# WP7 Task 3 Current Activities

—A Warm-Up—

Peter Steinbach[1], Felicita Purnama Dewi Gernhardt [1]

*[1] Helmholtz-Zentrum Dresden-Rossendorf, Core Facility for Digital Infrastructure and Computing*

11th October 2021

HZDR

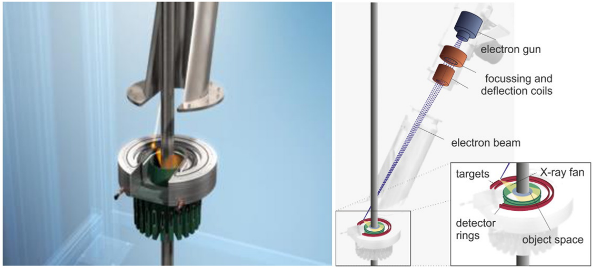
HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF

# Our Mandate

Adopt new algorithms and technologies for lossless and lossy data compression

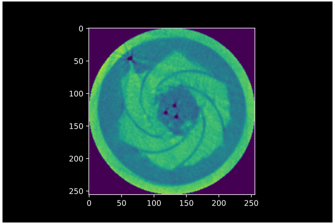
- Looking for collaborators!
- hired RSE/data scientist
- started to look at first datasets

# Data prototype: ROFEX



- Ultrafast electron beam X-ray computed tomography
- non-invasive investigation of dynamic processes
- electron beam is focussed towards a circular target
- periodically deflected with high frequency

# ROFEX details



one timepoint: $256 \times 256 \times 12500$ of `uint16` voxel intensities

- ROFEX-III raw data: $\approx 2\ GByte/s$
- one measurement campaign:
  25-50 samples of 15 $s$ each
- per year: max 10 campaigns
  15 $TB/year$
- reconstructed data $\approx$ raw data
- reconstructed data as `fxv` File Format

# Lossy Experiments: LibAPR[1]

https://github.com/AdaptiveParticles/LibAPR

- Library for producing and processing on the Adaptive Particle Representation (APR)
- APR replaces pixels with particles
- Particles are a generalization of pixels:
  - Points in space that carry intensity
  - Can be places wherever image content requires
  - May have different sizes in different parts of the image (size define resolution)

[1] possible alternatives: https://www.computationalimaging.org/publications/acorn/

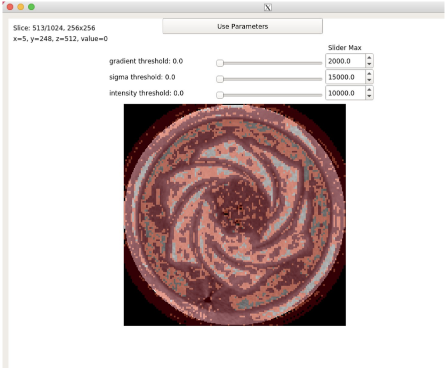
# ROFEX: rotating shovel

https://github.com/AdaptiveParticles/LibAPR

- Library for producing and processing on the Adaptive Particle Representation (APR)
- APR replaces pixels with particles
- Particles are a generalization of pixels:
  - Points in space that carry intensity
  - Can be places wherever image content requires
  - May have different sizes in different parts of the image (size define resolution)

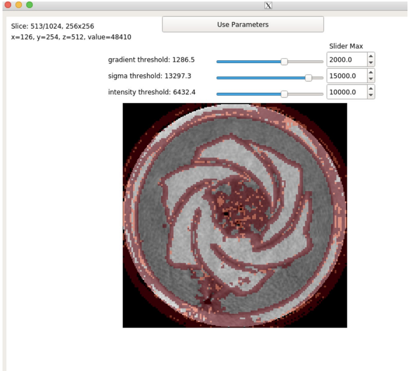
# LibAPR on ROFEX: chaos is troubling

## PyLibAPR - Compression



Slice: 513/1024, 256x256
x=5, y=248, z=512, value=0

gradient threshold: 0.0 2000.0
sigma threshold: 0.0 15000.0
intensity threshold: 0.0 10000.0



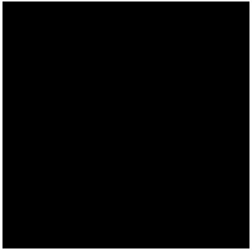
Slice: 513/1024, 256x256
x=126, y=254, z=512, value=48410

