Data Privacy in der Praxis

Datenschutz und Sicherheit in Daten- und KI-Projekten

KAPITEL 1

Data Governance und einfache Datenschutzansätze

Privacy ist ein großes und langlebiges Feld. Stellen Sie es sich wie eine alte Straße vor – voller interessanter Seitenstraßen und Abzweigungen, auf der man sich aber nur schwer zurechtfindet, wenn man den Weg nicht kennt. Dieses Kapitel soll Ihnen eine erste Orientierung auf dieser Straße bieten. In diesem Kapitel und im gesamten Buch helfe ich Ihnen, wichtige Abschnitte der Datenschutzlandschaft zu kartieren, und Sie werden Bereiche finden, in denen Sie mehr erfahren und vom ursprünglichen Weg abweichen wollen. Diese Landkarte in Ihrem Unternehmen anzuwenden, heißt, herausfinden, wer was tut, wer wofür verantwortlich ist und welche Anforderungen hinsichtlich des Datenschutzes in Ihrem Unternehmen bestehen.[1]

Den Begriff *Data Governance* (im engeren Sinne im Deutschen auch bekannt als Datenqualitätsmanagement) haben Sie vielleicht schon einmal oder auch Hunderte Male gehört, aber oft wird er nicht näher erläutert oder lässt einen gewissen Interpretationsspielraum zu. In diesem Kapitel erfahren Sie, wo sich Data Governance und Datenschutz für praktische Zwecke der Data Science überschneiden, und lernen relativ simple Ansätze zur Lösung von Datenschutzproblemen im Zusammenhang mit der Arbeit mit Daten kennen, wie z. B. die Pseudonymisierung. Darüber hinaus erfahren Sie, wie Data-Governance-Methoden wie die Dokumentation und Nachverfolgung der Datenhistorie – das sogenannte Data-Lineage-Tracking – dabei helfen können, Datenschutzprobleme zu identifizieren oder Methoden zur Implementierung des Datenschutzes zum richtigen Zeitpunkt zu implementieren.

1 In der deutschen Übersetzung des englischsprachigen Originals ist meist von *Unternehmen* die Rede, um Ihren Arbeitsplatz zu beschreiben. Je nachdem, ob Sie in einem kleinen agilen Data-Science-Beratungsunternehmen, einem großen Konzern, einer staatlichen Behörde bzw. Einrichtung oder auch in einer mittelgroßen gemeinnützigen Organisation arbeiten, sollen Sie sich gleichermaßen angesprochen fühlen und werden sicherlich ganz unterschiedliche Erfahrungen machen. Dieses Buch soll für alle Zielgruppen von Nutzen sein. Übernehmen Sie die Ratschläge und Erkenntnisse, wenden Sie Ihr eigenes Wissen über Ihre Tätigkeit an und bringen Sie sie mit Ihrer Unternehmensgröße und -kultur in Einklang.

Wenn Sie bereits mit Data Governance vertraut sind oder in diesem Bereich arbeiten, empfehle ich Ihnen, dieses Kapitel nur zu überfliegen oder gänzlich zu überspringen. Sind Ihnen Governance und Datenmanagement jedoch noch nicht geläufig, werden in diesem Kapitel die Grundlagen gelegt, die Sie benötigen, um die fortgeschrittenen Methoden anzuwenden, die Sie in den späteren Kapiteln kennenlernen werden.

In diesem Kapitel werden Werkzeuge und Systeme vorgestellt, mit denen Sie sensible Daten identifizieren, nachverfolgen und verwalten können. Ohne diese Grundlage wird es schwierig sein, Datenschutzrisiken zu bewerten und entsprechende Bedenken auszuräumen. Es ist sinnvoll, mit der Governance zu beginnen, da sich der Datenschutz gut in Governance-Rahmen und -Paradigmen einfügt und diese Arbeitsbereiche in Datensystemen ineinandergreifen.

Data Governance: Was ist das?

Der Begriff *Data Governance* wird oft als »allumfassende« Art und Weise verwendet, wie wir über unsere Entscheidungen in Bezug auf Daten nachdenken, z. B. ob Sie einem Dienst erlauben, Sie zu kontaktieren, oder zu entscheiden, wer Zugriffsrechte auf eine bestimmte Datenbank hat. Doch was bedeutet der Begriff nun wirklich, und wie können Sie ihn in die Praxis umsetzen?

Data Governance bedeutet im wörtlichen Sinne, Daten zu »regieren«. Governance kann einerseits durch die Übertragung von Rechten erfolgen, die Menschen individuell und kollektiv besitzen. Diese Rechte werden an Bevollmächtigte übertragen, die Aufgaben und Verantwortlichkeiten für Personen übernehmen, die dafür nicht die Zeit, das Fachwissen oder das Interesse haben. Bei der Data Governance überträgt der Einzelne Rechte, wenn er Daten an ein Unternehmen oder eine Organisation weitergibt. Mutzen Sie eine Webseite, einen Dienst oder eine Anwendung, erklären Sie sich mit den Datenschutzbestimmungen, -bedingungen und -vereinbarungen einverstanden, die Ihnen zu diesem Zeitpunkt von diesen Datenverarbeitern (oder denjenigen, die die Daten erheben) auferlegt werden. Das ist vergleichbar mit der Tatsache, dass Sie in einem bestimmten Land leben und damit implizit zustimmen, sich an die dort geltenden Gesetze zu halten.

Data Governance hilft Ihnen dabei, zu steuern, wessen Daten Sie erheben (bzw. sammeln), wie Sie sie erheben und anreichern und was Sie anschließend mit den erhobenen Daten machen. Abbildung 1-1 veranschaulicht, in welchem Zusammenhang Datenschutz und Sicherheit mit Data Governance stehen – und zwar anhand einer imaginären Insel, auf der die Nutzer und ihre Daten sowohl durch Datenschutz- als auch durch Sicherheitsinitiativen (engl. *Security Initiatives*) angemessen geschützt sind. In der Illustration sind die sensiblen Daten in einem Turm platziert. Sicherheitsinitiativen werden durch Privacy by Design unterstützt.[2] Regulierung und Compliance bilden eine Art Graben, durch den die sensiblen Daten abgetrennt

werden. Die Datenschutztechnologien, die Sie in diesem Buch kennenlernen werden, dienen als Brücke für die Nutzenden und die an den Daten Interessierten (*Data Stakeholders*), damit diese aus sensiblen Daten Erkenntnisse gewinnen und Entscheidungen treffen können, ohne die Privatsphäre einer oder eines Einzelnen zu verletzen.

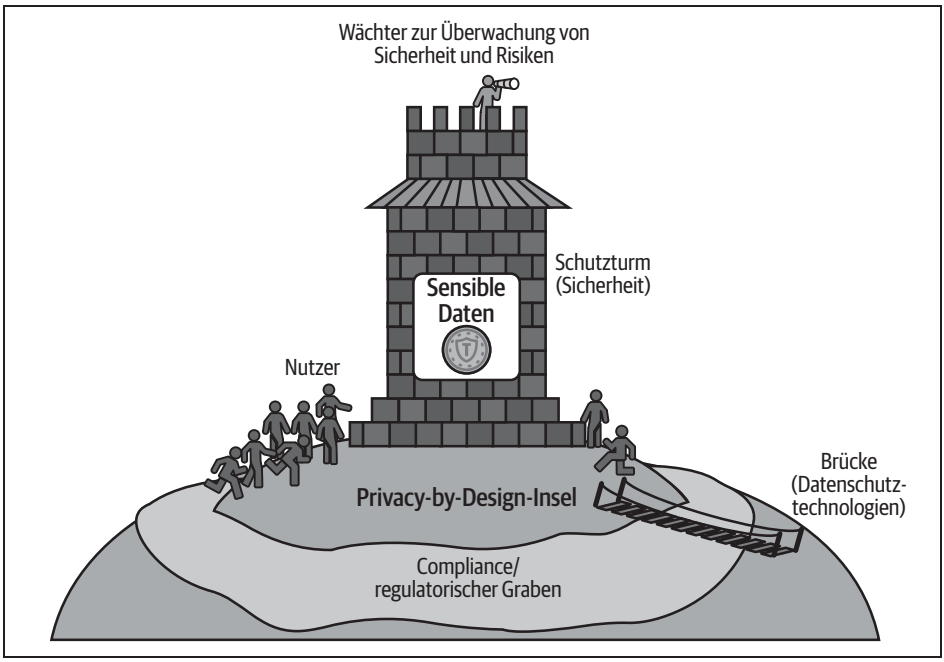

Abbildung 1-1: Data Governance veranschaulicht

Data Governance kann als eine Mischung aus Menschen, Prozessen und Technologien beschrieben werden. Unabhängig von der Größe Ihres Unternehmens existiert immer ein gewisses Maß an Data Governance, das erfüllt werden muss. In einem großen Unternehmen gibt es wahrscheinlich ein großes Team oder ein Gremium, das Standards erarbeitet, die dann in Form von Richtlinien und Verfahrensanweisungen umgesetzt und in die entsprechende technische Infrastruktur des Unternehmens implementiert werden müssen. Wenn Ihr Unternehmen klein ist, könnte dies die Aufgabe Ihres technischen oder juristischen Leiters sein. Sehen wir uns den technischen Bereich etwas genauer an, denn hier werden Sie wahrscheinlich gebeten, bei der Umsetzung dieser Richtlinien und Verfahrensanweisungen zu helfen und dafür zu sorgen, dass sie auch tatsächlich in die laufende Datenverarbeitung einfließen.

2 Privacy by Design steht für eine Reihe von Grundsätzen, die von Ann Cavoukian ausgearbeitet wurden und die sicherstellen sollen, dass die Architektur von Systemen und die Entwicklung von Softwarelösungen von Anfang an auf den Datenschutz bzw. den Schutz der Privatsphäre ausgerichtet sind. Den Ausdruck werden Sie häufig in Gesprächen mit erfahrenen Governance-Fachleuten hören. Ich empfehle Ihnen, sich die Zeit zu nehmen, diese Grundsätze durchzulesen und zu überlegen, wie sie sich auf Ihre eigene Arbeit mit Daten anwenden lassen. Sie finden diese Grundsätze in Kapitel 11.

Welche Elemente der Governance-Standards und -Richtlinien und deren Umsetzung in Technologie sind für Data Scientists von Bedeutung? Tabelle 1-1 umreißt wichtige Bereiche und damit verbundene Fragen innerhalb der Data Governance, mit denen Sie sich als Data Scientist auseinandersetzen werden.

Tabelle 1-1: Data Governance in der Data Science

Data Lineage/Datenhistorie	Richtlinien und Bestimmungen
Woher stammen die Daten?	Welche Gesetze oder internen Richtlinien gelten für diese Daten?
Wessen Daten sind das? Ist es möglich, mit ihnen Kontakt aufzunehmen?	Wo, wann und wie wurden die Daten erhoben?
Wurden diese Daten von jemand anderem erworben? Wenn ja, hat derjenige dokumentiert, wie sie verarbeitet wurden und wem sie gehören?	Welchen Bedenken hinsichtlich des Datenschutzes oder der Sicherheit müssen Sie bei der Verwendung dieser Daten Rechnung tragen?
Wie wurden die Daten durch die Verarbeitung verändert?	Wie lauteten die Datenschutzerklärung und die Nutzungsbedingungen zum Zeitpunkt der Erhebung bzw. Sammlung?
Sind die Metadaten, die Informationen zur Historie (*Lineage*) enthalten, leicht zugänglich und abzufragen?	Stammen die Daten von einem Dritten? Wenn ja, welche vertraglichen oder sonstigen Beschränkungen und Verpflichtungen bestehen für diese Daten?

