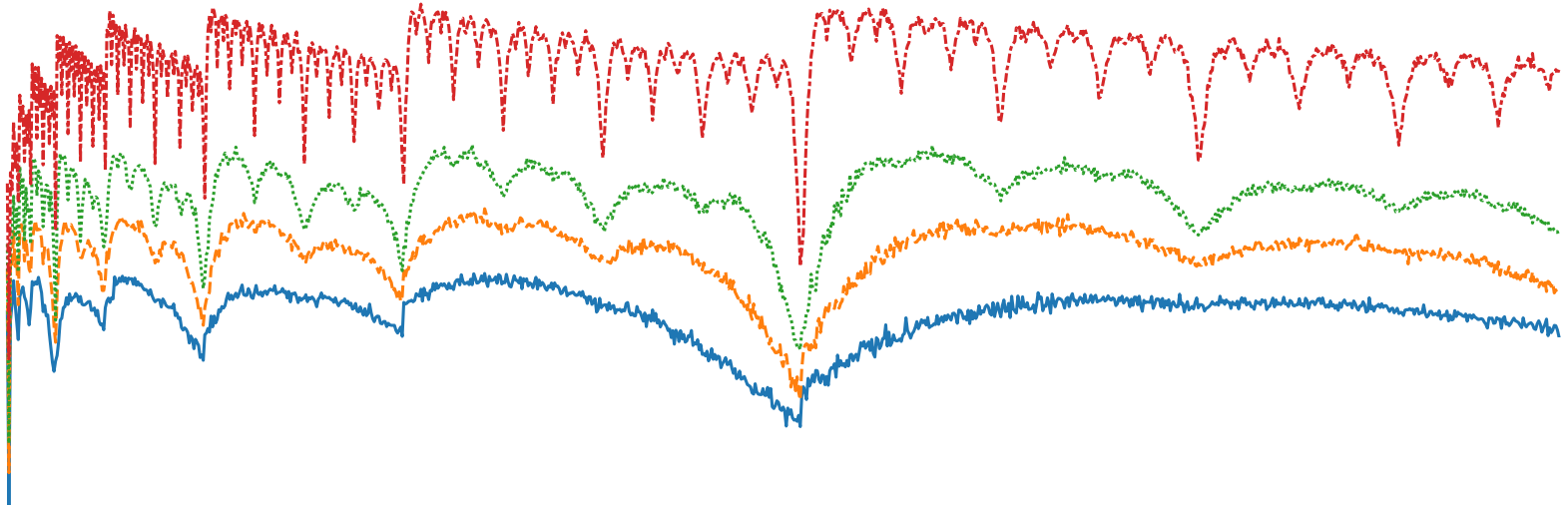
Karlsruher Institut für Technologie

Kompressionsmethoden für strukturierte Gleitkommazahlen und ihre Anwendung in den Klimawissenschaften

von Uğur Çayoğlu

STEINBUCH CENTRE FOR COMPUTING (SCC) und
INSTITUT FÜR METEOROLOGIE UND KLIMAFORSCHUNG (IMK-ASF)



Verlustfreie Kompression von Klimadaten

Karlsruher Institut für Technologie

Problem Hohes Datenaufkommen durch Klimasimulationen

ERA5 Datensatz für die Initialisierung und Validierung von Simulationsläufen umfasst 10.89 PiB

IMK-ASF Einer der größten Speicherplatzbenutzer am SCC mit >770 TiB (steigend)

Verlustfreie Kompression von Klimadaten

Karlsruher Institut für Technologie

Problem

Hohes Datenaufkommen durch Klimasimulationen (ERA5, 10.89 PiB)

Aktuelle Lösung

Reduzierung der zeitlichen Auflösung und gespeicherten Variablen

Folgen

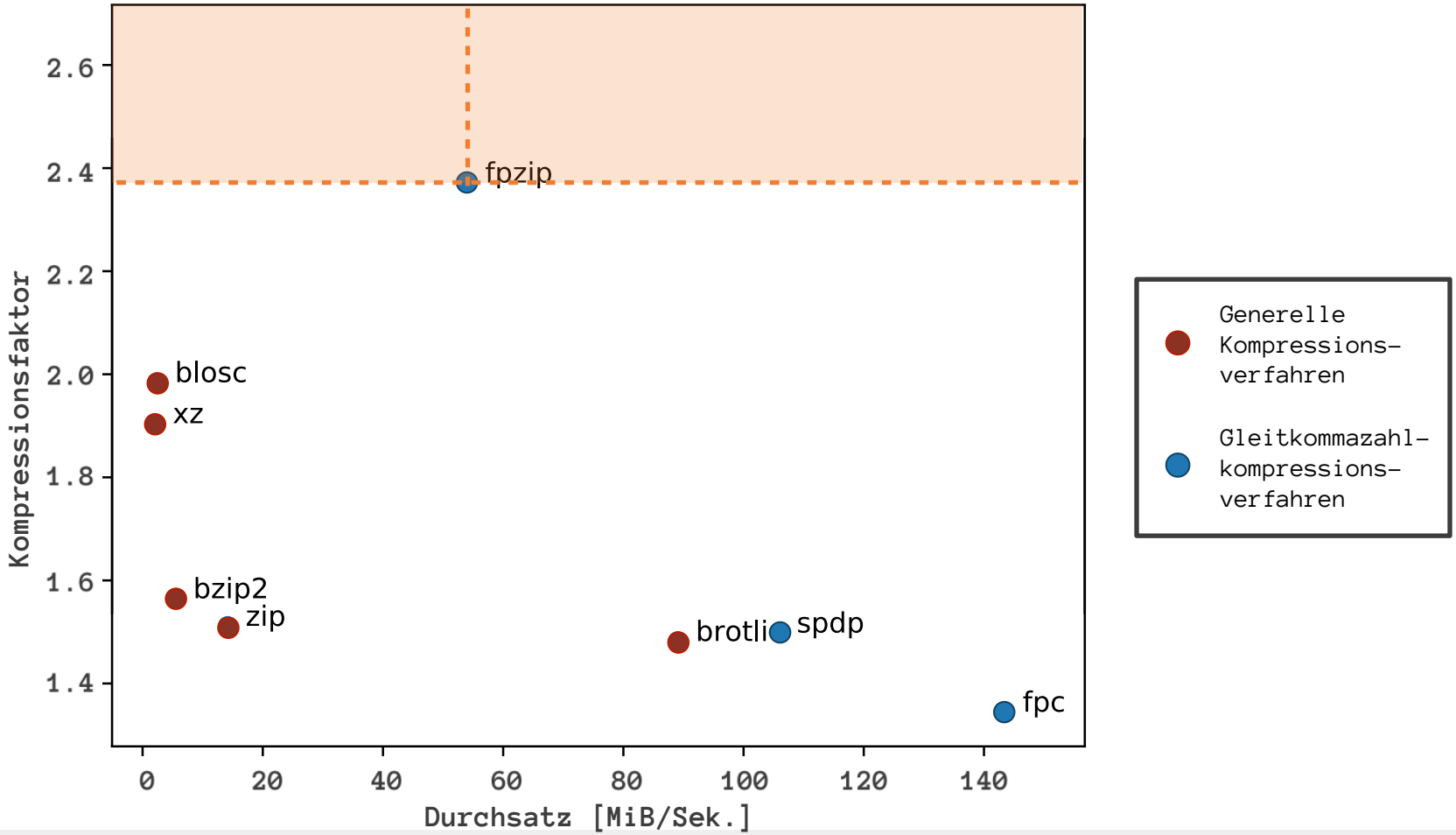
- Benutzung von Interpolationen
- Klimaereignisse (z.B. Entstehung von Stürmen) möglicherweise nicht abgebildet
- Neuberechnung von Simulationen

Ziel

Verlustfreies Kompressionsverfahren mit hohem Kompressionsfaktor

Kompressionsfaktor und Durchsatz

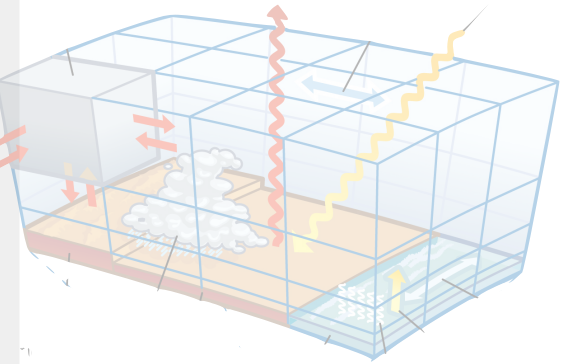
$$\text{Kompressionsfaktor} = \frac{\text{Ursprüngliche Dateigröße [Bytes]}}{\text{Komprimierte Dateigröße [Bytes]}} \qquad \text{Durchsatz} = \frac{\text{Ursprüngliche Dateigröße [Bytes]}}{\text{Kompressionszeit [Sek.]}}$$

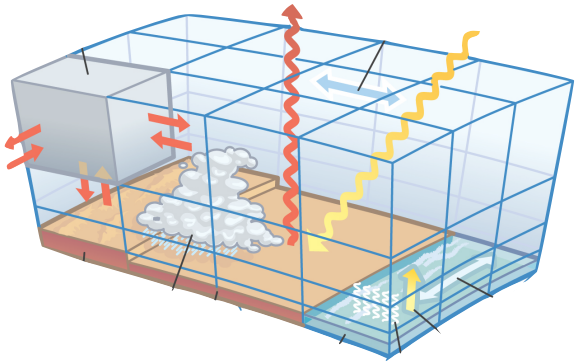


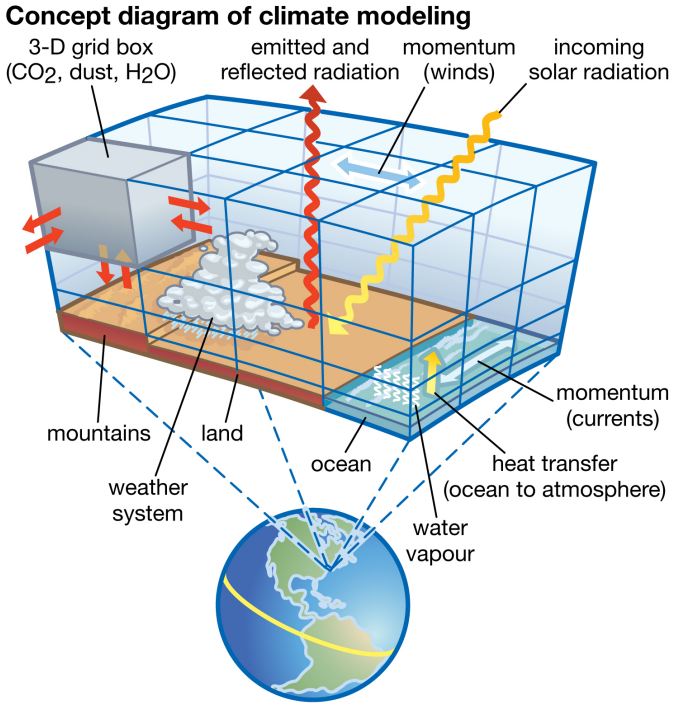
Verlustfreie Kompression von strukturierten Gleitkommazahlen

Karlsruher Institut für Technologie

Klimadaten 4D Daten (Längen- u. Breitengrad, Höhe, Zeit)

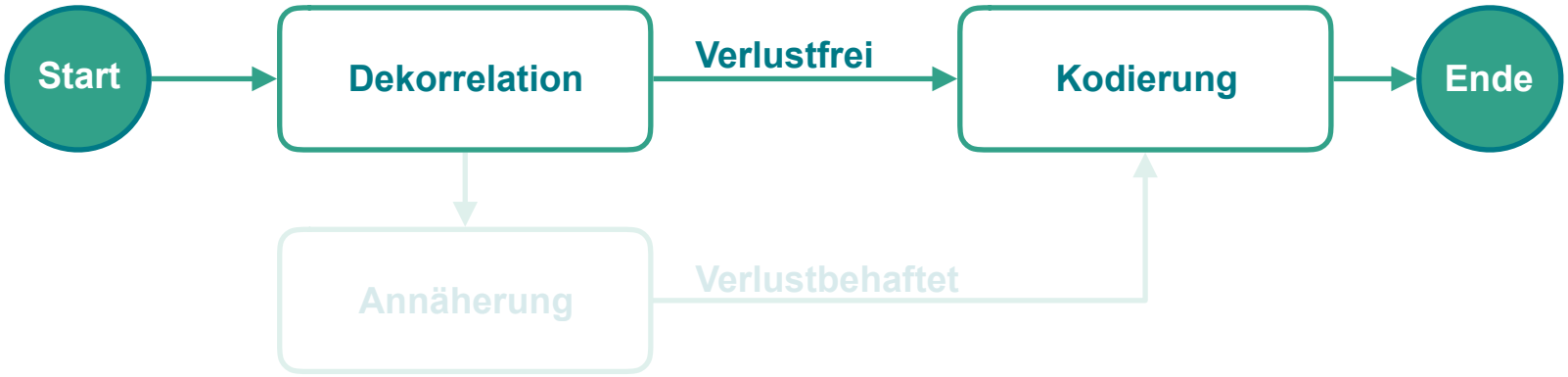






Quelle[1]

Verlustfreie Kompression von strukturierten Gleitkommazahlen



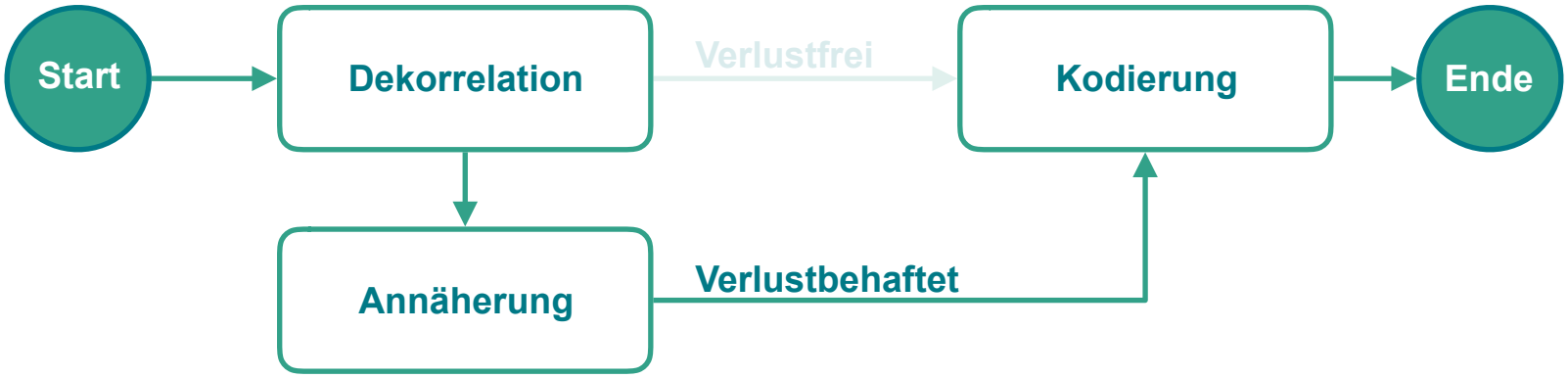
Dekorrelation: Entfernung von redundanter Information

