Karlsruhe Institute of Technology

Compression IOI

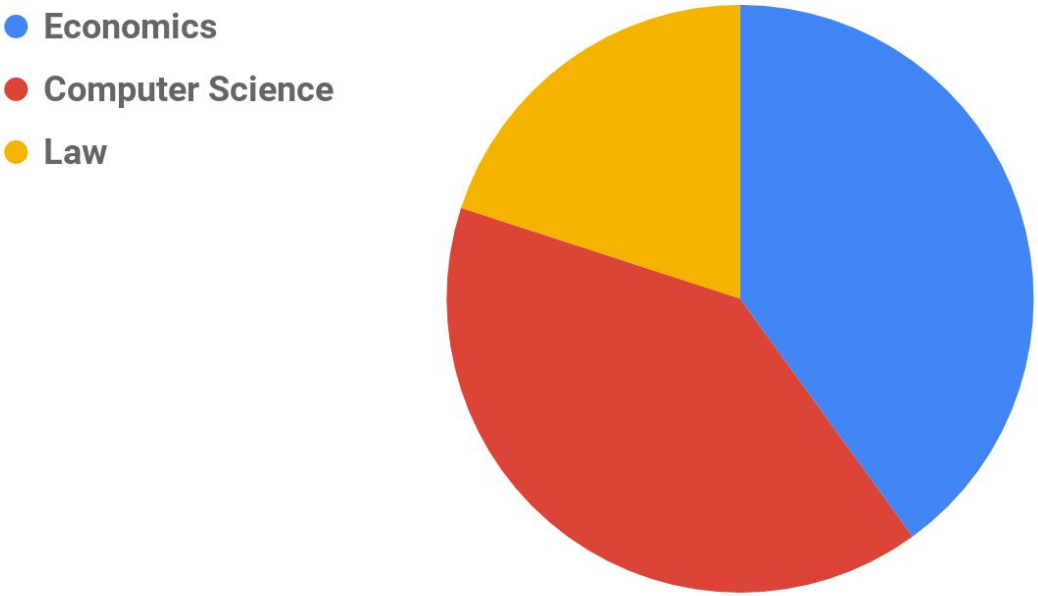
by Uğur Çayoğlu

13. February 2020

STEINBUCH CENTRE FOR COMPUTING (SCC) &
INSTITUTE FOR METEOROLOGY AND CLIMATE RESEARCH (IMK-ASF)

Me, myself and I

- Studied Information Management and Engineering @ KIT



Me, myself and I

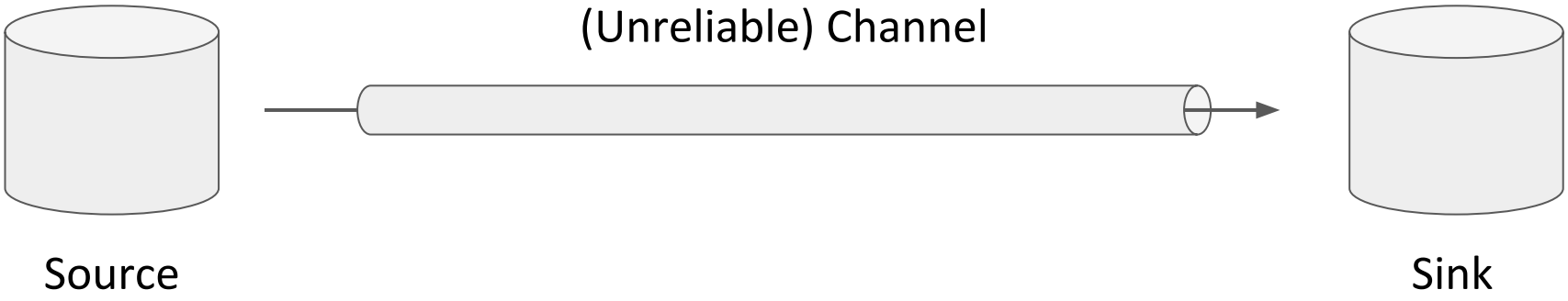
- Studied Information Management and Engineering @ KIT
- Bachelor thesis:
  - *Improve Collaboration Between **Marketing and Sales** and Analyze the Impact on the Quality of **Customer Data** in the Enterprise*

Me, myself and I

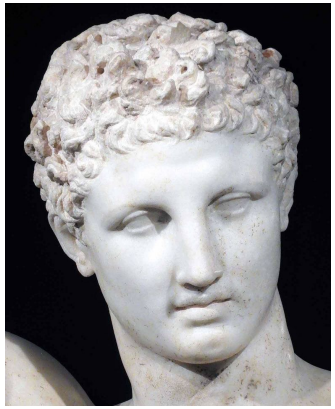
- Master thesis:
  - *Conceptualization and Prototypical Implementation of a Tool to Determine the Optimal **XML Similarity Measurement***
- Doctoral thesis:
  - ***Compression Methods** for Structured Floating Point Data and their Application in Climate Research*

History

- Compression has its origin in communication technology
- Hence compression == source coding

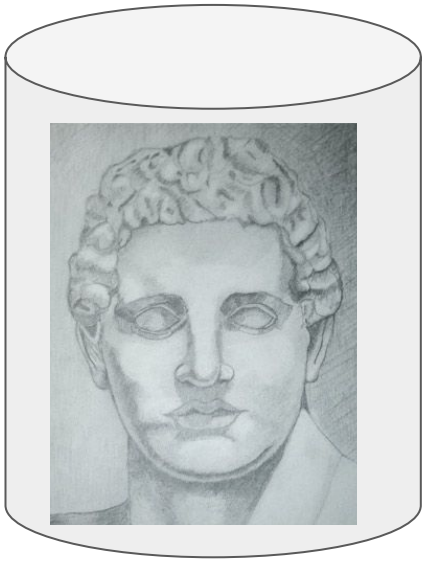


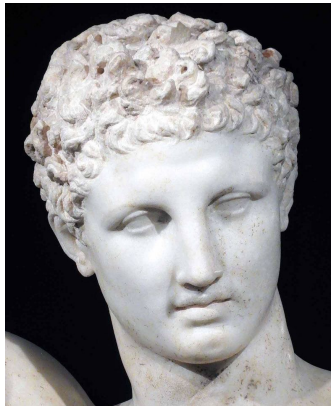
Compression IOI



Source [A]

