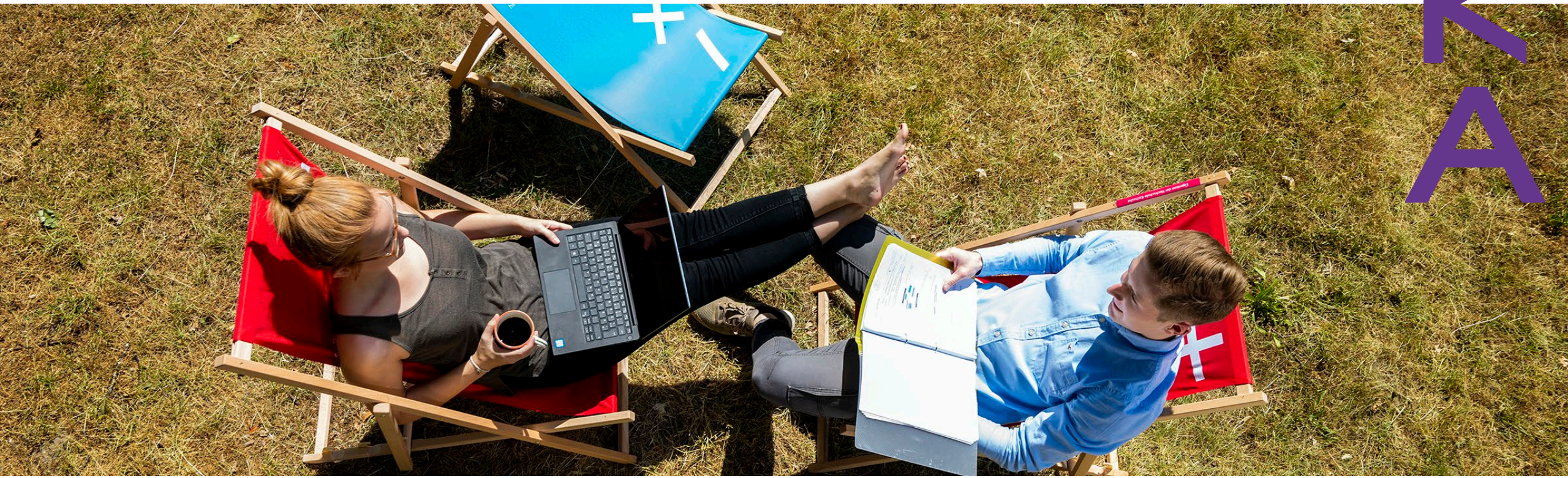
Hochschule Karlsruhe
University of
Applied Sciences

Fakultät für
Informatik und
Wirtschaftsinformatik

DSCB130 | 1.3 Nachricht und Interpretation



Matthias Mruzek-Vering M.Sc.

Foto: Quirin Leppert

Themenblock 1 Informationsverarbeitung

Lehrinhalte

1.3 Nachricht und Interpretation

- Begriffsklärung
- Kodieren und Interpretieren
- Informationsgehalt
- Auftrittswahrscheinlichkeiten von Buchstaben

1.3 Nachricht und Interpretation

Begriffsklärung

Nachricht

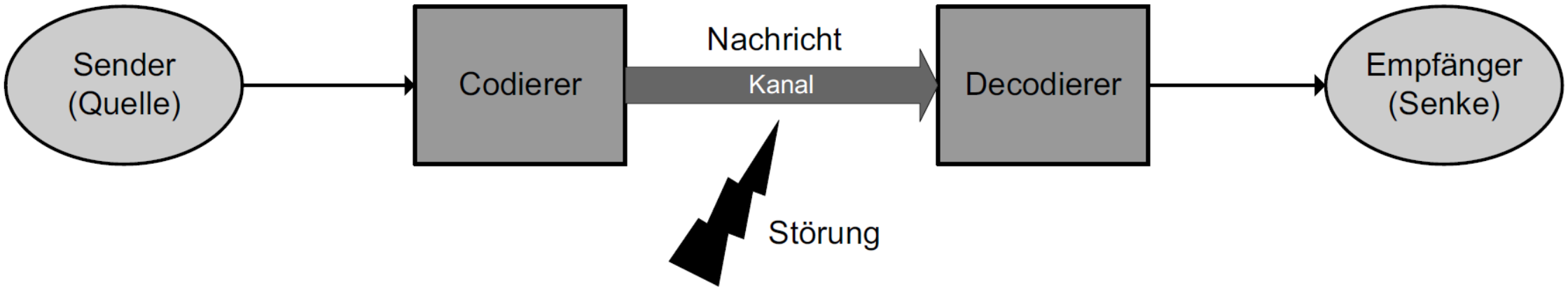
+ Folge von Zeichen, die von einem Sender (*Quelle*) ausgehend, in irgendeiner Form einem Empfänger (S*enke*) übermittelt wird
+ Nachrichtenübermittlung erfolgt dabei im technischen Sinne über einen *Kanal* (Kabel, Funk, TV, Briefpost, etc.)
+ Umformung der Originalnachricht in das für die Übertragung erforderliche Medium durch einen *Codierer*
+ Empfangene Signale wird durch einen *Decodierer* wieder in eine durch den Empfänger lesbare Form umgewandelt
+ Möglichkeit einer *Störung* der Nachricht während der Übermittlung