Kodierung: Darstellung von Informationen in kompakter Form

Annäherung: Zusammenfassung oder Entfernung von unwichtigen Informationen

Quelle[2]

Verlustfreie Kompression von strukturierten Gleitkommazahlen



Dekorrelation: Entfernung von redundanter Information

Kodierung: Darstellung von Informationen in kompakter Form

Annäherung: Zusammenfassung oder Entfernung von unwichtigen Informationen

Quelle[2]

Überblick zum Forschungsfeld Datenkompression

Karlsruher Institut für Technologie

Kompressionsarten

- Verlustbehaftete
- Verlustfreie

Datentypen

- Integer
- Gleitkommazahlen
- Bytes

Anwendungsgebiete

- Audio
- Video
- Bild
- Zeitreihen
- Strukturierte Daten
- Text

Kompressionsverfahren

- Kodierung mit variabler Länge
- Lauflängenkodierung
- Wörterbuch-Verfahren
- Transformation
- Vorhersagebasierte Verfahren
- Entropie-basierte Verfahren

Überblick zum Forschungsfeld Datenkompression

Karlsruher Institut für Technologie

JPEG

Kompressionsarten

- **Verlustbehaftete**
- Verlustfreie

Datentypen

- **Integer**
- Gleitkommazahlen
- Bytes

Anwendungsgebiete

- Audio
- Video
- **Bild**
- Zeitreihen
- Strukturierte Daten
- Text

Kompressionsverfahren

- Kodierung mit variabler Länge
- **Lauflängenkodierung**
- Wörterbuch-Verfahren
- **Transformation**
- Vorhersagebasierte Verfahren
- **Entropie-basierte Verfahren**

Überblick zum Forschungsfeld Datenkompression

Karlsruher Institut für Technologie

Kompressionsarten

- Verlustbehaftete
- Verlustfreie

Datentypen

- Integer
- Gleitkommazahlen
- Bytes

Anwendungsgebiete

- Audio
- Video
- Bild
- Zeitreihen
- Strukturierte Daten
- Text

Kompressionsverfahren

- Kodierung mit variabler Länge
- Lauflängenkodierung
- Wörterbuch-Verfahren
- Transformation
- Vorhersagebasierte Verfahren
- Entropie-basierte Verfahren

Beiträge zur Informatik

Karlsruher Institut für Technologie

pzip

Kompressionsarten

- Verlustbehaftete
- Verlustfreie

Datentypen

- Integer
- Gleitkommazahlen
- Bytes

Anwendungsgebiete

- Audio
- Video
- Bild
- Zeitreihen
- Strukturierte Daten
- Text

Kompressionsverfahren

- Kodierung mit variabler Länge
- Lauflängenkodierung
- Wörterbuch-Verfahren
- Transformation
- Vorhersagebasierte Verfahren
- Entropie-basierte Verfahren

Direkter Beitrag | Einflussbereich

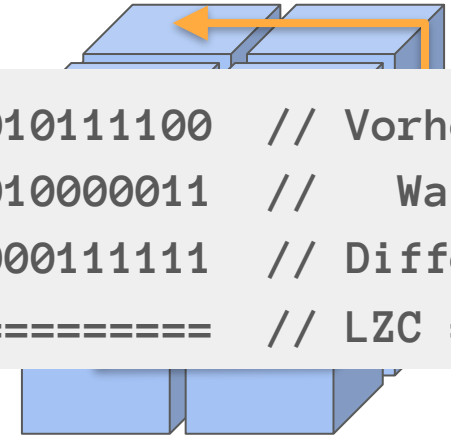
Pascal Zip (pzip)

Karlsruher Institut für Technologie

Vorhersagebasiertes Kompressionsverfahren

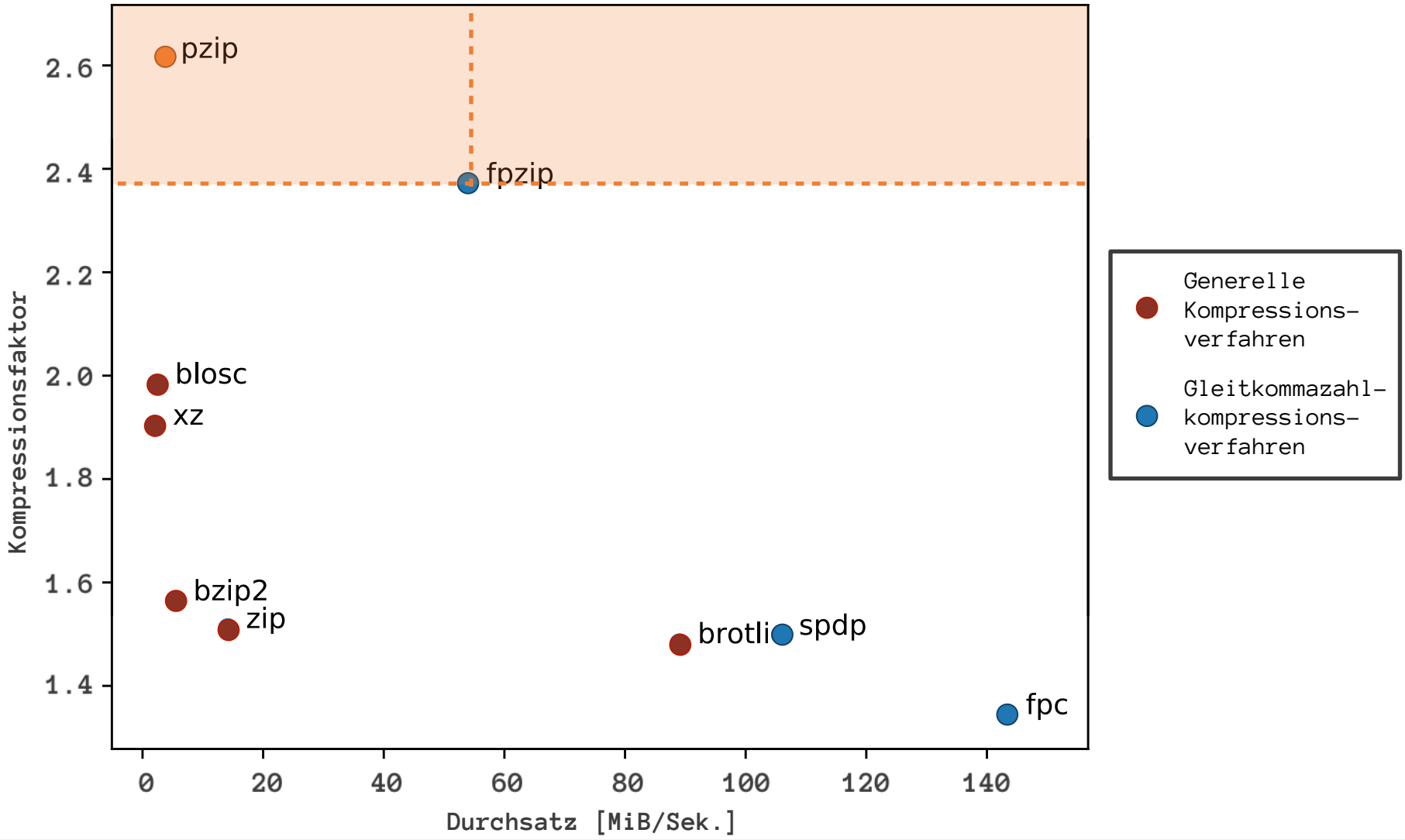
Methode

Für jeden einzelnen Datenpunkt wird (basierend auf vorhergehenden Werten) eine Vorhersage gegeben und die Differenz zum wahren Wert (Residuum) gespeichert

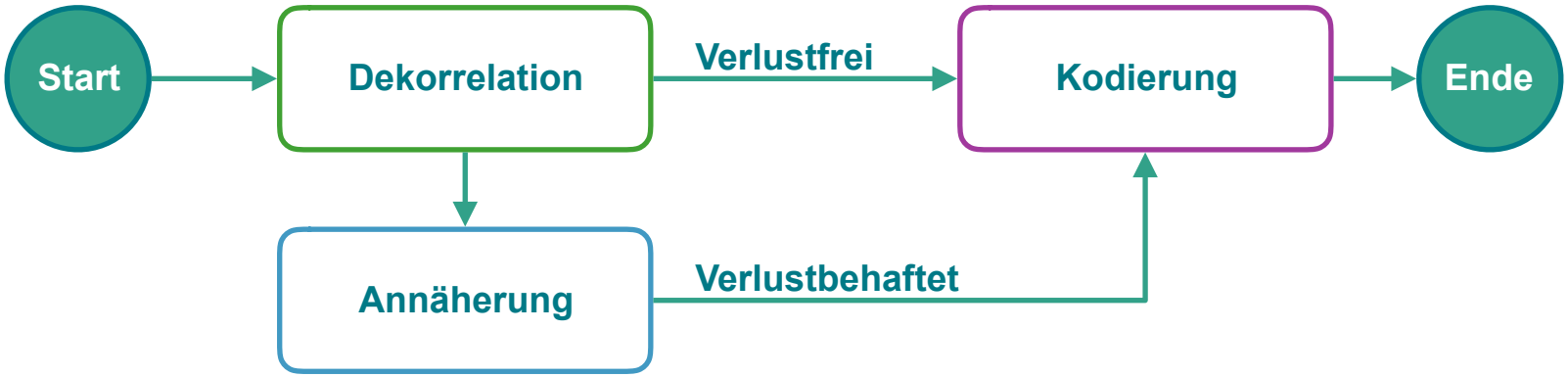


```
01000010100001111011010010111100  // Vorhersage = 67.853
01000010100100101000000010000011  //   Wahrheit = 73.251
00000000000101010011010000111111  // Differenz
          *====================  // LZC = 11 -> 32 - 11 - 1 = 20
```

Kompressionsverfahren im Vergleich



Publikationen und Konferenzbeiträge



Cayoglu et al. (2018a). Towards an optimised environmental data compression method for structured model output
EGU 2018

Cayoglu et al. (2019). On Advancement of Information Spaces to Improve Prediction-Based Compression
GI INFORMATIK 2019

Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression
IEEE Big Data 2018

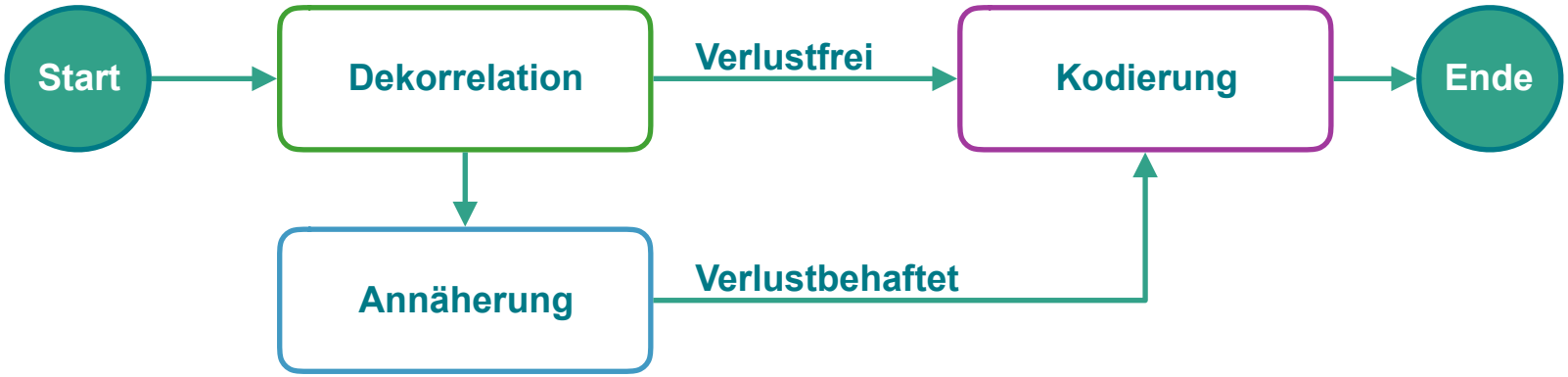
Cayoglu et al. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms
IEEE eScience 2019

Cayoglu et al. (2018b). A Modular Software Framework for Compression of Structured Climate Data
ACM SIGSPATIAL 2018

Cayoglu et al. (2017). Adaptive Lossy Compression of Complex Environmental Indices Using Seasonal Auto-Regressive Integrated Moving Average Models
IEEE eScience 2017

Kerzenmacher et al. (2017). QBO influence on the ozone distribution in the extra-tropical stratosphere
EGU 2018

Publikationen und Konferenzbeiträge



Cayoglu et al. (2018a). Towards an optimised environmental data compression method for structured model output
EGU 2018

Cayoglu et al. (2019). On Advancement of Information Spaces to Improve Prediction-Based Compression
GI INFORMATIK 2019

Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression
IEEE Big Data 2018

Cayoglu et al. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms
IEEE eScience 2019

