Kapitel 1

Einführung

1.1 Der Begriff Informatik

Das Wort **Informatik** setzt sich aus zwei Teilen zusammen: *Information* und *Automatik*. Ursprünglich wurde der Begriff in den 1950er Jahren in Frankreich geprägt („informatique“) und fand später auch im deutschsprachigen Raum Verbreitung. In anderen Sprachen wird der Bereich meist „Computer Science“ genannt, was die technische Seite stärker betont.

1.1.1 Information

Der erste Bestandteil, *Information*, bezieht sich auf Daten, die eine Bedeutung tragen. Information ist also nicht einfach nur eine Ansammlung von Zeichen oder Zahlen, sondern sie entsteht erst durch die *Interpretation* von Daten in einem bestimmten Kontext. Beispielsweise ist die Zahl „42“ zunächst nur ein einzelner Wert. Wird sie jedoch im Zusammenhang mit einer Temperaturangabe, einem Alter oder einem Ergebnis verstanden, wird daraus eine Information. Die Informatik beschäftigt sich also damit, Informationen *darzustellen*, *zu speichern*, *zu übertragen* und *zu verarbeiten*.

1.1.2 Automatik

Der zweite Bestandteil, *Matik* (von Automatik), bedeutet, dass diese Informationsverarbeitung durch Maschinen – insbesondere Computer – automatisiert geschieht. Ein zentrales Ziel der Informatik ist es, Verfahren zu entwickeln, die es ermöglichen, Informationen mithilfe von Computern effizient und zuverlässig zu verarbeiten. Dazu gehören das Erstellen von Programmen, das Entwerfen von Algorithmen sowie die Entwicklung von Systemen, die Informationen ohne manuelle Eingriffe verarbeiten können.

1.2 Bedeutung der Informatik

Die Informatik kann somit als *Wissenschaft von der systematischen Verarbeitung von Informationen, insbesondere mit Hilfe von Computern*, verstanden werden.

Sie ist nicht nur eine technische Disziplin, sondern verbindet Elemente aus Mathematik, Ingenieurwissenschaften, Logik und zunehmend auch Sozial- und Geisteswissenschaften. Heute prägt die Informatik nahezu alle Bereiche des täglichen Lebens: von Smartphones und dem Internet über moderne Autos bis hin zu Medizin, Wirtschaft und Wissenschaft.

1.3 Was tun eigentlich Computer?

Computer *verarbeiten Daten*. Sie führen Berechnungen aus, speichern, übertragen und strukturieren Daten — aber sie „verstehen“ keine Bedeutung im menschlichen Sinn. Bedeutung (Information) entsteht erst beim Menschen (oder in einem Modell), wenn Daten in *Kontext* gesetzt werden. Der Kernprozess ist daher zweistufig:

1. **Repräsentation:** Aus *Information* werden *Daten*, indem wir festlegen, wie etwas als Zeichen/Zahlen (Bits) dargestellt wird.

2. **Abstraktion:** Aus *Daten* wird (wieder) *Information*, indem wir Details weglassen, strukturieren und die Daten in einem Modell deuten.

1.3.1 Von Information zu Daten: Repräsentation

Repräsentation bedeutet: Wir legen eine **Abbildung** fest, die etwas Bedeutungsvolles (Information) in ein **Datenformat** überführt, das der Computer verarbeiten kann. Formal kann man das als Funktion auffassen:

$$\text{rep} : \text{Information} \times \text{Kontext} \rightarrow \text{Daten (Bits)}.$$

1.3.2 Von Daten zu Information: Abstraktion

Abstraktion bedeutet: Wir **interpretieren** Daten in einem passenden Modell und **lassen Details weg**, die für die Fragestellung nicht nötig sind. So entsteht Bedeutung. Formal:

$$\text{abs} : \text{Daten (Bits)} \times \text{Modell/Kontext} \rightarrow \text{Information}.$$

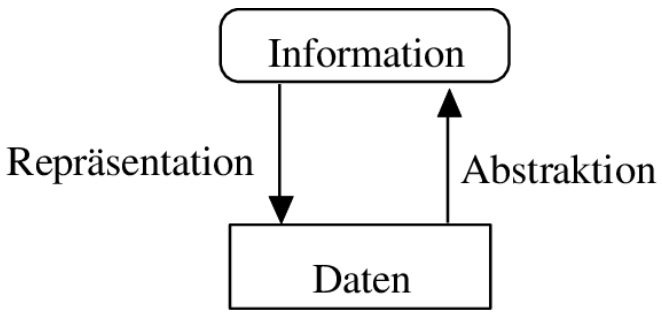

Abbildung 1.1: Wechselspiel zwischen Repräsentation (Information Daten) und Abstraktion (Daten Information).

