10	145.000
1	25	0	0	65.000

wird versucht genau diese Redundanzen zu beseitigen. Redundanzen sind dabei verschiedene Attribute, aus denen sich exakt die selben Informationen ableiten lassen. Dies erhöht die zu verarbeitende Datenmenge, ohne einen zusätzlichen Nutzen zu bieten. Außerdem wird versucht Attribute zu identifizieren, die in keinem Zusammenhang mit dem Resultat stehen. Die Bestimmung des Resultats für die gegebenen Attribute wird oft von einem Experten vorgenommen. [48] In unserem Fall bedeutet das, ein Experte bestimmt den Verkaufspreis, anhand der zur Verfügung stehenden Daten.
Der gesamte Prozess der Datengewinnung und -aufbereitung gestaltet sich in der Praxis oft schwierig. Auch für Experten sind die Beziehungen zwischen den Attributen und dem Resultat oft nicht ersichtlich und damit nicht einschätzbar. Attribute bei denen der Experte keinen Zusammenhang zum Problem vermutet können tatsächlich einen Einfluss haben. Auch die Auswahl der Klassen bei Klassifizierungsproblemen durch den Experten kann zu Problemen führen, wenn es Datensätze gibt, die sich in Grauzonen befinden und sich daher keiner Klasse eindeutig zuordnen lassen. [30] Deswegen ist es schwierig bei den riesigen zu Verfügung stehenden Datenmengen, die eigentlich relevante Teilmenge zu selektieren und diese bei Klassifizierungsproblemen der richtigen Klassen zuzuweisen. Kommt es vor, dass bei den vorhandenen Daten ein wichtiges Attribut fehlt, von welchem das endgültige Resultat abhängt, kann dieses fehlende Attribut oft nicht im Nachhinein bestimmt werden. Ein weiteres Problem stellen die sogenannten unstrukturierten Daten dar. Diese kommen in komplexen Formen, wie beispielsweise Graphen oder Videos vor. Um diese für den maschinellen Lernalgorithmus nutzen zu können, sind spezielle Techniken erforderlich, die zurzeit nur bedingt brauchbare Ergebnisse liefern. [3] Ein weiterer wichtiger Aspekt ist die Datenqualität. Bei den riesigen Datenmengen ist es oft nicht ersichtlich, wie die Daten ermittelt wurden und ob es sich um verlässliche Quellen handelt. Weiter ist es wichtig, dass die Attribute mit denen der Algorithmus trainiert wird weitestgehend mit den Attributen bei der eigentlich Verwendung übereinstimmen. Wird der Algorithmus mit mangelhaften oder einseitigen Daten trainiert, entstehen bei der späteren

Anwendung viele Fehleinschätzungen. [36] Da Daten die Grundlage des ganzen Modells sind und das Resultat maßgeblich beeinflussen, ist es wichtig die Aufbereitung der Daten sorgfältig durchzuführen.

2.2 Supervised Learning

Beim Supervised Learning wird die Beziehung zwischen Inputdaten und abhängigen Outputvariablen betrachtet, bei denen die vorhandenen Outputvariablen schon vor Implementierung des Algortithmus feststehen. Dabei hat jeder dieser Inputdaten einen korrekt zugehörigen Output. Dieser Output kann entweder kategorisch oder kontinuierlich sein. Das Modell entsteht also unter „Supervision“. [45] Das Hauptziel ist es einen Algorithmus zu entwickeln der neue Inputdaten dem richtigen Output zuordnet und somit die Anzahl an Fehlvorhersagen minimiert. Dieser Algorithmus kann anschließend dazu verwendet werden, eine Vorhersage über den Output neuer Inputdaten zu treffen, bei denen die Beziehung nicht bekannt ist. [48] Der Algorithmus wird mit Hilfe von Trainingsdaten trainiert, welche schon einem korrekten Output zugeordnet sind. Diese korrekte Zuordnung wird in den meisten Fällen von menschlichen Experten vorgenommen. Demnach lernt die Maschine Muster aus bekannten Beispielen zu erkennen und dieses Muster auf unbekannte Daten zu übertragen. Es wird also angestrebt ein allgemeingültiges Modell zu entwickeln. Je mehr unterschiedliche Trainingsdaten der Maschine zur Verfügung stehen, desto besser wird der Algorithmus generalisiert und desto höher ist die Erfolgsquote richtiger Zuordnungen.
Supervised Learning kommt demnach immer zum Einsatz, wenn ein genauer Zielwert gefunden werden soll und dieser Zielwert für Trainingsdaten von Experten bestimmt werden kann. Die Lernverfahren des überwachten Lernens unterscheiden sich anhand dieses Zielwertes. Ist dieser diskret handelt es sich um eine Klassifizierung, bei einem analogen Zielwert um eine Regression. [3] Die Verfahren des Supervised Learning finden in der Praxis bislang die meiste Anwendung.

2.2.1 Klassifikation

Bei der Klassifikation wird ein Modell erstellt, das die Eingangsdaten bestimmten diskreten Kategorien zuweist. Dabei errechnet eine Funktion eine Wahrscheinlichkeit, mit welcher die Daten zu einer bestimmten Kategorie zugeordnet werden können und weist die Kategorie mit der höchsten Wahrscheinlichkeit zu. [48] Diese Kategorie wird auch Label oder Klasse genannt. Da die Trainingsdaten bereits den richtigen Kategorien zugeordnet sind, wird von gelabelten Trainingsdaten gesprochen. Die Datensätze bestehen aus beliebig vielen Attributen, die einen Einfluss auf die zugehörige Kategorie haben. Die

einfachste Art der Klassifikation ist die binäre Klassifikation, auch Konzept-Lernen genannt, bei der es nur zwei zu unterscheidende Kategorien gibt. Beispielsweise wird diese Art von Klassifikation bei Banken verwendet, um zu entscheiden, ob ein Kredit gewährt werden soll oder nicht. Bei diesem Klassifikationsproblem können Attribute zum Beispiel das Alter, das Einkommen, Höhe des Kredites und Anzahl und Höhe vergangener Zahlungsausfälle sein. Gibt es mehr als zwei zu unterscheidende Klassen wird von mulit class Klassifikation gesprochen. Ein solches Problem könnte beispielsweise sein, handgeschriebene Zahlen zu erkennen und auszugeben. Dabei gibt es zehn verschiedene Klassen, 0 bis 9, in welche die zu erkennende Ziffer fallen kann.
In Abbildung 2.1 sind Datensätze einer binäre Klassifikation dargestellt. Dabei werden die Datenpunkte durch die Attribute 'Höhe des Kredites' und 'Jährliches Einkommen' beschrieben. Die Datenpunkte wurden bereits in zwei Klassen eingeteilt. Dabei sind die Kreise die gewährten Kredite und die Kreuze die nicht gewährten Kredite. Die Aufgabe ist es nun eine bestimmte Funktion zu finden und deren Parameter so anzupassen, sodass die zwei Klassen möglichst gut durch diese Funktion trennbar sind. In Wirklichkeit kommen sehr viel mehr Attribute zum Einsatz. Die Anzahl dieser Attribute ist unbegrenzt, nur steigt mit steigender Anzahl der Attribute auch die Komplexität des Modells und die Anforderung an die Rechenleistung der Maschine. Bei n Attributen muss eine (n-1)-dimensionale Trennfläche gefunden werden. [19] Für die Entscheidung welche Funktion verwendet werden soll, stehen eine Vielzahl von Möglichkeiten und unterschiedlichen Verfahren zu Verfügung. Für die Auswahl des richtigen Verfahren gibt es keine eindeutige Regel, sondern es ist vielmehr die Erfahrung des Anwenders von Bedeutung.

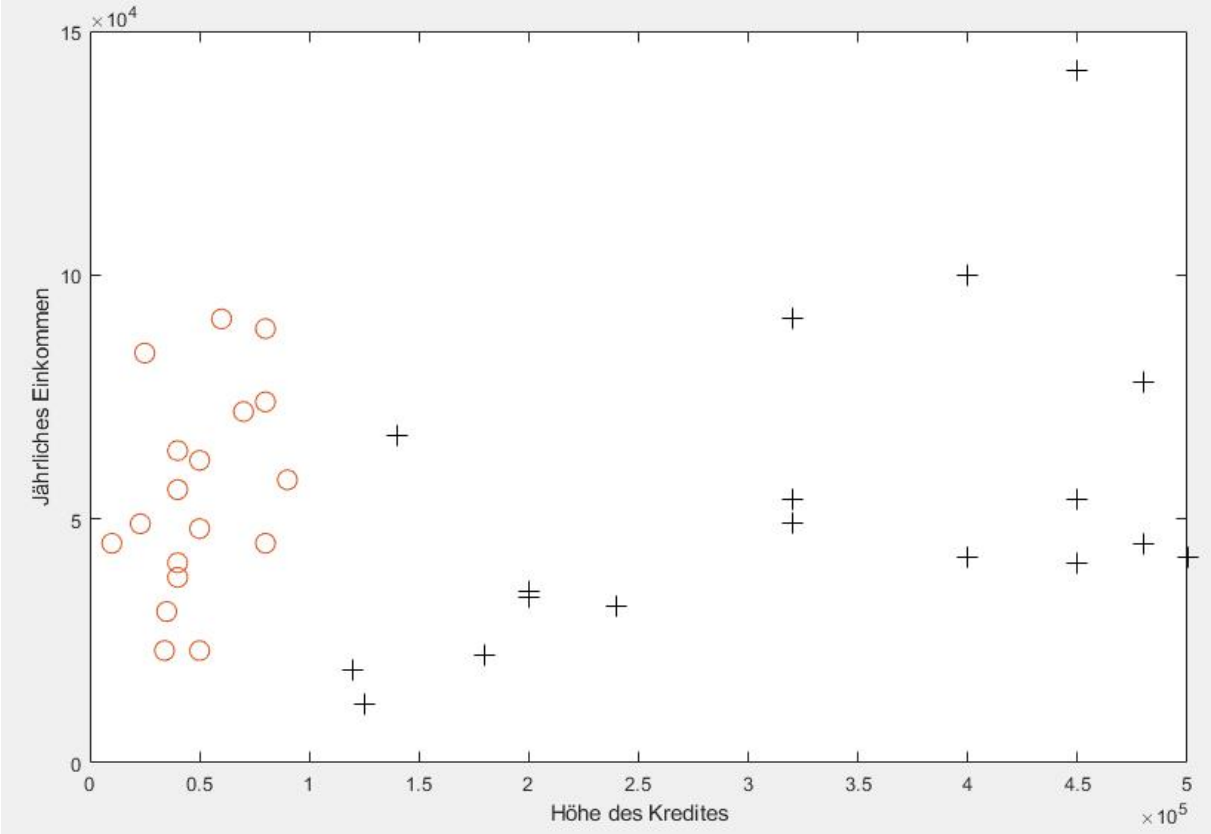

Abbildung 2.1: Beispiel Klassifikation

2.2.2 Regression

Bei der Regression ist das Ziel ein Modell zu erstellen, dass für den gegebenen Input einen genauen numerischen Ausgangswert bestimmt. Wie auch bei der Klassifikation werden hier Trainingsdaten mit schon richtig zugeordnetem Output verwendet, um den Algorithmus zu trainieren. Anstatt diese Daten aber in unterschiedliche Klassen zu trennen, sind den Daten konkrete Zahlenwerte zugeordnet. Nun kann eine Hypothesenfunktion gebildet werden, die diese konkreten Zahlenwerte anhand der Eingangsdaten berechnet. Eine Hypothesenfunktion mit nur einem Attribut kann folgendermaßen aussehen:

$$h_\theta(x) = \theta_0 + \theta_1 x \tag{2.1}$$

Hierbei ist $h_\theta(x)$ der Hypothesenwert, x sind die Werte des Attributs und θ_0 und θ_1 sind Parameter. Diese Parameter werden über einen Alogrithmus bestimmt. Ziel ist es die Parameter so zu bestimmen, dass der Vorhersagefehler minimal ist. Dazu wird eine sogenannte Kostenfunktion gebildet, welche in Abhängigkeit der Parameter θ_0 und θ_1 die kumulierte Abweichung zwischen dem Vorhersagewert $h_\theta(x)$ und dem realen Wert für alle gegebenen x-Werte berechnet. Ziel ist es nun für diese Kostenfunktion die Parameter zu

finden, sodass die Funktion ihren minimalen Wert erreicht. Hierfür stehen unterschiedliche Algorithmen zur Verfügung. Der am häufigsten verwendete ist das Gradientenverfahren, bei welchem die Parameter so angepasst werden, dass der Funktionswert in Richtung des steilsten Abstieges geändert wird. Dies wird solange wiederholt, bis ein weiteres Fortschreiten zu keiner Verminderung des Funktionswertes führt. [30] Sind die Parameter bestimmt, kann mit der Hyphothesenfunktion der Ausgangswert für unbekannte x-Werte berechnet werden. Das hier dargestellte Verfahren heißt lineare Regression. Dabei wird eine Gerade ermittelt, welche zu den Punkten der Trainingsdaten den kleinsten Abstand hat. Dabei ist θ_0 der Ordinatenabschnitt und θ_1 die Steigung der Geraden. Ein Beispiel einer solchen linearen Regression ist in Abbildung 2.2 dargestellt. In diesem Beispiel hängt der Wohnungspreis alleine von der Wohnfläche in m^2 ab. Die Gerade bildet dabei die Hypothesenfunktion ab, welche mit Hilfe der bereits bekannten Datenpunkte ermittelt wurde. Gibt es nun beispielsweise eine neue Wohnung auf dem Markt, zu der die vorhandene Wohnfläche bekannt ist und für die der Verkaufspreis bestimmt werden soll, kann über die Geradengleichung der geschätzte Verkaufspreis bestimmt werden.