Verlässlichkeit/Kenntnis der Daten	Datenschutz und -sicherheit
Welche Bedenken gibt es hinsichtlich des Einblicks in die Daten und Systeme (d. h. einschließlich der Datenerhebung, -verarbeitung und der nachgelagerten Systeme)?	Wie wird der Zugriff auf sensible Daten verwaltet und überwacht?
Ist die Dokumentation der Daten nachvollziehbar, und zwar bereits ab dem Zeitpunkt ihrer Beschaffung bzw. Erhebung?	Weiß das Unternehmen, ob und wann Daten unberechtigt abfließen? Wie?
Falls die Qualität der Daten beeinträchtigt ist, wissen Sie, wie Sie diese Probleme erkennen und beheben können?	Wer ist für die Verwaltung der Datenschutzkontrollen zuständig? Und wer ist für die Sicherheitskontrollen zuständig?
Gibt es eine Infrastruktur zur Datenspeicherung oder alte Datenspeicher, die nicht dokumentiert oder gar unbekannt sind?	Wenn sich jemand auf seine Datenrechte beruft (beispielsweise auf die DSGVO), gibt es dann ein gut dokumentiertes und verständliches System, um diese Rechte geltend zu machen?
Sind die Daten gut dokumentiert und verstanden? (»Kennen Sie Ihre Daten?!«)	Welche Technologien zum Schutz vor Datenverlust und zum Schutz der Privatsphäre setzen Sie ein und auf welche Weise?

Wahrscheinlich beschäftigen Sie sich bereits mit vielen dieser Fragen, da Daten ja ein wesentlicher Teil Ihrer Arbeit sind. Vielleicht haben Sie persönlich schon einmal darunter gelitten, dass Daten nicht dokumentiert wurden, dass Sie nicht verstanden haben, wie eine bestimmte Datenbank entstanden ist, und dass es Probleme mit den Labels und der Qualität der Daten gab. Ab jetzt haben Sie ein neues Wort, mit dem Sie diese Dinge beschreiben können: *Governance*!

Bei der Arbeit auf der Governance-Seite der Datenverwaltung bzw. des Datenmanagements geht es vor allem darum, wie Informationen über die Daten während ihres gesamten Lebenszyklus gesammelt und aktualisiert werden. Die rechtlichen, datenschutzrechtlichen und sicherheitsrelevanten Belange prägen diese Informationen und sorgen dafür, dass Governance-Entscheidungen und -Rahmenwerke Maßnahmen wie die Durchsetzung der Rechte von Einzelpersonen und die angemessene Nutzung von Daten vorantreiben. Wenn Ihre Daten nicht von Einzelpersonen stammen, gibt es möglicherweise andere Bedenken in Bezug auf urheberrechtlich geschützte Daten oder damit verbundene Sicherheitsbelange, die Einfluss auf Governance-Initiativen haben.

Wenn Sie sich konkret mit Data Governance befassen, denken Sie an Aufgaben wie die Dokumentation der sich ständig ändernden Datenflüsse in Ihrem Unternehmen. Das klingt zunächst einfach, ist es aber nicht.

Angenommen, Sie hätten einen riesigen Datenspeicher, der aus zehn verschiedenen Quellen gespeist wird, einige davon extern, andere intern. Wie können Sie anfangen, diese Daten zu verwalten? Wie würde eine skalierbare und einfach zu verwendende Lösung aussehen? Was passiert, wenn sich diese Datenflüsse ändern? Es mag ausreichend sein, nur den Code oder die Workflows zu dokumentieren, die tatsächlich ausgeführt bzw. verwendet werden, und sich ansonsten den Rest für die Zukunft aufzusparen. Doch wie gehen Sie mit Daten von Partnern oder anderen externen Datenerfassungssystemen um? Die Dokumentation dieser Daten muss koordiniert werden, damit die Rechts-, Datenschutz- und Risikoabteilungen sie für Audits und Risikobewertungen nutzen können. Dieser Prozess sollte nicht mit vorübergehenden Lösungen stückweise gelöst, sondern möglichst ganzheitlich angegangen werden.

Widmen wir uns zunächst der Frage, welche Daten im Sinne des praktischen Datenschutzes besonders schützenswert sind. Was ist unter sensiblen Daten zu verstehen, und wie können Sie sie identifizieren?

Sensible Daten identifizieren

Im Zusammenhang mit Datenschutz werden sensible Daten in der Regel als personenbezogene Daten oder einfach nur als persönlich identifizierende Daten definiert.[3] Darunter fallen Ihr vollständiger Name, Ihre E-Mail-Adresse, Ihr Geschlecht, Ihre Postanschrift, Ihre IP-Adresse, Ihr Profil in den sozialen Medien, Ihre Telefonnummer, Ihre Sozialversicherungsnummer oder andere nationale Identifikationsnummern, Ihre Kreditkartennummer, Ihr Geburtstag, Ihre Gesundheitsdaten oder auch biometrische Daten (z.B. Fingerabdrücke, Iris-Scan oder sogar Ihre Gangart (*https://oreil.ly/-ykxk*)!).

Alle diese Daten fallen unter die Kategorie der persönlich identifizierenden Daten, was viele als *persönlich identifizierende Informationen* (PII) bezeichnen. Dies sind

3 In Ihrem Unternehmen wird möglicherweise eine eigene Definition dazu verwendet, was unter sensiblen Daten zu verstehen ist, die von der in diesem Buch verwendeten Definition abweicht. Stellen Sie sicher, dass Sie intern die passenden Termini verwenden, wenn Sie das Thema ansprechen.

Daten, die sich spezifisch auf Ihre Person beziehen. Sie können allein oder in Kombination mit anderen über Sie verfügbaren Informationen verwendet werden, um Sie direkt zu identifizieren, indirekt zu identifizieren oder zu re-identifizieren. Dies sind die sensibelsten und am häufigsten regulierten Daten, da sie eine nahezu oder absolut eindeutige Identifizierung ermöglichen.

Die in diesem Buch verwendete Definition von sensiblen Daten erstreckt sich auf:

Persönlich identifizierende Daten
: Daten, die für eine bestimmte Person einmalig oder fast einmalig sind. Diese sind in der Regel im Rahmen von Richtlinien und Vorschriften definiert und können Informationen umfassen, die Sie vielleicht nicht erwarten, wie z. B. Ihre IP-Adresse, Ihr Geburtsdatum oder Ihren Arbeitsplatz.

Personenbezogene Daten
: Daten, die sich auf eine Person beziehen, aber nicht unter PII fallen. Dies kann alles sein, was mit der Person zu tun hat, einschließlich Interessen, Überzeugungen, Aufenthaltsorte sowie Online- und Offlineverhalten und -aktivitäten.

Urheberrechtlich geschützte und vertrauliche Daten
: Daten, die aus vertraglichen oder geschäftlichen Gründen als sensibel erachtet werden. Ihre Veröffentlichung würde eine Geschäftsbeziehung oder eine andere rechtliche Beziehung oder Vereinbarung gefährden.

Ich hoffe, dass dieses Buch Ihre Definition von sensiblen Daten erweitern wird. Halten Sie zum Beispiel den Standort Ihres Telefons für sensibel? Hängt dies davon ab, wo Sie sich befinden? Wenn Sie zu Hause sitzen, verrät Ihr Telefonstandort auch Ihre Postanschrift oder Ihren Wohnsitz, was wiederum eine persönlich identifizierende Information (PII) darstellt. Und wie verhält es sich, wenn Sie bei der Arbeit sind? Oder in einem Kinosaal? Oder im Haus eines guten Freunds?

Wie sieht es mit Ihrer politischen Orientierung und Ihren Interessen aus? Wie mit Ihrem Wahlverhalten? Und wie mit Ihren religiösen Ansichten und Praktiken? Was ist mit Ihren Freundschaften, Partnerschaften und anderen Kontakten? Und wie verhält es sich mit Ihrem Tagesablauf, den Nachrichten, der Musik und den Unterhaltungsangeboten, die Sie konsumieren, und wie mit den Endgeräten, die Sie besitzen?

Diese Fragen zeigen die Bandbreite unterschiedlicher Präferenzen bezüglich der Privatsphäre. Einige Menschen sind vielleicht damit einverstanden, den Standort ihrer Arbeit preiszugeben, oder sind sogar dazu angehalten, das zu tun. Andere sehen darin vielleicht einen Eingriff in ihre Privatsphäre. Während die eine Person sehr offen über ihre persönlichen Beziehungen, ihre politischen Ansichten und ihre religiösen Überzeugungen spricht, empfindet eine andere Person diese Themen vielleicht als sehr intim und vertraulich. Dies entspricht dem Konzept der kontextabhängigen Privatsphäre (*Contextual Privacy*) und dem der sozialen Dimension von Privatsphäre (*Social Privacy*), das in der *Einleitung* erörtert wurde. Hier greifen Vorschriften, die dem Einzelnen mehr Wahlmöglichkeiten hinsichtlich seiner Präferenzen in Bezug auf Privatsphäre einräumen und ihm die Möglichkeit geben, ebendiese den Datensammlern mitzuteilen, indem er seine Datenschutzeinstellungen und Einwilligungen ändert.

Das bedeutet aber auch, dass Sie, wenn Sie mit Daten arbeiten, die Bandbreite von Daten kennen, die als *sensibel* gelten. Sie sollten sich des zusätzlichen Privatsphärerisikos bewusst sein, das entsteht, wenn personenbezogene Daten auf eine neue Art und Weise miteinander verknüpft werden, durch die eine Person unbeabsichtigt identifizierbar wird. Wenn ich beispielsweise den ganzen Tag über Ihren Standort verfolge, würde ich wahrscheinlich erfahren, wo Sie arbeiten, wo Sie essen, was Sie täglich tun und wo Sie wohnen. Auch wenn ich nur Daten erheben würde, während Sie sich bewegen (d. h. Fahrdaten), könnte ich bereits einige dieser Merkmale identifizieren. Und selbst wenn ich lediglich dann Daten erheben würde, wenn Sie in Anwesenheit anderer Personen unterwegs sind, könnte ich wahrscheinlich immer noch Rückschlüsse auf Ihre Person ziehen, z. B. ob Sie mit Ihrer Familie oder einem Freund unterwegs sind oder ob Sie gern in einem bestimmten Geschäft einkaufen oder zu gewissen Zeiten auf einer bestimmten Strecke pendeln.

Ganz ähnlich haben Forscher gezeigt, dass aus einer Reihe von Facebook-Likes Rückschlüsse auf das Geschlecht, die sexuelle Orientierung und die politischen Ansichten gezogen werden können – und sogar auf den Familienstand Ihrer Eltern.[4] Dabei handelt es sich um Rückschlüsse, nicht notwendigerweise um Fakten. Es ist jedoch klar, dass das Onlineverhalten und das Verhalten in sozialen Netzwerken eindeutige Spuren hinterlässt und Muster offenbart, die die individuellen und privaten Eigenschaften einer Person offenlegen. Jeder Mensch, der Daten generiert, ist also möglicherweise identifizierbar.

Die Macht der Schlussfolgerung in Verbindung mit großen Informationsmengen kann Personen identifizieren, auch wenn dies nicht beabsichtigt ist. In der zielgerichteten Werbung (engl. *Targeted Advertising*) werden oft sensible Merkmale abgeleitet und ohne Zustimmung miteinander verknüpft, was zu Empfehlungen führt, die sensible Informationen wie die sexuelle Orientierung oder politische Ansichten preisgeben könnten. Je mehr Faktoren ein Werbetreibender bei der Auswahl von Zielgruppen einbezieht, desto eher könnte er fälschlicherweise eine einzelne Person oder eine sehr kleine Zielgruppe ansprechen. Handelt ein Werbetreibender in böswilliger Absicht, kann er die ihm vorliegenden Informationen nutzen, um herauszufinden, wie er diese Person gezielt ansprechen oder ihr ziemlich nahekommen kann.

Aus diesen Gründen könnte der Begriff *sensible Daten* jegliche personenbezogenen Daten umfassen, unabhängig davon, ob sie direkt einer Person zugeordnet werden können oder nicht. Personenbezogene Daten, insbesondere in großen Mengen oder in aggregierter Form, sind einer Person zuordenbar. Wenn ich in diesem Buch von sensiblen Daten spreche, beziehe ich mich nicht nur auf PII, sondern auch auf ein breiteres Spektrum personenbezogener Daten, die in Kombination mit anderen Informationen dazu verwendet werden könnten, eine Person oder eine kleine Gruppe von Personen zu identifizieren.

4 Diese Studie wurde von einigen der Forscher veröffentlicht, die später an Cambridge Analytica gearbeitet haben. Siehe Kosinki et al., »Private traits and attributes are predictable from digital records of human behavior« (*https://oreil.ly/ZZCnR*), 2013.