1.3 Nachricht und Interpretation

Begriffsklärung

Nachricht

+ Schematische Darstellung des Modells einer Nachrichtenübermittlung



Vgl. [5] S. 37

+ Nachrichten sind konkrete, wenn auch idealisiert immaterielle Objekte, die von einem Sender zu einem Empfänger übertragen werden können.
+ Allerdings wird die Nachricht meist nicht in ihrer ursprünglichen Form, sondern in einer technisch angepassten Art und Weise übertragen, z. B. akustisch, optisch oder mithilfe elektromagnetischer Wellen.

1.3 Nachricht und Interpretation

Begriffsklärung

Nachricht

+ Exakte Definition einer Nachricht baut auf dem Begriff des Alphabets auf

Alphabet: Ein Alphabet A besteht aus einer abzählbaren Menge von Zeichen (Zeichenvorrat) und einer Ordnungsrelation, d. h. eine Regel, durch die eine feste Reihenfolge der Zeichen definiert ist.
Meist betrachtet man Alphabete mit einem endlichen Zeichenvorrat, gelegentlich auch abzählbar unendliche Zeichenvorräte wie die natürlichen Zahlen.

Nachricht: Eine Nachricht ist eine aus den Zeichen eines Alphabets gebildete Zeichenfolge.
Diese Zeichenfolge muss nicht endlich sein, aber abzählbar (d. h. man muss die einzelnen Zeichen durch Abbildung auf die natürlichen Zahlen durchnummerieren können), damit die Identifizierbarkeit der Zeichen sichergestellt ist.

Vgl. [5] S. 37

1.3 Nachricht und Interpretation

Begriffsklärung

Alphabete

Die folgenden geordneten, abzählbaren Mengen sind nach Definition 2.1 Alphabete:

a) {a, b, c, . . . , z} Die Menge aller Kleinbuchstaben in lexikografischer Ordnung.
b) {0, 1, 2, . . . , 9} Die Menge der ganzen Zahlen 0 bis 9 mit der Ordnungsrelation „<“.
c) {♦, ♥, ♠, ♣} Die Menge der Spielkartensymbole in der Reihenfolge ihres Spielwertes.
d) {2, 4, 6, . . . } Die unendliche Menge der geraden natürlichen Zahlen mit Ordnung „<“.
e) {0, 1} Die Binärziffern 0 und 1 mit $0 < 1$.

Vgl. [5] S. 38

1.3 Nachricht und Interpretation

Begriffsklärung

Nachrichtenraum

+ Die Menge aller Nachrichten, die mit den Zeichen eines Alphabets A gebildet werden können, heißt Nachrichtenraum $N(A)$ oder A^* über A.
+ Beschränkung auf Zeichenreihen mit einer maximalen Länge s
+ Der eingeschränkte Nachrichtenraum $N(A^s)$ umfasst nur endlich viele Elemente, sofern das zu Grunde liegende Alphabet endlich ist.

1.3 Nachricht und Interpretation

Begriffsklärung

Diskretisierung von Nachrichten

+ Nachrichten müssen vor ihrer digitalen Verarbeitung aus der für gewöhnlich kontinuierlichen Form durch Diskretisierung (Digitalisierung) in eine diskrete Form überführt werden.
+ Voraussetzung: Die Nachricht liegt als reelle Funktion vor, die stetig oder zumindest von beschränkter Schwankung (lebesgue-integrierbar*) ist.
+ Insbesondere darf die entsprechende Funktion keine Unendlichkeitsstellen (Pole) haben.
+ Anschaulich bedeutet dies, dass in der Nachricht wohl Sprünge vorkommen dürfen, dass aber die jeweils zugeordneten physikalischen Werte, z. B. Helligkeiten oder Tonhöhen, immer endliche Beträge aufweisen müssen.

