Handbuch Data Science und KI

Mit Machine Learning und Datenanalyse Wert aus Daten generieren

2
Das A und O der KI

Stefan Papp

„Unzureichende Fakten bergen immer Gefahren.“

Spock

„Wir haben zwei Ohren und einen Mund, also sollten wir mehr zuhören als sagen.“

Zenon von Kitium

Fragen, die in diesem Kapitel beantwortet werden:

- Wie könnte das menschliche Ego Einzelner den Erfolg oder Misserfolg von KI- und Dateninitiativen beeinflussen?
- Was ist Bias (Voreingenommenheit), und warum ist es wichtig, dieses Thema zu adressieren?
- Wie finden Sie objektiv Ihre Anwendungsfälle und Ziele für Data Science und KI?
- Wie können Workshops durchgeführt werden, um datenorientierter zu werden, und wie können gemeinsame Herausforderungen angegangen werden?

Lassen Sie uns mit einer Hypothese beginnen: Jede Form von Data Science und schließlich auch der künstlichen Intelligenz entwickelt sich aus einer evidenzbasierten Entscheidungsfindung, die auf kritischem Denken beruht. Selbst wenn die außergewöhnlichsten Köpfe zusammenkämen und ihnen unendlich viele Ressourcen zur Verfügung stünden, würden sie scheitern, wenn ihre Arbeit auf falschen Schlussfolgerungen, Fakten und Annahmen beruhte.

In Kapitel 1 stellen wir Tom, den Verkäufer, vor. Als Technologieskeptiker will er Entscheidungen aus dem Bauch heraus treffen. Stellen wir uns vor, der Personalleiter von Halford stellt eine neue Verkäuferin, Cherry, ein, die sich auf Fakten und Logik verlässt. Angenommen, sie ist ebenso qualifiziert und erfahren wie Tom, wird sie dann erfolgreicher sein als er?

In der Wissenschaft geht es um Beweise. Die wissenschaftliche Methode soll subjektive Eindrücke oder Intuition bei der Erforschung eines Themas eliminieren. Die Aufgabe des Data Scientists besteht darin, Beweise für die vom Unternehmen vorgeschlagenen Unternehmungen zu erbringen, und dann in einem zweiten Schritt diese Beweisführung durch statistische Schlussfolgerungen unter Verwendung von Datenpipelines und möglicherweise maschinellen Lernmodellen zu automatisieren. Wenn wir uns eine Organisation vorstellen, die vollständig datengesteuert und faktenbasiert ist, könnten wir uns eine Einheit vorstellen, die ihren Betrieb vollständig darauf ausrichtet, Hypothesen darüber zu sammeln, was gut oder schlecht für das Unternehmen ist, und die auf hochautomatisierte Weise Beweise und Widerlegungen liefert. Sich auf Bauchgefühl und Intuition zu verlassen, könnte als veraltet und als mittelalterlicher Aberglaube abgetan werden.

In späteren Kapiteln werden wir zeigen, dass generative KI nicht immer evidenzbasierte Informationen liefert und dass künstliche Intelligenz voreingenommen sein kann, genau wie Menschen. Bedeutet dies, dass Cherry, die Logik und Fakten liebt, genauso voreingenommen sein kann wie Tom, wenn sie sich auf KI verlässt?

Beginnen wir mit einer Geschichte, um diese Hypothese aus einem anderen Blickwinkel zu betrachten.

2.1 Die Datenverwendungszwecke

Der Personalleiter von Halford hat gerade zwei junge Data Scientists eingestellt: Mahsa und Rachid. Beide beginnen damit, potenzielle analytische Anwendungsfälle zu untersuchen, um einen ersten Anwendungsfall zu ermitteln, der dem gesamten Unternehmen den Wert von Daten aufzeigen kann. Da es sich um ein kleines Team handelt, teilen sie sich die Arbeit auf, befragen verschiedene Abteilungen getrennt und untersuchen sie auf Data-Science-Anwendungsfälle. Irgendwann treffen sich Rachid und Mahsa im Besprechungsraum, um ihre Zwischenergebnisse zu besprechen.

2.1.1 Bias

„Wir sind ein Produktionsunternehmen“, sagt Rachid. „Wenn ich mir all die Fälle ansehe, die ich untersucht habe, sehe ich ein enormes Potenzial, die Zahl der fehlerhaften Teile in unseren Produktionsprozessen zu verringern. Ich weiß, dass das Produktionsteam unserer Arbeit immer noch ablehnend gegenübersteht, aber sie müssen den Wert dessen erkennen, was wir hier tun.“

„Vielleicht ist diese Feindseligkeit ein guter Grund, unseren ersten Anwendungsfall woanders zu starten“, antwortet Mahsa. „Wenn wir unser erstes Projekt vermasseln und uns

einen schlechten Ruf erarbeiten, wird es lange dauern, bis wir das Vertrauen der Belegschaft zurückgewinnen. Wir haben Verbündete in der Abteilung für Kundenbeschwerden; sie haben eine neue Chefin, die sich beweisen will, und sie hat mir gesagt, dass sie die erste Abteilung in Halford sein will, die vollständig datengesteuert ist. Wir sollten mit der Entwicklung eines KI-Chatbots beginnen."

„Komm schon", sagt Rachid, „Beschwerdemanagement ist nicht unser Kerngeschäft. Konzentrieren wir uns auf etwas, das Einnahmen bringt."

„Wir verdienen Geld mit zufriedenen Kunden", entgegnet Mahsa.

„Ja, und *meine* Data-Science-Fallstudien werden dafür sorgen, dass unsere Kunden bessere Produkte bekommen."

„Du hast es in den ersten Tagen selbst gesagt: Wenn es um die Datenqualität geht, heißt es: ‚Garbage in, garbage out'. Meinem ärgsten Feind würde ich die Fabrikdaten nicht geben wollen. Wir werden einen Monat brauchen, um die Werksdaten auf ein akzeptables Niveau zu bringen."

„Aber kein langweiliger Chatbot löst ein Geschäftsproblem", betont Rachid. „Ich habe schon mit Fabrikdaten gearbeitet. Ich weiß, wovon ich spreche. Soweit ich weiß, gehörst du zu den besten Absolventen deiner Universität, aber was hast du außer der Erforschung von Sprachmodellen in einem Universitätskurs gemacht? Hast du irgendwelche praktischen Erfahrungen?"

Mahsa schüttelt den Kopf und fügt laut hinzu: „Ich habe hart daran gearbeitet, diese Strategie zusammen mit dem neuen Leiter der Beschwerdeabteilung zu entwickeln. Wenn wir mit etwas anderem anfangen, ist meine ganze Mühe umsonst gewesen."

* * *

Obwohl Mahsa und Rachid noch am Anfang ihrer Laufbahn stehen, kommt diese Art von Gespräch auch bei erfahreneren Fachleuten vor. Beide werben für ihre Ideen und ignorieren die Meinung des anderen. Beide versuchen, fachlich zu argumentieren, aber sie liefern keine schlüssigen Beweise für ihre Aussagen. Letztlich geht es ihnen darum, einen persönlichen Vorteil zu erlangen: Für Mahsa geht es darum, ihre gute Beziehung zum Leiter der Reklamationsabteilung zu pflegen, und für Rashid darum, seine Erfahrungen aus früheren Projekten zu nutzen. Die Diskussion wird mehr und mehr von Emotionen bestimmt. Sie könnten sich in einem Teufelskreis eines Confirmation Bias befinden.

Vielleicht schlafen die beiden über diesen Konflikt, und vielleicht wachen sie am nächsten Tag mit frischem Geist auf und ändern ihre Meinung. Mahsa könnte anerkennen, dass Rashid Recht hatte, als er sagte, dass Produktionsdaten für den Erfolg des Unternehmens entscheidend sind. Rashid könnte zustimmen, dass Mahsas Begründung, in einer anderen Abteilung anzufangen, der risikoärmere Ansatz sein könnte. Im schlimmsten Fall werden beide auf ihren Ansichten beharren, und ihre Differenzen könnten zu einer Fehde führen, wenn sie nicht einen Manager mit ausreichenden Führungsqualitäten haben, der ihnen hilft, ihren Konflikt zu lösen.

Confirmation Bias

Als Confirmation Bias bezeichnet man das Phänomen, dass wir dazu neigen, Dinge zu ignorieren, die unseren Ansichten widersprechen, aber alle Beweise, die unsere Ansichten bestätigen, in Erinnerung behalten.

Bild 2.1 Confirmation Bias[1)]

Bias wird oft bereits in den ersten Statistikvorlesungen für Studenten behandelt. Ohne zu lernen, Daten korrekt zu interpretieren, können wir keine fundierten Aussagen über unsere Umwelt machen.

Der Lebenszyklus von Data Science besteht in der Regel aus den folgenden Schritten:

1. Erstellen einer Hypothese darüber, was anhand von Daten geändert werden kann
2. Erstellen eines Konzeptnachweises (PoC)
3. Automatisieren einer Lösung

Allerdings kann es zu Diskussionen kommen, bis eine Lösung in Produktion ist, nachdem sie zuvor mehrere Runden positiver Bewertungen und Rückmeldungen erhalten hat. Nehmen wir an, Mahsa hat einen Weg gefunden, einen KI-Chatbot zur Verwaltung der Kundenkommunikation einzuführen, und sie hat gezeigt, wie Kunden mit diesem Chatbot kommunizieren können. Rashid könnte immer noch einwenden und behaupten, er könne mehr Wert schaffen als Mahsa, wenn er bis dahin das Budget für die ersten Anwendungsfälle von Fabrikdaten erhalten hätte.

