Big Data und KI

Roman Kollatschny

Matr.Nr. **30796**

rk056@hdm-stuttgart.de

mail@kollatschny.net

Hochschule der Medien

Computer Science and Media

Abstract

Diese Arbeit beschäftigt sich mit den Grundlagen von Big Data und wie Künstliche Intelligenz im Kontext von Big Data genutzt werden kann. Es wird dabei die Grundlagen von Big Data erklärt und mit welchen Herausforderungen bei der Nutzung der Daten zu beachten sind. In einem einfachen Architekturmodell werden die einzelnen Schichten beschrieben und erläutert, welche Komponenten und Funktionen diese beherbergen. Neben einer Einführung in die KI werden die Anwendungsgebiete dieser im Kontext Big Data definiert.

Inhalt

Einleitung

Die Menge an digitalen Daten in unserer Welt wächst mit einer exponentiellen Rate. Rund alle zwei Jahre verdoppelt sich dabei die erzeugen Daten, sodass die Menschheit bis zum Jahr 2020 ein Datenvolumen von über 40.000 Exabyte erzeugt hat. Zur besseren Verdeutlichung dieser Menge reicht schon die Umrechnung von Exabyte in die für Anwender üblicheren Gigabyte, in Gigabyte wären dies 4e*10 Gigabyte oder auch ausgeschrieben 40.000.000.000.000 Gigabyte.

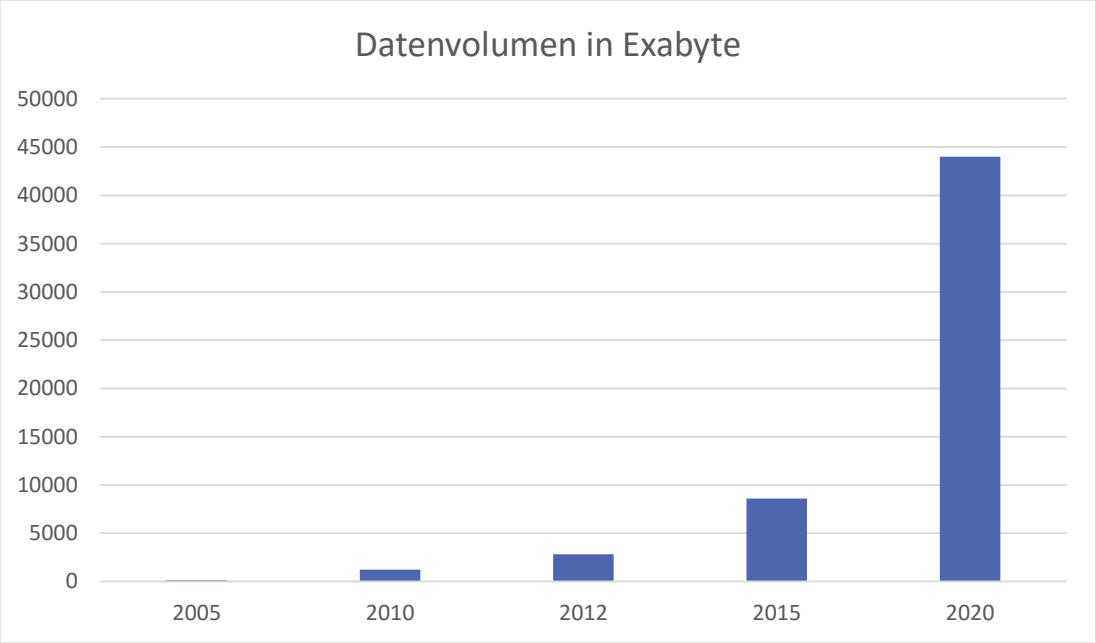

Abbildung 1: Diagramm mit dem geschätzten Datenvolumen bis 2020[1]

Die Geschwindigkeit, mit der die Daten erzeugt werden überholt dabei jedoch unsere Fähigkeiten, diese zu analysieren. Schon heute werden etwa nur 0.5%[2] aller Daten, die auf der Welt generiert werden, analysiert und untersucht. Und je mehr Daten in Zukunft generiert werden, desto geringer wird dieser Prozentsatz noch werden.

Die Frage ist also, wie wir dieser Flut an Daten entgegentreten könne und die generierten Daten, in denen eine Unmenge von Informationen stecken sinnvoll und effizient für uns nutzten können. Das Schlüsselwort hierzu lautet Big Data. Mit Big Data öffnen sich fast unbegrenzte Möglichkeiten. Big Data kann dazu eingesetzt werden DNA Stränge zu kodieren um Krankheitsbilder besser hervorsagen zu können[3], es

[1] Statista: Prognose zum Volumen der jährlich generierten digitalen Datenmenge weltweit in den Jahren 2005 bis 2020 (in Exabyte); unter http://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/
[2] Technologyreview: Big Data: Creating the Power to Move Heaven and Earth; https://www.technologyreview.com/s/530371/big-data-creating-the-power-to-move-heaven-and-earth/
[3] IBM Research: Doctors will routinely use your DNA to keep you well; unter http://www.research.ibm.com/cognitive-computing/machine-learning-applications/targeted-cancer-therapy.shtml

kann dabei helfen Energiekosten in Gebäuden einzusparen oder aber auch um Produktionen effizienter gestalten zu können.

