DATA SCIENCE ZUSAMMENFASSUNG

WOCHE 2 – AGILE DATA PRODUCT ENTWICKLUNG

Lernziele: Sie

- kennen Grundprinzipien agiler Arbeitsmethoden
- und die Phasen des Data Science-Prozesses
- können die Herausforderungen in der Entwicklung datenbasierter Produkte mit agilen Prinzipien verbinden
- beherrschen grundlegende Funktionen in Excel für Datenanalyse-Aufgaben

PRINZIPIEN AGILE

Agile Manifesto

Ein Manifesto von Softwareentwicklern, die eine andere Arbeitsmethode verwenden wollen, das heute weit verbreitet ist. Agile ist neuer und unterscheidet sich von der alten Wasserfallmethode, die heute auch noch eingesetzt wird.

- Individuen und Interaktionen
  - Vs Prozesse und Werkzeuge
- Funktionierende Software
  - Vs umfassende Dokumentation
- Zusammenarbeit mit Kunden
  - Mehr als Vertragsverhandlungen
- Reagieren auf Veränderung
  - Mehr als Befolgen eines Pans

Selbstorganisation

- Es braucht motivierte Mitarbeiter, damit die Arbeit erfolgreich ist
- Management soll unterstützen und eine optimale Umgebung schaffen, nicht herumkommandieren und Vorgaben machen
- Management soll dem Team vertrauen, nicht das Team überwachen
- Trotzdem braucht es Regeln und Hierarchien, einfach nicht so fest, dass es dann gar keine Flexibilität mehr gibt → Spagat zwischen Freiraum und Struktur

Nutzerbedürfnisse

- Nutzer des Endprodukts und ihre Erfahrung und Bedürfnisse werden in den Fokus gerückt über den ganzen Entwicklungsprozess hinweg

Unsicherheit als Chance

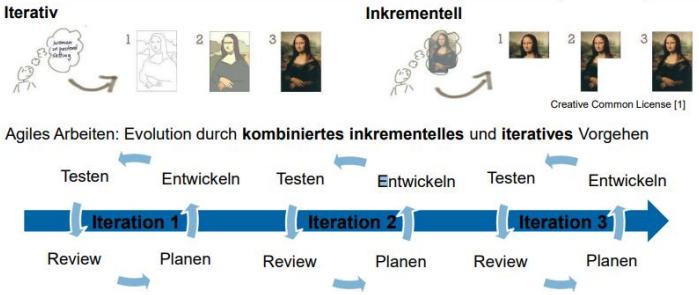
- Es gibt immer Unsicherheiten in der Planung von komplexen Prozessen, wo man nicht genau weiss, was passieren wird → das ist aber okay so
- Unsicherheit ist nicht etwas Schlechtes, das man nur eliminieren oder vermeiden muss. Mut zu Fehler.
- Veränderungen sind ein Chance für Wettbewerbsvorteil → man kann schnell reagieren auf neue Nachfragen und Bedürfnisse
- Das heisst auch Veränderungen im fortgeschrittenen Entwicklungsstadium des Produktes

Priorisierung beim Vorgehen

- Agile:
    - erste Prio sind die Ressourcen: Zeit und Mitarbeiter. Das wird zu Beginn gefixt
    - als zweites wird der Umfang des Projektes bestimmt und zwar abhängig davon, was mit der gegebenen Zeit und den gegebenen Mitarbeitern möglich ist
- Wasserfall (die traditionelle Arbeitsweise)
    - Erste Prio ist der Umfang des Projektes
    - Als zweites wird, abhängig vom Umfang des Projektes bestimmt, wie viel Zeit und wie viele Mitarbeiter notwendig sind

Inkrementelles und iteratives Vorgehen

- Die agile Methode kombiniert inkrementelle und iterative Arbeitsweise, indem sie zwar das Projekt in Teilstücken angehen (**inkrementell**) aber in diesen Teilstücken immer wieder Schritt für Schritt planen, entwickeln, testen, reviewen, und planen, … (**iterativ**)



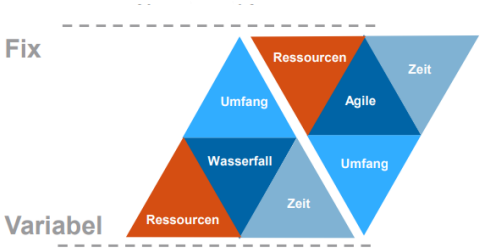
AGILE VS WASSERFALL

Für was ist Agile geeignet?

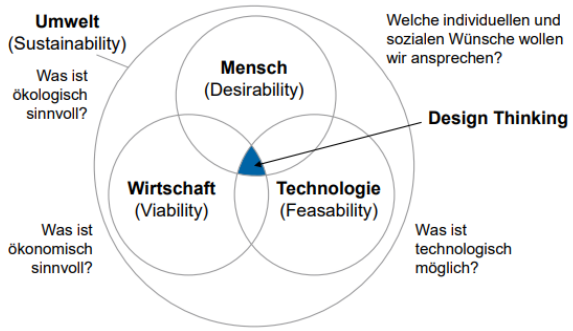
- Je mehr man über den Prozess noch nicht weiss, desto komplexer ist das Problem
- Je mehr man über die Bedingungen und technologische Umsetzung weiss, desto simplet ist das problem
- Agile ist geeignet für komplexere Probleme
- Bei simplen Problemen, wo die Bedingungen und Umsetzung glasklar ist (zum Beispiel weil es gesetzliche Vorgaben gibt, wie bei der Buchhaltung) bringt es nichts agil zu arbeiten, sondern besser Wasserfall

Wasserfall

- Alles ist viel organisierter, hierarchischer und streng vorgegeben
- Sequentiell / lineare Schritte
- Wasserfall (die traditionelle Arbeitsweise)
    - Erste Prio ist der Umfang des Projektes
- Als zweites wird, abhängig vom Umfang des Projektes bestimmt, wie viel Zeit und wie viele Mitarbeiter notwendig sind
- Keine Änderungen während dem Prozess, keine Unsicherheiten
- Benutzerbedürfnisse werden nur ganz am Anfang abgefragt und nicht immer wieder während dem Prozess durch Testing und Feedback. Jedes Mal wenn Nutzer etwas anders wollen, das von den ursprünglichen Anforderungen abweicht, ist das ein neuer Auftrag der mehr kostet

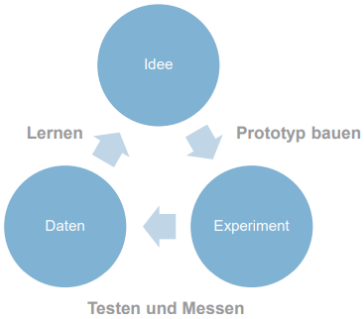


MODELLE AGILE

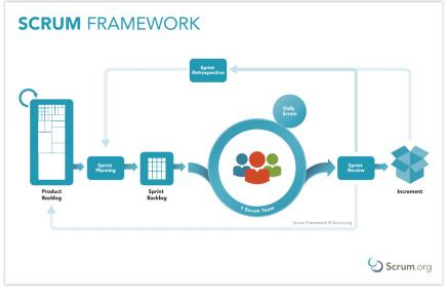


Design Thinking

Lean Startup



Scrum

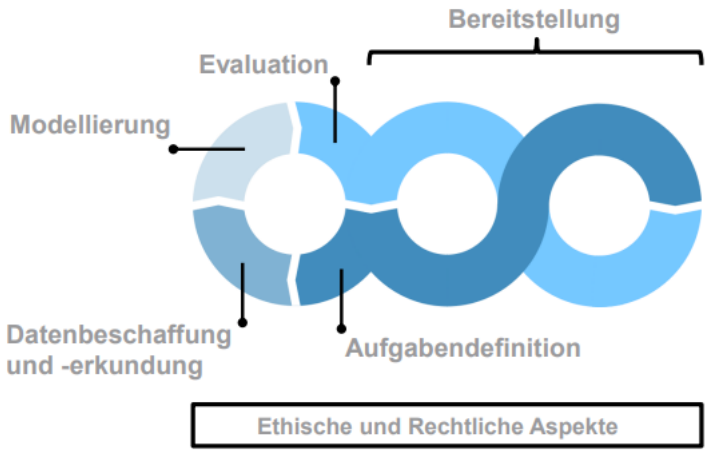


Kanban

Ein Kanban-Board hilft
- Aufgaben zu visualisieren
- Die Anzahl laufender Aufgaben (WIP) zu begrenzen
- Den Workflow zu optimieren

WIP-Limit:	Backlog	In progress 2	Done
Swimlane 1	Aufgabe Aufgabe Aufgabe	Aufgabe	
Swimlane 2	Aufgabe	Aufgabe	Aufgabe
Swimlane 3	Aufgabe Aufgabe Aufgabe Aufgabe		

DATA PRODUKTENTWICKLUNG



DATA PRODUKT ENTWICKLUNG

Lernziele: Sie

- kennen die Ziele, Tätigkeiten und Artefakte des Schrittes "Aufgabendefinition" im Data Science-Prozess

- können in einem Projekt
  - die Problemstellung beschreiben
  - Beschränkungen, Randbedingungen und Risiken erfassen
  - Projektziele formulieren, Kenngrössen und passende Zielwerte ableiten
  - einen Zeitplan erstellen
  - und in einer Projektcharta dokumentieren