Selbst wenn die ersten Tests darauf hindeuten, dass Mahsas Innovation den Bedarf an Callcenter-Mitarbeitern verringern könnte, könnte es Monate dauern, um zu über-

[1] Copyright: *https://www.simplypsychology.org/confirmation-bias.html*

3
Cloud-Dienste

Stefan Papp

„Vollkommenheit ist nicht erreicht, wenn es nichts mehr hinzuzufügen gibt, sondern wenn es nichts mehr wegzulassen gibt.“

Antoine de Saint-Exupéry

Fragen, die in diesem Kapitel beantwortet werden:

- Welche Systemumgebungen sind für Data-Science-Projekte erforderlich?
- Warum sind Cloud-Plattformen ideal für experimentelle Data-Science-Projekte?
- Welche Bedeutung haben GPUs und andere Hardwarekomponenten für Data-Science-Projekte?
- Wie können Sie Plattformen dynamisch aufbauen, indem Sie Infrastructure as Code verwenden, und sie mit Versionsverwaltungswerkzeugen verwalten?
- Was unterscheidet eine Microservice-Architektur von einer monolithischen Architektur?
- Wie können Linux-Systeme effizient für Data Science genutzt werden?
- Was sind die jeweiligen Merkmale, Stärken und Schwächen der verfügbaren „Everything-as-a-Service“-Modelle (XaaS)?
- Welche Cloud-Dienste können Datenexperten dabei helfen, Cloud-native Lösungen zu entwickeln?

3.1 Einführung

Dieses Buch beschreibt, wie **künstliche Intelligenz**, **Machine Learning** und **Deep Learning** unser Leben beeinflussen werden. Künstliche Intelligenz ist in letzter Zeit zu einem beliebten Thema in den Medien geworden, und das aus gutem Grund. Die KI-Revolution ist heute in aller Munde, und Sam Altmann ist fast allen, die sich für dieses Thema interessieren, ein Begriff, ein bekannter Name geworden. Sein Unter-

nehmen, OpenAI, wird oft als entscheidender Wegbereiter der generativen KI dargestellt. Unternehmen wie Microsoft oder Google werden als diejenigen angesehen, die die KI schließlich der Bevölkerung zugänglich machen werden.

Während die Unternehmen, die sich mit den softwarebezogenen Aspekten der KI befassen, weltweit Beachtung finden, sind die Unternehmen, die die Hardware herstellen, auf der die KI läuft, und die somit ebenfalls für viele der jüngsten Erfolge der KI verantwortlich sind, die unbesungenen Helden. Zumindest solange, bis wir den Erfolg solcher Unternehmen wie Nvidia an der Börse betrachten. Aber während Nvidia seine Millionen mit der Produktion von Grafikprozessoren verdient hat, gehen die Hardwareanforderungen für KI und Data Science weit darüber hinaus. Bei datengesteuerten Ansätzen untersuchen Datenexperten enorme Datenmengen, manchmal bis zu Petabytes. Um diese Daten zu hosten und zu verarbeiten, ist eine beträchtliche Menge an Hardware-Ressourcen erforderlich.

In diesem Kapitel erörtern wir, wie wir Cloud-Infrastrukturen für unsere Datenprojekte nutzen können. Wir werden uns mit **Infrastructure-as-a-Service (IaaS)** Lösungen beschäftigen, die uns mehr Freiheit bei der Installation der von uns gewünschten Anwendungen und Tools geben. Wir werden uns auch ansehen, wie wir über eine **Platform as a Service (PaaS)** vorausgewählte Pakete eines Cloud-Anbieters nutzen können.

3.2 Cloud-Essentials

In ihrer einfachsten Definition bedeuten Cloud-Dienste die Nutzung des Computers eines anderen. Anstatt Rechenzentren und Serverfarmen zu bauen, können IT-Unternehmen die Anschaffung von Hardware vermeiden und für das bezahlen, was sie nutzen.

Warum „On Premise“ schwierig ist

In Kapitel 1 haben wir ein IT-Unternehmen mit übermütigen IT-Mitarbeitern vorgestellt. Sie schlagen eine „No-Cloud-Strategie“ vor, weil sie glauben, dass sie die Datenressourcen des Unternehmens besser verwalten können als andere Cloud-Anbieter. Leider sind in vielen Fällen wie diesem die Lehren aus einer solchen Strategie bitter. Um ein Rechenzentrum in Eigenregie zu betreiben, muss ein Unternehmen viele Experten mit unterschiedlichen Fähigkeiten beschäftigen, um einen 24/7-Service anbieten zu können:

- Systemarchitekten entwerfen die Lösung auf der Grundlage der Anforderungen.
- Operations Engineers ersetzen defekte Hardware.
- Netzwerktechniker bauen das Netzwerk auf, einschließlich Router und Verkabelung.
- Betriebssystemexperten installieren und konfigurieren Betriebssysteme.
- Facility Manager kümmern sich um Systeme wie Klimaanlagen und Brandschutz.
- Das Sicherheitspersonal sichert den Zugang zum Rechenzentrum gegen unbefugtes Eindringen.

Data-Science-Projekte können eine Menge Hardware-Ressourcen benötigen, und Fehler bei der Dimensionierung – egal ob CPU, GPU, RAM, Festplatten oder Netzwerk – können teuer werden, da sie zu Engpässen führen können.

Darüber hinaus erfordern einige Data-Science-Projekte teure GPUs. Sobald diese gekauft sind, kann die nächste Generation besserer GPUs verfügbar sein. Da eine Cloud ständig mit der neuesten Hardware aufgerüstet wird, gerät ein Nutzer von Cloud-Diensten nie ins Hintertreffen, wohingegen es nicht garantiert ist, dass On-Premises-Installationen regelmäßig aufgerüstet werden.

Betriebssysteme und ihre Treiber werden regelmäßig aktualisiert. Linux-Betriebssysteme beispielsweise sind mit moderneren Kerneln – der Kernfunktionalität jedes Betriebssystems – und Treibern oft für moderne Hardware optimiert und bieten schnellere Durchsätze bei Datenübertragungen und eine bessere Leistung durch spezielle Algorithmen. Die Hersteller von Betriebssystemen versuchen, ihre Software so lange wie möglich abwärtskompatibel zu halten. Das bedeutet, dass Software, die auf älteren Versionen eines Betriebssystems lief, auch auf neueren Versionen ausgeführt werden kann. Das Ziel der Softwarehersteller ist es, so viele Kunden wie möglich zu bedienen. Wenn die Programmierer die Software mit den modernsten Funktionen der neuesten Kernel und Treiber optimiert haben, kann es sein, dass viele Kunden diese Software nicht ausführen können, weil die Richtlinien die Installation der neuesten Versionen eines Betriebssystems verbieten. Dies ist wahrscheinlich, da die IT-Politik vieler Unternehmen eher konservativ ist und die Unternehmen ihre Server erst spät auf neuere Versionen aktualisieren, da sie die Anzahl der parallel gepflegten Betriebssystemversionen begrenzen wollen.

Cloud-Anbieter können die neuesten Treiber und Betriebssystem-Kernelversionen für ihre nativen Cloud-Dienste verwenden. Sie sind nicht aus Kompatibilitätsgründen an alte Betriebssystemkomponenten gebunden. Dadurch können sie die neuesten Funktionen nutzen und den Nutzern ein insgesamt besseres Erlebnis bieten.

Viele Experten sehen die Cloud nicht nur als eine neue Technologie, sondern vielmehr als ein alternatives Geschäftsmodell für die Beschaffung und Verwaltung von Hardware selbst.[1)] Es muss nicht unbedingt ein bestimmter Dienst sein wie eine Cloud-basierte Datenbank oder ein Cloud-basierter Dateispeicher, der neue Kunden anzieht. Ein Cloud-Anbieter gewinnt wahrscheinlich einen neuen Kunden, wenn ein Unternehmen beginnt, an den Vorteil des Mietens von IT-Ressourcen gegenüber deren Besitz zu glauben.

Statistiken zeigen, dass die meisten Unternehmen die Cloud nutzen oder irgendwann nutzen werden.[2)] Auch wenn es für einige Unternehmen immer noch besser sein mag, betriebliche Kernsysteme in lokalen Umgebungen abzuschirmen, bevorzugen die meisten Unternehmen bereits die Cloud für analytische Workloads. Darüber hinaus ist die Cloud auch ein perfektes Experimentierlabor für Data Scientists, da sie die benötigten Ressourcen dynamisch erzeugen und stilllegen können.

[1] *https://medium.com/@storjproject/there-is-no-cloud-it-s-just-someone-else-s-computer-6ecc37cdcfe5*

[2] *https://www.cloudzero.com/blog/cloud-computing-statistics/*

Jeder Cloud-Anbieter definiert den Begriff „Cloud“ etwas anders.[3][4][5] Einige Wertversprechen richten sich auch an unterschiedliche Zielgruppen. Buchhalter beispielsweise hören gerne, dass sie ihre Fixkosten senken können, da sie nicht viel Geld für den Kauf einer neuen Plattform ausgeben wollen. Ein Pay-as-you-go-Modell macht die Sache für sie einfacher. Technische Teams werden den Vorteil sehen, dass sie neue Lösungen schnell einsetzen und bei Bedarf skalieren können.

