MPI-Parallel Machine Learning Algorithms for the Analysis of High-Speed Video Data

ECCOMAS Congress 2022
June 5th – 9th 2022

Alexander Rüttgers
Institute for Software Technology
German Aerospace Center (DLR)

Joint work with Anna Petrarolo
and Philipp Knechtges (all DLR)

DLR

Outline

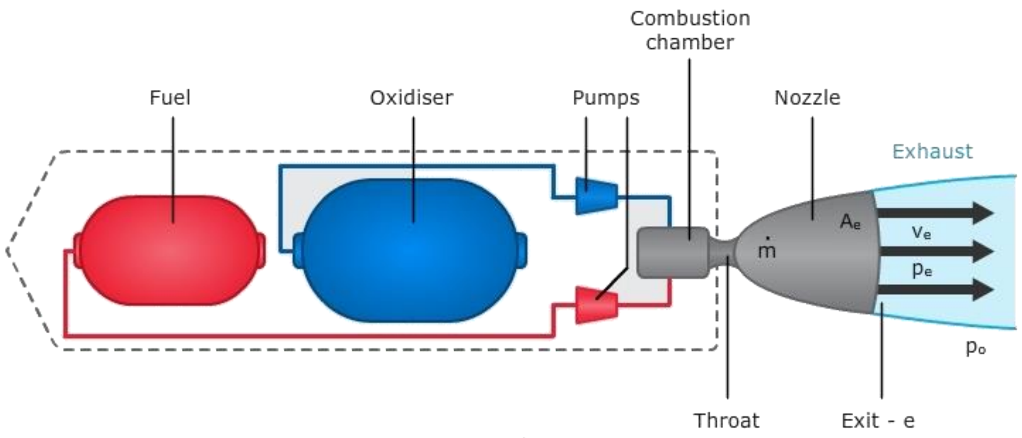
1. Rocket engine combustion analysis at DLR

2. Helmholtz Analytics Toolkit (Heat) for distributed ML

3. Results
   a) Spectral Clustering
   b) Anomaly Detection

Rocket engine combustion analysis

- **Aim:** Cost reduction of rocket engines, be competitive with e.g. Space-X

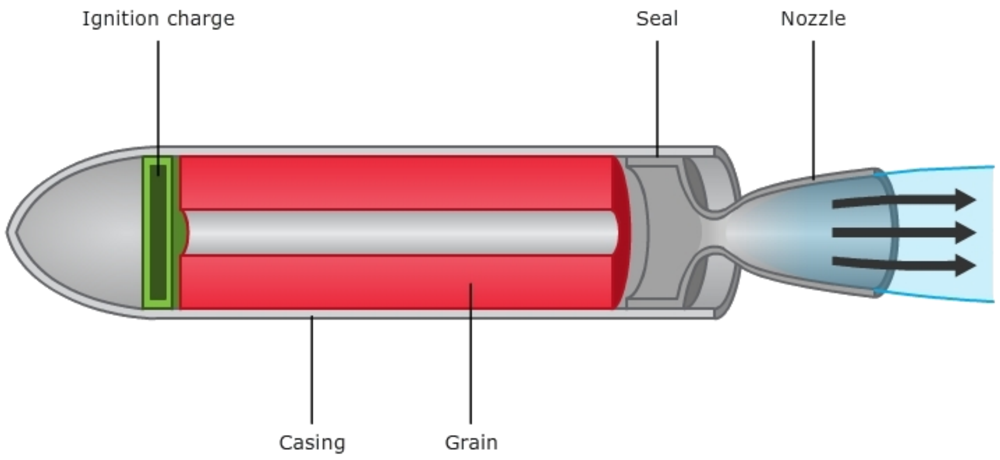


Traditional liquid rocket engine:

- 2 pumps transporting fluid fuel and oxidizer at very high pressure and flow
- Advantages
  - Burning rate can be controlled precisely
- Disadvantages
  - Pumps are mechanically very complex
  - Expensive

Rocket engine combustion analysis

- **Aim:** Cost reduction of rocket engines, be competitive with e.g. Space-X

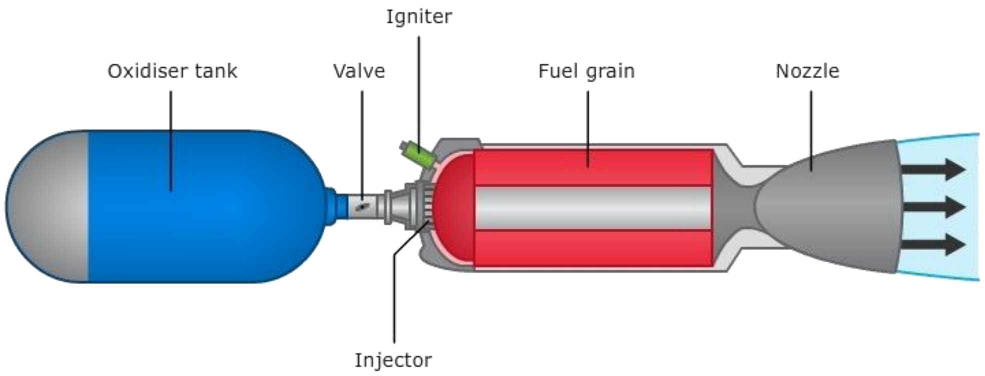


Solid propellant rocket engine

- Fuel and oxidizer are mixed in solid form
- Advantage
  - Cheap
- Disadvantage
  - Burning rate can not be varied during flight

Rocket engine combustion analysis

- **Aim:** Cost reduction of rocket engines, be competitive with e.g. Space-X



Hybrid rocket engine

- Pressurized fluid oxidizer
- Solid fuel
- A valve controls, how much oxidizer gets into the combustion chamber

- Advantages
  - Cheap
  - Controllable

Experiments on new hybrid rocket fuels at DLR

- DLR investigates new hybrid rocket fuels on a paraffin basis at Institute of Space Propulsion in Lampoldshausen.
- About 300 combustion tests were performed with single-slab paraffin-based fuel with 20° forward facing ramp angle + gaseous oxygen.
- Combustion is captured with high-speed video camera with 10 000 frames / second