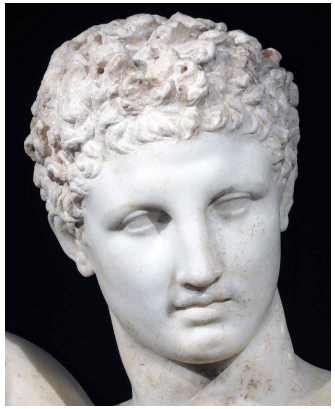
Compression IOI

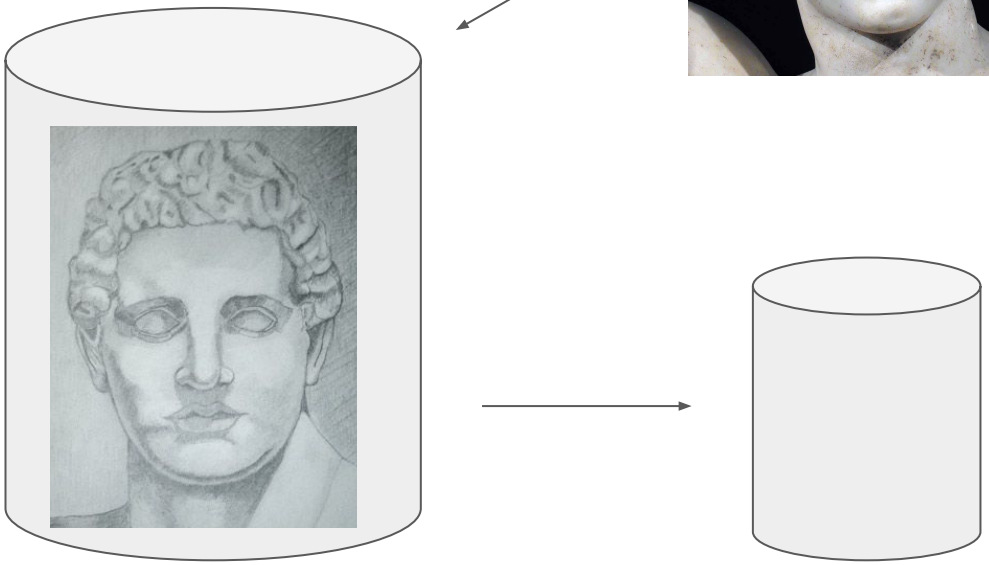




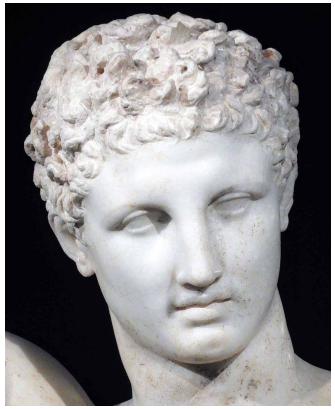
Source [B]