Bevor wir uns näher mit den funktionalen Diensten eines Cloud-Anbieters befassen, ist es wichtig, die Anzahl der Rechenzentren der Cloud-Anbieter weltweit als einen entscheidenden Faktor für die Leistungsfähigkeit eines Cloud-Anbieters hervorzuheben. Mit den Big 3 – Amazon, Google und Microsoft – können Kunden ihre Dienste global skalieren.[6][7][8] Vor allem für kleinere Cloud-Anbieter ist es schwer, mit ihnen zu konkurrieren, da sie nicht über die Ressourcen verfügen, um weltweit Rechenzentren aufzubauen.

Dieses Wertversprechen ist allgemein gehalten und kann auch auf andere Cloud-Anbieter angewendet werden, die möglicherweise ihre eigenen Varianten davon haben.

3.2.1 XaaS

Das Modell **„Everything-as-a-Service“ (XaaS)** skizziert verschiedene Szenarien und definiert, wer – der Kunde oder der Cloud-Anbieter – für was verantwortlich ist. Die vier bekanntesten XaaS-Modelle sind:

- Infrastructure-as-a-Service (IaaS)
- Platform-as-a-Service (PaaS)
- Software-as-a-Service (SaaS)
- Function-as-a-Service (FaaS)

Traditionelle On-Premises-IT bedeutet, dass Sie alles selbst machen, einschließlich des Baus Ihres Rechenzentrums mit Serverräumen, Kühlung, Racks und unterbrechungsfreier Stromversorgung (USV). Dazu gehören auch Themen im Zusammenhang mit dem Gebäudemanagement, z. B. Feueralarm und Zugangskontrolle zu den Serverräumen.

Colocation ermöglicht es den Kunden, das Facility Management auszulagern. Die Kunden mieten einen Serverraum, bauen ihre Hardware auf und sind für alle Themen im Zusammenhang mit der Computerhardware verantwortlich. Sie müssen nach wie vor spezialisierte Fachleute für alles von Netzwerken bis zu Betriebssystemen beschäftigen,

---

[3] *https://docs.aws.amazon.com/whitepapers/latest/aws-overview/six-advantages-of-cloud-computing.html*

[4] *https://microsoft.firstdistribution.com/wp-content/uploads/2021/08/Microsoft-Azure-Value-Proposition_2-1.pdf*

[5] *https://cloud.google.com/why-google-cloud*

[6] *https://aws.amazon.com/about-aws/global-infrastructure/regions_az/*

[7] *https://azure.microsoft.com/en-us/explore/global-infrastructure/geographies/#geographies*

[8] *https://cloud.google.com/about/locations*

4
Datenarchitektur

Zoltan C. Toth, Sean McIntyre

Fragen, die in diesem Kapitel beantwortet werden:

- Wie lässt sich das Datenreifegradmodell eines Unternehmens beschreiben?
- Was sind die wichtigsten Anforderungen an eine gut funktionierende Datenarchitektur?
- Welches sind die gängigsten Datei- und Speicherformate, die heute verwendet werden?
- Wie unterscheiden sich Data Warehouses, Data Lakes und Lakehouses in ihrer Funktionalität?
- Was sind die Vor- und Nachteile von Cloud-basierten gegenüber On-Premises-Architekturen?
- Wie können Lakehouses und Apache Spark eine skalierbare Plattform für die Datenanalyse bieten?

4.1 Übersicht

Sobald Unternehmen einige Anwendungsfälle für die Datenanalyse identifizieren, stellen sich mehrere Fragen:

- Sollten wir mit einem einfachen Tool wie Excel beginnen oder gleich eine vollwertige KI-Lösung entwickeln?
- Wollen wir unsere Infrastruktur selbst verwalten oder stattdessen Managed Services von Cloud-Anbietern kaufen?
- Sollten wir die Daten so verwenden, wie sie sind, z. B. als CSV-Dateien oder einfache Textdateien, oder müssen wir sie in einem Data Warehouse speichern, um anspruchsvollere Analysen zu ermöglichen?

- Reichen Ad-hoc- oder regelmäßige analytische Abfragen für unseren Anwendungsfall aus oder benötigen wir Echtzeitanalysen?
- Wie viele Daten haben wir? Wie gut müssen unsere analytischen Fähigkeiten skalierbar sein?
- Haben wir ein gutes Gespür dafür, welche Art von Antworten wir von der Datenanalytik erwarten?

Um datengesteuerte Spitzenleistungen zu erzielen, müssen analytische Systeme durch eine robuste, gut funktionierende und leistungsstarke Datenarchitektur unterstützt werden. Das wichtigste Merkmal einer guten Datenarchitektur ist, dass sie auf Ihre spezifischen Bedürfnisse zugeschnitten ist, damit Sie geschäftliche Fragen so effizient wie möglich beantworten können. In diesem Kapitel beschreiben wir die grundlegenden methodischen und technologischen Eckpfeiler solcher Architekturen, damit Sie die oben genannten Fragen beantworten können.

4.1.1 Maslowsche Bedürfnishierarchie für Daten

Bild 4.1 veranschaulicht die Maslowsche Bedürfnishierarchie, eine Idee aus der Psychologie, die von Abraham Maslow 1943 vorgeschlagen wurde. Diese Idee umreißt die Hierarchie der menschlichen Motivationen: Das Bedürfnis auf jeder Stufe muss im Individuum befriedigt werden, bevor die folgende Stufe angestrebt werden sollte oder wird.

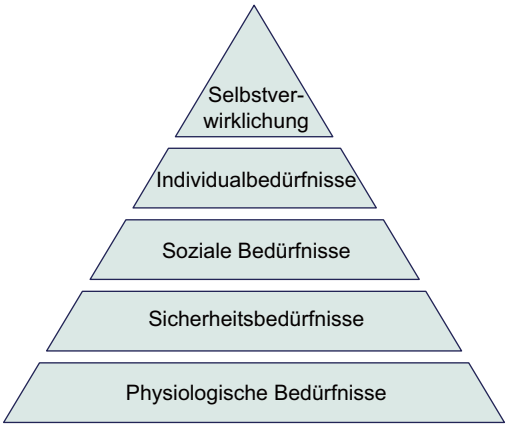

Bild 4.1
Maslowsche Hierarchie der Bedürfnisse

Dieses Konzept gilt nicht nur für die menschliche Psychologie, sondern lässt sich auch direkt auf die Reise eines Unternehmens in Richtung Data Science Readiness anwenden. Die Gliederung einer solchen Hierarchie der Datenbedürfnisse ist in folgender Abbildung dargestellt Bild 4.2.

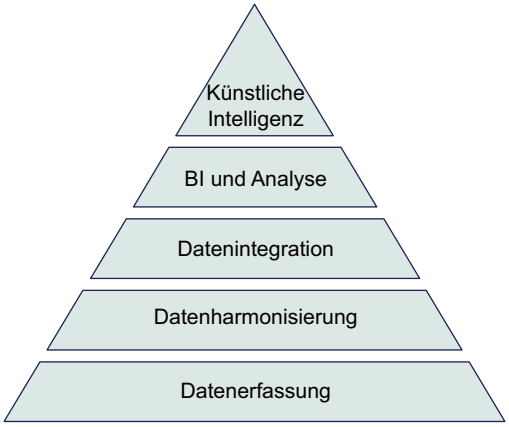

Bild 4.2
Hierarchie der Datenbedürfnisse

- Um sinnvolle Datenbereinigungs- und Integrationsabläufe zu erstellen, müssen Sie die Rohdatenerfassung richtig durchführen.
- Nur wenn saubere Daten in ein Data Warehouse oder einen Data Lake integriert und relevante Datenpunkte miteinander verknüpft werden, kann ein Unternehmen eine aussagekräftige Business-Intelligence (BI)- und Reporting-Ebene effizient implementieren.
- Das Geschäft muss durch Standard-BI und Datenanalyse verstanden werden, um den Raum für Machine Learning und fortgeschrittene analytische Anwendungsfälle zu öffnen.
- Die Forschung zeigt,[1)] dass KI-Anwendungen wie Large Language Models (LLMs) durch gut organisierte Daten verbessert werden.

Alles in allem gibt es keinen einfachen Weg zu einer exzellenten Datenanalytik. Für den Aufbau einer zuverlässigen Datenarchitektur, die einen maximalen geschäftlichen Nutzen bietet, müssen Sie die Grundlagen richtig anwenden. Im folgenden Abschnitt werden die grundlegenden Anforderungen für die Erstellung einer Datenarchitektur für Analyse-, Berichts- und ML-Anwendungen zusammengefasst.

4.1.2 Anforderungen an die Datenarchitektur

Eine gute Architektur ist konzeptionell einfach, leicht zu bedienen und zu ändern und erfüllt die drei Grundanforderungen: Verlässlichkeit, Skalierbarkeit und Wartbarkeit [1].

- Verlässlichkeit: Es wird erwartet, dass das System auch bei bestimmten Arten von Fehlern, einschließlich Hardware- und Softwarefehlern sowie menschlichem Versagen, weiterhin korrekt funktioniert. Während selbst das am besten konzipierte Sys-

[1] *https://arxiv.org/pdf/2311.07509.pdf*

tem nicht in der Lage sein wird, alle möglichen Arten von Fehlern zu bewältigen, werden zuverlässige Datenarchitekturen – zumindest bis zu einem gewissen Grad – in bestimmten vorhersehbaren Fehlerszenarien korrekt weiterarbeiten. Dazu gehören Netzwerkausfälle, Hardwarefehler, Systemneustarts, unerwartete Datenpunkte und fehlerhafter Datenbeladungs- oder Transformationscode.