Fig. 1: Fuel slap configuration before (top) and after (bottom) combustion test

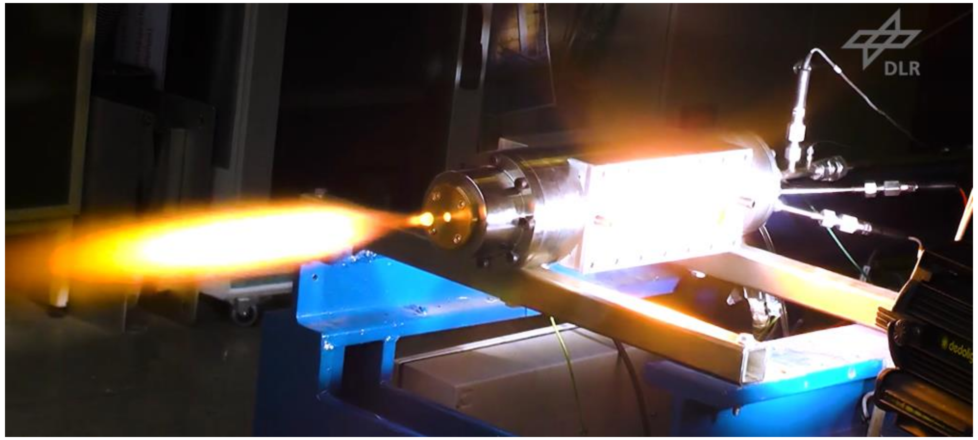


Fig. 2: Hybrid rocket engine combustion chamber

Video extract of test 284	fuel	oxidizer mass flow	CH*-filter	duration
Ignition, steady combustion, extinction	pure paraffin 6805	50 *g*/*s*,	yes, i.e. only wavelengths emitted from CH* are filmed	3 s = 30 000 frames

Outline

1. Rocket engine combustion analysis at DLR
2. Helmholtz Analytics Toolkit (Heat) for distributed ML
3. Results
   a) Spectral Clustering
   b) Anomaly Detection

Heat

- Heat = Helmholtz Analytics Toolkit

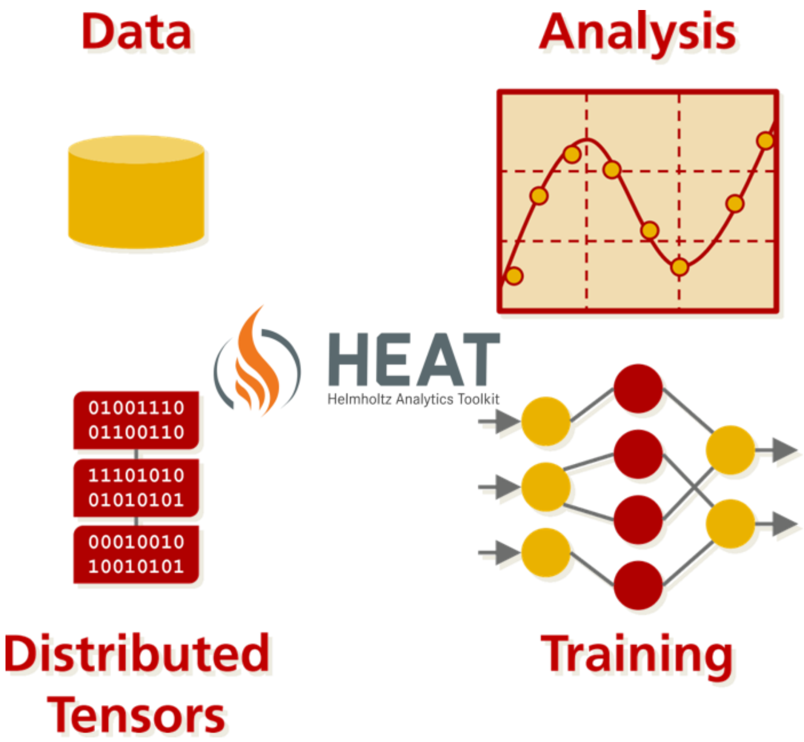
- Developed by three Helmholtz research organizations in Germany:
    - Research Center Juelich (FZJ)
    - Karlsruhe Institute of Technology (KIT)
    - German Aerospace Center (DLR)

- Python library for **parallel**, **distributed** data analytics and machine learning

- **Aim:** Bridge data analytics and **high-performance computing**

- Open Source licensed, MIT

https://github.com/helmholtz-analytics/heat



Scope

Facilitating analysis of Helmholtz applications

Bringing HPC and Machine Learning / Data Analytics closer together

Ease of use

Design

HEAT
Helmholtz Analytics Toolkit

- k-means
- SVM
- Deep Learning

And more machine learning algorithms

PyTorch

- Tensor Linear Algebra
- Automatic Differentiation
- NumPy-like interface
- GPU support

mpi4torch

- Automatically Differentiated Node-Node Communication using MPI

Core Idea: Data Distribution

HeAT Tensor

Server#1 PyTorch Tensor#1	Server#2 PyTorch Tensor#2	Server#3 PyTorch Tensor#3

split=1

Server#1 PyTorch Tensor#1
Server#2 PyTorch Tensor#2
Server#3 PyTorch Tensor#3

HeAT Tensor

split=0

Example:

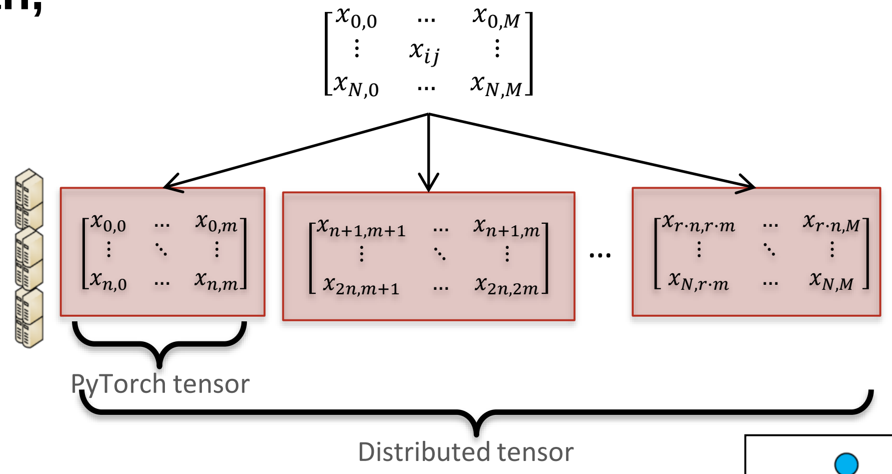
```
import heat as ht
# construct a range tensor
>>> range_data = ht.arange(6, split=1)
```