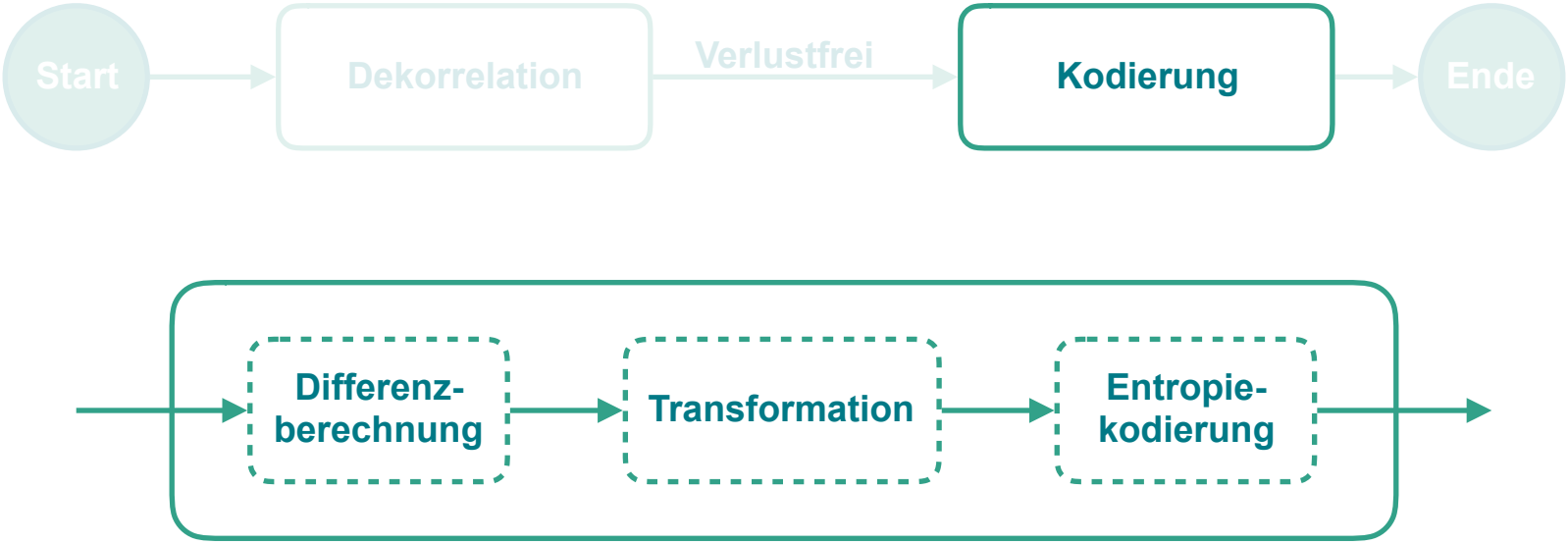
Cayoglu et al. (2018b). A Modular Software Framework for Compression of Structured Climate Data
ACM SIGSPATIAL 2018

Cayoglu et al. (2017). Adaptive Lossy Compression of Complex Environmental Indices Using Seasonal Auto-Regressive Integrated Moving Average Models
IEEE eScience 2017

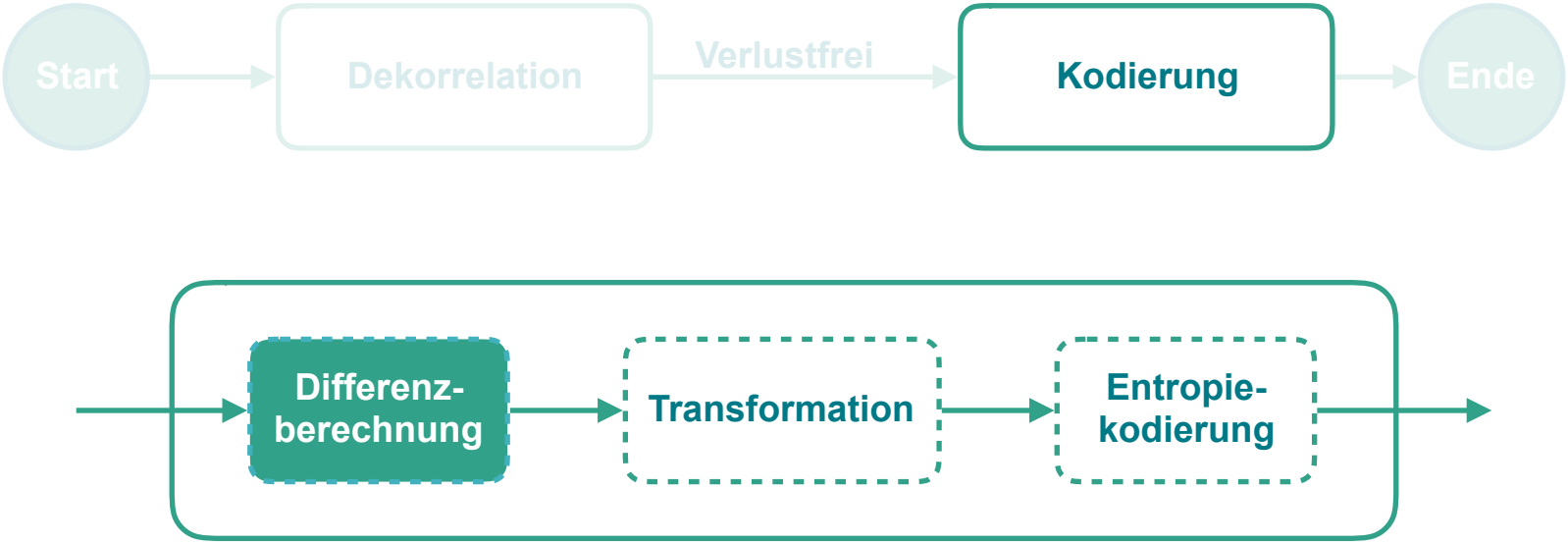
Kerzenmacher et al. (2017). QBO influence on the ozone distribution in the extra-tropical stratosphere
EGU 2018

Datenkodierung bei der verlustfreien vorhersagebasierten Kompression

Datenkodierung bei der verlustfreien vorhersagebasierenden Kompression



Datenkodierung bei der verlustfreien vorhersagebasierenden Kompression



Zwei Arten der Berechnung von Residuen

Karlsruher Institut für Technologie

Abs. Differenz

$$d = |v - w|$$

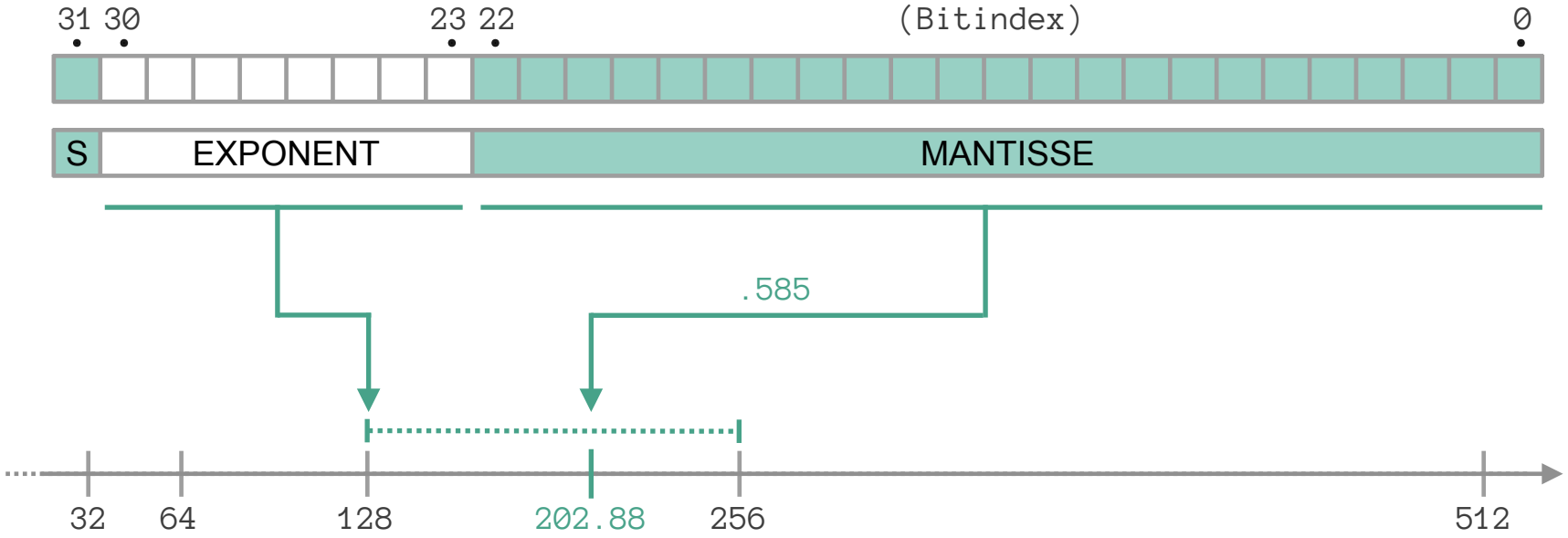
- ⊕ Kleine Residuen
- ⊖ Underflow
- ⊖ Zwei Operationen
- ⊖ Bit für Vorzeichen

XOR

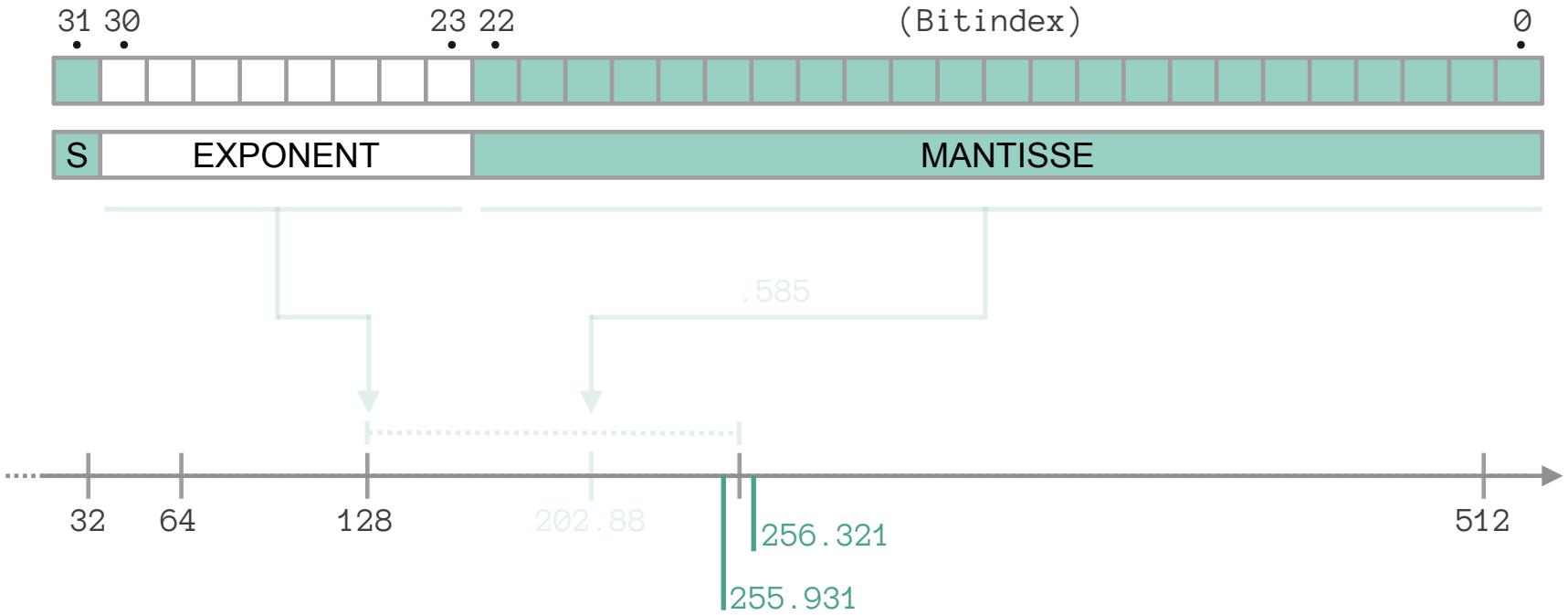
$$d = v \oplus w$$

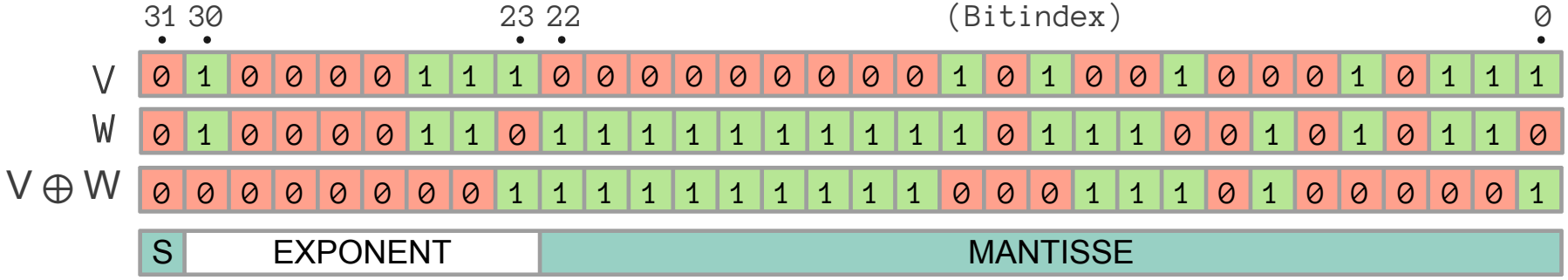
- ⊕ Eine Operation
- ⊕ Kein Underflow
- ⊖ Bitflip-Problem (große Residuen)

