DLR project “Big-Data-Plattform” – BigData@DLR

Guest lecture at FOM University of Applied Sciences
Cologne, February 13th, 2020

Dr. Alexander Rüttgers*, Dr. Paola Rizzoli**, Dr. Manfred Zink**,
Christian Linder***, Kay Gimm***
and all other project members

*Institute for Simulation and Software Technology
** Microwaves and Radar Institute
***Institute of Transportation Systems
German Aerospace Center (DLR)

Dr. Alexander Rüttgers

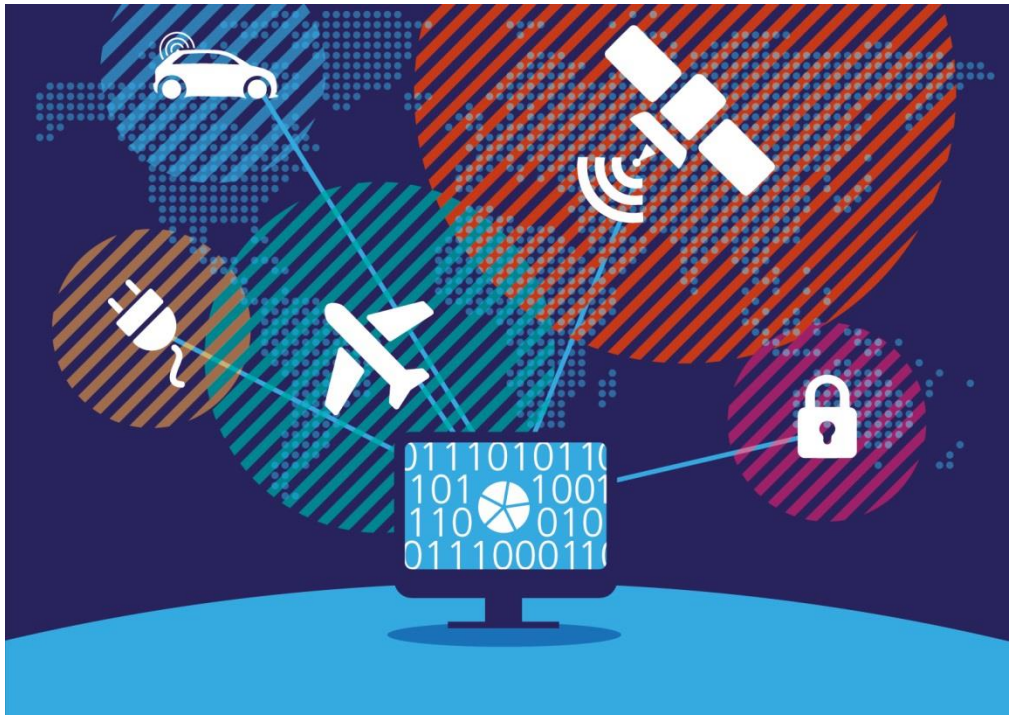
University of Bonn, Germany

- Diploma in mathematics in 2010
- PhD in mathematics in 2016
- Research assistant from 2011-2016

Institute for Simulation and Software Technology

German Aerospace Center (DLR), Cologne

- Since 2017: Data analysis and software development in space transportation projects ATEK and STORT
- Since 2018: Big-Data-Plattform project leader
- Since 2018: Organization of Machine Learning workshops at DLR

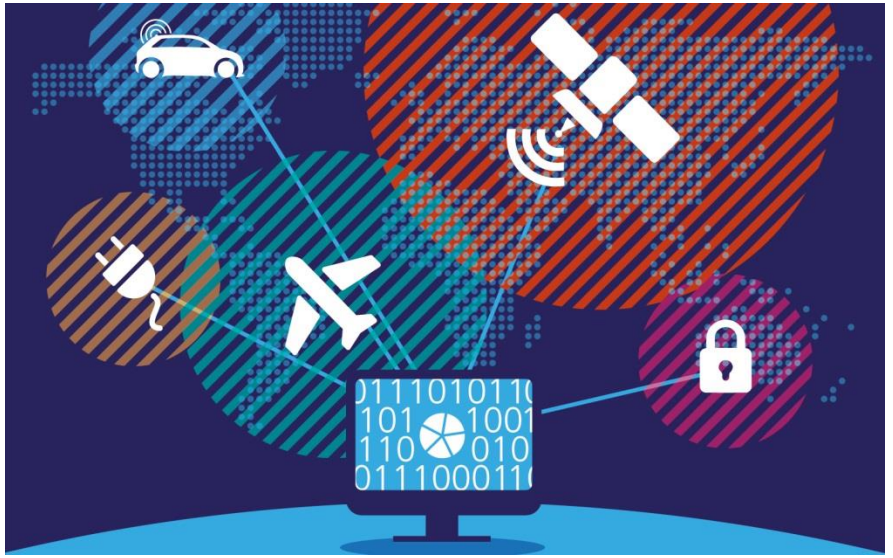


DLR project „Big-Data-Plattform“

Outline

- German Aerospace Center (DLR)
  - DLR at a glance
  - Institute for Simulation and Software Technology

- DLR project „Big-Data-Plattform“
  - Key facts of the project
  - Hybrid rocket data analysis
  - Global TanDEM-X forest map
  - Traffic monitoring and analysis



Deutsches Zentrum für Luft- und Raumfahrt (DLR)
German Aerospace Center

- Research Institution
- German Space Agency
- Project Management Agency

Location and employees

- approx. 9000 employees across 47 institutes and facilities at 27 locations
- Offices in Brussels, Paris, Tokyo and Washington
- Confirmed from BMWi in November 2018:
  - Institute for Maritime Energy Systems in Geesthacht
  - Institute for Systems Engineering in Oldenburg

DLR in Cologne

- DLR headquarter is located in Cologne (close to Cologne Bonn Airport)
- Approx. 1600 members of staff are employed across nine research institutes
- DLR campus in Cologne also includes
  - European Transonic Wind Tunnel (operated by Germany, France, Netherlands)
  - European Astronaut Centre (operated by ESA)

Institute for Simulation and Software Technology

- Research and development of software technologies and incorporation into DLR projects.

- Organized in three departments:

Intelligent and Distributed Systems

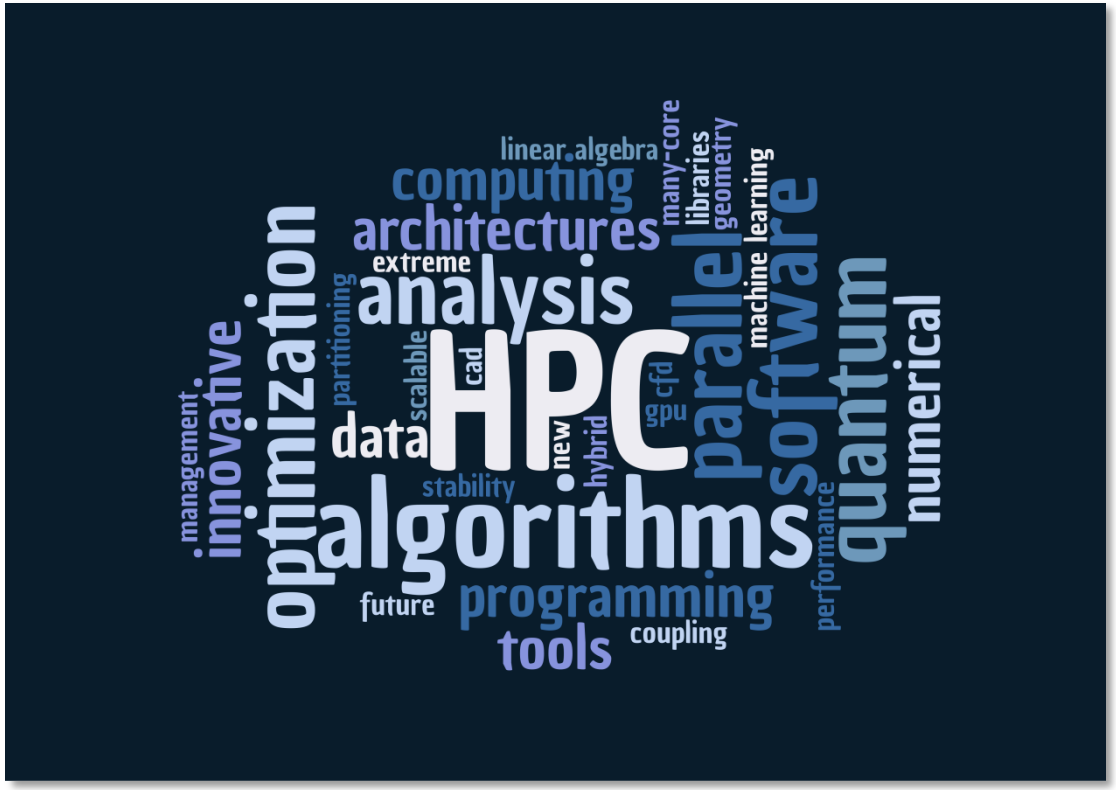
Software for Space Systems and Interactive Visualization