- Skalierbarkeit: Selbst wenn eine Datenarchitektur heute gut funktioniert, bedeutet das nicht unbedingt, dass sie auch in Zukunft zuverlässig funktioniert. Unternehmen wachsen, und damit wächst auch die Menge der gesammelten und verarbeiteten Daten. In vielen Anwendungsfällen muss eine gut funktionierende Datenarchitektur auf ein zunehmendes Wachstum vorbereitet sein. Glücklicherweise sind die meisten Cloud-basierten Datendienste auf Skalierbarkeit ausgelegt. Auch lokale Anwendungsfälle können von skalierbaren Lösungen profitieren; die beliebtesten Open-Source-Lösungen sind hier das Hadoop Distributed File System (HDFS) oder Apache Spark.
- Wartbarkeit: Mit der Zeit wird die Komplexität Ihrer Datenarchitektur in datengesteuerten Unternehmen mit sich ständig ändernden Anforderungen wahrscheinlich zunehmen. Damit sie weiterhin gut funktioniert, müssen Sie erhebliche Anstrengungen unternehmen, um sie wartbar zu halten. Wenn sich die Datenbereitschaft eines Unternehmens verbessert, muss diese Anforderung in mehreren Bereichen berücksichtigt werden: Verschiedene Ingenieure müssen das System verstehen, Funktionen und Änderungen implementieren, Fehler beheben und die Architektur kontinuierlich betreiben.

4.1.3 Die Struktur einer typischen Datenarchitektur

Die meisten Datenarchitekturen haben ähnliche Arbeitsabläufe:

- Datenbeladung: Daten aus verschiedenen Quellen werden aufgenommen und in der Dateninfrastruktur als Dateien oder Cloud-Datei-ähnliche Objekte gespeichert. Zu diesen Quellen gehören betriebliche Protokolldateien, CSV-Dateien und andere Datendateien, Anwendungen von Drittanbietern und Datenbanken.
- Datenbereinigung und -standardisierung:[2)] Die eingelesenen Daten werden bereinigt; redundante Datensätze werden entfernt, und Datenqualitätsprobleme werden erkannt und behoben. Die bereinigten Daten werden an einer zentralen Stelle in einem standardisierten Format integriert, um sicherzustellen, dass Verbindungen zwischen verschiedenen Datenpunkten hergestellt werden können.
- Datentransformation: Die bereinigten Daten werden in Datensätze umgewandelt, die von den Geschäftsbereichen direkt genutzt werden können und als Grundlage

[2] In den Begriffen des Modern Data Stack wird Data Ingestion oft als Datenintegration bezeichnet.

6
Data Governance

Victoria Rugli, Mario Meir-Huber

„Wenn man Daten lange genug quält, gestehen sie alles."

Ronald H. Coase

Fragen, die in diesem Kapitel beantwortet werden:

- Warum ist Data Governance für Unternehmen wichtig?
- Wie sehen die drei Dimensionen der Data Governance – Menschen, Prozesse und Technologien – aus?
- Wie kann die Rolle des Change Management, der Data Stewards und der Data Governance Boards definiert werden?
- Was sind die fünf Säulen der Prozesse bei der Data Governance?
- Wie können Technologien wie Cloud-Lösungen und Open-Source-Tools die Bemühungen um Data Governance unterstützen?

6.1 Warum brauchen wir Data Governance?

In einer datengetriebenen Welt, in der die Informationswellen unaufhörlich fließen, stehen wir an der Schwelle zu einer außergewöhnlichen Ära. Das Potenzial von Data Science und künstlicher Intelligenz (KI), unser Leben zu revolutionieren, ist grenzenlos und verspricht beispiellose Fortschritte im Gesundheitswesen, im Bildungswesen, in der Industrie und in vielen anderen Bereichen. Doch es gibt eine bittere Ironie, die einen langen Schatten auf diese strahlende Zukunft wirft: die Vernachlässigung und Unterfinanzierung von Data Governance.

Data Governance ist im Grunde genommen der Hüter der digitalen Seele unserer Gesellschaft. Sie ist der unsichtbare Wächter, der die Integrität, Sicherheit und ethische Nutzung von wertvollen Daten bewacht, die die Motoren von Data Science und KI antrei-

ben. Ohne eine angemessene Data Governance segeln wir ohne Kompass durch tückische Gewässer und haben keine Möglichkeit, die raue See der datengetriebenen Entscheidungsfindung zu navigieren.

Stellen Sie sich einen Moment lang ein Schiff vor, das auf dem Meer treibt. Die Segel sind zerfleddert, die Mannschaft müde und verzweifelt. Dieses Schiff steht für die zahllosen Data-Science- und KI-Projekte, die sich ohne die notwendige Unterstützung durch eine solide Data Governance auf eine gefährliche Reise begeben. Die Folgen dieser Vernachlässigung sind verheerend, denn sie gefährden nicht nur die Projekte, sondern auch die Privatsphäre, die Sicherheit und das Vertrauen der Menschen. Die Vernachlässigung von Data Governance in Unternehmen auf der ganzen Welt führt zu falschen Entscheidungen oder zum Scheitern eines datengesteuerten Projekts.[1)]

Data Governance dient als Bollwerk gegen den Missbrauch von Daten. Sie führt nicht nur zu einer wertvollen Entscheidungsfindung in großen Organisationen, sondern schützt auch die sensiblen Informationen von Einzelpersonen vor Missbrauch. Ohne eine angemessene Finanzierung von Data Governance ist unsere Gesellschaft den Schrecken von Datenschutzverletzungen, Identitätsdiebstahl und der Aushöhlung der persönlichen Privatsphäre ausgesetzt.

Darüber hinaus ist Data Governance der Leuchtturm für Transparenz und Verantwortlichkeit in der Welt der Data Science und KI. Sie stellt sicher, dass Algorithmen fair und unvoreingenommen sind, dass Entscheidungen erklärbar sind und dass das Potenzial für Diskriminierung minimiert wird. Die Vernachlässigung der Finanzierung von Data Governance ebnet den Weg für die unkontrollierte Verbreitung von voreingenommenen Algorithmen, die zu ungerechten Ergebnissen führen und Ungleichheiten aufrechterhalten.

Data Governance ist der am meisten vernachlässigte Teil eines jeden Datenprojekts, obwohl die Datenqualität einen der größten, wenn nicht sogar den größten Einfluss auf die gelieferte Endqualität hat. Vielen Unternehmen fällt es schwer, die notwendigen Investitionen zu rechtfertigen. Wenn ein Unternehmen jedoch nicht von Anfang an eine Governance-Struktur einführt, steigen die Kosten für eine spätere Implementierung dramatisch an. Auf den nächsten Seiten geben wir einen Überblick über die verschiedenen Faktoren und Kosten, die zum Scheitern von Data-Governance-Implementierungen beitragen.

6.1.1 Beispiel 1: Mit Data Governance Klarheit schaffen

SteelMaster, ein in mehreren Ländern tätiges Stahlunternehmen, hatte mit dem Schritt zu einer datengesteuerten Organisation zu kämpfen. Unterschiedliche Software führte zu Datensilos in verschiedenen Abteilungen wie Produktion, Qualitäts-

[1] Brous, P.; Janssen, M.; Krans, R.: Data Governance as Success Factor for Data Science. Responsible Design, Implementation and Use of Information and Communication Technology. 2020 Mar 6;12066:431–42. doi: 10.1007/978-3-030-44999-5_36. PMCID: PMC7134294.

kontrolle, Logistik und Finanzen, was die Zusammenarbeit erschwerte. Ein Beispiel war die Interpretation des „Wärmeindex“, ein Begriff, der je nach Definition der einzelnen Abteilungen unterschiedliche Bedeutungen hatte. In einigen Abteilungen bezog sich der „Wärmeindex“ auf die Intensität des Stahlproduktionsprozesses, während er in anderen Abteilungen die klimatischen Bedingungen der Stahllagerung bezeichnete. Dies führte zu Verwirrung, Ineffizienz und widersprüchlichen Berichten. Dies ist ein häufiges Problem in großen Unternehmen, da die einzelnen Abteilungen KPIs je nach ihren Bedürfnissen unterschiedlich definieren. Dies ist meist darauf zurückzuführen, dass die Informationen nicht zentral weitergegeben werden.

Data Lineage: Die Einführung von Data Lineage trug dazu bei, den Weg der SteelMaster-Daten von ihrer Entstehung bis zu ihrer endgültigen Nutzung aufzuzeigen. Diese neu gewonnene Transparenz zeigte alle Vorgänge von der Datenerfassung und -umwandlung bis zur Nutzung auf und ermöglichte die Identifizierung und Lösung von Diskrepanzen, Ungenauigkeiten und Widersprüchen im Zusammenhang mit dem „Wärmeindex“. Die Abstammung der Daten spielte eine wichtige Rolle bei der Angleichung von Terminologien und Interpretationen.

Datenkatalog: SteelMaster führte auch einen Datenkatalog ein, der als zentrale Ablage für alle datenbezogenen Informationen diente. Dieser Katalog bot ein benutzerfreundliches, durchsuchbares Inventar von Datenbeständen, einschließlich einheitlicher Definitionen, Nutzungsrichtlinien und Eigentumsangaben in Bezug auf den „Wärmeindex“ und andere wichtige Geschäftsbegriffe. Das Ergebnis war eine konsistente Interpretation des „Wärmeindex“ über alle Abteilungen und Standorte hinweg, die Verwirrung minderte und ein gemeinsames Verständnis förderte. Wenn nun eine Abteilung spezifische Daten in Bezug auf den „Wärmeindex“ benötigte, konnte sie mühelos den Datenkatalog konsultieren, die Übereinstimmung mit anderen Abteilungsmetriken sicherstellen und auf die gemeinsame Wissensbasis verweisen.