AUFGABENDEFINITION

Sie kennen die Ziele, Tätigkeiten und Artefakte des Schrittes "Aufgabendefinition" im Data Science-Prozess

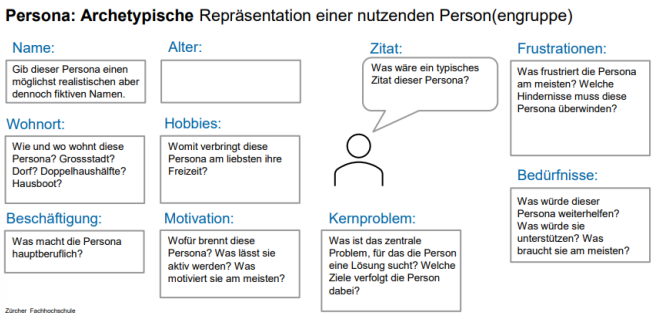
Zur Aufgabendefinition gehören	Output/Artefakt
1. Beschreibung der Problemstellung 2. Situationsbewertung 3. Projektziele und Erfolgskriterien 4. Modellierungsziele 5. Projektplan	Projektcharta

1 BESCHREIBUNG DER PROBLEMSTELLUNG

Problem aus gesellschaftlicher Sicht verstehen

- Am Anfang ist alles sehr schwammig und es ist unklar, was überhaupt das Problem ist und wofür man eine Lösung sucht.
- Darum muss man:
    - **Domänenwissen** aufbauen → Bescheid wissen über den Bereich vom Problem (zum Beispiel Bescheid wissen über verschiedene Restaurant Typen, Preisklassen)
    - **Eigenschaften der Nutzerinnen** analysieren (zum Beispiel verschiedene Konsumentengruppen für Resaurants, ihre Kaufkraft, ihre Präferenzen)
    - **Anspruchsgruppen** identifizieren
- Nutzer*innen:
    - Man muss die Nutzer*innen auch wirklich kennen, damit man Produkte erstellt, die auf sie zugeschnitten sind. Sonst entwickelt man an ihren Wünschen vorbei. Darum erstellt man eine Persona: der/die typische Nutzer*in ist die Persone

Persona



- Umfassende Beschreibung der **nutzenden** Personen
- zB: Firmenintern / Firmenextern (wenn man ein Produkt für eine Firma erstellt)
- zB: jung/ alt → Altergruppen, urban/ländlich, reich/alt → Konsumverhalten, etc.
- anhand von dieser Persona, die man schafft, kann man sich überlegen, was für diese Person wichtig ist, wie sie sich verhält (zB eine ältere Person ist nicht so digital unterwegs und man muss ein Produkt vielleicht sehr einfach gestalten, oder eine firmeninterne Person ist bestens vertraut mit Fachbegriffen und Konzepten und man kann ein Produkt auch komplizierter gestalten, etc.)

Stakeholder

- die unterschiedlichen Gruppen, welche irgendwie von dem Produkt beeinflusst, werden in ihrem (Berufs-) Alltag.
- Man definiert, welche Gruppen alles irgendwie mit dem Produkt zu tun haben werden und kann so auch Schnittstellen Fragen angehen
- Das schafft auch Transparenz: wenn man sich gründlich überlegen und auflisten muss, wer alles davon betroffen sein wird, hilft das versteckte Beziehungen und Motivationen offenzulegen. So vermeidet

man **«non-technical failure»**, das heisst wenn das Projekt aus Gründen scheitert, die mit dem zwischenmenschlichen oder politischen zu tun haben, und nicht der Technologie.

- Der Output ist die Projektcharta

Projektcharta

- So wie die ursprüngliche Abmachung. Wird vom Entwicklerteam geschrieben und von Sponsor*in unterschrieben
- Muss immer für alle transparent zugänglich sein
- Unsicherheiten als Chance → man kann die Projektcharta auch immer wieder mal anpassen, aber in Absprache mit allen und transparent.
- Die Projektcharta ist das **Output der Aufgabendefinition** und wird aber laufend verwendet während dem Plan

2 SITUATIONSBEWERTUNG

- Welche Ressourcen (Budget, Personal, Material, Software-Tools, Infrastruktur) stehen zur Verfügung?
- Wie viel Zeit hat man?
    - → ACHTUNG wenn man Agile arbeitet, sind Zeit und Ressourcen die erste Priorität und der Umfang des Produktes wird anhand von dem abgleitet
- Was sind die Einschränkungen und Rahmenbedingungen für das Projekt?
- Was sind die potenziellen Risiken?

3 PROJEKTZIELE UND ERFOLGSKRITERIEN

Anhand der Problemstellung und Situationsbewertung kann man nun Projektziele und Erfolgskriterien festlegen. Diese teilen sich auf in:

- **Zielwerte**
    - Ergeben sich aus **Projektzielen** (die qualitativ sind) und **Kenngrössen** (die quantitativ sind). Kombiniert erhält man Zielwerte. Das sind die Ziele, die mit dem Projekt am Schluss erfüllt werden müssen.
- Angestrebte Bereitstellungsform
- **Out of scope:** es ist wichtig zu festhalten, welche Sachen NICHT zum Projekt gehören, damit man effizient bleibt

Kenngrössen

- Das sind Zahlenwerte, die den Projektfortschritt abbilden.
- Müssen mit vertretbarem Aufwand messbar sein
- → durch Kenngrössen wird der Fortschritt vom Projekt messbar. Das schafft Transparenz und ist wichti als Grundlafe für Entscheidungen an den Meilensteinen. Das ermöglicht das iterative-inkrementelle Arbeiten
- → z.B. Güte eines Modells (MAE; RMSD, Accuracy, Precision, Recall, etc.)

Von Kenngrössen zu Zielwerten

- Zielwerte müssen einen Mehrwert für Nutzer*innen ergeben
- Zielwerte können zum Beispiel Verbesserungen sein. Also wenn das Projekt eine Erneuerung oder ein Update eines bestehenden Produkts ist, können die Zielwerte, welche man erreichen will, eine quantitativ festhaltbare Verbesserung in gewissen Bereichen für die Endnutzer*innen und Stakeholder sein.

- SMART Ziele: specific, measurable, achievable, relevant, time-bound

4 ZIELE DER DATENMODELLIERUNG

In diesem Schritt entscheidet man sich, was man mit den Daten machen will. Wie will man die Daten analysieren oder welche Modelle will man damit machen? Wie sieht die Modellierung aus? Solche **Datenanalyse und Modellierungsmethoden** sind:

- Regression
- Klassifikation
- Clusteranalyse
- Anomalie-Erkennung

Festlegung von Kenngrössen zur Bestimmung der Güte eines Modells

- Das heisst, wie gut fittet mein Modell meine Daten? Wie gut kann mein Modell erklären, wie die Daten verteilt sind?
- Für Regression: **MAE, RMSD**
- Für Klassifikation: **Accuracy, Precision, Recall**
- → Kenngrössen der Güte eignen sich gut als quantitative Kenngrösse für die Zielwerte!

5 PROJEKTPLAN

Zusammenarbeit im Team

- Rollen und Aufgaben klar verteilen
  - Sponsor*in → diese Person trifft am Schluss die Entscheidungen: gibt das Projekt frei, entscheidet bei den Meilensteinen, wie es weitergeht
  - Entwicklerteam
    - Teamsprecher*in → diese Person ist der „Chef“: achtet darauf, dass alle ihre Rollen, Aufgaben und Projektplan einhalten und kommuniziert gegen aussen für das Team
- **Kollaborationsmodus / Kommunikationsmodus**
  - Sitzungsryhtmus definieren → Form (Sitzungen digital oder vor Ort? Kurz oder lang?) und Frequenz (1x die Woche? Täglich?)
  - Wie oft geben die Teammitglieder einander Updates? → Austausch mit Product Owner
  - Wie oft präsentiert man seinen Fortschritt dem/ der Sponsor*In?
  - Wie und wann kommuniziert man mit Stakeholder?
  - Kontaktangabe: wie erreiche ich die Personen? Wann sind sie erreichbar?

Planung

- Zeitliche Grobplanung der Projektphase, z.B. mit Gantt Chart
- Meilensteine (Teilerfolge) festlegen → Meilensteine werden an Sponsor*in und/oder Stakeholders vorgestellt
- Benötigte Infrastruktur und Tools festlegen

Projektcharta

- So wie die ursprüngliche Abmachung. Wird vom Entwicklerteam geschrieben und von Sponsor*in unterschrieben
- Muss immer für alle transparent zugänglich sein

- Unsicherheiten als Chance → man kann die Projektcharta auch immer wieder mal anpassen, aber in Absprache mit allen und transparent.
- Die Projektcharta ist das **Output der Aufgabendefinition** und wird aber laufend verwendet während dem Plan

DATA ACQUISITION AND EXPLORATION

Data Acquisition, Lernziele: Sie

- kennen zentrale Begriffe in der Kategorisierung von Daten und Information
- können zwischen strukturierten und unstrukturierten Daten unterscheiden
- kennen typische Herausforderungen in der Datenbeschaffung

Data Exploration, Lernziele: Sie

- kennen statistische Kennzahlen und Graphiken einer explorativen Datenanalyse (EDA)
- können auf einem gegebenen Datensatz eine EDA durchführen
- können die Ergebnisse einer explorativen Datenanalyse interpretieren und Aussagen über die Datenqualität machen
- die Ergebnisse der EDA in einem Datenbericht zusammenfassen
- kennen Verfahren der Datenaufbereitung und können diese auf einen gegebenen Datensatz anwenden

DATEN UND INFORMATIONEN

Was sind Daten?