Gleitkommazahlen und der Bitflip

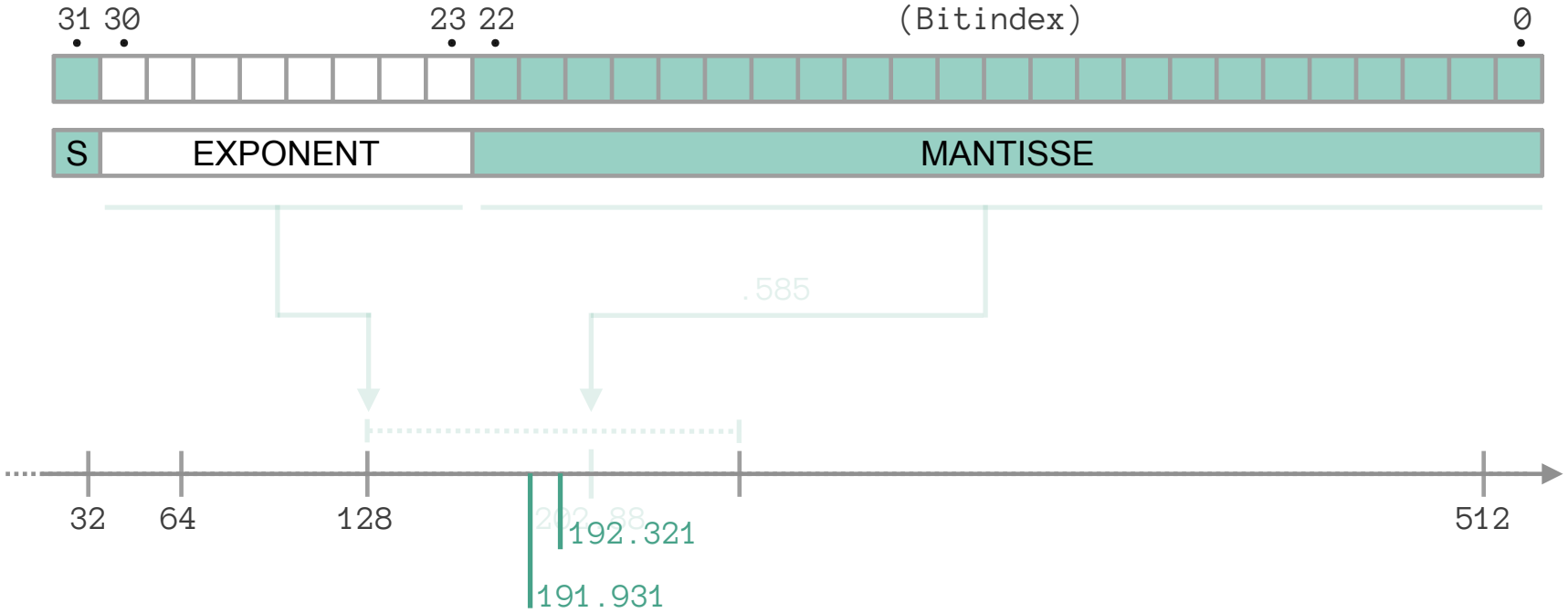


Gleitkommazahlen und der Bitflip

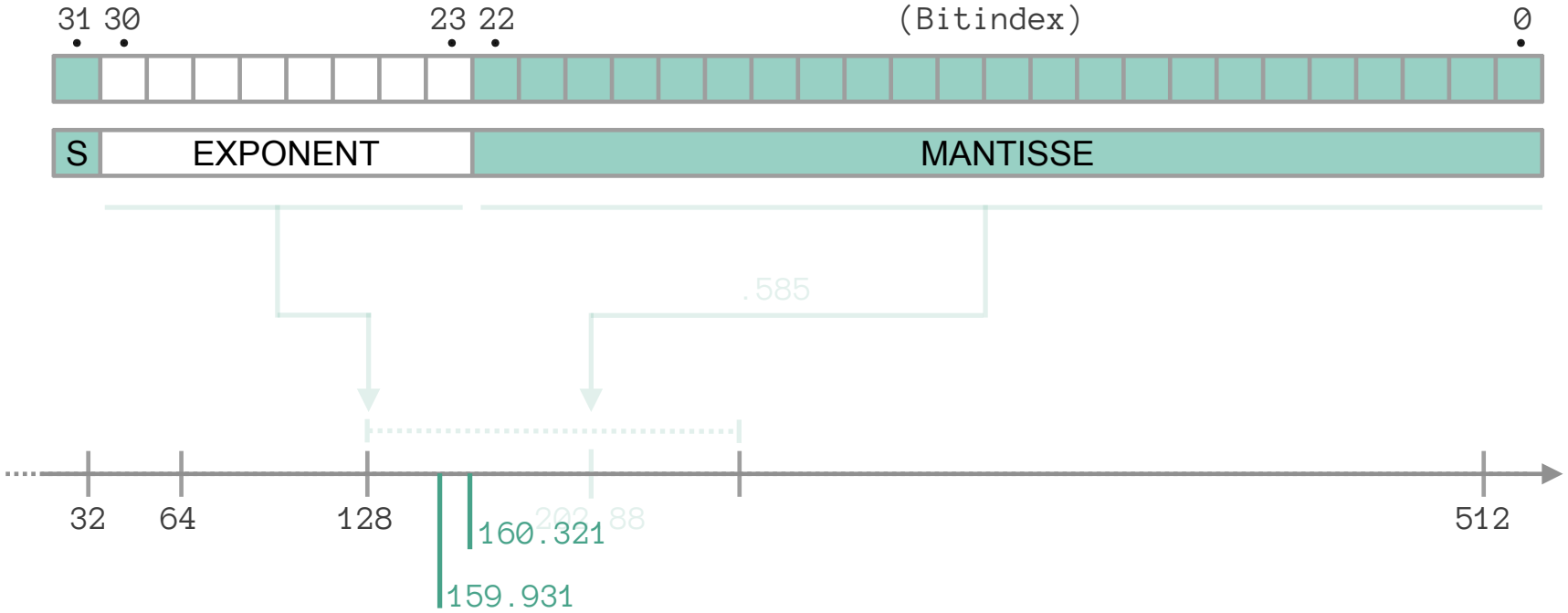




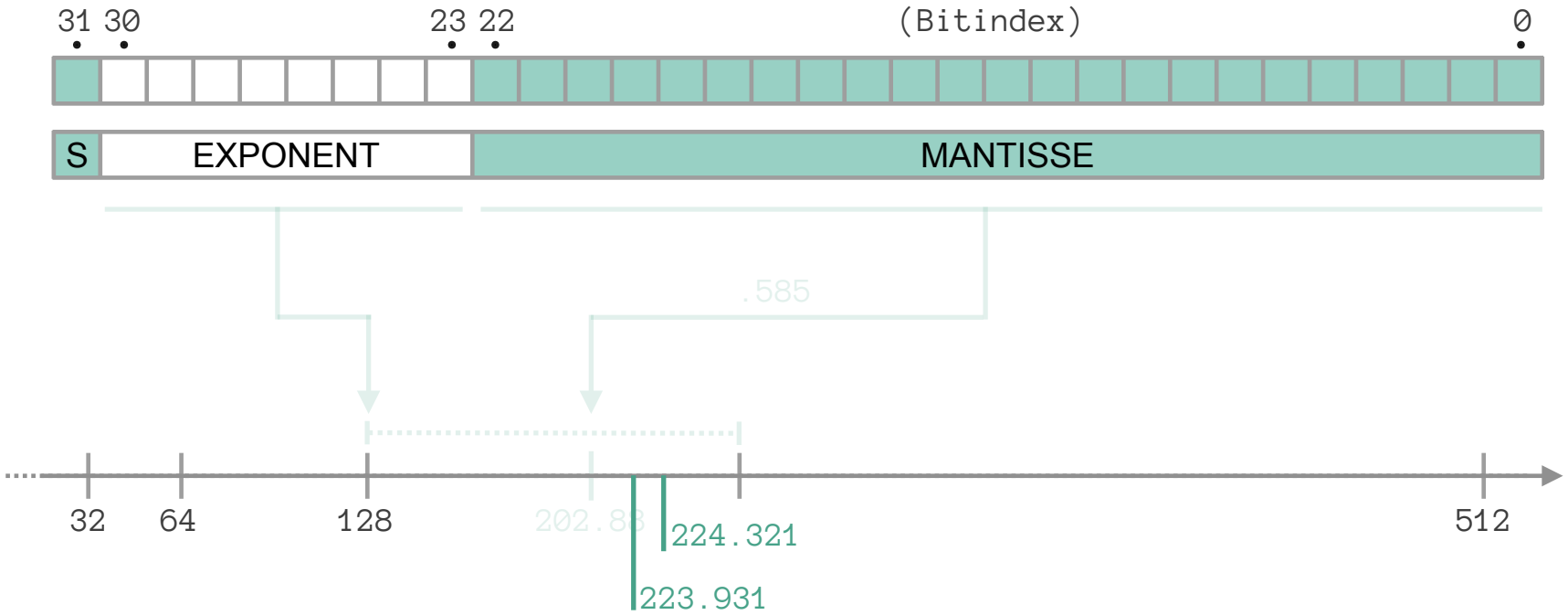
Gleitkommazahlen und der Bitflip



Gleitkommazahlen und der Bitflip



Gleitkommazahlen und der Bitflip



Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

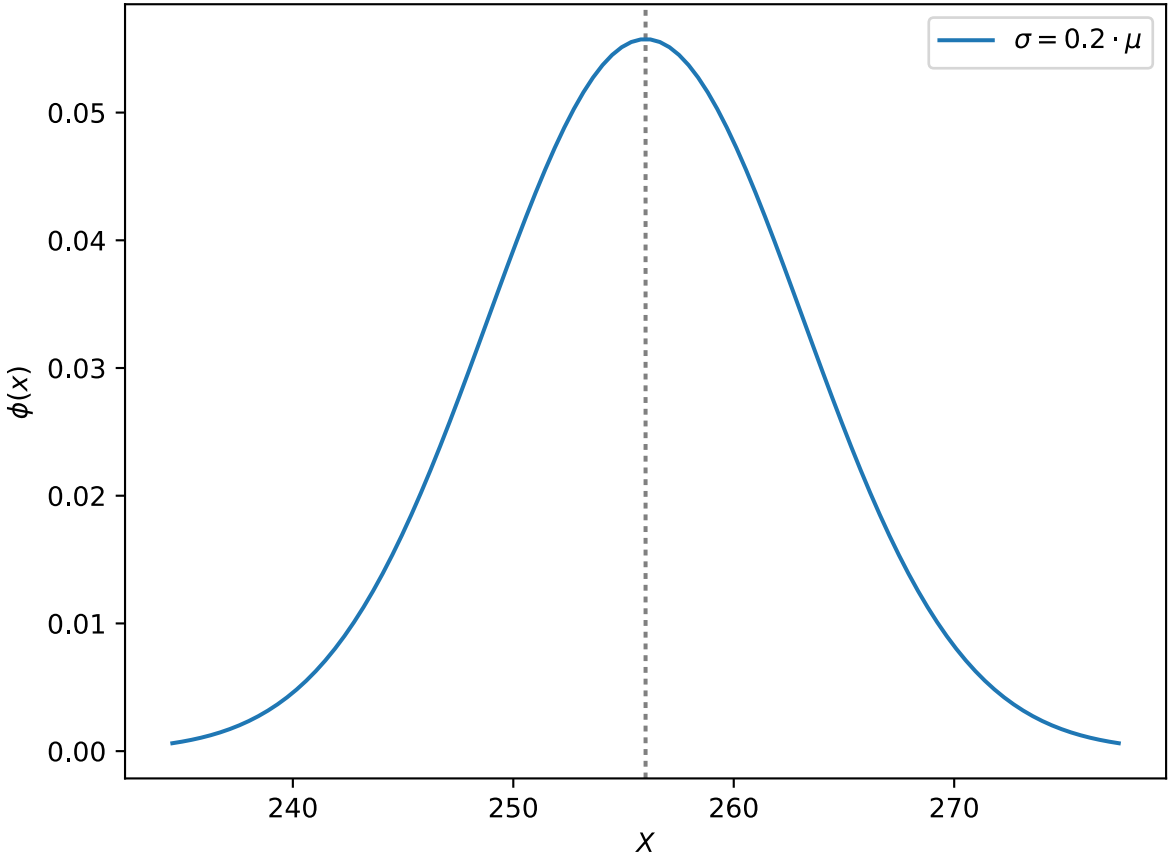
Karlsruher Institut für Technologie

Frage

Kann ich im Vorfeld bestimmen wie stark die Kompression vom Bitflip betroffen sein wird?

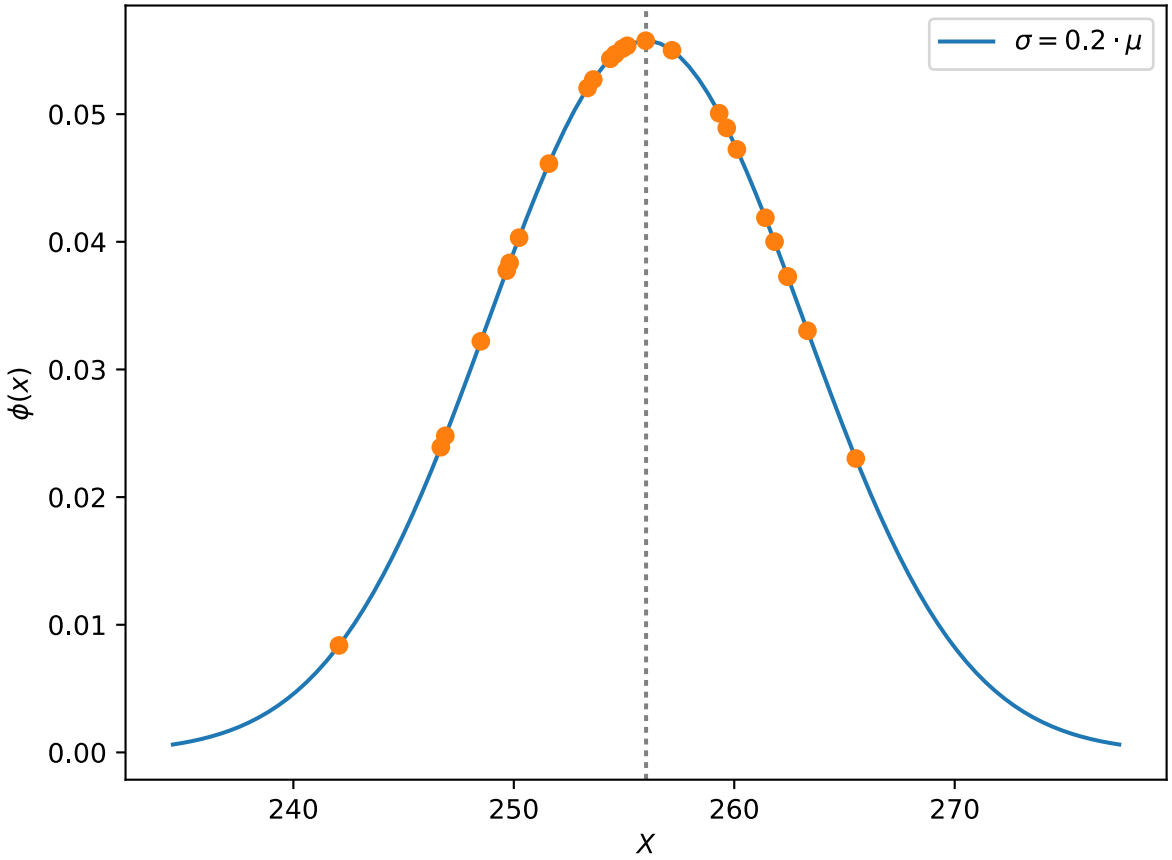
Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

