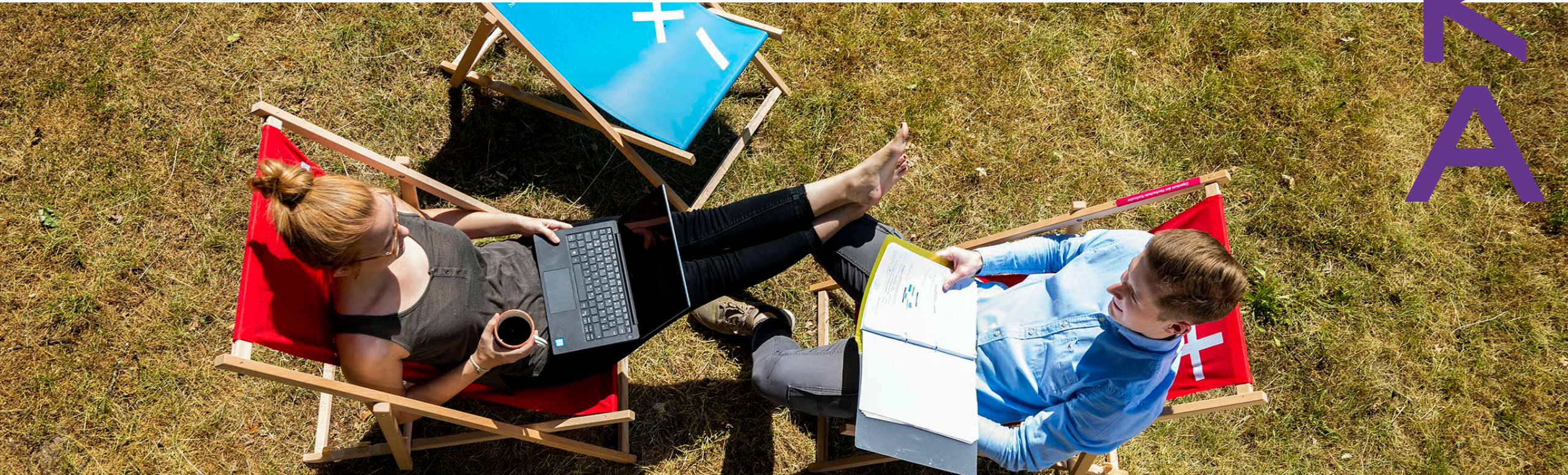


DSCB130 | 1.4 Datenverarbeitung



Themenblock 1 Informationsverarbeitung

Lehrinhalte

1.4 Datenverarbeitung

- + Intro
- + Begriffsklärung
- + Typische Probleme
- + Fazit
- + Gruppenarbeit



1.4 Datenverarbeitung

Intro



*„Daten sind der Treibstoff und das Schmiermittel der vernetzten Wirtschaft. Ohne sie geht fast gar nichts mehr. Wir alle speisen den Ozean der Daten mit jedem Klick, jedem Einkauf, jedem Griff zur Bonuskarte oder der Eingabe auf dem Navigationsgerät. Experten haben für diese unerhörte Flut von Informationen, die Menschen wie Maschinen erzeugen, einen imposanten Begriff geprägt: **Big Data**.*

Das klingt nach Größe, Vehemenz, Tragweite, nach einer Zahl mit zu vielen Nullen, als dass sie ein Mensch noch verstehen könnte. Und „Big Data“ ist zugleich vage genug, um jede Menge Erklärungsversuche herauszufordern.“

Heuer, Steffan: **Kleine Daten, große Wirkung - Big Data einfach auf den Punkt gebracht**; In: Digitalkompakt #06, Hrsg. Landesanstalt für Medien Nordrhein-Westfalen (LfM), S. 3, 2013.

Big Data: Die vier Kriterien für Big-Data-Anwendungen in Anlehnung an die Anfangsbuchstaben der englischen Begriffe **V**olume, **V**elocity, **V**ariety und **V**alue. Dabei geht es um die Datenmenge, die Geschwindigkeit, mit der Daten eingehen und verarbeitet werden, die Vielfalt der Datensätze und schließlich den Wert, der sich aus ihrer Aufbereitung und Analyse gewinnen lässt.