High-Performance Computing

High-Performance Computing Department

Research topics

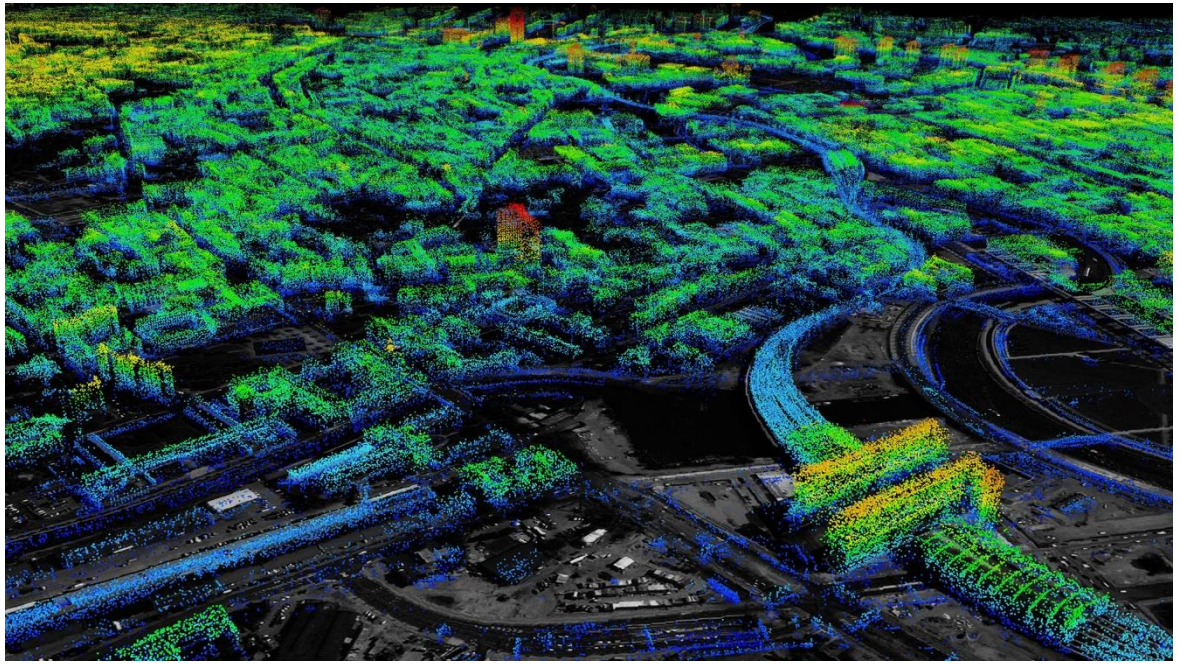
- Efficient treatment of large datasets and Big Data
- Parallel Numerical Mathematics
- Modeling and shape optimization
- Software for future architectures



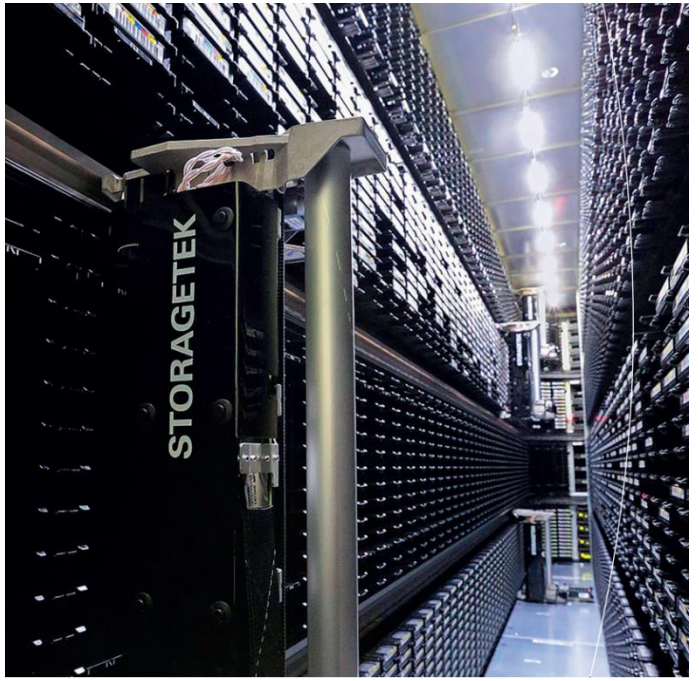
Examples for BigData@DLR

German Satellite Data Archive (DSDA)

- Data archive is operated by DLR‘s German Remote Sensing Data Center (DFD).
- The archive currently stores over 15 000 terabytes.
- Estimated growth is 12 000 terabytes / year from 2020 onwards.

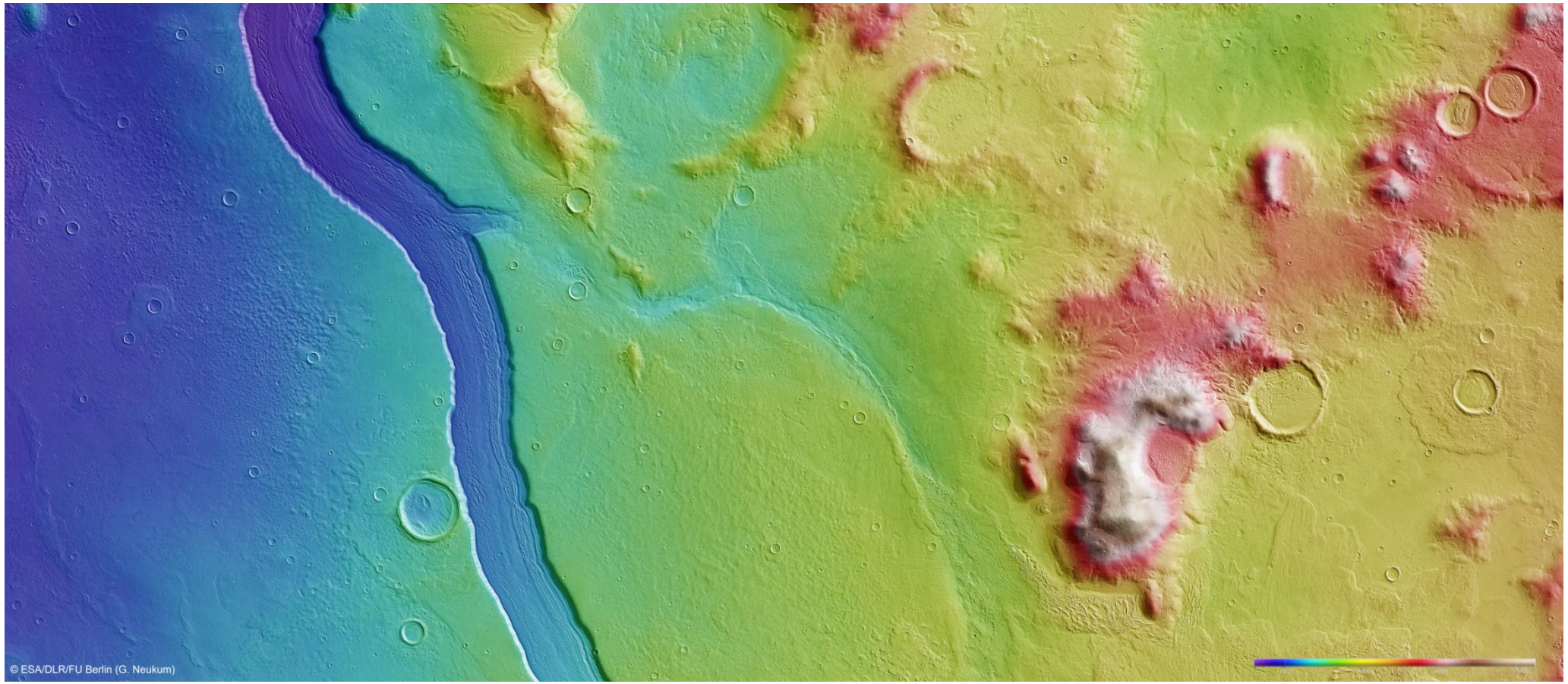


TanDEM-X mission data showing Mitte district of Berlin.