Prämisse: Normalverteilung der Vorhersagen



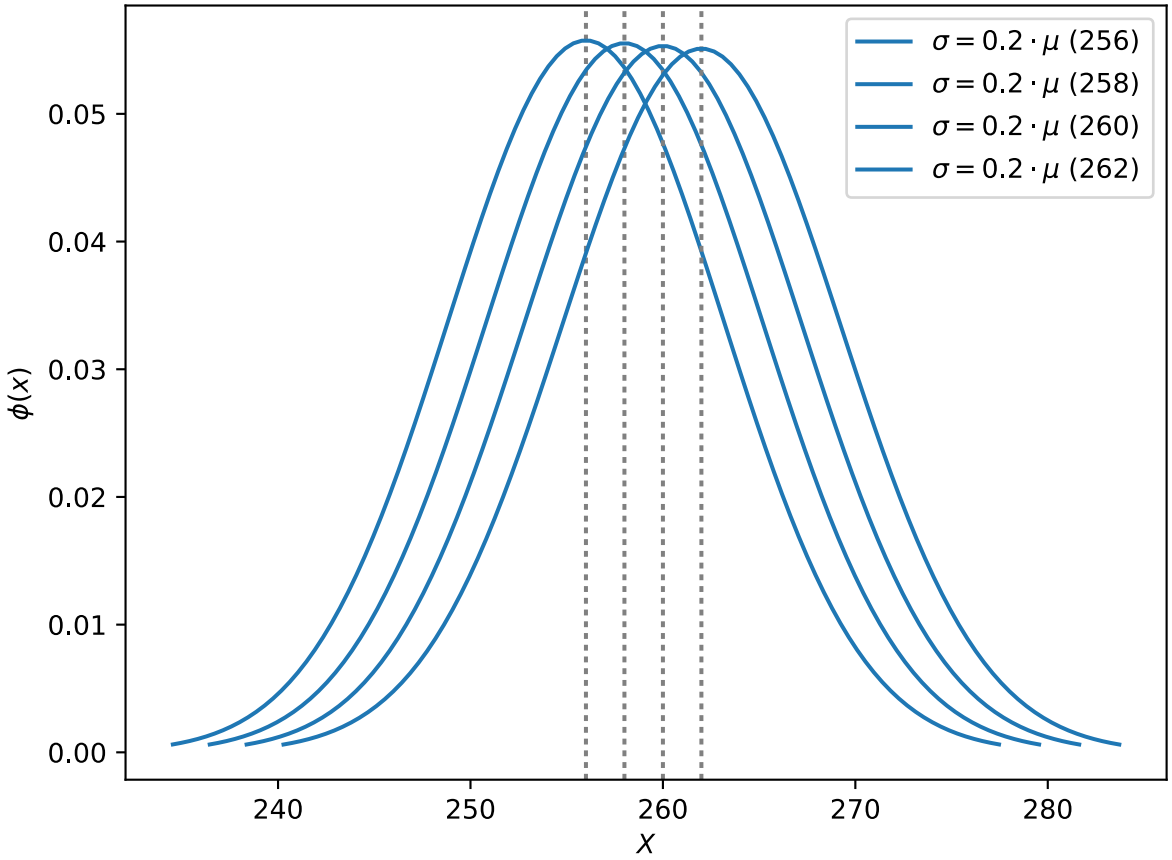
Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

Prämisse: Normalverteilung der Vorhersagen



Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

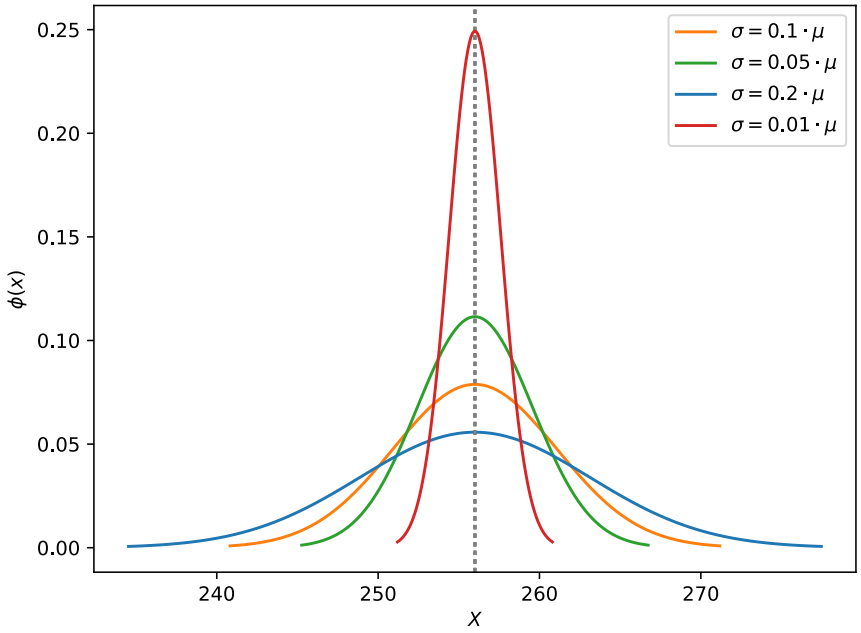
Prämisse: Normalverteilung der Vorhersagen



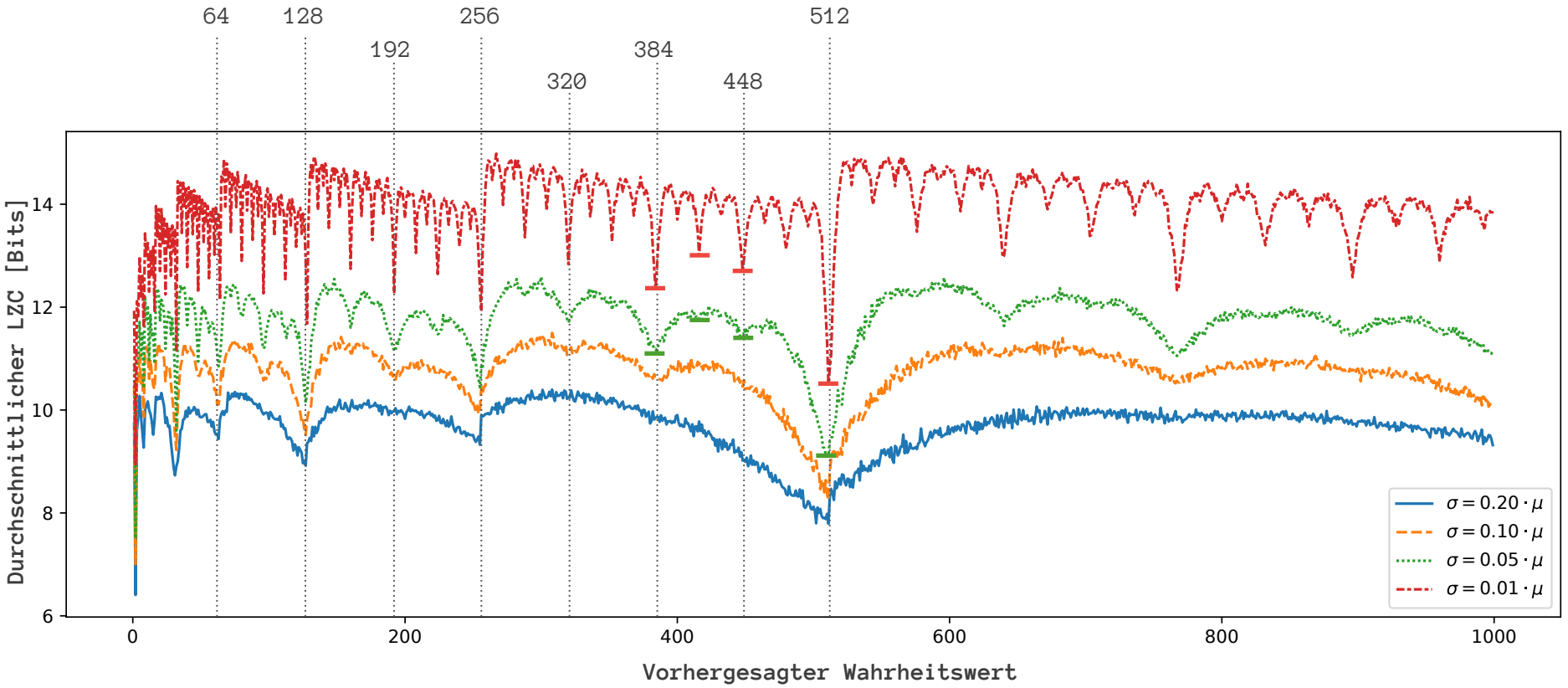
Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor

Karlsruher Institut für Technologie

- Verschiedene normalverteilte Datensätze (100 Datenpunkte)
  - Erwartungswert (Wahrheit) $\mu \in [0; 1000]$ mit $\mu \in \mathbb{R}$
  - Standardabweichung (Vorhersagen) $\sigma \in \{0.2\mu, 0.1\mu, 0.05\mu, 0.01\mu\}$
- Berechnen des durchschnittlichen LZC



Untersuchung der Auswirkung von Bitflips auf den Kompressionsfaktor



Lösung zum Bitflip-Problem

Karlsruher Institut für Technologie

- Verschiebung der Differenzberechnung in einen anderen Wertebereich

$$\begin{aligned} \text{Vorhersage} &= V \\ \text{Wahrheit} &= W \\ V \oplus W &= R \end{aligned} \qquad \begin{aligned} V + \text{Shift} &= V' \\ W + \text{Shift} &= W' \\ V' \oplus W' &= R' \end{aligned} \qquad V + \text{Shift} = ?$$

- Eigenschaften vom Zielwert
  - **Größtmögliche Distanz** zu Zweierpotenzen
  - **Minimale Fortpflanzung** von Bitflips
    (durch Addition/Subtraktion einer beliebigen Zahl)
- Verschiebung darf keinen **großen Overhead** erzeugen
- Verschiebung muss **reproduzierbar sein** für den Dekompressor

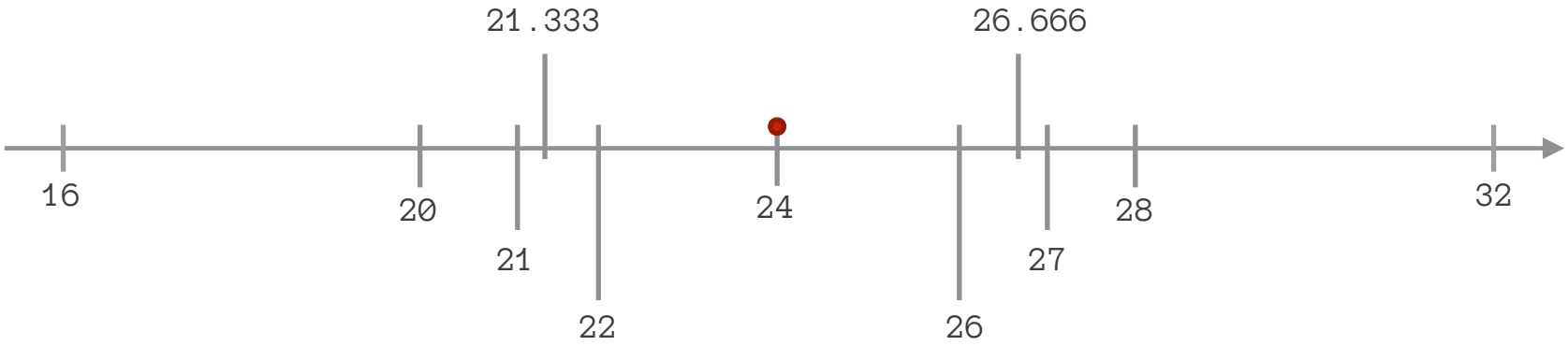
Lösung zum Bitflip-Problem

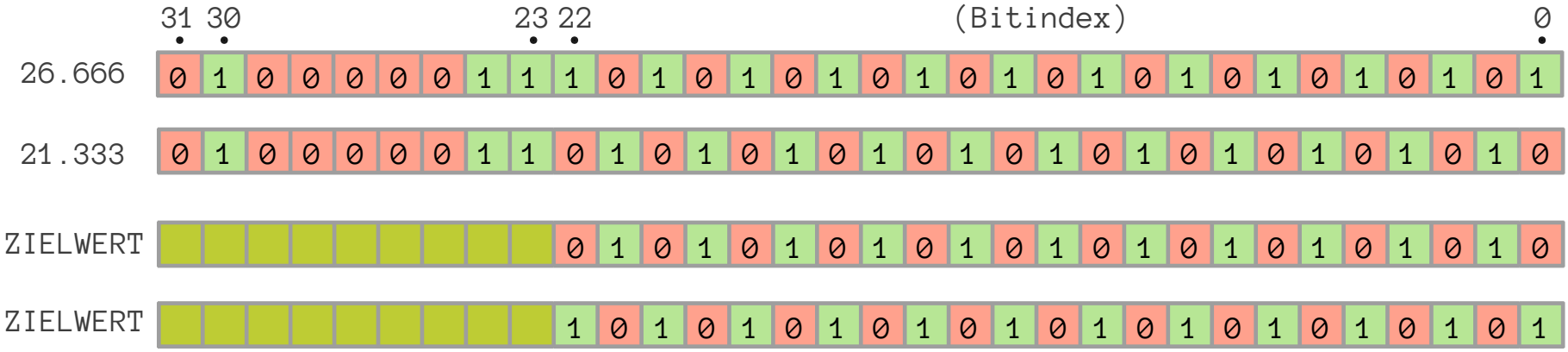
- Verschiebung der Differenzberechnung in einen anderen Wertebereich

```
Vorhersage = V        V + Shift = V'
  Wahrheit = W        W + Shift = W'        V + Shift = ?
     V ⊕ W = R          V' ⊕ W' = R'
```

- Eigenschaften vom Wertebereich
  - **Größtmögliche Distanz** zu Zweierpotenzen
  - **Minimale Fortpflanzung** von Bitflips
    (durch Addition/Subtraktion einer beliebigen Zahl)
- Verschiebung darf keinen **großen Overhead** erzeugen
- Verschiebung muss **reproduzierbar sein** für den Dekompressor