Im folgenden Teil dieser Arbeit werden die Themengebiete Big Data und KI beschrieben. Das Hauptaugenmerkmal liegt dabei auf den Grundlagen beider Themengebiete sowie wie sie zusammen eingesetzt werden können, um die zukünftigen Herausforderungen bei der Verarbeitung und Analyse im Hinblick auf immer größer werdendes Datenaufkommen zu begegnen.

Big Data

In der Zeit von Big Data haben Unternehmen Zugriff auf eine schier unendliche Anzahl von verschiedenen Daten. Das Datenaufkommen ist dabei jedoch so enorm, dass der Mensch und die derzeitigen Technologien nicht mehr damit mithalten können, diese auf herkömmlichen Wegen zu verarbeiten und auszuwerten. Durch das immer weitere Zurückfallen der Unternehmen verpassen diese unzähligen Möglichkeiten aus diesen Daten zu lernen und anzuwenden, was sie gelernt haben.

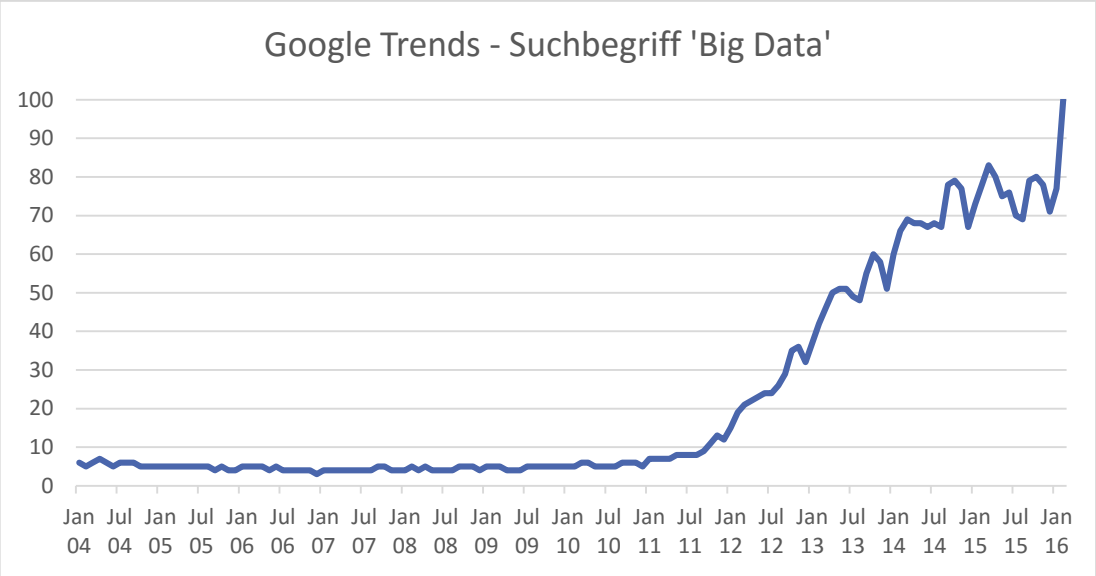

Abbildung 2: Diagramm mit der relativen Anzahl von Suchanfragen zum Begriff 'Big Data'[4]

Das Big Data dabei jedoch nicht nur eine temporäre Erscheinung oder eines von vielen Hype[5] Themen ist, sondern immer mehr Unternehmen planen dieses einzusetzen oder sogar verschiedene Big Data Systeme in Verwendung haben, lässt sich beispielweise auch anhand der Google Trends Services ablesen. Zwar sind für den Suchbegriff keine diskreten Angaben zu der Anzahl der Suchanfragen vorhanden, jedoch lässt sich auch so sehr gut anhand des Diagramms ablesen. Im Zeitraum der 2000 Jahre blieb das Interesse an diesem Thema auf einer relativ stabilen Nachfrage. Ab Anfang des Jahres 2011 gewann es jedoch immer mehr an Bedeutung und die Suchanfragen schnellten diesbezüglich in neue Höhen. So hat sie die Anzahl der Suchanfragen bis heute um mehr als das zwanzigfache wie noch vor fünf Jahren erhöht.

[4] Die Zahlen stellen das Suchinteresse relativ zum Höchstwert im Diagramm dar. Das höchste Suchinteresse ist dabei immer 100%.
[5] datanami: Why Gartner Dropped Big Data Off the Hype Curve; unter http://www.datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve/

Zwar ist nun das Interesse an Big Data bei vielen geweckt, jedoch wissen vermutlich nur wenige, was damit gemeint ist. In einem Post auf Facebook hat es dabei der amerikanische Psychologe und Hochschullehrer Dan Ariely im Jahr 2013 passend formuliert:

> *Big Data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.*[6]

In Anlehnung an dieses Zitat werden im folgenden Kapitel die Fragen was Big Data ist, welche Daten verwendet werden, woher diese Daten stammen und wie man Big Data nutzen kann beantwortet.