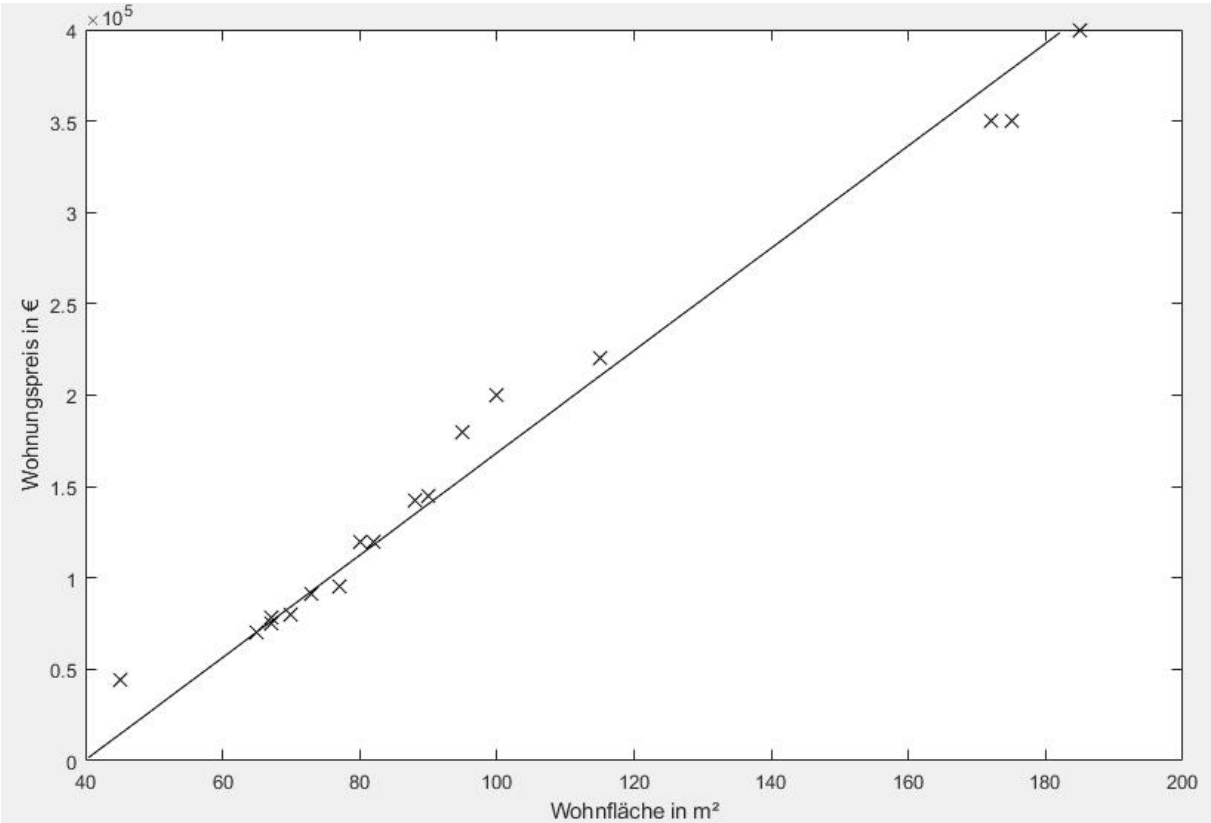

Abbildung 2.2: Beispiel Regression

In der Realität spielen viel mehr Attribute eine Rolle, dadurch wird das zu lösende Problem auch mehrdimensional und es muss für eine komplexere Kostenfunktion der Tief-

punkt gefunden werden. Ist die Hypothesenfunktion nicht linear sondern beinhaltet weitere Attribute, wird von polynomialer Regression gesprochen.

2.2.3 Probleme beim Supervised Learning

Um geeignete Modelle erstellen zu können, müssen auch auftretende Schwierigkeiten verstanden werden. Bei der Wahl der richtigen Funktion für Separation der Klassen oder für die Funktion für die Regression können beliebig große Polynome verwendet werden. Je größer das Polynom, desto weniger Fehler werden beim Bestimmen der Parameter mit Hilfe der Trainingsdaten erlaubt. [30] Abbildung 2.3 veranschaulicht diesen Sachverhalt.

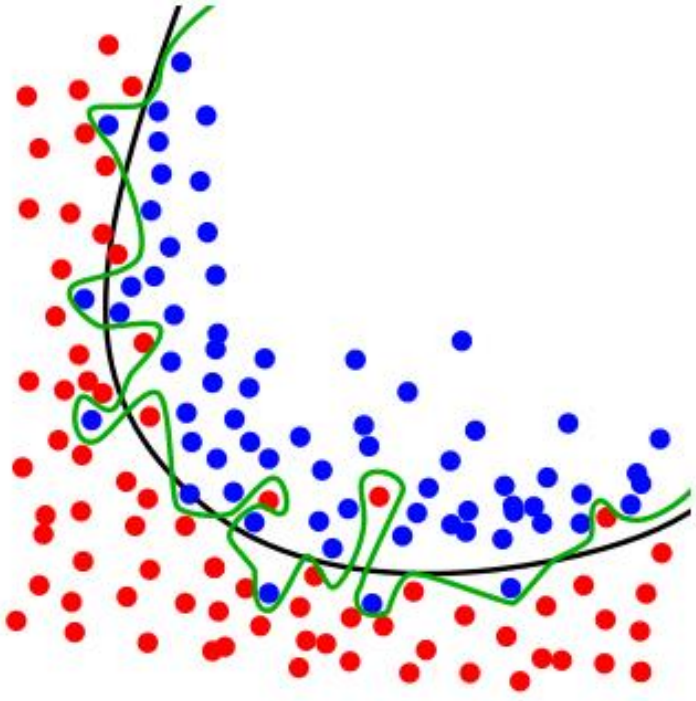

Abbildung 2.3: Problem des Overfitting [61]

Während die Funktion A, dargestellt durch die schwarze Linie, einige Fehler zulässt, versucht die Funktion B, dargestellt durch die grüne Linie, alle Trainingsdaten der richtigen Klasse zuzuordnen. Die Funktion B führt für die gegebenen Trainingsdaten zu einer sehr hohen Erfolgsquote und damit Performance, kann aber beim Anwendung auf andere Daten zu verminderter Performance führen. Tom Mitchell definiert diesen Sachverhalt, der als Overfitting bekannt ist folgendermaßen: *„Given a hypothesis space H, a hypothesis h ϵ H is said to overfit the training data if there exists some alternative hypothesis h' ϵ H, such that h has smaller error than h' over the training examples, but h' has a smaller error than h over the entire distribution of instances.“* [36] Dieser Grat zwischen einer zu einfachen und einer zu komplexen Funktion kann nur durch Erfahrung und Testen des Modells

herausgefunden werden. In der Praxis wird das vorhandene Datenset in Trainingsdaten und Testdaten getrennt. Die Testdaten dienen dazu, das trainierte Modell mit Hilfe neuer Daten zu testen, zu denen der richtige Output bekannt ist. Dadurch kann die Vorhersagen des Modells verifiziert werden und überprüft werden, ob das Modell auch für neue Daten eine gute Performance liefert.
Ein weiterer großer Nachteil des Supervised Learning ist das Labeln der Daten. Ein menschlicher Experte muss die Datensätze dem richtigen Output manuell zuordnen. Dies ist, wie in Kapitel 2.1 bereits beschrieben sehr zeitintensiv und daher kostspielig.

2.3 Unsupervised Learning

In den Modellen des Supervised Learning sind wir immer davon ausgegangen, dass Trainingsdaten zur Verfügung stehen, bei denen bereits Klassen von menschlichen Experten gebildet wurden und dass bei diesen Trainingsdaten bereits eine Zuordnung zu den definierten Klassen oder zu kontinuierlichen Werten erfolgt ist. Dies ist allerdings nicht immer möglich oder aufgrund des großen Aufwandes der Aufbereitung der Trainingsdaten nicht gewünscht. Die Methoden des Unsupervised Learnings benötigen keine vorher preperierten Trainigsdaten um Wissen aus einer Datenmenge zu extrahieren. Es werden Strukturen und Muster von der Maschine gefunden, ohne dass eine Vorstrukturierung, beziehungsweise eine „Supervision“ durch einen Menschen erfolgt. [3] Beim Unsupervised Learning geht es im Allgemeinen weniger um das Vorhersagen bestimmter Outputs, es geht vielmehr um das Darstellen von Beziehungen und Mustern der Daten und damit um das Extrahieren von Wissen oder Information aus einer Datenmenge, die für den Menschen beim Betrachten der Datenmenge nicht ersichtlich sind. [48] Dazu werden beispielsweise Ähnlichkeiten einzelner Datensätze berechnet. Das mit am häufigsten verwendete Unsupervised Learning Verfahren hierzu ist das Clustering, welches im Folgenden genauer erläutert wird.

2.3.1 Clusteranalyse

Die Clusteranalyse ist eine der meistgenutzten Verfahren im Unsupervised Learning. Dabei werden die Daten in Untergruppen, beziehungsweise Cluster oder Klassen segmentiert, in welchen sich die Daten möglichst ähneln. Die einzelnen Cluster müssen dabei aber genügend große Unterschiede aufweisen. So können Zusammenhänge der Daten erkannt werden. Um die Segmentierung vorzunehmen, wird mit einem Abstandsmaß die Ähnlichkeit der Daten bestimmt. Es wird also berechnet, wie weit die einzelnen Datenpunkte auseinanderliegen. Das Ziel ist, die Abstände innerhalb eines Clusters zu minimieren, während der Abstand der Cluster untereinander maximiert wird. [25] Für diese

Abstandsmaßberechnung stehen unterschiedliche Verfahren zur Verfügung. Auch unterscheiden sich die Verfahren hinsichtlich der Erzeugung der Cluster. Es gibt Verfahren, bei denen ist die Anzahl der Cluster vorbestimmt.
Die Clusteranalyse wird oft als ein Vorprozess für die Supervised Learning Verfahren der Klassifikation verwendet. So können Cluster beziehungsweise Klassen ermittelt werden, welche verwendet werden können um einen Klassifikationsalgroithmus zu trainieren. [1]

2.3.2 Outlier Detection

Ein weiteres häufig verwendetes Verfahren ist Outlier Detection. Dabei wird versucht Datenpunkte ausfindig zu machen, die sich von dem Rest der Daten ungewöhnlich stark unterscheiden. Dies wird manchmal als Gegenstück zu der Clusteranalyse gesehen, bei dem die Ähnlichkeit der Daten untersucht wird. Beispielsweise werden bei Outlier Detection Algorithmen Datenpunkte gesucht, die sich keinem Cluster eindeutig zuordnen lassen. [1] Outlier Detection wird oft für die Datenbereinigung verwendet. Da zu große Ausreißer oft fehlerhafte Daten darstellen und daher zu einem falsch trainierten Algorithmus führen können. Ein realer Anwendungsfall ist das Erkennen von Kreditkartenmissbrauch.

2.4 Künstliche neuronale Netze

Neuronen im menschlichen Gehirn sammeln Signale von anderen Neuronen über sogenannte Dendriten. Die Signale werden in Form von elektrischen Impulsen über Axone weitergeleitet. Am Ende eines Axons, teilt es sich in viele kleine Verzweigungen, die Dendriten auf. Am Ende der Dendriten kann über eine Synapse das elektrisches Signal an weitere Neuronen weitergegeben werden. Erhält eine Synapse ausreichend große erregende Eingangssignale, im Vergleich zu hemmenden Eingangssignalen, dann sendet es elektrische Signale zu weiteren Neuronen. [25] Auch künstliche neuronale Netzwerke (KNN) bestehen aus vielen eng miteinander verbundenen Einheiten, welche ebenfalls als Neuronen bezeichnet werden. Jeder dieser Neuronen nimmt eine verschieden große Anzahl an Eingangswerten entgegen, führt mit diesen Berechnungen durch und produziert daraus einen Ausgangswert, welcher wiederum ein Eingangswert für ein weiteres Neuron sein kann. Mit KNN können sowohl Regressions- als auch Klassifikationsprobleme, sowie Probleme des Unsupervised Learning gelöst werden. [36] Auch wenn diese Einheiten den Neuronen in einem Gehirn ähneln, wird bei den KNN nicht versucht das menschliche Gehirn nachzubilden. Es geht lediglich um eine Modellbildung, bei der versucht wird über mehrere Ebenen hinweg eine Abstraktion des Wissens zu erzielen. [3] Jede dieser Ebenen besteht aus einem oder mehrerer Neuronen, die untereinander vernetzt sind. Jede Verbindung zwischen zwei Neuronen ist mit einem Gewicht versehen, welche die Stärke der Verbindung darstellt

und bei der Berechnung des Outputs eine Rolle spielt. Diese Gewichte werden bei dem Lernprozess angepasst, sodass die optimalen Gewichte für alle Verbindungen ermittelt werden. Der Lernprozess besteht auch hier aus dem verarbeiten von Trainingsdaten. Dies kann als Supervised oder Unsupervised Learning erfolgen. Neben der Qualität und der Menge der Trainingsdaten, trägt auch die Architektur des Netzwerkes maßgeblich zur der Performance des künstlichen neuronalen Netzwerkes bei. Hierfür gibt es eine Vielzahl an unterschiedlichen Architekturen, die über die Jahre entwickelt wurden, welche von einfachen Netzwerken mit nur einer Ebene und wenigen Neuronen pro Schicht, bis hin zu Netzwerken mit vielen Schichten und mehreren Neuronen pro Ebene variieren. [1] Künstliche neuronale Netze werden als sogenannte Black Box Modelle bezeichnet. Das bedeutet es ist nicht ersichtlich wie genau das KNN „denkt“, also die Daten verarbeitet. [25]

2.4.1 Aufbau

Wie bereits beschrieben, besteht ein KNN aus mehreren Ebenen, wie in Abbildung 2.4 gezeigt. In der Inputebene werden die Eingangssignale verarbeitet. In der Outputschicht wird das Resultat ausgegeben, also zum Beispiel die Klasse der Eingangsdaten. Zwischen der Input und der Outputschicht können weitere Ebenen liegen, welche als Hidden Layers, also versteckte Ebenen bezeichnet werden. Diese werden als versteckt bezeichnet, da ihr Output nur innerhalb des Netzwerkes und nicht als Gesamtoutput verfügbar ist. [36] Die Anzahl der künstlichen Neuronen innerhalb einer Schicht, hängt ebenfalls stark von dem zu lösenden Problem ab.