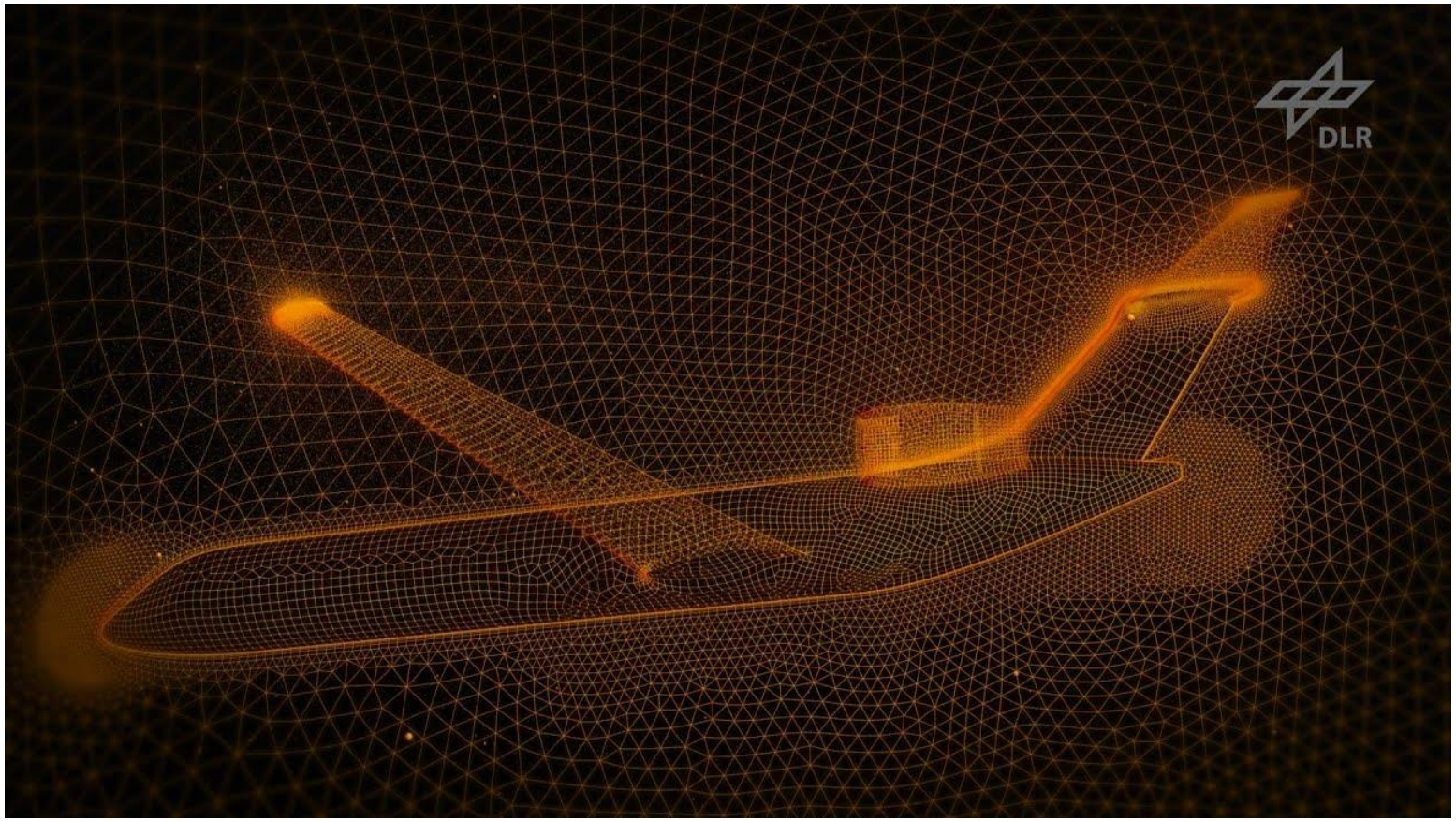
DSDA in Oberpfaffenhofen.

Satellite data from space missions



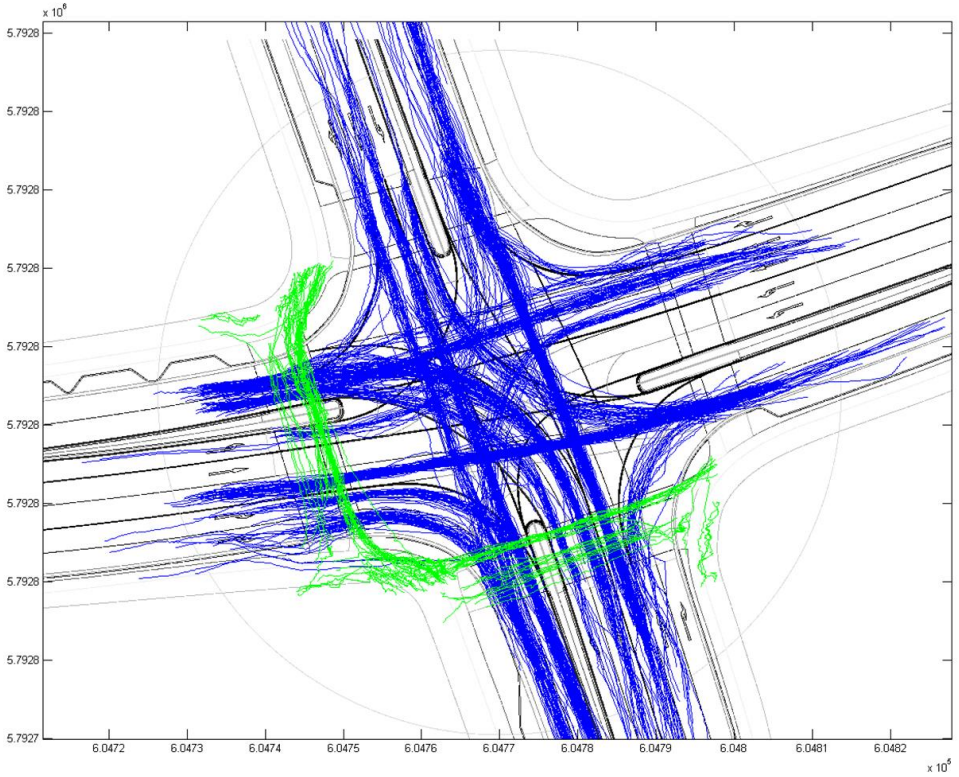
Topographic view of Reull Vallis on Mars. The height information is given by the color of the map.