Definition

Die elektronische Verarbeitung von Daten gibt es schon sehr lang und wird auch praktisch in jedem Unternehmen für verschiedenen Prozesse eingesetzt. Wie unterscheidet sich nun also die klassische EDV von der Datenverarbeitung in Kontext von Big Data. Die Datensätze, die in Big Data verwendet werden unterscheiden sich nicht nur in der Größe oder dem Umfang, sondern auch der Komplexität oder in der Form, dass sie sich während der Zeit ändern können von den Daten auf die bisher zugegriffen wurde. Ihre Eigenschaften verhindern es daher sie mit manuell oder mit klassischen Methoden der Datenverarbeitung auswerten zu können.

Neben der eigentlichen Bezeichnung der Datensätze wird der Begriff ‚Big Data' auch oft als der Komplex der Technologien beschrieben der dafür notwendig ist. Dieser Komplex beinhaltet dabei verschiedene Technologien die zum Sammeln, Speichern und Auswerten der Datenmengen benötigt werden. Die Herausforderungen umfassen dabei vor das Erfassen, das Speichern, den Transfer, die Analyse, die Datenpflege, das Abfragen, das Suchen, das Teilen, die Visualisierung und der Informationsdatenschutz.

Merkmale

Mit Hilfe von Merkmalen lassen sie die in Big Data verwendeten Datensätze genauer charakterisieren. Es wird dabei zwischen mehreren verschiedenen Merkmalen unterschieden. Mit der Zeit wurden dabei immer mehr dieser Merkmale definiert, sodass

[6] Facebook: Dan Ariely; unter https://www.facebook.com/dan.ariely/posts/904383595868

die Anzahl dieser von drei[7] am Anfang auf nun bis zu sieben[8] Merkmalen angewachsen ist. Im Folgenden werden die fünf wichtigsten und meistgenannten Merkmale aufgelistet und beschrieben.

- **Volume**: Mit dem Volume wird die Skalierung der Daten bezeichnet. Es wird dazu zum einen die Größe der Daten an sich als auch die Masse oder die Anzahl der Datensätze die zu Verfügung stehen beschrieben.
- **Velocity**: Mit Velocity ist das Tempo der Daten gemeint. Die sich ständig erhöhende Geschwindigkeit bezieht sich dabei auf die Erzeugung, das Speichern, die Verarbeitung und die Analyse der Daten.
- **Variety**: Mit Variety wird die Verschiedenheit der Daten bezeichnet. Die Daten bei Big Data unterscheidet man in strukturiert, halb oder semistrukturiert oder unstrukturiert. 90% der Daten liegt dabei unstrukturiert vor. Die Datensätze können dabei in Form von üblichen Text Dokumenten oder Tabellen, audiovisuellen Daten wie Bild und Video oder Audiodatensätzen, oder Maschinengenerierte Daten wie Serverlogfiles vorkommen.
- **Veracity**: ‚Veracity' beschreibt die Glaubwürdigkeit der Daten. Die Daten haben nur einen Wert, wenn sie weder falsch noch ungenau sind. Dies gilt vor allem im Zusammenhang mit Diensten die unüberwachte Machine Learning Algorithmen verwenden, da die Ergebnisse solcher Anwendungen nur so gut sind, wie die Daten die sie zu Verfügung haben.
- **Variability**: Mit ‚Variability' ist die Varianz der Daten gemeint. Die Bedeutung der Daten ist nicht fest, sondern kann sich mit der Zeit oder mit dem Kontext ändern. Dies ist vor allem bei von Menschen generierten Datensätzen wie Texten der fall, da Worte keine statischen Definitionen haben und ihre Bedeutung je nach Kontext oder Betonung variieren kann.

In manchen Quellen werden noch ‚Visibility' oder ‚Value' als Merkmale von Big Data genannt. Diese stellen jedoch an sich keine Merkmale dar, da sie nicht die Daten in Big Data charakterisieren. Sie sind viel mehr als die Ergebnisse oder Endprodukte zu verstehen, die mit Hilfe von Big Data erzeugt werden.

Big Data Life Cycle

Bevor damit begonnen werden kann eine Architektur für ein Big Data System zu entwickeln ist es wichtig die Anforderungen zu berücksichtigen. In Abbildung 3 wird dargestellt, dass die Daten als erstes erfasst, dann organisiert und integriert werden. Wurde diese Phase erfolgreich durchgeführt, können die Daten nach dem definierten

[7] Gartner Blogs: 3D Data Management: Controlling Data Volume, Velocity, and Variety; unter http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[8] Dataconomy: Understanding Big Data: The Seven V's; unter http://dataconomy.com/seven-vs-big-data/

Wert hin analysiert werden. Letztlich werden die Ergebnisse dazu genutzt um auf Basis der Erkenntnis der Analyse zu agieren.