- Beobachtungen, Messungen, Zahlenwerte
- Auf Zahlenwerte beruhende Angaben oder formulierbare Befunde
- Zeichen, die eine Information darstellen,
- Alles, was sich für eine Datenverarbeitungsanlage (=Computer), auf erkennbare Art und Weise codieren, speichern, verarbeiten und transportieren lässt.
- → abstrahierte und computergerecht aufbereitet Informationen
- → es gibt analoge, digitale und digitalisierte Daten

Analoge Daten

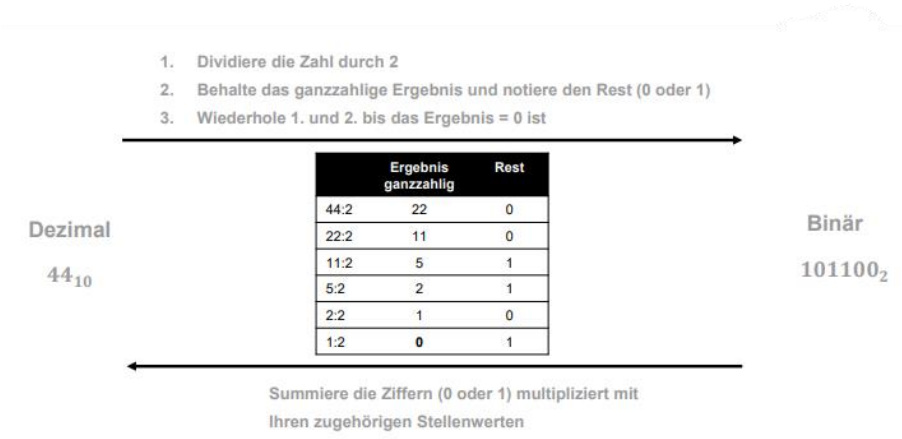
- Z.B. Zeigeruhren
- Z.B. Farbvariationen auf einer Leinwand, Prägungen auf einer Schallplatte
- Analoge Daten werden durch kontinuierliche Funktionen dargstellt, das heisst sie können jeden beliebigen Wert in der reellen Zahlenmenge einnehmen
- → ein Computer kann analoge Daten nicht lesen
- Reproduktionskosten hoch, Qualitätsverlust durch Reproduktion

Digitale Daten

- Digitale Daten sind **diskret**
- Mit digitalen Daten kann man Informationen auf Medien speichern
- Medien sind:
    - Nicht elektronisch: DNA, Rauchsignale, Morse Code, Braille Schrift, Verkehrszeichen
    - Elektronisch: Halbleitspeicher, Backup Tapes, etc.
- Elektronische Schaltkörper können nur zwei Zustände repräsentieren: 0, 1 → „bit“ (binary digit)

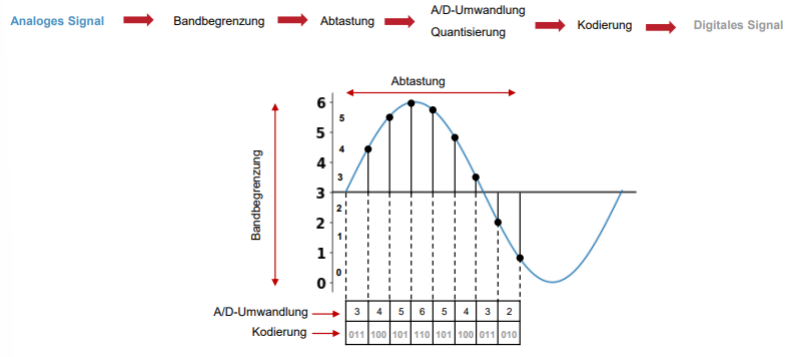
    - Darum muss man alle Zahlen mit 0 und 1 darstellen können → binäres Zahlensystem
    - 1 Byte = 8 bits
    - 1 Word = 4 Bytes = 32 bits
- **Herstellungskosten** für digitale Daten höher als für analoge Daten
- **Reproduktionskosten** gering, kein **Qualitätsverlust** durch Reproduktion
- **Nicht Rivalität:** das heisst, wenn man digitale Produkte zur Verfügung stellt, dann können alle darauf zugreifen, sofort und immer. Bei einem Buch: wenn ich es gerade am Lesen bin, kannst du es nicht gleichzeitig lesen. Bei pdf Slides auf Moodle: ich und du können es gleichzeitig lesen.
- **Granularität**: kleinste mögliche Auflösung kann zu Quantisierungsfehlern führen (?)
- **Kompressibilität**

Umwandlung Dezimalzahlen (normale Zahlen) → zu Binarzahl



Digitalisierte Daten

- Analoge Daten müssen digitalisiert werden, damit man sie speichern kann.
- Dabei werden kontinuierliche Daten in diskrete Daten umgewandelt
- Entscheidend für die Qualität, also den Informationsgehalt dieser Daten sind:
    - Bandbegrenzung
    - Abtastrate
    - Quantisierungsskala
- Digitalisierung von Bildern
    - Diskretisieren über Bildmatrix mit Pixeln, die Werte haben können abhängig von Farbraum
- Digitalisierung von Texten
    - Diskretisierung mit **ASCII Zahlensystem, Unicode** oder andere Codierungsschema
    - Scanning von Papier mit Text → digitales Bild → Texterkennungssoftware → Zeichensatz (menschenlesbar) zu binäre Respräsentation (computerlesbar)
    - Oder ein armer Heini tippt alles ab



Datei

- Besteht aus einer Reihe von bits in byte-Blöcken
- Dateiformat: „Wörterbuch“ / „Grammatikregeln“ wie man dieses bit-Reihen interpretieren / übersetzen muss. Je nach Inhalt braucht es ein anderes Dateiformat

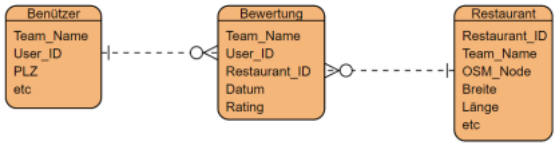
Beispiele und gängige Dateiformate

Inhalte	Formate
Texte	txt, odt, docx, html
Bilder	jpeg, png
Audio	wave, aiff, flac, mp3, aac
Video	mp4, mov, wmv

- Anwendungsprogramm: damit kann Dateiformat auf bit-Reihe angewendet und verarbeitet werden. Also zB man braucht Photoshop, um eine tiff-Datei zu öffnen.

DATENMODELLE

Entitäten



- Ein Datenmodell besteht aus Entitäten und die Beziehung der Entitäten untereinander
- **Entity Relationship Diagramme:** sind Standard für relationale Datenbanken. Das sind Diagramme, welche Entitäten und deren Beziehungen zueinander abbilden. Z.B.: Entität1 = User, Entität 2 = Restaurant, Beziehung zwischen beiden Entitäten = Bewertung des Restaurants durch User
- **Attribut**: Entität wird in einer relationalen Datenbank als Tabelle dargestellt, wo jede Spalte ein Attribut der Entität abbildet.
- **Foreign keys:** bildet die Beziehungen der Entitäten untereinander ab

Strukturierte Daten

- Besitzen ein Datenmodell (mit Entitäten)
- Verschiedene Datensätze folgen demselben Datenmodell
- Z.B.: Daten in Tabellenformat sind strukturiert. Jede Spalte ist ein Attribut
- Dateiformate : csv, ods, xlsx, HDF, Apache Parquet
    - Je nach Inhalt braucht es ein anderes Dateiformat

Unstrukturierte Daten

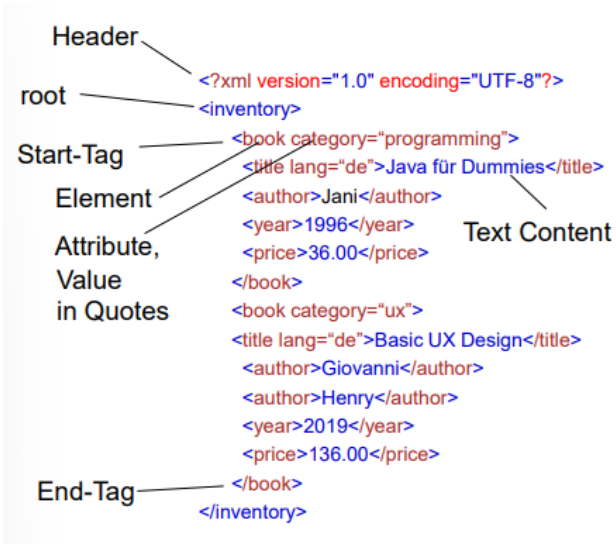
- Haben kein festes Datenmodell
- Haben keine explizite Struktur, aber vielleicht eine implizite Struktur, z.B. Grammatikregeln in Textdaten
- Bsp: Bilder, Videos, Sound Files, Textdaten, etc. → sind nicht in Tabellenformat wiedergegeben sondern vielleicht in einem Pixelraster, etc.