Numerical simulation of an aircraft in the air

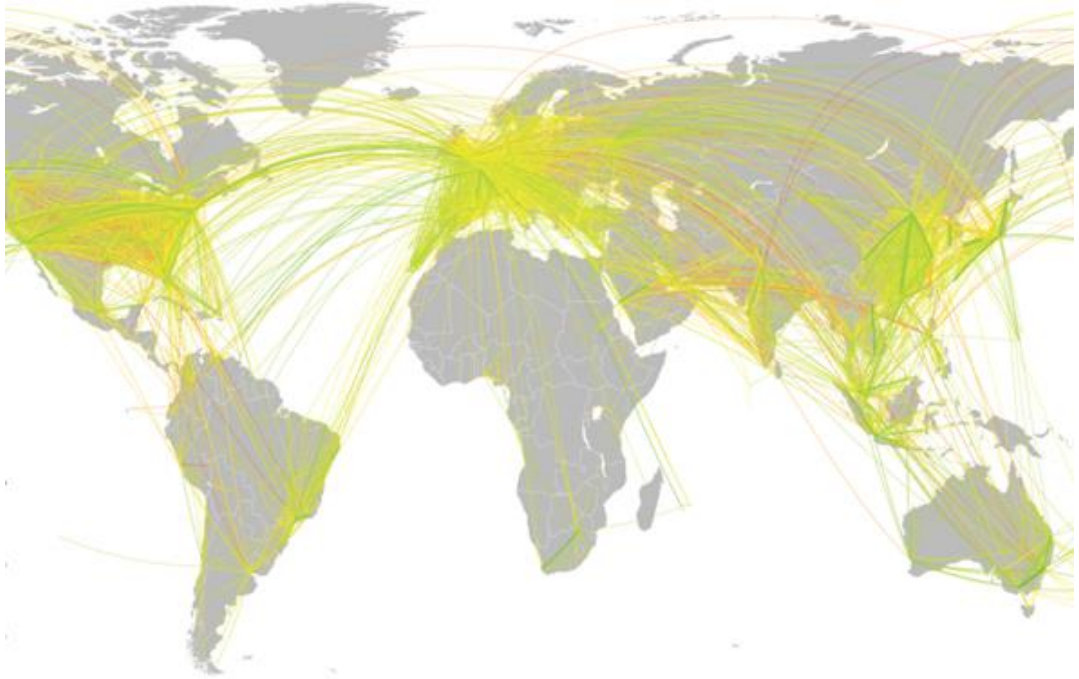


CFD simulation on the CASE2-cluster at DLR in Braunschweig with billions of unknowns for the velocity, density and pressure of the air.

Traffic data



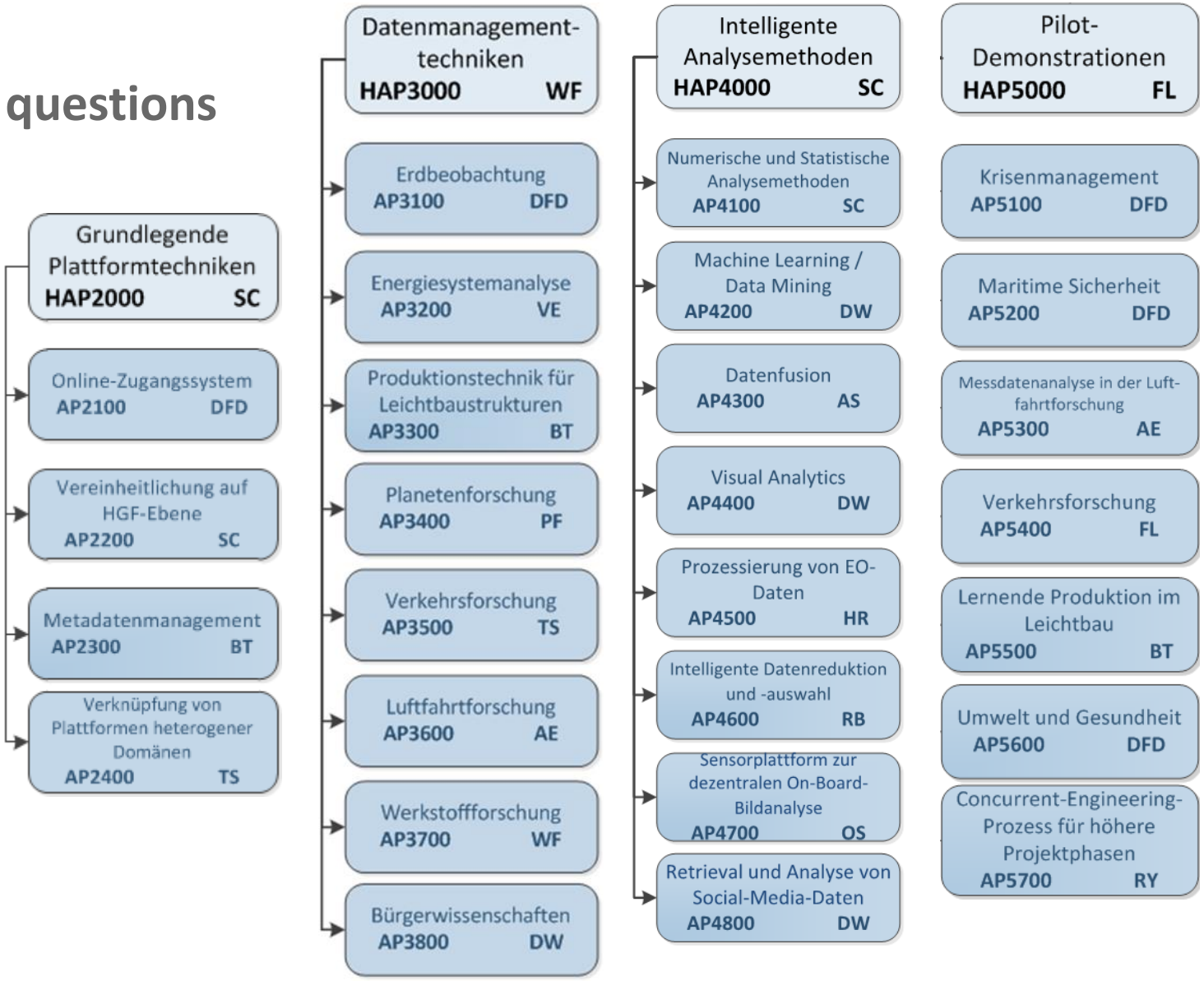
Car and pedestrian movement at road junction in Braunschweig.



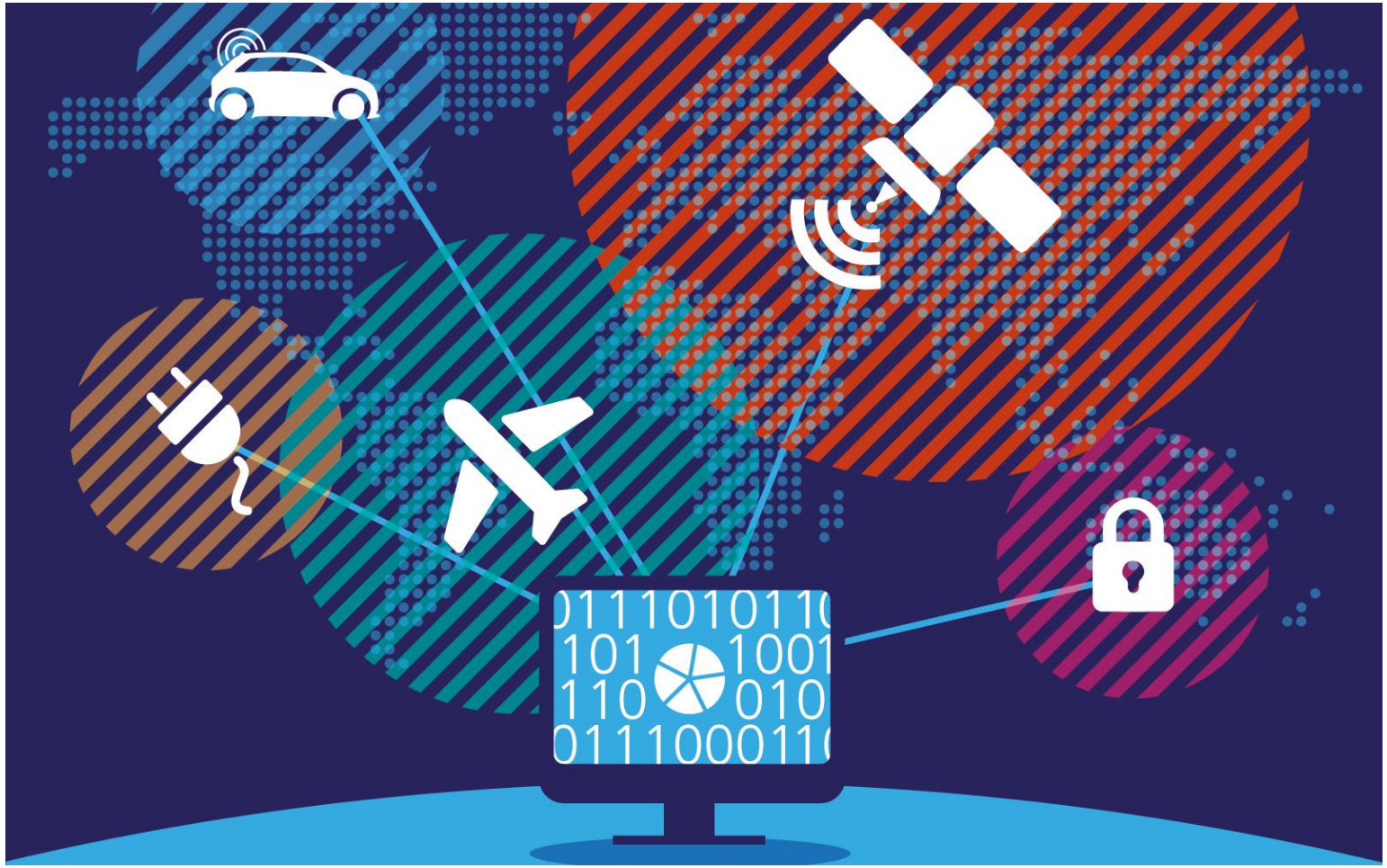
Global air traffic control.