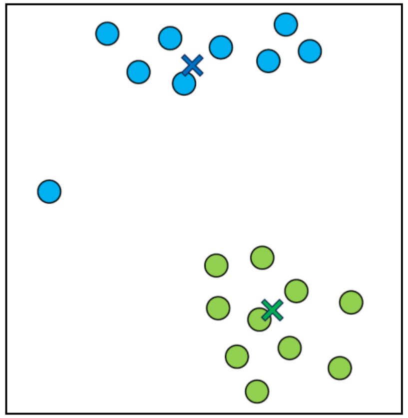
Server#1 [0, 1]	Server#2 [2, 3]	Server#3 [4, 5]

```
>>> range_data.mean()
2.5
>>> range_data.argmax()
5
```

Functionality achieved

- Implementation of a **distributed parallel tensor math,** NumPy-compatible, based on PyTorch
- Some linear algebra routines
- **Parallel data I/O** via HDF 5 and NetCDF
- Development of **mpi4torch** to enable **automatic differentiation** of distributed PyTorch code
- Multiple methods (clustering, classification, regression)
- Data-parallel training of neural nets (new DASO algorithm)



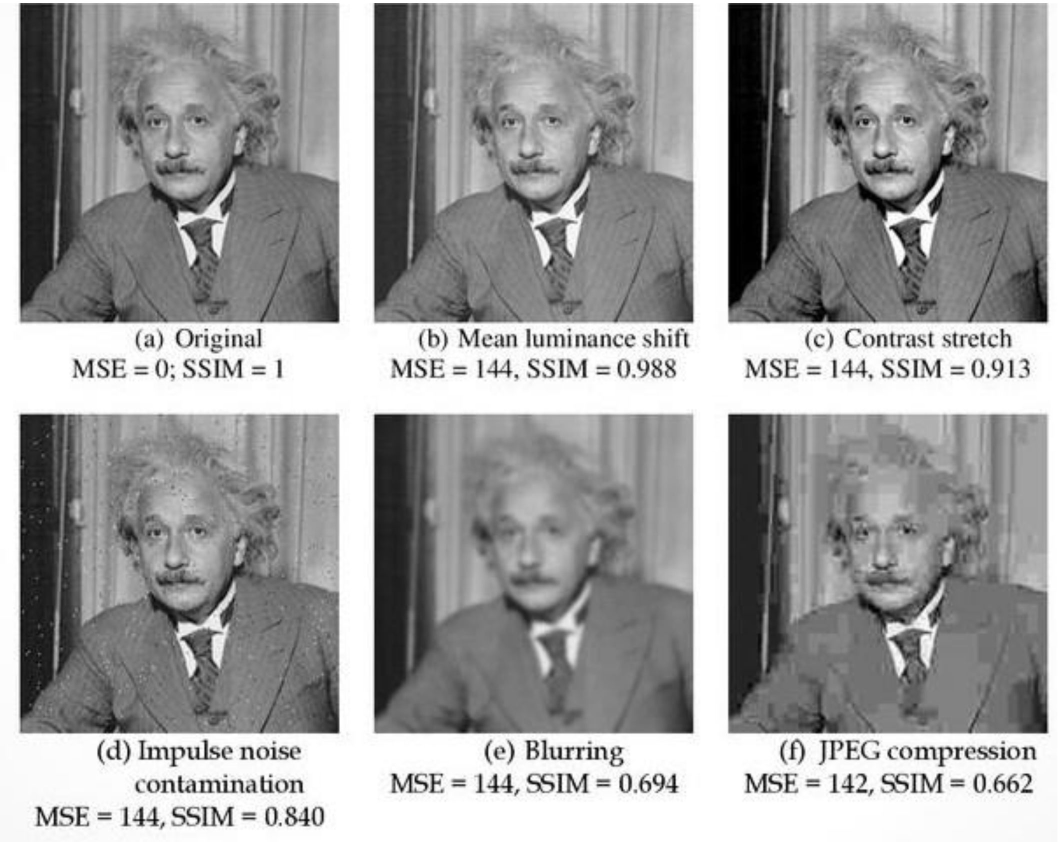


Outline

1. Rocket engine combustion analysis at DLR
2. Helmholtz Analytics Toolkit (HeAT) for distributed ML
3. Results
   a) Spectral Clustering
   b) Anomaly Detection