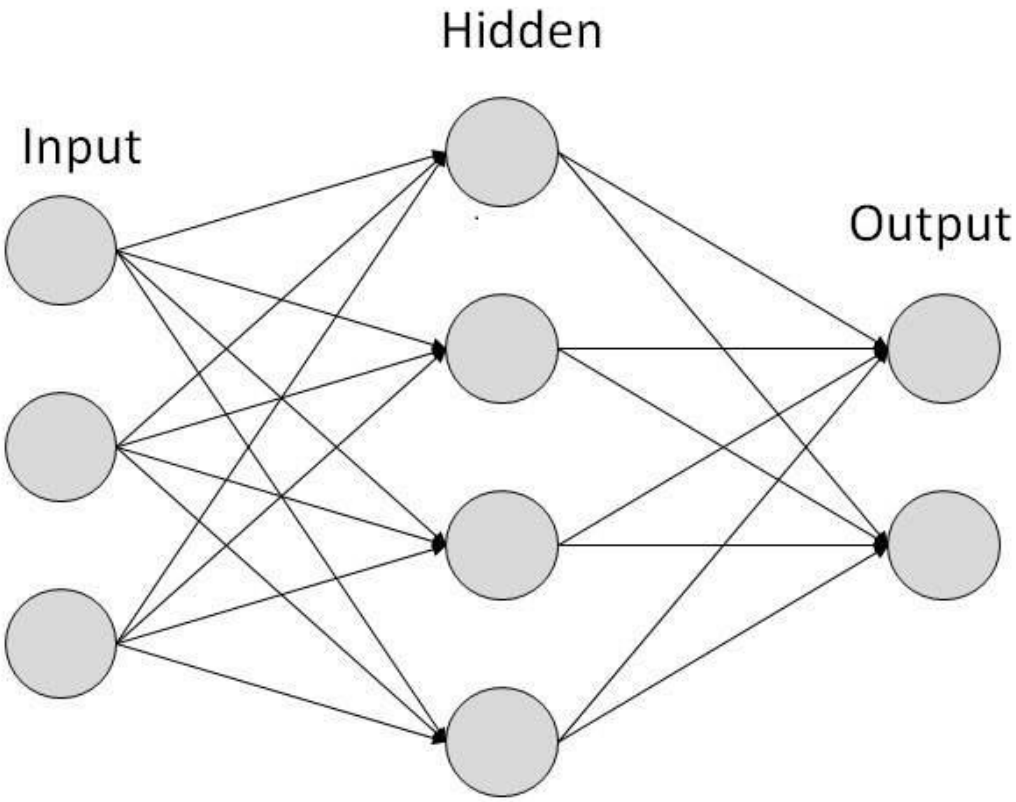

Abbildung 2.4: Aufbau künstlichen neuronales Netzwerk [56]

Die Funktionsweise eines Neurons in einem KNN ist in Abbildung 2.5 dargestellt. Jedes dieser künstlichen Neuronen hat eine bestimmte Anzahl p an Eingangssignalen x_i, welche mit einem bestimmten Gewicht w_i gewichtet werden. Die Eingangssignale werden mit dem zugehörigen Gewicht multipliziert und anschließend addiert. Zusätzlich zu diesem Wert wird ein sogenannter Bias b hinzu addiert. Dieser Bias ist ein fester Wert und kann positiv oder negativ sein und kann damit den Output verstärken oder hemmen. Der Bias wird, wie auch die Gewichte, während des Lernprozesses angepasst. Der Resultierende Wert wird an eine Aktivierungsfunktion θ weitergeben, welche nun ein bestimmtes Ausgangssignal y berechnet, welches an weitere Neuronen geschickt wird. Alle Neuronen innerhalb einer Ebene arbeiten hierbei parallel. [3]

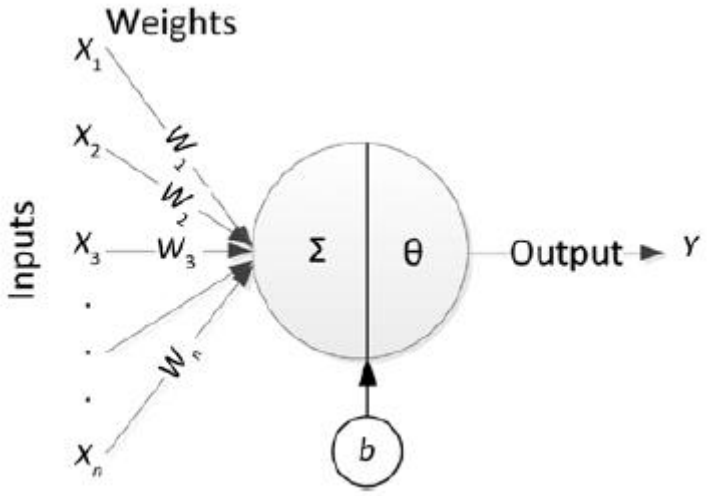

Abbildung 2.5: Aufbau künstlichen Neuron [3]

Es gibt mehrere Arten von Aktivierungsfunktionen, die verwendet werden können. So kann der Output beispielsweise einen binären (1 oder 0) oder einen analogen Wert annehmen.

2.4.2 Einlagiges Perceptron

Das einlagige Perceptron ist die einfachste Form eines KNN. Es besteht lediglich aus einem einzigen künstlichen Neuron in der Outputebene, das mehrere Eingangssignale der Inputebene mit den jeweiligen Gewichten und einen Bias entgegennimmt und daraus eine binäre Entscheidung trifft. [25] Der generelle Aufbau ist demnach wie in Abbildung 2.5 bereits gezeigt. Dabei beträgt die Menge der Eingangssignale exakt der Menge der Attribute der verwendeten Daten. Jedes Neuron der Inputebene empfängt genau ein Attribut und leitet es ohne eine Berechnung durchzuführen an das Neuron in der Outputebene weiter. Nur das Neuron in der Outputschicht führt eine Berchnung aus. [1]
Das Perzeptron ist ein Lernverfahren, welches zwei linear separable Mengen trennen kann. Demnach gibt es zwei Mengen M1 und M2, die sich durch eine lineare Funktion in zwei Klassen unterteilen lassen. Die Eingangsdaten werden also durch das Perzeptron einer der beiden Klassen zugeordnet. Dies geschieht, indem der Ausgabewert der Aktivierungsfunktion entweder den Wert 1 für die Menge M1 oder 0 für die Klasse M2 annimmt. Der Ausgabewert wird mit folgender Formel [19] berechnet:

$$\theta(x) = \begin{cases} 1 & \text{für } w * x^T + b > 0 \\ 0 & \text{für } w * x^T + b \leq 0 \end{cases} \tag{2.2}$$

Dabei ist x der Vektor der Eingangssignale. Bei dem einlagigem Perzeptron entspricht das dem Attributsvektor der Daten. w ist der Vektor der Gewichte, welche die Verbindungen

Tabelle 2.2: Trainingsdaten des Perceptron

$\mathbf{x_1}$	-4	5	6	4	2	-3	-6	-2	4	11	10	8
$\mathbf{x_2}$	-2	1	3	4	6	4	-5	-5	-4	-1	-5	-3
$\mathbf{y}$	0	0	0	0	0	0	1	1	1	1	1	1

von Inputebene und Outputebene gewichten. Demnach ist das Produkt $w * x^T$ nichts anderes als $\sum_{i=1}^{n} w_i * x_i$, wobei n die Anzahl der Attribute ist.
Nehmen wir nun an wir haben bereits gelabelte Daten, welche durch zwei Attribute x_1 und x_2 beschrieben werden, siehe Tabelle 2.2. Die vorhandenen Datenpunkte sind in Abbildung 2.6 mit der dazugehörigen Klasse abgebildet.

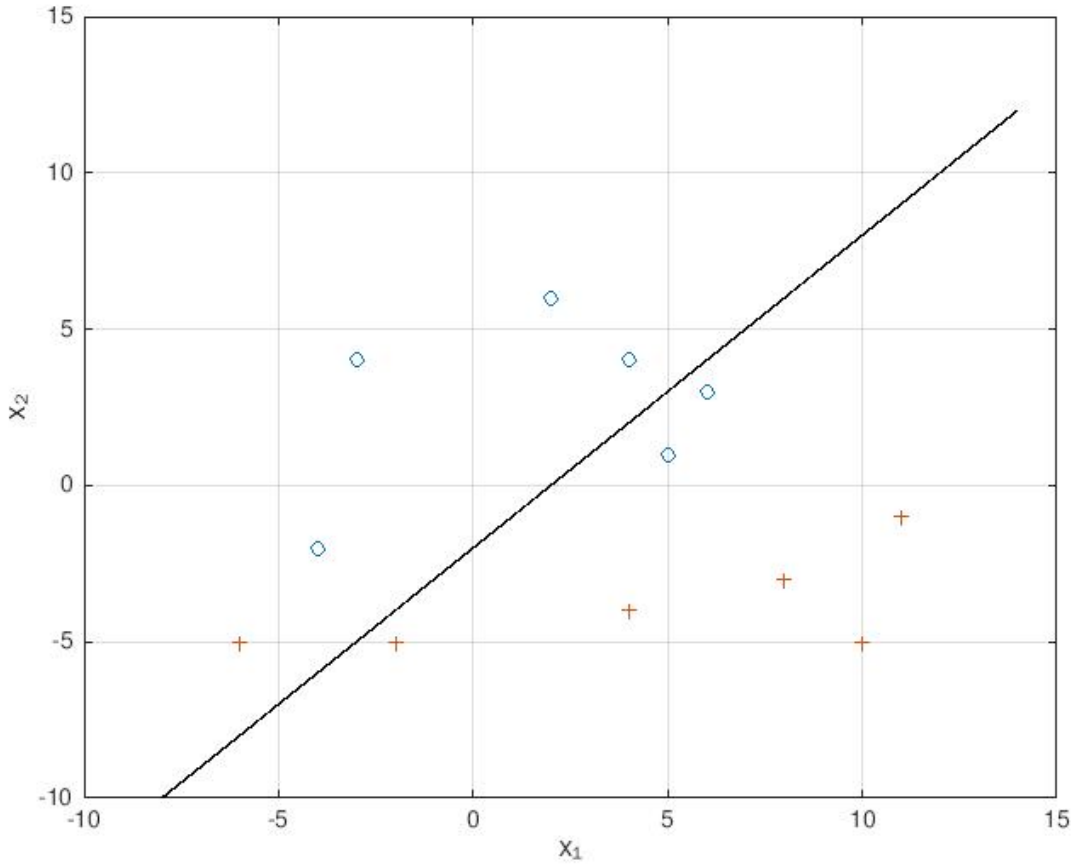

Abbildung 2.6: Beispiel Perzeptron Klassifizierung

Die Gleichung der Trennebene der beiden Klassen kann durch $w * x^T + b = 0$ dargestellt werden (hier eine Gerade). Zu Beginn werden die Gewichte und der Bias zufällig gewählt (hier: $w_1 = 1$, $w_2 = -1$, $b = -2$). In diesem Beispiel ergibt sich die Geradengleichung:

$$1 * x_1 - 1 * x_2 - 2 = 0$$

Wie sich erkennen lässt, schaffen die gewählten Gewichte und der Bias es nicht die zwei

Klassen vollständig zu trennen. Ziel ist es nun die Gewichte und den Bias so anzupassen, dass die abgebildete Gerade die beiden Klassen vollständig separiert. Bei dem Lernprozess des Perceptron werden nun der Bias und die Gewichte für die Datenpunkte angepasst, bei denen eine falsche Klassifizierung durchgeführt wurde. [1] Eine falsche Klassifizierung liegt vor wenn für einen Datenpunkt gilt:

$$\theta(x) = 1 \text{ und } y = 0$$

$$\theta(x) = 0 \text{ und } y = 1$$

Beispielsweise ergibt sich für den Datenpunkt (-6,-5) eine falsche Klassifizierung:

$$w * x^T + b < 0, \text{ da } 1 * (-6) - 1 * (-5) - 2 < 0$$

und damit $\theta(-6, -5) = 0$, $y(-6, -5) = 1$. In unserem Beispiel liegt somit für die Punkte $(-6, -5)$, $(5, 1)$ und $(6, 3)$ eine falsche Klassifizierung vor. Ein möglicher Algorithmus, um die gewünschten Gewichte und den gewünschten Bias zu erreichen ist folgender:

1. Für jeden Datenpunkt für den gilt y = 1 und $\theta(x) = 0$ rechne $w = w + \frac{x}{|x|}$ und $b = b + 1$
2. Für jeden Datenpunkt für den gilt y = 0 und $\theta(x) = 1$ rechne $w = w - \frac{x}{|x|}$ und $b = b - 1$
3. Wiederhole das ganze solange bis alle Datenpunkte richtig Klassifiziert sind.

Dies hat zur Folge, dass sich die Gerade nach jeder Iteration in eine Richtung verschiebt. War der Term $w * x^T + b$ negativ (Klassifiziert als 0), hätte aber positiv sein müssen (richtige Klasse 1), werden die Gewichte und der Bias in die positive Richtung geändert und anders herum. Des Teilen des Vektors x durch seinen Betrag hat zur Folge, dass jedes Attribut gleich viel zur Verschiebung der Trennfläche beiträgt. [19]

2.4.3 Deep Learning

Im vorherigen Abschnitt wurde die Funktionsweise eines einzelnen Neurons innerhalb des einlagigen Perceptrons beschrieben. In der Praxis gilt es aber oft weitaus komplexere Probleme als eine linearer Klassifikation zu lösen. Dazu werden KNN mit mehreren Ebenen verwendet. Besitzen diese KNN mindestens eine versteckte Ebene, wird von sogenannten Multilayer Neural Networks gesprochen. Ein Netzwerk mit mehreren verstecken Ebenen wird auch als Deep Neural Network (DNN) bezeichnet. Das Trainieren solcher Netzwerke heißt Deep Learning. Dabei muss die Anzahl der Ebenen und die Anzahl der Neuronen innerhalb der Ebenen von dem Analyst bestimmt werden. [1] Das Training des einlagigen Perceptron war dahingegen einfach, da es nur einen Output zu bestimmen gab und die

Möglichkeiten dieses Outputs bekannt waren. Daher konnte dem Neuron genau mitgeteilt werden, ob seine Einteilung der Gewichte und des Bias richtig war. Bei DNN ist dies nicht in direkter Weise möglich. Jede Verbindung zwischen zwei Neuronen besitzt ein eigenes Gewicht und einen Bias. Beim Trainieren eines DNN ist es allerdings nicht möglich jedem einzelnen Neuron direktes Feedback darüber zu geben, ob sein Output zur richtigen Lösung beigetragen hat oder nicht. Ein Beispiel soll dies verdeutlichen.
Ziel der DNN ist es ein sehr komplexes Problem, wie beispielsweise das Klassifizieren von Bildern nach ihrem Inhalt, durch das Herunterbrechen des Problems in einfachere Teilprobleme zu lösen. Ein Beispiel zur dieser Funktionsweise ist in Abbildung 2.7 gezeigt. Hier wird in der ersten Ebene des DNN die Kanten der gezeigten Objekte klassifiziert, indem die Helligkeitsunterschiede einzelner Pixel miteinander verglichen werden. In den weiteren Ebenen wird aus diesen einfacheren Erkenntnissen ein zunehmend komplexerer Sachverhalt abgebildet. In der zweiten Ebene werden durch die klassifizierten Kanten, Konturen und Formen zusammengesetzt. In der dritten Ebene werden diese Konturen und Formen dazu verwendet einzelne Teile der Objekte zu identifizieren. Durch die Erkenntnisse, welche Teile von Objekten auf dem Bild abgebildet sind, lässt sich schlussendlich eine Entscheidung treffen, um was für ein Objekt es sich handelt.
Als Anwender kann nicht bestimmt werden, welche Einteilung des DNN in den versteckten Ebenen richtig war, da es für diese Teilschritte keine gelabelten Daten gibt, mit denen eine solche Einschätzung möglich wäre. [1] Es kann also einem einzelnen Neuron der ersten verstecken Ebene, welches mehrere Pixel als eine bestimmte Kante klassifiziert hat, nicht mitgeteilt werden, ob dieser Schritt zur richtigen Klassifizierung des Bildes beigetragen hat. Wie das Netzwerk die einzelnen Teilschritte handhabt und entscheidet, welche Unterteilung sinnvoll ist und welche Verbindungen zwischen den Neuronen zur richtigen Klassifizierung beitragen, muss das Modell durch den Lernalgorithmus selbst bestimmen. Es ist lediglich möglich die Daten der Inputebene und der Outputebene zu überwachen, daher werden diese Ebenen auch als sichtbare Ebenen bezeichnet. [24] Der meistgenutzte Algorithmus für DNN ist der Backpropagation Algorithmus, auf welchen im Rahmen dieser Arbeit nicht weiter eingegangen wird.