Different applications – but similar questions

- How can standardized access to data be enabled?
- How can high data quality be ensured?
- How can actual knowledge be derived from the data?
- What are the potential benefits for society?



DLR project „Big-Data-Plattform“

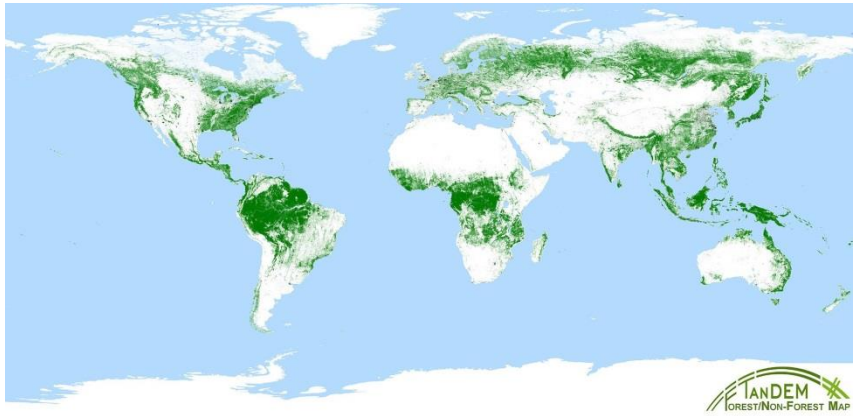


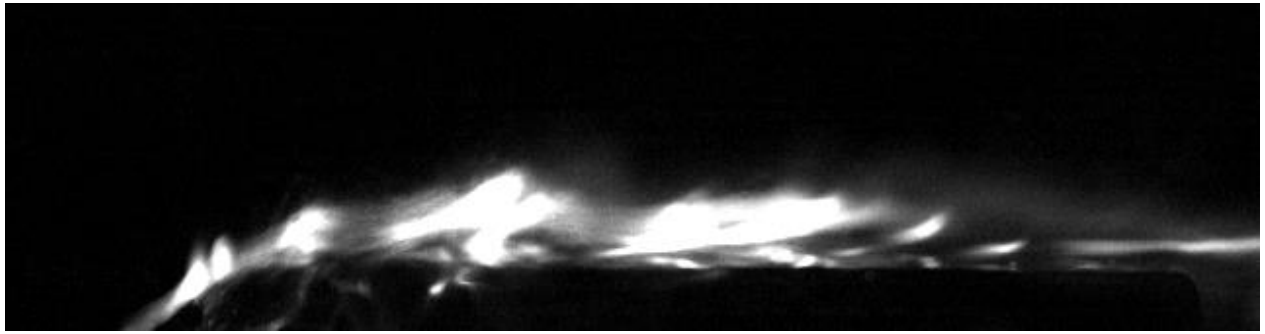
Key facts of DLR‘s „Big-Data-Plattform“

- Cross-sectoral project consists of 21 institutes from all of DLR‘s research areas
  - space
  - aeronautics
  - transport
  - energy
  - security
- Duration: 4 years (2018-2021)
- Project leader:
  - Dr. Achim Basermann
  - Dr. Alexander Rüttgers
- Main idea: develop general methods to perform data analytics on huge datasets in more than 50 synergy projects

Selected research from the Big-Data-Plattform

- Hybrid rocket data analysis
- Global TanDEM-X forest map
- Traffic monitoring and analysis





Hybrid rocket data analysis

Motivation (ATEK research rocket flight in June 2019)

https://www.youtube.com/watch?v=JlcReUwZXFU

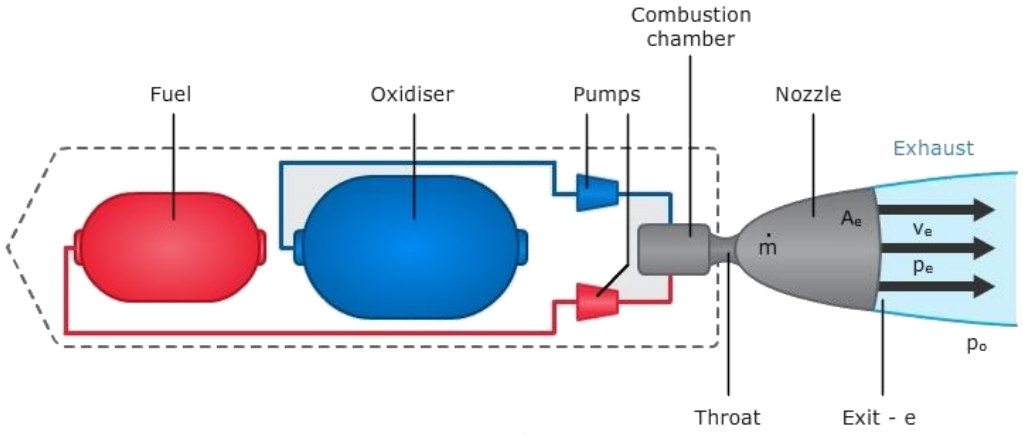
Project ATEK

- Project aim:
  - Cost reduction of spacecraft systems by using *reusable* or *less complex* propulsion technologies
  - Be competitive with e.g. Space-X

- Tasks:
  - Numerical simulation and data analysis
  - Experiments / technical construction
  - Flight operation (June 2019 in Kiruna, Sweden)
  - Next flight operation: October 2021 (project STORT)