1.4 Datenverarbeitung

Begriffsklärung



Daten: Zum Zweck der Verarbeitung zusammengefasste Zeichen, die aufgrund bekannter oder unterstellter Abmachungen, d.h. Angaben über Sachverhalte und Vorgänge, Informationen darstellen.

Klassifikation in strukturierte, semistrukturierte und unstrukturierte Daten

+ Strukturierte Daten

- Haben klar definierte Eigenschaften
- Sind einem festen Feld in einer Datenbank oder einer Tabelle zugeordnet
- Beispiel: USER_ID, NAME, IP-ADDRESS, etc.

+ Semistrukturierte Daten

- Unterliegen keiner festen Struktur, sondern tragen Strukturinformation mit sich
- Müssen keinem Typenmodell unterworfen werden. Somit kann sich eine Datensammlung aus semistrukturierten Daten beliebig erweitern. Ein Strukturmodell kann impliziert werden.
- Beispiel: HTML-Webseiten, XML-Dokumente mit und ohne DTD (Document Type Definition)

1.4 Datenverarbeitung

Begriffsklärung



Klassifikation in strukturierte, semistrukturierte und unstrukturierte Daten

+ Unstrukturierte Daten

- Entsprechen weder einem Datenmodell noch haben sie eine Struktur
- Können nicht in Form von Zeilen und Spalten wie in Datenbanken gespeichert werden
- Folgen keiner Semantik und keinen Regeln
- Haben kein bestimmtes Format und keine bestimmte Reihenfolge
- Haben keine leicht identifizierbare Struktur
- Da sie keine erkennbare Struktur haben, können sie von Computerprogrammen nicht einfach verwendet werden. Folge: Manuelle Datenaufbereitung.
- Beispiele: Bilder (JPEG, GIF, PNG, etc.), Videos, Memos, Reports, Word-Dokumente und PowerPoint Präsentationen, Umfragen

1.4 Datenverarbeitung

Begriffsklärung

Dynamische Ein- und Ausgabedaten

- + Generiert durch Programme oder Rückgabewerte von Kontextdaten oder einer **API**-Antwort
- + Verändern sich zur Laufzeit in Programmen

Beispiele

Einkäufe online wie offline | Finanznachrichten | Social Media Nachrichten | Wetterstationen | Finanzielle Transaktionen | Web-Protokolle aus einem Browser oder einer mobilen App | Verbindungsdaten von SMS und Telefonaten | Standortdaten von vernetzten Geräten vom Smartphone bis zur Digitalkamera | Verkehrsdaten aus einem Navigationsgerät einem Fahrzeug oder in die Straße eingebetteten Sensoren und Mautstationen | Sensordaten aus ganzen Fertigungsstraßen oder Warenlagern | von Containern und einzeltem Stückgut Biometrische und Vitaldaten vom Fitness-Studio bis zum Krankenhaus | Einträge in sozialen Medien | Video- und Tondateien

API: Application Programming Interface. Programmier-Schnittstelle, die es einem Stück Software erlaubt, bei entsprechender Zugangsberechtigung selbstständig Daten von einem anderen Programm abzurufen.



1.4 Datenverarbeitung

Begriffsklärung

Statische Ein- und Ausgabedaten

- + Fest vorgegebene Daten bspw. durch manuelle Eingabe von Textdaten durch Benutzer
- + I.d.R. strukturiert abgespeichert, oft als Tupel von Werten in Relationen in einer Datenbank
- + Ohne Bezug zu einem Zeitfaktor
- + Beispiele: Stammdaten, Festwerte, Konstanten, Klassenvariablen



1.4 Datenverarbeitung

Begriffsklärung

Datenverarbeitung – *Technische Merkmale*

- + Erfassung/Erhebung (Lesen): Eingabewerte von Anwender oder von Programmen
- + Speicherung (*Persistieren, dt. festschreiben*): in einem nichtflüchtigen Speicher wie bspw. Dateien und Datenbanken.
- + Verarbeitung: Wertmäßige Veränderung oder Veränderung der Information