Durch die Implementierung dieser Data-Governance-Tools wurde die Kommunikation verbessert, die Entscheidungsfindung auf der Grundlage vergleichbarer Werte präzisiert und die betriebliche Effizienz gesteigert.

6.1.2 Beispiel 2: Die (negativen) Auswirkungen einer mangelhaften Data Governance

ConnectY, ein internationales Telekommunikationsunternehmen, hatte anfangs Probleme mit der Datenqualität. Die Datenstrategie des Unternehmens basierte auf schnellen Geschäftsimplementierungen, was bedeutete, dass im Laufe der Jahre viele schnelle Korrekturen an den Datenmodellen und -strukturen vorgenommen worden waren. Das Ergebnis? Mehrere Datenmodelle, die sich gegenseitig widersprachen, sodass nur Top-Experten eine einzige Quelle der Wahrheit schaffen konnten. Dies führte zu verschiedenen Problemen, und die Geschwindigkeit der Implementierung nahm mit jedem Iterationszyklus ab.

Irgendwann beschloss die Geschäftsleitung, etwas dagegen zu unternehmen, da die anstehenden Probleme so drängend waren.

Als eines der Hauptprobleme wurde festgestellt, dass die KPIs in den einzelnen Abteilungen unterschiedlich definiert waren. Dieselben KPIs wurden nicht einheitlich berechnet, was zu unterschiedlichen Zahlen bei Kennzahlen führte, z. B.:

Kundenabwanderungsrate:

- Definition A: Berechnet als der Prozentsatz der Kunden, die ihr Abonnement in einem bestimmten Zeitraum gekündigt haben.
- Definition B: Berechnet als Prozentsatz der inaktiven SIM-Karten, für die in den letzten 30 Tagen keine Nutzungsaktivitäten stattgefunden haben.

Durchschnittlicher Umsatz pro Nutzer (Average Revenue Per User, ARPU):

- Definition A: Gesamteinnahmen geteilt durch die Gesamtzahl der aktiven Abonnenten.
- Definition B: Gesamtumsatz geteilt durch die Gesamtzahl der ausgegebenen SIM-Karten.

Durch die Einrichtung eines Data Governance Board wurde die Harmonisierung der wichtigsten KPIs erreicht. Dies führte zu einer besseren Steuerung des Unternehmens und zu einem gemeinsamen Verständnis der Unternehmensleistung. In einem weiteren Schritt wurden die Datenmodelle neugestaltet und ein einziges, gemeinsames Datenmodell erstellt. Dies führte zu schnelleren Entwicklungszeiten und Daten als Service. Mit „Data as a Service“ konnten die Mitarbeiter des Unternehmens Daten in einer Art Selbstbedienung abrufen.

6.2 Die Bausteine der Data Governance

Aus branchenspezifischen Governance-Techniken wie der Konzentration auf Datenqualität oder Datensicherheit und -schutz, die im Bankensektor stark verankert sind, haben sich verschiedene Ansätze zur Data Governance entwickelt. Der modernere und branchenneutrale Ansatz sieht Data Governance als ein Instrument zur Erleichterung der Datennutzung innerhalb einer Organisation.[2)]

Data Governance ist der Prozess zur Festlegung der Verantwortlichkeit für Datenbestände und datenbasierte Entscheidungen. Sie umfasst die Verwaltung der Verfügbarkeit, Integrität, Nutzbarkeit und Sicherheit von Daten innerhalb einer Organisation. Sie regelt die Prozesse, Richtlinien, Standards und Technologien sowie die verantwortlichen Personen, die die effektive Verwaltung von Daten sicherstellen.

[2] Madsen, Laura B. (2019): Disrupting Data Governance: A Call to Action.

10
Statistik – Grundlagen

Rania Wazir, Georg Langs, Annalisa Cadonna

"All models are approximations. Assumptions, whether implied or clearly stated, are never exactly true. All models are wrong, but some models are useful. So the question you need to ask is not 'Is the model true?' (it never is) but 'Is the model good enough for this particular application?"[1)]

Fragen, die in diesem Kapitel beantwortet werden:

- Wie können wir verschiedene Arten von Daten klassifizieren?
- Was ist der Unterschied zwischen Regression und Klassifikation?
- Was ist lineare Regression? Wie interpretiert man die Parameter in einem linearen Regressionsmodell?
- Was ist logistische Regression? Wie interpretiert man die Parameter in einem logistischen Regressionsmodell?
- Welche Leistungskennzahlen verwenden wir bei Regression und Klassifizierung?
- Was sind Kreuzvalidierung und Bootstrapping?

Wenn man sich zum ersten Mal mit Data Science beschäftigt, stellt sich oft die Frage: Was ist der Unterschied zwischen Statistik und maschinellem Lernen? Sie werden eine Vielzahl unterschiedlicher Antworten auf diese Frage hören, die sich in einer Aussage zusammenfassen lassen: Der Unterschied zwischen Statistik und maschinellem Lernen ist ihr Zweck.

Die Statistik konzentriert sich darauf, Rückschlüsse auf Beziehungen zwischen Variablen zu ziehen. Statistische Modelle ermöglichen es uns, neue Beobachtungen vorherzusagen, aber das ist nicht ihr Schwerpunkt. Der Zweck des maschinellen Lernens hingegen ist es, Vorhersagen über neue Beobachtungen so präzise wie möglich zu treffen.

[1] Box, G. E. P.; Luceño, A.; del Carmen Paniagua-Quiñones, M. (2009), Statistical Control By Monitoring and Adjustment.

Einfache Modelle wie die lineare Regression oder die logistische Regression werden häufig als statistische Modelle betrachtet. Dies ist insbesondere dann der Fall, wenn wir testen, ob ein Prädiktor einen signifikanten Einfluss auf die Antwort hat oder ob ein Modell besser ist als ein anderes. Statistische Modelle sind interpretierbar, d. h., wir können die Rolle der einzelnen Parameter im Modell verstehen. Bei der Verwendung von Modellen des maschinellen Lernens hingegen ist die Interpretation oft nicht von primärem Belang.

Denken wir an ein Geschäftsproblem, das definiert wurde, und für dessen Lösung alle erforderlichen Daten gesammelt und organisiert wurden. Data Scientists durchlaufen in der Regel drei Phasen, bevor sie entscheiden, welches Modell in einem solchen Zusammenhang in Produktion gehen soll.

1. In der ersten Phase werden die Daten untersucht. Wir stellen uns Fragen wie: Gibt es fehlende oder falsche Datenpunkte? Gibt es Ausreißer? Welche Art von Attributen gibt es, und wie viele davon?
2. In der zweiten Phase müssen die Daten aufbereitet werden. Wir stellen uns folgende Fragen: Wie sollen wir mit fehlenden Daten umgehen? Wie und woran erkennen wir fehlerhafte Daten? Wie gehen wir mit Ausreißern um? Sind sie für den Entscheidungsprozess wichtig, oder sollten sie entfernt werden? Je nachdem, welche Modelle wir anwenden wollen, müssen die Daten möglicherweise auch normalisiert werden.
3. In der dritten Phase schließlich wird mit verschiedenen statistischen und maschinellen Lernmodellen experimentiert.

Die drei oben beschriebenen Phasen werden in einem Data-Science-Projekt nicht nur einmal, sondern mehrmals umgesetzt werden. Warum ist dies der Fall? Wir könnten zum Beispiel feststellen, dass die Modellergebnisse stark von der Art und Weise abhängen, wie wir mit Ausreißern und fehlenden Werten umgehen, und beschließen, die Daten anders aufzubereiten. Oder wir stellen fest, dass einige Merkmale besser ausgearbeitet werden könnten, um dem Modell bessere Vorhersagen zu erlauben.

Dieses Kapitel ist wie folgt gegliedert. Abschnitt 10.1 beschreibt verschiedene Arten von Daten. In Abschnitt 10.2, Abschnitt 10.3 und Abschnitt 10.4 werden die einfache lineare Regression, die multiple lineare Regression und die logistische Regression behandelt. Schließlich werden in Abschnitt 10.5 einige Zugänge vorgestellt, um die Genauigkeit eines Modells zu bewerten.

10.1 Daten

Daten können je nach Anwendung und Erfassungsmechanismus in verschiedenen Formen vorliegen. Die erste Unterscheidung, die wir treffen können, ist die zwischen strukturierten und unstrukturierten Daten. Ein einfaches Beispiel für strukturierte Daten ist

eine Tabellenkalkulation: Jede Datenzeile steht für eine Beobachtung; jede Spalte für eine Variable. Unstrukturierte Daten sind dagegen zum Beispiel Text- oder Bilddaten. In der Statistik haben wir es hauptsächlich mit strukturierten Daten zu tun.