Weiterführendes Material:

- *https://de.wikipedia.org/wiki/Lebesgue-Integral
- Shannonsches Abtasttheorem, siehe C. Shannon. A mathematical theory of communication. Bell System Techn. Journ., 27:379–423, 1948.

1.3 Nachricht und Interpretation

Begriffsklärung

Abtastung

+ Als *Abtastung* oder *Sampling* bezeichnet man die Abtastung der Werte einer Funktion an bestimmten vorgegebenen Stellen, also die Diskretisierung des Definitionsbereichs der Funktion.
+ Der kontinuierliche Verlauf des Funktionsgraphen wird dann durch eine **Treppenfunktion** oder eine Anzahl von Pulsen angenähert, die im Allgemeinen äquidistant auf dem Definitionsintervall der Funktion angeordnet sind.
+ Der Vorgang der Abtastung ist in der folgenden Abbildung veranschaulicht. Stellt man sich die zu verarbeitende Funktion $f(t)$ als eine Funktion der Zeit t vor, so bedeutet die Abtastung, dass der Funktionswert $f(t)$ in äquidistanten Zeitschritten t_s (gleichen Zeit-Abständen) bestimmt wird.

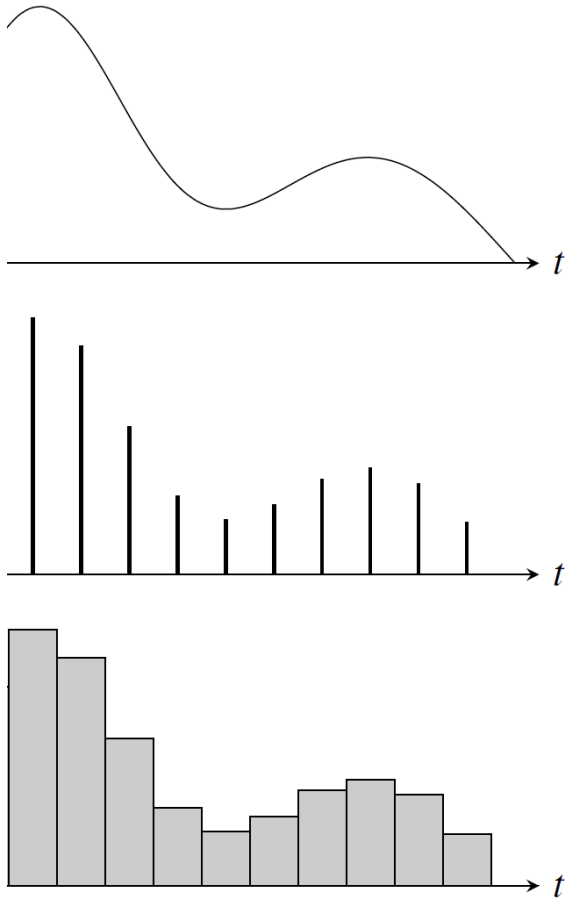


Abb. 2.3 Abtastung einer Funktion
Oben: Kontinuierliche Funktion $f(t)$
keit von der Zeit
Mitte: Abgetastete Funktion: Darstell
ne Folge äquidistanter Pulse
Unten: Abgetastete Funktion: Altern
lung durch eine Treppenfunktion

Vgl. [5] S. 44

1.3 Nachricht und Interpretation

Begriffsklärung

Quantisierung (Verfahren)

+ Der Übergang von einer kontinuierlichen zu einer digitalen Nachricht erfordert nach der Abtastung noch einen **zweiten Diskretisierungsschritt**, die Quantisierung.
+ Dazu wird der Wertebereich der zu diskretisierenden Funktion in eine Menge von Zahlen abgebildet, die das Vielfache einer bestimmten Zahl sind, des sogenannten Quantisierungsschritts.
+ Hierbei wird wieder vorausgesetzt, dass die zu quantisierende Funktion beschränkt ist, denn nur dann führt die Quantisierung schließlich auf eine endliche Menge von Zahlen – und nur endliche Mengen von Zahlen können technisch verarbeitet werden.
+ **Fazit:** Quantisierung ermöglicht die Umwandlung einer beliebigen Nachricht zu einer digitalen Nachricht, die aus einer endlichen Folge von natürlichen Zahlen besteht, die ihrerseits wieder in ein beliebiges Alphabet abgebildet werden können.
+ Den hier beschriebenen Vorgang der digitalen Abtastung einer analogen Nachricht bezeichnet man auch als *Pulscode-Modulation*. Dies ist die Grundvoraussetzung für die Verarbeitung von Nachrichten mithilfe einer digitalen Datenverarbeitungsanlage. Dieser Vorgang der Quantisierung wird auf der folgenden Seite verdeutlicht.

