Einführung

Proseminar Datenkompression

Sommersemester 2020

PD Dr. Ralf Schlüter

Lehrstuhl Informatik 6
RWTH Aachen
52056 Aachen
mailto:schlueter@cs.rwth-aachen.de

1 Überblick

Einführung	• **Anwendungsbereiche** • **Motivation** • **Beispiele**
Methodik	• **Verlustlos vs. verlustbehaftet** • **Performanzbewertung** • **Ansätze** • **Vorgehensweise**
Statistik	• **Grundbegriffe und Konzepte** • **Stochastische Prozesse**
Information	• **Definition** • **Modellierung** • **Kodierung**

2 Einführung

Anwendungen für Datenkompression:

- Internet
- Telekommunikation
- Videokommunikation

Undenkbar ohne Datenkompression:

- Bild-, Audio- und Videoinformation im Internet
- Hochqualitative Mobiltelephonie
- Digital-TV

Einführung

Warum Datenkompression?

- Ergänzung zu Verbesserungen in Speichertechnologie und Datenübertragung
- Informationsbedarf steigt stärker als verfügbare Resourcen
- Physikalische Grenzen für Speicher- und Übertragungskapazitäten

Jedoch:

- Komprimierbarkeit ist auch begrenzt (Entropie!)

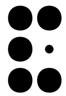
Einführung

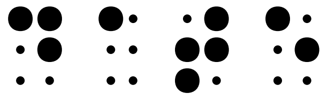
Morse Kodierung:

- kürzere Repräsentationen für häufigere Zeichen
- z.B.:

SOS	... — ...	(9 bit)
AND	.- -. -..	(7 bit)
DATE	-.. .- - .	(7 bit)

Braille Kodierung:

- 6 bit pro Zeichen sowie häufigste Wörter
- z.B.: AND  (6 bit)

  DATE  (24 bit)

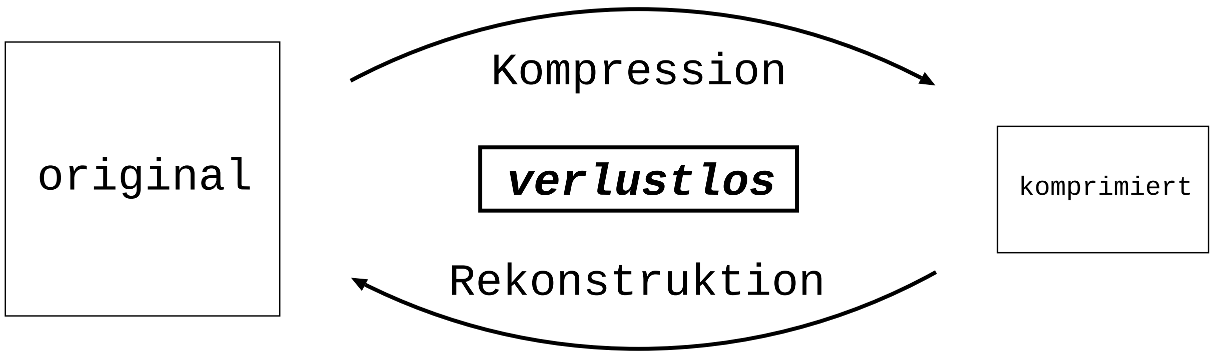
- Anwendung: Blindenschrift

3 Methodik

Verlustlose Kompression:

notwendig für z.B.:

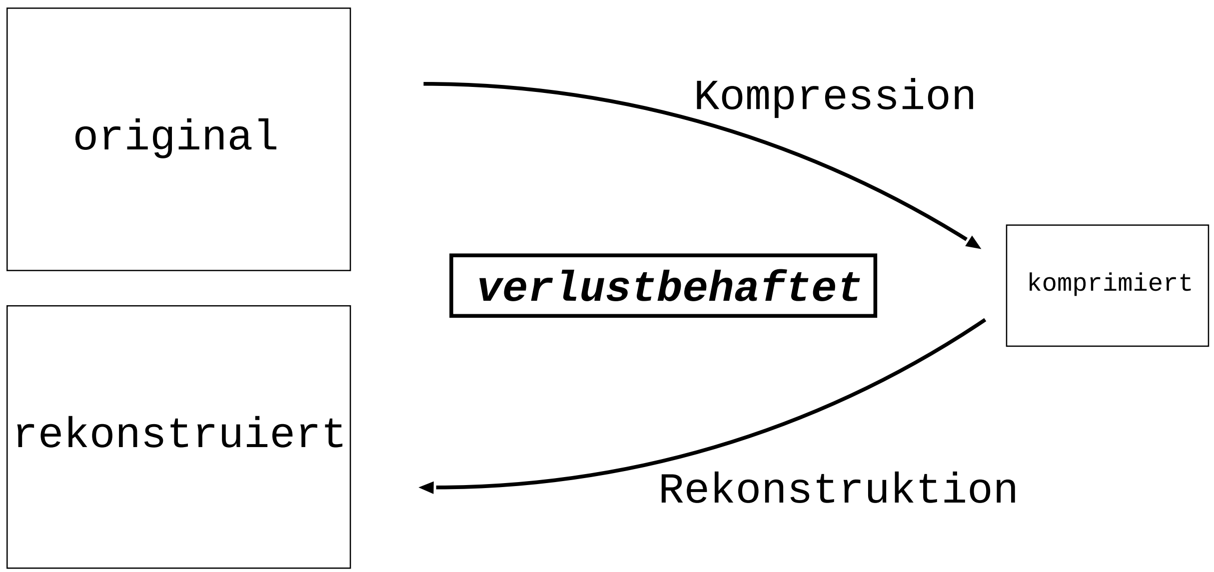
- **Text-Daten ("Do not jump!" vs. "Do now jump!")**
- **System-Daten**
- **Bank-Daten**
- **Verhinderung von Artefakten bei Weiterverarbeitung**



Methodik

Verlustbehaftete Kompression:

- **höhere Kompressionsraten möglich**
- **Redundanz der Originaldaten**
- **Informationsgehalt vs. Perzeptionsgrenzen bzw. Akzeptanz**



Methodik

Performanzbewertung

- **Maß der Kompression:**
  - **Kompressionsrate:** $\dfrac{\textbf{Anzahl Bits im Original}}{\textbf{Anzahl Bits nach Kompression}}$
  - **Normiert: z.B. für Bilder in Bits pro Pixel**

- **Verlustbehaftet: zusätzlich Qualitätsmaße**
  - **Verzerrung: Ähnlichkeit zum Original**
  - **Sprache, Video: menschliche Perzeption**
  - **Mathematische Modellierung der menschlichen Perzeption**

Methodik

Ansätze für Kompression:

- **Statistik, z.B.:**
  - **Häufigkeit einzelner Symbole (z.B. Huffman, arithmetisch)**
  - **Häufigkeit von Symbolfolgen/Wörtern (z.B. string-basiert)**
  - **Berücksichtigung von Kontext (z.B. prädiktiv)**

- **Physikalische Strukturierung, z.B.:**
  - **Sprache: Vokaltraktparameter statt Abtastwerte**

- **Wahrnehmungs-Orientierung, z.B.:**
  - **Sprache: Abtastrate angepasst an Verständlichkeit**
  - **Bilder: Auflösung angepasst an Perzeptionsgrenzen**
  - **Film: Bildrate angepasst an Fähigkeit, aufeinanderfolgende Bilder explizit zu unterscheiden**

Methodik

Wahl der Kompressionsmethode:

- Finden von Redundanzen
- Modellierung (von Redundanzen), z.B.:
  - Statistisch
  - Gruppierung
  - Prädiktion
  - Funktional
  - Transformation
- Kodierung, z.B.:
  - Statistisch, z.B.:
    - variable Kodewortlänge
    - Gruppierung
  - Modellparameter
    + Abweichung vom Modell (Residuum)