Semistrukturierte Daten

- Kein festgeschriebenes Datenmodell mit Entitäten, aber Strukturinformationen sind in Form von Tags und Markers enthalten
- Entitäten derselben Klasse können unterschiedliche Attribute haben und die Reihenfolge ist beliebig: zB User1 und User2 haben untershiedliche Attribute (Email anstatt Handynummer, oder einmal gibt es das Attribut Hobby, ein anderes mal nicht)
- Wichtige Dateienformate: XML, JSON
- Anwendungen:
    - Email
    - Objektbasierte Datenbanken (No SQL), wie MongoDB, Elasticsearch
- Vorteile:
    - Objekte können ohne festgelegtes Schema gespeichert werden
    - Mapping auf ein relationales Datenmodell unnötig
    - Einfache Abbildung verschachtelter Objekte oder hierarchischer Abhängigkeiten
- Nachteile
    - Keine performante Abfragensprache wie SQL
    - Fehleranfällig durch fehlendes Datenmodell

Extensible Markup Language (XML)

- Keine vordefinierten Tags wie HTML
- Extensible: die Applikation funktioniert auch dann, wenn Dokumenten Tags fehlen oder wenn es zusätzliche hat
- Enthält Daten aber ohne die Information, wie man diese Daten darstellen soll
- Hierarchische Struktur




```
{
  "inventory": [
    {
      "category": "programming",
      "title": {
        "lang": "de",
        "text": "Java für Dummies"
      },
      "author": "Jani",
      "year": 1996,
      "price": 36.00
    },
    {
      "category": "programming",
      "title": {
        "lang": "en",
        "text": "Python für Dummies""
      },
      "author": "John",
      "year": 2015,
      "price": 22.00
    }
  ]
}
```

Syntaxvalidierung: https://jsonlint.com/

JavaScript Object Notation (JSON)

- Ist

sprachenunabhängig, auch wenn es von Javascript abgeleitet ist

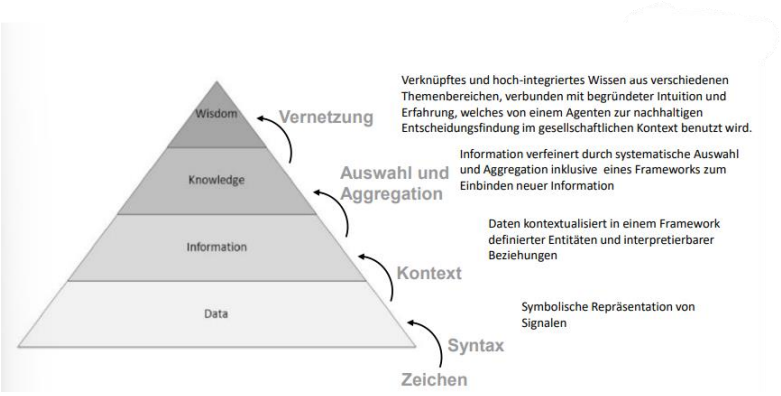
- Format ist reiner Text → Speicherung und Austausch ist einfacher
- Key/Value Paare: Daten werden so abgebildet
- Kommas als Trennung
- Keys sind immer strings
- Übersichtlicher und einfacher zu parsen als XML

Metadaten, Anwendungsdaten, Rohdaten

- Metadaten sind Daten über Daten (zB bei einer Bilddatei → wie gross ist die Bilddatei? Wann wurde sie aufgenommen? Etc.)
- Anwedungsdaten sind die Daten selbst
- Rohdaten sind Daten, die nicht weiter prozessiert wurden → das heisst sie haben noch keine Metadaten, sie sind noch unstrukturiert, etc.

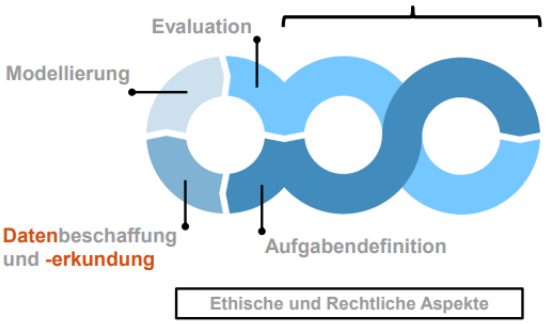
Von Daten zu Wissen

- Informationen, Signale, etc. werden durch Daten abgebildet
- Den Daten wird durch Strukturisierung, das heisst Datenmodelle mit Entitäten und Attributen Kontext gegeben
- Daten werden systemaitsch aggregiert und ausgewählt um mehr spezifisches Wissen zu schaffen
- Wissen aus verschiedenen Daten werden vernetzt und verknüpft um noch mehr Wissen zu schaffen



DATENBESCHAFFUNG

Throwback Data Product Entwicklung, die Datenbeschaffung kommt nach der Aufgabendefinition und vor der Datenexploration, Modellierung und Evaluation.



Datenbericht

- Im Datenbericht wird alles zur **Datenbeschaffung** und **Datenexploration** protokolliert
- Der Datenbericht gewährleistet
    - **Nachvollziehbarkeit** (Wieso haben wir das damals schon wieder so gemacht?) und
    - **Reproduzierbarkeit** der Ergebnisse (wie haben wir das schon wieder hingekriegt?)
- Datenbericht beinhält
    - Data Governance Regeln
    - Datenquellen
    - Datenablage

Datenquellen identifizieren

- Welche Daten brauche ich, um die Aufgabendefinition zu erfüllen?
- Wo finde ich diese Daten?
- Unterschiedliche Datenquellen:
    - (firmen)**intern** / (firmen)**extern**
    - **Öffentlich** verfügbar (also gratis) / **kommerziell** (kostet etwas)
    - Persönlich oder geschäftskritisch, vertraulich
- → Hier muss man viel über das Thema recherchieren und Domänen-Expert*innen herbeiziehen. Falls die Datenquellen, die man identifiziert hat, nicht ausreichen:
    - Eventuell mit weiteren Teams zusammenarbeiten
    - Bestehendes System allenfalls erweitern
    - Neue Experimente durchführen zur Datengewinnung

Beispiele für Datenquellen

- Datensätze aus Forschungsprojekten
- Datensätze von öffentlichen Ämtern (zB bfs Bundesamt für Statistik)
- Wikipedia
- Google Datasets
- Projectgutenberg.org
- Websiteninhalte → durch Scraping
- Kommerzielle Anbieter
- Firmeninterne Datenquellen:
    - Zentrales Data Warehouse
    - Applikationsspezifische Datenbanken
    - Sharepoint, shared drives
    - Internes Wiki
    - Email Server

Aspekte der Datenbeschaffung

- Erfordleriche Frequenz der Datenabfrage:

    - Daten werden immer wieder aktualisiert (zB jährliche Umfragen / Meswerte, etc.) → wie oft muss ich diese neuen Daten einbeziehen?
    - Ich habe Datensätze gefunden, aber sie passen nicht exakt auf meine Aufgabenstellung → öfters Datenabfrage machen
- Push vs Pull
- Batch-Download oder ad-hoc Zugang
- Lokal speichern oder nur streamen?
- Werden Daten persistiert? Das heisst erlaubt mir eine Website nur ein Mal die Daten runterzuladen?
- Anforderungen an die Infrastruktur der Datenquelle zur Bereitstellung der Daten
    - Verfügbarkeit
    - Throughput / Maximale Last der ANfrage
    - Latenz: Wartezeit für Daten
    - Stabilität des Schnittstellendesign (Wie oft muss ich meine Arbeitsweise, Datenbeschaffungspipeline, an Veränderungen der Schnittstelle anpassen?

Daten ablegen und organisieren

- Thema der Vorlesung Data Engineering
- Optimale Organisation ist wichtig für die späteren Phasen der Data Product Entwicklung
- Es gibt verschiedene Ablagesysteme. Welches Ablagesystem am besten für die Daten, mit denen man arbeitet, geeignet ist, hängt ab von folgenden Faktoren:
    - Datenmenge
    - Zugriffsfrequenz
    - Latenz
    - Verfügbarkeit
    - Datenmodell
    - Sicherheit und Governance
    - Kosten
- Beispiele für Ablagesysteme:
    - Datenbanken
        - Relational (Oracle, mysql, etc.)
        - Non-relational (NoSQL, MongoDB)
        - Graphen (zB GraphDB)
    - Event Streaming Platform
    - File Shares
    - Object Stores

Data Governance

- Daten sind wertvoll. Darum gibt es Regeln für den Umgang mit Daten, bezüglich
    - Schutz von Daten
    - Zugriff auf Daten
    - Verwendungszweck von Daten
- Die Strenge und Ausmass dieser Anforderungen hängen ab von:
    - Gesetzen (Datenschutzgesetze)
    - Erwartungen von Gesellschaft und Nutzer*innen
    - Industrie Standards
    - Organisationsinterne Vorschriften
- Governance Frameworks werden unterteilt in: (je weiter unten, desto strenger)
    - Öffentlich
    - Geschätsrelevant

- Geschäftskritisch
- Persönlich (PID: personally identifiable data)

DATA EXPLORATION

Lernziele: Sie

- kennen statistische Kennzahlen und Graphiken einer explorativen Datenanalyse (EDA)
- können auf einem gegebenen Datensatz eine EDA durchführen
- können die Ergebnisse einer explorativen Datenanalyse interpretieren und Aussagen über die Datenqualität machen
- die Ergebnisse der EDA in einem Datenbericht zusammenfassen
- kennen Verfahren der Datenaufbereitung und können diese auf einen gegebenen Datensatz anwenden