1.3 Nachricht und Interpretation

Begriffsklärung

Quantisierung

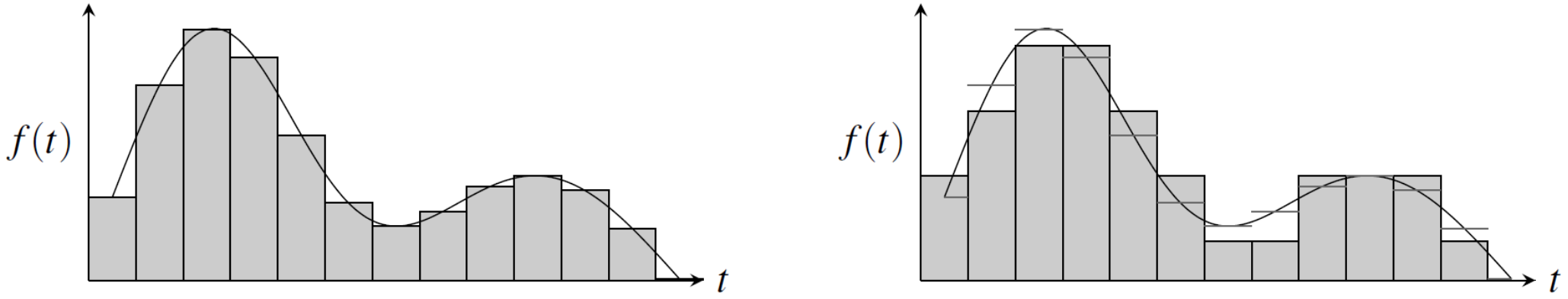


Abb. 2.4 Quantisierung einer abgetasteten Funktion $f(t)$.
Links: Abgetastete Funktion $f(t)$ in Abhängigkeit von der Zeit, jedoch mit kontinuierlichem Wertebereich
Rechts: Die gleiche Funktion mit quantisiertem Wertebereich und abgetasteten Definitionsbereich. Die ursprünglichen Werte sind als Balken mit eingezeichnet