4 Statistische Grundlagen

Wahrscheinlichkeit:

- Beschreibung von Ereignissen
- Erwartungsmaß; z.B. Häufigkeit
- Positivität, Normierung und Additivität
- Unabhängigkeit / Bedingtheit
- Bezeichnungen:
  - a-priori Wahrscheinlichkeit: $p(B)$
  - a-posteriori Wahrscheinlichkeit: $p(A|B)$
  - Verbundverteilung: $p(A,B) = p(A \cup B) = p(A|B) \cdot p(B)$
  - Randverteilung: $p(B) = \sum_A p(A,B)$

Statistische Grundlagen

Ereignisse:

- mögliche Werte einer Zufallsvariablen, z.B.:
  - Ergebnisse eines Würfelwurfs
  - Zeichen, Wörter oder ganze Sätze
  - Orte eines Meteoriteneinschlags
- diskret oder kontinuierlich
- Gruppierung, z.B. $\{a, b\} \rightarrow \{aa, ab, ba, bb\}$

Statistische Grundlagen

Wichtige Begriffe und Konzepte:

- **Bayessche Identität:** $$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A \cup B)}{p(B)}$$

- **kumulative Verteilung:** $$p(x \le x_0) = \int_{-\infty}^{x_0} p(x)dx$$

- **Erwartungswerte:** $$E\{f(x)\} = \int_{-\infty}^{\infty} f(x)p(x)dx \quad \textbf{(kontinuierlich)}$$

  $$E\{f(x)\} = \sum_i f(x_i)p(x_i) \quad \textbf{(diskret)}$$

- **Mittelwert:** $$\mu = E\{x\}$$

- **Varianz:** $$\sigma^2 = E\{(x-\mu)^2\} = E\{x^2\} - E\{x\}^2$$

Statistische Grundlagen

Stochastische Prozesse:

- statistische Modellierung von Zeitreihen, z.B.:
  - Niederschlagsmenge
  - Stromverbrauch
  - Radioaktiver Zerfall
  - Sprache
  - Video-Sequenzen
- zeitabhängige Zufallsvariable
- Autokorrelation:

$$R_{xx}(t_1, t_2) = E\{x(t_1) \cdot x(t_2)\}$$

- Stationarität: statt expliziter nur noch relative Zeitabhängigkeit

$$R_{xx}(t_1, t_2) = R_{xx}(t_2 - t_1)$$

5 Informationstheorie

Was ist eigentlich Information?

- **Datenmenge:**
  - n **verschiedene Zeichen** z**:** $i(z) = \log_2 n$ **(Bits pro Zeichen)**
  - **Gleichverteilung:** $p(z) = \frac{1}{n} \quad \Rightarrow \quad i(z) = -\log_2 p(z)$
  - **Allgemeine Verteilung:** $i(z) = -\log_2 p(z)$
  - **C. E. Shannon: “Eigeninformation”**
  - **Einheit:**
    - **bestimmt durch Basis des Logarithmus**
    - **z.B. Anzahl Bits, Zeichen, Wörter, Seiten, etc.**
- **Informationsgehalt:**
  - **Bezug zu Datenmenge?**
  - **Intuitiv: minimal mögliche Datenmenge ohne Verluste**
  - **Vorsicht: In diesem Sinne enthält eine zufällige Zeichenfolge mehr Information als z.B. eine Seminararbeit gleicher Länge!**

5 Informationstheorie

Was ist eigentlich Information?

- **Datenmenge:**
  - n **verschiedene Zeichen** z**:** $i(z) = \log_2 n$ **(Bits pro Zeichen)**
  - **Gleichverteilung:** $p(z) = \frac{1}{n} \quad \Rightarrow \quad i(z) = -\log_2 p(z)$
  - **Allgemeine Verteilung:** $i(z) = -\log_2 p(z)$
  - **C. E. Shannon: "Eigeninformation"**
  - **Einheit:**
    - **bestimmt durch Basis des Logarithmus**
    - **z.B. Anzahl Bits, Zeichen, Wörter, Seiten, etc.**
- **Informationsgehalt:**
  - **Bezug zu Datenmenge?**
  - **Intuitiv: minimal mögliche Datenmenge ohne Verluste**
  - **Vorsicht: In diesem Sinne enthält eine zufällige Zeichenfolge mehr Information als z.B. eine Seminararbeit gleicher Länge!**

  ... d.h.: Quantität ist nicht gleich Qualität!