Beurteilung der Datenqualität

- Mittels deskriptiver Statistik kann man die Datenqualität beschreiben und daraus ableitend die notwendige Datenaufbereitung bestimmen:
- **Datenqualität**
  - **Unvollständigkeit:** Es fehlen Attribute oder einzelne Werte
  - **Rauschen und Ausreisser**: gibt es Werte, die random sind aufgrund von Fehlern bei der Aufnahme (Rauschen) oder Ausreisser?
  - **Inkonsistenz:** widersprüchliche Werte, Einträge oder unerklärliche Abweichungen
- **Datenaufbereitung:** was kann man machen, um die Datenqualität zu verbessern? Was für Modellierungen sind mit diesem Datenüberhaupt möglich?
  - Einschränkungen für Modellierung
  - Braucht es noch mehr Datensätze? → Qualität verbessern
- → Falls Datenqualität ungenügend, muss man im Extremfall das Projekt abbrechen

Deskriptive Statistik

- Kennzahlen, die uns sagen wie unsere Daten verteilt und verortet sind
- Masse der zentralen Tendenz
  - Mean, median, etc.
- Masse der Streuung
  - Verteilungen (z.B. Normalverteilung mit mü und sigma)
  - Standardabweichung

Variablentypen (Skalas) – "Elemente der explorativen Datenanalyse"

- Kategorisch
  - Nominal (zB Länder nach Kontinent kategorisieren)
  - Ordinal (zB ZGlücksempfinden nach hoch, mittel, tief (also Reihenfolge, order) kategorisieren)
- numerisch
  - disrekt (1, 2, 3, 4, 5)
  - stetig (1, 1.00001. 1.00002, 1.00003, …)

Datenmatrix

- Eigentlich die Form, in der man eine Entität ablegt, mit Spalten für Attribute. Das ist mathematische gesehen eine Matrix
- Spalte ID (für die Nummerierung der Beobachtungen), dann weitere Spalten für verschiedene Attribute (Variablen)
- Anzahl Zeilen = Anzahl Beobachtungen
- Anzahl Spalten = Anzahl Variablen

$$X_{N\times D}: \begin{pmatrix} \vec{x}^{(1)^T} \\ \vec{x}^{(2)^T} \\ \vdots \\ \vec{x}^{(10)^T} \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} & x_5^{(1)} & x_6^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} & x_5^{(2)} & x_6^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^{(10)} & x_2^{(10)} & x_3^{(10)} & x_4^{(10)} & x_5^{(10)} & x_6^{(10)} \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 25443 \\ 0 & 1 & \cdots & 22367 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 26444 \end{pmatrix}$$

$$\vec{x}^{(1)} = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_6^{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 25443 \end{pmatrix}, \vec{x}^{(10)} = \begin{pmatrix} x_1^{(10)} \\ x_2^{(10)} \\ \vdots \\ x_6^{(10)} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 26444 \end{pmatrix}$$

EXPLORATIVE DATENANALYSE ANWENDEN

Siehe Excel!!

Datensatz Übersicht		
Anzahl Merkmale	10 (ohne Index)	
Anzahl Beobachtungen	1000	
Leere Zellen	576	9.6%
Duplizierte Zeilen	3	0.3%
Kategorische Merkmale	7	
Numerische Merkmale	3	

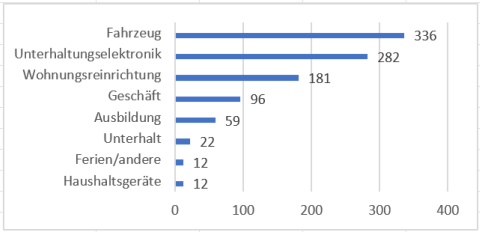
Univariate Deskription

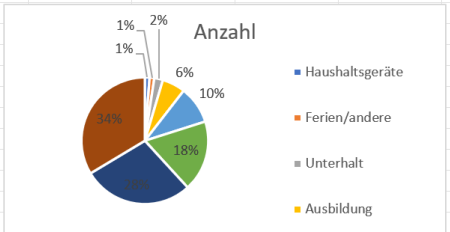
- Für jede einzelne Variable (Attribut) eine Übersicht erstellen
- Konsistenz heisst hier, dass Anzahl Zeilen = Anzahl leere Zellen + Anzahl Zahlenwerte + Anzahl Textwerte
- Konsistenz ist ein Kriterium der Datenqualität
- Häufigkeit berechnen für die einzelnen Beobachtungen pro Variable

Merkmal:	Alter		
Variablentyp	Numerisch		
Anzahl Zeilen	1000		
Eindeutige Werte	56	5.6%	CHECK: ok
Leere Zellen	0	0.0%	
Text-Werte	0	0.0%	Check, falls die Spa
Zahlenwerte	1000	100.0%	
Nullwerte	0	0.0%	
Negative Werte	0	0.0%	

Merkmal:	Verwendungszweck		
Variablentyp	Kategorisch		
Anzahl Zeilen	1000		Konsistenz:
Eindeutige Werte	8	0.8%	CHECK: ok
Leere Zellen	0	0.0%	
Text-Werte	1000	100.0%	
Zahlenwerte	0	0.0%	

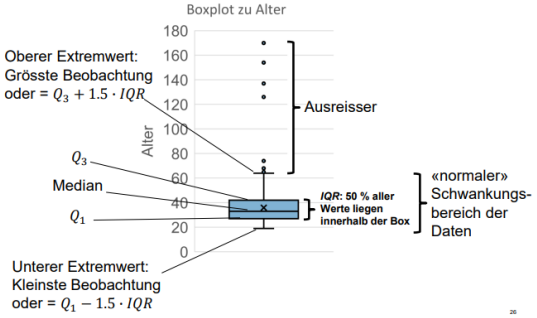
Ausprägungen	**Anzahl**	**Anteil [%]**
Haushaltsgeräte	12	1.20%
Ferien/andere	12	1.20%
Unterhalt	22	2.20%
Ausbildung	59	5.90%
Geschäft	96	9.60%
Wohnungsreinrichtung	181	18.10%
Unterhaltungselektronik	282	28.20%
Fahrzeug	336	33.60%
Gesamtergebnis	**1000**	**100.00%**





Deskriptive Statistik

- **Lagemasse**
  - Arithmetisches Mittel
  - Modus
  - Quantilstatistik
    - Minimum
    - Q1
    - Median

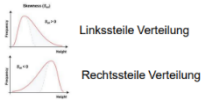


- Q3
- Maximum
- → Ausreisser anhand der Quantilstatistik kann man mit Boxplot darstellen

- **Streumasse**
  - Range (x) = max(x) – min(x)
  - IQR = Q3-Q1
  - Varianz
  - Standardabweichung = Wurzel (Varianz)
  - Histogramm
  - Kummulative Häufigkeit
  - Schiefe: rechtsschief / linksschief
  - Wölbung: wie gross / klein ist die Standardabweichung

Momentkoeffizient der Schiefe:

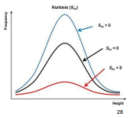
$$g_m = \frac{m_3}{s_x^3} \text{ mit } m_3 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^3$$



$g_m = 0$ für symmetrische Verteilungen
$g_m > 0$ für linkssteile Verteilungen
$g_m < 0$ für rechtssteile Verteilungen

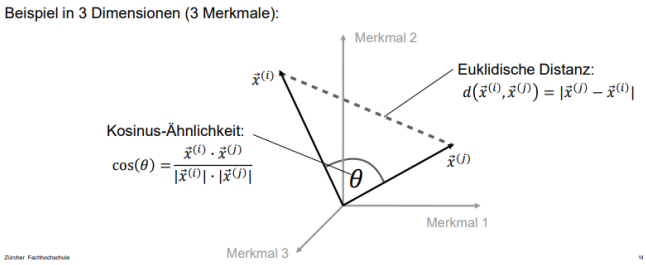
$$\gamma = \frac{m_4}{s_x^4} - 3 \quad \text{mit} \quad m_4 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \bar{x})^4$$

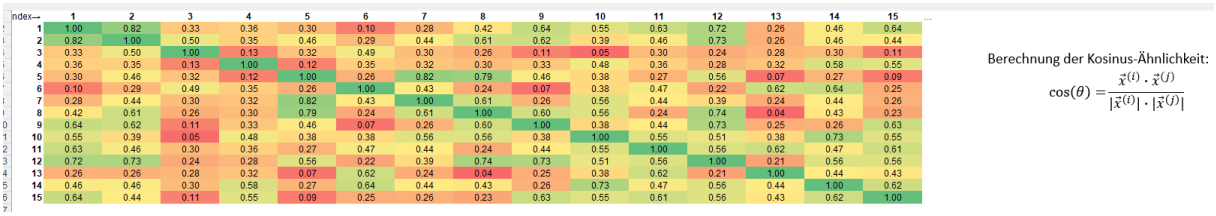
$\gamma = 0$ bei Normalverteilung
$\gamma > 0$ bei spitzeren Verteilungen. D.h. die Verteilung weist weniger (extreme) Outlier auf, als die Normalverteilung.
$\gamma < 0$ bei flacheren Verteilungen. D.h. die Ränder gehen langsamer gegen Null als die Normalverteilung