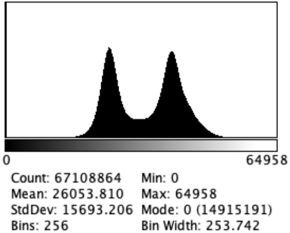
gradient threshold: 1286.5 2000.0
sigma threshold: 13297.3 15000.0
intensity threshold: 6432.4 10000.0

Size of APR file: 55.1MB

# LibAPR on ROFEX: intensities

## PyLibAPR - Decompression

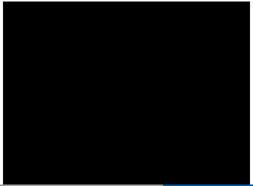


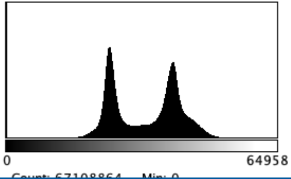


```
Count: 67108864    Min: 0
Mean: 26053.810    Max: 64958
StdDev: 15693.206  Mode: 0 (14915191)
Bins: 256          Bin Width: 253.742
```

Original tiff

134.4MB





```
Count: 67108864    Min: 0
```

Decompressed APR

APR: 55.1MB

# LibAPR on ROFEX: Current Status of Lossy Compression

- currently working with downstream pipeline
- need to find metrics/observables/checks that show how we broke the dataset
- LibAPR = part of a pipeline
- how to encode pipeline?
  - software: hdf5 vs zarr vs custom
  - data curation: which format will exist in 5-10-15 years?
  - decompress pipeline on any OS

# Talking about Datasets?

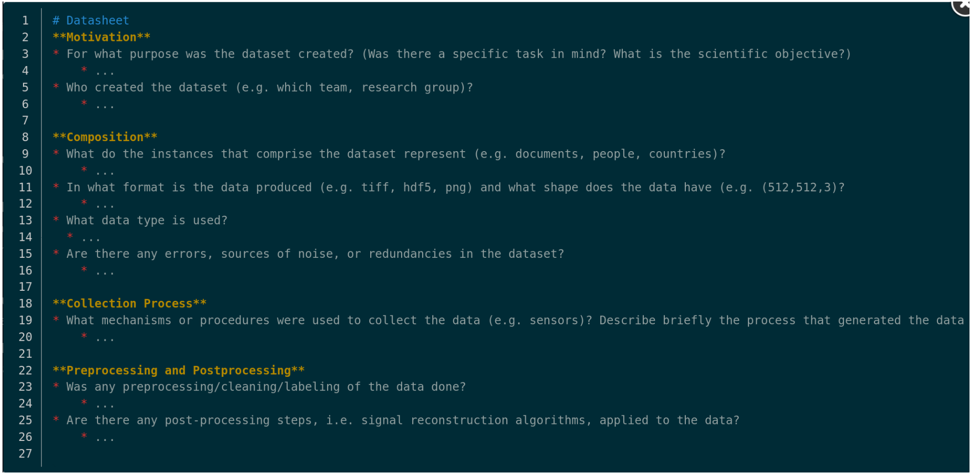
## Challenges

- bridging the communication gap: RSE / data scientists <-> domain scientist
- clearing out misconceptions, expectations and reinventions
- identify domain specific terminology

## Advantages

LEAPS community has datasets available:
https://zenodo.org/record/4558708#.YMhzwSaxVrM

# Mini Data-Sheets[2] by Felicata Gernhardt



```
# Datasheet
**Motivation**
* For what purpose was the dataset created? (Was there a specific task in mind? What is the scientific objective?)
  * ...
* Who created the dataset (e.g. which team, research group)?
  * ...

**Composition**
* What do the instances that comprise the dataset represent (e.g. documents, people, countries)?
  * ...
* In what format is the data produced (e.g. tiff, hdf5, png) and what shape does the data have (e.g. (512,512,3)?
  * ...
* What data type is used?
 * ...
* Are there any errors, sources of noise, or redundancies in the dataset?
  * ...

**Collection Process**
* What mechanisms or procedures were used to collect the data (e.g. sensors)? Describe briefly the process that generated the data
  * ...

**Preprocessing and Postprocessing**
* Was any preprocessing/cleaning/labeling of the data done?
  * ...
* Are there any post-processing steps, i.e. signal reconstruction algorithms, applied to the data?
  * ...

```

[2] inspired by https://arxiv.org/abs/1803.09010

# Summary

- started to work on datasets close to us (local, Expands dataset)
- technology, methods and social challenges
- identification of downstream goals/quality/performance crucial

Questions, Comments or Feedback ?!
Welcoming collaborators!