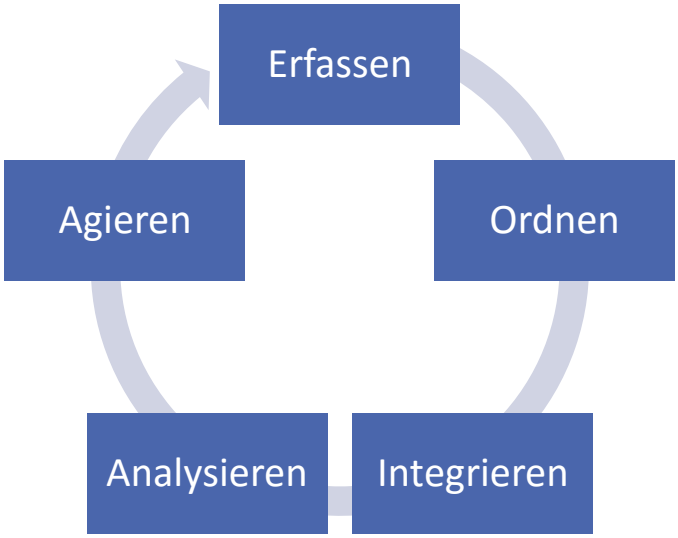

Abbildung 3: Big Data Management Life Cycle

Auch wenn dies auf den ersten Blick relativ einfach aussieht, sind bestimmte Feinheiten von diesen Funktionen kompliziert. So ist beispielsweise die Validierung ein besonders wichtiges Thema. Wenn ein Unternehmen verschiedene Datenquellen kombinieren möchte ist es wichtig, dass man die Möglichkeit hat zu überprüfen, ob die Quellen überhaut zusammen einen Sinn ergeben. Auch können manche Datenquellen vertrauliche Informationen enthalten, sodass man für ausreichende Sicherheit Level und Governance sorgen muss.

Big Data Architektur

Um eine grobe Planung eines Big Data Systems zu ermöglichen bietet es sich an, eine Architektur auf logischen Schichten aufzubauen. Die logischen Schichten bieten die Möglichkeit, ihre Komponenten, die bestimmte Funktionen ausführen, zu organisieren. Da die schichten logisch aufgebaut sind, bedeutet das nicht, dass Funktionen, die jede Schicht unterstützten auf einer separaten Maschine oder Prozess laufen.

Eine Big Data Solution umfasst typischerweise folgende logischen Schichten.

1. Big Data Quellen Schicht
2. Daten Anpassung und Sicherung Schicht
3. Analyse Schicht
4. Verbraucher Schicht

Die erste Schicht bilden alle Quellen und Kanäle, die verfügbaren Daten für die spätere Analyse bereitstellen. Wie zuvor beschrieben, variieren die Daten in Format und der Herkunft.

- **Format**, strukturierte, halb oder unstrukturierte Daten
- **Geschwindigkeit** und **Volumen** mit der die Daten ankommen variiert je nach Datenquelle
- **Sammelpunk**, wo die Daten gesammelt werden; direkt oder durch Datenprovider; in Echtzeit als Stream oder in Batch Mode mit Datenpaketen.
- **Herkunft** der Daten innerhalb des Unternehmens oder von einem externen Dienstleister; zeitlich beschränkter Zugriff auf Datenquellen.

Im Folgenden werden mögliche Datenquellen aufgezählt, die für die Verwendung als Big Data Quellen Verwendung finden.

- Bereits **bestehende Systeme** wie Customer Relationship Management Systeme, Rechnungssysteme, Mainframe Anwendungen oder Web Anwendungen
- **Datenmanagement Systeme** wie Microsoft Word und Office, Dokumente die in strukturierte Daten konvertiert werden können
- **Datenspeicher** wie Data Warehouses, Operative Datenbanken oder Transaktionsdatenbanken, deren Daten in der Regel bereits strukturiert vorliegen und direkt zur Analyse verwendet werden können
- **Smart Devices**, die in Informationen erfassen, verarbeiten und über verschiedene Protokolle und Formate weitersenden können. Smartphones, Tablets, Smart Meters und Wearables zählen beispielsweise zu diesen Geräten und ermöglichen eine Echtzeit Analyse.
- **Andere Datenquellen**, die unterschiedliche Arten von Daten bereitstellen wie beispielsweise geografische Daten wie Karten oder Regionale Details, von Menschen erzeugter Content wie Social Media, Emails oder Blogs, Sensor Daten über die Umgebung, Elektrizität, Position, Akustik, Optisch, etc.

In der Daten Anpassung und Sicherung Schicht werden die Daten aus der darunterliegenden Quellen Schicht erfasst und, sofern es erforderlich ist, in ein Format konvertiert, das für die Sicherung und die Analyse benötigt wird. Die Schicht beinhaltet dabei drei wichtige Komponenten zur Datenerfassung, Datenvorbereitung und Datensicherung.

- Datenerfassung: Erfasst Daten von den unterschiedlichen Datenquellen und leitet die Daten zur Komponente zur Datenverarbeitung oder zur Komponente zur Datensicherung. Die Komponente muss dabei intelligent genug sein um zu entscheiden, ob die Daten zuerst von der anderen Komponente angepasst und verarbeitet werden müssen oder direkt an die Analyse Schicht weitergegeben werden kann.

- Datenvorbereitung: Zuständig für Anpassung der Daten in das für die Analyse benötigte Format. Diese Komponente kann dabei einfache logische Transformationen oder komplexe statistische Algorithmen aufweisen um die Quelldaten zu konvertieren. Die größten Herausforderungen sind unstrukturierte Datenformate wie beispielsweise Bild, Video, Audio oder andere binäre Formate.
- Datensicherung: Speichert die Daten aus den Quellen. Oftmals sind dabei verschiedene Optionen zur Speicherung der Daten wie Distributed File Systems, Cloud, NoSQL oder andere vorhanden.