1.4 Datenverarbeitung

Begriffsklärung



Datenverarbeitung – Technische Merkmale

- + Analyse: Verwendung von statistischen Methoden, um aus vorliegenden numerischen Einzeldaten zusammenfassende Informationen (Kenngrößen) zu gewinnen und tabellarisch oder grafisch aufzubereiten und zu dokumentieren. Datenanalyse kann in 6 Arten unterteilt und organisiert werden, die nach ihrem Schwierigkeitsgrad geordnet sind.
 - **Deskriptive Analyse:** Empirische Daten durch Tabellen und Grafiken übersichtlich darstellen und ordnen, sowie durch geeignete grundlegende Kenngrößen zahlenmäßig beschreiben
 - **Explorative Analyse:** Bisher unbekannte Strukturen und Zusammenhänge (Korrelationen und Beziehungen zwischen den Variablen) in den Daten zu finden und hierdurch neue Hypothesen zu generieren
 - **Inferentielle Analyse:** Verallgemeinerung einer größeren Population mit einer kleineren Stichprobengröße
 - **Prädiktive Analyse:** Anhand von Daten Vorhersagen über die Zukunft treffen
 - **Kausale Analyse:** Konzentriert sich auf die Suche nach der Ursache für eine Korrelation zwischen Variablen
 - **Mechanistische Analyse:** Genaue Veränderungen in den Variablen messen, die zu anderen Veränderungen in anderen Variablen führen

Vgl. Grellmann, Martin: <https://martin-grellmann.de/die-sechs-arten-der-datenanalyse>, abgerufen am 19.10.2022

1.4 Datenverarbeitung

Begriffsklärung



Dateiformat – *Merkmale*

- + Definiert die Syntax und Semantik von Daten innerhalb einer Datei
- + Stellt eine bidirektionale Abbildung von Information auf einen eindimensionalen binären Speicher dar
- + I.d.R. durch Software-Hersteller (proprietäres Dateiformat) oder durch ein standardisierendes Gremium festgelegt
- + Auch aus proprietären Dateiformaten können sich Standardformate entwickeln, wenn sie dokumentiert sind und von anderen aufgegriffen werden.
- + Standardformate ermöglichen, dass Software verschiedener Hersteller mit den gleichen Dateiformaten arbeiten
- + In einer Spezifikation sollte die Art der Codierung und Anordnung von Daten innerhalb eines Dateiformats genau beschrieben werden.
- + Für viele Dateiformate sind die Spezifikationen veröffentlicht, andere Spezifikationen werden als Geschäftsgeheimnisse behandelt.
- + Beispiele: .txt | .md | .ipynb | .docx | .pdf | etc.

1.4 Datenverarbeitung

Begriffsklärung



Datenverarbeitung – Juristische Merkmale

- + „Verarbeitung“ schließt eine Vielzahl unterschiedlicher mit oder ohne Hilfe automatisierter Verfahren ausgeführter Vorgänge im Zusammenhang mit personenbezogenen Daten ein. Sie umfasst das **Erheben, das Erfassen, die Organisation, das Ordnen, die Speicherung, die Anpassung oder Veränderung, das Auslesen, das Abfragen, die Verwendung, die Offenlegung durch Übermittlung, Verbreitung** oder eine andere Form der Bereitstellung, den **Abgleich oder die Verknüpfung, die Einschränkung, das Löschen oder die Vernichtung** personenbezogener Daten.
- + Die Datenschutz-Grundverordnung (DSGVO) gilt für die vollständig oder teilweise automatisierte Verarbeitung personenbezogener Daten sowie für die nichtautomatisierte Verarbeitung personenbezogener Daten, die in einem strukturierten Dateisystem gespeichert sind.

https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-constitutes-data-processing_de, abgerufen am 19.10.2022

- + **Grundsätzlich: Gebot der Datenminimierung**
- + Datenschutz ist grundsätzlich kein „*Spielverderber*“, sondern kümmert sich um die eigenen Grundrechte. Das Recht auf informationelle Selbstbestimmung gibt jedem von uns das Recht, grundsätzlich selbst über die Preisgabe und Verwendung unserer personenbezogenen Daten zu bestimmen. Datenschutz ist ein **Freiheitsrecht**.
- + Mehr dazu im 2. Semester in DSCB230