Compression IOI

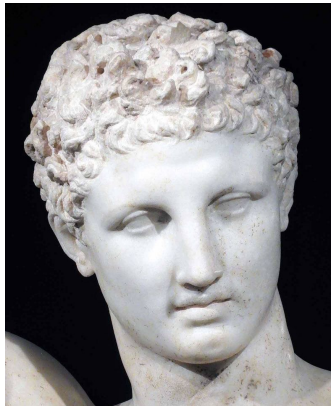


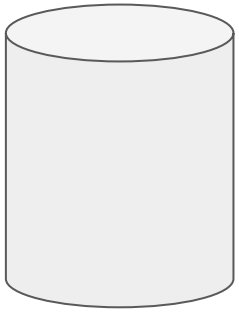


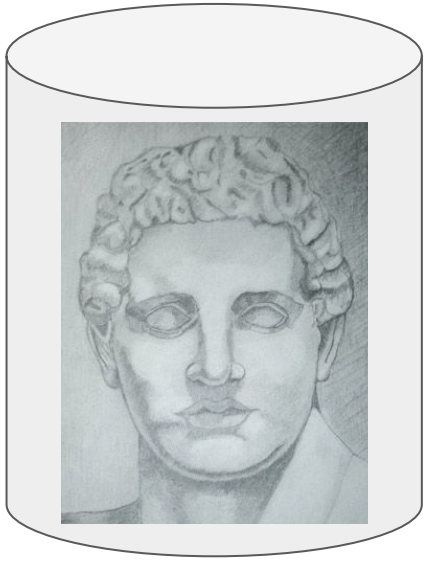
Compression IOI



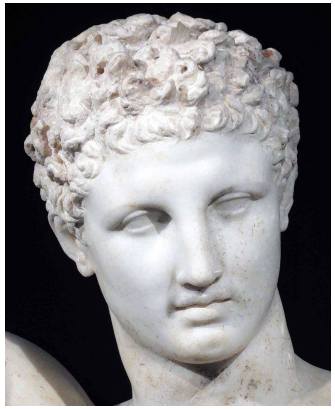
Compression IOI

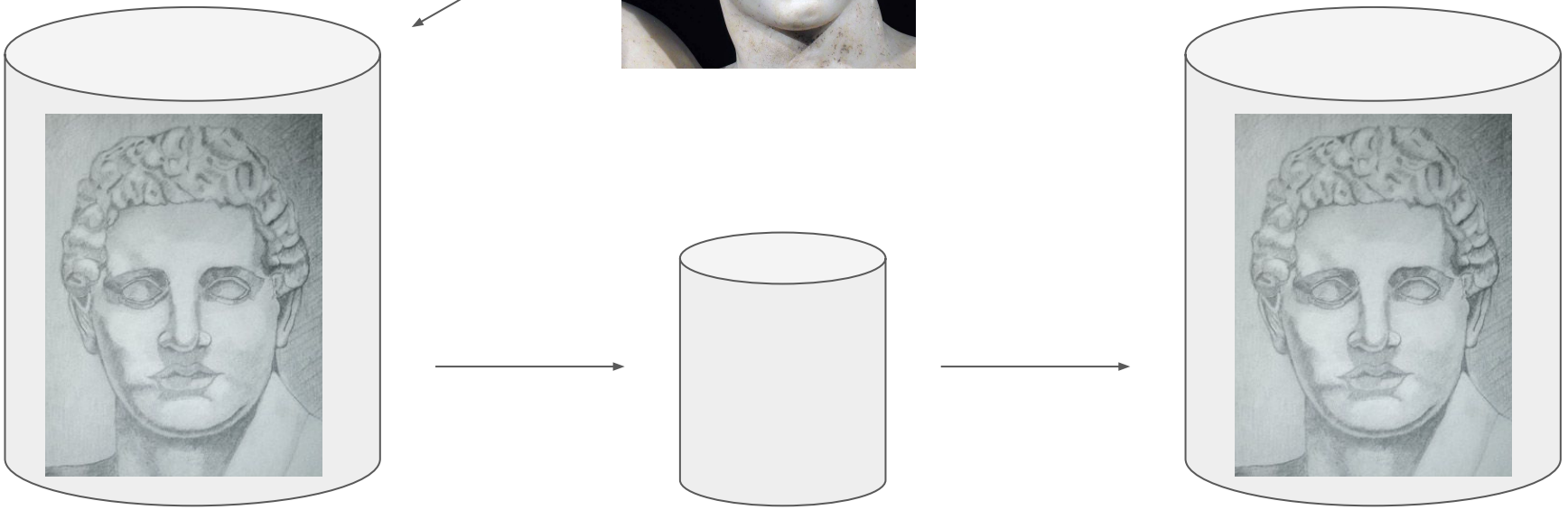






Compression IOI

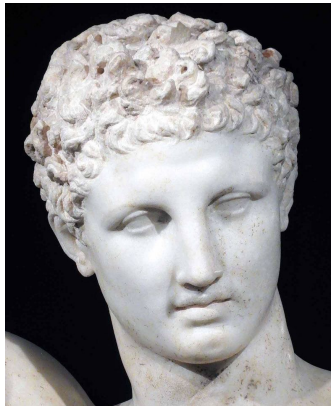


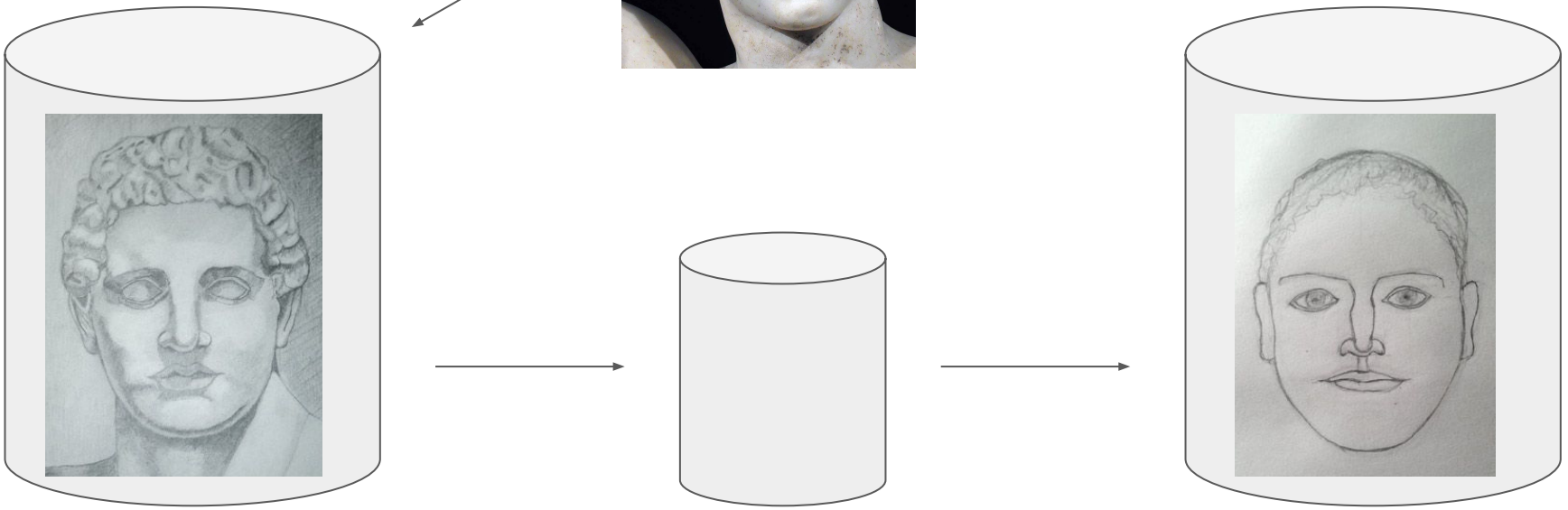


Lossless compression

STEINBUCH CENTRE FOR COMPUTING (SCC)
INSTITUTE FOR METEOROLOGY AND CLIMATE RESEARCH (IMK-ASF)

Compression IOI

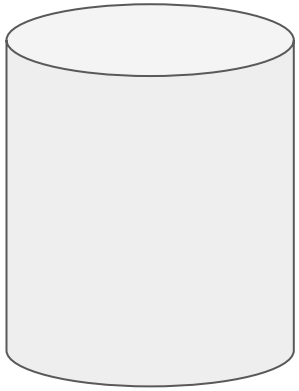




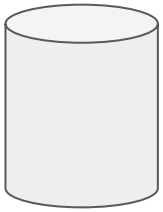
Lossy compression

Source [B]

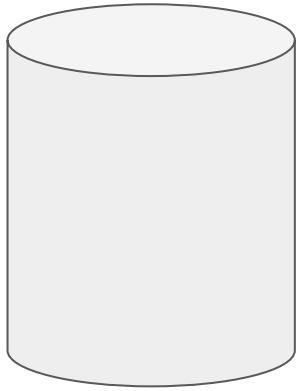
Nomenclature



Original data



Encoded data

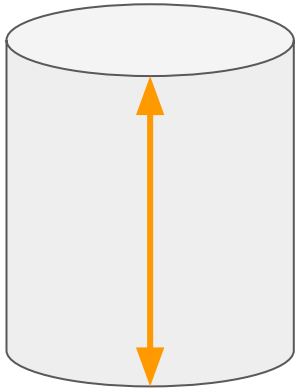


Decoded data

STEINBUCH CENTRE FOR COMPUTING (SCC)
INSTITUTE FOR METEOROLOGY AND CLIMATE RESEARCH (IMK-ASF)