Die Analyse Schicht liest die von der vorherigen Schicht vorbereiteten Daten aus. In manchen Fällen kann die Analyse Schicht auch direkt auf die Datenquellen zugreifen. Dazu müssen die Daten jedoch bereits in einem für die Analyse passenden Format vorliegen. Entscheidungen bei der Planung dieser Schicht müssen im Hinblick auf folgende Aufgaben gemacht werden:

- Erzeugen der gewünschten Analysen
- Ableiten der Erkenntnis aus den Daten
- Finden der benötigten Entitäten
- Ermitteln der Datenquellen, die Daten für diese Entitäten bereitstellen
- Verstehen, welche Algorithmen und Anwendungen benötigt werden, um die Analysen durchzuführen

In der letzten Schicht wird die Ausgabe der Analysen bereitgestellt. Die Ergebnisse der Analyse werden dabei von verschiedenen Nutzern innerhalb oder auch außerhalb des Unternehmens verwendet. Externe Nutzer stellen dabei Kunden, Verkäufer, Partner oder Zulieferer dar.

KI – Künstliche Intelligenz

In den letzten Jahren erlebt die KI als ein Teilgebiet der Informatik eine weitere Renaissance. Viele große namenhafte Firmen aus der Tech Branche wie Google, Apple, Facebook, Amazon, Microsoft oder IBM, investieren enorm um im Wettlauf um das beste und intelligentestes System nicht zurückzufallen[9].

Warum es derzeit eine erneute große Nachfrage nach Technologien dieser Art gibt liegt darin, das praktisch universell eingesetzt werden können und als Allheilmittel für viele Problemen in der Informatik angesehen werden. Das diese Technologien bereits in vielen verschiedenen Bereichen erfolgreich eingesetzt werden zeigt, dass sie nicht nur einfache Hype Themen[10] sind.

Definition

Künstliche Intelligenz oder auch einfach KI genannt, ist ein Forschungsgebiet in der Informatik das sich damit beschäftigt intelligentes menschliches Verhalten auf maschinelle Systeme zu übertragen. Es wird dabei versucht das menschliche Gehirn virtuell in einem System nachzubilden, damit dieses sowohl denken als auch lernen kann wie es in intelligenter Mensch tun würde.

In Bezug auf Big Data werden dabei Systeme gemeint, die die riesigen und anwachsenden Speicher von Daten analysieren und verarbeiten können. Dazu werden verschiedene Methoden und Algorithmen eingesetzt um die Daten zu sortieren, Informationen aus diesen zu extrahieren, die Informationen zu analysieren und in Verbindung zueinander zu bringen.

Was KI aber für Big Data besonders attraktiv macht ist eine Eigenschaft der Lernfähigkeit. Eine KI Anwendung kann nicht nur das ausführen, was man ihr einprogrammiert hat, sie ist auch in der Lage sich auf Änderungen einzustellen und selbst Anpassungen basierend auf dem was sie bisher gelernt hat an sich vorzunehmen.

Machine Learning

Machine Learning, im Deutschen auch Maschinelles Lernen genannt, ist ein Teilgebiet der Informatik, das sich mit der künstlichen Generierung von Wissen aus Erfahrung beschäftigt. Im Allgemeinen lernt dabei ein System aus Testdaten verschiedene Muster und Gesetzmäßigkeiten, um diese dann später an neuen Daten anwenden zu können und diese beurteilen zu können.

[9] Financial Times: Investor rush to artificial intelligence is real deal; unter http://www.ft.com/cms/s/2/019b3702-92a2-11e4-a1fd-00144feabdc0.html
[10] wired: Myth Busting Artificial Intelligence; unter http://www.wired.com/insights/2015/02/myth-busting-artificial-intelligence/

In Machine Learning gibt es dabei mehrere verschiedenen Ansätze, wie die Systeme aufgebaut sind und funktionieren. Man unterscheidet dabei hauptsächlich zwischen drei Ansätzen, dem überwachten und dem unüberwachten Lernen sowie dem verstärkenden Lernen.

- **Überwachtes Lernen**: Bei Systemen die auf überwachte Lernalgorithmen setzten werden die Modelle gelernt, die verschiedene Datensätze untereinander unterscheiden. Diese Algorithmen sind später beispielsweise im Stande unterschiedliche Daten einer definierten Kategorie zuzuordnen. Um dies zu ermöglichen werden solche Systeme mit Testdaten trainiert, die bereits einer entsprechenden Kategorie zugeordnet sind. Dadurch kann das System die Merkmale und Charakteristiken, die eine Kategorie ausmachen lernen und ein Modell dieser Kategorie erzeugen. Ist das System trainiert, können die erzeugten Modell dazu verwendet werden, die eigentlichen Datensätze zu sortieren.