- Participants: 8 DLR institutes

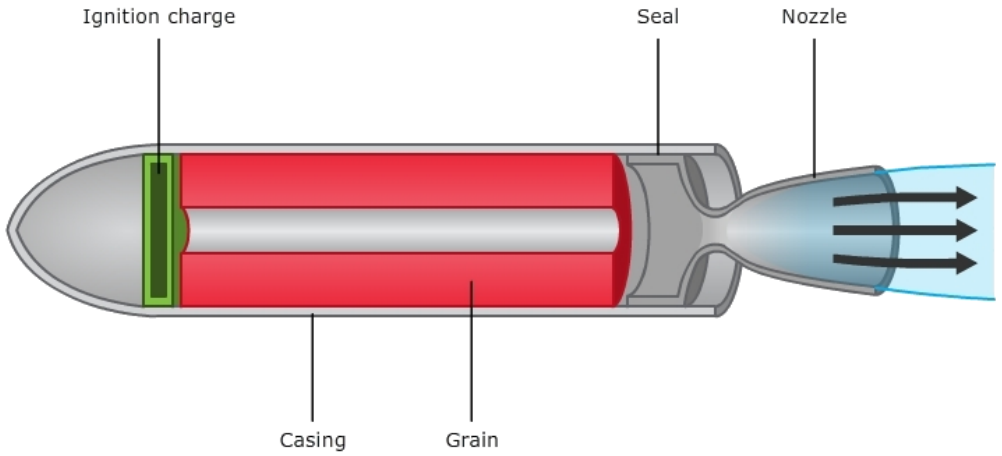
Rocket engine combustion analysis



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

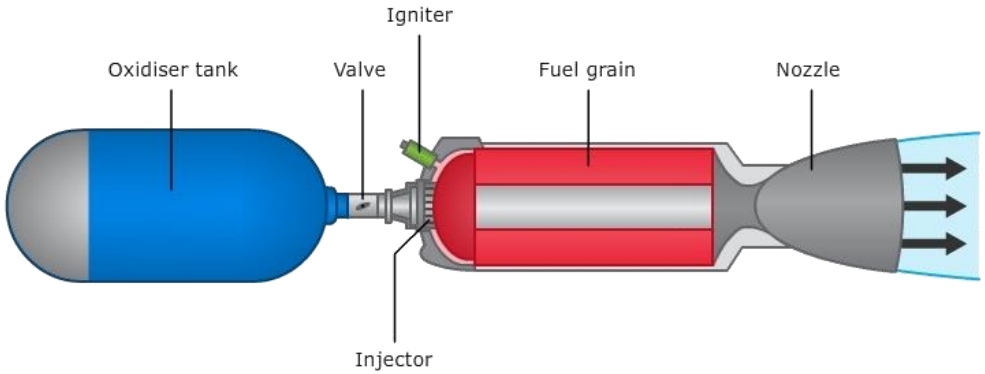
Rocket engine combustion analysis



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



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

- DLR investigates new hybrid rocket fuels on a paraffin basis at Institute of Space Propulsion in Lampoldshausen.

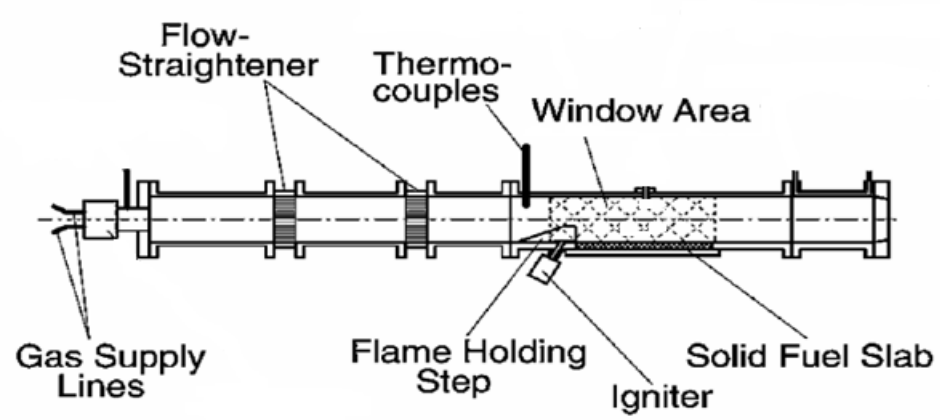
- About 300 combustion tests were performed with single-slab paraffin-based fuel with 20° forward facing ramp angle + gaseous oxygen.

- Two different fuel compositions:
  - pure paraffin 6805
  - paraffin 6805 + 5% polymer

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

Combustion chamber set-up

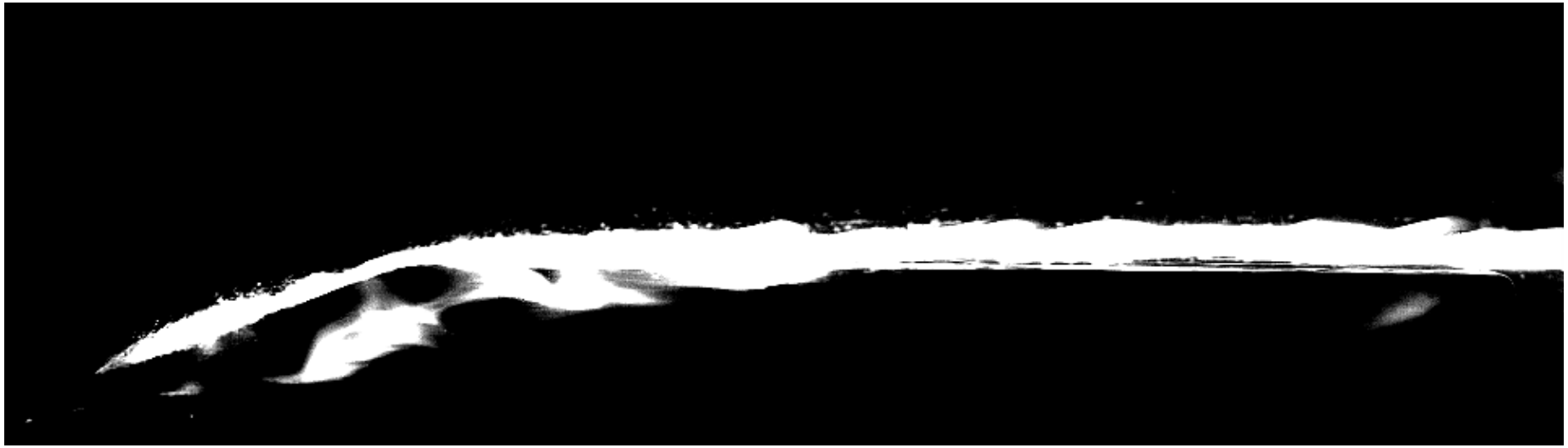
- Optically accessible combustion chamber is 450 mm long, 150 mm wide and 90 mm high.
- Tests were performed with different configurations (e.g. fuel, oxidizer mass flow, filters)
- Combustion is captured with high-speed video camera with 10 000 frames / second



Side view of combustion chamber

Test no.	Fuel		$\dot{m}_{Ox}[g/s]$		CH* filter
	6805	6805+5% polymer	10	50	
284	✓			✓	✓
289		✓		✓	✓
296		✓	✓		✓
243		✓	✓		

Test matrix used for data analysis



note: video has been replaced by this image

Video extract of test 284	fuel	oxidizer mass flow	CH*-filter	duration
Ignition, steady combustion, extinction	pure paraffin 6805	50 *g/s*,	yes, only wavelengths emitted from CH* are filmed	10 000 frames per second over 3 s

Identify different combustion phases

- **Aim:** Use clustering to identify short term turbulences and characterize combustion process

- 2D images can be interpreted as 1D vectors

- Time frame $i = 1, \dots, \ m = 30\,000$

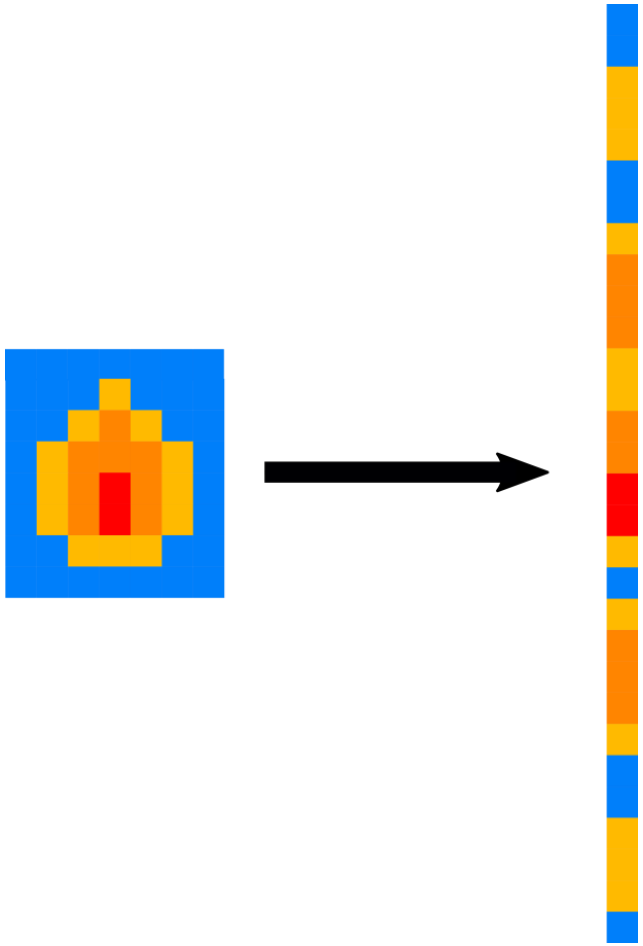
  $\vec{x}^{(i)} = (x_1{}^{(i)}, \ x_2{}^{(i)}, \dots, x_n{}^{(i)})$

  with $n = 1024 \ \times 187$ pixels

- Data set $X = \left\{\vec{x}^{(1)}, \ \vec{x}^{(2)}, \dots, \ \vec{x}^{(m)}\right\}$