Vermeidung des Bitflip-Problems

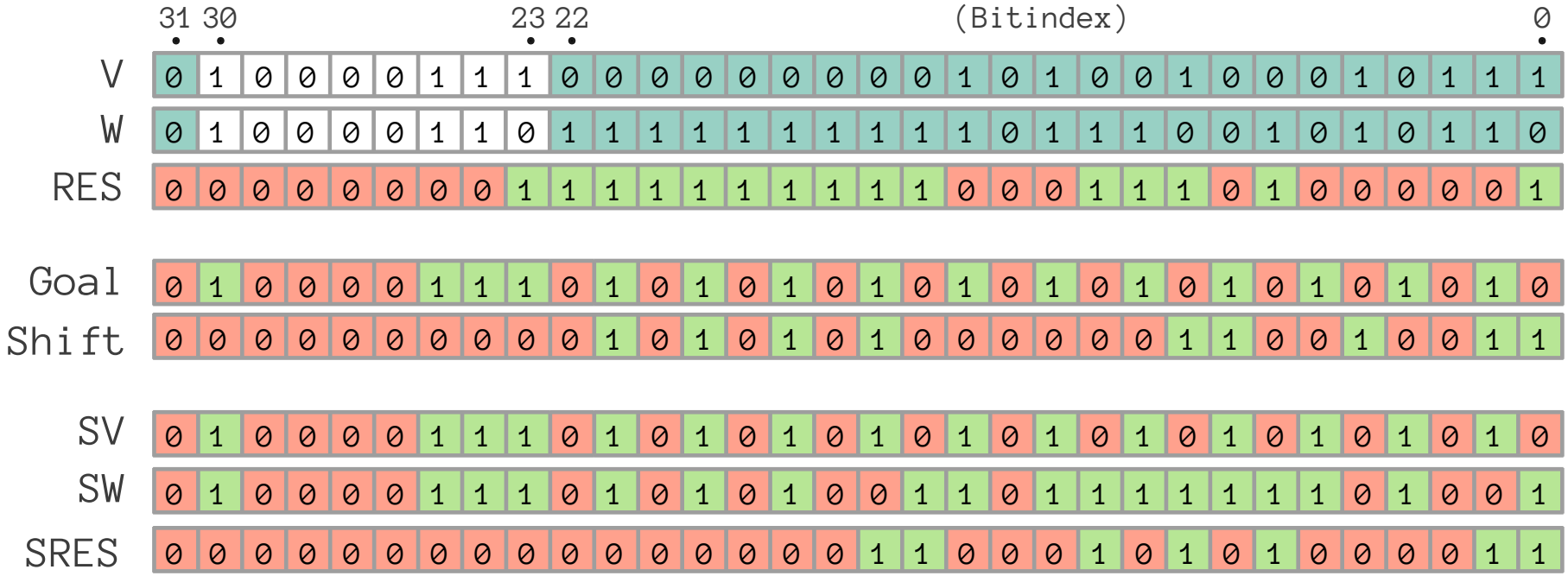




Verschiebung anhand eines Beispiels

Karlsruher Institut für Technologie

```
Vorhersage (V):   256.321
Wahrer Wert (W):  255.931
```



LZC: 8 -> 16

Verschiebung anhand eines Beispiels

Karlsruher Institut für Technologie

```
 Vorhersage (V):  256.321
Wahrer Wert (W):  255.931          LZC: 8 -> 16
```

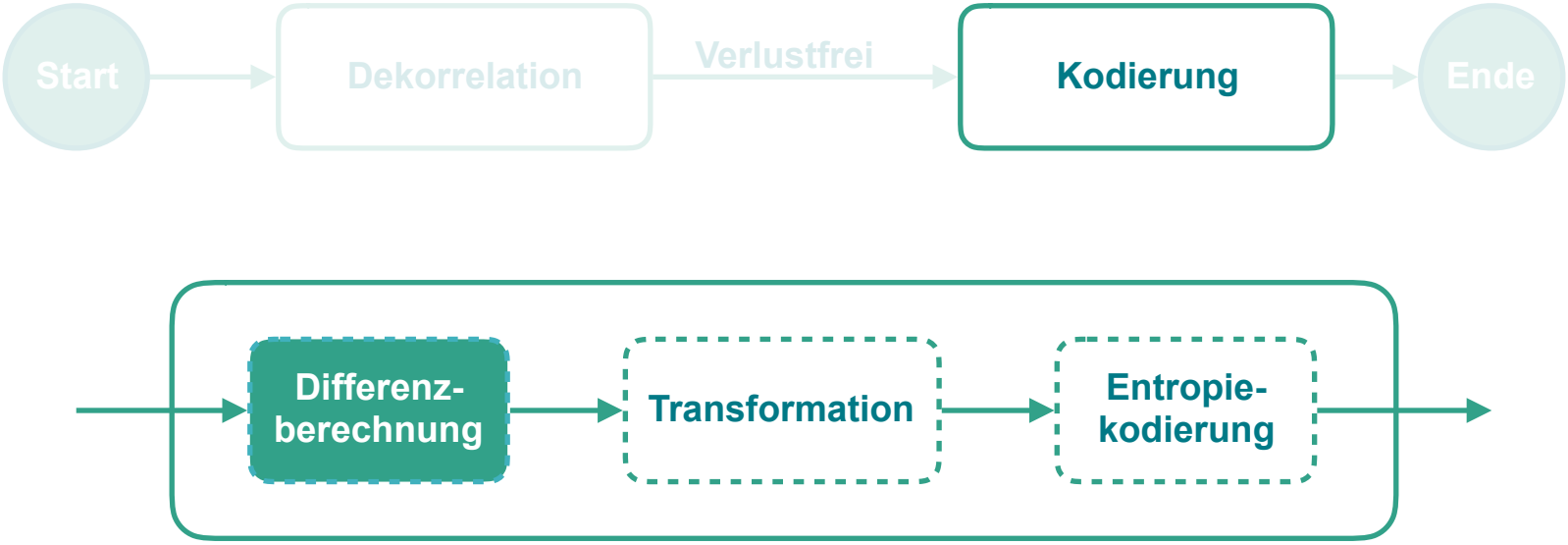
Es funktioniert besser, je näher die Zahlen an Zweierpotenzen liegen

```
 Vorhersage (V):  256.002
Wahrer Wert (W):  255.991          LZC: 8 -> 21
```

Es funktioniert besser, je größer die Zahlen sind

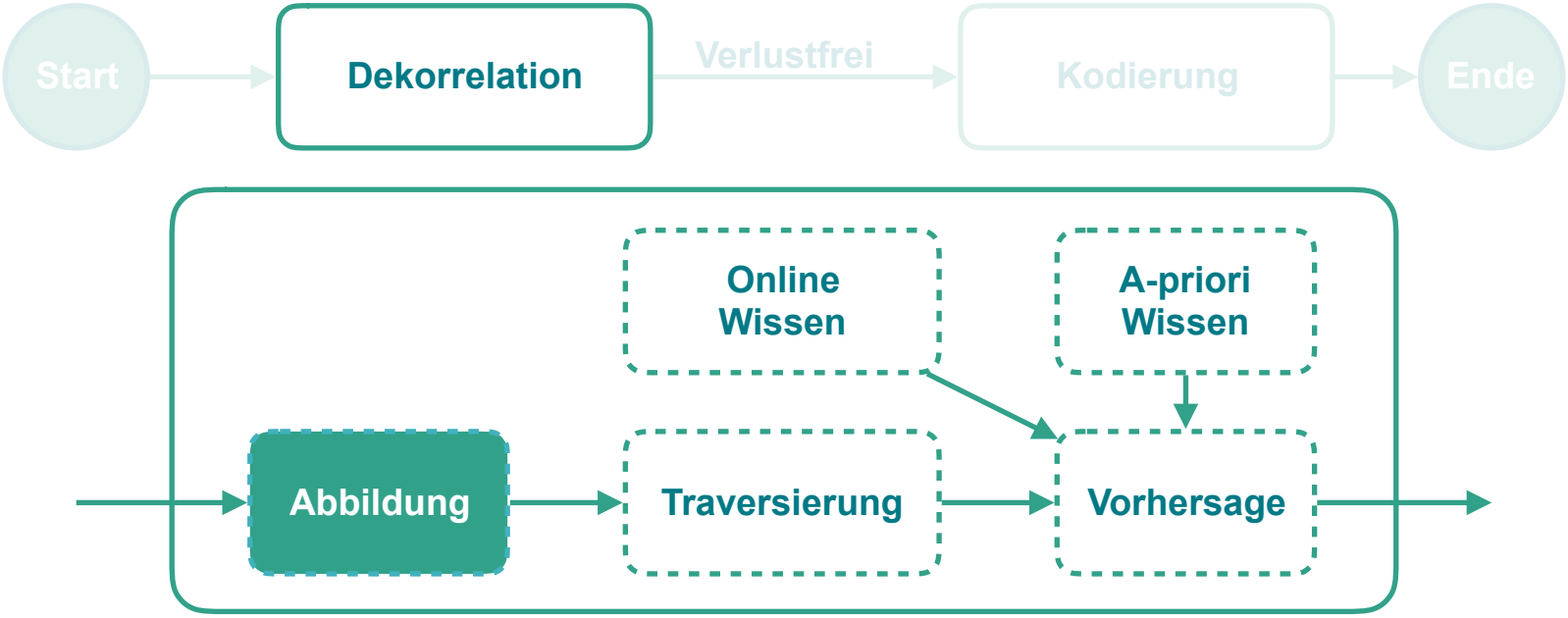
```
 Vorhersage (V): 1024.002
Wahrer Wert (W): 1023.991          LZC: 8 -> 24
```

Datenkodierung bei der verlustfreien vorhersagebasierenden Kompression



Pascal Zip (pzip)

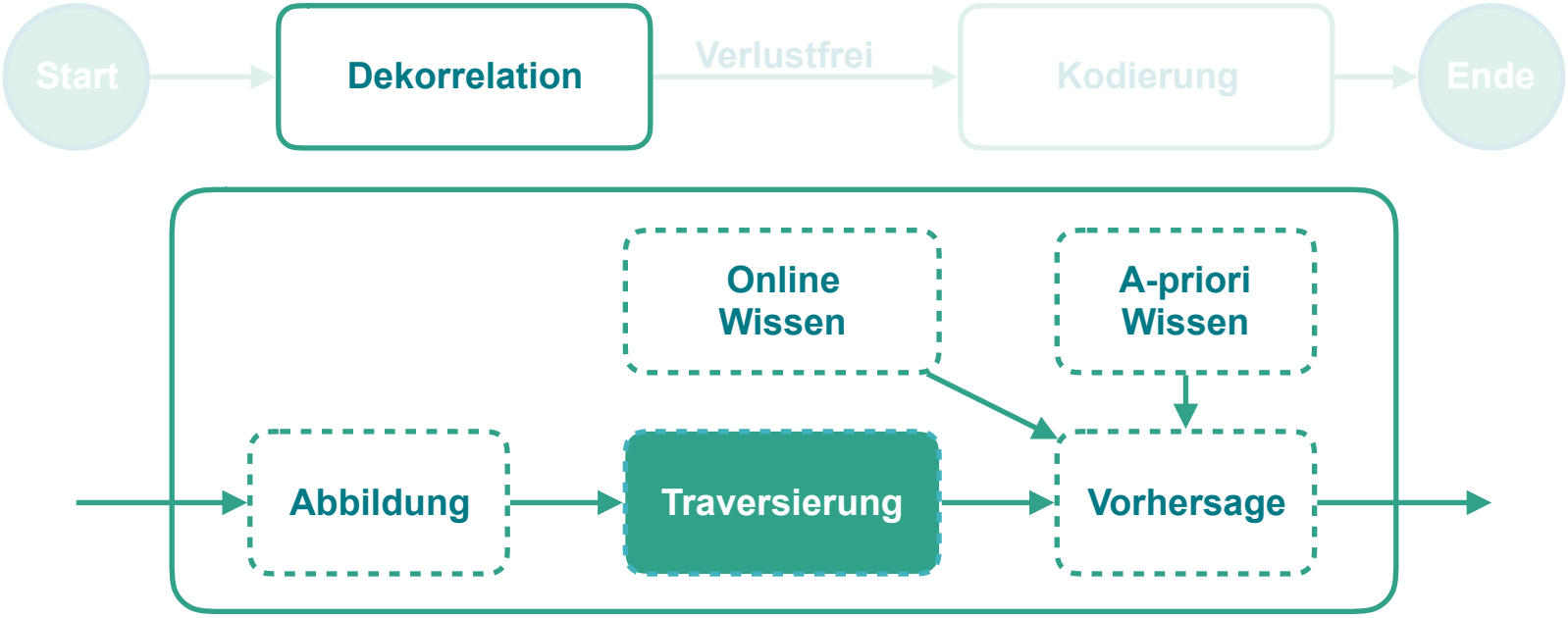
Pascal Zip (pzip)



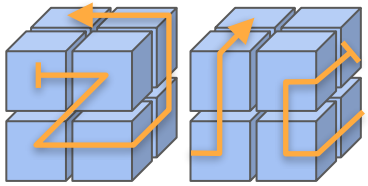
$$m : \mathbb{R} \rightarrow \mathbb{N}$$

```
256.321 -> 1132472599
255.931 -> 1132457558
    . . .
```

Pascal Zip (pzip)