Duplikate erkennen

- Auf Excel, mit Verketten Funktion: So erkennt man Zeilen, die potenziell copy paste sind, wenn zwei Zeilen wirklich komplett identisch sind.
- Ähnlichkeit von Vektorzeilen generell:
  - Jede Beobachtung ist eine Zeile mit D Werten pro Spalte (Attribut / Varaible). Das heisst jede Beobachtung kann mit einem Vektor abgebildet werden → Zeilenvektor
  - Wie ähnlich sind zwei Beobachtungen? Man misst das mit
    - der euklidischen Distanz zwischen zwei Zeilenvektoren
    - Kosinus Ähnlichkeit → 1: ähnlich; 0: senkrecht abhängig; -1: unähnlich





Index→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.00	0.82	0.33	0.36	0.30	0.10	0.28	0.42	0.64	0.55	0.63	0.72	0.26	0.46	0.64
2	0.82	1.00	0.50	0.35	0.46	0.29	0.44	0.61	0.62	0.39	0.46	0.73	0.26	0.46	0.44
3	0.33	0.50	1.00	0.13	0.32	0.49	0.30	0.26	0.11	0.05	0.30	0.24	0.28	0.30	0.11
4	0.36	0.35	0.13	1.00	0.12	0.35	0.32	0.30	0.33	0.48	0.36	0.28	0.32	0.58	0.55
5	0.30	0.46	0.32	0.12	1.00	0.26	0.82	0.79	0.46	0.38	0.27	0.56	0.07	0.27	0.09
6	0.10	0.29	0.49	0.35	0.26	1.00	0.43	0.24	0.07	0.38	0.47	0.22	0.62	0.64	0.25
7	0.28	0.44	0.30	0.32	0.82	0.43	1.00	0.61	0.26	0.56	0.44	0.39	0.24	0.44	0.26
8	0.42	0.61	0.26	0.30	0.79	0.24	0.61	1.00	0.60	0.56	0.24	0.74	0.04	0.43	0.23
9	0.64	0.62	0.11	0.33	0.46	0.07	0.26	0.60	1.00	0.38	0.44	0.73	0.25	0.26	0.63
10	0.55	0.39	0.05	0.48	0.38	0.38	0.56	0.56	0.38	1.00	0.55	0.51	0.38	0.73	0.55
11	0.63	0.46	0.30	0.36	0.27	0.47	0.44	0.24	0.44	0.55	1.00	0.56	0.62	0.47	0.61
12	0.72	0.73	0.24	0.28	0.56	0.22	0.39	0.74	0.73	0.51	0.56	1.00	0.21	0.56	0.56
13	0.26	0.26	0.28	0.32	0.07	0.62	0.24	0.04	0.25	0.38	0.62	0.21	1.00	0.44	0.43
14	0.46	0.46	0.30	0.58	0.27	0.64	0.44	0.43	0.26	0.73	0.47	0.56	0.44	1.00	0.62
15	0.64	0.44	0.11	0.55	0.09	0.25	0.26	0.23	0.63	0.55	0.61	0.56	0.43	0.62	1.00

Berechnung der Kosinus-Ähnlichkeit:

$$\cos(\theta) = \frac{\vec{x}^{(i)} \cdot \vec{x}^{(j)}}{|\vec{x}^{(i)}| \cdot |\vec{x}^{(j)}|}$$

Streumasse und Korrelationskoeffizient

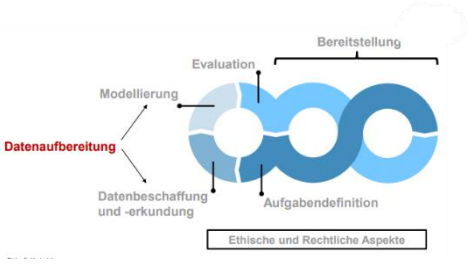
- Ein Streudiagramm gibt einen ersten Hinweis auf Beziehungen von zwei Variablen (wie beeinflusst Variable a die Variable b?)
- Wertebereich: -1 <= r => 1
  - 0: random, 1: perfekt korreliert, -1: perfekt negativ korreliert
- Berechnung von Korrelationskoeffizient
  - Nach Bravais-Pearson
  - Nach Spearman

Bravais-Pearson: $r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$

Spearman: $r_{SP}(x,y) = \frac{\sum_{i=1}^{N}(rg(x_i) - \overline{rg}_x)(rg(y_i) - \overline{rg}_y)}{\sqrt{\sum_{i=1}^{N}(rg(x_i) - \overline{rg}_x)^2}\sqrt{\sum_{i=1}^{N}(rg(y_i) - \overline{rg}_y)^2}}$

DATENAUFBEREITUNG

Die Datenaufbereitung heisst das Anpassen / Standardisieren von Daten, oder umwandeln, in Entitäten, etc. Die Daten so aufbereiten, dass man damit Modellieren kann. Die Erkenntnisse



aus der Datenexploration kann man verwenden, um die Datenaufbereitung zu machen.

Fehlende / Falsche Werte

- Entweder ersetzen durch eine Konstante, z.B. „NA“: not available
    - Vorteil: es geht einfach
    - Vorteil: ersetzte Werte bleiben als markiert als ursprünglich mal leer
    - Nachteil: verändert / verzerrt die Werteverteilung
    - Nachteil:ignoriert Korrelation zwischen Attributen
- Oder ganze Zeile (ganze Beobachtung) aus dem Datensatz löschen
    - Vorteil: es geht einfach
    - Nachreil: Informationsverlust
- Oder den Wert manuell bestimmen (also leere Zeile nachträglich füllen)
    - Vorteil: maximal mögliche Genauigkeit
    - Vorteil: effektiv für kleine Datensätze
    - Nachteil: ineffizient (weil mega aufwendig) für grosse Datensätze
- Oder ersetzen durch ein Lagemass, zB den Durchschnitt
    - Vorteil: es geht einfach
    - Nachteil: veränderte / verzerrte Werteverteilung
    - Nachteil: igmoriert Korrelation zischen Attributen
    - Nachteil: systematisch unterschätzte Standardabweichung (weil man die Verteilung so manipuliert, dass sie schmaler wird → mehr Daten rund um das mü)
- Oder den Wert mithilfe eines Modells bestimmen (z.B. nächste Nahcbar Heuristik und Ähnlichkeitsmass für ein Modell verwenden)
    - Vorteil: bestmögliche Näherung des Wertes
    - Nachteil: Genauigkeit abhängig vom eingesetzten Modell
    - Nachteil: Rechenaufwand

Kodierung kategorischer Merkmale

- Wenn meine Werte für eine Variable (ein Attribut / eine Spalte) ordinal- / nominalskaliert sind, muss ich trotzdem eine Zahl einsetzen können, damit ich damit rechnen kann. Hierfür gibt es zwei Varianten
- Ordinalkodierung
    - Jeder Wert bekommt eine Zahl zugewiesen zwischen [0, (Anzahl Werte – 1)]
    - → eignet sich für Ordinalskalierte Daten
- One-Hot Kodierung
    - Jeder Wert wird binär dargestellt. Das heisst, wenn man z.B. nur eine Spalte mit „Automarke“ hatte und darin vier verschiedene Werte, also vier verschiedene Marken, vorkamen, muss man jetzt für jeden Wert eine separate Spalte machen mit 0=nein und 1=nein
    - → eignet sich für nominalskalierte Daten

Skalierung numerischer Merkmale

- Skalierung dient der Transformation aller Merkmale auf einen gemeinsamen Wertebereich
- Das heisst, alle Zahlen werden als Prozent des maximalen Wertes angegeben → Wertebereich für alle Zahlen von allen Spalten ist [0, 1]
- Skalierungsverfahren:
    - Linerare Nominalisierung
    - Standardisierung

$$x_{\text{normalisiert}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$x_{\text{standardisiert}} = \frac{x - \bar{x}}{\sigma_x}$$

DATEN MODELLIEREN

Lernziele: Sie

- kennen die Elemente einer Machine Learning-Pipeline
- können zwischen überwachtem und unüberwachtem Lernen unterscheiden
- kennen erste Methoden der Regression und Clusteranalyse
- können eine lineare Regression in Excel auf einen gegebenen Datensatz ausführen
- kennen Evaluationsmetriken für Regressionsprobleme und können diese in Excel berechnen

REGRESSION

MACHINE LEARNING

KLASSIFIKATION

NACHBARSCHAFTSHEURISTIK

RECHTLICHE ASPEKTE

Lernziele sind: –

- Kurz Überblick Rechtssystem /Rechtsordnung Schweiz –
- Überblick Schweizerisches Strafrecht: Computerdelikte –
- Einführung Datenschutz und Terminologie
- Grundkenntnisse des schweizerischen Datenschutzrechts
- Anwendung

Gesetz/ Verordnung	Gilt für:	Verfahren	Datenschutzbeauftragte/-r
Bundesgesetz über den Datenschutz (DSG)	Bundesbehörden/ Private	Verwaltungs-verfahrensgesetz/ Zivilgesetzbuch z.T. im DSG	Eidg. Datenschutz- und Öffentlichkeits-beauftragter (EDÖB)
Gesetz (des Kt. ZH) über die Information und den Datenschutz (IDG)	Behörden im Kt. ZH/ Private, soweit sie öffentliche Aufgaben für den Kt. ZH erfüllen	Im IDG geregelt	Datenschutzbeauf-tragte/-r des Kt. ZH
Datenschutzverordnung (DSV) der Stadt Zürich	Behörden der Stadt Zürich	IDG und DSV	Datenschutzbeauf-tragte/-r der Stadt ZH