Bei strukturierten Daten ist es wichtig zu verstehen, mit welcher Art von Variablen wir es zu tun haben. Variablen können in quantitative und qualitative Variablen unterteilt werden:

- **Quantitative Variablen,** auch numerische Variablen genannt, können gemessen werden und nehmen numerische Werte an. Beispiele für quantitative Variablen sind Alter, Einkommen, Wert eines Grundstücks und Temperatur. Quantitative Variablen können diskret oder kontinuierlich sein. In der Regel handelt es sich um eine diskrete Variable, wenn ihr Wert die Frage „Wie viele?“ beantwortet, wie z. B. „Wie viele Schüler sind in einer Klasse?“ oder „Wie viele Zimmer hat ein Haus?“. Diskrete kontinuierliche Variablen können Dezimalzahlen in ihrem Wert haben. In der Regel hängt die Anzahl der Ziffern von der Genauigkeit des Messinstruments ab. Ein Thermometer kann zum Beispiel 37,1 °C anzeigen, ein anderes 37,12 °C. Ein Trick, um zu verstehen, ob eine Variable diskret oder kontinuierlich ist, besteht darin zu fragen, ob das Hinzufügen einer Dezimalstelle sinnvoll ist. Können wir sagen, dass wir 5,5 (lebende) Elefanten haben? Nein? Dann ist die Anzahl der Elefanten diskret.
- **Qualitative Variablen,** auch kategorisch genannt, nehmen Werte in einer begrenzten Anzahl von Kategorien an: Zufriedenheitsbewertungen in einer Umfrage („sehr“, „ausreichend“, „überhaupt nicht“), Altersgruppen oder die Marke eines gekauften Produkts sind nur einige Beispiele. Qualitative Variablen können ordinal oder nominal sein. Ordinale Variablen sind Variablen mit einer „logischen“, wohl verstandenen Reihenfolge. Betrachtet man zum Beispiel die Ergebnisse eines Wettbewerbs, so kann man sagen, dass ein zweiter Platz besser ist als ein dritter Platz. Bei den T-Shirt-Größen gibt es eine wohlverstandene Reihenfolge: S – M – L – XL. Nominale Variablen hingegen haben keine Ordnung. Die Augenfarbe ist ein Beispiel für eine nominale kategorische Variable. Man mag vielleicht braune Augen mehr mögen als grüne, aber wir können nicht sagen, dass braune Augen besser, höher oder größer sind als grüne Augen.

Abhängig von den verfügbaren Daten kann es zwei Arten von statistischen Methoden des maschinellen Lernens geben:

- Überwachte Lernmethoden werden verwendet, wenn wir sowohl die unabhängigen als auch die abhängigen Variablen beobachten können. Dies wird als „überwachtes Lernen“ bezeichnet. Wir können ein solches Modell weiter klassifizieren:
  - **Regressionsmodelle,** welche verwendet werden, wenn die abhängige Variable quantitativ ist und deren Ziel darin besteht, einen quantitativen Wert vorherzusagen.
  - **Klassifizierungsmodelle** welche verwendet werden, wenn die abhängige Variable qualitativ ist und deren Ziel darin besteht, eine Klasse vorherzusagen.

- Unüberwachte Lernmethoden werden verwendet, wenn wir keine abhängige Variable haben oder wenn diese Variablen für unsere Daten einfach nicht verfügbar sind. Die beliebteste Methode in diesem Zusammenhang ist das Clustering. Clustering-Algorithmen suchen nach Ähnlichkeitsgruppen in einem Datensatz, ohne genau zu wissen, was die zugrunde liegenden „wahren" Gruppen sind. Dies ist z. B. bei einem Marktsegmentierungsproblem nützlich, bei dem wir Kunden gruppieren und nach ähnlichen Verhaltensmustern einstufen wollen, ohne dass wir im Voraus eine Kategorisierung vornehmen können. Wenn Sie ein Beispiel sehen möchten, finden Sie in Abschnitt 19.4.2 eine Erläuterung zum Clustering von Dokumenten mithilfe des K-Means-Algorithmus.

10.2 Einfache lineare Regression

Viele in Data Science verwendete Methoden sind Verallgemeinerungen oder Erweiterungen der linearen Regression. Da die lineare Regression eine der einfachsten und am weitesten verbreiteten Methoden ist, bietet sie uns einen schnellen Einstieg in statistische Modelle.

Beginnen wir mit der Untersuchung der Beziehung zwischen zwei Variablen. Eine davon, X, ist unabhängig, und wir nennen sie auch „Regressor" oder „Einflussgröße", die andere, Y, ist vom Regressor abhängig, und wir nennen sie „Regressand" oder „Zielgröße". Zunächst nehmen wir vereinfachend an, dass diese Beziehung linear ist.

Ein einfaches lineares Regressionsmodell sieht wie folgt aus:

$$Y = mX + n + e \tag{10.1}$$

wobei m die „Steigung" genannt wird, n als „Achsenabschnitt" bezeichnet wird und e ein Fehlerterm ist, der bei null zentriert ist und dessen Varianz nicht abhängt von X. Die Einführung des Fehlerterms bedeutet, dass wir davon ausgehen, dass unsere Daten nicht einer perfekten geraden Linie folgen, sondern um sie gestreut sind.

Wie lassen sich die Parameter Achsenabschnitt und Steigung interpretieren?

Der Achsenabschnitt ist der erwartete Wert der abhängigen Variablen, wenn die unabhängige Variable null ist. Wenn der Wert der Steigung m genau null ist (d. h., X hat keinen Einfluss auf Y), dann bliebe uns $Y = n + e$, was bedeutet, dass alle Datenpunkte um den Achsenabschnitt herum verteilt sind. Der Steigungsparameter hat folgende Bedeutung: Wenn wir den Wert X um eine Einheit ändern, würden wir eine Änderung von Y um m Einheiten erwarten. Diese Art der Änderung wird als „absolute Änderung" bezeichnet. Die Auswirkung von X auf den Erwartungswert von Y ist linear, und er ist positiv, wenn m positiv ist, und negativ, wenn m negativ ist. Man beachte, dass der Erwartungswert von Y unter der Bedingung von X gleich $mX + n$ ist: Dies ist die Regressionsgerade, die wir zu schätzen versuchen. Mit anderen Worten, wir wol-

21
Modellierung und Simulation – Erstellen Sie Ihre eigenen Modelle

Günther Zauner, Wolfgang Weidinger

„Verlieben Sie sich nicht in Ihr Simulationsmodell.“

F. Breitenecker

„Alle Modelle sind falsch, aber einige sind nützlich.“

F. Breitenecker

Fragen, die in diesem Kapitel beantwortet werden:

- Was sind die grundlegenden Methoden der Modellierung und Simulation?
- Wo kann klassische Data Science mit Modellierung und Simulation kombiniert werden?
- Wie werden Modellierung und Simulation in verschiedenen Bereichen eingesetzt, z. B. im Verkehrswesen, bei der modernen Simulation von Lagerbestandsmaßnahmen und bei Modellen für Infektionskrankheiten und Pandemiesimulationen?
- Warum ist es so wichtig, die Methoden in Abhängigkeit von der Problem- oder Fragestellung zu wählen und nicht umgekehrt?

Die Zielgruppen dieses Kapitels sind:

- Personen mit einem starken Interesse an Data-Science-Anwendungen in dynamischen Systemen,
- Personen, die Interesse daran haben, die Methoden der dynamischen Modellierung kennenzulernen,
- Menschen mit realen Fragen zu Strategien für die Modellierung der Ausbreitung von Infektionskrankheiten und
- Personen, die an Strategieevaluierung interessiert sind.

21.1 Einführung

Ziel dieses Kapitels ist es, die Standards in der Modellierung und Simulation zu beschreiben mit besonderem Schwerpunkt auf den verschiedenen Modellierungsmethoden und ihrer Anwendung. Diese werden anhand einer Reihe von Anwendungsbeispielen veranschaulicht, von der Modellierung von Infektionskrankheiten (Covid-19) und der Verkehrssimulation, die sowohl die Modellkalibrierung als auch die Simulation diskreter Prozesse beleuchtet, bis hin zur Simulation der Lagerhaltungsstrategie. Um zu zeigen, wie Data Science in den Modellierungsprozess und in die Interpretation der Ergebnisse integriert wird, beginnen wir mit einem Überblick über einen Modellierungsprozess im Allgemeinen. Dann werden wir kurz verschiedene Modellierungsmethoden und ihre Vor- und Nachteile beschreiben. In den folgenden Abschnitten wird erläutert, wie ein Modell von der Parametrisierung und Kalibrierung über die Verifizierung und Validierung bis hin zu den Simulationsexperimenten und Szenarien, die Ergebnisse liefern, zu handhaben ist.

Neben der Erstellung des Modells ist auch die Ausführung von Simulationsmodellen von wesentlicher Bedeutung. Eine Simulation führt das Modell mit einer definierten Parametrisierung aus und ermöglicht es Ihnen, die Logik Ihres Verhaltensmodells zu validieren. Die Analyse der Simulationsergebnisse, ihre grafische Interpretation und die klassische Statistik sind Teil der Realisierung eines Modellierungs- und Simulationsprojekts. Dies gilt auch für die Erklärung der Simulationsergebnisse auf der Grundlage einer (hohen) Anzahl von Simulationsläufen eines Modells mit stochastischen Parametern.

Einschränkungen in Bezug auf Methoden im Fokus

Alle Systeme, sowohl natürliche als auch vom Menschen geschaffene, sind dynamisch in dem Sinne, dass sie in der realen Welt existieren, die sich mit der Zeit weiterentwickelt. Mathematische Modelle solcher Systeme würden natürlich als dynamisch angesehen werden, da sie sich im Laufe der Zeit entwickeln und daher die Zeit einbeziehen. Oft ist es jedoch sinnvoll, eine Annäherung vorzunehmen und die Zeitabhängigkeit in einem System zu ignorieren. Ein solches Systemmodell wird als „statisch“ bezeichnet.