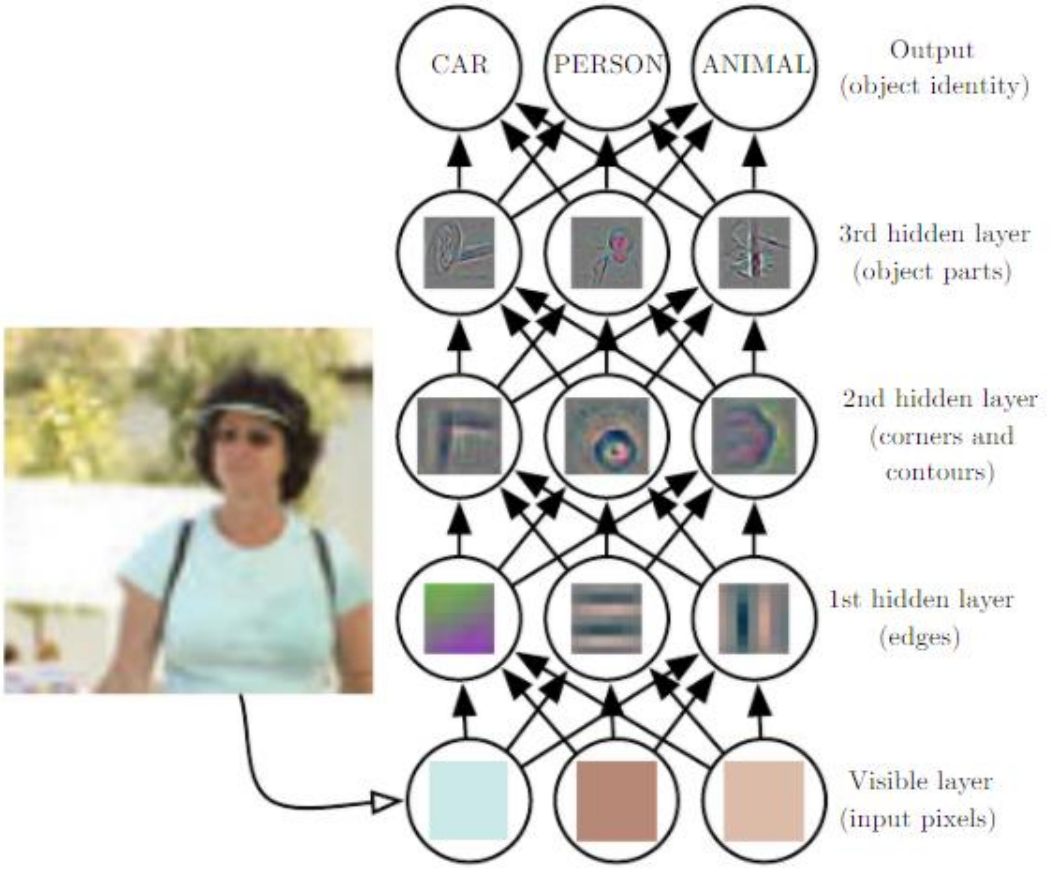

Abbildung 2.7: Funktionsweise eines Deep Neural Network [24]

2.4.4 Hardware-technische Umsetzung

Wie bereits erwähnt wurde, hängt die Performance und damit, wenn man so möchte die ‚Intelligenz' eines künstlichen neuronalen Netzwerkes von der Architektur des Netzwerkes ab. Während das lineare Perceptron einfache binäre Klassifizierungen durchführen kann, werden beispielsweise für ein Netzwerk, welches bei selbstfahrenden Autos verwendet wird, Netzwerke mit weit über Tausend Neuronen, 27 Millionen Verbindungen und 250 Tausend Parametern verwendet. [6] Während früher Central Processing Units (CPUs) für derartige Berechnungen verwendet wurden, wird heute auf Graphics Processing Units (GPUs) zurückgegriffen. Die GPUs wurden für die Gamingindustrie entwickelt, da dort in der Bildverarbeitung viele gleichzeitige Operationen ausgeführt werden. Diese Operationen bestehen aus vielen einfachen Berechnungen, welche unabhängig voneinander durchführbar sind. Durch die Einfachheit und Unabhängigkeit dieser Berechnungen, wird nicht viel Rechenleistung für die einzelnen Berechnungen benötigt und diese lassen sich parallel durchführen. Die CPUs hingegen können komplizierte Berechnungen ausführen. Dabei

können sie allerdings nur eine Berechnung zu einem bestimmten Zeitpunkt ausführen und die nächste Berechnung kann erst starten, wenn die vorherige Berechnung beendet wurde. KNN benötigen die gleichen Recheneigenschaften, wie sie auch in der Gamingindustrie bei der Bildverarbeitung benötigt werden. Die einzelnen Berechnungen die bei jedem Neuron ausgeführt werden, sind nicht kompliziert und die Berechnungen innerhalb einer Ebene hängen auch nicht voneinander ab, dafür müssen sehr viele dieser Berechnungen zur selben Zeit ausgeführt werden. Das einzige Problem, welches sich ergab war, dass die verwendeten GPUs für Berechnungen in der Bilderverarbeitung ausgelegt waren. Der richtige Erfolg der GPUs wurde durch sogenannte General Purpose GPUs (GP-GPUs), welche alle möglichen Berechnungen durchführen konnten, erzielt. Die Rechenleistung einer einzelnen Maschine reicht allerdings für größere Anwendungen nicht aus. Dafür werden die Rechenaufgaben auf mehrere Maschinen verteilt. Dabei kann beispielsweise jedes Inputattribut von einer anderen Maschine verarbeitet werden. [24]

2.4.5 Probleme

Schlechte Datenqualität ist das größte Problem, welches es bei einem profitablen Einsatz von Machine Learning Techniken gibt. Hier gibt es gleich zwei Gebiete, die beachtet werden müssen. Einmal die historischen Daten, die verwendet werden, um den Algorithmus zu trainieren und zum anderen die Daten, die es später gilt von dem trainierten Algorithmus zu verarbeiten. Die Anforderung an die historische Datenqualität sind besonders hoch. Die Daten müssen unter anderem korrekt, vollständig, unvoreingenommen und in den meisten Anwendungsfällen auch richtig gelabelt sein. Dazu müssen genügend Datensätze vorhanden sein, die diesen Qualitätsanforderungen entsprechen. Mehr Daten müssen dabei nicht immer eine bessere Performance bedeuten, es ist auch wichtig, dass die Daten ausreichend diversifiziert sind, um einen guten generalisierten Algorithmus zu entwickeln. Aber auch ein sehr gut trainierter Algorithmus kann schlechte Ergebnisse liefern, wenn er auf eine Datenmenge von schlechter Qualität angewendet wird. [14] Ein System welches Daten nutzt, um Vorhersagen zu treffen, kann nur so gut sein, wie die Daten mit dem es trainiert wurde.
Ein Faktor der nicht unbedingt an mangelnder Datenqualität liegen muss, ist die genannte Voreingenommenheit, im englischen als Bias bezeichnet. Dies kann sich auf zwei Arten zeigen. Zum einen wenn vorhandene Trainingsdaten zu einseitig sind. Ein Beispiel hierfür ist Gesichtserkennungssoftware. Eine kürzlich veröffentlichte Studie untersuchte, wie gut Software auf Bildern abgebildete Gesichter unterschiedlicher Menschen richtig, als Mann oder Frau identifizieren konnte. Dabei wurde festgestellt, dass die Software, darunter Software von Microsoft und IBM, hellere Gesichter besser klassifizieren konnte als dunklere Gesichter. [10] Ein Grund hierfür könnte beispielsweise sein, dass mehr gela-

belte Trainingsdaten von hellen Gesichtern vorliegen, als von dunklen Gesichtern und der Algorithmus daher eine größere Wissensbasis über helle Gesichter hat.
Eine andere Art des Bias liegt vor, wenn Zusammenhänge in die Entscheidungsfindung mitaufgenommen werden, welche von Menschen als rassistisch oder vorurteilsbehaftet angesehen werden. Ein Beispiel hierfür ist die Herkunft oder Hautfarbe eines Menschen, wenn es darum geht ob ein Darlehn gewährt werden soll oder nicht. Und da es sich bei den genutzten System um schon erwähnte Black Box Modelle handelt, ist nicht ersichtlich welche Faktoren genau für die Entscheidungsfindung genutzt werden und ob eventuell ethische bedenkliche Faktoren mit einbezogen wurden. [52] Es werden zwar Forschungen in diese Richtung betrieben, aber bis lang gibt es noch keine Möglichkeit Einblicke in den Entscheidungsprozess der DNN zu erlangen.
Ein weiteres großes Problem, das im Zusammenhang mit diesem Nichtwissen über den Entscheidungsprozess besteht wird folgend beschrieben. Wenn zum Beispiel ein autonomes Auto fährt, werden die Informationen der Sensoren direkt in ein riesen Netzwerk von künstlichen Neuronen geschickt und verarbeitet. Daraus wird eine Entscheidung getroffen was gemacht werden soll. Es kann allerdings nicht herausgefunden werden, wie genau das DNN zu der Entscheidung gekommen ist. Kommt es nun zu einer unvermeidbaren Unfallsituation bei der es um Leben und Tod geht, stößt dieses Unwissen teilweise auf ethische Bedenken. [55]
In der Medizin gibt es zum Beispiel schon Software, die Krebserkrankungen sehr zuverlässig erkennen können. Nur lassen sich durch die angewendeten DNN, die ausschlaggebenden Attribute, welche zur Klassifizierung des Patienten genutzt wurden, nicht erkennen. Somit kann der Arzt die Entscheidung des Algorithmus nicht Nachvollziehen oder überprüfen. [29] Allerdings gibt es auch Medikamente, bei welchen nicht komplett erforscht ist, warum diese wirken. Solang die angewendete Methode zum Erfolg führt, gibt es auch Ärzte die hier großes Potential sehen. [8]
Ein weiteres Problem das in der Literatur angesprochen wird, sind die hohen Rechenanforderungen, die an die Hardware gestellt werden. Die Algorithmen müssen in Echtzeit auf durch Sensoren erfasste Daten reagieren können. Für diese Anwendungen werden oft lokal installierte Maschinen, gegenüber Cloud-Lösungen bevorzugt. Gründe dafür sind Datensicherheit, Grenzen in der Übertragungsrate und Risiken durch Übertragungsabbruch. Die für diese Systeme benötigten Rechenleistungen müssen also lokal über Maschinen realisiert werden, beispielsweise bei autonomen Fahren oder Drohnen. Dabei stellen unter anderem auch die damit Verbundenen hohen Energiekosten ein Problem dar. [54]

3 Data Scientists

3.1 Berufsentwicklung

Es werden in den Unternehmen täglich viele Daten gesammelt. Um diese Daten sinnvoll zu nutzen, wurden sogenannte Data Warehouses eingeführt. Dies sind Umgebungen, in welche alle operativen Daten geladen werden. Innerhalb dieser Umgebung können nun Analysen erstellt werden. Früher wurden dazu sehr gut ausgebildete Arbeitskräfte benötigt, welche die Daten sammelten, bereinigten und in die Umgebung transferierten. Des Weiteren mussten auch Programme zur Analyse der Daten geschrieben werden und die gewünschte Analyse durchgeführt werden. Heutzutage sind die Tools im Bereich Date Warehousing schon so weit entwickelt, dass jeder der sich etwas in diese Thematik einarbeitet Analysen durchführen kann. Der Datenfluss der operativen Daten in die Analyseumgebung ist durch sogenannte Data Pipelines automatisiert und es gibt sehr benutzerfreundliche Tools, welche eine Analyse und Visualisierung der Daten durch Drag & Drop Befehle möglich machen.
In vielen Firmen sind diese Data Piepelines schon sehr gut entwickelt, nur werden in vielen dieser Firmen bereits heute so eine große Menge an Daten generiert, dass es für Menschen unmöglich ist all diese Daten zu verwerten und Analysen zu erstellen, deshalb werden Maschinen benötigt, die diese Aufgabe übernehmen können. [28] Des Weiteren liegen die Daten nicht nur in Tabellenform vor, sondern in komplexeren Formen, wie zum Beispiel Bilder oder Videos und lassen sich daher nicht mehr so einfach in die Data Warehouses integrieren. Um den Maschinen diese Fähigkeiten anzueignen und auch eine größere Menge komplexer Daten zu verstehen und zu analysieren, hat sich über die letzten Jahre das Berufsbild des Data Scientists entwickelt. Dies sind Arbeitskräfte, die eine Vielzahl an Kompetenzen vereinen, um neue Probleme im Reich der Daten zu lösen. Auf die genauen Kompetenzen dieser sogenannten Data Scientists wird zu einem späteren Zeitpunkt eingegangen.

Während der Begriff Data Science schon früher in einigen Publikationen auftauchte [44], begann sich der Begriff Data Scientist als Berufsbezeichnung im Jahre 2008 zu formen. Firmen wie IBM oder HP verwendeten den Jobtitel 'Research Scientist'. Dieser Research Scientist bezeichnete aber eher Personen, die in Laboren abgetrennt vom Tagesgeschäft arbeiteten. DJ Patil und Jeff Hammerbacher suchten nach einer anderen Bezeichnung, da ihre Teams Ergebnisse lieferten, welche sofortigen und großen Einfluss auf das Unterneh-

men hatten. Daher prägten sie den Namen Data Scientist. [42] Daten spielen schon lange eine Rolle in strategischen Entscheidungen. Daher wird von DJ Patil 'decision science' als ein großer Bereich der Data Scientists beschrieben, in welchem durch das Sammeln und Auswerten Unternehmens interner und externer Daten, Ergebnisse zur Entscheidungsunterstützung geliefert wurden.
Da die Datenmenge aber stetig anwuchs und immer komplexer wurde, war eine stärkere Spezialisierung in diese Richtung erforderlich. Diese enormen Datenmengen werden auch als Big Data bezeichnet. Dabei liegen die Daten in anderen Formen als Reihen und Spalten vor. Antworten auf gestellte Fragen können nur durch Kombination vieler verschiedener Datenquellen und analytischer Ansätze gefunden werden. Durch diese Spezialisierung auf das Verstehen, Bearbeiten, Visualisieren und Analysieren großer Datenmengen, sowie das Kommunizieren der gewonnenen Erkenntnisse entstand dieses neue Berufsfeld. [27]
Die steigende Nachfrage an Data Scientists wurde durch die größten Internetunternehmen wie Google, Facebook und Amazon vorangetrieben. Diese sind so erfolgreich, weil sie es schaffen die Daten sinnvoll zu nutzen. Das heißt nicht nur ein Data Warehouse zu nutzen, um bestimmte Informationen zu erlangen, sondern die Daten verwertbar zu machen also in etwas Monetäres zu verwandeln. Wie zum Beispiel gezielte, personalisierte Werbung platzieren oder Kaufangebote an die Kunden automatisiert zu verschicken. Bei Amazon ist dies zum Beispiel die Funktion „Personen die diesen Gegenstand kauften, kauften auch“. [42]
Mittlerweile ist das Berufsfeld Data Scientist weltweit anerkannt und findet daher auch immer mehr Einzug in verschieden Programme von Universitäten. Auch wenn der Bereich in Deutschland noch in den Anfängen steckt.