Die letzte Kategorie sensibler Daten bilden Daten, die aus nicht personenbezogenen Gründen urheberrechtlich geschützt oder vertraulich sind. Dabei kann es sich um Geschäftsgeheimnisse, geschützte Informationen über das Unternehmen, ein bestimmtes Produkt oder um Informationen handeln, die unter eine Geheimhaltungsklausel fallen. Das könnten Daten sein, die zwischen Mutter- und Tochtergesellschaften ausgetauscht werden und die aufgrund interner Vorschriften oder Vertraulichkeitsvereinbarungen geheim gehalten werden müssen. Oder es könnte sich um sensible interne Daten handeln, die, wenn sie nach außen dringen, der Konkurrenz einen Vorteil verschaffen oder das Unternehmen auf andere Weise gefährden würden. Auch diese Art von sensiblen Informationen profitieren von den Ansätzen und Technologien, die Sie in diesem Buch kennenlernen werden.

Ich bin eine Befürworterin von Whistleblowing. Wenn Sie über Daten verfügen, von denen Sie glauben, dass sie öffentlich bekannt sein sollten, die aber als sensibel eingestuft werden, sollten Sie anhand der in diesem Buch vorgestellten Techniken darüber nachdenken, wie Sie diese Daten auf verantwortungsvolle und wohlüberlegte Art und Weise öffentlich machen oder an die zuständigen Behörden weitergeben können.

Der erste Schritt, um sensible Daten schützen zu können, besteht darin, sie auf verlässliche Weise zu identifizieren. Sobald die Daten identifiziert und als sensibel dokumentiert sind, können Sie ermitteln, wie sie am besten geschützt werden können.

Persönlich identifizierende Informationen (PII) identifizieren

Persönlich identifizierende Informationen (engl. *Personally Identifiable Information*, PII) fallen in den meisten Datenschutzvorschriften unter eine bestimmte rechtliche Kategorie, der bei der Umsetzung von Data Governance in einem Unternehmen bzw. einer Organisation besondere Aufmerksamkeit geschenkt werden muss. Wenn Ihr Unternehmen persönliche Daten – insbesondere Daten von Angestellten – sammelt, gelten für diese Daten oftmals besondere Anforderungen an die Data Governance. Meist mangelt es daran, persönlich identifizierende Daten zu dokumentieren oder zu kategorisieren, da sie oft in Textdateien, Logdateien oder anderen unstrukturierten Daten auftauchen, die oftmals schlecht dokumentiert sind.

Es gibt mehrere Tools, die speziell dafür entwickelt wurden, persönlich identifizierende Daten in unstrukturierten Daten mithilfe einer Vielzahl verschiedener Methoden zu finden. Zudem habe ich schon Teams gesehen, die ihre eigenen Tools und Systeme entwickelt haben, mit denen sie PII erfolgreich ausfindig machen können. Bei vielen Tools kommen relativ anfällige Methoden zum Einsatz, wie z. B. reguläre Ausdrücke (Zeichenketten zum Abgleichen von Mustern) oder die sogenannte String-Entropie (zum Auffinden von Schlüsseln für Programmierschnittstellen bzw. APIs, kryptografischen Schlüsseln oder Passwörtern). Darüber hinaus habe ich bereits Deep-Learning-Modelle entwickelt, mit denen ich PII in Nachrichtentexten identifizieren konnte. Je nach Anwendungsfall werden Ihre Ergebnisse mit diesen Ansätzen unterschiedlich ausfallen und sollten dementsprechend evaluiert werden.

PII lassen sich nicht immer vollständig aufspüren – und das wird auch nie der Fall sein. Es ist wichtig, dass Sie mit Ihren Risikoteams (Datenschutz, Rechtsabteilung, Sicherheit) über diese Tatsache sprechen, egal ob Sie ein Toolkit zur Erkennung von PII kaufen oder Ihr eigenes entwickeln. Am sichersten ist es, Daten, die von Menschen eingegeben wurden, als äußerst sensibel zu behandeln (d.h. als PII), unabhängig davon, welche »Bereinigungsverfahren« angewandt wurden. Wenn Sie die von Menschen eingegebenen Daten nutzen möchten, ohne die mit PII verbundenen zusätzlichen Schutzmaßnahmen zu ergreifen, müssen Sie die Risiken korrekt identifizieren und sicherstellen, dass sie entsprechend berücksichtigt und überprüft werden.

Sollten Sie mit einem hohen Bestand an undokumentierten Daten arbeiten und befürchten, dass darin eine Menge PII enthalten sind, sollten Sie ein leicht zu verwendendes Open-Source-Tool zur Hand nehmen. Sobald Sie sich ein Bild davon gemacht haben, wie weit Sie damit kommen, können Sie entscheiden, ob Sie in einen fortschrittlicheren oder kostspieligeren Ansatz investieren sollten. Ich kann Ihnen Presidio von Microsoft (*https://oreil.ly/Tao7Z*) empfehlen, das auch einige grundlegende Pseudonymisierungsverfahren umfasst, die im weiteren Verlauf dieses Kapitels behandelt werden.

Der beste Ansatz für das Management und die Nachverfolgung von PII besteht darin, die Daten zu verfolgen, wenn sie eingehen, und diese Daten zu kennzeichnen bzw. zu labeln und zu verwalten, während sie das System durchlaufen, sodass Sie sie später nicht erst mühsam suchen bzw. aufspüren müssen. Eine der Möglichkeiten, wie Sie die Gewohnheit und Kultur der frühzeitigen und häufigen Erkennung von PII beginnen können, ist der Aufbau einer Dokumentationskultur rund um die Beschaffung und Verwendung von Daten. Tabelle 1-1 bietet Ihnen einen guten Ausgangspunkt. Um einen umfassenden Ansatz zu entwickeln, sollten Sie zahlreiche Bereiche Ihres Unternehmens einbeziehen, darunter das Sicherheitsteam, die für Informationen und Daten zuständigen Personen sowie die Infrastruktur- und IT-Abteilungen.

Datennutzung dokumentieren

Ein wesentlicher Bestandteil der Data Governance ist die Dokumentation Ihrer Daten, unabhängig davon, ob es sich um sensible Daten handelt oder nicht. Sollten Sie in einem größeren Unternehmen tätig sein, verfügen Sie möglicherweise auch über ein Klassifizierungssystem für Ihre Daten, bei dem bestimmte Richtlinien auf verschiedene Kategorien bzw. Klassen von Daten angewendet werden. Beispielsweise müssen Sie möglicherweise PII-Daten mit Tags und Labels versehen, um sicherzustellen, dass der Zugriff auf diese Daten eingeschränkt ist. In diesem Fall können Sie die Datenschutzklassifizierung als ersten Schritt in Ihrer Dokumentation nutzen.

Allerdings gehört zur Dokumentation weit mehr als die Klassifizierung sensibler Daten. In diesem Abschnitt erfahren Sie, wie Sie mit der Dokumentation der Daten beginnen können, wie Sie undokumentierte Daten finden können, wie Sie Informatio-

nen zu Herkunft (Lineage), Erhebung und Verarbeitung hinzufügen können und wann Sie eine Versionskontrolle der Daten einführen sollten. Viele dieser Systeme sind notwendig, um die Grundlage dafür zu schaffen, wie sensible Daten im Unternehmen verwendet und verwaltet werden. Auf diese Weise können Sie Anwendungsfälle erkennen, die sich für fortgeschrittene Datenschutztechnologien eignen, die Sie im weiteren Verlauf dieses Buchs kennenlernen werden.

Grundlagen der Datendokumentation

Daten müssen entsprechend der Art und Weise ihrer Nutzung dokumentiert werden. Versetzen Sie sich bei der Dokumentation von Daten in die Leserinnen und Leser – genau so, wie Sie es bei der Dokumentation von Code handhaben würden. Was wird der Adressat verstehen? Wie werden diese Personen die Dokumentation in ihrem täglichen Arbeitsablauf auffinden und darauf zugreifen? Wie werden sie die Dokumentation durchsuchen? Welche Wörter werden sie verwenden? Was sind die wichtigsten Informationen? Wie können Sie sie so prägnant und hilfreich gestalten, dass sie tatsächlich gelesen werden? Wie können Sie dafür Sorge tragen, dass sich die Daten möglichst selbst dokumentieren und leicht aktualisieren lassen?

Obwohl die Datendokumentation an sich nicht wirklich etwas Neues ist, ist sie auch keine weit verbreitete Praxis, vor allem nicht in Data-Science-Teams. Die Arbeit in der Data Science hat sich von forschungsorientierten Teams und analyseorientierten Dashboards hin zu Versuchen, Fehlern, agiler Entwicklung und Deployment-Standards verlagert. Umso wichtiger ist es, dass andere anhand einer gut geschriebenen Dokumentation verstehen, was in Datenworkflows und Experimenten vor sich geht.

Ähnlich wie bei der Versionskontrolle von Daten und Experimenten ermöglicht die Dokumentation von Daten anderen Teams, Datenquellen zu entdecken und zu nutzen, die ohne die richtige Dokumentation möglicherweise unübersichtlich oder schwer zu finden wären. In vielen Unternehmen sind die Datenquellen oft auf verschiedene Teams in unterschiedlichen Bereichen des Unternehmens verteilt oder mehrfach vorhanden. Es kann sich schon als schwierig erweisen, grundlegenden Zugang und Interoperabilität richtig zu gestalten, sodass die Erstellung der Dokumentation oft unter den Tisch fällt.

Doch das muss nicht zwangsläufig der Fall sein – vorausgesetzt, die Dokumentation wird als wesentlicher Bestandteil der Arbeit mit Daten angesehen. Hier sind einige Möglichkeiten, wie Sie das Datenmanagement oder die Geschäftsbereiche davon überzeugen können, dass es sich durchaus lohnt, mehr Zeit und Mühe in die Dokumentation von Daten zu investieren. Datendokumentation ...

- beschleunigt Datenexperimente, was zu neuen datengetriebenen Erkenntnissen und Entdeckungen führen kann.
- ermöglicht abteilungs- und teamübergreifende Zusammenarbeit.
- beschleunigt den Zugang zu Daten und deren Nutzung für alle Beteiligten.
- hilft dabei, unbekannte oder undokumentierte Daten zu entfernen.

- unterstützt Datenteams bei der Entscheidung, welche Daten für neue Vorhaben, Produkte und Modelle verwendet werden sollten.
- gibt den Produktteams Auskunft darüber, welche Daten zur Umsetzung neuer Ideen verfügbar sind.
- zeigt Analysten disparate Datensätze, die für neue Erkenntnisse und Reports genutzt werden könnten.
- verschafft den Compliance- und Audit-Teams einen guten Überblick und unterstützt sie bei neuen Maßnahmen zur Gewährleistung der Datensicherheit und des Datenschutzes.
- reduziert Compliance- und Datensicherheitsrisiken.

Data Governance kann in Ihrem Unternehmen nur gelingen, wenn es eine funktionierende und effektive Datendokumentation gibt. Eine ordnungsgemäße Dokumentation kann sogar in Identitätsmanagement- und Zugriffssysteme integriert werden, um Datenadministratoren, -eigentümern und -managern einfache Möglichkeiten zu geben, den Zugriff auf der Grundlage der Dokumentation zu gewähren bzw. zu entziehen.

Um ein KI-getriebenes Unternehmen verantwortungsvoll zu gestalten, benötigen Sie eine zusätzliche Dokumentation, in der alle in den Daten enthaltenen Stereotype oder bekannten Verzerrungen bzw. Voreingenommenheiten (sogenannte Bias) aufgezeigt werden. Ich empfehle Ihnen, sich die Zeit zu nehmen, etwas über Data Cards (*https://oreil.ly/zPPZu*), also auf Deutsch »Datenkarten«, zu erfahren – eine Möglichkeit, Daten für nicht technische Nutzer zu dokumentieren. Wenn Ihr Team an Machine-Learning-Systemen arbeitet, die sich an Verbraucher richten, empfehle ich Ihnen, sowohl Data Cards als auch Model Cards (*https://oreil.ly/_5NL5*) (»Modellkarten«) zu verwenden, um sicherzustellen, dass die Systeme, die Sie zur Verfügung stellen, akkurat, fair und zuverlässig arbeiten!