Überblick Rechtssystem Schweiz

- Es gibt: Privatrecht, Öffentliches Recht und als Teil davon Strafrecht
- Gesetze werden auf Bundes-, Kantons-, und Kommunalebene gemacht
    - Europäische Ebene: DSGVO. Weitere
    - Bundesebene: DSG, VDSG
        - Geltungsbereich: Private Personen, Bundesorgane
    - Kanton Zürich: IDG, ZH; IDV, ZH
        - Geltungsbereich: öffentliche Organe auf kantonaler und kommunaler Ebene

Art 143 StGB (Datendiebstahl)

- Wer in der Absicht, sich oder einen andern unrechtmässig zu bereichern, sich oder einem andern elektronische oder in vergleichbarer Weise gespeicherte oder übermittelte Daten beschafft, die nicht für ihn bestimmt und gegen seinen unbefugten Zugriff besonders gesichert sind, wird mit Freiheitsstrafe bis zu fünf Jahren oder Geldstrafe bestraft
- Die unbefugte Datenbeschaffung zum Nachteil eines Angehörigen oder Familiengenossen wird nur auf Antrag verfolgt.
- → es geht, um Datendiebstahl mit der Absicht sich oder einen anderen zu bereichern. Damit es Datendiebstahl ist, muss klar sein, dass diese Daten nicht für diese Person bestimmt sind und vor Zugriff dieser Person extra geschützt wurden.

Art 143 bis StGB (Hacking)

- Wer auf dem Wege von Datenübertragungseinrichtungen unbefugterweise in ein fremdes, gegen seinen Zugriff besonders gesichertes Datenverarbeitungssystem eindringt, wird auf Antrag, mit Freiheitsstrafe bis zu drei Jahren oder Geldstrafe bestraft
- Wer Passwörter, Programme oder andere Daten, von denen er weiss oder annehmen muss, dass sie zur Begehung einer strafbaren Handlung gemäss Absatz 1 verwendet werden sollen, in Verkehr bringt oder zugängliche macht, wird mit Freiheitsstrafe bis zu drei Jahren oder Geldstrafe bestraft

Art 144 bis Ziff. 1 StGB (Datenbeschädigung)

Art. 147 StGB (Computerbetrug)

Prüfschema

- Objektiver Tatbestand → Tatobjekt, Tathandlung
  - Angriffsobjekt:
    - Passwörter, Programme oder andere Daten
    - die gespeicherten oder übermittelten Daten selbst
    - die Berechtigung an die Daten (es ist nur ein Tatbestand, wenn die Täter*innen keine Berechtigung an die Daten haben)
    - Besondere Sicherung: Daten müssen gesichert worden sein vor unbefugten Zugriff
  - Tathandlung
    - Beschaffen der Daten
    - In Verkehr bringen der Daten, zugänglich machen der Daten
- Subjektiver Tatsbestand
  - Man muss eine Bereicherungsabsicht haben (StGB 144) oder man muss keine besondere Absicht haben (StGB 143 bis)
  - Man muss Bescheid wissen, dass man unbefugterweise auf die Daten zugreift („… von denen er annehmen muss…“)
- Rechtswidrigkeit
- Schuld
- Ergebnis

Strafantrag

- Einen Antrag stellen können nur diese Personen, die Berechtigt sind an die Daten, die gestohlen, vermittelt, etc. worden sind

Gesetz/ Verordnung	Gilt für:	Verfahren	Datenschutzbeauftragte/-r
Bundesgesetz über den Datenschutz (DSG)	Bundesbehörden/ Private	Verwaltungs-verfahrensgesetz/ Zivilgesetzbuch z.T. im DSG	Eidg. Datenschutz- und Öffentlichkeits-beauftragter (EDÖB)
Gesetz (des Kt. ZH) über die Information und den Datenschutz (IDG)	Behörden im Kt. ZH/ Private, soweit sie öffentliche Aufgaben für den Kt. ZH erfüllen	Im IDG geregelt	Datenschutzbeauf-tragte/-r des Kt. ZH
Datenschutzverordnung (DSV) der Stadt Zürich	Behörden der Stadt Zürich	IDG und DSV	Datenschutzbeauf-tragte/-r der Stadt ZH

Datenschutz (DSG)

- Gesetz für Bearbeiten von Daten natürlicher und juristischer Personen durch private Personen und Bundesorgane
- **Personendaten**: alle Angaben, die sich auf eine bestimmte oder bestimmbare natürliche Person beziehen
    - Besonders schützenswerte Personendaten → Daten über religiöse, weltanschauliche, politische oder gewerkschaftliche Ansichten oder Tätigkeiten; die Gesundheit, die Intimsphäre oder die Rassenzugehörigkeit; Massnahmen der sozialen Hilfe; administrative oder strafrechtliche Verfolgungen und Sanktionen; genetische Daten; biometrische Daten
- Es bezweckt: das Handeln der öffentlichen Organe transparent zu gestalten und damit die freie Meinungsbildung und die Wahrnehmung der demokratischen Rechte zu fördern, sowie die Kontrolle des staatlichen Handelns zu erleichtern; die Grundrechte von Personen zu schützen, über welche die öffentliche Organe Daten bearbeiten.
- **Persönlichkeitsprofil**: eine Zusammenstellung von Daten, die eine Beurteilung wesentlicher Aspekte der Persönlichkeit einer natürlichen Person erlaubt.
- **Profiling und Profiling mit hohem Risiko**: jede Art der automatisierten Bearbeitung von Personendaten, die darin besteht, dass diese Daten verwendet werden, um bestimmte persönliche Aspekte, die sich auf eine natürliche Person beziehen, zu bewerten, insbesondere um Aspekte bezüglich Arbeitsleistung, wirtschaftlicher Lage, Gesundheut, persönlicher Vorlieben, Interessen, Zuverlässigkeit, Verhalten, Aufenthaltsort oder Ortswechsel dieser natürlichen Person zu analysieren oder vorherzusagen. → also basically das, was Data Scientists so machen halt
- **Anonymisierung**: Personenbezug wird irreversibel entfernt
- **Pseudonymisierung**: Personenbezug wird reversibel entfernt (das heisst man kann immer noch nachschauen, wer dahintersteckt)

Grundsätze des Datenschutzes:

- Rechtsmässigkeit
- **Treu und Glauben**
    - Ein loyales und vertrauenswürdiges Verhalten im Rechtsverkehr ist grundlegend, ein widersprüchliches Verhalten läuft diesem zuwider

    - Der Grundsatz von Treu und Glauben stellt eine Generalklausel dar und kann in all denjenigen Konstellationen zum Zuge kommen, in denen die anderen Bearbeitungsgrundsätze nicht greifen
- **Verhältnismässigkeit**
    - Die Bearbeitung ist geeignet, um den Zweck zu erreichen
    - Die Daten werden nur bearbeitet, soweit es erforderlich ist um den Zweck zu erreichen
    - Der Zweck steht im vernünftigen Verhältnis zum Eingriff
    - → Datenvermeidung, Aufbewahrungsdauer minimieren, Datensparksamkeit
- **Zweckbindung**
    - Keine Datenbeschaffung auf Vorrat
    - Zweck ergibt sich aus den Umständen der Datenbeschaffung und dem Gesetz
- **Erkennbarkeit**
    - Eckpfeiler des ganzen Datenschutzsystems: betroffene Person soll entscheiden können, ob sie sich widersetzen/wehren will
    - Transparenz der Datenbeschaffung → wie im Datenbericht
    - Personendaten dürfen nicht grundlos abgefragt werden. Wenn jemand dich um Personendaten bittet, muss transparent gesagt werden, was der Zweck für die Beschaffung ist. Die Daten dürfen dann auch nur gemäss diesem Zweck bearbeitet werden.
- **Datenrichtigkeit**
    - Daten müssen richtig sein, unrichtige Daten müssen gelöscht werden
    - Je nach Thema ist Richtigkeit extrem wichtig, z.B. in der Medizin
    - Vergewisserungspflicht: diejenigen die die Daten beschaffen und bearbeiten müssen sich vergewissern, dass sie richtig sind
    - Anspruch auf Löschung
- **Datensicherheit**
    - Schutz gegen unbefugtes Bearbeiten von Personendaten durch angemessene technische oder organisatorische Massnahmen
    - Besonders schützenswerte Personendaten oder Persönlichkeitsprofile müssen mehr geschützt werden
    - Vertraulichkeit: nur die, die berechtigt sind, dürfen Zugriff auf Daten haben
    - Verfügbarkeit: Informationen müssen verfügbar sein, wenn gewünscht
    - Datenintegrität: Durch Bearbeitung der Daten dürfen Daten nicht verändert werden.