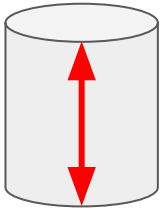
Nomenclature

Size of original data

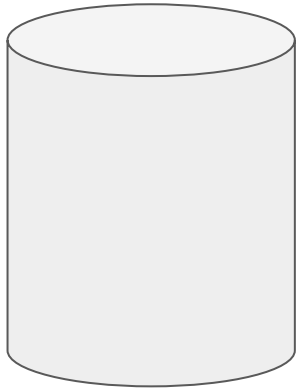


Original data

Size of encoded data



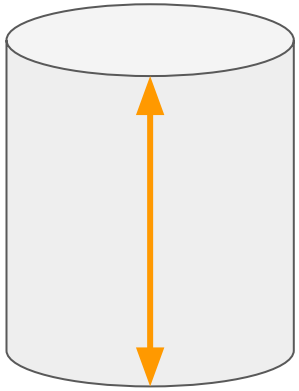
Encoded data



Decoded data

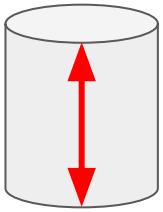
Nomenclature

Size of original data

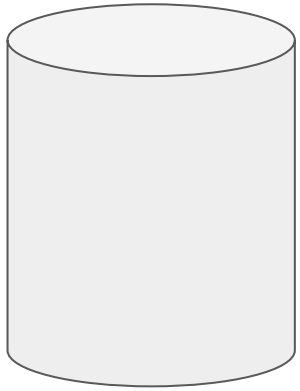


Original data

Size of encoded data



Encoded data



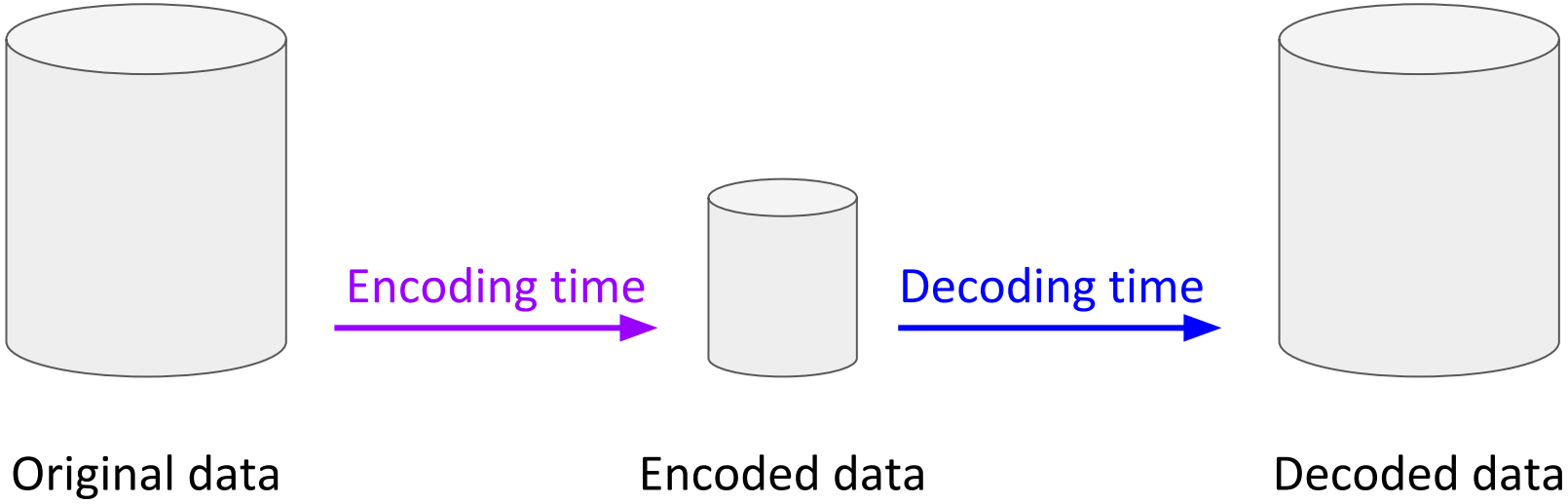
Decoded data

$$\frac{\text{Size of encoded data}}{\text{Size of original data}}$$

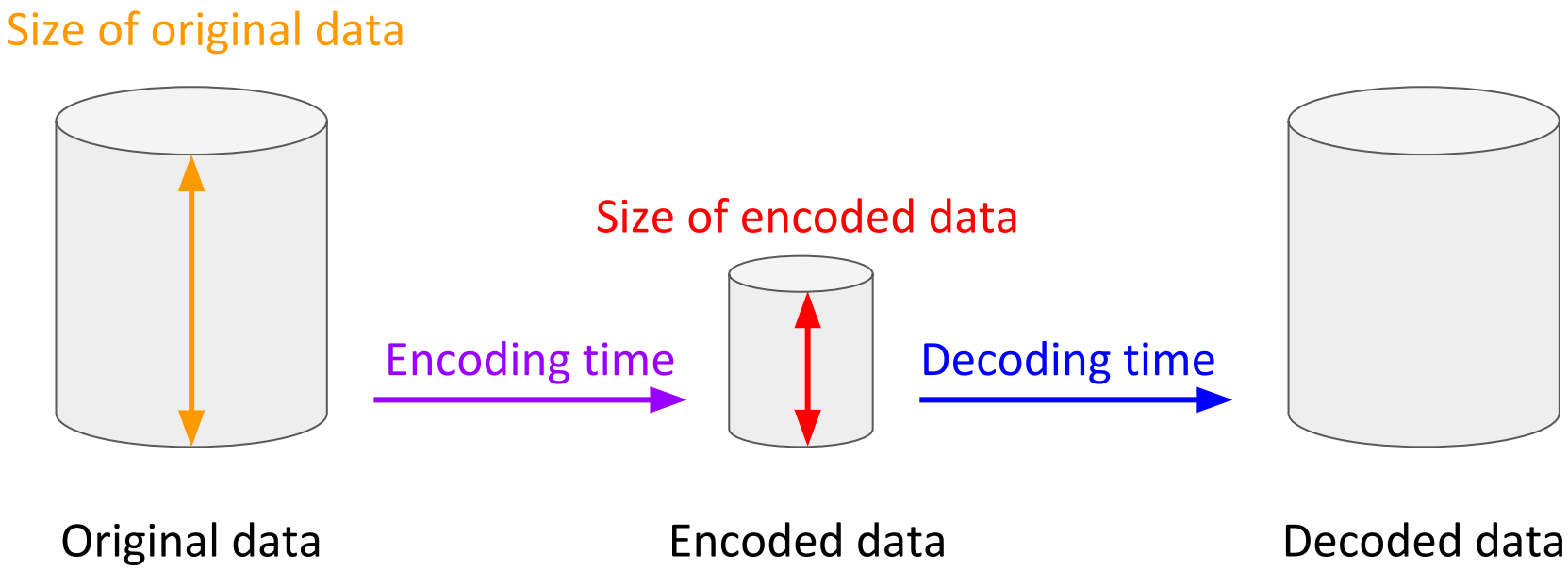
Compression Factor [no dimension]

STEINBUCH CENTRE FOR COMPUTING (SCC)
INSTITUTE FOR METEOROLOGY AND CLIMATE RESEARCH (IMK-ASF)