Die Dokumentation kann Angaben zu einem der folgenden Abschnitte oder zu allen enthalten und über das hinausgehen, was in diesem Kapitel behandelt wird. Wenn Sie eine neue Dokumentation erstellen, sollten Sie herausfinden, was angesichts Ihrer Rahmenbedingungen am besten geeignet ist. Setzen Sie Prioritäten bei den für Ihr Unternehmen wichtigsten Abschnitten und gehen Sie von dort aus weiter vor.

Datenerhebung

Erläuterung und damit verbundene Fragen

Wer, wo, wann, warum und wie wurden die Daten erhoben? Eine Beschreibung, wann die einzelnen Datensätze gesammelt bzw. erhoben wurden, welches Team oder mit welcher Software die Erhebung verwaltet wurde, welche Verarbeitung nach der Erhebung durchgeführt wurde und warum die Sammlung bzw. Erhebung der Daten (engl. *Data Collection*) stattfand (d.h. unter welchen Umständen und mit welchem rechtlichen Hintergrund).

Beispiel

Ein Diagramm, das den Datenfluss zeigt, wobei dokumentiert wird, wann und von wem er implementiert wurde, sowie ein Snapshot der Daten zu diesem Zeitpunkt. Das Diagramm sollte auch Informationen über die Zustimmung der Stakeholder in den Rechts-, Datenschutz- und Compliance-Abteilungen enthalten – oder andere Gründe für ein berechtigtes Interesse an der Sammlung der Daten, falls keine Zustimmung erteilt wird.

Datenqualität

Erläuterung und damit verbundene Fragen

Wie wurden die Daten standardisiert (falls vorgenommen)? Welche Qualitätskontrollen wurden durchgeführt? Wie viele Nullwerte oder Extremwerte sind in den Daten enthalten? Wurden die Daten auf Duplikate oder Inkonsistenzen geprüft bzw. entsprechend aufbereitet? In diesem Abschnitt können Sie auch dokumentieren, wenn ein Schema oder eine Einheit geändert wurde.

Beispiel

Eine Analyse der Qualität der Daten für einen bestimmten Zeitraum, einschließlich der Häufigkeit von Nullwerten, der Standardisierung und Harmonisierung von Werten (d.h. Werten, die in Prozentwerte umgewandelt oder in Standardeinheiten umgerechnet werden) sowie Bias und Varianz der Daten. Eine detailliertere Analyse könnte Histogramme über zahlreiche Dimensionen beinhalten, die Kovarianz oder Korrelation zwischen Attributen aufzeigen oder auch Informationen zu Ausreißern oder Extremwerten liefern. Sollten sich die Daten innerhalb eines kurzen Zeitraums erheblich verändern, können Sie ein Monitoring und ein entsprechendes Alarm- bzw. Benachrichtigungssystem einsetzen.

Alle diese Aspekte müssen einzeln und anschließend auch in ihrer Gesamtheit beachtet werden. Wenn Sie lediglich die Datenqualität testen, ohne den Datenschutz oder die Datensicherheit zu berücksichtigen, kann dies dazu führen, dass Sie versehentlich sensible Daten preisgeben. Stellen Sie sicher, dass Sie diese Governance-Mechanismen ganzheitlich anwenden und die verschiedenen Maßnahmen sorgfältig und vollständig integrieren.

Datensicherheit

Erläuterung und damit verbundene Fragen

Wie hoch ist das für diese Daten bestehende Sicherheitsrisiko? Welche Maßnahmen sollten ergriffen werden, wenn die Daten verwendet werden oder wenn darauf zugegriffen wird? Führen Sie eine Risikoanalyse in Bezug auf die Sensibilität der Daten und deren Infrastruktur, Architektur und Speicherdetails durch. Stellen Sie Informationen bereit, die anderen Teams (z.B. Security & Operations) helfen, das Risiko akkurat zu modellieren und zu bewerten und Entscheidungen bezüglich Zugriffsbeschränkungen zu treffen. Hier können Sie auch die Sicherheits- und Datenschutztechnologien dokumentieren, die zur Ri-

sikominderung eingesetzt werden. Diese Risikominderungen sollten ebenfalls dokumentiert werden, damit die Datenkonsumenten, d. h. die Personen, die die Daten verwenden, analysieren oder weiterverarbeiten werden, wissen, wie sie mit den Daten auf sinnvolle Weise arbeiten können. In Kapitel 4 erfahren Sie mehr darüber, wie Sie Sicherheitsrisiken bewerten, eindämmen und dokumentieren können.

Beispiel

Eine Evaluierung dazu, wie eine bestimmte Maßnahme zur Verringerung des Datensicherheitsrisikos beiträgt, und eine Empfehlung, sofern diese Schutzmaßnahme erfolgreich dazu beiträgt, das ermittelte Risiko zu verringern. Wenn die Maßnahme umgesetzt wird, sollte neben den erforderlichen Entscheidungsprotokollen auch Einzelheiten zur Umsetzung enthalten sein.

Data Privacy

Erläuterung und damit verbundene Fragen

Enthalten die Daten personenbezogene Informationen? Wenn ja, welches Privacy by Design (*https://oreil.ly/S13mS*) für rechtliche oder datenschutzrechtliche Vorgaben wurde durchgeführt, um sicherzustellen, dass die Daten ordnungsgemäß gehandhabt werden? Wenn PII oder personenbezogene Daten gespeichert werden, sollte die Dokumentation Informationen darüber enthalten, in wessen rechtlichen Zuständigkeitsbereich die Daten fallen und welche Datenschutzbestimmungen und Einwilligungsmöglichkeiten den Nutzern zum Zeitpunkt der Erhebung mitgeteilt wurden. Außerdem sollte eine klare Dokumentation aller Mechanismen zur Wahrung der Privatsphäre vorliegen, wobei der zugehörige Code und, wenn möglich, auch die Commit-Hashes angegeben bzw. verlinkt werden sollten. Stellen Sie sicher, dass alle personenbezogenen Daten als sensible Daten behandelt werden und dass die Schutzmaßnahmen angemessen dokumentiert werden.

Beispiel

Eine Auflistung aller Spalten, die personenbezogene Daten enthalten, sowie eine Spalte mit einem Zeitstempel, der angibt, wann die Daten auf der Grundlage der Richtlinie und der Gerichtsbarkeit, in der die Daten erhoben wurden, gelöscht werden sollen. Idealerweise ist die Löschung automatisiert.

Datendefinitionen

Erläuterung und damit verbundene Fragen

Wie sind die Daten aufgebaut? Wenn es sich um tabellarische Daten handelt, was bedeuten die Spaltennamen? Um welche Datentypen handelt es sich? Welche Maßeinheiten werden verwendet? Erläutern Sie die Bedeutung von Fachbegriffen und Codierungen, die in den Daten verwendet werden (z. B. Abkürzungen oder interne Zuordnungen). Beschreiben Sie die Datenfelder, Spaltennamen (falls vorhanden), Schlüssel und Werte, Codierungen, Maßeinheiten und sonstige Angaben, die einem neuen Nutzer der Daten dabei helfen, die Daten zu verstehen, die er verwendet. Wenn bestimmte Formatierungsstandards gewählt

werden, wie z.B. die Darstellung des Datums gemäß ISO-Format, die Angabe der Uhrzeit im 24-Stunden-Format oder andere Standardeinheiten, sollten Sie Angaben dazu machen, wie diese Verarbeitung erfolgt, für den Fall, dass andere nach möglichen Fehlern suchen müssen oder sich die Verarbeitung ändert.

Beispiel

Eine abfragbare und leicht zugängliche Liste aller Spalten mit ihren Beschreibungen und Datentypen. Für kategoriale Spalten beinhaltet dies auch eine leicht durchsuchbare tabellarische Darstellung, aus der hervorgeht, wie sie codiert sind.

Deskriptive Statistiken

Erläuterung und damit verbundene Fragen

Wie sehen die gängigen deskriptiven Statistiken für einen Datensatz aus, z.B. Varianz, Verteilung, Mittelwert usw.? Wie sind die Daten hinsichtlich bestimmter wichtiger Merkmale verteilt? Gibt es Hinweise darauf, dass die Daten gewisse Verzerrungen aufweisen oder dass die Klassen unausgewogen verteilt sind? Zusammenfassende statistische Beschreibungen der Daten können in schriftlicher oder grafischer Form beigefügt werden, damit andere Personen schnell beurteilen können, ob die Daten ihren Anforderungen entsprechen. Eine interaktive Darstellung ist hier von großem Vorteil. Diese Informationen können sehr sensibel sein, sodass Sie sie nur nach Abwägung der Sicherheits- und Datenschutzrisiken, die mit der Gewährung des Zugriffs verbunden sind, veröffentlichen sollten. Wenn Sie sich inspirieren lassen möchten, sollten Sie sich das Tool Facets (*https://oreil.ly/tV1PK*) von Google ansehen.

Beispiel

Ein Diagramm, aus dem die Perzentile numerischer Spalten hervorgehen (z.B. mithilfe von Box-Plots, im Deutschen auch als Kastengrafiken bekannt, bei denen Ausreißer entfernt wurden). Über eine dynamische Auswahl könnten auch andere Merkmale dargestellt werden, sodass Sie in dem Datensatz bestehende Korrelationen und Verzerrungen (Bias) selbst analysieren können.

Diese Aufzählung ist nicht erschöpfend. Sie kann Ihnen aber als Leitfaden dienen, wenn Sie in Ihrem Unternehmen mit der Dokumentation der Daten beginnen. Denken Sie daran, dass die Dokumentation nicht für Sie, sondern für die Datennutzenden bestimmt ist. Verwenden Sie daher Formulierungen, Visualisierungen und Beschreibungen, die für die Teams im gesamten Unternehmen verständlich sind, damit die Endnutzer die gesuchten Informationen finden und problemlos nutzen können.

Wie bei der Dokumentation von Code und Architektur werden Ihr Unternehmen und die Datennutzer von gut dokumentierten Daten profitieren. Projekte werden beschleunigt, die Art und Weise, wie das Unternehmen mit Daten arbeitet, wird standardisiert, und Herkunft sowie Qualität der Daten werden klargestellt. Das erleichtert die Entscheidungsfindung. Außerdem werden Datenschutz- und Sicherheitsfragen zu Beginn des Lebenszyklus der Daten und zu Beginn neuer Projekte geklärt – dann, wenn es am effektivsten ist!

Einführen einer Datendokumentation

Wenn Sie zum ersten Mal eine Dokumentation für Ihre Daten anlegen, sollten Sie ausprobieren, was am besten funktioniert, indem Sie kleinere Tests und Dokumentationen vornehmen und Feedback einholen. Hier sind ein paar weitere nützliche Tipps, die Ihnen dabei helfen:

Nutzererlebnis
: Wenn möglich, sollten Sie mit Experten in Kontakt treten, die sich mit Nutzererlebnis bzw. -erfahrung (UX) oder Produkten auskennen und Ihnen weiterhelfen können. Führen Sie Interviews mit Ihren Anwenderinnen und Anwendern durch und bringen Sie in Erfahrung, ob die Dokumentation von Nutzen ist. Überarbeiten und verbessern Sie die Dokumentation anhand dieses Feedbacks.

Standardisieren Sie erst, wenn erfolgreich erprobt
: Standardisieren Sie die vollständige Datendokumentation erst dann, wenn Sie ein System gefunden haben, das funktioniert und das Sie mit der Zeit auch pflegen und verbessern können.

Auf Genauigkeit achten
: Eine ungenaue Dokumentation ist oft schlimmer als gar keine Dokumentation, denn Ihre Nutzer denken, dass sie die Daten verstehen, was aber nicht der Fall ist. Wenn sie basierend auf diesem Missverständnissen Modelle erstellen oder Entscheidungen treffen, kann das in ein Desaster führen. Deshalb ist es wichtig, dass Sie eine nachhaltige Lösung entwickeln, die so einfach wie möglich zu pflegen ist.

Möglicherweise haben Sie Ihre Dokumentation bereits im Griff, aber ein quälendes Problem: undokumentierte Daten, von denen Sie nichts wissen. Es ist wichtig, dieses Problem in Angriff zu nehmen, denn undokumentierte Daten sind häufig sensibler Natur.