3.2 Gestaltung eines lernenden Systems

Um die nötigen Kompetenzen eines Data-Scientists besser verstehen zu können, ist es sinnvoll den Aufbau und Ablauf eines Projektes darzustellen. Laut einer Umfrage auf kdnuggets war 2007 und 2014 die beliebteste Methodik für Data Science Projkete jeweils das Cross Industry Standard Process for Data Mining Modell (CRISP-DM Modell). [43] Aber auch neuere Artikel bestätigen noch den Einsatz des Modells. [57]
Die Methodik wurde entwickelt, um eine allgemeine Herangehensweise an alle Data Mining (und heute auch Machine Learning und andere Datenanalyse Aufgaben) zu entwickeln und deren Ergebnisse im Unternehmenskontext zu nutzen. Dies soll helfen solche Projekte einfacher durchführbar zu machen. Dies beinhaltet die Kosten zu senken, die Geschwindigkeit, sowie die Verlässlichkeit zu erhöhen und den ganzen Prozess zu standardisieren und damit besser steuerbar zu machen. [62]

Die sechs verschiedenen Phasen des CRISP-DM Modells sind in Abbildung 3.1 abgebildet. Im Folgenden werden die verschiedenen Phasen und deren Inhalt erläutert.

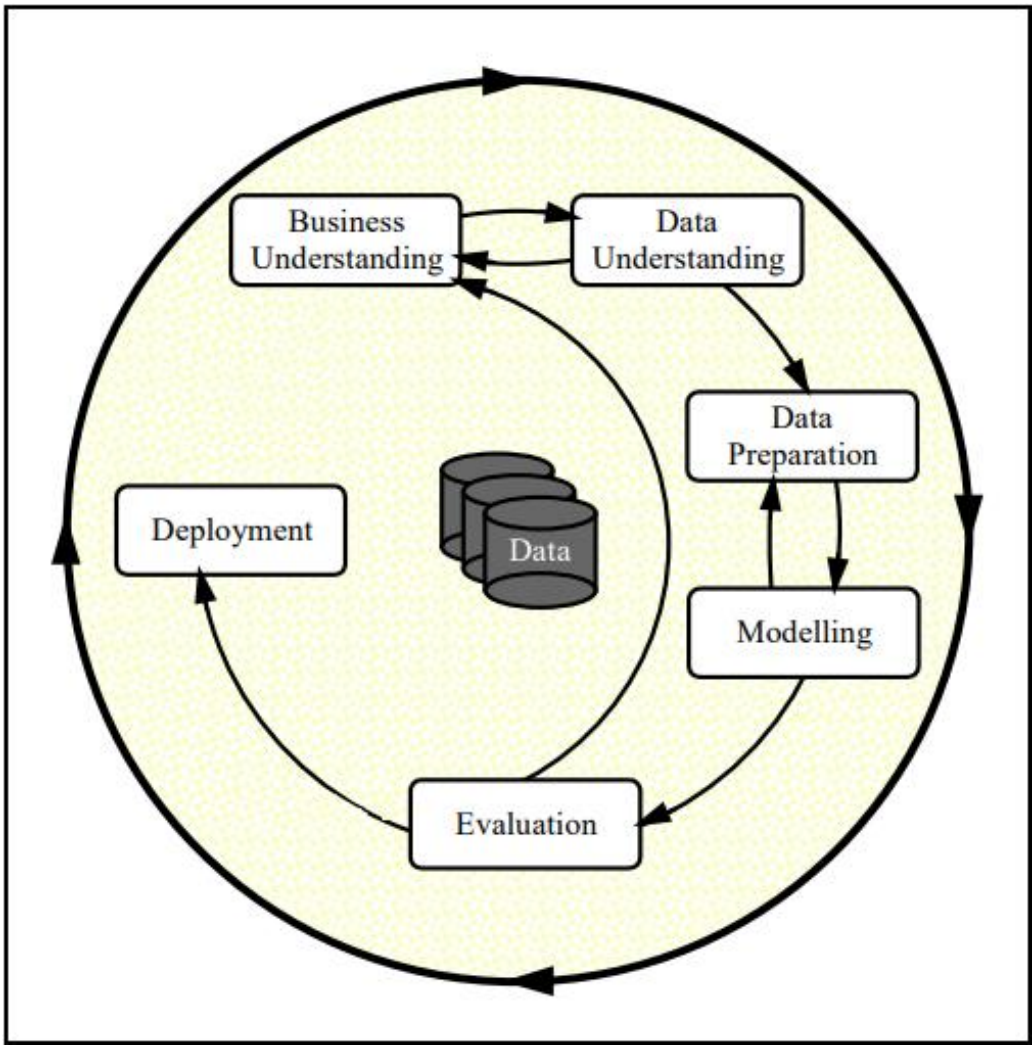

Abbildung 3.1: Die Phasen des CRISP-DM Modells [62]

In der ersten Phase 'Business Understanding' geht es darum zu verstehen, was für das Unternehmen wichtig ist und was aus Sicht des Unternehmens erreicht werden soll. Dazu gehören alle Faktoren die das Resultat des Projektes beeinflussen zu identifizieren. Dieser Schritt ist sehr wichtig, da durch ein klares Ziel verhindert werden kann, dass in die falsche Richtung gearbeitet wird.
In dieser Phase müssen die konkreten Ziele, die durch das Projekt erreicht werden sollen beschrieben werden. Es wird sich ein konkreter Plan überlegt, wie diese Ziele erreicht und gemessen werden sollen. Es sollte sich bereits Gedanken über die möglichen Verfahren die verwendet werden können gemacht werden. Zur Bewertung der Ziele, beziehungsweise zum Messen, ob das Projekt erfolgreich war, sollten sich passende Kriterien überlegt und beschrieben werden. Wie in normalen, nicht daten-getriebenen Projekten muss die

aktuelle Situation analysiert werden. Dazu gehören vorhandene Ressourcen, wie die vorhandenen Datenquellen, Personen und deren Kompetenzen, sowie vorhandene Hardware und Software. Außerdem müssen Risiken, Beschränkungen und Annahmen analysiert und beschrieben werden. Dazu gehören Fragestellungen zum Datenschutz oder der Größe der Datenmenge, aber auch andere nicht datenbezogene Risiken und Beschränkungen. [48]

In der zweiten Phase 'Data Understanding' geht es um das Verstehen der Daten. Dazu müssen in einem ersten Schritt erst einmal die vorhandenen Daten gesammelt werden. Wenn die Daten aus mehreren Quellen geladen werden, müssen Überlegungen angestellt werden, wann und wie diese zusammengeführt werden. Sind die Daten gesammelt, müssen Aussagen zu deren Format, Größe, Inhalt und anderen einfachen Eigenschaften getroffen und dokumentiert werden. Ein weiter Schritt ist es ein tieferes Verständnis der Daten zu erlangen. Dazu werden die Daten visualisiert, Datenbankabfragen erstellt oder andere Reports erstellt, die weitere Einblicke in die Daten ermöglichen. Aus diesen gewonnenen Einblicken können Einschätzungen zu der Datenqualität getroffen werden. Die Datenqualität, wie schon in Kapitel 2.1 angesprochen, beinhaltet Fragen zur Vollständigkeit und Richtigkeit der Daten. [48]

In Phase drei 'Data Preparation' werden die gesammelten Daten selektiert und aufbereitet. Der Prozess der Datenselektion und -aufbereitung wurde auch schon in Kapitel 2.1 genauer beschrieben. Bei der Selektion geht es um das Bestimmen der Daten, sowie der Attribute, die für die Modellbildung verwendet werden sollen. Beim Aufbereiten der Daten, wird versucht die Datenqualität auf ein verwendbares Level anzuheben. Ebenfalls zur Datenaufbereitung gehört der Prozess der Datenintegration, bei dem Daten aus unterschiedlichen Quellen und Formaten zusammengeführt werden. [62]

Die vierte Phase 'Modelling' beschäftigt sich mit der eigentlichen Modellbildung. Hier werden zu Beginn die Techniken beziehungsweise Algorithmen ausgesucht, die verwendet werden sollen oder die in der ersten Phase identifizierten verifiziert. Für die ausgesuchten Techniken werden anschließend die folgenden Schritte durchlaufen. Der ausgewählte Algorithmus wird mit Hilfe der Trainingsdaten trainiert und durch sogenannte Testdaten verifiziert. Dazu werden die vorhandenen Daten in Trainings- und Testdaten aufgeteilt. Für die Aufteilung gibt es keine genauen Vorgaben, aber ein beliebtes Verfahren hierbei ist Cross-Validation. Dieses Aufteilen hat den Hintergrund, das trainierte Modell mit neuen Daten zu überprüfen. So kann verhindert werden, dass der trainierte Algorithmus eine hohe Performance auf den Trainingsdaten liefert, aber zu schlecht generalisiert ist, um diese Performance auch bei neuen Daten zu erreichen. Anschließend kann der

trainierte Algorithmus oder die trainierten Algorithmen auf die aufbereiteten Daten angewandt werden, um ein oder mehrere gewünschte Modelle zu bilden. Diese Modelle können nun beschrieben und interpretiert werden. Durch diese Beschreibung und Interpretation kann nun der Erfolg des Modells, anhand der definierten Kriterien bestimmt werden. Bei schlechten Ergebnissen können Anpassungen hinsichtlich der Attribute vorgenommen werden und eine erneute Modellbildung erfolgen. Danach können die Resultate hinsichtlich der Verwendung im Unternehmenskotext besprochen werden. [48]

In der fünften Phase 'Evaluation' werden die Ergebnisse ausgewertet. Während in der vorherigen Phase die Performance des Modells oder der Modelle gemessen wurde, wird in dieser Phase die Anwendbarkeit der Ergebnisse hinsichtlich der Unternehmensziele bewertet. Das Modell kann zum Beispiel gute Ergebnisse liefern, die aber keinen verwertbaren Nutzen für das Unternehmen mit sich bringen. Es können aber auch ein Nutzen entdeckt werden, der anfänglich nicht geplant war.
Im nächsten Schritt kann der komplette Prozess bis zu diesem Punkt nochmal betrachtet und bewertet werden. Hier können Punkte identifiziert werden, die vergessen wurden oder wiederholt werden müssen. Resultierend daraus können die nächsten Schritte bestimmt werden, ob beispielsweise ein weiteres Projekt angestoßen werden soll, die identifizierten Punkte wiederholt werden sollen oder mit der Implementierung der Ergebnisse begonnen werden soll. [48]

In der letzten Phase 'Deployment' geht es um die Implementierung der Ergebnisse des Projektes. Dazu muss eine Strategie entwickelt werden, wie genau das gewonnene Wissen dazu eingesetzt werden kann, monetären Nutzen zu erzielen. Danach kann ein offizieller Abschlussbericht erstellt werden und die gewonnenen Erfahrungen und Erkenntnisse dokumentiert werden. [62]

3.3 Kompetenzprofil

Zu Beginn dieses Kapitels wurde das Entstehen des Berufsbildes Data Scientists erläutert. Im zweiten Unterkapitel wurde eine übliche Methodik vorgestellt, wie Projekte ablaufen können, in denen Data Scientsits arbeiten. Dieses Kapitel soll einen Überblick über die verschiedenen Kompetenzen, die im Tätigkeitsfeldes eines Data Scientists benötigt werden, darstellen. Da die Kompetenzliste der Data Scientists scheinbar endlos erscheint, wird sich hier in Teilen wieder auf das CRISP-DM Modell bezogen, um den Kompetenzen einen Praxisbezug zuzuordnen. Dazu wird Wissen aus Büchern verwendet, die schon durch Literatur- und Internetrecherche die Kompetenzen, die von Unterneh-

men gefordert werden oder von Hochschulen in ihrem Programmen genannt werden, zusammengetragen haben. [37] [46] [51] [21] Ein Abgleich wurde mit aktuellen Stellenanzeigen bei Facebook, Amazon und bei einer deutschen Jobbörse vorgenommen. [2] [20] [53]

Zu Beginn eines Projektes geht es erst einmal darum zu identifizieren, welche Projekte sich für das Unternehmen lohnen. Dazu müssen die Data Scientists gutes unternehmerisches Verständnis mitbringen. Sie müssen nicht nur Wissen welche Probleme sich durch Anwendung ihrer Methoden lösen lassen, sondern auch wann das Lösen eines Problems Mehrwert für das Unternehmen bietet. Sie müssen wissen was sowohl softwaretechnisch als auch hardwaretechnisch möglich ist. Dazu gehört zum Beispiel Wissen darüber welche Daten sich nutzen lassen und auch welche Daten überhaupt verwendet werden dürfen. Das involviert ein grobes Verständnis über Datenschutzaspekte und dazugehörige Gesetze. Da es sich hierbei um ein Projekt handelt, sollten die Data Scientists auch Wissen in Projektmanagement mitbringen.
In den nächsten Schritten geht es um das Beschaffen der Daten und um das Bearbeiten dieser. Hier müssen die Data Scientists Wissen über die verschiedenen Speicherorte der Daten besitzen und wie unterschiedliche Datenquellen zusammengeführt und genutzt werden können. Hierzu sind diverse Datenbankabfragen notwendig. Ebenso wichtig ist es, die zusammengeführten Daten zu Bearbeiten, sodass diese verwendbar sind. Sie sollten darüber hinaus auch Verständnis über die Daten haben, mit denen sie arbeiten. Das erfordert Wissen aus diversen Fachbereichen und das Verstehen der Zusammenhänge im Unternehmensprozess. Auch sollten Tools zur Visualisierung der Daten beherrscht werden und auch die Fähigkeit die Daten verständlich abzubilden.
Sind die Daten aufbereitet, geht es um die Generierung von Wissen aus den Daten. Hierzu müssen die Data Scientists diverse Tools beherrschen, um die Daten zu analysieren. Dies beinhaltet die in dieser Arbeit behandelten Verfahren des Machine Learnings, aber auch weitere Techniken wie Simulationen oder Techniken des Data Minings. Da immer wieder neue Tools und Methoden veröffentlicht werden, ist es sehr wichtig, dass der Data Scientist immer auf dem neusten Stand bleibt. Auch Programmierfähigkeiten müssen Data Scientists mitbringen. Diese werden benötigt um eigene Tools zu entwickeln der bestehende anzupassen. Dazu ist auch tiefgehendes Wissen in den Bereichen der Mathematik und der Statistik notwendig, da die Modelle auf mathematischen und statistischen Berechnungen beruhen.
Werden die Resultate ausgewertet, ist es wichtig, dass der Data Scientist die Ergebnisse in einen monetären Unternehmensnutzen umwandelt. Dazu ist neben dem unternehmerischen Verständnis auch der Wille wichtig, Vorgabewerte des Managements zu erreichen. Des Weiteren muss ein Data Scientist gute soziale und kommunikative Kompetenzen mit-