Nomenclature



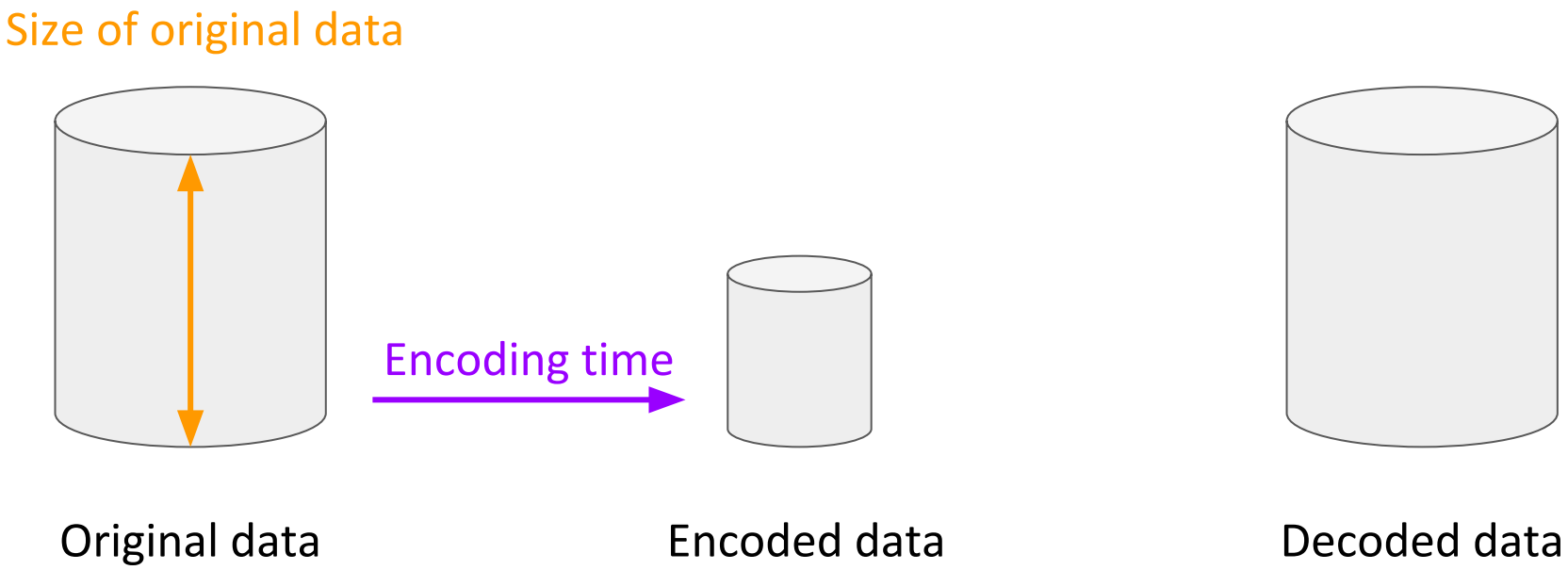
Nomenclature

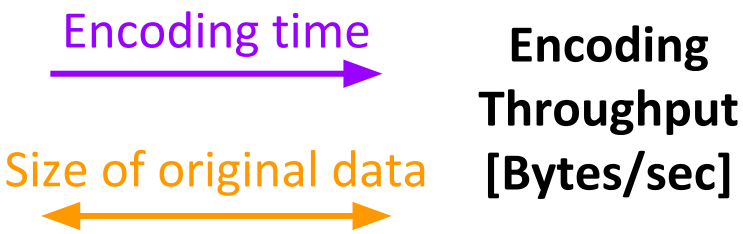


Encoding time / Size of original data → **Encoding Throughput [Bytes/sec]**

Decoding time / Size of encoded data → **Decoding Throughput [Bytes/sec]**

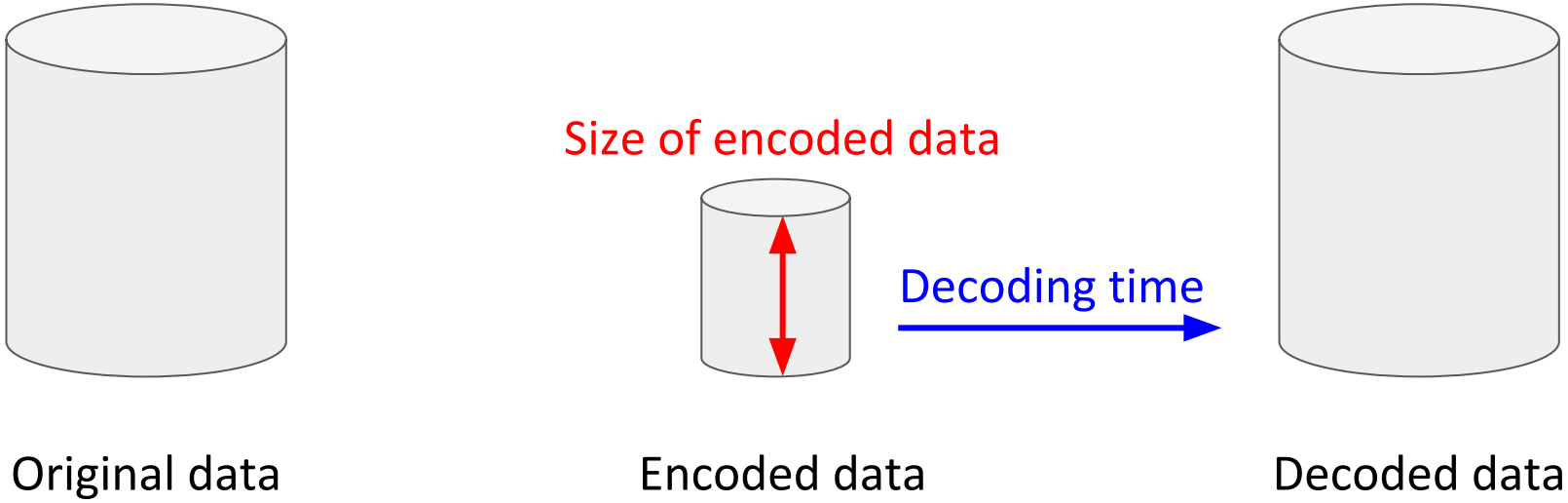
Nomenclature

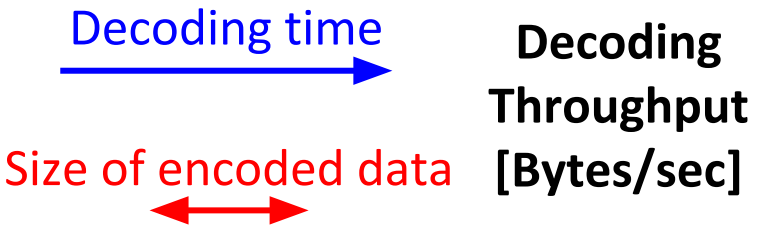




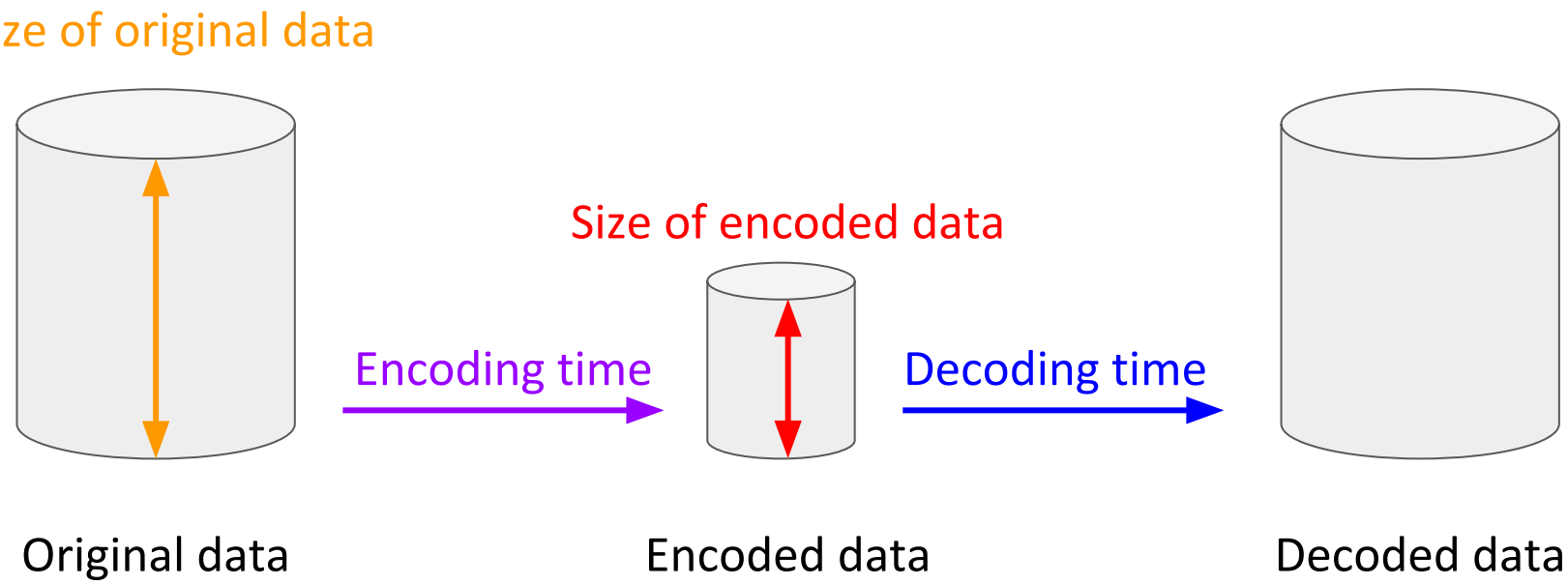
Encoding Throughput [Bytes/sec]

Nomenclature





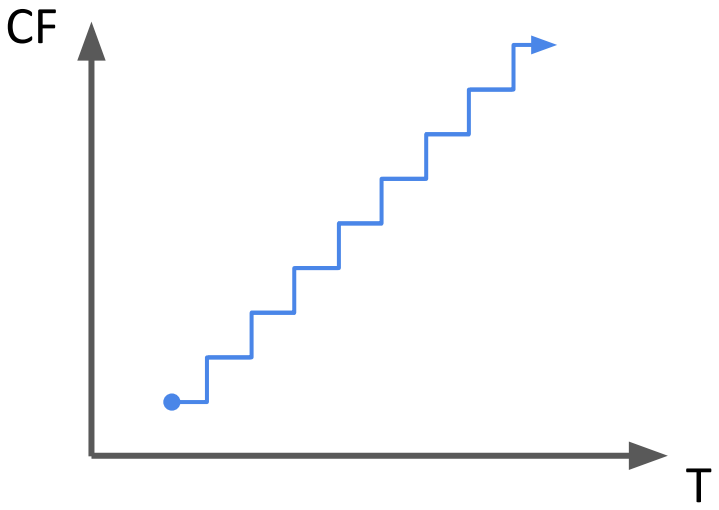
Nomenclature



Compression Factor (CF)
[no dimension]

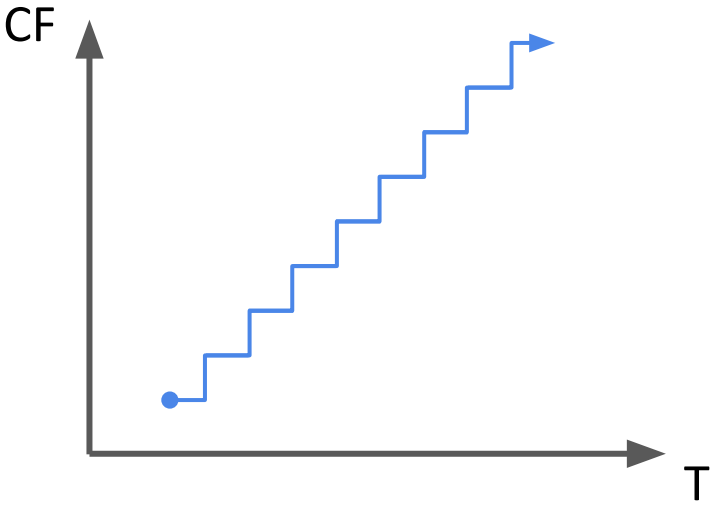
Throughput (T)
[Bytes/sec]