Unbekannte Daten aufspüren und dokumentieren

Unbekannte Daten sind Lücken in der Datendokumentation oder sogar im grundlegenden Wissen über die Daten und in deren Verständnis. Diese treten in großen Unternehmen dann zutage, wenn es an Best Practices mangelt, und zwar meist über viele Jahre hinweg. Daten aus früheren Anwendungen oder auslaufenden Produkten, Daten, die im Rahmen von Übernahmen oder Kooperationen gewonnen wurden, oder Daten, die vor langer Zeit gekauft wurden, häufen sich in Datenbanken oder Dateien an, die nicht dokumentiert sind. Mitunter werden diese Daten aktiv genutzt, ohne dass jemand weiß, wie sie dorthin gelangt sind oder wann sie gesammelt wurden. Zudem kommt es vor, dass Daten neu entdeckt werden und das Unternehmen sich über ihre Herkunft im Unklaren ist. Solch undokumentierte Daten können auch das Ergebnis von Verlust von Know-how sein – wenn Daten nicht dokumentiert wurden, bevor wichtige Mitarbeiter das Unternehmen verlassen haben.

Beim Umgang mit unbekannten Daten ist es wichtig, einen Prozess und eine Routine zu etablieren, damit die Daten nicht noch länger undokumentiert bleiben. Denn Daten, die Sie nicht kennen, die nicht verwaltet und nicht nachverfolgt bzw. getrackt werden, bergen große Risiken für den Datenschutz und die Sicherheit, wie Sie noch im Laufe dieses Kapitels und auch in Kapitel 4 erfahren werden. Im Folgenden schlage ich Ihnen einen Prozess vor, über den Sie gefundene unbekannte Daten untersuchen, dokumentieren und eine Entscheidung treffen können. Sie können dieses Vorgehen an Ihre Erfordernisse anpassen, indem Sie weitere Schritte hinzufügen oder klarere Anforderungen für jede Phase festlegen, wie z. B. den spezifischen Ort und die Technologie, die verwendet werden soll.

1. *Ermitteln Sie die mutmaßliche Herkunft.*
   Sehen die Daten wie bereits dokumentierte Daten des Unternehmens aus? Sind die Daten leicht über eine Suchmaschine zu finden (d. h., handelt es sich um öffentlich verfügbare Daten)? Weiß jemand in einem der verwandten Teams, woher die Daten stammen?
2. *Erkunden Sie die Daten.*
   Untersuchen Sie alle möglichen Datenquellen und setzen Sie sich mit anderen Teams und Abteilungen in Verbindung. Vielleicht kennt jemand die Herkunft der Daten oder sieht Ähnlichkeiten zu Daten, die er bereits verwendet hat. Gehen Sie den Inhalt der Daten auf mögliche Hinweise durch und dokumentieren Sie Ihre Erkenntnisse.
3. *Ermitteln Sie, wie sensibel die Daten sind.*
   Enthalten die Daten personenbezogene oder persönlich identifizierende Informationen? Können Sie das Datum, an dem die Daten erfasst wurden, anhand der entsprechenden Daten in der Datenbank oder dem Dokument ermitteln? Wenn es sich um personenbezogene Daten handelt, wie lassen sich diese Daten vor dem Hintergrund der Datenschutzerklärung Ihres Unternehmens einordnen? Gibt es eine bestimmte Einstufung in Bezug auf den Datenschutz oder die Vertraulichkeit, die beachtet und eingehalten werden sollte?
4. *Fangen Sie damit an, die Dokumentation an den Adressaten zu richten.*
   Wie können Sie eine Dokumentation erstellen, die von den Nutzerinnen und Nutzern gelesen und verwendet wird? Beginnen Sie damit, zu dokumentieren, wo die Daten gefunden wurden, was sie beinhalten und zu welchem Ergebnis Sie bezüglich der Herkunft und des Grads der Vertraulichkeit gekommen sind. Die Dokumentation sollte den Unternehmensstandards entsprechen sowie leicht auffindbar und zugänglich sein. An diesem Punkt sollten Sie die Entscheidungsträger des Datenmanagements einbeziehen, um die nächsten Schritte festzulegen.
5. *Löschen, archivieren oder pflegen?*
   Was sind die nächsten Schritte für die »jetzt bekannten« Daten? Um dies zu entscheiden, sollten Sie die betroffenen Parteien einbeziehen, einschließlich der Compliance- und Audit-Abteilungen, sofern diese in Ihrem Unternehmen vorhanden sind. Wenn die Daten nicht sensibel sind, sich nicht auf urheberrecht-

lich geschützte Details oder Personen beziehen und nützlich sind, können Sie sie wahrscheinlich einfach in die Dokumentation aufnehmen. Stellen Sie sicher, dass die Dokumentation gemeinsam genutzt und aktualisiert wird. Andernfalls können Sie die Daten archivieren, bis mehr Informationen verfügbar sind. Diese Option ist auch dann sinnvoll, wenn die Daten nicht mehr benötigt werden, Sie aber noch eine Weile warten möchten, bevor Sie sie löschen. Die Standards zur Datenminimierung im Bereich des Datenschutzes empfehlen die Löschung von Daten, insbesondere wenn sie personenbezogene Informationen enthalten und Sie nicht herausfinden konnten, unter welchen Umständen die Daten gesammelt wurden. Dennoch sollten Sie diese Entscheidung und die Untersuchung dokumentieren, bevor Sie die Daten löschen, damit Sie bei einer eventuellen Prüfung einen Nachweis haben. Ich würde fast immer dazu raten, Daten zu löschen, wenn sie zu alt und für datenwissenschaftliche oder geschäftliche Zwecke nicht mehr nützlich sind.

Wenn Sie sich Sorgen um unbekannte Daten in Ihrem Unternehmen machen, gibt es verschiedene Produkte, die Teams bei der Suche nach diesen Daten unterstützen. Diese Dienste bieten oft Scan-Software an, die Server durchsucht und versucht, Daten auch dort zu finden, wo man sie nicht vermuten würde. Wenn Sie jedoch Ihre Datenprozesse gut dokumentieren und in engem Austausch mit anderen Datenteams stehen, ist es ziemlich unwahrscheinlich, dass Daten ungenutzt und unentdeckt herumliegen.

Ein häufiges Problem mit unbekannten Daten ist, dass es sich um historische Daten aus Reports handelt, die häufig vor der Einführung von Data Science in Unternehmen von Entscheidungsträgern gesammelt und verwendet wurden. Wenn das Data-Science-Team oder die Ausrichtung auf Data Science in Ihrem Unternehmen noch relativ neu ist, könnte es sich lohnen, die von den Geschäftseinheiten verwendeten Reporting-Daten zu untersuchen, die in der Regel außerhalb Ihrer regulären Datenerfassungsmechanismen liegen. Beispielsweise könnte es diverse Tabellenkalkulationen oder andere Arten von dokumentenbasierten Reports geben, die jahrelang verwendet wurden, bevor bessere Kunden- oder Produktdaten zur Verfügung standen. Es kann auch Integrationen in Tools wie Mitarbeiter- oder Kundenverwaltungssystemen oder anderer Software geben, die Daten abrufen und sie auf internen Servern oder Dateisystemen speichern. Wenn Sie einem Team beitreten, das mit undokumentierten Datenquellen arbeitet, sollten Sie sich mit solchen Systemen vertraut machen.

Diese Praktiken haben sich manchmal aus einer »Schatten-IT« entwickelt, bei der sensible Daten aufgrund von Zugriffsbeschränkungen an viele verschiedene Speicherorte kopiert werden. Schatten-IT ist ein Begriff, der Vorgänge beschreibt, die sich außerhalb des IT-Managements oder des Zuständigkeitsbereichs befinden – oft als nützliche Abkürzung geschaffen – und die häufig zu einem Albtraum in Sachen Sicherheit und Auditing führen. Leider ist dieser Prozess nicht ungewöhnlich, da Mitarbeitende teils viele Stunden, Tage und Wochen damit verbringen, auf eine Zugriffsgenehmigung zu warten. Sobald der Zugriff gewährt wird, kopieren die Nutzer die Daten sofort oder entwickeln dafür automatisierte Verfahren, damit sie nicht er-

neut warten müssen. Ein Teil Ihrer Aufgabe wird es sein, diese Praktiken aufzudecken und bessere Datenschutztechnologien zu entwickeln, um die Zeit bis zum Datenzugriff zu verkürzen. Machen Sie Schluss mit der Schatten-IT und setzen Sie stattdessen auf transparente, nutzerfreundliche, gut dokumentierte und datenschutzfreundliche Zugriffssysteme!

Bei der Suche nach unbekannten Daten und der Empfehlung, diese zu löschen, werden Sie auf unterschiedliche Reaktionen stoßen. Einige werden sich zurückhaltend, ängstlich oder sogar ablehnend verhalten. Für das Unternehmen ist es jedoch wichtig, sicherzustellen, dass die Rechte an den Daten zuverlässig und nachweisbar gewahrt werden, unabhängig davon, wie gut die Absichten auch sein mögen. Es gibt Datenteams, die der Meinung sind, dass Daten ohne Wenn und Aber gespeichert werden sollten. Diese Art von Datenhortung stellt den Datenschutz vor große Herausforderungen.

Anstatt die kulturellen und kommunikativen Probleme allein anzugehen, sollten Sie die Entscheidungsträger über die Risiken, die mit der Aufbewahrung unbekannter Daten verbunden sind, aufklären. Tabellenkalkulationen voller Kunden- oder Mitarbeiterinformationen sind wertvolle Vermögenswerte, die entweder von fähigen Daten- und Sicherheitsteams dokumentiert und verwaltet werden sollten, die Best Practices befolgen, oder gelöscht werden sollten, um nicht zur Zielscheibe für interne oder externe Sicherheitsrisiken zu werden. Wenn sich solche Reporting-Daten für die Beantwortung von Fragen als nützlich erweisen, sollten sie aufbewahrt, dokumentiert und geprüft werden. Auf diese Weise erhalten Sie bessere Einblicke und können gezieltere Entscheidungen treffen!

Bei undokumentierten Daten handelt es sich oft um in Systemen verwaiste Daten, die nicht oder nur unzureichend dokumentiert sind. Ähnlich wie die grundlegende Dokumentation, über die Sie bereits einiges gelernt haben, ist die Rückverfolgung von Daten und ihrer Herkunft (das Data-Lineage-Tracking) ein wichtiges Instrument für die Data Governance.

Data-Lineage-Tracking

Mithilfe von *Data Lineage* (manchmal auch *Data Provenance* genannt) können Sie rückverfolgen, woher die Daten stammen, wie sie dorthin gelangt sind und wie die Daten seit ihrem Eingang in das System verarbeitet wurden. Diese Informationen sind, wie Sie sich vorstellen können, für Data Scientists äußerst nützlich, um sich ein Bild von der Qualität, dem Inhalt und dem möglichen Nutzen der Daten machen zu können.

Die Informationen zur Datenhistorie ermöglichen es Ihnen, Fragen wie die folgenden zu beantworten:

- Wann wurden diese Daten erfasst? Woher stammen sie?
- Inwieweit darf ich diese Daten nutzen? Welche Art von Einwilligung wurde bei der Erhebung der Daten erteilt?

- Wie wurden diese Daten verarbeitet, bereinigt und aufbereitet (z. B. durch Entfernen von Nullwerten, Standardisierung von Maßeinheiten usw.)?
- Wo werden diese Daten sonst noch verarbeitet, und wo werden zugehörige Daten gespeichert?
- Was gilt es hinsichtlich der Qualität und der Herkunft der Daten zu bedenken?

Als die Unternehmen ihre Dateninfrastrukturen entwickelten, gab es leider oft keine geeigneten Systeme, um die Herkunft der Daten zu ermitteln. Das Hauptaugenmerk lag darauf, die Daten möglichst effizient einzuspeisen und zu speichern – und nicht darauf, sie rückzuverfolgen und festzustellen, wie sie verarbeitet wurden. Datensysteme gehen mit sogenannten technischen Schulden (engl. *Technological Debt*) einher. Selbst wenn Sie keine Informationen über die Datenhistorie (Lineage-Daten) haben, ist es nicht zu spät, damit zu beginnen.