Dissimilarity measure for image data

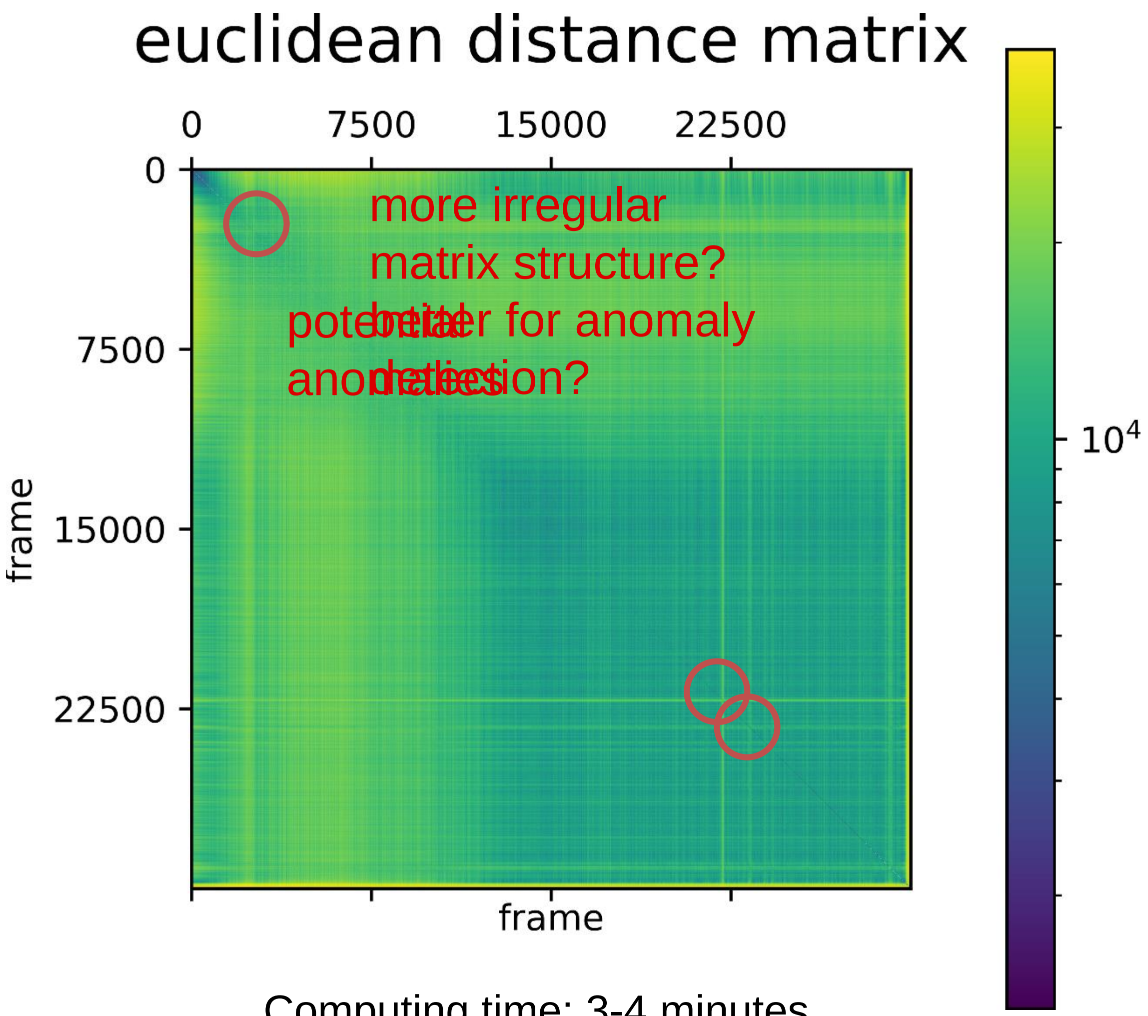
- Algorithms often require pairwise dissimilarity of images (matrix of size nr_of_images x nr_of_images).
- Standard approaches such as mean squared error (MSE) / discrete L^2-norm often differ from human recognition.
- Advanced dissimilarity measures such as structural similarity (SSIM) often perform better (considers luminance, contrast and structure) but are much more expensive.
- Structural similarity (SSIM)/ structural dissimilarity (DSSIM) is not a distance metric.



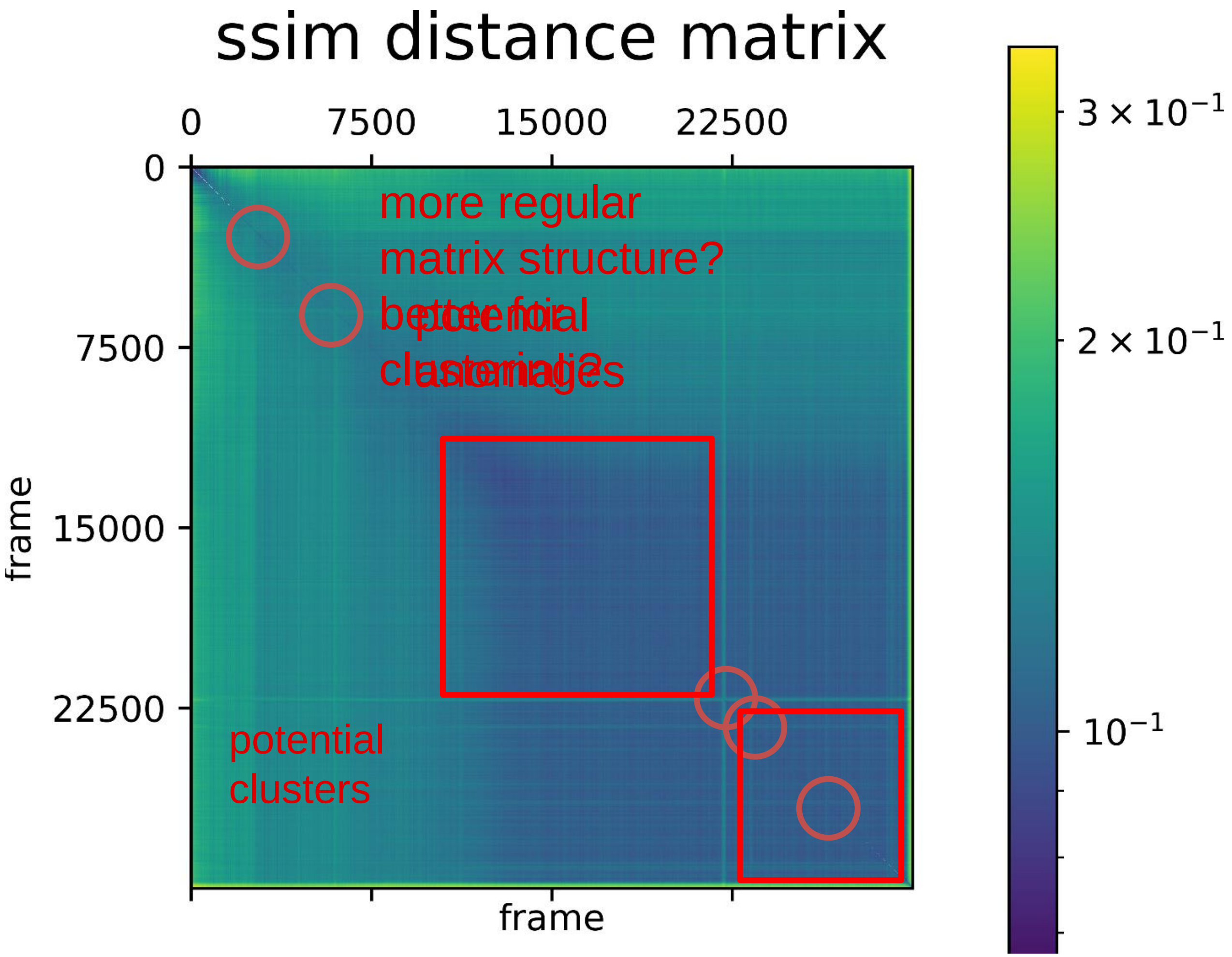
Example: (b)-(f) with same MSE, SSIM decreases*