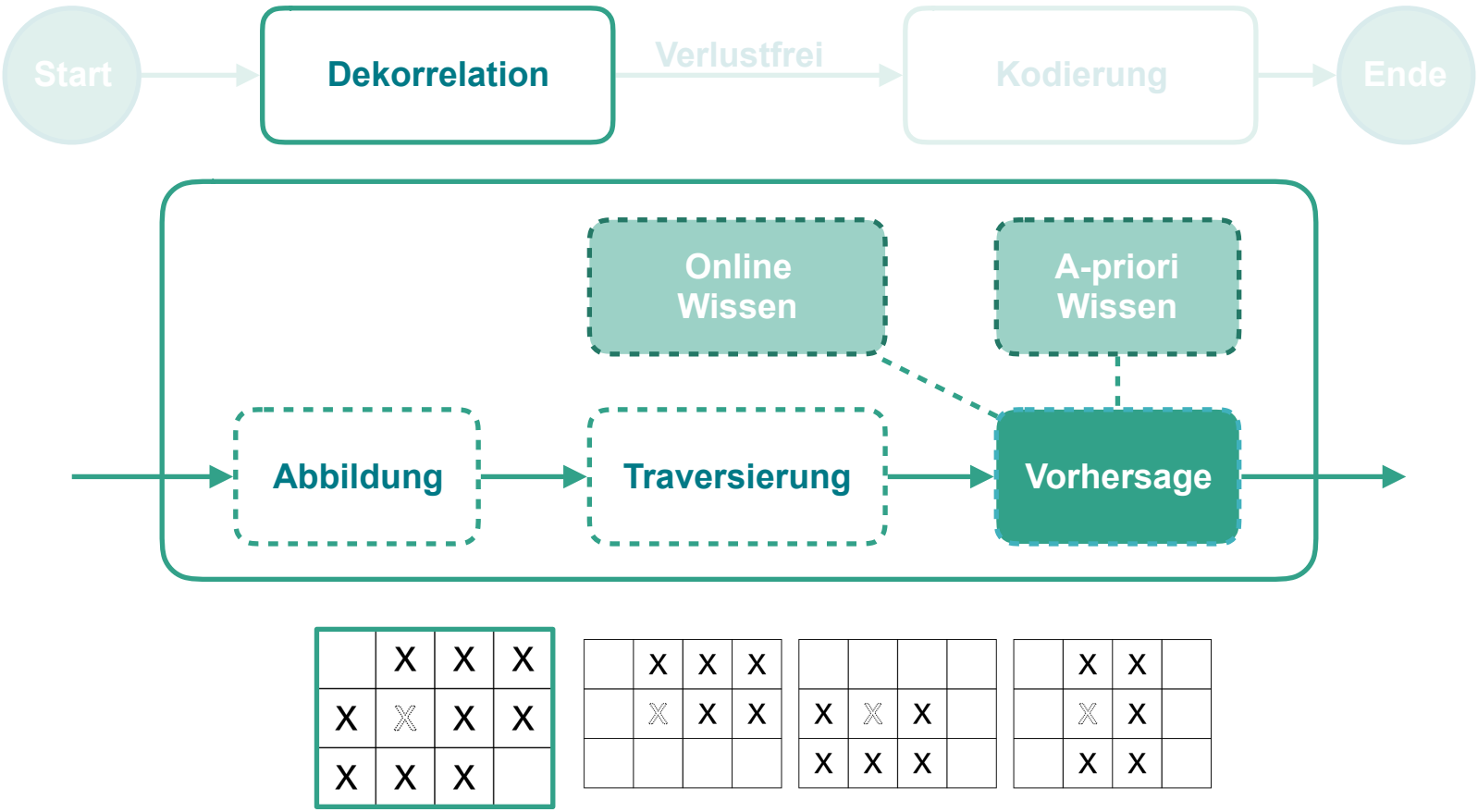


$$t : \mathbb{N} \rightarrow \mathbb{N}$$



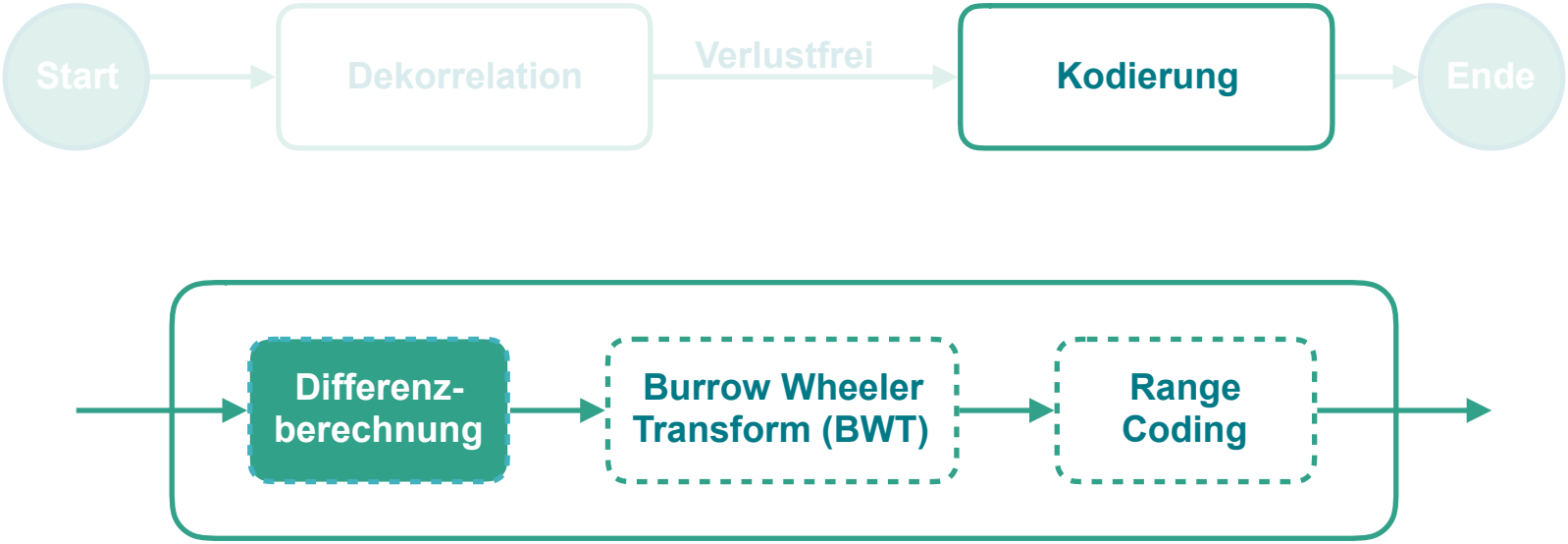
Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression **IEEE Big Data 2018**
Cayoglu et al. (2019). On Advancement of Information Spaces to Improve Prediction-Based Compression **GI INFORMATIK 2019**

Pascal Zip (pzip)



Cayoglu et al. (2018a). Towards an optimised environmental data compression method for structured model output **EGU 2018**
Cayoglu et al. (2017). Adaptive Lossy Compression of Complex Environmental Indices Using SARIMA **IEEE eScience 2017**
Cayoglu et al. (2018). Concept and Analysis of Information Spaces to improve Prediction-Based Compression **IEEE Big Data 2018**

Pascal Zip (pzip)

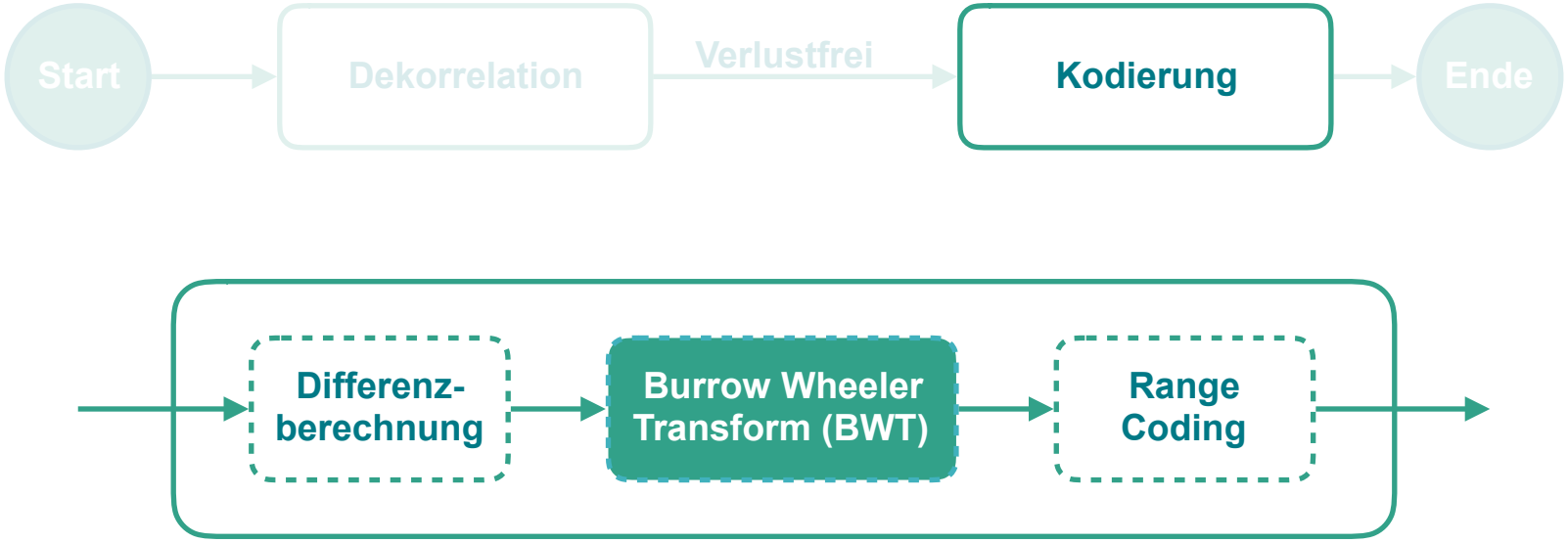



```
V + Shift = V'
W + Shift = W'
  V' ⊕ W' = R'
```

```
R' = 00000000000000111010111001000 1011
LZC = 14, FOC = 03, RES = 101110010001011
```

Cayoglu et al. (2019b). Data Encoding in Lossless Prediction-Based Compression Algorithms **IEEE eScience 2019**

Pascal Zip (pzip)



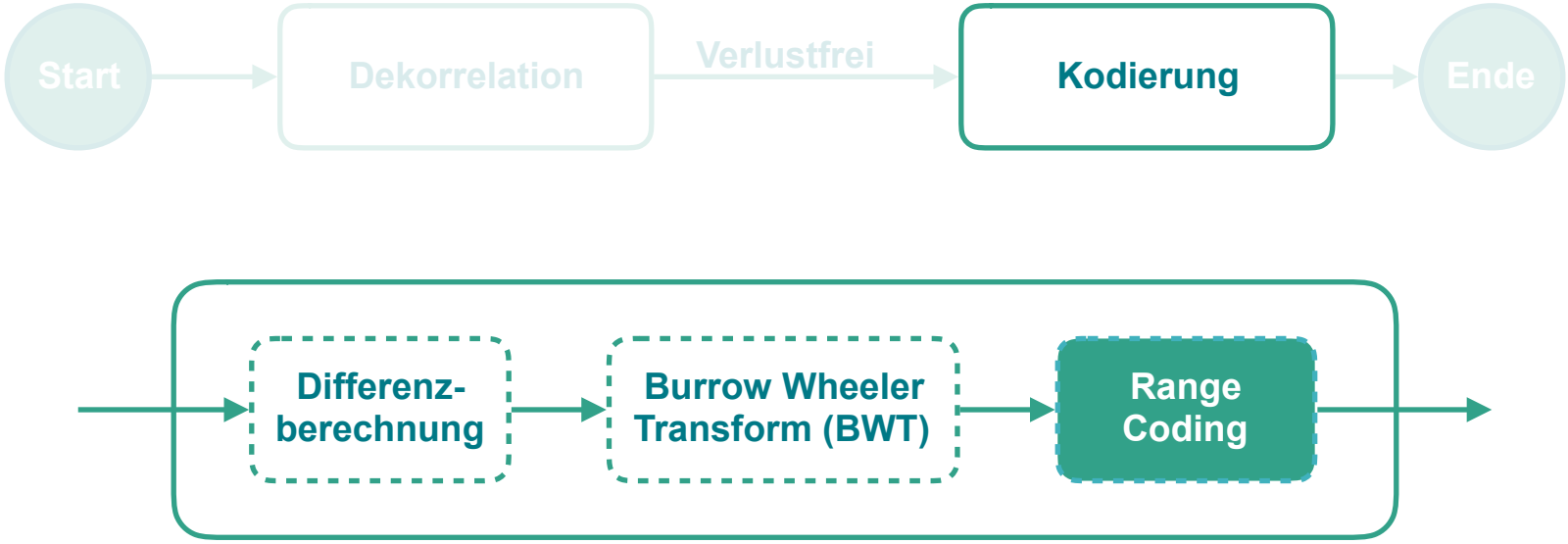
^BANANA$

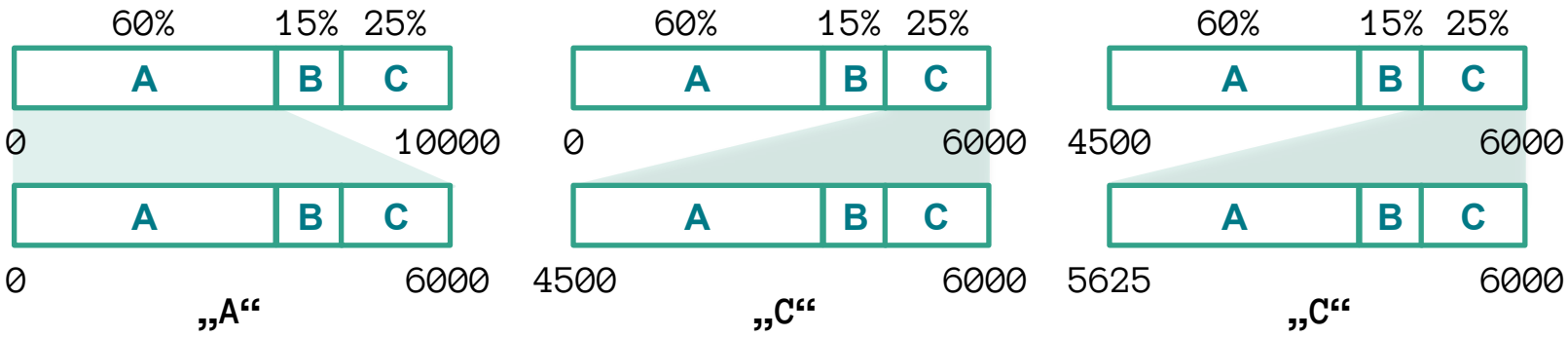
^BANANA$	**A**NANA$^B	ANANA$^**B**
$^BANANA	**A**NA$^BAN	ANA$^BA**N**
A$^BANAN	**A$**^BANAN	A$^BANA**N**
NA$^BANA	**B**ANANA$^	BANANA$^
ANA$^BAN	**N**ANA$^BA	NANA$^B**A**
NANA$^BA	**N**A$^BANA	NA$^BAN**A**
ANANA$^B	^BANANA$	^BANANA**$**
BANANA$^	**$**^BANANA	$^BANAN**A**