Informationstheorie

Information von Ereignissen

Ereignisse A und B seien unabhängig

- **dann gilt $p(A, B) = p(A) \cdot p(B)$**
- **Information des Verbundereignisses $A \cup B$:** $\quad i(A, B) = i(A) + i(B)$
- **kein Informationsgewinn durch Gruppierung**
- **Beispiele: Münzwurf**
  - **gleichverteilt:** $\quad p(Kopf) = p(Zahl) = 1/2$
    $\Rightarrow \quad i(Kopf) = i(Zahl) = 1\ Bit.$
  - **nicht gleichverteilt:** $\quad p(Kopf) = 7/8, \qquad p(Zahl) = 1/8$
    $\Rightarrow \quad i(Kopf) = 0.193\ Bits, \quad i(Zahl) = 3\ Bits.$

Informationstheorie

Information von Ereignissen

Ereignisse A und B seien abhängig

- **dann gilt $p(A, B) = p(A|B) \cdot p(B)$**
- **Information des Verbundereignisses $A \cup B$:** $\quad i(A, B) = i(A|B) + i(B)$
- **Informationsgewinn durch Gruppierung möglich, z.B.:**
  - **Betrachte Ziffernfolge:** $\quad f = 1\,2\,1\,2\,3\,3\,3\,3\,1\,2\,3\,3\,3\,3\,1\,2\,3\,3\,1\,2$
  - **Gleiche Kodewortlänge für alle Symbole: $20 \cdot 2$ Bits $= 40$ Bits**
  - **Annahme unabhängiger Einzelsymbole:**
    $p(1) = p(2) = 1/4, p(3) = 1/2 \;\Rightarrow\; i(f) = 5 \cdot 2 + 5 \cdot 2 + 10 \cdot 1$ **Bits** $= 30$ **Bits**
  - **Abhängigkeit durch Gruppierung zu Blöcken 12 und 33:**
    $p(1\,2) = p(3\,3) = 1/2 \qquad \Rightarrow\; i(f) = 5 \cdot 1 + 5 \cdot 1$ **Bits** $= 10$ **Bits**

Informationstheorie

Mittlere Information:

- betrachte Zufallsprozess $\mathcal{Z}$, mögliche Ereignisse: A_i
- Ereignis A_i tritt ein mit Wahrscheinlichkeit $p(A_i)$
- mittlere Information eines Ereignisses dieses Zufallsprozesses: Erwartungswert des Informationsgehalts

$$
\begin{aligned}
H(\mathcal{Z}) = E\{i(A)\} &= \sum_i p(A_i) \cdot i(A_i) \\
&= -\sum_i p(A_i) \cdot \log_2 p(A_i)
\end{aligned}
$$

> C. E. Shannon: Entropie H gibt minimale Anzahl von Bits zur verlustlosen Kodierung des Zufallsprozesses an

- Qualität einer Kompressionsmethode: Vergleich mit Entropie
- **Vorsicht:** Entropie abhängig vom Modell (bzgl. Kontextabhängigkeit)!