1.4 Bits und Bytes

Ein **Bit** (binary digit) ist die kleinste Informationseinheit im Rechner: Es kann genau zwei Zustände annehmen, meist als `0` und `1` notiert. Physikalisch werden Bits z. B. durch zwei Spannungsniveaus, magnetische Ausrichtungen oder Lichtimpulse realisiert.

Einzelne Bits sind für die Verarbeitung jedoch zu fein. Deshalb werden Bits zu **Gruppen** zusammengefasst und *gruppenweise* gelesen/geschrieben:

- **Byte** = 8 Bit (heute die gebräuchlichste Grundeinheit; in den meisten Architekturen zugleich die *kleinste adressierbare Einheit*).
- **Nibble** = 4 Bit (halbes Byte; nützlich z. B. bei Hexadezimaldarstellungen).
- **Wort (Word)** = *architekturabhängige* Arbeitsbreite der CPU (typisch 16, 32 oder 64 Bit).
- **Doppelwort/Quadwort** = Vielfache der Wortbreite (z. B. 32/64/128 Bit).

1.4.1 Wer bestimmt die Lesegruppen?

Hardware: Die *Register- und ALU-Breite* eines Prozessors legt fest, wie viele Bits er in einem Schritt besonders effizient verarbeiten kann (z. B. 32-Bit oder 64-Bit). Auch *Datenbus* und *Cache-Zeilengrößen* begünstigen das Holen/Speichern ganzer Bytes-, Wort- oder Mehrfach-Wort-Blöcke. Moderne CPUs können zusätzlich mit Vektor/SIMD-Einheiten noch größere Pakete (z. B. 128/256/512 Bit) auf einmal verarbeiten.

Software/Betriebssystem: Damit die Hardwarebreite genutzt werden kann, muss das *Betriebssystem* die Architektur unterstützen (32-Bit- oder 64-Bit-Modus, Treiber, Systembibliotheken, ABI). Ein 64-Bit-Prozessor entfaltet seine Vorteile erst vollständig mit einem 64-Bit-Betriebssystem und passenden Programmen; andernfalls arbeitet er im 32-Bit-Kompatibilitätsmodus.

Beispiele.

- **32-Bit-System:** Die CPU arbeitet effizient mit 32-Bit-Wörtern (z. B. `int32`); Zeiger/Adressen sind 32 Bit breit. Daten werden häufig in 32-Bit-Schritten geladen/geschrieben, obwohl der Speicher byteweise adressiert wird.
- **64-Bit-System:** Register und Zeiger sind 64 Bit breit. Das System kann größere Zahlenbereiche adressieren und pro Schritt breitere Daten verarbeiten; trotzdem bleiben Bytes (8 Bit) die kleinste adressierbare Einheit.

1.5 Größe der Daten

1 k	=	1024 Bit	=	2^{10}	(k = Kilo)
1 M	=	1024×1024 Bit	=	2^{20}	(M = Mega)
1 G	=	$1024 \times 1024 \times 1024$ Bit	=	2^{30}	(G = Giga)
1 T	=	$1024 \times 1024 \times 1024 \times 1024$ Bit	=	2^{40}	(T = Tera)
1 P	=	$1024 \times 1024 \times 1024 \times 1024 \times 1024$ Bit	=	2^{50}	(P = Peta)
1 E	=	$1024 \times 1024 \times 1024 \times 1024 \times 1024 \times 1024$ Bit	=	2^{60}	(E = Exa)
1 Z	=	$1024 \times 1024 \times 1024 \times 1024 \times 1024 \times 1024 \times 1024$ Bit	=	2^{70}	(Z = Zetta)
1 Y	=	$1024 \times 1024 \times 1024 \times 1024 \times 1024 \times 1024 \times 1024 \times 1024$ Bit	=	2^{80}	(Y = Yotta)

1.6 Textdarstellung

Von ASCII zu Unicode — warum überhaupt?