Vgl. [5] S. 45

1.3 Nachricht und Interpretation

Begriffsklärung

Interpretation

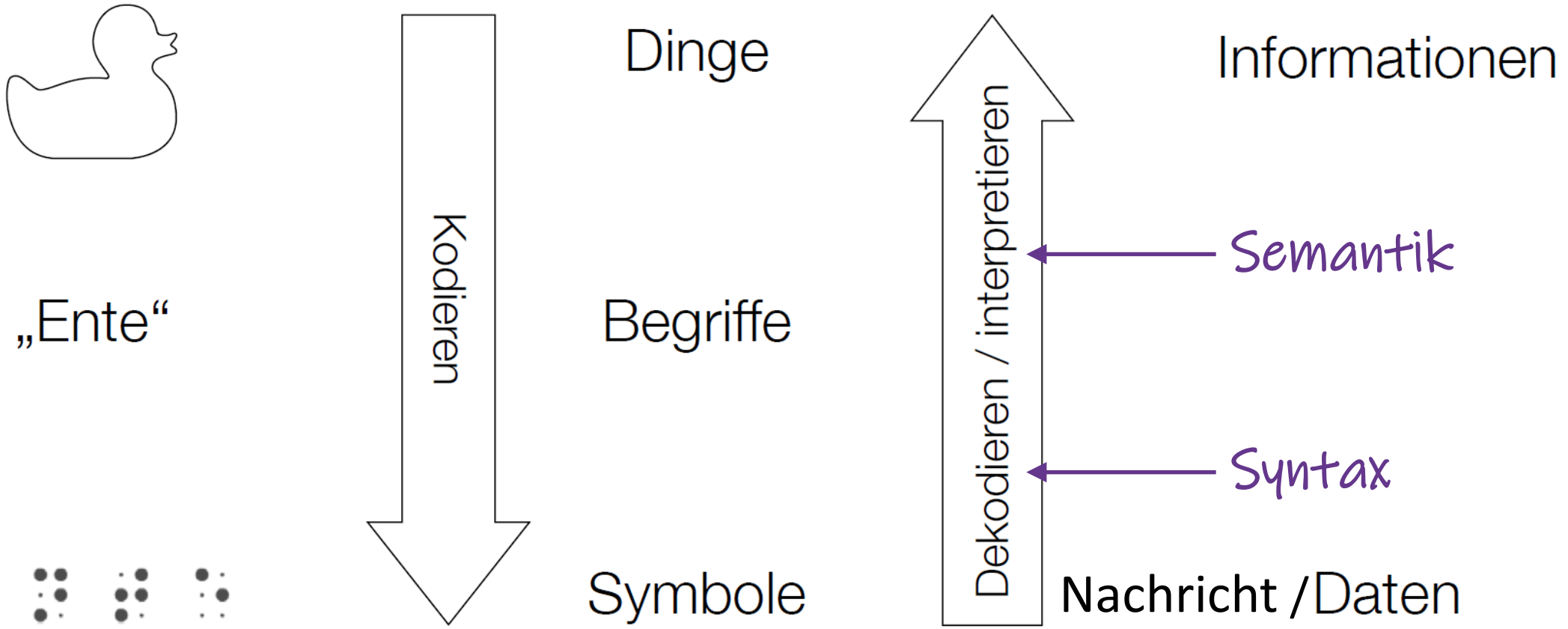
+ Um die Bedeutung von Daten (deren Information) zu erschließen, müssen wir mehr über die Daten wissen.
+ Die Extraktion von Information aus einer Nachricht setzt eine Zuordnung (Abbildung) zwischen Nachricht und Information voraus, die *Interpretation* genannt wird.
+ Die **Interpretation einer Nachricht ist** jedoch nicht unbedingt eindeutig, sondern **subjektiv**.
+ In noch stärkerem Maße gilt das für die Bedeutung, die eine Nachricht tragen kann. Ein und dieselbe Nachricht kann bisweilen auf verschiedene Weisen interpretiert werden.

Beispiel

+ Wie stellt man die Postleitzahl eines Orts in Deutschland dar und wie kann man diese Darstellung (auf der Landkarte) interpretieren?
+ Darstellung: Folge von genau 5 Ziffern
+ Interpretation: Gleiche Anfänge bedeuten (bedingt) geographische Nähe

1.3 Nachricht und Interpretation

Kodieren und Interpretieren



1.3 Nachricht und Interpretation

Informationsgehalt

Claude Shannon (1916-2001) formalisierte erstmals in seiner Informationstheorie

+ Der Informationsgehalt eines Zeichens ist seine *statistische Signifikanz*.
+ Für die mathematische Beschreibung des statistischen Informationsgehalts $I(x)$ eines Zeichens oder Wortes x, das in einer Nachricht mit der Auftrittswahrscheinlichkeit $P(x)$ vorkommt, stellt man einige elementare Forderungen:

1. Je seltener ein bestimmtes Zeichen x auftritt, d. h. je kleiner $P(x)$ ist, desto größer soll der Informationsgehalt dieses Zeichens sein. $I(x)$ muss demnach zu einer Funktion, die von $1/P(x)$ abhängt, proportional sein und streng monoton wachsen.
2. Eine Zeichenkette x_1x_2 aus voneinander unabhängigen Zeichen x_1 und x_2 hat die Auftrittswahrscheinlichkeit $P(x_1x_2) = P(x_1) \cdot P(x_2)$. Die Gesamtinformation dieser Zeichenkette soll sich aus der Summe der Einzelinformationen ergeben, also:

$$I(x_1x_2) = I(x_1) + I(x_2)$$

   Dieses Additionsgesetz lässt sich auf beliebig lange Zeichenketten erweitern.
3. Für den Informationsgehalt eines mit Sicherheit auftretenden Zeichens x, also für den Fall $P(x) = 1$, soll $I(x) = 0$ gelten.