1.4 Datenverarbeitung

Begriffsklärung



Darstellung/Visualisierung

- + Abbildung von Daten in einem Modell mit ihren Beziehungen zueinander durch einen Zeichenvorrat
- + Informationsdarstellung durch Ziffern, Buchstaben, Sonderzeichen
- + Visualisierung mittels geeigneter Methoden und grafischen Elementen
- + Beispiele: Tabelle, Abbildung, Grafik, Animation, Graph, Karte, Diagramm,interaktive Infografik, Logo, Label, Cartoon, Porträt, Schema, Piktogramm, Foto, Pinnwand



1.4 Datenverarbeitung

Begriffsklärung



Kontext Betriebswirtschaft

- + Organisationsdaten
 - Strukturelle Informationen, welches die Objekte eines Unternehmens beschreibt und in eine Ordnungsrelation darstellen kann z. B. Organigramm.

- + Stammdaten
 - Wichtige Grunddaten eines Unternehmens, die über einen Zeitraum, i.d.R. langfristig, nicht verändert werden
 - Repräsentieren meist Datensätze wie z. B. Kunden, Material, Kondition, Nachrichten

- + Bewegungsdaten
 - Vorgangsbezogene Daten, die kurzlebig sind und bestimmten Stammdaten zugeordnet werden z. B. Kundenauftrag.



1.4 Datenverarbeitung

Begriffsklärung

Kontext Betriebswirtschaft

+ Dokumente

- Datensätze, die generiert werden, wenn ein Geschäftsvorfall durchgeführt wurde
- Stellt ein Protokoll des Geschäftsvorfalls dar
- Enthält alle relevanten, vordefinierten Informationen aus den Organisationsdaten, Stammdaten und Bewegungsdaten

Beispiele

- + Verkaufsbeleg
- + Einkaufsbeleg
- + Materialbeleg
- + Buchungsbeleg
- + Zahlungsbeleg



1.4 Datenverarbeitung

Begriffsklärung

Kontext Statistik

- + Daten als Messwerte, die im Rahmen von Befragungen, Beobachtungen oder Experimenten erhoben werden.
- + Für gewöhnlich entweder numerisch oder kategorial
 - Numerische Daten: mathematische Bedeutung als Messwert, wie Anzahl oder Gewicht
 - Kategoriale Daten: bezeichnen Merkmale wie Geschlecht, Meinung oder Familienstand
- + Die Gesamtheit aller Daten, also numerische und kategoriale, wird als *Stichprobe* bezeichnet



1.4 Datenverarbeitung

Begriffsklärung



Metadaten: Metadaten oder Metainformationen sind strukturierte Daten, die Informationen über Merkmale anderer Daten enthalten. Dazu zählen auch Angaben von Eigenschaften eines einzelnen Objektes.

Merkmale

- + Der Zweck bestimmt das Ergebnis: Abhängig von der Art der Daten und ihrer Verwendung.
- + I.d.R. ausführlicher oder besser strukturiert als die inhaltlichen Daten, die sie beschreiben.

Beispiele

- + Computerdatei: Dateiname, Dateityp, Größe, Attribute: Zugriffsrechte, Erstellungsdatum, Änderungsdatum
- + Wissenschaftlich-bibliothekarische Regelwerke:
 - Buch: Autor, Titel, Verlag, Erscheinungsjahr, Anzahl der Seiten, ISB Nummer
 - Technische Zeichnung: Name des Konstrukteurs, Bauteilnummer, Zeichnungsnummer, Versionsnummer, Datum
 - Digitales Foto: Aufnahmedatum, Abmessungen (Pixel), Auflösung (DPI), Bittiefe (Bit), Farbdarstellung, Geräteinformationen, Beschreibung, Schlagworte (Tags)

1.4 Datenverarbeitung

Begriffsklärung



Datenkompetenz (engl. Data Literacy): Die Fähigkeit, Daten zu erkunden, Daten zu verstehen und mit Daten zu kommunizieren. Dazu gehört auch die Fähigkeit zu kritischem Denken bei der Anwendung bzw. Interpretation von Daten und bei deren Nutzung für die Entscheidungsfindung. Schließlich sollte man auch in der Lage sein, die Bedeutung einer Entscheidung zu vermitteln und sie gegenüber anderen zu vertreten.