- Large data sets = Long computing times

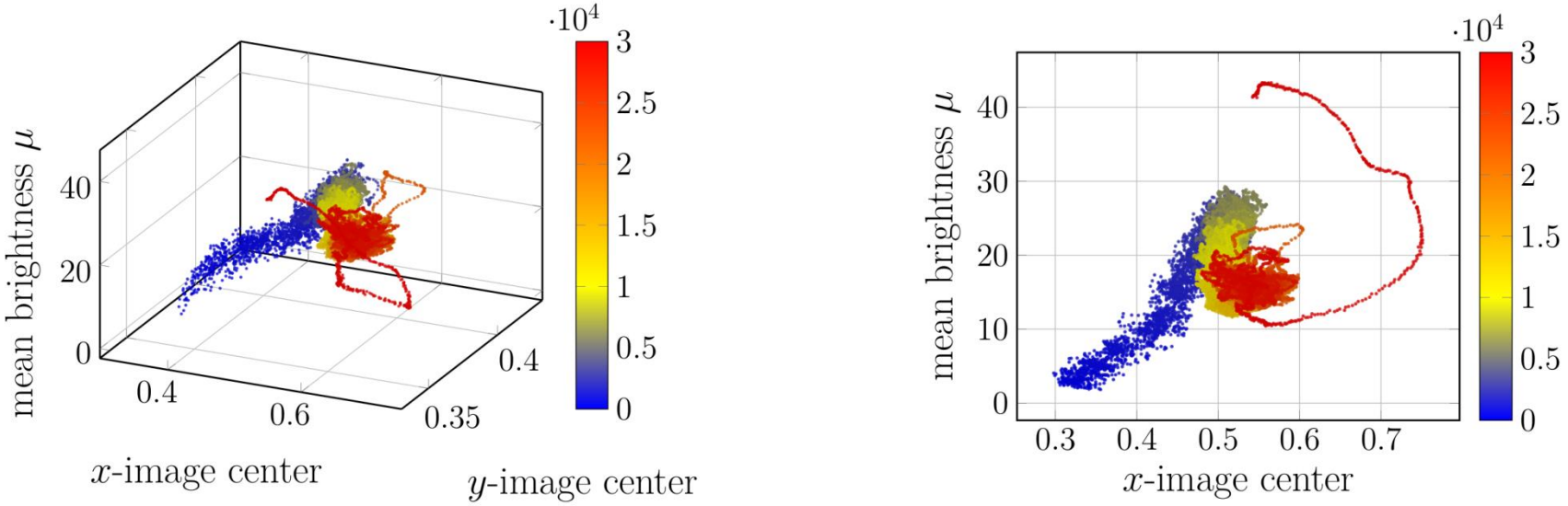
  ➡ High-performance computing



Choice of a clustering algorithm

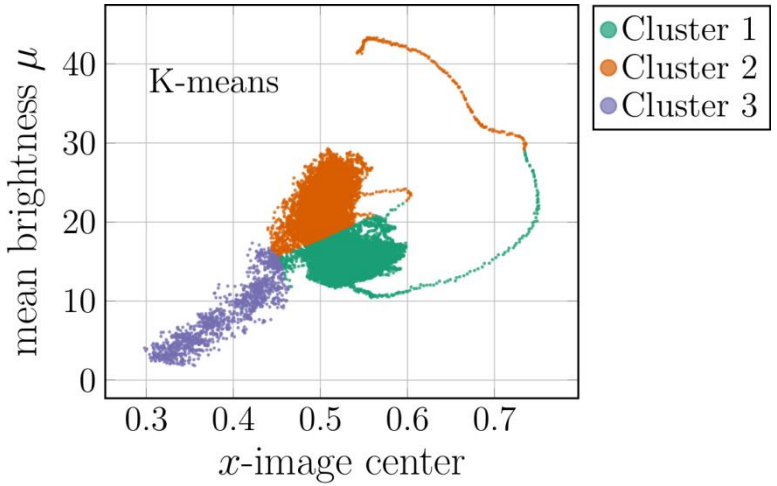
- Various clustering algorithms exist in the literature (DBSCAN, spectral clustering, k-means, …).

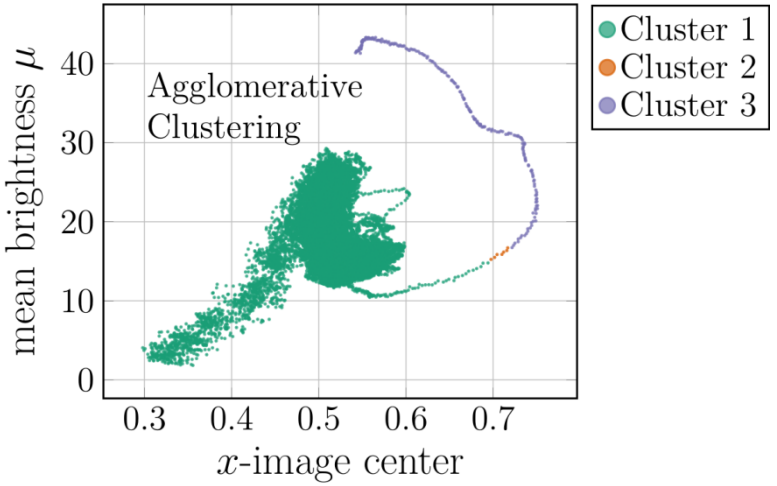
- **Test problem:** Comparison of algorithms on three features $(\mu, \overline{x}, \overline{y})_j$ for all $j = 1, \dots, 30000$ images of test 284.
  - μ: mean image brightness
  - $(\overline{x}, \overline{y})$: image barycenter

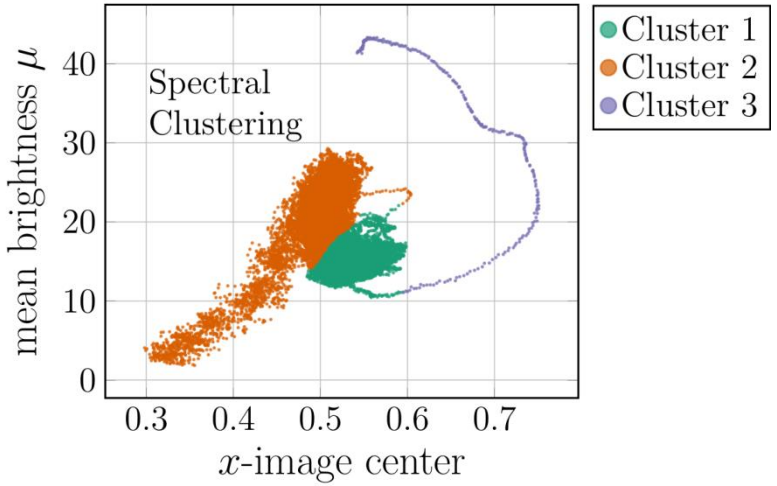


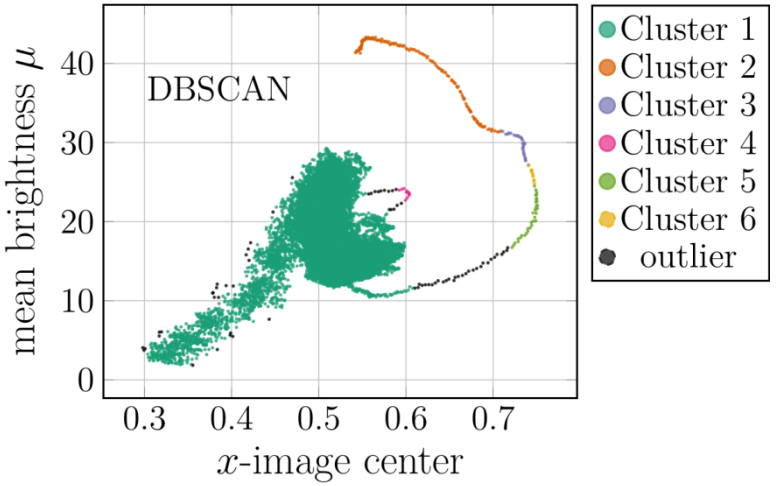
Low-dimensional approximation of test 284. The color represents time.