Das Konzept eines Modells kann als dynamisch bezeichnet werden, wenn es eine kontinuierliche zeitabhängige Komponente enthält. Das Wort „dynamisch“ leitet sich von dem griechischen Wort *dynamis* ab, das „Kraft“ und „Macht“ bedeutet, wobei Dynamik das zeitabhängige Zusammenspiel von Kräften ist. Die Zeit kann explizit als Variable in eine mathematische Formel einfließen oder indirekt vorhanden sein, z. B. als Zeitableitung oder als Ereignisse, die zu bestimmten Zeitpunkten auftreten. Im Gegensatz dazu werden statische Modelle ohne Einbeziehung der Zeit definiert. Statische Modelle werden häufig zur Beschreibung von Systemen in stabilen Zuständen oder Gleichgewichtszuständen verwendet.

In diesem folgenden Kapitel liegt der Schwerpunkt ausschließlich auf der dynamischen Modellierung von Systemen. Statische Modelle basieren häufig auf der klassischen Statistik und werden daher hier nicht behandelt. Auch das wissenschaftliche Gebiet der partiellen Differentialgleichungen (PDEs) sprengt den Rahmen dieses Buches, da die verschiedenen PDEs zu physikalischen Bereichen gehören, in denen ein tiefgreifendes theoretisches Verständnis erforderlich ist, um die Methoden zu verstehen. Dennoch werden hier eine kurze Zusammenfassung und einige grundlegende Hinweise gegeben [1 – 5].

Eine partielle Differentialgleichung (PDE) ist eine mathematische Gleichung, die mehrere unabhängige Variablen, eine unbekannte Funktion, die von diesen Variablen abhängt, und partielle Ableitungen der unbekannten Funktion in Bezug auf die unabhängigen Variablen umfasst. PDEs werden häufig verwendet, um das zeitliche Verhalten von mehrdimensionalen Systemen in der Physik und im Ingenieurwesen zu beschreiben. Es gibt aber auch Anwendungen im Finanzwesen und bei Marktanalysen.

Einige PDEs haben exakte Lösungen, aber im Allgemeinen sind numerische Näherungslösungen erforderlich. Welche numerische Standardmethode zu verwenden ist, hängt stark von dem zugrunde liegenden beschriebenen System ab, das durch eine PDE definiert ist, sowie von dem Bereich und dem erforderlichen Detailgrad der Forschungsfrage. Bei der Finite-Differenzen-Methode beispielsweise werden die Ableitungen in der PDE angenähert, und anschließend wird die unbekannte Funktion bei jedem dieser Werte unter Verwendung einer großen Anzahl von inkrementellen Werten der unabhängigen Variablen berechnet.

Die Finite-Differenzen-Methode wird oft als die am einfachsten zu erlernende und anzuwendende Methode angesehen. Die Finite-Elemente- und die Finite-Volumen-Methode sind in der Elektrotechnik und der Flüssigkeitssimulation weit verbreitet. Mehrgittermethoden sind ebenfalls eine Standardmethode in der Anwendung. Im Allgemeinen kann die PDE-Modellierung und -Simulation als ein separates Arbeitsgebiet und eine wissenschaftliche Disziplin betrachtet werden, die nicht Teil dieses Kapitels ist.

21.2 Allgemeine Aspekte

Wenn ein Problem auftritt, stellt sich in der Regel die Frage nach möglichen Lösungen oder Methoden für die Bewertung. Informationen werden gesammelt, analysiert, und zusammen mit der jeweiligen Erfahrung wird eine Entscheidung getroffen.

Modelle sind eine dieser möglichen Lösungen und können besonders geeignet sein, wenn es zu wenig Hinweise und Daten über das mögliche Ergebnis der zu lösenden Aufgabe gibt. Das Problem wird dann in eine abstrakte Vereinfachung des realen Systems übersetzt. Der gesamte Modellierungsprozess befasst sich mit dieser Formalisierung und Abstraktion des Problems sowie mit dem Ziehen von Schlussfolgerungen daraus für das ursprüngliche System in einer notwendigen und ausreichenden Weise.

Ein mathematisches Modell eines Systems ist eine symbolische Beschreibung unter Verwendung einer abstrakten Formulierung. Es verwendet mathematische Symbole und ist ohne korrekte Interpretation nutzlos. Bei der Manipulation von Symbolen werden ausschließlich mathematische Gesetze verwendet. Eine mathematische Formel bestätigt den Status eines Modells, wenn Symbole und Modellvariablen in Beziehung zueinander gesetzt werden. Ein Modell ist dann geeignet, wenn es die Fragen beantwortet oder in geeigneter Weise zur Lösung vorbereitet [6].

Was die Modellierung betrifft, so enthält die Außenwelt drei verschiedene Arten von „Dingen“:

- vernachlässigte Dinge
- Dinge, die das Modell beeinflussen, die aber nicht mit dem Modell untersucht werden sollen (exogene Werte)
- Dinge, die der Grund für die Erstellung des Modells sind (endogene Werte)

Die Unterscheidung zwischen exogenen und endogenen Werten hängt von der Betrachtungsweise ab. So wird beispielsweise ein Modell, das die Wirkungsweise eines bestimmten Medikaments im menschlichen Körper erklären soll, eine andere Struktur haben als ein Modell zur Untersuchung der Kosteneffizienz desselben Medikaments.

21.3 Modellierung zur Beantwortung von Fragen

Um sich dem Hauptnutzen zu nähern, der durch Modellierung und Simulation gewonnen werden kann, muss eine Formalisierung der Forschungsfrage definiert werden. Die folgende Struktur bietet einen Leitfaden:

1. Formulierung der Problemstellung: Welche Fragen sollen beantwortet werden?
2. Konzept des Modells:
   - Welche Werte sind wichtig?
   - Welche Werte beschreiben die Zustände des Modells?
   - Welches sind die Parameter?
   - Welche Werte beeinflussen das Modell im Allgemeinen?
   - Welche Beziehungen bestehen zwischen den Werten?
3. Ist das Modellkonzept nützlich? Gibt es genügend Wissen und Daten, um das Modell umzusetzen? Können die vorgeschlagenen Fragen mithilfe des vorgeschlagenen Modells beantwortet werden, wenn die Modellannahmen zutreffen?
4. Kann das Modell validiert werden?

Zunächst sollte die Verfügbarkeit von Daten zunächst keinen Einfluss auf das Modellkonzept haben. Erst wenn der Mangel an Informationen eine Umsetzung unmöglich macht,

sollte das Konzept angepasst werden. Das Modellkonzept schlägt oft eine geeignete Modellierungstechnik vor, aber letztendlich wählt der Modellierer sie aus.

Die Modellierer spielen in dem gesamten Prozess eine wichtige Rolle. Sie übersetzen das modellierte Objekt in das abstrakte Modell und vermitteln die Modelleigenschaften an andere. Daher können ihre Überzeugungen und ihr Vorwissen das Modell und die Interpretation der Ergebnisse beeinflussen [7]. Jedes Modell ist nur ein idealisiertes Abbild der Realität. Viele äußere und innere Einflüsse müssen vernachlässigt werden, um das Modell handhabbar zu machen. Es ist wichtig, jede Vereinfachung im Modellkonzept zu notieren und zu begründen, auch wenn es nicht möglich ist, einen wissenschaftlichen Beweis dafür zu erbringen, dass bestimmte Vereinfachungen keinen gravierenden Einfluss auf das Modellverhalten haben [8].

Die Übersetzung eines Problems in eine abstrakte mathematische Sprache besteht aus mehreren Schritten, die in Bild 21.1 dargestellt sind.

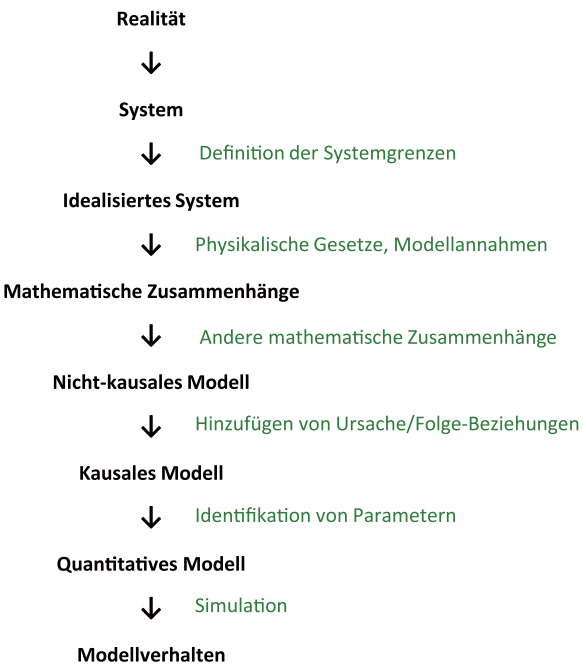

Bild 21.1
Übersetzung eines Problems in eine abstrakte mathematische Sprache

Wenn einer der in Bild 21.1 dargestellten Schritte nicht durchgeführt werden kann oder nicht die gewünschten Ergebnisse liefert, muss der Modellierer einen oder mehrere Schritte zurückgehen, um das Modell zu überarbeiten. Alle aufeinanderfolgenden Schritte müssen erneut durchgeführt werden. Modellierung ist ein iterativer Prozess [9], der in Bild 21.2 dargestellt ist.