Vgl. [5] S. 60

1.3 Nachricht und Interpretation

Informationsgehalt

Claude Shannon (1916-2001)

+ Der Informationsgehalt einer Nachricht ist eine **logarithmische Größe**, die angibt, wie viel Information in dieser Nachricht übertragen wurde.
+ Für die Abhängigkeit des Informationsgehalts eines Zeichens x von seiner Auftrittswahrscheinlichkeit $P(x)$ schreibt man daher:

$$I(x) = log_b \frac{1}{P(x)} = -log_b P(x)$$

+ Da eine Wahrscheinlichkeit $P(x)$ immer zwischen 0 und 1 liegt, ist der Informationsgehalt stets positiv.
+ Die Basis b des Logarithmus bestimmt lediglich den Maßstab, mit dem man Informationen schließlich messen möchte. Zur Festlegung dieses Maßstabes geht man von dem einfachsten denkbaren Fall einer Nachricht aus, die nur aus einer Folge der beiden Zeichen 0 und 1 besteht, wobei die beiden Zeichen mit der gleichen Wahrscheinlichkeit $p_0 = p_1 = 0{,}5$ auftreten sollen.
+ Dem Informationsgehalt eines solchen Zeichens wird nun per Definitionem der Zahlenwert 1 mit der Maßeinheit **Bit** (von *binary digit*) zugeordnet. Daraus ergibt sich $log_b\left(\frac{1}{0{,}5}\right) = log_b(2) = 1$ und folglich durch Auflösung dieser Gleichung nach b die Basis $b = 2$.

Vgl. [5] S. 60

1.3 Nachricht und Interpretation

Informationsgehalt

Claude Shannon (1916-2001)

+ Man erhält also schließlich für den statistischen Informationsgehalt eines mit der Wahrscheinlichkeit$P(x)$ auftretenden Zeichens x den Zweierlogarithmus (Logarithmus Dualis, **ld**) aus der reziproken (durch Vertauschung von Zähler und Nenner) Auftrittswahrscheinlichkeit:

$$I(x) = ld\frac{1}{P(x)} = -ldP(x)\ [Bit]$$

+ Die binäre Darstellung von Nachrichten verdeutlicht auch, dass die Maßeinheit Bit eine sinnvolle Wahl ist, denn der - auf die nächstgrößere ganze Zahl gerundete - Informationsgehalt eines Zeichens ist gerade die **Anzahl der Stellen des Binärwortes**, das man für eine eindeutige binäre Darstellung des Zeichens verwenden muss.

1.3 Nachricht und Interpretation

Informationsgehalt

Berechnung des Informationsgehaltes

+ Je kleiner die Auftretenswahrscheinlichkeit eines Zeichens ist, desto höher ist sein Informationsgehalt.
+ Andersherum ist der Informationsgehalt eines Zeichens sehr gering, wenn es sehr oft vorkommt.

Die Berechnung des Informationsgehaltes lässt sich ohne weiteres auf nicht-binäre Nachrichten, etwa das lateinische Alphabet, übertragen, wie das folgende Beispiel zeigt.
In einem deutschsprachigen Text tritt der Buchstabe b mit der Wahrscheinlichkeit 0,016 auf. Wie groß ist der Informationsgehalt dieses Zeichens? Die Lösung dafür lautet:

$$I(\mathrm{b}) = \mathrm{ld}\,\frac{1}{0{,}016} = \frac{\log \frac{1}{0{,}016}}{\log 2} \approx \frac{1{,}79588}{0{,}30103} \approx 5{,}97\,\mathrm{Bit}.$$

Für die tatsächliche binäre Codierung müsste man also – notwendigerweise aufgerundet auf die nächst größere natürliche Zahl – die Stellenzahl 6 wählen.

Vgl. [5] S. 62

1.3 Nachricht und Interpretation

Auftrittswahrscheinlichkeiten von Buchstaben

Einzelne Buchstaben und Kombinationen von zwei Buchstaben in einem typischen deutschen Text

Tabelle 2.3 Wahrscheinlichkeiten für das Auftreten von Buchstaben in einem typischen deutschen Text. Zwischen Groß- und Kleinbuchstaben wird dabei nicht unterschieden.