bringen. Zum einen arbeiten Data Scientists im Team. Zum anderen müssen die Ergebnisse der Analysen auch den Entscheidungsträgern vermittelt werden. Dazu müssen Präsentationen erstellt werden und die Erkenntnisse in verständlicher Sprache vorgetragen werden. Das heißt, sie müssen die Fähigkeit besitzen sich einfach auszudrücken und die Ergebnisse auch nicht Experten zu erklären.
Dann gibt es noch allgemeine Fähigkeiten, welche die Data Scientists besitzen sollten. Dazu gehören zum einen Neugierde. Sie wollen Unternehmensprobleme lösen und sich Wissen aus unterschiedlichen Bereichen aneignen. Weiter brauchen sie gute analytische Fähigkeiten und Kreativität. Sie müssen Wissen aus den unterschiedlichen Bereichen kombinieren, um Problemen aus unterschiedlichen Sichtweisen zu betrachten. [42]
Eine gute Übersicht der genannte Kompetenzen bietet Abbildung 3.2. Die Kompetenzen sind hier nach fünf verschiedenen Bereichen gegliedert. Des Weiteren gibt es fünf allgemeinere Fähigkeiten die ein Data Scientist besitzen sollte. Unter den Punkt 'Techonolgy' fallen die genannten Punkte der Prgrammierfähigkeiten, des hardwaretechnischen Verständnisses sowie der Aspekt der Datensicherheit. 'Analytics' beschreibt die Fähigkeiten, die benötigt werden, um aus den Daten Wissen zu generieren. Unter Data Management werden die Fähigkeiten für die Datengewinnung, Datenbearbeitung und Datenspeicherung genannt. Die Visualisierung der Daten und die Fähigkeit die Ergebnisse auch zu präsentieren und zu kommunizieren sind hier unter 'Art & Design' zusammengefasst. Der letzte Bereich des 'Entrepreneurship' bezeichnet die Fähigkeiten, die Sichtweise eines Unternehmens und dessen Teilbereiche zu verstehen und auch zu verstehen, welche Tätigkeiten dem Unternehmen Mehrwert bieten. Ein Data Scientist sollte darüber hinaus auch mit Kreativität, Neugierde und einer wissenschaftlichen Herangehensweise Probleme lösen. Dazu sollte er immer den Nutzen für das Unternehmen im Blick halten. Ein Data Scientist alleine kann kaum alle Kompetenzen abdecken. Es müssen auf das Unternehmen abgestimmte Typen von Data Scientists mit unterschiedlichen Kompetenzschwerpunkten eingesetzt werden. Dazu sollte analysiert werden, welche spezifischen Kompetenzanforderungen für die gewünschte Tätigkeit gefordert sind. So kann bestimmt werden wie die Stärke der Ausprägungen der einzelnen Bereiche bei dem gesuchten Data Scientist sein muss. Eine Kompetenz die alle Data Scientist besitzen müssen, ist der Umgang mit Daten. [42]

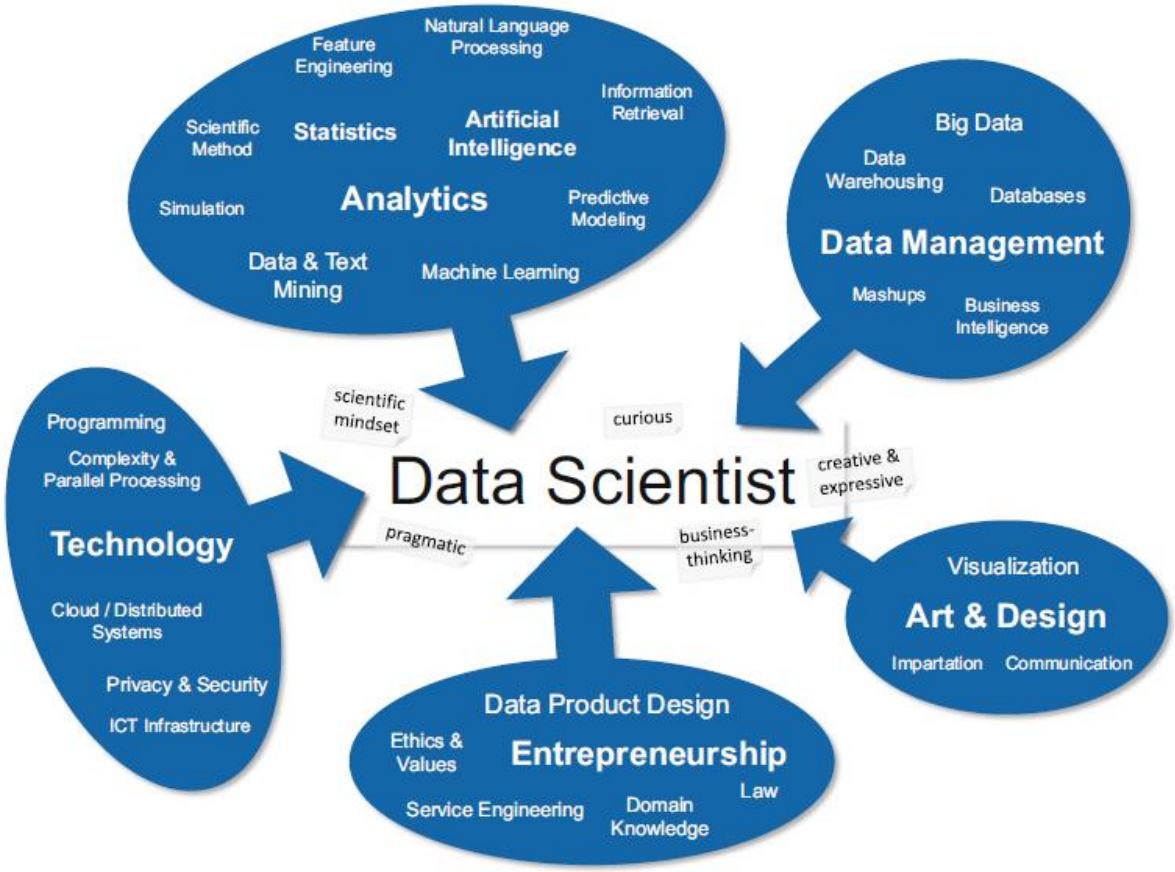

Abbildung 3.2: Data Scientist Kompetenzübersicht [21]

3.4 Herausforderungen für Unternehmen

Machine Learning Verfahren werden in Amerika bereits in vielen Bereichen verwendet. Das erklärt auch den akuten Fachkräftemangel von Data Scientists der dort herrscht. In Deutschland stecken die Anwendungen und auch das Know How noch in einer frühen Entwicklungsphase. Eine Studie von McKinsey bestätig, dass amerikanische Unternehmen mehr als doppelt so viel Geld in Digitalisierung investieren als deutsche. [34] Es gibt einige Herausforderungen, welchen sich die Unternehmen stellen müssen. Diese Herausforderungen, aber auch die Chancen die sich bieten, sollen in diesem Kapitel aufgezeigt werden.

Das wohl größte Problem das besteht ist schlechte Datenqualität. Daten sind der Treibstoff aller Machine Learning Anwendungen, deswegen wird diesen so einer hoher Wert zugesprochen. Inwiefern schlechte Daten die Performance von Machine Learning Anwendungen beeinflussen wurde schon in Kapitel 2.4.5 beschrieben. Oft ist in Unternehmen zwar bekannt, dass eine mangelnde Datenqualität im Unternehmen herrscht, allerdngs ist

der Einfluss dieser schlechten Daten nicht bekannt. Eine Untersuchung mit Hilfe der Friday Afternoon Measurement Method in Irland zeigte, dass nur etwa 3% aller untersuchten Datensätze in einem akzeptablen Bereich lagen. Dazu wurden 75 Manager angewiesen 100 verschiedene Datensätze in ihrem Unternehmen zu betrachten und die Fehler hervorzuheben. Der akzeptable Bereich wurde von allen Beteiligten ab 97 korrekten Datensätzen festgelegt. Weiter wird davon ausgegangen, dass unter der Berücksichtigung der 'rule of ten' inkorrekte Datensätze die Kosten eines Arbeitspaketes verzehnfachen. [38] Gartner, eins der führenden Beratungs- und Forschungsunternehmen aus Amerika, schätzt die Kosten mangelnder Datenqualität ebenfalls sehr hoch. Schlechte Datenqualität soll ein Unternehmen im Schnitt pro Jahr 13,5 Millionen Dollar kosten. Als Hauptproblem wird hier kein einheitliches Verständnis, was der Begriff einer guten Datenqualität bedeutet, gesehen. So entstehen widersprüchliche Definitionen von selben Daten. [22] Eine weitere Zahl liefert IBM. IBM schätzt, dass schlechte Datenqualität den gesamten US Markt 3,1 Billionen $ im Jahr 2016 kostete. Weitere Gründe warum schlechte Datenqualität so hohe Kosten mit sich bringt, ist die verschwendete Zeit die von Managern, Entscheidungsträgern oder anderen qualifizierten Arbeitskräften aufgebracht werden muss, um die korrekte Datensätze zu bekommen. Oft werden die Datensätze in eigener Hand korrigiert, da Zeitdruck herrscht. Es wird dadurch vermieden dem wahren Grund der schlechten Datenqualität auf den Grund zu gehen. Korrigieren mangelnder Daten wird bereits als Alltagsgeschäft angesehen und akzeptiert. Und auch durch das Korrigieren werden nicht alle Fehler behoben, was wiederum kostspielige Folgen hat. [13]

Daher sollte bevor irgendein Machine Learning Projekt angestoßen wird, zu aller erst überprüft werden, ob ausreichend Daten in geeigneter Qualität vorliegen. Es sollten sich auch Gedanken darüber gemacht werden, wie diese Datenqualitätsprobleme angegangen werden. Ein erster Ansatz ist es eine direkte Kommunikation zwischen den Datenerstellern und den Datenkonsumenten sicherzustellen. Die Datenersteller wissen in vielen Fällen gar nicht welche Folgen eine schlechte Datenqualität bei den Datenkonsumenten mit sich bringt oder was die genauen Anforderungen an die Daten sind, also wie die Verwendung der Daten aussieht. Durch direkte Kommunikation kann direktes Feedback gegeben werden und Wege zur Verbesserung eingeleitet werden. Des Weiteren sind die Datenkonsumenten oft weniger qualifiziert die Daten zu korrigieren. Es sollte das Überprüfen der Entstehungswege der Daten sichergestelllt werden. Dabei sollte darauf geachtet werden, dass korrekte Daten erzeugt werden. [12]

Ein weiter wichtiger Aspekt ist ein allgemeines Verständnis für die Daten zu schaffen. Die Mitarbeiter müssen die Wichtigkeit und den Nutzen guter Daten wissen. DJ Patil geht sogar soweit, dass jeder in einem Unternehmen soviel Zugriff auf so viele Daten wie möglich haben sollte. Nur so können die Mitarbeiter diese auch effizient nutzen und neue Ideen zur

Verwertung der Daten einbringen. [42] Diese Meinung eines zentralen Data Warehouses, auf das jeder Zugriff hat, wird ebenfalls von Andrew Ng, einem der führenden Köpfe im Bereich Machine Learning und Artificial Intelligence, vertreten. [39] Als Gegenargument kann hier die Datensicherheit genannt werden. Umso mehr Leute auf die Daten Zugriff haben, umso leichter können diese nach außen gelangen. Allerdings reicht es nicht nur eine große Menge an Daten zu haben und diese für jeden zugänglich zu machen. Schlussendlich kommt es darauf an, die vorhandenen Daten auch sinnvoll zu nutzen. Dazu bieten Machine Learning Verfahren vielseitige Anwendungsmöglichkeiten, beispielsweise autonome Fahrzeuge in der Logistik, Qualitätskontrolle, Predictive Maintenance, optimiertes Marketing oder optimierte Produktpreisbestimmung.
Diese Anwendngsmöglichkeiten und deren Vorteile werden immer wieder in Forschungsberichten aufgezeigt. Allerdings ergab eine Befragung von der Staufen AG, dass sich nur 30% der Unternehmen, von den Unternehmen die sich bereits mit Industrie 4.0 beschäftigen auch mit Machine Learning auseinander setzen. [26] Die Staufen AG sieht diese Zahl sogar als hoch an. Zu einem anderen Ergebnis kommt eine Studie von Crisp Research. Hier beschäftigen sich 64% mit der Thematik Machine Learning und 34% setzen Machine Learning bereits in ausgewählten Bereichen im Unternehmen ein. [7] Klar ist auf jeden Fall, dass Machine Learning bereits teilweise Einzug in deutsche Firmen oder zumindest in die Köpfe der Entscheidungsträger gefunden hat.
In Deutschland ist allerdings das Know How in diesen Bereichen noch nicht so verfügbar. Mit ein Grund dafür ist, dass es in Deutschland kaum Universitäten gibt, die Studiengänge im Bereich Data Science, Machine Learning oder Big Data anbieten. Dadurch sind Kosten für Beratungen oder das Einstellen von Experten entsprechend hoch. Deswegen ist es umso wichtiger, dass Entscheidungsträger die Einsatzmöglichkeiten und auch den Nutzen beziehungsweise das Einsparungspotential kennen. Auch sind beispielsweise Spracherkennungssysteme in der deutschen Sprache noch nicht so gut entwickelt wie in der englischen Sprache. Auch die bereits vorhandenen Lösungen bei der Predictive Maintenance überzeugt laut der Befragung der Staufen AG 74% der Unternehmen noch nicht. Bei Predictive Maintenance werden Maschinendaten in Echtzeit ausgewertet um Vorhersagen darüber zu machen, wann Teile der Maschine ausfallen, sodass proaktiv gehandelt werden kann.
Durch diese noch am Anfang stehende Entwicklung ergeben sich aber eben auch Chancen für die Unternehmen sich Wettbewerbsvorteile zu erarbeiten. Heute werden durch die großen Internetfirmen wie Google oder Amazon schon gute Cloud Lösungen zu zahlbaren Preisen angeboten. Dadurch sind die Hürden für erste Implementierungen von Machine Learning Projekten um einiges niedriger als noch vor einigen Jahren. Crisp Research empfiehlt beispielsweise mit Cloud Provider zu starten, aber sich nach und nach eigene Systeme aufzubauen. [7] Allerdings zeigt die Studie von McKinsey, dass