Frühe Computersysteme nutzten **ASCII** (American Standard Code for Information Interchange, 1960er Jahre). ASCII ist ein 7-Bit-Zeichensatz mit 128 Zeichen: lateinische Grundbuchstaben A–Z/a–z, Ziffern, Satzzeichen und Steuerzeichen (z. B. Zeilenumbruch). Für englischen Text genügte das, *aber*: Umlaute (ä, ö, ü), Akzente (é), das Euro-Zeichen (€), kyrillisch (Ж), griechisch (Ω), arabisch (م), asiatische Schriften (日), Emojis (😊) *fehlten*. „é“ (NFC) vs. „é“ (NFD) bzw. „e + ◌́“. *fehlten*.

Als Übergang entstanden viele **„erweiterte ASCII“-Codepages** (8-Bit, 256 Zeichen), z. B. `ISO-8859-1`, `Windows-1252`, `KOI8-R`. Jede deckte *einen* Sprachraum ab. Ergebnis: Inkompatibilitäten (sogenanntes *Mojibake*), weil dieselben Bytewerte je nach Codepage andere Zeichen bedeuten.

Unicode (seit 1991) löst dieses Grundproblem: *ein* weltweiter Zeichensatz für *alle* Schriftsysteme, Symbole, technische Zeichen und Emojis. Ziel: *„Jedem Zeichen ein eindeutiger Codepunkt“* — unabhängig von Sprache, Plattform oder Anwendung.

Begriffe sauber trennen

- **Codepunkt** (Unicode): eine Nummer in der Form `U+XXXX` (z. B. `U+00E4` = „ä“, `U+20AC` = „€“, `U+1F60A` = „□“).
- **Kodierung** (Encoding): konkrete Regel, wie Codepunkte in *Bytes* umgesetzt werden (z. B. UTF-8, UTF-16, UTF-32).
- **Graphem-Cluster**: was der Mensch als „ein Zeichen“ wahrnimmt (z. B. „e“ + kombinierender Akzent `U+0301` → „é“, oder Familien-Emoji aus mehreren Codepunkten).

Wichtige Unicode-Kodierungen

- **UTF-8** (variabel, 1–4 Byte pro Codepunkt): ASCII-Zeichen bleiben 1 Byte (abwärtskompatibel), alle anderen werden als 2–4 Byte kodiert. Heutzutage Standard im Web, in Dateien und Schnittstellen.

- **UTF-16** (variabel, 2 oder 4 Byte): Basis-Mehrsprachige Ebene (BMP) meist 2 Byte; Supplementärzeichen (z. B. viele Emojis) als Surrogatpaare (4 Byte). Achtung auf Byte-Reihenfolge (*Endianness*) und optionales *BOM*.
- **UTF-32** (fix 4 Byte): einfacher, aber speicherintensiv; praktisch v. a. intern in manchen Systemen.

Historische Entwicklung in Kürze

1963 ASCII (7-Bit) standardisiert Grundzeichen und Steuerzeichen.

1980er Viele 8-Bit-Codepages (`ISO-8859-x`, `Windows-125x`) — regionale Lösungen, wenig kompatibel.

1991+ Unicode-Projekt: *ein* universeller Zeichensatz; Trennung von *Zeichen* (Codepunkte) und *Kodierung* (UTFs).

2000er+ UTF-8 setzt sich global durch (Internet, Linux/Unix, moderne Apps und Protokolle).

Wie wird Unicode praktisch genutzt?

- **Dateien und Protokolle:** Textdateien, JSON, HTML, E-Mails, Datenbanken — fast überall ist UTF-8 üblich. Wichtig: *Encoding angeben* (z. B. HTTP `Content-Type`, HTML `<meta charset=ütf-8»`, DB-Kollation).
- **Betriebssysteme:** Dateinamen und Konsolen sind (je nach System) Unicode-fähig; moderne Terminals verstehen UTF-8.
- **Programmiersprachen:** Python, Java, JavaScript, C# u. a. arbeiten intern mit Unicode-Zeichenketten; I/O nutzt meist UTF-8.