Hal Varian, [The McKinsey Quarterly](#), Januar 2009:

"The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids."

Heidrich, Jens; Bauer, Pascal; Krupka, Daniel: **Google's Chief Economist Hal Varian on Statistics and Data**

<https://flowingdata.com/2009/02/25/googles-chief-economist-hal-varian-on-statistics-and-data/>, abgerufen am 19.10.2022

1.4 Datenverarbeitung

Begriffsklärung



Datenkompetenz – *Merkmale*

- + Vielfältige Kompetenzen hinsichtlich des Umgangs mit und der Nutzung von Daten, die heute in allen Bereichen der Gesellschaft von Bedeutung sind.
- + Neben der Ausbildung in Data Science, die ein vertieftes Wissen im Umgang von Daten bedingt, muss auch die Ausbildung von Data-Literacy-Kompetenzen für alle Menschen als wichtig angesehen werden.
- + Grundlegende Aspekte der Datenerfassung, -speicherung, -verarbeitung und -analyse sowie der Arbeit mit entsprechenden Ergebnissen (und deren Verständnis) wird in alle Disziplinen und Lebensbereiche Einzug halten.
- + Damit rückt Data Literacy (auch politisch) mehr in die Notwendigkeit des generellen „Lehren und Erlernen von Datenkompetenzen“.

1.4 Datenverarbeitung

Begriffsklärung



Datenqualität: Die Gesamtheit der Ausprägungen von Qualitätsmerkmalen eines Datenbestandes bezüglich dessen Eignung, festgelegte und vorausgesetzte Erfordernisse zu erfüllen.

- + Der Begriff der **Informationsqualität** wird häufig synonym verwendet
- + Ansatz zur Strukturierung der Datenqualitäts-Merkmale in drei Hauptmerkmale nach **Larry English**:
 - Qualität der Datendefinition
 - Inhaltliche Datenqualität (Korrektheit der Datenwerte)
 - Qualität der Datenpräsentation
- + Maßnahmen und Methoden zur Verbesserung der Datenqualität werden unterschieden in technische und organisatorische Bereiche.
- + Etablierung von Datenqualitätsmanagement (DQM), ein qualitätsorientiertes Management der Daten
- + Mehr dazu im 2. Semester in DSCB230

1.4 Datenverarbeitung

Typische Probleme

Inkompatibilität

- + Darstellung von Zeichenketten
- + Quellencodierung: Blockcodes und Huffman-Codierung
- + Zeichensätze: ASCII und Unicode
- + Byte-Reihenfolge im Speicher: Little- und Big-Endian, siehe [4] S. 210
- + Strukturierte Datentypen: Basis-Datentypen, Tabellen, relationale Datenmodelle
- + Semistrukturierte Datentypen: Metadaten in digitalen Fotos (ein oder mehrere Klassifizierungsattribute), JSON
- + Unstrukturierte Datentypen: liegen in absoluter Rohform vor, ohne bestimmtes Format und sind qualitativ und nicht quantitativ