Zu prüfen ist eine Strafbarkeit nach Art. 143 bis StGB Sachverhalt: Hans Muster fragt zum Zweck eines Lebensmitteleinkaufes für die Geldbörse seines Onkels. Der Onkel erteilte ihm Absage mit der Begründung, dass sich nebst Bargeld und Kreditkarten all seine Passwörter und diverse Logindaten ebenfalls in der Geldbörse befinden und diese seien geheim. Sehr interessant, dachte sich Hans. Er wartete einen günstigen Moment ab und entwendete heimlich die Geldbörse des Onkels, fotografierte alle Passwörter und Login Daten und ging wieder nach Hause. Hans weiss, dass sein Onkel Kunde beim Finanzinstitut B ist. Er loggt sich mit Hilfe der fotografierten Zugangsdaten erfolgreich ein.	Hans könnte sich des Tatbestands des unbefugten Eindringens in ein Datenverarbeitungssystem strafbar gemacht haben, indem er sich mit den erhaltenen Zugangsdaten seines Opfers Zugang zum Konto beim Finanzinstitut B verschafft hat. Beim E-Banking-System handelt es sich um ein für Hans fremdes und gegen seinen Zugriff besonders gesichertes Datenverarbeitungssystem. Hans ist in dieses Datenverarbeitungssystem auf dem Wege von Datenübertragungseinrichtungen eingedrungen. Fraglich ist, ob dieses Eindringen unbefugt erfolgt ist. H überwindet die Zugangsschranken des Online-Portals des Finanzinstituts B, indem er die durch List erschlichenen Zugangsdaten verwendet. Hans wusste darum, dass sein Onkel nicht wollte, dass er Zugriff auf die Geldbörse hat. Hans wusste, dass er sich ohne Erlaubnis des Onkels

•	die Zugangsdaten durch eine List beschafft hat. Damit fehlt es bei Hans an einer Zugangsberechtigung, womit sein Eindringen als unbefugt einzustufen ist. A handelt mit Wissen und Willen und damit vorsätzlich. Rechtfertigungs- oder Schuldausschlussgründe sind keine ersichtlich. Im Ergebnis hat sich Hans Muster sich nach Art. 143 bis Abs. 1 StGB schuldig gemacht.

ETHISCHE ASPEKTE

Lernziele: Sie

- haben ein grundlegendes Verständnis für die ethischen Herausforderungen bei der Entwicklung datenbasierter Produkte
- kennen die Grundstruktur des Ethik-Kodex der Swiss data innovation alliance
- können ethische Fragestellungen in der Entwicklung datenbasierten Produkten anhand des Ethik-Kodex diskutieren

ETHISCHE HERAUSFORDERUNGEN DATA PRODUCT ENTWICKLUNG

Ethik: Wissenschaft der Moral

Skalierungspotenzial durch Digitalisierung

- unmoralisches Verhalten betrifft schnell extrem viele Menschen, wenn es online geschieht
- digitale Plattformen sind global → Konsequenzen von unmoralischem Verhalten gehen über Grenzen hinaus

Leitgedanken für Datenethik

- Digitalisierung soll dem **Wohl der gesamten Gesellschaft** dienen und ihren Zusammenhalt stärken
- **Menschenzentrierte** und **werteorientierte** Gestaltung der Technologie
- Förderung digitaler Kompetenzen und kritische Reflexion in der digitalen Welt
- Stärkung des Schutzes von persönlicher Freiheit, Selbstbestimmung und Integrität
- Risikoadaptierte Regulierung und wirksame Kontrolle algorithmischer Systeme
- Wahrung und Förderung von Demokratie
- Ausrichtung digitaler Strategien an Zielen der Nachhaltigkeit

ETHIK KODEX SWISS DATA INNOVATION ALLIANCE

Code findet man vollständig hier: https://data-innovation.org/data-ethics/

Grundorientierungen

- Schadenvermeidung
  - Schutz vor Datenverlust
  - Sicherheit von Daten vor Hacker

    - Nachhaltigkeit (im Bezug auf Umwelt)
- Gerechtigkeit
    - Gleichheit → Schutz vor Diskriminierung
    - Fariness → Gegenleistung für das Sammeln von Kundendaten
    - Solidarität → Daten der Öffentlichkeit zur Verfügung stellen für gemeinschaftlichen Nutzen
- Autonomie
    - Freiheit → Wahlfreiheit bei der Konfiguration von digitalen Dienstleistungen
    - Privatssphäre → nur Daten sammeln, die notwendig sind,
    - Würde → Kunden ernst nehmen

Prozedurale Werte

- Kontrolle
    - Interne Prozesse mit Daten müssen wohldefiniert und steuerbar sein
- Transparenz
    - Dokumentieren und kommunizieren, was mit den Daten gemacht wird
    - Transparenz für Kund*innen und Auditor
- Rechenschaft
    - Klare Zuständigkeiten definieren
    - Verantwortung übernehmen bei Regelverletzungen

Struktur Ethik Kodex

- Vier Phasen
    - Datenerzeugung und Akquirierung
        - Ergebnisse: digitale und elektronische Daten
        - → Leute die man hierfür anstellt auch richtig bezahlen
        - → um Erlaubnis bitten, nicht Menschen ohne ihr Wissen im privaten Raum durch Internet of Things überwachen
        - → um Zustimmung für Cookies bitten
        - → Formulare für Datenerfassung machen → Transparenz → Zweck muss klar sein etc.
    - Datenspeicherung und Management
        - Ergebnisse: Datenbank inklusive Zugriffsregeln und Sicherheitsmechanismen
        - → Cybersicherheit
        - Bereinigen von Daten (Richtigkeit gewährleisten)
    - Datenanalyse und Wissensgenerierung
        - Ergebnisse: gewonnene Erkenntnisse aus den Daten (die eine Wertschöpfung zum Ziel haben) und die zu diesem Zweck angewandte Mechanismen
        -
    - Produkte und Dienstleistungen
        - Ergebnisse: die auf Mechanismen und Erkenntnissen beruhende Dienstleistungen und Produkte, welche auf die reale Welt wirken

Bias

- Ein Algrorithmus oder Klassifizierungssystem kann einen Bias haben gegenüber einigen Bevölkerungsgruppen
- Ein Bias kann entstehen, wenn die Daten, auf welche der Algorithmus basiert, nicht die Grundgesamtheit repräsentativ abdeckt und somit systematisch gegen eine Bevölkerungsgruppe diskriminiert
- Ein Bias kann entstehen, weil ein Modell Korrelationen und Kausalitäten nicht auseinanderhalten kann
- Ein Bias kann in jedem Schritt der AI / ML Pipeline entstehen:
    - **Data Creation → Sampling Bias**: Daten sind nicht vielfältig genug (Bias bei denen, die Daten sammeln)
    - **Problem formulation → Framing Effect Bias:** Problem wird falsch verstanden
    - **Data Analysis → Sample Selection Bias**: wegen Daten, die nicht vielfältig genug sind, ist die Analyse auch biased.
        - Proxy Variable bias, ommitted variable bias → eine andere Variable erklärt besser – Korrelation und Kausalität verwechseln
    - **Validation and Testing → sample treatment bias**: menschliche Bias bei der Evaluation wegen Psychologie und kultureller Hintergrund

Beispiel Prüfungsaufgabe:

Ausgangslage und Problemstellung:	**Lösungsvorschlag:**
Ein Kreditanbieter nutzt ein maschinelles Lernmodell mit Informationen über Antragsteller von Konsumkrediten, um vorherzusagen, ob sie über einen Zeitraum von zwei Jahren pünktliche Zahlungen leisten werden. Die Vorhersage des Modells wird genutzt um zu entscheiden, ob eine antragstellende Person für einen Kredit in Frage kommt oder nicht. Zu den erklärenden Variablen gehören Alter, Einkommen, Vermögen etc. Es stellt sich heraus, dass ein grösserer Anteil von Anträgen von Personen mit Migrationshintergrund aufgrund dieser Merkmale abgelehnt werden. **Aufgabe:** Welche ethischen Grundorientierungen aus dem Ethik	Besonders die Grundorientierung der **Gerechtigkeit** wird in diesem Fall tangiert: Es besteht das Problem von indirekter Diskriminierung, in diesem Fall von Personen mit Migrationshintergrund. Dies ist in diesem Fall besonders relevant, da das erstellte Modell die Entscheidungsfindung beim Kreditanbieter unterstützt. Empfehlungen • **zur Gerechtigkeit**: ○ Es wird sichergestellt, dass der **Trainingsdatensatz vielfältig genug** ist, sodass die indirekte Diskriminierung nicht aufgrund schlechter Trainingsdaten herrührt. ○ Das Modell wird auf indirekte **Diskriminierung** bezüglich sensibler

Kodex sehen Sie in diesem Fall betroffen, bzw. verletzt, und warum? Machen Sie zwei konkrete Vorschläge für die Stärkung der ethischen Grundorientierungen und Umsetzung der prozeduralen Werte in diesem Fall. Dabei muss nicht zwischen den vier Phasen des Daten-Lebenszyklus' unterschieden werden.	Merkmale der antragstellenden Personen **untersucht** und die Ergebnisse der Audit-Stelle zur Verfügung gestellt. ○ Es wird erklärt, welche Formen indirekter **Diskriminierung** aufgrund der Minimierung ökonomischer Risiken **in Kauf genommen werden**. • **zur Transparenz:** Den antragstellenden Personen ○ wird zu Beginn des Antragsprozesses der **Einsatz des statistischen Modells erklärt.** ○ werden die **Einflussfaktoren** auf die Entscheidung des Modells **offengelegt** ○ wird mit dem **Entscheid automatisch** von einer Software berechnete, individuellen Änderungen an ihrem Profil aufgezeigt, die die Entscheidung des KI-Modells verändert hätten.