*https://nsf.gov/news/mmg/mmg_disp.jsp?med_id=79419&from=

Pairwise distance matrices for test 284

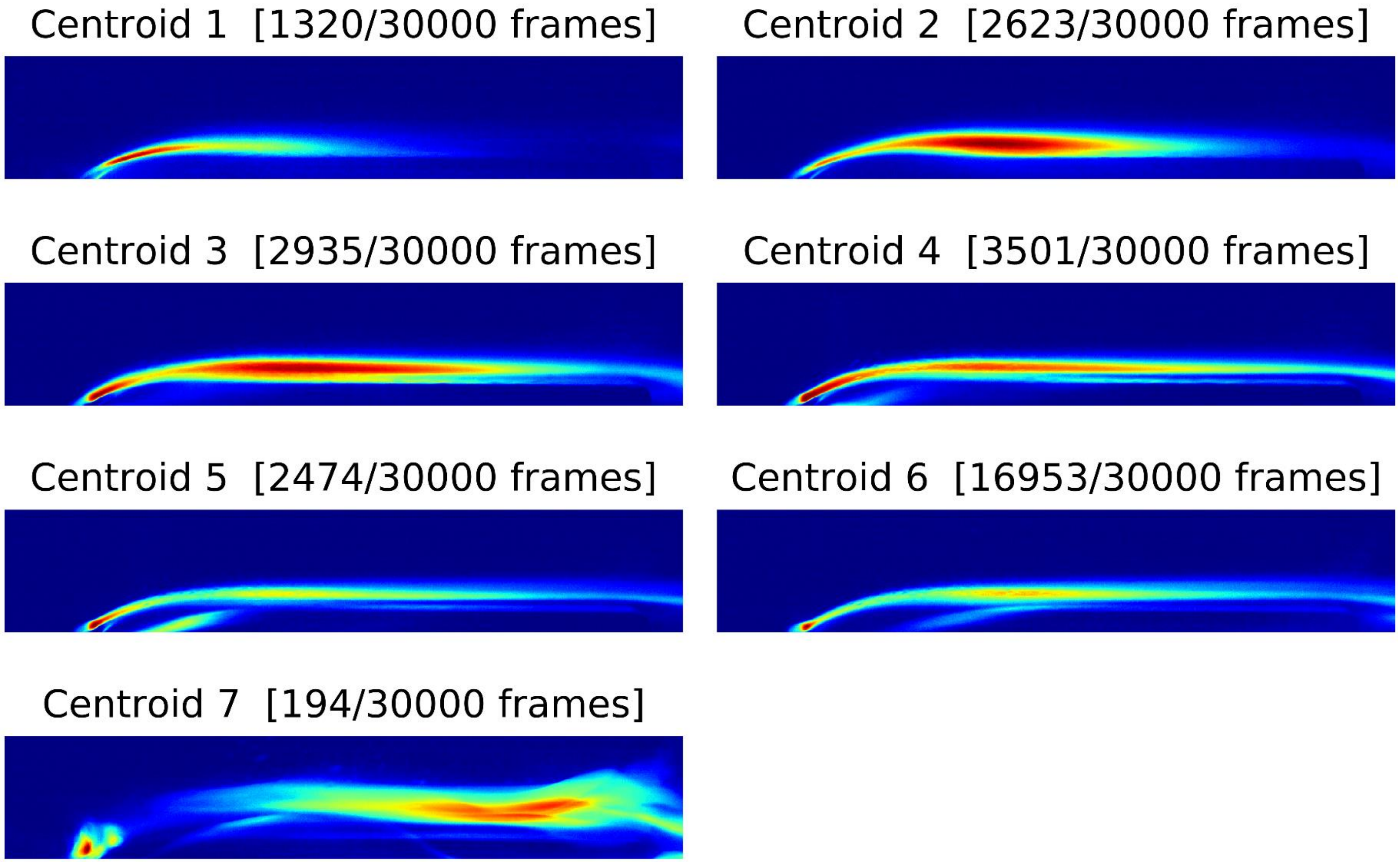


Computing time: 3-4 minutes



Computing time: 5 days (OpenMP parallel, 56 cores)
one comparison ≈ 0.1 s (scikit-image)