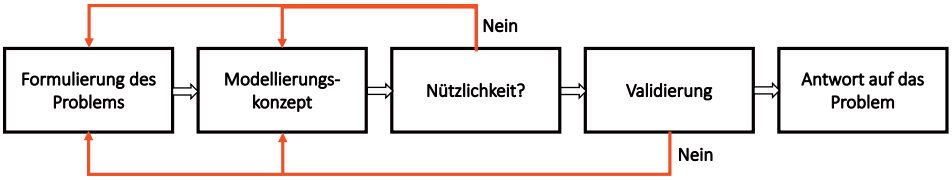

Bild 21.2 Iterativer Zyklus der Modellierung

21.4 Reproduzierbarkeit und Lebenszyklus des Modells

Wenn ein Simulationsmodell reproduzierbar ist, sollte es von einem anderen Modellierungsexperten erstellt werden können, der das Experiment unter Verwendung derselben experimentellen Daten und Methoden, unter denselben Betriebszuständen, in derselben oder einer anderen Umgebung und in mehreren Versuchen wiederholt. Die Reproduzierbarkeit ist einer der wichtigsten Werte für Modelle, die in großen Projekten eingesetzt werden. Während der Projektlebenszyklus als Ausgangspunkt dient, sind Parameter- und Ergebnisdefinition, Dokumentation, Verifizierung und Validierung allesamt Aspekte von großer Bedeutung. Um die Reproduzierbarkeit und damit auch die Glaubwürdigkeit eines Modells zu verbessern, kann eine Vielzahl von Aufgaben durchgeführt werden.

Ein grundlegendes Prinzip wissenschaftlicher Arbeit ist, dass Wissen transparent sein sollte, was bedeutet, dass die Verfügbarkeit für einen professionellen Diskurs gegeben sein sollte. Für die Reproduzierbarkeit ist es notwendig, sich darauf zu konzentrieren, wie man Beiträge von anderen Forschern erhält, was eine Erklärung über die Grenzen und Annahmen innerhalb eines Modells erfordert. Mögliche Unzulänglichkeiten werden aufgedeckt, und Annahmen können infrage gestellt werden. Neben all den anspruchsvollen und möglicherweise kostenintensiven Aufgaben, die mit dem Erreichen der Reproduzierbarkeit verbunden sind, sollte man sich den Nutzen dieser Bemühungen vor Augen halten. Besonderes Augenmerk muss auf die Dokumentation, Visualisierung, Parameterformulierung, Datenaufbereitung [10], Verifizierung und Validierung gelegt werden. Detaillierte Informationen zu diesem Prozess finden sich in der Arbeit von Popper [11].

Das Verständnis des Lebenszyklus des Entwicklungsprozesses hinter einem Modellierungs- und Simulationsprojekt ist für die Diskussion über die Reproduzierbarkeit von wesentlicher Bedeutung. Der Grund dafür ist, dass man genau wissen muss, welche Informationen in welcher Phase des Entwicklungsprozesses erzeugt werden. In Verbindung mit der Formulierung von Parametern, der Dokumentation und der Verifizierung sowie der Validierung ist das Verständnis des Lebenszyklus von entscheidender Bedeutung, um zuverlässige und verwertbare Ergebnisse zu erzielen. Nur so können Erkennt-

23
Datengetriebene Unternehmen

Mario Meir-Huber, Stefan Papp

„Information is the oil of the 21st century, and analytics is the combustion engine."

Peter Sondergaard, SVP Gartner

Fragen, die in diesem Kapitel beantwortet werden:

- Welche strategischen Entscheidungen sind für eine datengestützte Arbeit unerlässlich?
- Was ist eine Datenstrategie, und wie kann man sie entwickeln?
- Wie bauen Sie ein Datenteam auf? Zentralisiert oder dezentralisiert?

„Daten sind der Rohstoff des 21. Jahrhunderts" ist eine gängige Aussage in den Chefetagen großer Unternehmen. Viele CEOs schenken diesem Thema erhöhte Aufmerksamkeit. Unternehmen wiederum, die den Trend zur Wertschöpfung aus Daten durch analytische Prozesse ignorieren, gefährden ihre Existenz.

Viele Unternehmen stehen daher vor der Frage, wie sie eine nachhaltige unternehmensweite Datenstrategie aufbauen können. Diese neue Ausrichtung kann manchmal bestehende Geschäftsmodelle infrage stellen. Manchmal bedeutet es sogar, dass die Datenstrategie ein Unternehmen grundlegend verändern kann und völlig neue Geschäftsmodelle erfordert.

Für viele Unternehmen ist der Einstieg nicht einfach, da sie nicht über die Erfahrung von Google oder Facebook verfügen, die seit vielen Jahren Petabytes an Daten auswerten. Dieses Kapitel beschreibt aus unternehmerischer Sicht, wie ein Unternehmen eine nachhaltige Datenstrategie aufbauen kann.

In diesem Kapitel wird ein Modell vorgestellt, das aus den drei Bereichen „Technologie", „Kultur" und „Business" besteht.

23.1 Die drei Ebenen eines datengesteuerten Unternehmens

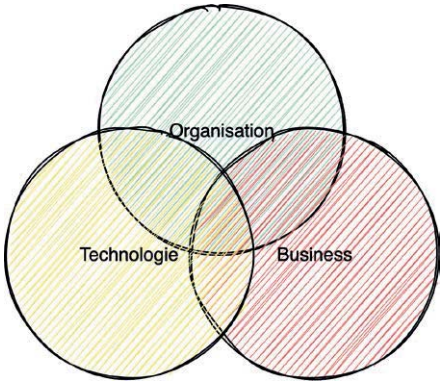

Bild 23.1
Der Schnittpunkt zwischen Business, Technologie und Organisation (Kultur)

Daten im Unternehmenskontext sind multidisziplinär. Es handelt sich nicht nur um eine technische, sondern auch um eine organisatorische und kulturelle Herausforderung. Um die Datennutzung in einem Unternehmen zu steigern, müssen die Verantwortlichen die Unternehmensrichtlinien und -kultur ändern. Dies wird in Abschnitt 23.2, „Kultur“, näher erläutert.

In Abschnitt 23.3, „Technologie“, werden die wirtschaftlichen Aspekte von Datenplattformen ausführlicher erläutert. Technische Aspekte werden hier nicht behandelt, da sich im Prinzip das gesamte Buch mit technischen Aspekten befasst.

In Abschnitt 23.4, „Business“, werden bestimmte Grundvoraussetzungen für die Durchführung datengesteuerter Projekte erörtert. Spezifische Anwendungsfälle werden hier nicht erwähnt, da Sie hierzu eine umfassende Darstellung in Kapitel 24 finden, das nach Branchen und Bereichen gegliedert ist und in dem konkrete Anwendungsfälle gezeigt werden.

23.2 Kultur

> *„Jede Strategie dauert bis zum ersten Kontakt mit dem Feind. Danach gibt es nur noch ein System von Stellvertretern.“*
>
> Helmuth Graf von Moltke

Ein zentraler Aspekt der Unternehmensstrategie ist die Organisation der Unternehmenseinheiten. Die folgenden Abschnitte konzentrieren sich auf die Kernaspekte:

1. Unternehmensstrategie für Daten
2. Kultur und Organisation

Im ersten Abschnitt, Unternehmensstrategie für Daten, geht es in erster Linie um das Reifegradmodell und die Erstellung einer Datenstrategie. Der zweite Abschnitt, Kultur und Organisation, befasst sich mit der Unternehmensentwicklung. Der wesentliche Punkt dabei ist, dass die beiden Aspekte nicht iterativ durchgeführt werden müssen, sondern parallel und gleichzeitig laufen können.

23.2.1 Unternehmensstrategie für Daten

Jede Wanderung beginnt mit einer Positionsbestimmung. Wenn man einen Berg besteigen will, muss man sicherstellen, dass man geeignete Wanderkarten (in digitaler oder nichtdigitaler Form) dabeihat und seine Position kennt. Ähnlich verhält es sich, wenn man eine Datenstrategie im Unternehmen aufstellen will, dann muss man zunächst den Ist-Zustand kennen. Deshalb wird bei der Strategieentwicklung in der Regel eine Ist-Analyse durchgeführt. Mithilfe dieser Analyse ist es möglich, den Reifegrad der eigenen Organisation zu bestimmen. Im Folgenden werden die vier Reifegradphasen vorgestellt.

Unternehmerische Datenreife

Die Ermittlung des Reifegrades des eigenen Unternehmens ist unerlässlich, um sinnvolle Maßnahmen zur Steigerung der analytischen Kompetenz abzuleiten. Zur Bestimmung der Datenreife eines Unternehmens sind die folgenden vier Phasen vorgesehen:

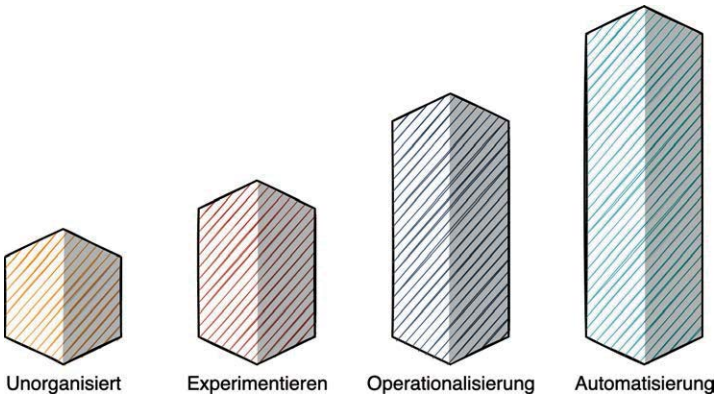

Bild 23.2 Das Reifegradmodell

Phase 1: Unorganisiert

Streng genommen analysiert ein Unternehmen bereits Daten, wenn es Informationen in Excel-Tabellen speichert und diese Daten als Grundlage für Diskussionen in Besprechungen verwendet. Wir wollen hier aber nicht die Informationen in Tabellenkalkulationen als analytischen Prozess betrachten.