Buchstabe x_i	$P(x_i)$	Buchstabe x_i	$P(x_i)$
andere Zeichen	0,1515	o	0,0177
e	0,1470	b	0,0160
n	0,0884	z	0,0142
r	0,0686	w	0,0142
i	0,0638	f	0,0136
s	0,0539	k	0,0096
t	0,0473	v	0,0074
d	0,0439	ü	0,0058
h	0,0436	p	0,0050
a	0,0433	ä	0,0048
u	0,0319	ö	0,0025
l	0,0293	j	0,0016
c	0,0267	y	0,0002
g	0,0267	q	0,0001
m	0,0213	x	0,0001

Tabelle 2.4 Auftrittswahrscheinlichkeiten für die 20 häufigsten Kombinationen von zwei Buchstaben in einem typischen deutschen Text. Zwischen Groß- und Kleinbuchstaben wird nicht unterschieden.

Gruppe g_i	$P(g_i)$	Gruppe g_i	$P(g_i)$
en	0,0447	ge	0,0168
er	0,0340	st	0,0124
ch	0,0280	ic	0,0119
nd	0,0258	he	0,0117
ei	0,0226	ne	0,0117
de	0,0214	se	0,0117
in	0,0204	ng	0,0107
es	0,0181	re	0,0107
te	0,0178	au	0,0104
ie	0,0176	di	0,0102
un	0,0173	be	0,0096

Vgl. [5] S. 65

1.3 Nachricht und Interpretation

Auftrittswahrscheinlichkeiten von Buchstaben

Korpusbasierte Zeichenhäufigkeitslisten

+ Institut für Deutsche Sprache (IDS), Mannheim
+ *DeReChar-v-uni-XXX-2018-02-28-1.0*
+ Häufigkeitsverteilung der verschiedenen Zeichen im Sprachgebrauch, insb. der Buchstaben des dt. Alphabets
+ Verschiedene Auswertungen der Sammlung authentischer Texte, des Deutschen Referenzkorpus DeReKo.

Beispiel

+ Zeichenhäufigkeitsliste *DeReChar-v-uni-030-b-l-2018-02-28-1.0*
  - nur deutsches Alphabet
  - Groß-/Kleinschreibung ignorieren
  - ohne „andere Zeichen“

RF▼	AF	G	N	B
0.160061266048	23552726179	e	LATIN LETTER E	Basic Latin
0.096608077356	14215704079	n	LATIN LETTER N	Basic Latin
0.077526004178	11407811479	i	LATIN LETTER I	Basic Latin
0.077377610953	11385975684	r	LATIN LETTER R	Basic Latin
0.063693071514	9372320424	t	LATIN LETTER T	Basic Latin
0.063439603408	9335023050	s	LATIN LETTER S	Basic Latin
0.060067477706	8838820844	a	LATIN LETTER A	Basic Latin
0.047182137047	6942766241	d	LATIN LETTER D	Basic Latin
0.042497859499	6253483261	h	LATIN LETTER H	Basic Latin
0.038209944546	5622524321	u	LATIN LETTER U	Basic Latin
0.037871919697	5572784575	l	LATIN LETTER L	Basic Latin
0.030642574549	4508999495	g	LATIN LETTER G	Basic Latin
0.027983071757	4117658463	m	LATIN LETTER M	Basic Latin
0.026900980520	3958430692	c	LATIN LETTER C	Basic Latin
0.026841202334	3949634440	o	LATIN LETTER O	Basic Latin
0.021480751076	3160853720	b	LATIN LETTER B	Basic Latin
0.018323709896	2696300813	f	LATIN LETTER F	Basic Latin
0.015368289907	2261416100	k	LATIN LETTER K	Basic Latin
0.014275418694	2100602075	w	LATIN LETTER W	Basic Latin
0.012376156182	1821129027	z	LATIN LETTER Z	Basic Latin
0.010497044240	1544621099	p	LATIN LETTER P	Basic Latin
0.009188013866	1351999644	v	LATIN LETTER V	Basic Latin
0.006835955398	1005898489	ü	LATIN LETTER U WITH DIAERESIS	Latin-1 Supplement
0.005489485951	807767942	ä	LATIN LETTER A WITH DIAERESIS	Latin-1 Supplement
0.002978121104	438225141	j	LATIN LETTER J	Basic Latin
0.002698198599	397035050	ö	LATIN LETTER O WITH DIAERESIS	Latin-1 Supplement
0.001706185925	251062177	ß	LATIN LETTER SHARP S	Latin-1 Supplement
0.001079873609	158901451	y	LATIN LETTER Y	Basic Latin
0.000517091308	76089052	x	LATIN LETTER X	Basic Latin
0.000282903133	41628685	q	LATIN LETTER Q	Basic Latin