Je nachdem, wie fortschrittlich Ihre Dateninfrastruktur und Ihre technischen Systeme sind, gibt es vielleicht schon gute Voraussetzungen, um die Data Lineage zu dokumentieren. Diese Informationen können aus Systemen wie Apache Spark (oder Beam, Flink, Kafka und Airflow) oder anderen Pipeline-Automatisierungssystemen abgerufen und in die von Ihnen verwendeten Dokumentations- oder Tracking-Systeme integriert werden. Falls Sie sich nicht sicher sind, inwiefern dies bereits der Fall ist, sollten Sie sich mit den Teams in Verbindung setzen, die für die Verwaltung von Datenkatalogen und Datenschemata zuständig sind. Ein Datenkatalog (engl. *Data Catalog*) ist ein Verzeichnis mit der Dokumentation von Datenquellen, die im gesamten Unternehmen verwaltet und zur Verfügung gestellt werden. Dazu gehören oft auch die Dokumentation der Daten, die Zugriffsanforderungen und sogar Hinweise darauf, wie die Daten verarbeitet und gespeichert werden, sowie Informationen zur Qualität. Wenn Ihr Unternehmen derzeit keine Daten katalogisiert bzw. keine Informationen zur Datenhistorie nachverfolgt, sollten Sie sich mit einer Gruppe von Datenexperten zusammensetzen und bestimmen, welcher Ansatz für den Anfang am besten geeignet ist.

Als Data Scientist haben Sie es vermutlich schon einmal mit einem Datensatz zu tun gehabt, bei dem Sie Zweifel daran hatten, dass die Daten valide sind. Sollte es tatsächlich zu Fehlern bei der Erfassung oder Verarbeitung gekommen sein, können Sie dies nur herausfinden, wenn Sie die Datenhistorie (die Data Lineage) der Daten sorgfältig tracken. Mit dem Wissen, wann und wo der Fehler aufgetreten ist, können Sie und andere Teammitglieder feststellen, wodurch er verursacht worden sein könnte und ob es sich um einen Bug handelt. Insbesondere bei Streaming- oder echtzeitnahen Systemen ist es unerlässlich, die Herkunft der Daten und ihre Verarbeitung nachzuverfolgen, da sich Fehler bei der Datenerfassung und -transformation sehr rasch auf Modelle oder andere nachgelagerte Produkte auswirken können.

Es gibt inzwischen viele Tools, die Sie bei der Dokumentation Ihrer Daten, dem Lineage-Tracking und der Versionskontrolle unterstützen. Im Idealfall finden Sie ein geeignetes Toolset oder sogar ein einzelnes Tool, das Sie bei Ihren Governance-Bemühungen unterstützt. Obwohl ständig neue Produkte auf den Markt kommen, sind mir folgende Tools bekannt, die von Teams gut angenommen werden: DBT

(*https://oreil.ly/gBfKh*), CKAN (*https://ckan.org*), AWS Glue (*https://aws.amazon.com/glue*) und Tableau (mit seinem Data Catalog (*https://oreil.ly/rQzle*)).

Zur Überwachung von Änderungen in den Datenströmen empfehle ich Tools wie Great Expectations (*https://oreil.ly/F-SPq*), mit denen Sie Daten während des Vorgangs testen und feststellen können, ob es wesentliche Veränderungen gegeben hat. Great Expectations ermöglicht Ihnen, sogenannte *Data Unit Tests* zu schreiben, mit denen Sie überprüfen können, ob die Daten Ihren Erwartungen entsprechen. Sie können beispielsweise testen, dass ein bestimmter Wert kein Nullwert ist, ob ein Wert über oder unter einem bestimmten Wert liegt, ob ein Datumsstring richtig standardisiert wurde, oder auch, ob ein Wert ein String oder eine Ganzzahl (engl. *Integer*) ist. Diese Tests können – sofern sie von den Datenteams in geeigneter Weise durchgeführt werden – einen sofortigen Hinweis auf die oben genannten Bugs liefern.

Außerdem bietet Ihnen die Nutzung von Lineage-Daten klare Vorteile für den Datenschutz. In Kapitel 3 erfahren Sie noch mehr über die Rückverfolgung von Einwilligungen bzw. Einverständniserklärungen. Eine Pipeline für die Rückverfolgung von Daten zur Einwilligung von Nutzern und damit verbundenen Lineage-Daten könnte so aussehen wie in Abbildung 1-2. Zunächst wird dem Anwender eine Benutzeroberfläche angezeigt, auf der in verständlicher Form die einzelnen, fein abgestuften Datenschutzeinstellungen erklärt werden. Während die Daten erhoben werden, werden Einzelheiten zur Datenherkunft, z. B. wo, wann und wie die Daten erhoben wurden, als Datenbankfelder auf gleicher Ebene wie die Daten in die Datenstruktur aufgenommen statt als Anhang in einem separaten JSON-Dokument, das niemand verwendet.[5] Die Daten durchlaufen die üblichen Bereinigungs- und Transformationspipelines. Diese Daten werden dann im Hinblick auf Qualität und andere Governance-Standards analysiert. Darüber hinaus wird die Sensibilität der Daten und auch das Vorhandensein von PII auf halb automatische Weise analysiert. Dieser Schritt muss gegebenenfalls der erste Schritt sein, je nachdem, um welche Datenquelle es sich handelt und wie die Protokollierung bzw. das Logging in den Transformationsschritten erfolgt. Achten Sie bitte darauf, dass Sie keine sensiblen Daten loggen! Anschließend werden die Daten gespeichert, wobei die zusätzlichen Informationen in verknüpfbaren Datenstrukturen gespeichert werden. Dadurch wird sichergestellt, dass die zusätzlichen Governance-Daten leicht referenzierbar sowie stets auf dem neuesten Stand und mit den Nutzerdaten verknüpft sind, bis die Daten aus dem System entfernt werden.

5 Im Idealfall handelt es sich dabei entweder um eine separate, leicht verknüpfbare Tabelle, oder die Zeilen selbst haben zusätzliche Spalten oder Attribute, die das Auffinden erleichtern. In Kapitel 3 werden Sie noch ein konkretes Beispiel dafür kennenlernen.

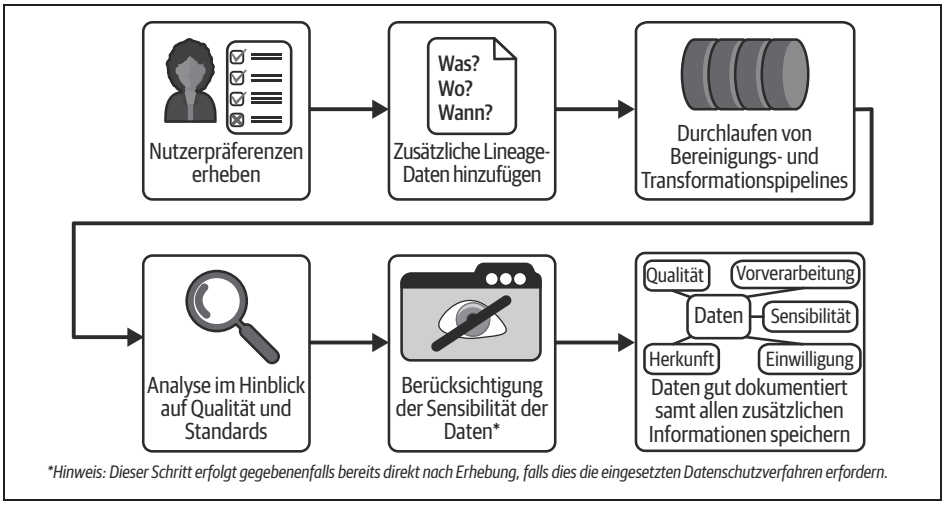

Abbildung 1-2: Eine Pipeline, mit der Daten zur Einwilligung und damit verbundene Lineage-Daten rückverfolgt werden können

Lineage adressiert auch die Gesetzgebung und Bestimmungen zur Datenhoheit, die im Laufe der Zeit immer mehr zuzunehmen scheinen. Diese Gesetze zielen darauf ab, die Daten von Einwohnerinnen und Einwohnern innerhalb eines bestimmten Hoheitsgebiets zu halten – z. B. die Daten von Bürgerinnen und Bürgern der EU auf einer Infrastruktur zu hosten, die sich in der EU befindet. Die Einführung von Datenhoheitskontrollen kann Ihre Rechts-, Compliance- und Sicherheitsteams bei der Überprüfung der Konformität Ihrer Dateninfrastruktur unterstützen.

Die Rückverfolgung der Datenherkunft und der Einwilligungen bedeutet zwar einen zusätzlichen Aufwand, vereinfacht aber die Fehlersuche und beschleunigt den späteren Zugriff auf die Daten. Wenn Sie eng mit einem Data-Engineering- oder Infrastrukturteam zusammenarbeiten, sollten Sie mit diesen klären, wie Sie den Arbeitsaufwand zur Einrichtung und Pflege der Lineage-Informationen untereinander aufteilen können. Genau wie bei Ihrer Datendokumentation sollten Sie sicherstellen, dass die Informationen lesbar, nachvollziehbar und nutzbar für die die Anwender sind. Wenn die Informationen in einer Datei gespeichert sind, in die niemand hineinschaut, lohnt es sich nicht, sie überhaupt zu erstellen und zu pflegen!

Versionskontrolle für Daten

In Ihrer täglichen Arbeit mit Daten werden Sie festgestellt haben, dass es immer wieder zu inkrementellen Änderungen an Daten kommt. Bei einem sehr großen Datensatz fallen diese Änderungen nicht auf. Sie haben jedoch erhebliche Auswirkungen auf Workflows, Analysen und Modelle, sei es, weil sich Datenerfassung und -transformation wesentlich verändert haben oder weil externe Faktoren die eingehenden Daten beeinflussen, z. B. während einer globalen Pandemie.

Genau an dieser Stelle kommt die Datenversionierung ins Spiel. Unter Datenversionierung versteht man die Möglichkeit, Versionen bzw. sogenannte Checkpoints für Ihre Daten zu erstellen, die an einen bestimmten Zeitpunkt gebunden sind. Ähnlich wie bei Systemen zur Versionskontrolle von Code (z. B. `git`) können Sie mit der Datenversionskontrolle alle Änderungen verfolgen, indem Sie zu einem bestimmten Zeitpunkt einen Snapshot bzw. einen *Commit* erstellen. Dieser kann dann mit früheren oder späteren Versionen verglichen werden, um nachzuvollziehen, wie sich die Daten verändert haben.

Ähnlich wie in der agilen Softwareentwicklung, bei der Tests, Continuous Integration (CI), Continuous Delivery (CD) und eine Versionskontrolle für Software zum Einsatz kommen, profitiert Data Science von einer Datenversionskontrolle. Stellen Sie sich vor, Sie könnten genau feststellen, wann eine Änderung oder ein Fehler aufgetreten ist, und so herauszufinden, warum ein Modell nicht wie erwartet arbeitet oder warum ein bestimmter Report oder ein Analysetool nicht mehr funktioniert.

Wenn Sie wissen, wie sich die von Ihnen gesammelten Daten im Laufe der Zeit verändern, kann dies für das Verständnis des Verhaltens oder der Systeme, die Sie modellieren, extrem hilfreich sein. Wie verändern sich die Nutzer Ihrer Anwendung im Laufe der Zeit? Wie haben sich die Ergebnisse Ihrer Experimente verändert? Welche Annahmen haben Sie ursprünglich in Bezug auf die Daten getroffen, und wie haben sich diese bewährt? Bei vielen der von Ihnen aufgeworfenen Fragen zu den Daten kann es von Vorteil sein, diese regelmäßig zu überprüfen, um die Daten selbst und ihre Veränderungen im Laufe der Zeit besser zu verstehen. Änderungen in der Software, an den Pipelines oder anderen Verarbeitungsschritten können zu erheblichen Veränderungen in den Daten und somit zu Fehlern in den Daten führen. Daher ist es wichtig, ein Monitoring der Daten und der Software vorzunehmen, da diese Veränderungen negative Auswirkungen auf andere Datenmodelle, Analysen und Systeme haben können.