- **Unüberwachtes Lernen**: Systeme, die auf unüberwachtes Lernen setzten, ermöglichen eine Sortierung und Strukturierung von Datensätzen. Im Gegensatz zu überwachten Lernverfahren, werden hier keine Modelle trainiert, sondern die Daten aufgrund von Regelmäßigkeiten und Mustern sortiert. Bei diesem Ansatz untersuchen die Methoden die Daten auf den jeweiligen Merkmalen und Unterschieden hin um sie dann aufgrund dieser Informationen in Gruppen einzuteilen. Im Gegensatz zu überwachten Lernverfahren sind die Klassen oder Gruppen im vornherein noch unbekannt.

- **Verstärkendes Lernen**: Bei diesem Ansatz muss sich die Anwendung in einer dynamischen Umgebung orientieren, in dem es ein gewisses Ziel erreichen muss. Eine Rückmeldung, wie beim Überwachten Lernen von einem Lehrer gibt es hier nur in der Form, dass ein sogenannter Kritiker nach mehreren Aktionen der Anwendung ein positives oder negatives Feedback gibt.

Neben den Ansätzen im Machine Learning können Anwendungen auch nach ihrer Funktion und Aufgabe, beziehungsweise nach deren Ergebnisse eingeteilt werden. Man unterscheidet dazu unter Klassifikation, Regression, Clustering und der Reduktion von Dimensionen.

- **Klassifikation** setzt meist auf überwachte Lernverfahren und ermöglicht die Eiteilung der Daten in verschiedene Klassen. Solche Anwendungen werden beispielsweise bei der Sortierung von Mails in Spam verwendet, die die Mails in die Klassen ‚Spam' und ‚Nicht Spam' einteilt.

- **Regression**, bei der der Zusammenhang zwischen den einzelnen Datensätzen geschätzt wird. Auch hier werden vor allem überwachte Lernmethoden verwendet.

- **Clustering** werden hingegen unüberwachte Lernverfahren eingesetzt, um die Datensätze in verschiedenen Gruppen einteilen zu können, die im Gegensatz zur Klassifikation noch nicht bekannt sind.
- **Dimensionsreduktion** wird bei hochdimensionalen Datensätzen verwendet, indem diese in einem niedriger dimensionalen Raum transformiert werden.

Natural Language Processing

Natural Language Processing, kurz NLP, ist ein Teilgebiet der Informatik, das sich mit der menschlichen Sprache, der Natural Language, befasst. Es wird dabei verstärkt auf Machine Learning Methoden gesetzt, um Computern das Verstehen einer natürlichen Sprache zu ermöglichen. Mit ‚Verstehen' ist dabei nicht nur das Erkennen der unterschiedlichen Worte gemeint, sondern auch das Verstehen der Bedeutung dieser in einem bestimmten Kontext. Als Eingabe kann dabei sowohl die Sprache in gesprochener Form als Audio Daten vorliegen als auch in Form von Dokumenten in Textform.

Natural Language Processing wird dabei für verschiedenen Aufgaben verwendet, dies umfasst beispielsweise das automatische Zusammenfassen von Texten, die maschinelle Übersetzung, Spracherkennung oder OCR um nur einige zu nennen.

Im Kontext von Big Data werden diese Technologien hauptsächlich dafür eingesetzt, um von Menschen generierte Daten verarbeiten zu können, wie sie beispielsweise in Textdokumenten in Unternehmen oder in Posts und Artikeln im Social Media und dem Internet vorkommen.

Computer Vision

Computer Vision beschäftigt sich mit der Entwicklung von Fähigkeiten des menschlichen Sehens für maschinelle Systeme. Das Ziel von Computer Vision ist es Computern visuelle Daten wie Bildern und Videos effizient wahrzunehmen, zu verarbeiten und zu verstehen.

Häufige Aufgaben der Computer Vision sind dabei die Recognition, Motion Analysis, Scene Reconstruction und die und Image Restoration.

Dimensionsreduktion wird bei hochdimensionalen Datensätzen verwendet, indem diese in einem niedriger dimensionalen Raum transformiert werden.

- **Recognition**: Hierbei unterschiedet man unter drei Arten: Object Recognition beziehungsweise Object Classification, Identification und Detection. Object Recognition dient zum Erkennen von verschiedenen zuvor definierten Objekten in einer Szene, wie beispielsweise Menschen, Autos und Bäume. Identification hingegen dient zum Erkennen von spezifischen Objekten wie einer bestimmten Personen oder dessen Fingerabdrücken. Detection durchsucht das Bild nach einer bestimmten Bedingung und gibt dabei Rückmeldung, ob diese erfüllt oder nicht erfüllt ist.
- **Motion Analysis** benötigt eine Sequenz von Bildern um auf dieser eine Bewegung erkennen zu können. Dies kann zum einen die Bewegung von einzelnen Körpergliedern eines Menschen sein oder die Bewegungen mehrere Menschen auf einem größeren Platz.
- **Scene Reconstruction** dient zur Rekonstruktion von Szenen. Um eine Rekonstruktion eines 3D Raumes zu ermöglichen werden mehrere Bilder oder Bildsequenzen benötigt
- **Image Restoration** beschäftigt sich mit Entfernen von Rauschen und damit der Verbesserung der Qualität der Bilddaten.