81% der befragten Unternehmen ihre Systeme allerdings nur innerhalb Deutschlands outsorcen würden. Als Grund dafür wird die Datensicherheit genannt. [34] Hier müssen Unternehmen das Risiko abwiegen. Die Unternehmen sollten auf jeden Fall einen eigenen Geschäftsbereich durch Zukauf von Experten aufbauen. Der wichtigste Schritt, wie schon angedeutet wurde, ist es erstmal die Unternehmensbereiche zu identifizieren, bei denen eine schnelle Implementierung mit vorhandenen Daten möglich ist. Durch die Experten lassen sich Projekte innerhalb der Firma identifizieren und umsetzen, wodurch der interne Geschäftsbereich weiter ausgebaut und Erfahrung gesammelt werden kann. Dieser zentrale Geschäftsbereich kann nach und nach in die Prozesse der anderen Geschäftsbereiche eingegliedert werden. Auch die Schulung der vorhandenen Mitarbeiter ist dabei wichtig. Hier gibt es zahlreiche Möglichkeiten sich online Wissen anzueignen. [40]

Machine Learning kann auch dazu verwendet werden Produkte zu verbessern. In Zukunft werden Softwarelösungen um die Produkte herum immer wichtiger. Bei beispielsweise Fertigungsmaschinen kommt es neben der Einsatzmöglichkeit und der Fertigungstechnologie eben auch darauf an, wie gut die dazugehörige Software Fehler oder verschleißbedingte Ausfälle im Vorhinein voraussagen kann. Das Ziel ist es hier Produkte zu entwickeln bei denen der Kunde hilft Daten zu generieren. Diese Daten können dann wiederum dazu verwendet werden, die verwendeten Algorithmen zu verbessern, sodass mehr Kunden das Produkt kaufen und somit noch mehr Daten erzeugen. Hier müssen allerdings Themen des Datenschutzes beachtet werden, wozu es in Deutschland bislang Andrew Ng erklärt, dass die neusten Machine Learning Algorithmen, also die DNN, sehr stark über die Menge der Daten skalieren, im Vergleich zu herkömmlichen Machine Learning Algorithmen. [39] In Abbildung 3.3 ist dieser Sachverhalt abgebildet. Hier bestehen große Chancen für Unternehmen sich Wettbewerbsvorteile zu sichern. Schafft ein Unternehmen sich eine Datenvorherrschaft in einem Bereich aufzubauen, ist es für Wettbewerber sehr schwer bis unmöglich Systeme oder Produkte mit gleicher Performance zu entwickeln. Hier haben natürlich große Unternehmen Vorteile. So wird auch davon ausgegangen, dass die meisten Machine Learning Anwendungen über die Cloud erfolgen, da die großen Tech Unternehmen aus Amerika aber auch aus China schon sehr gut trainierte Systeme und rieseige Rechenanlagen besitzen, um diese zur Verfügung stellen. [11] Machine Learning Verfahren können aber auch bei kleinen und mittelständischen Unternehmen, bei denen keine Datenvorherrschaft aufgebaut werden kann, zu Verbesserungen führen. Das wichtigste ist die vorhandenen Daten effizient zu nutzen. [40] [42] [9]

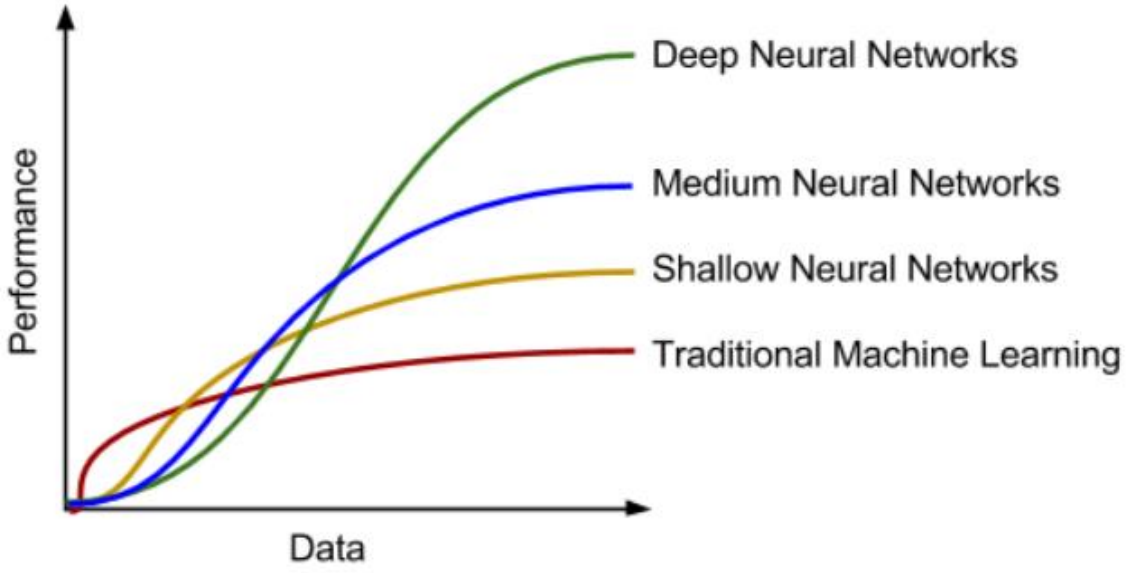

Abbildung 3.3: Skalierbarkeit der Algorithmen mit Daten [4]

Ein weiterer wichtiger Punkt um langfristig Wettbewerbsvorteile zu erlangen oder zumindest nicht an Wettbewerbsfähigkeit zu verlieren ist es, sich nicht nur auf die Machine Learning Verfahren und einzelne Anwendungsbereiche zu konzentrieren, sondern auch auf interne Prozesse und wie Machine Learning in diese Prozesse eingegliedert werden kann. Machine Leraning darf nicht als Forschungsobjekt betrachtet, sondern sollte als integrierte Businessfunktion gesehen werden. Machine Learning ist damit nicht als Quick Win zu betrachten, sondern als fundamentaler Baustein für eine langfristige Strategie. [7] [50]
Hier wird von vielen Seiten auch die aktuell vorwiegend herrschende Unternehmens- und Führungskultur als Problem angesehen. Kritikpunkte sind zu starre Hierarchien, zu lange Kommunikationswege, restriktive Datenkultur und auch das fehlende Delegieren der Verantwortung nach unten. Unternehmen sollten verstärkt agile Methoden einsetzen und eine Organisationsstruktur aufbauen, die Inoovationen erleichtern und fördert. [49] [26] [7]

3.5 Aussichten

Data Science und Machine Learning hat in den letzten Jahren einen extremen Aufschwung erlebt und ist ein noch junger und agiler Bereich. Daher gibt es in diesem Feld ständig Veränderungen und neue Entdeckungen. In diesem Kapitel werden aktuelle Trends und Themen diskutiert und verschiedene Ansichten dazu dargelegt.

Ein Report von CrowdFlower zeigt, dass Data Scientits aktuell die meiste Zeit damit

verbringen, Daten zu sammeln, zu bereinigen, zu labeln und zu strukturieren. Dafür verwendet ein Data Scientists durchschnittlich 53% seiner Zeit. [15] In Blogbeiträgen wird sogar von Prozentzahlen bis zu 80% gesprochen. Da die Datenmenge und die Geschwindigkeit mit der die Daten erzeugt werden ständig weiter anwächst, haben sich immer mehr Unternehmen und Startups mit der Fragen beschäftigt, wie sich diese Prozesse automatisieren lassen. Ein großer Teil der Arbeit eines Data Scientists wird als banal angesehen. Das zusammenführen verschiedener Datenquellen, das Bereinigen dieser Daten und auch das Testen verschiedener Modelle, nimmt viel Zeit in Anspruch, aber erfordert keine große Denkleistung und kein tiefgehendes Wissen in der Data Science. Durch die Automatisierung sich ständig wiederholender Aufgaben mit hohem manuellem Aufwand, wird für Data Scientist mehr Zeit für andere Aufgabenbereiche frei, bei denen ihre Expertise wirklich benötigt wird. Die Frage die sich hier stellt ist, wie sich der Beruf des Data Scientists in Zukunft entwickelt wenn Aufgaben, welche die meiste Arbeitszeit der Data Scientists in Anspruch nehmen, automatisiert werden. Gartner sagt voraus, dass bereits 2020 40% der Aufgaben eines Data Scientists automatisiert sein werden. Dabei wird sich ebenfalls auf die Aufgaben der Datenintegration und der Modellbildung bezogen. [23] Auf diese Weise soll der Mangel an Fachkräften kleiner werden, da erstens nicht mehr so viele Data Scientsts benötigt werden, aber auch die Kompetenzen, welche ein Data Scientist besitzen muss geringer werden. Eine Liste zu verschiedenen Datenbearbeitungs-Tools gibt es auf der Homepage von Gartner ebenfalls. Neben den Datenbearbeitung-Tools wie beispielsweise von Trifacata oder IBM gibt es auch Firmen, welche sich auf die Modellbildung spezialisiert haben, wie beispielsweise DataRobot. DataRobot beschreibt Data Science als einen Zusammenschluss von den Bereichen Mathematik, Programmieren und Wissen aus den Geschäftsbereichen. Ihre Software soll die Bereiche Mathematik und Programmieren übernehmen. Es müssen nur die richtigen Daten eingelesen werden. Das Programm findet dann zu den gegebenen Daten das Machine Learning Modell, welches am besten für die Daten geeignet ist, trainiert es und gibt die Ergebnisse aus. Auch Amazon und Google haben Ende 2017 neue Cloud Lösungen auf den Markt gebracht. Dabei soll Amazons SageMaker Entwicklern die Möglichkeit bieten, einfach auf verschiedene Datenquellen zuzugreifen und die Daten zu visualisieren. Weiter können schnell verschiedene Algorithmen trainiert und direkt in einer „produktionsfertigen Umgebung“ angewendet werden. Dabei können schon vorgefertigte Algorithmen verwendet werden, aber auch selbst entwickelte. Googles äquivalente Software nennt sich AutoML.

Durch diese Automatisierung werden auch die in Kapitel 3.3 besprochenen Kompetenzen der Data Scientists anders gewichtet. In unterschiedlichen Beiträgen über die Automatisierung des Data Scientists, werden immer wieder die gleichen Aufgaben genannt, welche sich (noch) nicht automatisieren lassen und damit stark an Wichtigkeit

gewinnen. [47] [35] [5] Eine der Hauptaufgabe der Data Scientists ist es zu aller erst Unternehmensprobleme ausfindig zu machen, die sich mit Hilfe von Machine Learning und andern Datenanalyseverfahren lösen lassen und zu ermitteln, welcher Effekt damit erzielt werden kann. Also es geht darum die richtigen Fragen zu stellen. Werden automatisierte Tools zu Datenbearbeitung und Modellbildung verwendet, bedarf es immer noch der Kontrolle eines erfahrenen Data Scientists die Ergebnisse zu validieren, bevor damit wichtige Entscheidungen getroffen werden. Auch müssen die Ergebnisse Entscheidungsträgern verständlich präsentiert und schlussendlich implementiert werden. Sogar DataRobot schreibt in einem Blogeintrag auf ihrer Homepage, dass trotz der ganzen Automatisierung weiterhin Data Scientists benötigt werden, die sich dank der Automatisierung mit wichtigeren Dingen beschäftigen können. [18] Ein weiterer Bereich der laut William Vorhies, einem Data Scientist mit 15 Jahren Erfahrung, noch nicht gut automatisert werden kann, ist der Bereich des Feature Engineerings, bei dem geeignete Attribute aus den Daten ausgewählt werden, welche die betrachtete Variable beeinflussen. [58] Allerdings sind auch hier schon genügend Programme auf dem Markt, die ständig weiterentwickelt werden. Ein Report, bei dem auch IBM mitwirkte, berichtet von einem weiteren Anstieg von Data Scientists bis 2020. [32] Das zeigt, dass trotz der ganzen Automatisierung immer noch Experten gebraucht werden, um die Projekte zu überblicken und die Machine Learning Anwendungen in die Unternehmensprozesse einzugliedern.

Durch die Automatisierung von großen Kompetenzbereichen der Data Scientists wurde in den letzten Jahren aber auch zunehmend der Begriff des Citizen Data Scientist geprägt. Die erste richtige und laut vieler Blogeinträgen auch in der Community anerkannte Definition dieses Citizen Data Scientists liefert Gartner. „Ein Citizen Data Scientist ist eine Person, die Modelle erstellt oder generiert, die fortgeschrittene diagnostische Analysen oder prädiktive und präskriptive Fähigkeiten verwenden, deren Hauptaufgaben jedoch außerhalb des Bereichs Statistik und Analytik liegt.“ [23] Das heißt Citizen Data Scientists können qualitativ hochwertige Analysen durchführen, ohne die Fähigkeiten zu besitzen, die einen Data Scientist ausmachen. Gartner meint in seinem Report außerdem, dass bereits 2019 mehr Analysen von Citizen Data Scientists erstellt werden als von „richtigen“Data Scientists. Ein Grund für den Anstieg dieser Data Scientists sind die Möglichkeiten sich online Wissen anzueignen. Des Weiteren ist für viele Firmen das Einstellen von mehreren Data Scientists zu teuer. Einer der neuen Kernkompetenzen der Data Scientists ist das Verständnis der Daten mit denen gearbeitet wird. Durch die Vereinfachung der Tools kann es einfacher sein Fachleuten, welche bereits das nötige Wissen über die Daten besitzen, die notwendigen Tools und das Data Science Wissen anzueignen, als teure Data Scientsits einzustellen.

IBM veröffentlicht auf ihrer Homepage sogar einen Bericht zu einer 'Anleitung' wie ein Unternehmen Citizen Data Scientists integrieren und erziehen kann. Auf diese Weise versucht auch IBM den Mangel an Data Scientists zu verringern. Dabei sollen aber auf keinen Fall die Data Scientists ersetzt werden. Im ersten Schritt soll die Kommunikation während Entwicklungsprozessen zwischen den IT-Abteilungen und den Business-Abteilungen gestärkt werden. Dabei sollen alle Abteilungen Zugriff auf verschiedene Tools haben, um die Möglichkeit zu bieten, das Wissen in Data Science zu vertiefen. In Verbindung damit sollte auch die Möglichkeit gegeben werden, das Data Science Wissen anzuwenden, auch wenn zu Beginn noch kein großer Beitrag einzelner Personen geleistet wird. Am Anfang ist darauf zu achten, dass einfache Tools zur Verfügung gestellt werden. Neben den Citizen Data Scientists, welche die Analysen durchführen, ist es auch wichtig Leute im Unternehmen zu haben, welche die neu gewonnenen Informationen implementieren. Dazu ist auch die Zusammenarbeit dieser Personen zu regeln, genauso wie eine Strategie zu entwickeln, wie vielversprechendes Wissen implementiert werden soll. Es sollte aber bedacht werden, dass die Entwicklung der Citizen Data Scientists Zeit in Anspruch nimmt und für die Erstellung guter Analysen Erfahrung notwendig ist. [17] Sinnvoll kann es hier sein, einen erfahrenen Data Scientist als übergeordnete Instanz einzustellen, die bei der Entwicklung der Citizen Data Scientists hilft.