Die Datenversionierung trägt auch zu Datenschutz und vertrauenswürdigen KI-Praktiken bei. Wenn Sie feststellen können, welche Daten für ein bestimmtes Modell verwendet wurden, wenn Sie die Zeitpunkte vor und nach der Löschung der Daten einer Person genau bestimmen können und wenn Sie Änderungen, die Sie zum Schutz der Privatsphäre vorgenommen haben, transparent nachweisen können, können Sie sicherstellen und überprüfen, dass die Daten im System verantwortungsvoll verwendet werden. Eine Versionskontrolle der Daten ermöglicht Ihnen, die Ergebnisse Ihrer Datenschutzmaßnahmen nachzuvollziehen, und unterstützt Sie bei der Beantwortung der oben genannten Fragen. Dies hilft Ihnen, die Ergebnisse im Fall eines Fehlers zu debuggen.

Wie sollten Sie nun mit der Versionierung der Daten beginnen? Es gibt zahlreiche Tools. Mithilfe der folgenden Fragen können Sie die für Sie geeigneten Tools finden:

- Wie verwaltet das Tool Snapshots und Checkpoints für die Daten? Dies sollte auf eine Weise geschehen, die sowohl programmierbar als auch leicht nachzuvollziehen ist. Wenn Sie Daten umgehend wiederherstellen müssen, möchten Sie sich nicht erst durch die Dokumentation wühlen!

- Beanspruchen die Snapshots bzw. Versionen möglichst wenig Speicherplatz? Wie können Sie den erforderlichen Speicherplatz verwalten? Ein eher unbedarfter Ansatz wäre, Ihre gesamte Datenbank jeden Tag zu kopieren und eine alte Kopie zu speichern. Das ist gut, wenn Sie über unbegrenzten Speicher und Platz verfügen und nur ein paar Hundert Zeilen haben, allerdings ist das eher die Ausnahme. Dementsprechend sollten Sie abwägen, wie viele Snapshots Sie speichern, wie viel zusätzlichen Speicherplatz und Rechenleistung diese benötigen und wann Sie ältere Snapshots löschen.
- Lässt sich die Software gut mit anderen Teams und deren Workflows integrieren? Wie bei allen Entscheidungen in Bezug auf Software sollten Sie sicherstellen, dass das, was für Ihr Team funktioniert, auch anderen Teams nützt, die mit den Daten arbeiten. Klären Sie mit den Data Engineers und den Softwareentwicklern, ob sie ebenfalls wissen, wie man Daten wiederherstellt. Sie sollten sicherstellen, dass sie sowohl die Programmierschnittstelle (API) als auch die verwendete Programmiersprache verstehen und wissen, wie das Tool verwendet wird.
- Können Sie sich vorstellen, wie dieses Tool zum Einsatz kommen würde? Geben Sie Ihrem Team die Möglichkeit, das Tool auszuprobieren, und formulieren Sie ein paar Fallbeispiele, mit denen Sie relevante Anwendungsfälle vorprogrammieren können. Wie würde eine Wiederherstellung von Daten nach einer Änderung des Schemas erfolgen? Wie kann beantwortet werden, welche Daten für ein bestimmtes Modell genutzt wurden, das trainiert und eingesetzt wurde? Was ist, wenn ein Löschungsantrag gemäß DSGVO eingeht und Sie nachweisen möchten, dass die Daten ordnungsgemäß gelöscht wurden? Nehmen Sie sich Zeit, um gemeinsam mit dem Datenteam eine umfassende Aufstellung zu erarbeiten, und stellen Sie sicher, dass die Anwendungsfälle, für die Sie die Daten verwenden möchten, gut verstanden werden und gegebenenfalls sogar programmiert wurden.
- Kann das Tool im Zusammenhang mit Data Lineage genutzt werden, um Daten für bestimmte Anwendungsfälle besser auswählen zu können? Wie Sie bereits in diesem Kapitel gelernt haben, können Sie mithilfe der Datenhistorie feststellen, ob ein bestimmter Datensatz für eine gegebene Aufgabe wirklich geeignet ist. Dank dieser Informationen und der Versionskontrolle können Sie Modelle und Experimente schneller erstellen, da Sie die Daten besser einschätzen und systematische Veränderungen bzw. Abweichungen in den Daten (engl. *Shifts*) so früh wie möglich aufdecken können.

Die Versionierung von Daten ist eng mit der Versionierung von Modellen verbunden. Die oben genannten Fragen können auch in Bezug auf die Auswahl von Tools für die Modellversionierung gestellt werden, und mehrere Open-Source-Bibliotheken bieten inzwischen beide Möglichkeiten an. Wie auch immer Ihr Vorhaben aussieht und wie auch immer Ihr Team am besten arbeitet – denken Sie darüber nach, die Versionierung von Daten und Modellen in Ihre normalen Workflows zu integrieren. Diese Praktiken sind aus der Softwareindustrie bekannt und können dazu beitragen, dass die Data-Science-Arbeit in Ihrem Unternehmen vorhersehbarer, fehler-

freier und besser verstanden wird. Ich würde jedem Team empfehlen, sie jetzt einzuführen und zusammen mit der Data Governance weiterzuentwickeln, auch wenn die ersten Jahre einen Lernprozess bedeuten.

Außerdem wird die Möglichkeit zur Versionskontrolle von Datensätzen zunehmend auch direkt in Data-Lake- und Data-Warehousing-Tools unterstützt. Wenn Ihr Unternehmen bereits umfassende Tools zum Datenmanagement einsetzt, sollten Sie zunächst prüfen, inwieweit diese unterstützt werden bzw. integrierbar sind, ehe Sie eine weitere Bibliothek einführen.[6]

Wie Sie wahrscheinlich wissen, ändert sich die Landschaft der Datentools relativ schnell. Daher ist es wichtig, dass Sie sich über neue Tools informieren und einige davon vergleichen, um herauszufinden, welches am besten für Ihre Zwecke geeignet ist! Verwenden Sie diese Fragen als Leitfaden für Ihre Bewertung und Ihre Auswahlkriterien und scheuen Sie sich nicht, einige Proof-of-Concept-Implementierungen durchzuführen, bevor Sie sich für ein bestimmtes Tool als Standard entscheiden.

Grundlegender Datenschutz: Pseudonymisierung beim Privacy by Design

Sie haben inzwischen gelernt, was Data Governance ist, wie Sie sensible Daten finden und bewerten, wie Sie Daten im Hinblick auf ihre Sensibilität dokumentieren und wie Sie mithilfe von Data-Lineage-Tracking und Versionskontrolle herausfinden, wenn sich etwas ändert. Damit sind Sie nun in der Lage, Datenschutztechniken für personenbezogene Daten auf gut dokumentierte und wiederholbare Weise anzuwenden.

Beginnen Sie mit einfachen Maßnahmen. Manchmal ist der einfachste Ansatz bereits die Lösung des Problems und kann viele interne und externe Bedenken ausräumen. Im weiteren Verlauf werden Sie noch erfahren, wie Sie ermitteln können, welche Datenschutztechnologien und -verfahren für die jeweiligen Risiken geeignet sind und welche Vorteile sich daraus ergeben (z. B. eine Erweiterung des Datenzugriffs).

Die *Pseudonymisierung* eignet sich hervorragend für grundlegende Datenschutzanforderungen, z. B. wenn Sie mit Daten arbeiten, die niemals jemandem außerhalb einer Gruppe vertrauenswürdiger Mitarbeitenden zugänglich gemacht werden sollen. Pseudonymisierung ist eine Methode, die es ermöglicht, »Pseudonyme« anstelle von echten Namen und Daten zu verwenden. Es gibt verschiedene Ansätze zur Pseudonymisierung, die in Tabelle 1-2 aufgeführt sind.

6 Wenn Ihr Unternehmen ein älteres On-Premise-System nutzt, das diese Funktionen nicht bietet, könnten Sie über eine einfache Lösung nachdenken, z. B. regelmäßige Snapshots von Daten, die für bestimmte Aufgaben verwendet werden können, und Tools, mit denen Sie die Snapshots bei Bedarf einfach laden oder austauschen können. Viele Versionierungstools, wie etwa DVC, erweitern auch ihre Unterstützung für selbst gehostete und On-Premise-Systeme.

Tabelle 1-2: Verschiedene Ansätze zur Pseudonymisierung

Pseudonymisierungsansatz	Beschreibung	Beispiel
Maskierung	Die Daten werden mit einer »Maske« versehen, d. h., die Werte werden meist durch eine Reihe von Standardwerten ersetzt.	888-23-5322 → <ID-NUMMER> oder <XXX-XX-5322>
Tokenisierung (tabellenbasiert)	Tokens, durch die eine eindeutige Identifizierung möglich ist, werden mittels einer Nachschlagetabelle, die eine Eins-zu-eins-Ersetzung ermöglicht, ersetzt.	Mondo Bamber → Fiona Molyn
Hashing	Die Daten werden mithilfe eines Hashing-Mechanismus weniger interpretierbar gemacht, können aber weiterhin zugeordnet werden.	foo@bar.com (*mailto:foo@bar.com*) → 32dz22945nzow
Formaterhaltende Verschlüsselung	Mithilfe eines Verschlüsselungs- oder eines anderen kryptografischen Verfahrens werden die Daten durch ähnliche Daten ersetzt. Oftmals können diese auch miteinander verknüpft werden.	(0)30 4344 3333 → (0)44 4627 1111

Wie Sie vielleicht schon in Tabelle 1-2 gesehen haben, können diese Ansätze die Qualität Ihrer Daten sowie die Privatsphäre des Einzelnen erheblich beeinträchtigen. Der Hashing-Mechanismus nimmt beispielsweise etwas, das leicht als E-Mail-Adresse zu erkennen ist, und wandelt es in etwas um, das nicht mehr interpretiert werden kann. Dadurch wird zwar ein Mindestmaß an Datenschutz gewährleistet, aber gleichzeitig die Möglichkeit genommen, nützliche Informationen abzuleiten (z. B. die Zuordnung von E-Mail-Konten anhand der Domain). Bei der Maskierung (engl. *Masking*) werden je nach Art der Implementierung entweder alle Informationen entfernt, die die Identifizierung einer Person ermöglichen, oder es verbleiben zu viele Informationen, die leicht mit anderen Datensätzen verknüpft werden können, um personenbezogene Informationen aufzudecken. Bei der tabellenbasierten Tokenisierung (engl. *Table-based Tokenization*) wird eine Lösung verwendet, die möglicherweise nicht mit Ihren Daten skaliert, aber eine geeignete und für Menschen lesbare Zuordnung ermöglicht, wenn Sie verschiedene Datensätze miteinander verknüpfen bzw. verlinken müssen.

Wie Sie noch in Kapitel 4 erfahren werden, ist die Verknüpfung (engl. *Linking*) von Daten, auch Daten-Linkage genannt, ein wesentlicher Angriffsvektor, mit dem die Identität einer Person ermittelt werden kann. Je mehr Daten Sie verknüpfen können, desto einfacher ist es, mithilfe dieser verknüpften Informationen auf die Person zu schließen oder ausreichend über sie zu erfahren, um eine konkrete Vermutung hinsichtlich der Identität anstellen zu können. Bei der formaterhaltenden Verschlüsselung (engl. *Format-preserving Encryption*) besteht weiterhin die Möglichkeit, Verknüpfung herzustellen, allerdings lässt sie sich besser skalieren, da ein auf kryptografischen Verfahren basierender bidirektionaler Standardmechanismus verwendet wird. Häufig lässt sich die Verknüpfbarkeit in festgelegten Zeitabständen beseitigen, indem die geheimen Schlüssel ausgetauscht werden. Dies kann im Rahmen interner Anwendungs-

fälle ausreichend Sicherheit bieten, vorausgesetzt, es werden geeignete Standardwerte verwendet. Wenn Sie Wert darauf legen, dass Ihre Daten verknüpfbar sind, sollten Sie auch die verschiedenen Techniken im Bereich Privacy-preserving Record Linkage (PPRL) (*https://oreil.ly/wjU2i*), d.h. datenschutzfreundlicher Verfahren zum Daten-Linkage, in Betracht ziehen, die neben diesen Pseudonymisierungsverfahren auch verschiedene wahrscheinlichkeitsbasierte Hashing-Verfahren umfassen.

In Tabelle 1-3 sind die Vor- und Nachteile der Pseudonymisierung zusammengefasst.