Im Kontext von Big Data werden Computer Vision Methoden dazu eingesetzt, um die Daten aus visuellen Quellen zum einem effizient speichern zu können und für die weitere Analyse vorzubereiten. Des Weiteren können die Bilddaten durch Object Recognition Methoden auf deren Inhalt, auf das was sie Abbilden, hin analysiert werden. Dies kann beispielsweise in Kombination mit Natural Language Processing verwendet werden um beispielsweise abgebildete Zeichen und den Text zu erkennen.

Big Data und KI

Wie passen Big Data und KI nun zusammen. Wirft man einen Blick auf die beiden vorherigen Kapitel haben wir einige wichtige Erkenntnisse gewonnen.

- Der Grund warum der Zugriff und die Verwaltung der Daten so schwer ist, da sie aus so vielen verschiedenen Quellen stammen. Anstelle eines einzigen großen Speichers kommen die Daten aus vielen verschiedenen Quellen und sind dabei noch fragmentiert, überflüssig und nicht für eine direkte Analyse geeignet. Es muss erst ein Weg gefunden werden die Daten zu konsolidieren und aufzubereiten bevor diese weiter untersucht werden müssen
- Die bisherigen Techniken und Methoden zur Datenverarbeitung sind für solche Massen von komplexen Daten nicht geeignet um diese effizient verarbeiten zu können. Der Mensch kann der Rate, mit der die Daten anwachsen, nicht mehr standhalten
- Ansätze auf der KI, Machine Learning, Natural Language Processing und der Computer Vision bieten effiziente Methoden, um Daten wie Menschen zu erfassen, verstehen und analysieren zu können. Sie haben dabei die Eigenschaft wie Menschen zu lernen, aus dem bisher Gelernten ihre Aktionen anzupassen.

Werden diese Erkenntnisse kombiniert kommt man dabei auf das Ergebnis, dass Big Data auf die Methoden von Ki, Machine Learning, Natural Language Processing und Computer Vision angewiesen ist, um auch noch in Zukunft die verschiedenen Daten effizient verarbeiten zu können.

Neben den Vorteilen für die KI ergeben sich aber auch Vorteile für die eingesetzten Methoden, da diese je mehr Daten sie verarbeiten immer besser und effizienter werden. Beide Technologien sind somit gegenseitig aufeinander angewiesen und profitieren dabei gelichermaßen.

Anwendungsgebiete

Durch Big Data und KI werden viele unterschiedliche Anwendungsgebiete ermöglicht die vorher so nicht realisierbar waren.

- Die aus Social Media, von Nutzerkonten und Positionsdaten gewonnen Erkenntnisse können beispielsweise dafür genutzt werden um Kunden spezielle

Angebote zu bieten. Läuft ein Kunde an einem Laden vorbei, kann das Unternehmen über die Kundenpräferenzen und Standorterkennung ihm personalisierte Angebote geben.

- Eine weitere Möglichkeit stellt auch die Fraud Detection, die Erkennung von Betrug, dar. Hierbei werden Transaktionen in Echtzeit überprüft und mit den Daten die bereits erfasst und gespeichert wurden verglichen. Stimmen diese nicht überein und es kommt zu Abweichungen von den bisherigen Transaktionen, kann der Kunden über einen möglichen Betrug noch während dieser passiert informiert werden um sofort handeln zu können.

- Ein Empfehlungssystem in einem Online Shop kann auf die Daten zugreifen um Kunden verschiedene Vorschläge zu bieten. Dabei werden zum einen die Daten, die der Nutzer eingibt mit den Daten, die andere Nutzer eingegeben haben und letztendlich gekauft haben verglichen. Auf Basis dieser Erkenntnis kann das System so dem Kunden in Echtzeit die Produkte oder Dienstleistungen vorschlagen die ihm am meisten interessieren.

- Für Unternehmenskunden können beispielsweise verschiedenen Berichte und Übersichten erstellt werden. Diese erlauben es fundierte Entscheidungen und entsprechende Strategien aktuelle Angelegenheiten zu entwerfen. Dabei kann die operative Effizienz durch die in Echtzeit generierten Daten und überwachten Kennzahlen gesteigert werden.

Neben diesen Anwendungsbeispielen können aber auch noch weitere Anwendungsgebiete erschlossen werden. Dies beinhaltet beispielsweise die Medizin, in der ein solches System zur Diagnose von Krankheiten eingesetzt werden kann.

Prinzipiell kann überall dort, wo eine Menge von Daten anfallen diese Big Data in Verbindung mit KI eingesetzt werden, um die Prozesse zu vereinfachen, zu beschleunigen oder effizienter zu gestalten.

Aussicht

Wie zum Anfang dieser Arbeit bereits thematisiert wurde, wird das Volumen in den nächsten Jahren auf über 44 Zettabyte an Daten anwachsen. Für einen Teil dieses Wachstum ist zum einen der Mensch selbst verantwortlich. Zum einen durch das eigenständige erzeugen von Daten in Social Media aber auch durch die Nutzung von verschiedenen Diensten und Geräten, wie Wearables, um seinen Tagesablauf überwachen zu können. Zum anderen Teil werden aber auch noch Technologien wie das Internet of Things[11] oder die Automatisierung von Prozessen und alltäglichen Routinen in Form von selbstfahrenden Fahrzeugen an einem enormen Wachstum dieses Volumens beitragen.