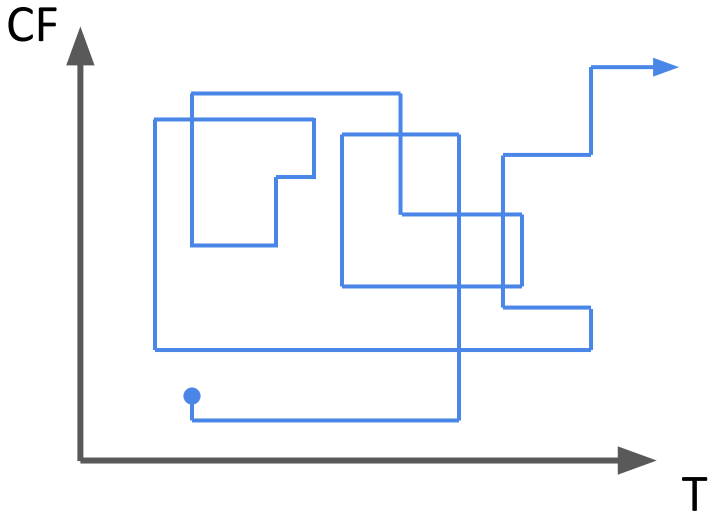
Iterative Process



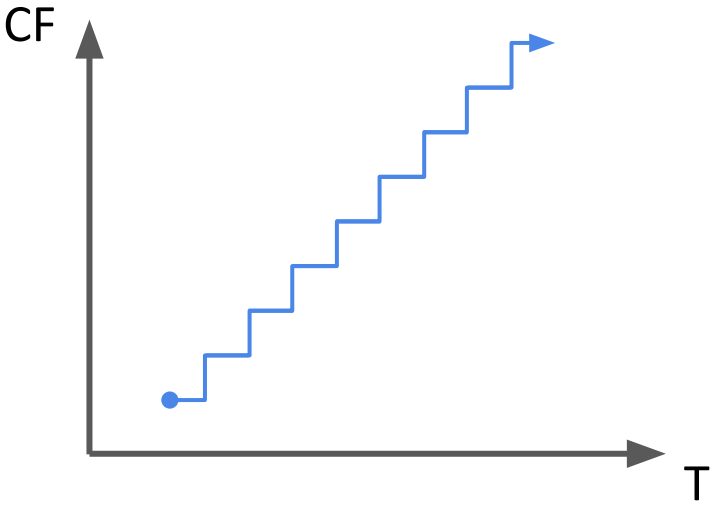
STEINBUCH CENTRE FOR COMPUTING (SCC)
INSTITUTE FOR METEOROLOGY AND CLIMATE RESEARCH (IMK-ASF)

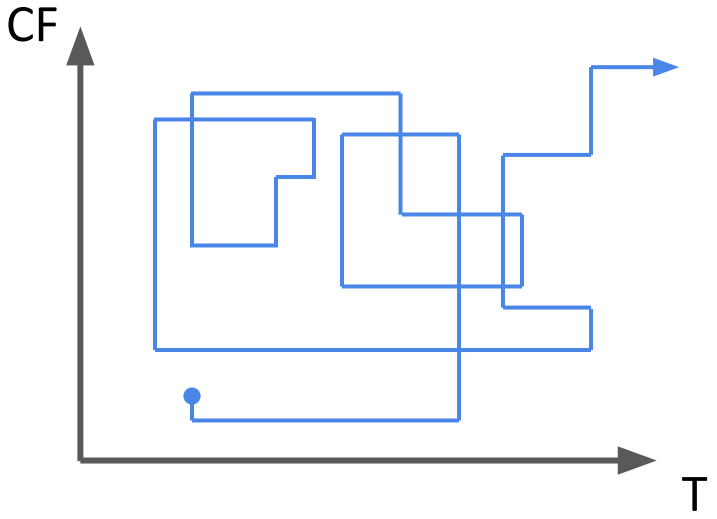
Iterative Process



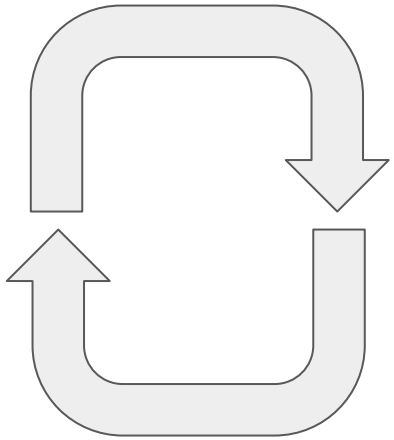


Iterative Process





Compression Factor (CF)
[no dimension]



Throughput (T)
[Bytes/sec]

Timeline

Past	Present	Future	Far Future
Compression Factor	Throughput	Compression Factor	Throughput
Understanding the information and intrinsic structure of the data	Research of data structures and algorithms	AI/ML methods for understanding of interdependencies	…
Custom scientific field	Applied computer science	Custom + related scientific field	…

Timeline

Past	Present	Future	Far Future
Compression Factor	Throughput	Compression Factor	Throughput
Understanding the information and intrinsic structure of the data	Research of data structures and algorithms	AI/ML methods for understanding of interdependencies	…
Domain knowledge	Applied computer science	Custom + related scientific field	…

STEINBUCH CENTRE FOR COMPUTING (SCC)
INSTITUTE FOR METEOROLOGY AND CLIMATE RESEARCH (IMK-ASF)