BNN^AA$A

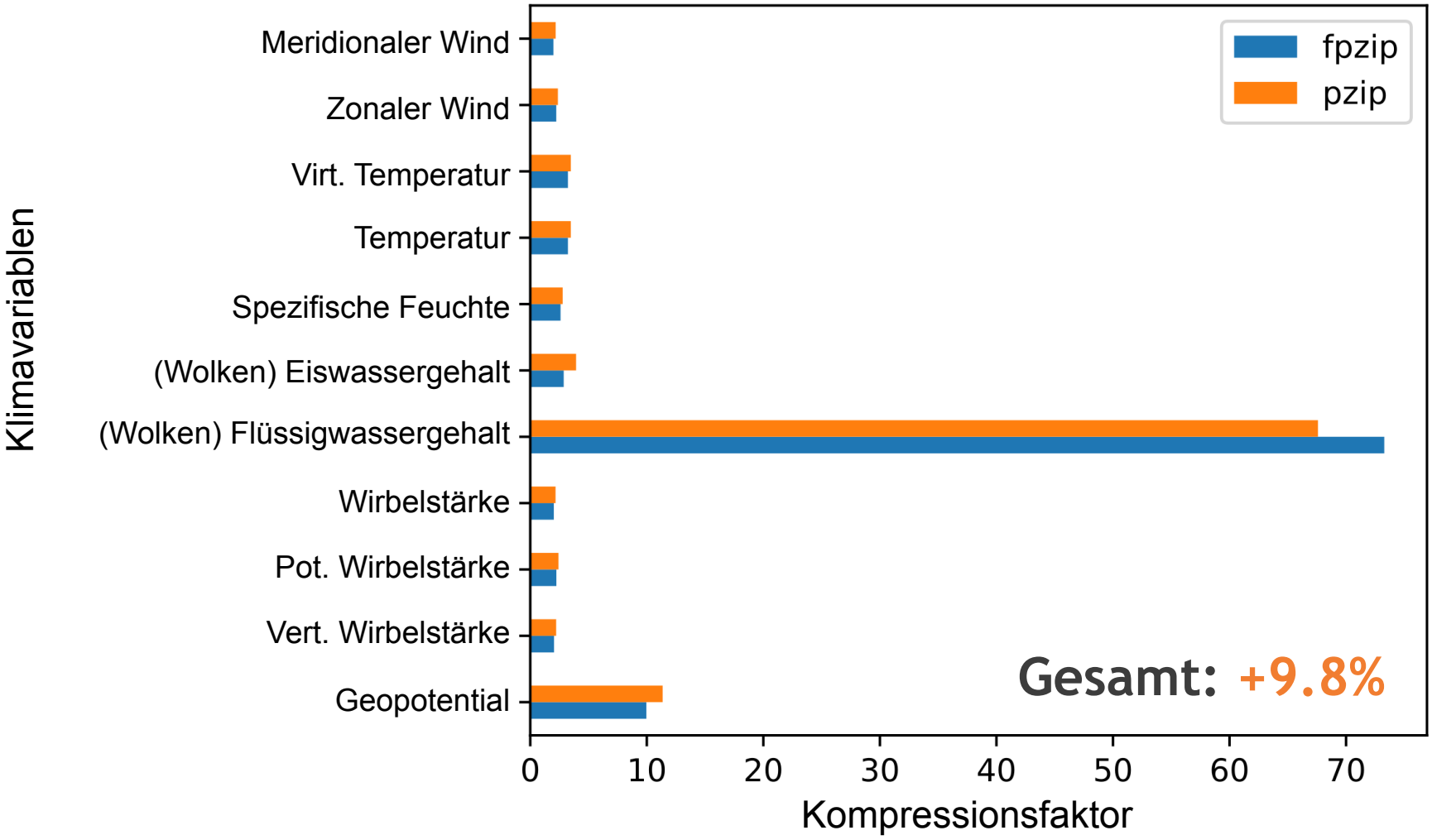
Quelle[3]

Pascal Zip (pzip)

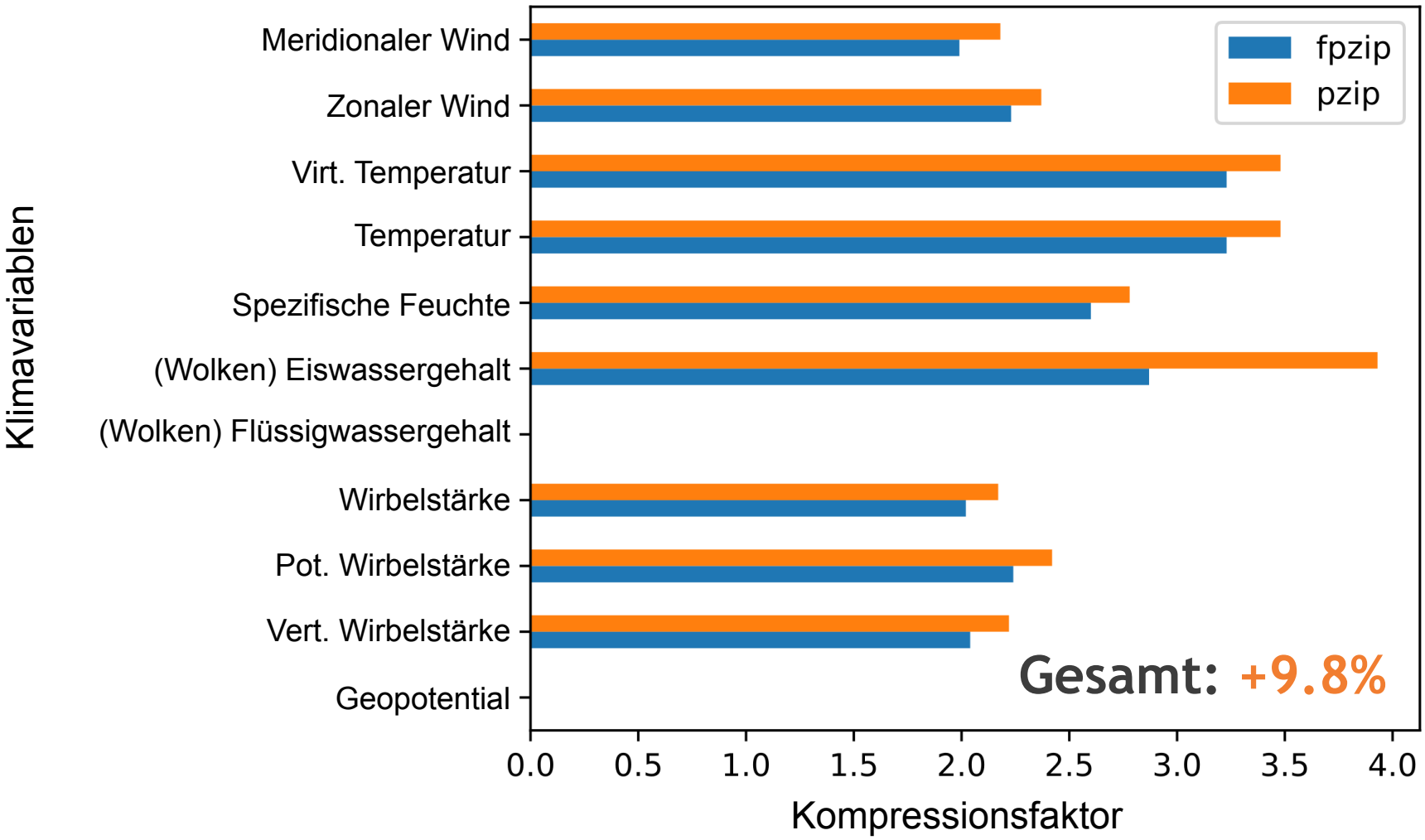




Kompressionsfaktor

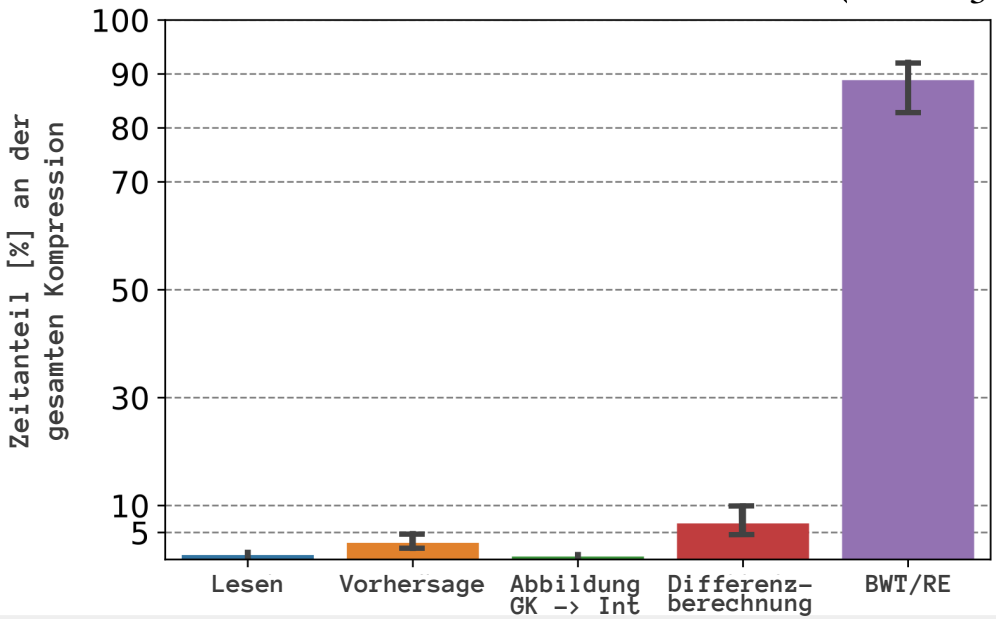


Kompressionsfaktor



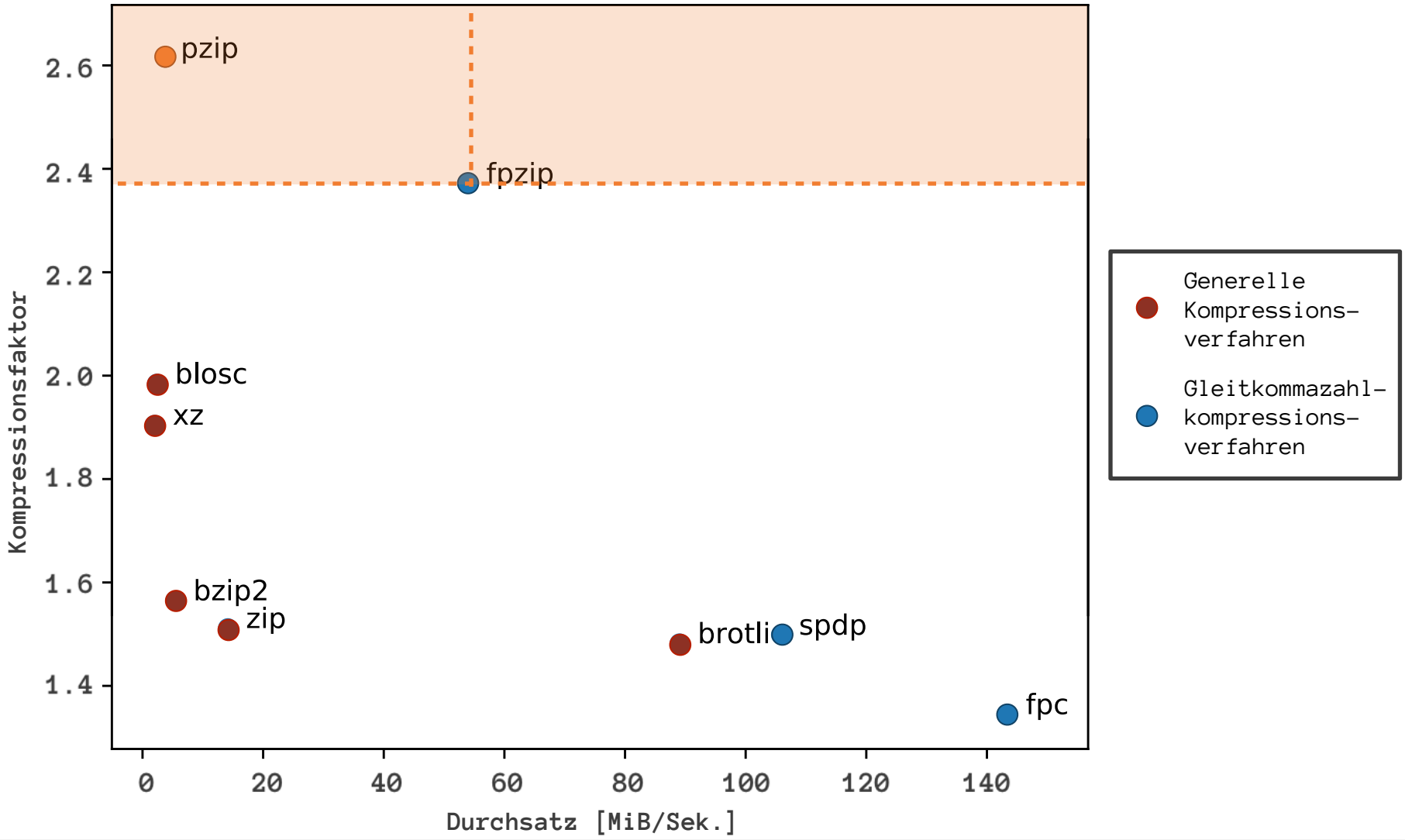
Durchsatz und Komplexität

- Engpass in der aktuellen Implementierung ist BWT/RE
- Laufzeit- und Speicherplatzkomplexität von fpzip $\mathcal{O}(n)$
- Laufzeitkomplexität $\mathcal{O}(n + 4 \cdot \tau \cdot n + n)$
- Speicherkomplexität $\mathcal{O}(\tau \cdot \left(1 + \frac{n}{d_3}\left(\frac{1}{d_2}\left(\frac{1}{d_1} + 1\right) + 1\right)\right) + n \log \sigma)$



τ = Nachbarschaft
$n = d_0 d_1 d_2 d_3$
σ = |Alphabet|

Kompressionsverfahren im Vergleich



Verlustfreie Kompression von Klimadaten

Karlsruher Institut für Technologie

Reduktion

		~2.6	
ERA5:	10.89 PiB	⇨	4.19 PiB
IMK-ASF:	770 TiB	⇨	296 TiB

Open Source

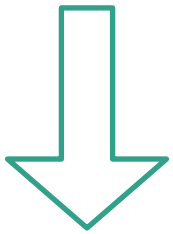
Alle Programmbeispiele und Daten sind öffentlich zugänglich (github.com, GPLv3)

Beitrag

Andere Kompressionsverfahren können einzelne Entwicklungen aus der Arbeit aufgreifen und einbauen

Ziel

Verlustfreies Kompressionsverfahren mit hohem Kompressionsfaktor erfüllt ✔



Verlustfreie Kompression von Klimadaten

Vielen Dank