Spectral Clustering of test 284

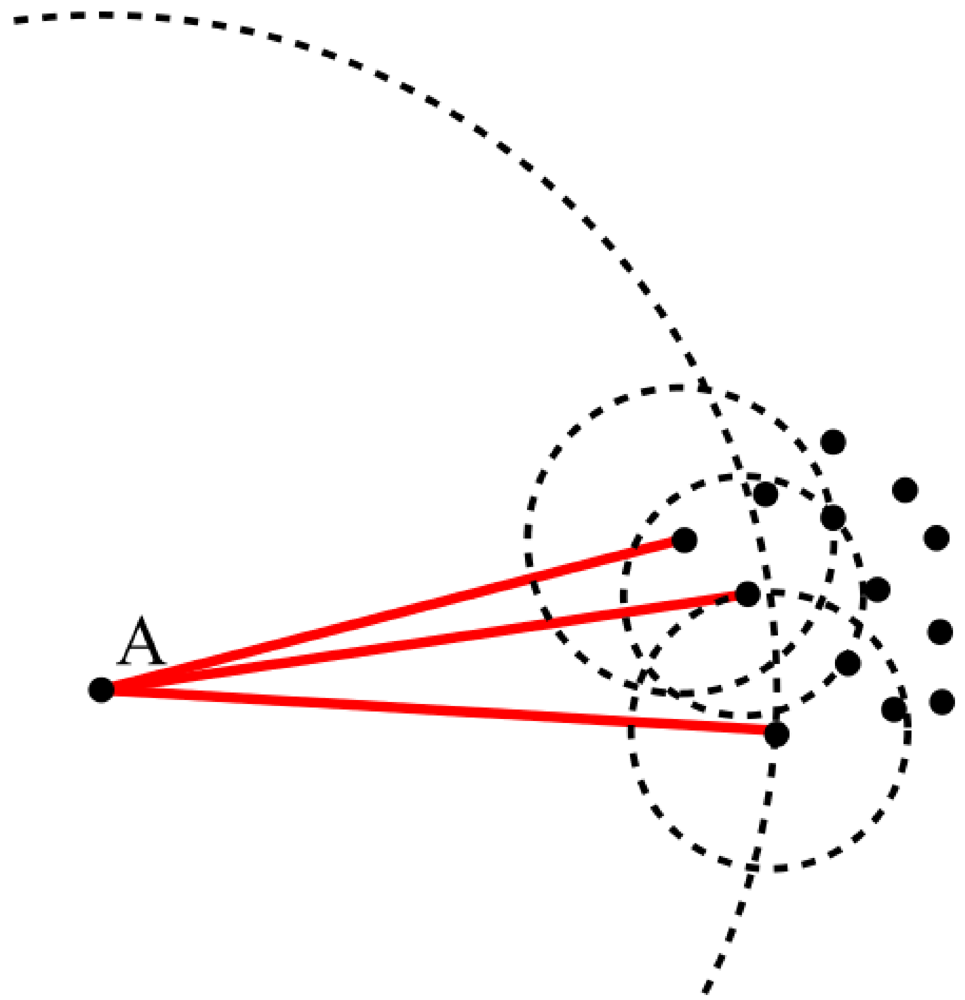


- **Fig. 1**: Results of spectral clustering with ssim affinity matrix.
- Using an Euclidean affinity matrix leads to a separation of the extinction phase into various clusters.
- Note that the number of clusters *k* is a hyperparameter of the clustering algorithm.

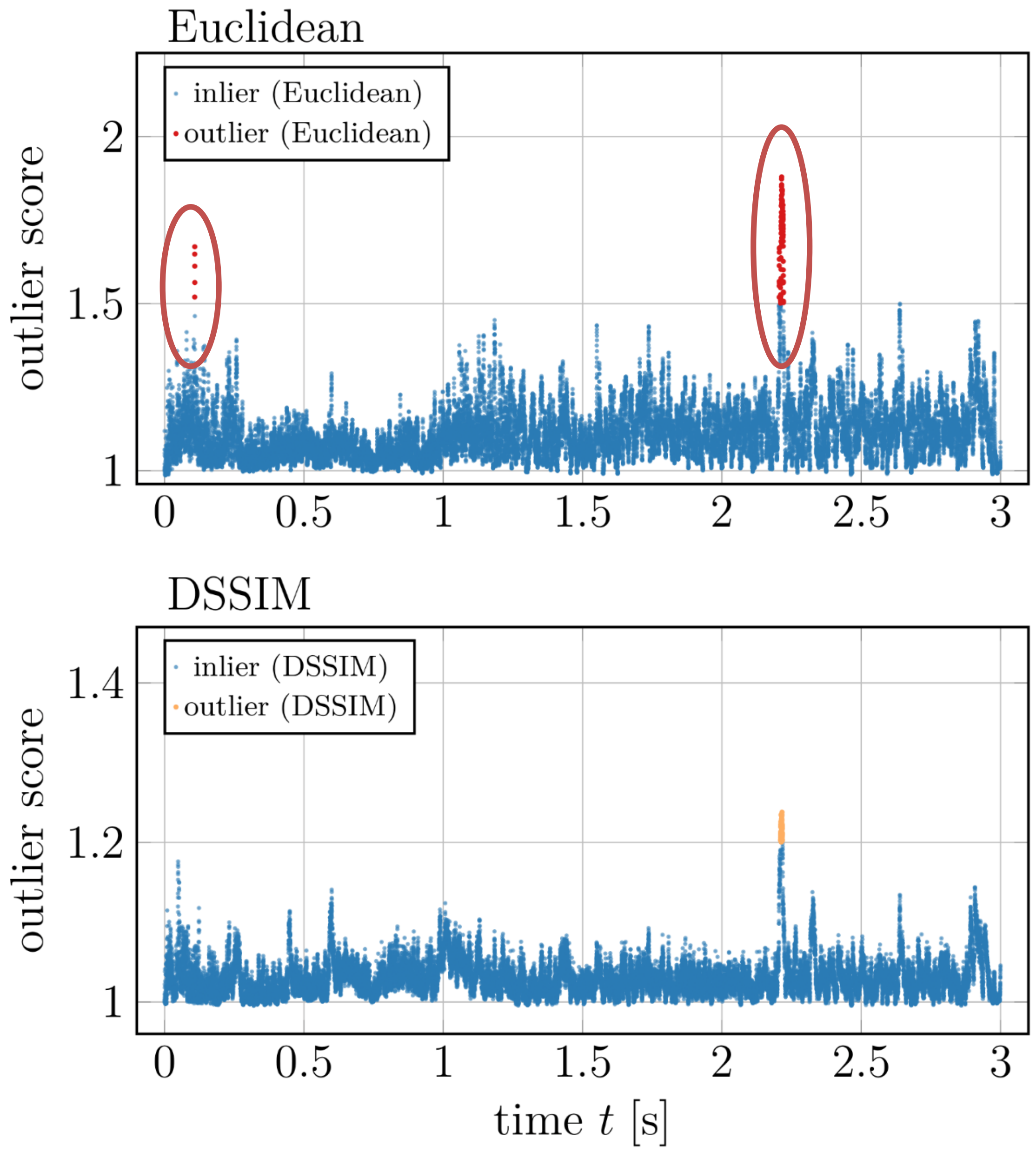
A. Rüttgers, A. Petrarolo, M. Kobald, Clustering of paraffin-based hybrid rocket fuels combustion data. *Exp. Fluids*, 61:4 (2020)

Anomaly Detection: Local Outlier Factor (LOF)