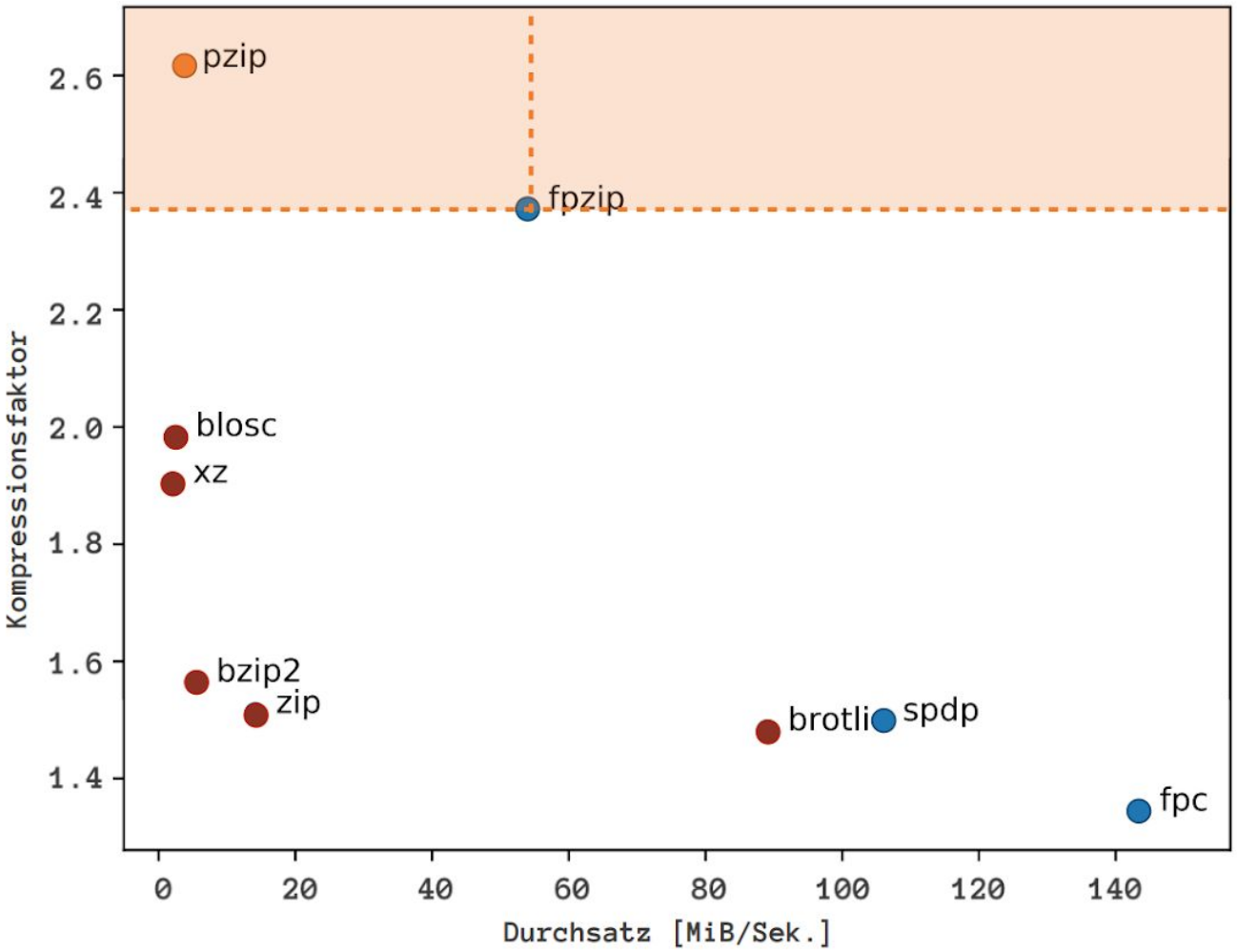
Timeline

Past

Compression Factor

Understanding the information and intrinsic structure of the data

Domain knowledge

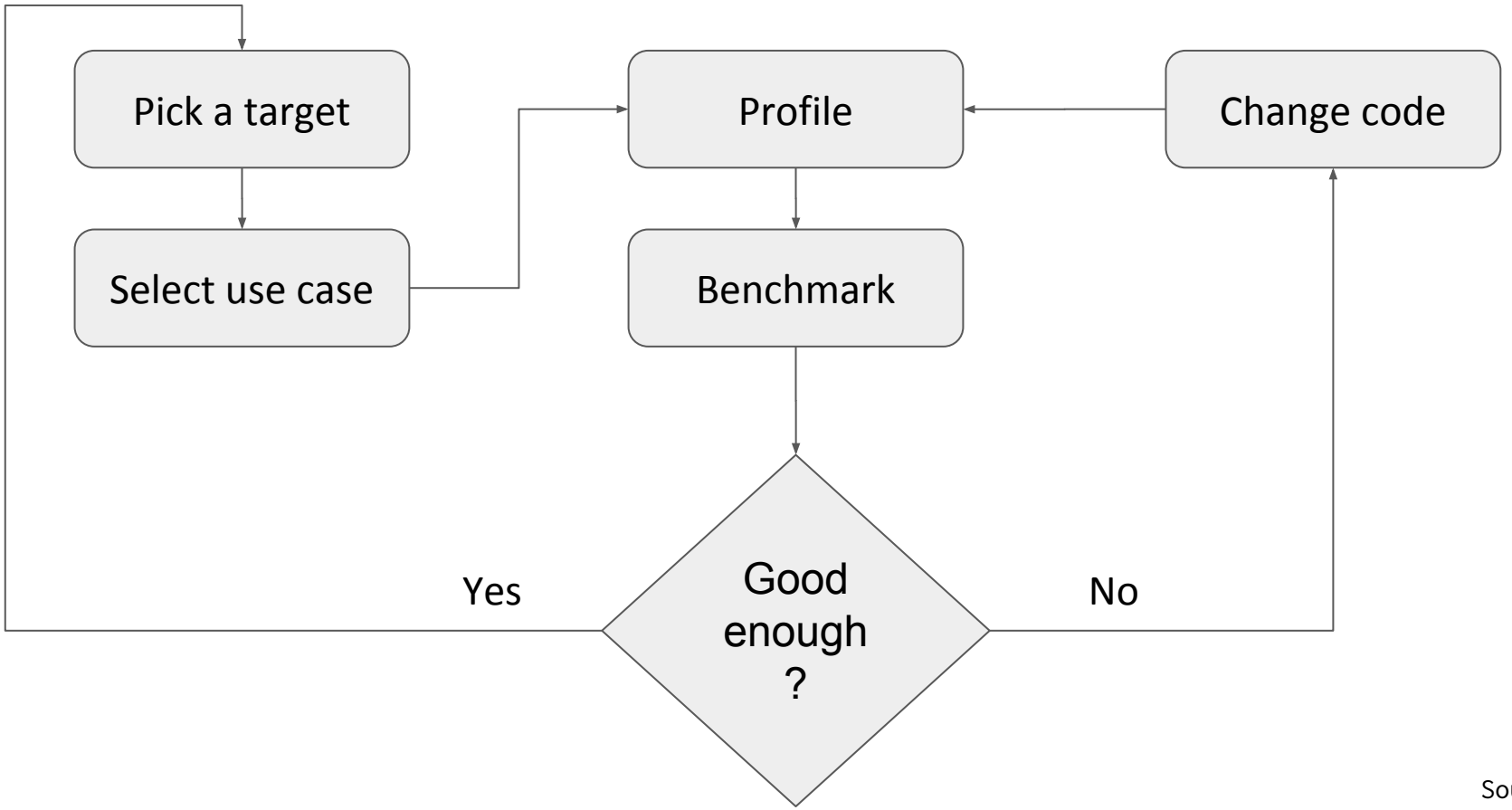


Timeline

Past	Present	Future	Far Future
Compression Factor	Throughput	Compression Factor	Throughput
Understanding the information and intrinsic structure of the data	Research of data structures and algorithms	AI/ML methods for understanding of interdependencies	…
Domain knowledge	Applied computer science	Custom + related scientific field	…

STEINBUCH CENTRE FOR COMPUTING (SCC)
INSTITUTE FOR METEOROLOGY AND CLIMATE RESEARCH (IMK-ASF)

Workflow for optimizing throughput or compression factor



Source [C]