1.4 Datenverarbeitung

Typische Probleme

Interoperabilität

- + Die Fähigkeit unterschiedlicher Systeme, möglichst nahtlos an der gleichen Sache gemeinsam zu arbeiten.
- + Standards für interoperable Metadaten haben die Aufgabe, Metadaten aus unterschiedlichen Quellen nutzbar zu machen. Sie umfassen dazu zunächst die Aspekte Semantik, Datenmodell und Syntax.
- + Austauschformate (Auswahl)
 - **TXT**: Eine TXT-Datei ist ein Standard-Textdokument, das nicht formatierten Text enthält. Es wird von jedem Textverarbeitungsprogramm erkannt und kann von den meisten anderen Programmen verarbeitet werden.
 - **CSV (Comma-separated Values)**: Beschreibt den Aufbau einer Textdatei zur Speicherung oder zum Austausch einfach strukturierter Daten.
 - **XML (Extensible Markup Language)**: Darstellung hierarchisch strukturierter Daten im Format einer Textdatei, die sowohl von Menschen als auch von Maschinen lesbar ist.
 - **JSON (JavaScript Object Notation)**: Programmiersprachenunabhängig, für Menschen einfach zu lesen/schreiben
 - **PDF (Portable Document Format)**: Plattformunabhängiges Dateiformat, das 1992 vom Unternehmen Adobe Inc. entwickelt und veröffentlicht wurde und aktuell von der PDF Association weiterentwickelt wird.



1.4 Datenverarbeitung

Typische Probleme

Unbegründete und ungeprüfte Größenannahmen

- + Wertebereiche und Überläufe
- + Länge von Datenstrukturen



1.4 Datenverarbeitung

Typische Probleme

Fehlinterpretation

- + Postleitzahlen oder Telefonnummern?
- + Das sieht aus wie ein
- + Das wird schon ein ... sein!



1.4 Datenverarbeitung

Typische Probleme

Technische Überforderung

- + Überlaufprobleme
- + Ungenauigkeit bei Gleitkommawerten
- + Überschreiben bei Update von Werten in Dateien und Datenbanktabellen
- + Dateinamen-Gleichheit: Gefahr des Überschreibens
- + Datei-Zugriff
 - Datei nicht vorhanden
 - Dateiformat inkompatibel mit Anwendung
 - Zugriffsrechte für Lesen fehlen
 - Zugriffsrechte für Schreiben fehlen



1.4 Datenverarbeitung

Fazit

„Aufhalten lässt sich der Wandel zum datengetriebenen Leben nicht. Menschen passen ihren Lebenswandel den ständig datenhungrigen Geräten an, die ihr Leben einfacher machen sollen, aber bemerken erst später, welchen hohen Preis sie für den Komfort der immer verfügbaren Dienste bezahlen.

Nämlich die Gefahr, quer durchs Netz und von ihnen unbekannten Anbietern verfolgt, umworben und ausgespäht zu werden. Auf der Strecke bleibt oft die Gelegenheit, informierte Entscheidungen zu treffen und die Einspeisung der eigenen Daten zu verweigern.“

Heuer, Steffan: **Kleine Daten, große Wirkung - Big Data einfach auf den Punkt gebracht**; In: Digitalkompakt #06, Hrsg. Landesanstalt für Medien Nordrhein-Westfalen (LfM), S. 38, 2013.



1.4 Datenverarbeitung

Gruppenarbeit



Welche Erfahrungen haben Sie selbst bei der Verarbeitung von Daten gemacht?

- + Nutzen Sie als Orientierungshilfe die angegebenen Kontexte
- + Verwenden Sie [Collaboard](#)

Kontexte

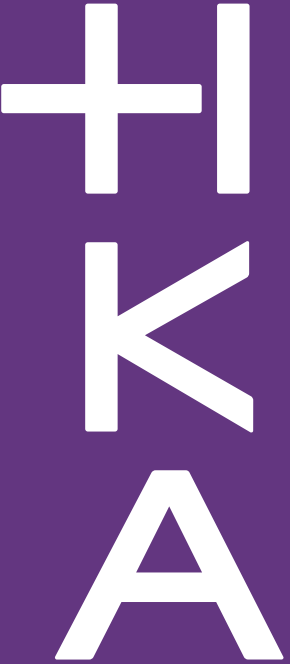
Inkompatibilität | Interoperabilität | Unbegründete und ungeprüfte Größenannahmen | Fehlinterpretation | Technische Überforderung | Struktur | Validität | Aktualität | Fehlerfreiheit | Menge | Schnittstellen | Lokalität: On Premises vs. Cloud | Datenhoheit (juristisch) | Datenschutz (DSGVO)





Hochschule Karlsruhe
University of
Applied Sciences

Fakultät für
Informatik und
Wirtschaftsinformatik



www.h-ka.de