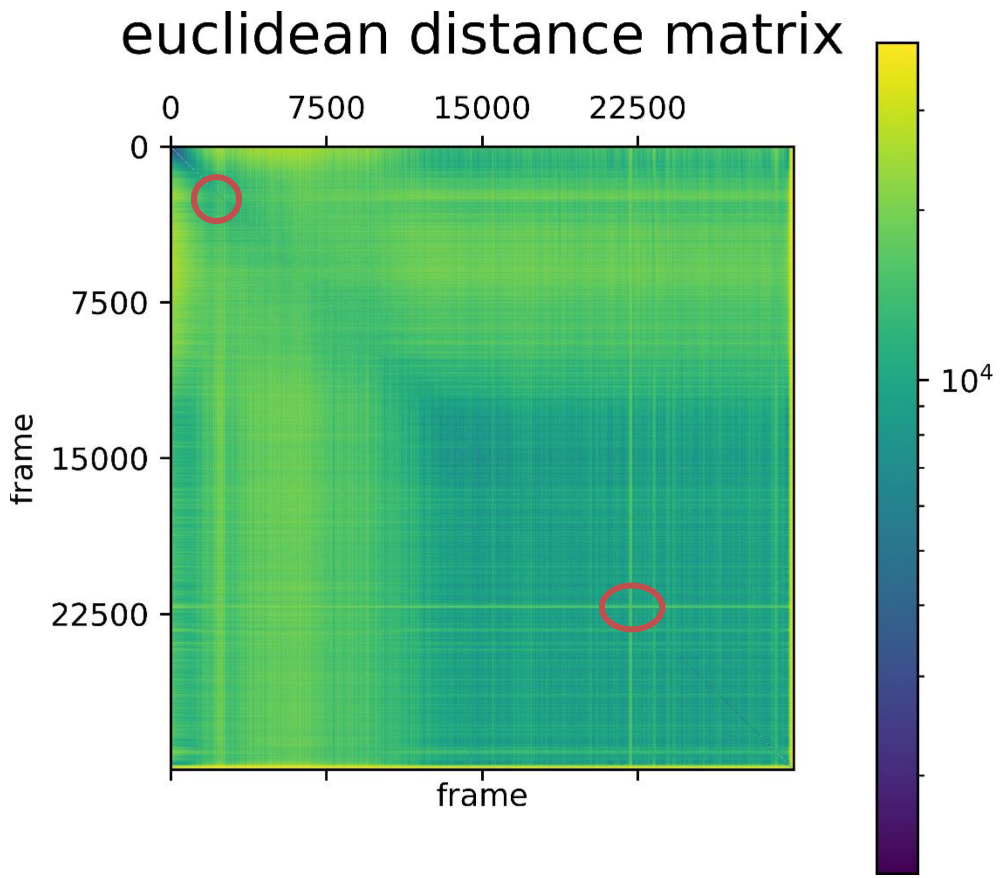
- Algorithm that bases on local density of data points.
- Shares some concepts with clustering algorithms such as DBSCAN and OPTICS.
- Does not show a decision boundary, i.e. cannot directly be used on new data (not necessary here).
- **Core idea:** Compare local density of an object to the local densities of its neighbors.
  → distance matrices from clustering are reused
- Ratio „Density of neighbors / local density of an objects"
  - ≈ 1.0 means similar density as neighbors
  - > 1.0 means lower density than neighbors (outlier candidate)



Point density with respect to k=3 closest neighbors

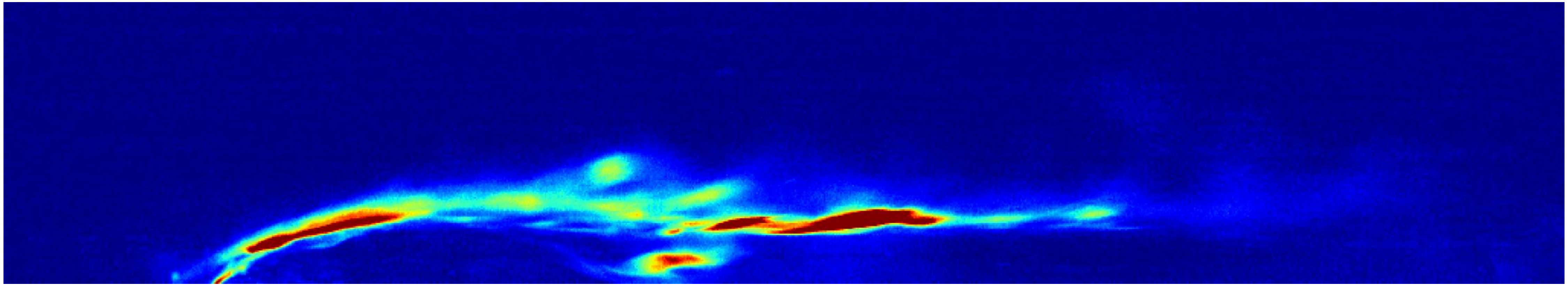


- Euclidean distance norm returns larger outlier score values (due to irregular matrix?).
- SSIM and Euclidean distance share some anomalies but there are differences.

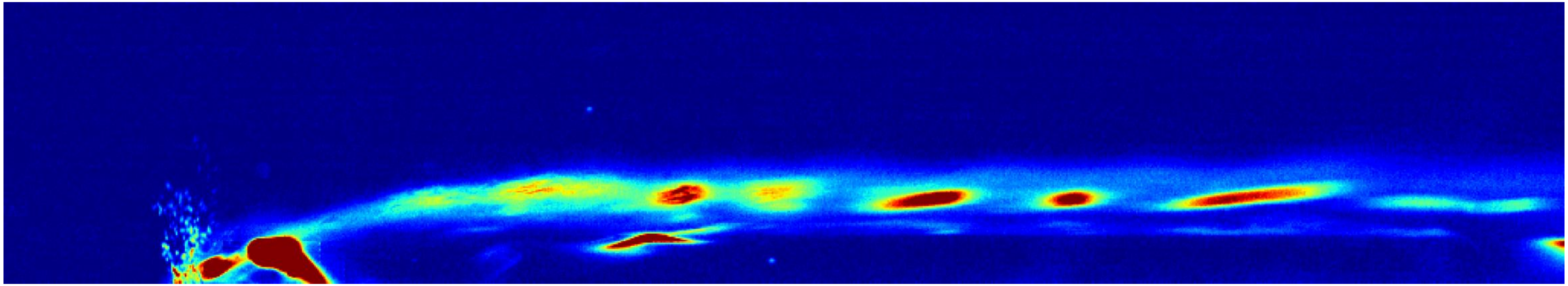


A. Rüttgers, A. Petrarolo, Local Anomaly Detection in Hybrid Rocket Combustion Tests. *Exp. Fluids*, 62:136 (2021)

Peak outliers of Euclidean metric (test 284)