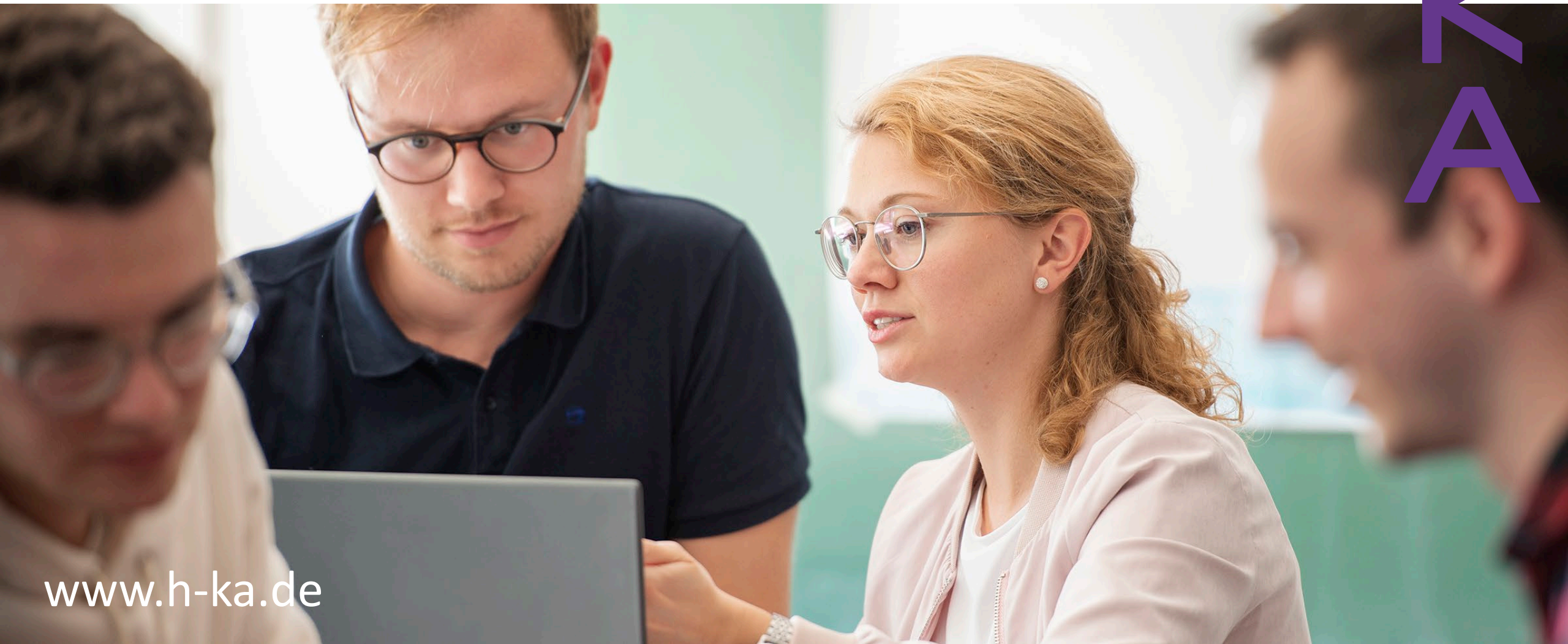
Anzahl gezählter Zeichen: 147.148.193.692

Buchstabenhäufigkeit, Zeichenhäufigkeit, German character frequency, German letter frequency

Vgl. Zeichenhäufigkeitsliste DeReChar-v-uni-030-b-l-2018-02-28-1.0

Hochschule Karlsruhe
University of
Applied Sciences

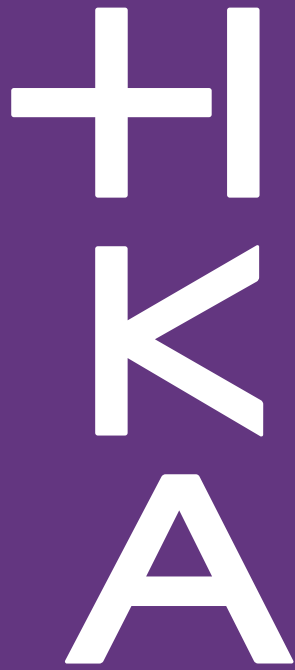
Fakultät für
Informatik und
Wirtschaftsinformatik



www.h-ka.de

Hochschule Karlsruhe
University of
Applied Sciences

Fakultät für
Informatik und
Wirtschaftsinformatik



www.h-ka.de