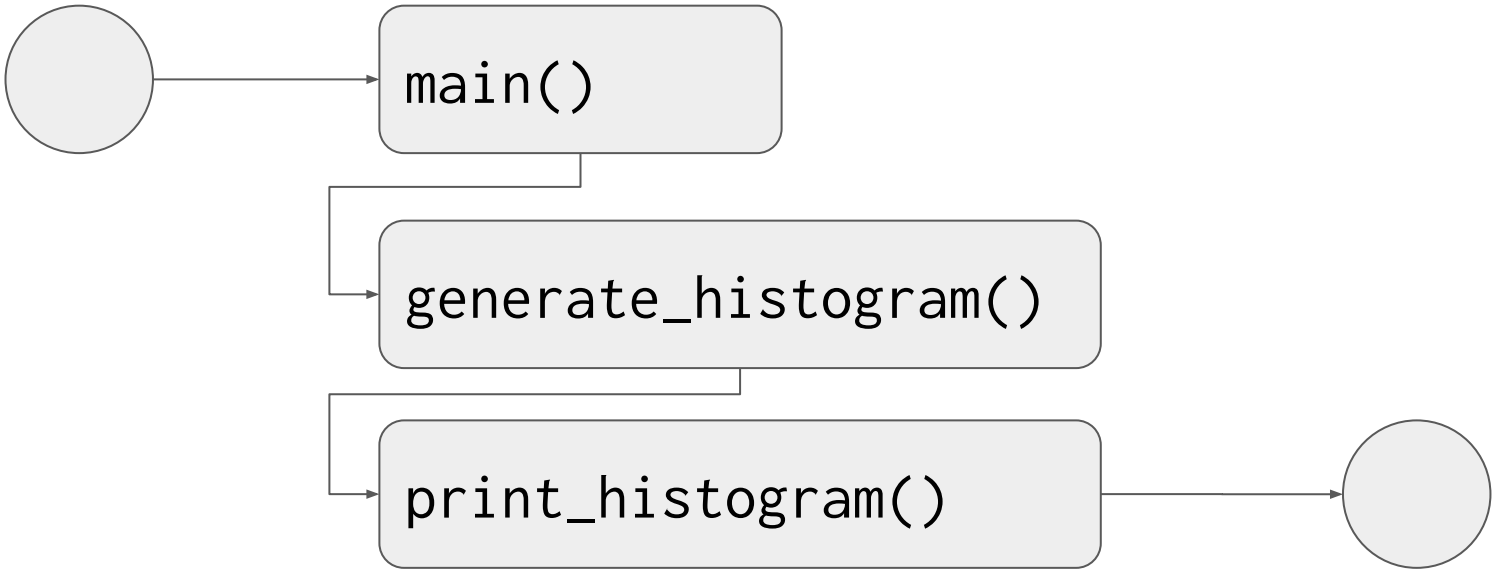
Profiling and Benchmarking in this context

Karlsruhe Institute of Technology

- Profiling: Analysis of the **program as a whole** (might be split down to function level)
- Benchmarking: Comparison of **evolutions/versions** of functions

Profiling and Benchmarking in this context

- Profiling: Analysis of the **program as a whole** (might be split down to function level)
- Benchmarking: Comparison of **evolutions/versions** of functions



Profiling and Benchmarking in this context

- Profiling: Analysis of the **program as a whole** (might be split down to function level)
- Benchmarking: Comparison of **evolutions/versions** of functions

```
generate_histogram()
```

```
fast_histogram()
```

```
instant_histogram()
```


STEINBUCH CENTRE FOR COMPUTING (SCC)
INSTITUTE FOR METEOROLOGY AND CLIMATE RESEARCH (IMK-ASF)

Profiling and Benchmarking in this context

Karlsruhe Institute of Technology

- Profiling: Analysis of the **program as a whole** (might be split down to function level)
  - `hyperfine` (time)
  - `valgrind` (time + memory)
  - `gnu time` (time + memory)
- Definitely worth looking into …
  - `flamegraph`
  - `perf`
  - `*not-perf`
  - `*cargo-instruments`

based on [C]

Profiling and Benchmarking in this context

Karlsruhe Institute of Technology

- Benchmarking: Comparison of **evolutions/versions** of functions
  - Very specific for each programming language
  - No general recommendation possible
  - Rust: `criterion`

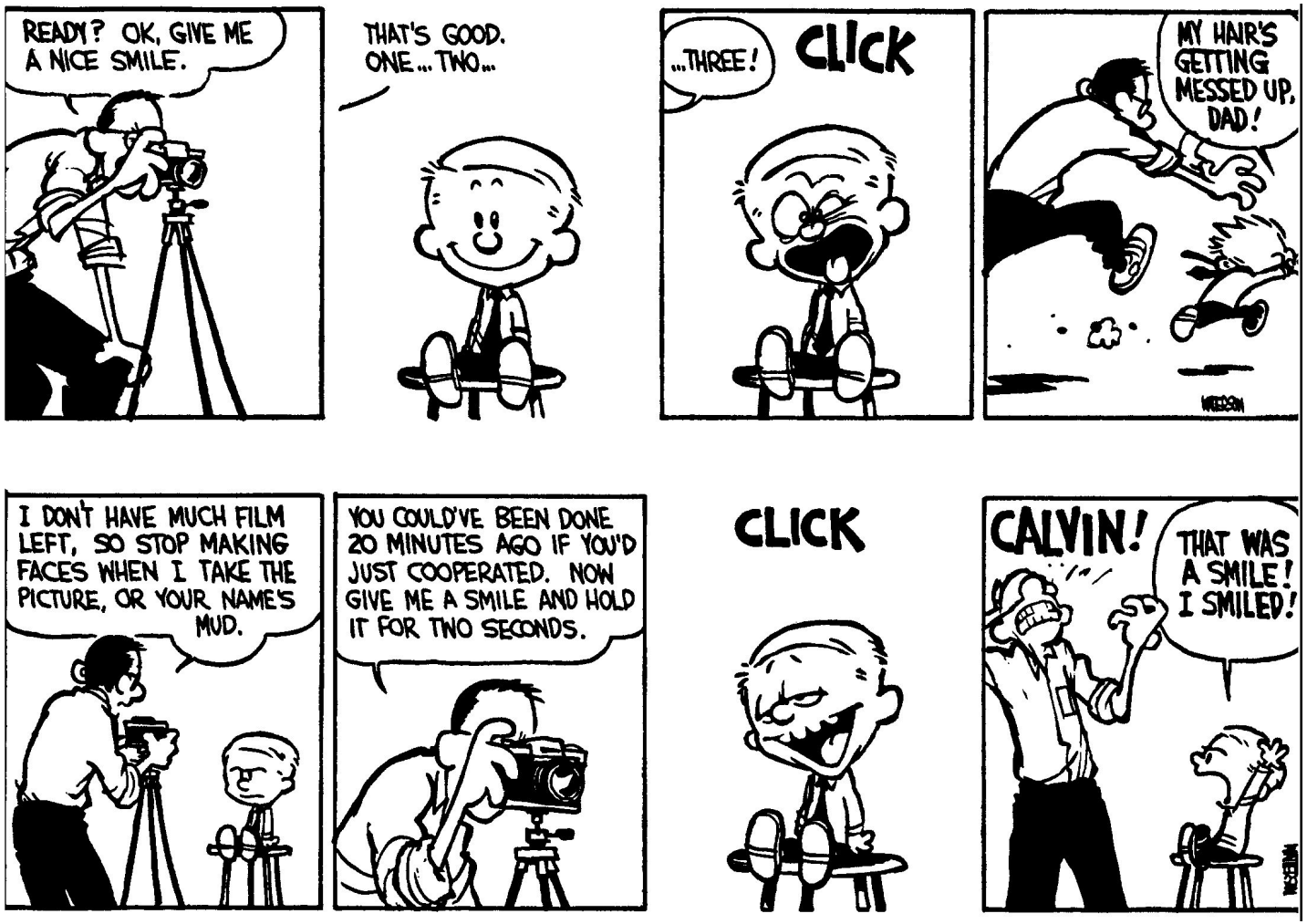
Demo time

Karlsruhe Institute of Technology

- Profile I/O of my laptop
- First generate some random data
- Copy this data from a to b
- Profile different copy commands
  - cp
  - dd
  - rsync
- Applications we will use
  - hyperfine (**time**)
  - valgrind (time + **memory**)
  - gnu time (time + **memory**)

STEINBUCH CENTRE FOR COMPUTING (SCC)
INSTITUTE FOR METEOROLOGY AND CLIMATE RESEARCH (IMK-ASF)

That’s all folks



Source [Z]

Sources

Karlsruhe Institute of Technology

[Z] Calvin Photoshoot
[A] Greece Olympia Marble Statue
[B] https://drawingacademy.com/drawing-progress
[C] https://fosdem.org/2020/schedule/event/rust_optimizing_rav1e/