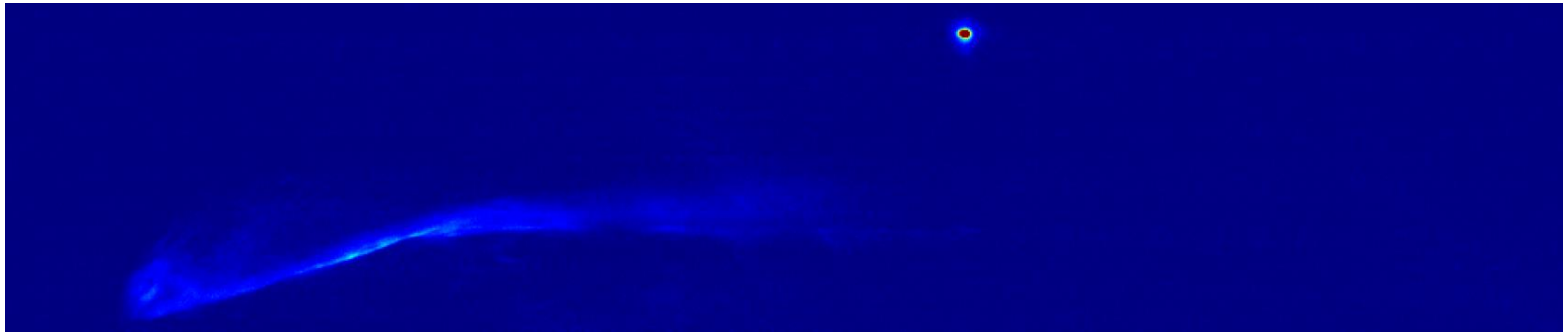


Flame fluctuations in ignition phase at t = 0.1078 s



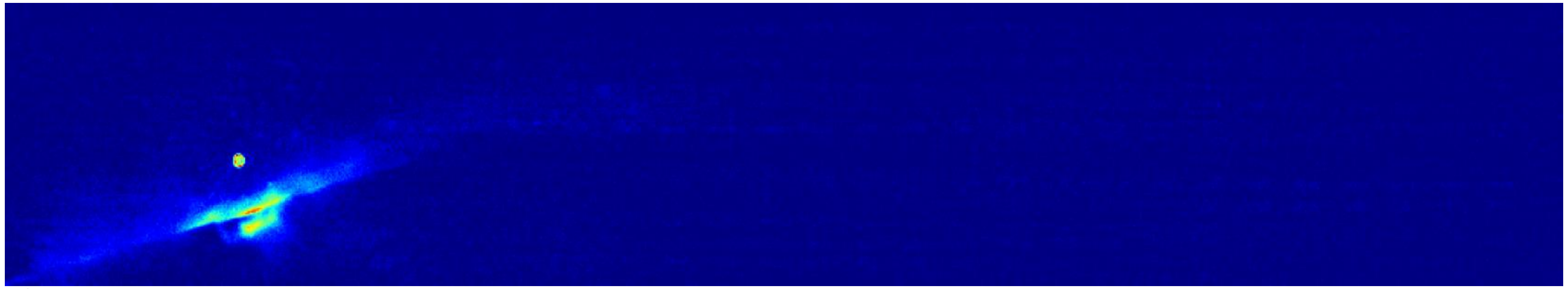
Droplet detection towards end of combustion at t = 2.2055 s

Some outliers found in other combustion tests



Test 291:
satellite droplet at t = 0.0253 s

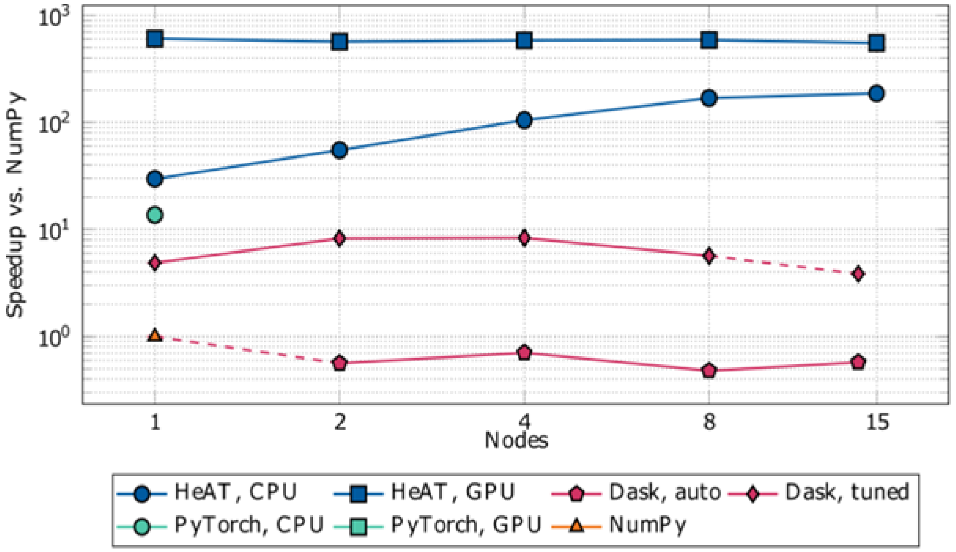
Test 296:
satellite droplet at t = 0.0017 s



Test 296:
satellite droplet at t = 0.0223 s

Conclusion and outlook

- Compute intensive clustering and anomaly detection on large data (e.g. rocket combustion image data) is possible using our software Heat
- Outperforms DASK, PyTorch and Scikit-Learn on distributed data
- Allows deep insights into the combustion process, e.g. to identify different phases and irregularities during combustion
- further insights are possible if datasets are combined (e.g. anomaly detection in spectral and image data).
- Heat currently used for a variety of applications, e.g.
  - Structural prediction of Proteins and RNA (project ProFiLe)
  - Classification of Land-Cover
  - Temporal prediction of physical system with Reservoir Computing



Runtime Speed-Up on distributed data

M. Götz et al., HeAT - a Distributed and GPU-accelerated Tensor Framework for Data Analytics. *2020 IEEE International Conference on Big Data* (2020) pp. 276-287

Thank you for your attention!

DLR