Bei einer Expertendiskussion von The Cube unter dem Stichpunkt 'IBM Data Science for all' wird die Meinung vertreten, dass ein großer Bereich der Data Scientist weiterhin die Forschung sein wird. Es müssen auch die Tools zur Automatisierung der Aufgaben entworfen und entwickelt werden. Data Scientists entwerfen die Tools, welche von den Anwerndern, also den Citizen Data Scientists verwendet werden. Hier kommt auch der eigentliche wissenschafltiche Bereich, der schon im Namen steckt zum Vorschein. Es müssen immer weitere Wege entworfen werden, wie die Daten erforscht und ausgewertet werden können. Weiter wird gesagt, dass es bei diesem Hype um den Citizen Data Scientist mehr darum geht, der Belegschaft die Wichtigkeit der Daten klar zu machen und die Organisation in Richtung Data-Driven zu steuern. Die Enduser der automatisierten Tools sollen trotzdem Grundkenntnisse in der Statistik haben und sie sollten wissen, was die Algorithmen machen, um die Ergebnisse richtig einschätzen zu können. Nichtdestotrotz werden hier große Chancen gesehen. Wenn die Tools immer weiter vereinfacht werden und immer mehr Leute sich durch die unzähligen online Ressourcen weiterbilden, gibt es bald eine noch viel rasantere Entwicklung als wir es bisher gesehen haben. [28]

Literaturverzeichnis

[1] C. C. Aggarwal. *Data mining: The textbook*. Springer, Cham, 2015.

[2] Amazon. Jobs. https://www.amazon.jobs/de/search?base_query=data+scientist&loc_query=, 2018. Letzte Einsichtnahme: 28.04.2018.

[3] M. Awad and R. Khanna. *Efficient learning machines: Theories, concepts, and applications for engineers and system Designers*. The expert's voice in machine learning. Apress Open, New York, 2015.

[4] A. C. Bahnsen. Building ai applications using deep learning. http://blog.easysol.net/building-ai-applications/, 2017. Letzte Einsichtnahme: 27.04.2018.

[5] J. Berkman. Automated machine learning won't replace data scientists. https://www.datascience.com/blog/automated-machine-learning-wont-replace-data-scientists, 2017. Letzte Einsichtnahme: 19.04.2018.

[6] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba. End to end learning for self-driving cars. https://arxiv.org/pdf/1604.07316.pdf, 2016. Letzte Einsichtnahme: 12.04.2018.

[7] B. Böttcher, D. Klemm, and C. Velten. Machine learning im unternehmenseinsatz: Künstliche intelligenz als grundlage digitaler transformationsprozesse. https://www.unbelievable-machine.com/downloads/studie-machine-learning.pdf, 2017. Letzte Einsichtnahme: 24.04.2018.

[8] M. Brouillette. Deep learning is a black box, but health care won't mind. https://www.technologyreview.com/s/604271/deep-learning-is-a-black-box-but-health-care-wont-mind/, 2017. Letzte Einsichtnahme: 14.04.2018.

[9] E. Brynjolfsson and A. Mcafee. The business of artificial intelligence. https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence, 2017. Letzte Einsichtnahme: 28.04.2018.

[10] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf, 2018. Letzte Einsichtnahme: 13.04.2018.

[11] P. Burrows. How the ai cloud could produce the richest companies ever. https://www.technologyreview.com/s/610554/how-the-ai-cloud-could-produce-the-richest-companies-ever/?utm_campaign=Artificial%2BIntelligence%2BWeekly&utm_medium=email&utm_source=Artificial_Intelligence_Weekly_77, 2018. Letzte Einsichtnahme: 14.04.2018.

[12] T. C. Redman. Data's credibility problem. https://hbr.org/2013/12/datas-credibility-problem, 2013. Letzte Einsichtnahme: 22.04.2018.

[13] T. C. Redman. Bad data costs the u.s. $3 trillion per year. https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year, 2016. Letzte Einsichtnahme: 20.04.2018.

[14] T. C. Redman. If your data is bad, your machine learning tools are useless. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless%20https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards, 2018. Letzte Einsichtnahme: 13.04.2018.

[15] CrowdFlower. 2017 data scientist report. https://visit.crowdflower.com/WC-2017-Data-Science-Report_LP.html, 2017. Letzte Einsichtnahme: 20.04.2018.

[16] DeepL. https://www.deepl.com/home. Letzte Einsichtnahme: 06.04.2018.

[17] M. Denham. Grow your own citizen data scientists with these 5 tips. https://www.ibm.com/information-technology/grow-your-own-citizen-data-scientists-with-these-5-tips, 2017. Letzte Einsichtnahme: 28.04.2018.

[18] T. Dinsmore. Automated machine learning: A short history. https://blog.datarobot.com/automated-machine-learning-short-history, 2016. Letzte Einsichtnahme: 19.04.2018.

[19] W. Ertel. *Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung*. Computational Intelligence. Vieweg+Teubner Verlag / GWV Fachverlage GmbH Wiesbaden, Wiesbaden, 2., überarbeitete auflage edition, 2009.

[20] Facebook. Jobs. https://www.facebook.jobs/jobs/?q=data+scientist, 2018. Letzte Einsichtnahme: 28.04.2018.

[21] D. Fasel and A. Meier, editors. *Big Data: Grundlagen, Systeme und Nutzungspotenziale*. Edition HMD. Springer Vieweg, Wiesbaden, 2016.

[22] Gartner. Gartner says cios and cdos must 'digitally remaster' their organizations. https://www.gartner.com/newsroom/id/2975018, 2015. Letzte Einsichtnahme: 30.04.2018.

[23] Gartner. More than 40 percent of data science tasks will be automated by 2020. https://www.gartner.com/newsroom/id/3570917, 2017. Letzte Einsichtnahme: 21.04.2018.

[24] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[25] F. Gorunescu. *Data Mining: Concepts, Models and Techniques*, volume 12 of *Intelligent Systems Reference Library*. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2011.

[26] W. Goschy and T. Rohrbach. Industrie 4.0. https://www.staufen.ag/fileadmin/HQ/02-Company/05-Media/2-Studies/STAUFEN.-studie-deutscher-industrie-4.0-index-2017-de_DE.pdf, 2017. Letzte Einsichtnahme: 24.04.2018.

[27] T. H. Davenport and D. J. Patil. Data scientist: The sexiest job of the 21st century. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century, 2012. Letzte Einsichtnahme: 08.04.2018.

[28] B. Hayes, D. Hinchcliffe, J. Shin, J. Caserta, and McKendrick Jow. Data science: Present and future — ibm data science for all. https://www.thecube.net/ibm-data-for-all-2017/content/Videos/3TqYuK25bjqmkKyCE, 2017. Letzte Einsichtnahme: 14.04.2018.

[29] W. Knight. The dark secret at the heart of ai. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/, 2017. Letzte Einsichtnahme: 08.04.2018.

[30] M. Kubat. *An introduction to machine learning*. Springer, Cham, 2015.

[31] R. Kumar. *Machine learning and cognition in enterprises: Business Intelligence Transformed*. Apress, Berkeley, CA, 2017.

[32] W. Markow, S. Braganza, B. Taska, S. M. Miller, and D. Hughes. The quant crunch: How the demand for data science skills is disrupting the job market. https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/analytics-analytics-platform-im-analyst-paper-or-report-iml14576usen-20171229.pdf, 2017. Letzte Einsichtnahme: 14.04.2018.

[33] B. Marr. The top 10 ai and machine learning use cases everyone should know about. https://www.forbes.com/sites/bernardmarr/2016/09/30/what-are-the-top-10-use-cases-for-machine-learning-and-ai/#1c37a0b194c9, 2016. Letzte Einsichtnahme: 12.04.2018.

[34] McKinsey. Mckinsey-studie zu industrie 4.0. https://www.mckinsey.de/mckinsey-studie-zu-industrie-40-deutsche-unternehmen-trotz-wachsender-konkurrenz- 2017. Letzte Einsichtnahme: 14.04.2018.

[35] MIT Laboratory for Information and Decision Systems. Ml 2.0: Machine learning for many: Automated data science tools developed by mit and feature labs deliver their first ai product. http://news.mit.edu/2018/ml-20-machine-learning-many-data-science-0306, 2018. Letzte Einsichtnahme: 18.04.2018.

[36] T. M. Mitchell. *Machine learning*. McGraw-Hill series in computer science. McGraw-Hill, New York, NY, international ed., [reprint.] edition, 2010.

[37] S. Mohanty, M. Jagadeesh, and H. Srivatsa. *Big Data Imperatives: Enterprise Big Data Warehouse, BI Implementations and Analytics*. The expert's voice in big data. Apress, Berkeley, CA and s.l., 2013.

[38] T. Nagle, T. C. Redman, and D. Sammon. Only 3% of companies' data meets basic quality standards. https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards, 2017. Letzte Einsichtnahme: 08.04.2018.

[39] A. Ng. The state of artificial intelligence. https://www.youtube.com/watch?v=NKpuX_yzdYs, 2017. Letzte Einsichtnahme: 14.04.2018.

[40] A. Ng. How artificial intelligence and data add value to businesses. https://www.mckinsey.com/global-themes/artificial-intelligence/how-artificial-intelligence-and-data-add-value-to-businesses, 2018. Letzte Einsichtnahme: 17.04.2018.

[41] M. Paluszek and S. Thomas. *MATLAB machine learning*. For professionals by professionals. Apress, New York, 2017.

[42] D. J. Patil. Building data science teams. http://radar.oreilly.com/2011/09/building-data-science-teams.html, 2011. Letzte Einsichtnahme: 08.04.2018.

[43] G. Piatetsky. Crisp-dm, still the top methodology for analytics, data mining, or data science projects. https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html, 2014. Letzte Einsichtnahme: 10.04.2018.

[44] G. Press. A very short history of data science. https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#1348ea855cfc, 2013. Letzte Einsichtnahme: 07.04.2018.

[45] K. Ramasubramanian and A. N. Singh. *Machine learning using R*. Apress, New York, 2017.

[46] Á. Rocha, A. M. Correia, H. Adeli, L. P. Reis, and S. Costanzo, editors. *Recent Advances in Information Systems and Technologies: Volume 2*, volume 570 of *Advances in Intelligent Systems and Computing*. Springer International Publishing, Cham and s.l., 2017.

[47] S. Saitta. Data science automation: Debunking misconceptions. https://www.kdnuggets.com/2016/08/data-science-automation-debunking-misconceptions.html, 2016. Letzte Einsichtnahme: 16.04.2018.

[48] D. Sarkar, R. Bali, and T. Sharma. *Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems*. 100 facts you should know. Apress, Berkeley, CA, 2018.

[49] W. Schnober and R. Geiger. Digitalisierung: Herausforderungen und wie unternehmen sie bewältigen. https://blog.tci-partners.com/digitalisierung-herausforderungen-und-wie-unternehmen-sie-bewaeltigen/, 2017. Letzte Einsichtnahme: 12.04.2018.

[50] B. Schreck, M. Kanter, K. Veeramachaneni, S. Vohra, and R. Prasad. Getting value from machine learning isn't about fancier algorithms — it's about making it easier to use. https://hbr.org/2018/03/getting-value-from-machine-learning-isnt-about-fancier-algorithms-its-about-makir 2018. Letzte Einsichtnahme: 22.04.2018.

[51] C. Schumann, P. Zschech, and A. Hilbert. Das aufstrebende berufsbild des data scientist: Vom kompetenzwirrwarr zu spezifischen anforderungsprofilen. pages 453–466.

[52] J. Snow. New research aims to solve the problem of ai bias in "black box" algorithms. https://www.technologyreview.com/s/609338/new-research-aims-to-solve-the-problem-of-ai-bias-in-black-box-algorithms/, 2017. Letzte Einsichtnahme: 14.04.2018.

[53] Stepstone. Jobs. https://www.stepstone.de/5/ergebnisliste.html?gclid=Cj0KCQjw2pXXBRD5ARIsAIYoEbf0qvaCSXx6dh55h1dO61Kxcph5citWEAiDHTq1_cFfv4wL8fWxqyoaAockEALw_wcB&stf=freeText&ke=Data%20Scientist&ws=&loc_interest=&loc_physical=9042160&cid=SEAdvert_Google_SEARCH_DE_1000000-E_c_Jobs-Data-Scientist_jobs%20data%20scientist_RLd_EtaId2-L1_-&s_kwcid=AL!523!3!213380912361!e!!g!!jobs%20data%20scientist&ef_id=WWDCvAAAAbFNpBBO:20180429173009:s, 2018. Letzte Einsichtnahme: 28.04.2018.

[54] V. Sze, Y.-H. Chen, J. Emer, A. Suleiman, and Z. Zhang. Hardware for machine learning: Challenges and opportunities. https://arxiv.org/pdf/1612.07625.pdf, 2017. Letzte Einsichtnahme: 14.04.2018.

[55] T. Szent-Ivanyi. Autonomes fahren ethiker sehen probleme bei selbstfahrenden autos. https://www.berliner-zeitung.de/wirtschaft/autonomes-fahren-ethiker-sehen-probleme-bei-selbstfahrenden-autos--29563794, 2018. Letzte Einsichtnahme: 10.04.2018.

[56] tutorialspoint. Artificial intelligence - neural networks. https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_neural_networks.htm. Letzte Einsichtnahme: 06.04.2018.

[57] W. Vorhies. Crisp-dm – a standard methodology to ensure a good outcome. https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome, 2016. Letzte Einsichtnahme: 10.04.2018.

[58] W. Vorhies. Data science is changing and data scientists will need to change too – here's why and how. https://www.datasciencecentral.com/profiles/blogs/data-science-is-changing-and-data-scientists-will-need-to-change-, 2018. Letzte Einsichtnahme: 24.04.2018.

[59] K. Wang, Y. Wang, J. O. Strandhagen, and T. Yu, editors. *Advanced Manufacturing and Automation VII*, volume 451 of *Lecture Notes in Electrical Engineering*. Springer Singapore, Singapore, 2018.

[60] D. Wellers, T. Elliott, and M. Noga. 8 ways machine learning is improving companies' work processes. https://hbr.org/2017/05/8-ways-machine-learning-is-improving-companies-work-processes, 2017. Letzte Einsichtnahme: 12.04.2018.

[61] wikipedia. Overfitting. https://en.wikipedia.org/wiki/Overfitting. Letzte Einsichtnahme: 06.04.2018.

[62] R. Wirth and J. Hipp. Crisp-dm: Towards a standard process model for data mining. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf, 2000. Letzte Einsichtnahme: 16.04.2018.