Informationstheorie

Mittlere Information im allgemeinen Fall:

- betrachte stochastischen Prozess $\mathcal{S}$, der Folge von Ereignissen A_i erzeugt
- Ereignisse seien aus Alphabet $\{A_1, ..., A_m\}$
- Entropie:

$$H(\mathcal{S}) = -\lim_{n\to\infty} \frac{1}{n} \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \cdots \sum_{i_n=1}^{m} p(A_{i_1}, \ldots, A_{i_n}) \cdot \log_2 p(A_{i_1}, \ldots, A_{i_n})$$

- Reichweite von Korrelationen/Redundanzen unbekannt:
  Betrachtung im Limes unendlich langer Folgen
- Verteilung bzw. Strukturierung der Daten
  im Allgemeinen nicht (exakt) bekannt
- Notwendigkeit der Modellierung

Informationstheorie

Herleitung der Entropie

- **Informationstheoretische Basis: C. E. Shannon**
- **Herleitung der mittleren Information allein über Axiome**
- **Betrachte unabhängige Ereignisse A_i mit Wahrscheinlichkeiten $p_i = p(A_i)$**
- **Axiome:**

1. **Mittlere Information H ist stetige Funktion der Wahrscheinlichkeiten p_i; kleine Änderungen in den Wahrscheinlichkeiten führen zu kleinen Änderungen in der mittleren Information.**
2. **Für gleichverteilte Ereignisse mit $p_i = 1/n$ ist mittlere Information eine monotone Funktion von n, der Anzahl der möglichen Ereignisse.**
3. **Konsistenz der mittleren Information unter Gruppierung. Betrachte $A_2 \vee A_3$ als neues Ereignis:**

$$H(p_1, p_2, p_3) = H(p_1, p_2 + p_3) + p_1 \cdot H(\frac{p_1}{p_1} = 1) + (p_2 + p_3) \cdot H(\frac{p_2}{p_2 + p_3}, \frac{p_3}{p_2 + p_3})$$

Informationstheorie

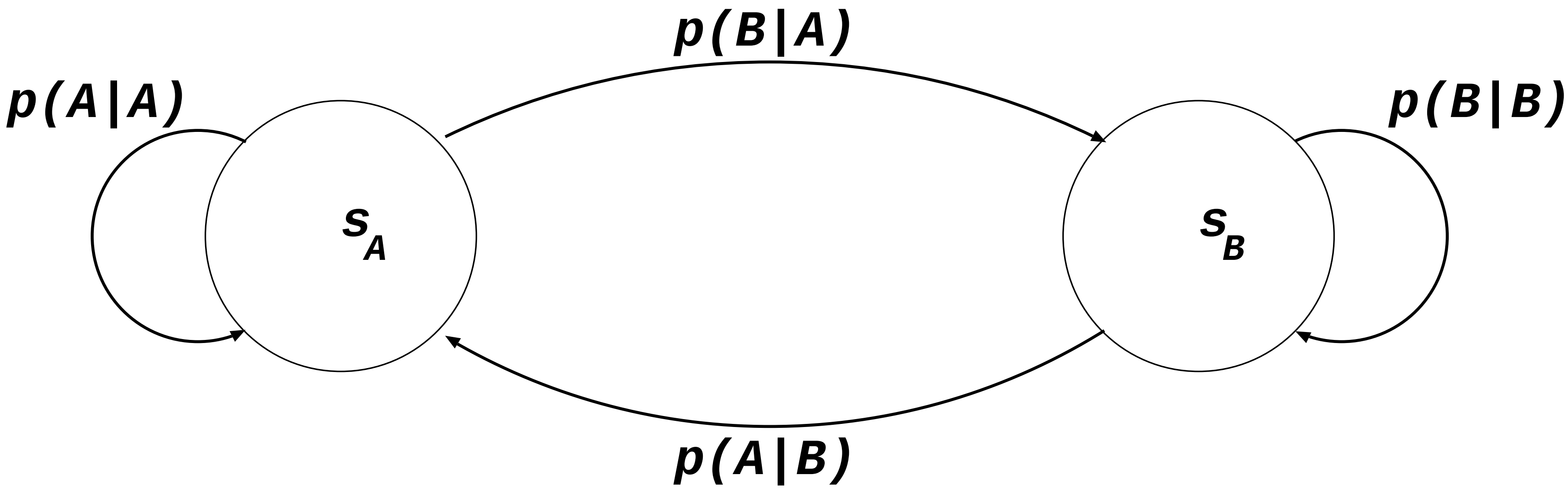
Modellierung:

- **Physikalisch**
  - **Wissen über die Strukturierung der Quelle**
  - **Vorhersage von Werten mittels Modell**
  - **Kodierung des Residuums: Abweichung vom Modell**
- **Statistisch**
  - **Steuerung von Kodewortlänge, Gruppierung, etc. anhand der Wahrscheinlichkeiten**
  - **Beispiele:**
    * **Unabhängigkeit:** $$p(A_1, \ldots, A_n) = \prod_{i=1}^{n} p(A_i)$$
    * **Markov Annahme: Abhängigkeit endlicher Reichweite** m**,**

$$\begin{aligned} p(A_1, \ldots, A_n) &= \prod_{i=1}^{n} p(A_i | A_1, \ldots, A_{i-1}) \\ &= \prod_{i=1}^{n} p(A_i | A_{i-m}, \ldots, A_{i-1}) \end{aligned}$$

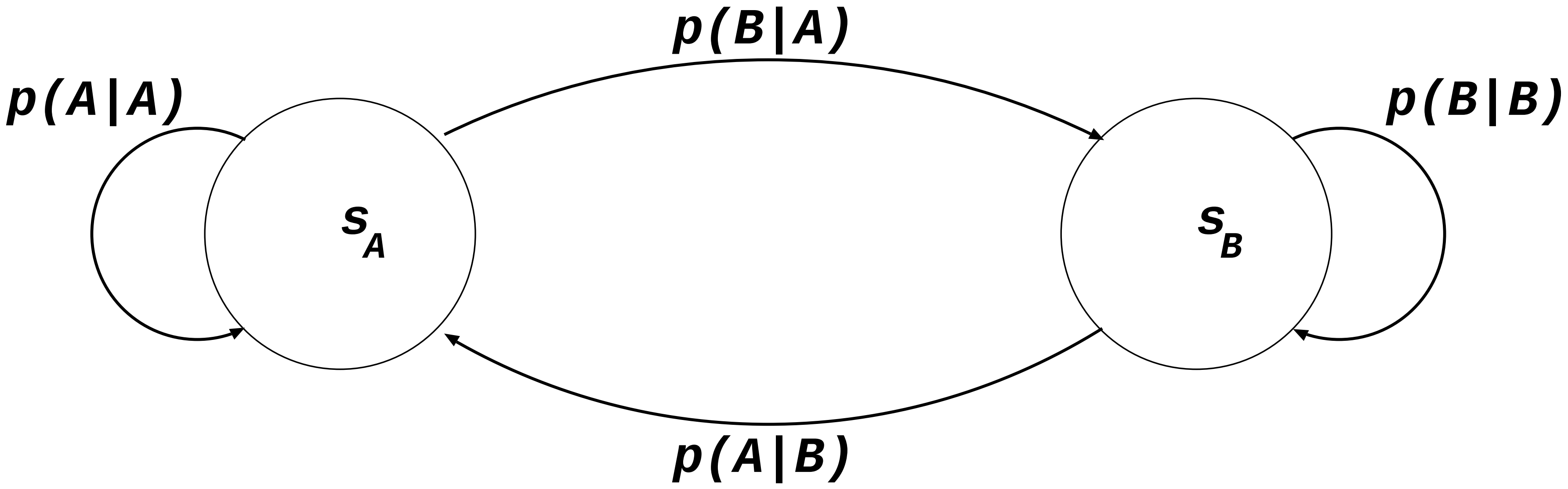
Markov-Prozesse:

- **Einfachster nicht-trivialer Fall:**
  Abhängigkeit allein vom direkt vorhergehenden Ereignis
- $p(A_i|A_1, \ldots, A_{i-1}) = p(A_i|A_{i-1})$
- **Vgl. stochastischen endlichen Automaten, z.B.:**



Markov-Prozesse:

- **Vgl. stochastischen endlichen Automaten, z.B.:**



  - **Binärer Zufallsprozess, Ereignisse $X \in \{A, B\}$**
  - **Zustände des Automaten: s_A, s_B**
  - **Zustand s_X "emittiert" Ereignis X**
  - **Übergangswahrscheinlichkeiten: $p(X_i|X_{i-1})$**
- **Bezüge: Sprachmodell in der Spracherkennung, Hidden Markov Modelle**

Informationstheorie

Kodierung

- Zuweisung binärer Folgen zu Elementen eines Alphabets
- Kode: Menge der binärer Folgen
- Zeichen: Element eines Alphabets
- Kodewörter: Elemente eines Kodes

- Problem: **Welche Kodewörter bzw. Kodewortlängen sind den Elementen des Alphabets zuzuordnen, um eine möglichst hohe Kompressionsrate auf den zu erwartenden Datensätzen zu erreichen?**

Informationstheorie

Eindeutige Dekodierbarkeit: Beispiele

Zeichen	Wahrscheinlichkeit	Kode 1	Kode 2	Kode 3	Kode 4
a_1	0.5	0	0	0	0
a_2	0.25	0	1	10	01
a_3	0.125	1	00	110	011
a_4	0.125	10	11	111	0111
mittlere Länge		1.125	1.25	1.75	1.875

- **mittlere Länge:** $\sum_i p(a_i)n(a_i)$, **mit Kodewortlänge** $n(a_i)$
- **Entropie:** 1.75
- **Kode 1: Identische Kodierung für a_1 und a_2: Kode 1 nicht eindeutig!**
- **Kode 2: Dekod. von 100 liefert a_2a_3 oder $a_2a_1a_1$: Kode 2 nicht eindeutig!**
- **Notwendig: Eindeutige Dekodierbarkeit**
- **Kode 3: Präfix-Kode – Ende des Kodeworts direkt erkennbar!**
- **Kode 4: ähnlich Kode 3, Kodewort-Ende aber erst zu Beginn des Folge-Kodeworts erkennbar!**

Eindeutige Dekodierbarkeit: Beispiele (2)

Zeichen	Wahrscheinlichkeit	Kode 5	Kode 6
a_1	0.5	0	0
a_2	0.25	01	01
a_3	0.125	11	10
...	...	...	...

- **Kode 5: Dekodierung erst am Ende eindeutig**
  - **z.B.: Dekodiere** $0\,\underbrace{1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1}_{\text{17 Einsen}}$
  - a_1 **mit 8 folgenden** a_3**: “baumelndes” Bit am Ende – nicht möglich!**
  - **Korrekt:** a_2 **gefolgt von 8** a_3
- **Kode 6: Dekodierung von** $0\,1\,0$ **liefert** $a_1 a_3$ **oder** $a_2 a_1$**: nicht eindeutig!**

Informationstheorie

Präfix-Kodes

- **Definition: Kein Kodewort darf Präfix eines anderen Kodeworts sein.**
- **Binärbaum-Darstellung:**
  - **Verzweigung: rechts entspricht 1, links 0**
  - **mögliche Kodewörter: Knoten**
  - **Präfix-Kode: Kodewörter nur an Blättern (Endknoten)**
- **Kraft-McMillan Ungleichung:** **Für jeden eindeutig dekodierbaren Kode gibt es einen entsprechenden Präfix-Kode mit gleichen Kodewortlängen.**

6 Zusammenfassung

Grundlagen

- Motivation
- Anwendungen
- Beispiele

Kompressionskonzepte

- Kompressionsmethoden
- Unterscheidung verlustlos/verlustbehaftet
- Statistische Betrachtungsweise, Grundkonzepte
- Bewertungsmöglichkeiten

Informationstheorie

- Grundkonzepte, Shannon
- Kodierung

7 Weiterer Ablauf

Ausführung
- Literatur
- Rechercheschulung/Info-Bibliothek
- Bearbeitungshinweise
- LaTeX, Vorlagen

Organisatorisches
- Ablauf/Fristen
- Probevortrag
- Vortrag
- Korrektur/Benotung

Themen
- Kurzvorstellung
- Wünsche
- Vergabe