Neben dem Wachstum der Datenmengen werden aber auch die Entwicklungen im Bereich der Künstlichen Intelligenz weiter zu nehmen. Bereits in den vergangenen Jahren konnten so enorme Fortschritte erzielt werden, die davor einer bloßen Fantasie entsprachen. Wer glaubte vor nicht einmal 10 Jahren, dass wir heute einen intelligenten persönlichen Assistenzen namens Siri, Cortana oder Google now in unserer Hosentasche mit uns herumtragen oder Fahrzeuge verwenden, die voll autonom durch den Straßenverkehr navigieren können.

In der Zukunft wird das Bedürfnis von intelligenten Lösungen die auf verschiedenen Methoden der Künstlichen Intelligenz aufbauen immer wichtiger für Big Data werden, da die Komplexität und Masse an Daten immer weiter zunehmen wird. Die Künstliche Intelligenz wird dabei jedoch auch auf Big Data angewiesen sein, um entsprechend intelligente werden zu können und effiziente Methoden liefern zu können.

Ohne KI kein Big Data und ohne Big Data keine KI.

[11] Computer Weekly Data set to grow 10-fold by 2020 as internet of things takes off; unter http://www.computerweekly.com/news/2240217788/Data-set-to-grow-10-fold-by-2020-as-internet-of-things-takes-off

Quellenverzeichnis

Literatur

Sawant, Nitin: Big Data Application Architecture Q&A : A Problem - Solution Approach. New York: Apress, 2014.

Weisman, Loren: Music Business For Dummies. 1. Aufl.. New York: John Wiley & Sons, 2015.

Warden, Pete: Big Data Glossary. Sebastopol: O'Reilly Media, 2011.

Dean, Jared: Big Data, Data Mining, and Machine Learning : Value Creation for Business Leaders and Practitioners. 1. Aufl.. New York: John Wiley & Sons, 2014.

Internetquellen

Zheng, Alice: Evaluating Machine Learning Models: A Beginner's Guide to Key Concepts and Pitfalls. O'Reilly, 2015, unter http://www.oreilly.com/data/free/evaluating-machine-learning-models.csp?intcmp=il-data-free-lp-lgen_free_reports_page

Croll, Alistair: Data: Emerging Trends and Technologies: How sensors, fast networks, AI, and distributed computing are affecting the data landscape. O'Reilly, 2014, unter http://www.oreilly.com/data/free/data-emerging-trends-and-technologies.csp?intcmp=il-data-free-lp-lgen_free_reports_page

Big Data Now: 2015 Edition: Current Perspectives from O'Reilly Media. O'Reilly, 2016, unter http://www.oreilly.com/data/free/big-data-now-2015-edition.csp?intcmp=il-data-free-lp-lgen_free_reports_page

AI meets Big Data: (and eTailers Live Happily Ever After). Umbel, 2015, unter https://www.umbel.com/go/ai-meets-big-data/download/

Mysore, Divakar; Khupat, Shrikant; Jain, Shweta: Big data architecture and patterns, Part 2: How to know if a big data solution is right for your organization. IBM, 2013 unter http://www.ibm.com/developerworks/library/bd-archpatterns2/index.html

Mysore, Divakar; Khupat, Shrikant; Jain, Shweta: Big data architecture and patterns, Part 3: Understanding the architectural layers of a big data solution. IBM, 2013, unter http://www.ibm.com/developerworks/library/bd-archpatterns3/index.html

Quellen

Fußnoten

1. Statista: Prognose zum Volumen der jährlich generierten digitalen Datenmenge weltweit in den Jahren 2005 bis 2020 (in Exabyte); unter http://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/

2. Technologyreview: Big Data: Creating the Power to Move Heaven and Earth; https://www.technologyreview.com/s/530371/big-data-creating-the-power-to-move-heaven-and-earth/

3. IBM Research: Doctors will routinely use your DNA to keep you well; unter http://www.research.ibm.com/cognitive-computing/machine-learning-applications/targeted-cancer-therapy.shtml

4. Die Zahlen stellen das Suchinteresse relativ zum Höchstwert im Diagramm dar. Das höchste Suchinteresse ist dabei immer 100%.

5. datanami: Why Gartner Dropped Big Data Off the Hype Curve; unter http://www.datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve/

6. Facebook: Dan Ariely; unter https://www.facebook.com/dan.ariely/posts/904383595868

7. Gartner Blogs: 3D Data Management: Controlling Data Volume, Velocity, and Variety; unter http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

8. Dataconomy: Understanding Big Data: The Seven V's; unter http://dataconomy.com/seven-vs-big-data/

9. Financial Times: Investor rush to artificial intelligence is real deal; unter http://www.ft.com/cms/s/2/019b3702-92a2-11e4-a1fd-00144feabdc0.html

10. wired: Myth Busting Artificial Intelligence; unter http://www.wired.com/insights/2015/02/myth-busting-artificial-intelligence/

11. Computer Weekly Data set to grow 10-fold by 2020 as internet of things takes off; unter http://www.computerweekly.com/news/2240217788/Data-set-to-grow-10-fold-by-2020-as-internet-of-things-takes-off