Tabelle 1-3: Die Vor- und Nachteile der Pseudonymisierung

Vorteil	Nachteil
Verknüpfbar (engl. *Linkable*): Pseudonymisierungstechniken lassen oft die Möglichkeit offen, Daten miteinander zu verknüpfen. Dies ist nützlich, wenn Sie Datensätze mit persönlichen Identifikatoren oder anderen Spalten mit sensiblen Attributen verknüpfen möchten.	Pseudonymisierung ist nicht mit Anonymisierung gleichzusetzen. Die Re-Identifizierung pseudonymisierter Daten durch Linkage Attacks ist eine der größten und beständigsten Bedrohungen für den Schutz der Privatsphäre und kann umso leichter durchgeführt werden, je mehr Daten verfügbar sind (mehr dazu in Kapitel 4).
Formaterhaltende Eigenschaft: Verschiedene Pseudonymisierungsverfahren ermöglichen Ihnen, das Format der Daten beizubehalten oder den ursprünglichen Zweck der Daten zu erkennen (handelt es sich z.B. um eine E-Mail?). Dies ist besonders dann von Vorteil, wenn Sie nicht genau wissen, woher die Daten ursprünglich stammen oder wie sie aufgebaut sind.	Alle Informationen, die in den pseudonymisierten Daten enthalten sind, erhöhen den Informationsgehalt und damit das Risiko, falls die Daten veröffentlicht oder versehentlich offengelegt werden. Wäre eine Datendokumentation ein besserer Ansatz, um das zugrunde liegende Schema zu verstehen?
Privacy-by-Design-Verfahren: Pseudonymisierung ist eine Alternative zur Verwendung von Rohdaten, die von Rahmenwerken wie Privacy by Design vorgeschlagen wird, insbesondere wenn es sich um sensible Daten handelt.	Simple Verfahren wie die Pseudonymisierung können ein falsches Gefühl der Sicherheit vermitteln, sodass die Wahrscheinlichkeit steigt, dass die Daten weitergegeben werden oder dass behauptet wird, die Daten seien »anonymisiert«, weil persönliche Identifikatoren entfernt bzw. pseudonymisiert wurden.

Meiner Erfahrung nach ist das Hauptargument gegen die Pseudonymisierung, dass sie fälschlicherweise das Gefühl vermittelt, dass die Daten sicher und geschützt seien. Ich habe erlebt, wie Teams mit Datenschutzlösungen Schwierigkeiten hatten und zu dem Schluss gekommen sind, dass eine Pseudonymisierung völlig ausreichend ist, da die Daten dadurch »anonymisiert« werden. Dabei handelt es sich leider um einen weitverbreiteten Irrtum. In Kapitel 2 erfahren Sie, was Anonymisierung tatsächlich bedeutet – und was nicht. Aber ich kann Ihnen versichern, dass keine noch so gute Pseudonymisierung Ihnen die gewünschte Anonymisierung bringen wird, sofern diese das empfohlene Verfahren ist.

Wenn Sie jedoch sicherstellen können, dass die Daten nur intern von einer begrenzten Anzahl von Personen verwendet werden, die gegebenenfalls einen privilegierten Zugang benötigen, könnte die Pseudonymisierung eine gute Lösung sein. Denkbar wäre dies z.B. für interne Vertriebs- oder Kundenbetreuungsteams, die Zugang zu bestimmten Kundendaten benötigen, aber wahrscheinlich nicht zu allen. Weitere

Anwendungsfälle könnten interne Business-Intelligence-(BI-) und Analyse-Dashboards sein, die Daten verknüpfen müssen, aber keinen direkten Zugriff auf sensible Informationen haben sollten.

In beiden Fällen gibt es jedoch klare Alternativen zur Pseudonymisierung. Denken Sie z. B. an ein BI-Dashboard, bei dem eine Übersicht der Bestellungen über mehrere Regionen hinweg dargestellt wird. Die dafür erforderlichen Abfragen zielen nur auf die aggregierten Werte der einzelnen Regionen ab, weshalb ein gewisses Maß an Datenschutz – sofern die Regionen eine gewisse Größe aufweisen – gewahrt bleibt. Oder stellen Sie sich ein Kundensupportsystem vor, bei dem nur die Felder sichtbar sind, die für die Ausführung der jeweiligen Aufgabe erforderlich sind.

Ich habe oft gesehen, dass Pseudonymisierung als Mechanismus verwendet wird, um Daten aus Produktionssystemen zu extrahieren und sie in Testumgebungen (zur Softwareentwicklung und für Analysetools) zu verwenden oder um etwaige Fehler in einem sicheren System zu loggen. Dies hat den Vorteil, dass die Daten den Daten in der Produktion stark ähneln, es sich aber nicht um die tatsächlichen Rohdaten, die in der Produktion verarbeitet werden, handelt. Es ist extrem riskant, diese Produktionsdaten in einer Testumgebung zu verwenden, denn sie sind nur wenig geschützt, und Testumgebungen sind häufig ungesichert. Wenn es bei einem Test auf die Merkmale der Produktionsdaten ankommt (z. B. bei Tests von Modellen), empfehle ich Ihnen, einen Weg zu finden, diese Merkmale synthetisch in Ihren Testdaten zu erzeugen, anstatt echte Produktionsdaten zu verwenden, die nur minimal geschützt sind.

Sollte die Pseudonymisierung von Ihren internen Stakeholdern befürwortet werden und Sie das Risiko als gering genug erachten, um sie zu verwenden, gibt es mehrere Tools und Bibliotheken, die Sie ausprobieren können. Sie finden im zum Buch gehörigen Repository (*https://github.com/kjam/practical-data-privacy*) ein Notebook, in dem ich Ihnen einige einfache und unterhaltsame Beispiele für Pseudonymisierung zeige.

An dieser Stelle stelle ich einen Beispielworkflow mit Hashicorp Vault (*https://oreil.ly/xvvlt*) vor. Hashicorp Vault ist ein Dienst, der von Infrastrukturteams zum anwendungsübergreifenden Management von Secrets verwendet wird. Bei der Verschlüsselung werden Daten in einen geheimen Code (den sogenannten Chiffretext) und wieder zurück (Klartext) übersetzt, um so einen sicheren Zugriff zwischen zwei oder mehr Parteien zu ermöglichen. Mit gemeinsam genutzten Protokollen und Verschlüsselungsalgorithmen können Nutzende die Dateien oder Nachrichten verschlüsseln, die nur für andere ausgewählte Clients zugänglich sind. Obwohl es verschiedene Verschlüsselungsverfahren gibt, haben alle die Fähigkeit gemeinsam, Daten mithilfe eines kryptografischen Schlüssels zu ver- und zu entschlüsseln. Dieser eindeutige Schlüssel ist eine zufällige Zeichenfolge, die speziell für die Verschlüsselungstransaktion erzeugt wird. Je mehr Bits und je komplexer das Verfahren, desto besser ist es. Dies ist ein gängiges Muster für den Aufbau von Microservices, bei denen viele Anwendungen und Dienste in Containern bereitgestellt werden und auf sensible Daten wie API-Schlüssel, Chiffrierschlüssel oder Identitätsdaten auf eine skalierbare und sichere Weise zugreifen müssen.

Wenn Sie das formaterhaltende Verfahren verwenden möchten, erstellen Sie zunächst ein Muster für einen regulären Ausdruck (engl. *Regular Expression*), das dem gewünschten Format entspricht. Ein regulärer Ausdruck für eine Kreditkartennummer könnte wie folgt aussehen:

```
\d{4}-\d{2}(\d{2})-(\d{4})-(\d{4})
```

Dieses Muster können Sie als Vorlage für eine bestimmte Transformation in Hashicorp hinterlegen und verschiedene Rollen zuweisen (d. h. festlegen, wer diese Transformation verwenden darf). Hashicorp unterstützt bereits unterschiedliche Arten der formaterhaltenden Verschlüsselung. Achtung! Manche dieser Methoden können rückgängig gemacht werden, andere nicht!

Jetzt können Sie die Transformation über das Kommandozeilen-Interface testen, und Sie werden sehen, dass Sie eine falsche, aber dem Muster nach gültige neu generierte Kreditkartennummer zurückerhalten:

```
$ vault write transform/encode/payments value=1111-2222-3333-4444

Key              Value
---              -----
encoded_value    1111-2200-1452-4879
```

Wenn Sie eine umkehrbare Methode verwendet haben, können Sie auch überprüfen, ob Sie den Wert mit der geeigneten Rolle und der entsprechenden Berechtigung korrekt entschlüsseln können:

```
$ vault write transform/decode/payments value=1111-2200-1452-4879

Key              Value
---              -----
decoded_value    1111-2222-3333-4444
```

Da sich die API-Aufrufe ändern können, sollten Sie dies mit dem für Ihr Unternehmen zuständigen Infrastruktursupport für Hashicorp abklären. Ich empfehle Ihnen, einen Blick in die aktuelle Hashicorp-Dokumentation zu werfen, um herauszufinden, ob Hashicorp die richtige Lösung für Ihre Bedürfnisse im Bereich Pseudonymisierung ist.

Es gibt auch mehrere Open-Source-Bibliotheken, die formaterhaltende Verschlüsselung sowie andere Pseudonymisierungstechniken wie Hashing, Maskierung oder Tokenisierung unterstützen. Zum Zeitpunkt der Erstellung dieses Buchs sind unter anderem folgende Bibliotheken verfügbar, die eine gute Dokumentation und nützliche Funktionen bieten:

- Kodex von KIProtect (*https://heykodex.com*) verfügt über eine Open-Source-Community-Edition, die mehrere Pseudonymisierungstechniken unterstützt.[7]

7 Disclaimer: Ich war Mitgründerin von KIProtect und war an den ersten Implementierungen dieser Bibliothek beteiligt. Mittlerweile wirke ich jedoch nicht mehr an dem Unternehmen oder der Bibliothek mit.

- Die Python-basierte Format-preserving-Encryption-Bibliothek von Mysto (*https://oreil.ly/RKvpV*) ermöglicht Ihnen, verschiedene formaterhaltende Algorithmen über eine einfach zu bedienende Python-Schnittstelle aufzusetzen.
- Presidio von Microsoft (*https://oreil.ly/Tao7Z*) bietet Ihnen eine Reihe an Möglichkeiten zur Maskierung und Tokenisierung sowie Methoden, die es Ihnen ermöglichen, PII in Textdaten zu identifizieren.
- Private Input Masked Output (PIMO) (*https://oreil.ly/LDA_3*) nutzt eine Go-basierte Engine und verfügt über zahlreiche Vorlagen, mit denen Daten pseudonymisiert und maskiert werden können.

Wenn Sie sich mit den fortgeschrittenen Techniken und den potenziellen Gefahren, die in diesem Buch beschrieben werden, auseinandersetzen, werden Sie in der Lage sein, die Bedeutung von Risiko und Benutzerfreundlichkeit für Ihr Team und Ihre Arbeit besser einzuschätzen. Sie werden verstehen, wann Pseudonymisierung angemessen ist und wann es besser ist, das Problem mit einer fortschrittlicheren und schützenderen Technik anzugehen.

Zusammenfassung

In diesem Kapitel haben Sie erfahren, was Data Governance ist und wie Sie als Data Scientist davon Gebrauch machen können und sollten. Sie haben gelernt, wie Data Governance mit Data Science und Datenschutz zusammenhängt, z. B. wie Sie undokumentierte Daten aufspüren, sensible Daten identifizieren, die Datendokumentation verwalten und die Datenhistorie (Lineage) nachverfolgen können. Darüber hinaus haben Sie einige einfache Datenschutzansätze bei der Arbeit mit sensiblen Daten kennengelernt, wie z. B. die Pseudonymisierung.

Sie sollten nun das Gefühl haben, den ersten Schritt in die richtige Richtung gegangen zu sein. Sie sind dabei, sich einen Überblick über Privacy zu verschaffen und zu erkennen, was für Ihre Arbeit wichtig und relevant ist. Vielleicht haben Sie bereits Fragen dazu, wie Sie Datenschutzrisiken in der Praxis erkennen und handhaben können oder wie Sie sicherstellen können, dass Sie Datenschutztechniken auch wirklich effektiv einsetzen. Dann habe ich gute Nachrichten für Sie: Im nächsten Kapitel erfahren Sie, wie Sie diese Fragen mit wissenschaftlichen Methoden beantworten können. Widmen wir uns nun also der Differential Privacy!