Hyperparameter study on 2D test problem – choice of clustering algorithm









Comparison of clustering algorithms for presented application

	K-means	Spectral clustering
approach	• Iteratively minimize the within-cluster sum of squares $J = \sum_{i=1}^{K} \sum_{\boldsymbol{x}_j \in C_i} \|\boldsymbol{x}_j - \boldsymbol{\mu}_i\|^2$	• Construct similarity matrix $\boldsymbol{A}$ of size (nr_of_points)x(nr_of_points) with e.g. Gaussian kernel • Build graph Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ with diag. matrix $D_{ii} = \sum_j A_{ij}$ • Compute first K eigenvectors of $\boldsymbol{L}$ • Cluster low-dimensional data representation with e.g. K-means
pros*	• Scales to large data sets • Guarantees convergence • Very simple	• Reduces curse of dimensionality • Does not make strong assumptions on clusters (e.g. spherical shape)
cons*	• Choosing K manually • Local optimal solutions • Curse of dimensionality • Similar-size clustering	• Choosing K manually • Expensive for large datasets • Number of hyperparameters

Large datasets (combustion images) = Long computing times ➡ High-performance computing with HeAT

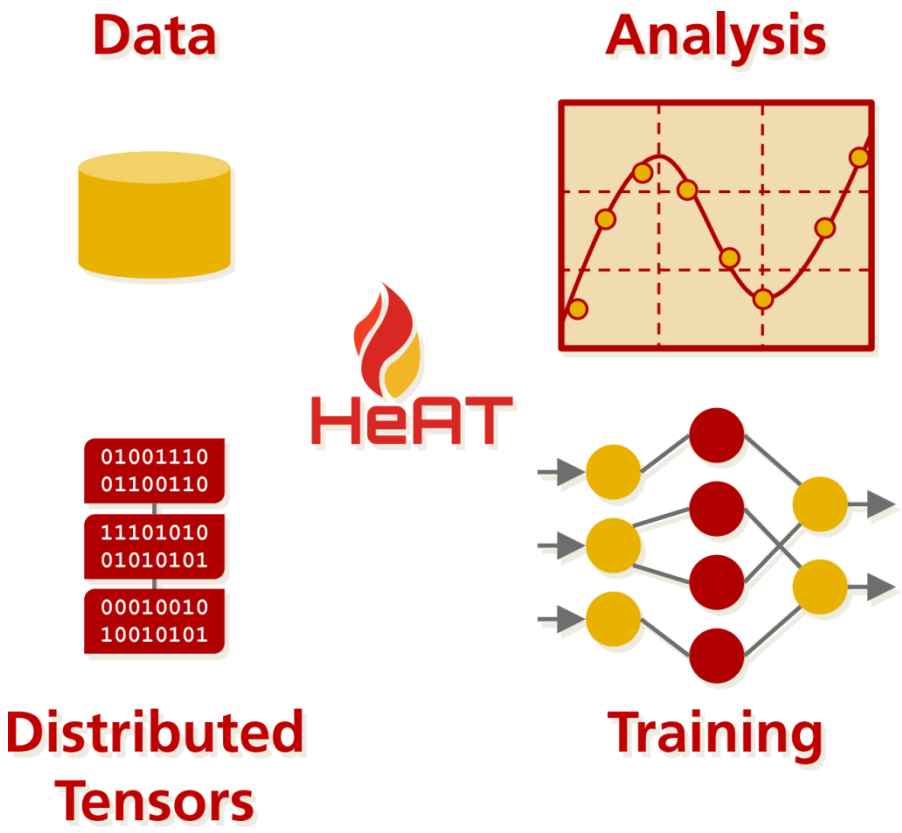
*incomplete list

The Helmholtz Analytics Toolkit (HeAT)

Bridge data analytics and high-performance computing

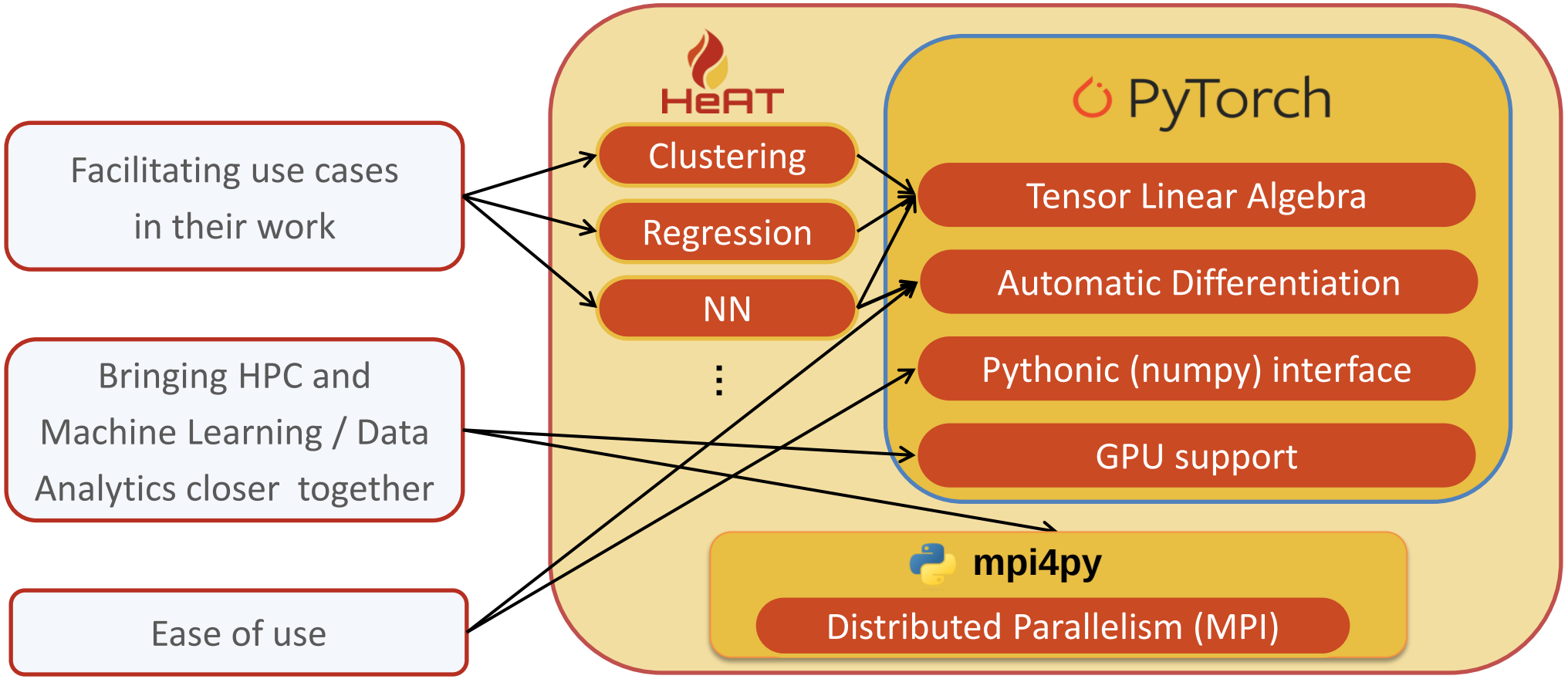
- HeAT = Helmholtz Analytics Toolkit
- Python framework for **parallel**, **distributed** data analytics and machine learning
- Developed within the Helmholtz Analytics Framework Project since 2018
- **Aim:** Bridge data analytics and **high-performance computing**
- Open Source licensed, MIT

helmholtz-analytics/heat

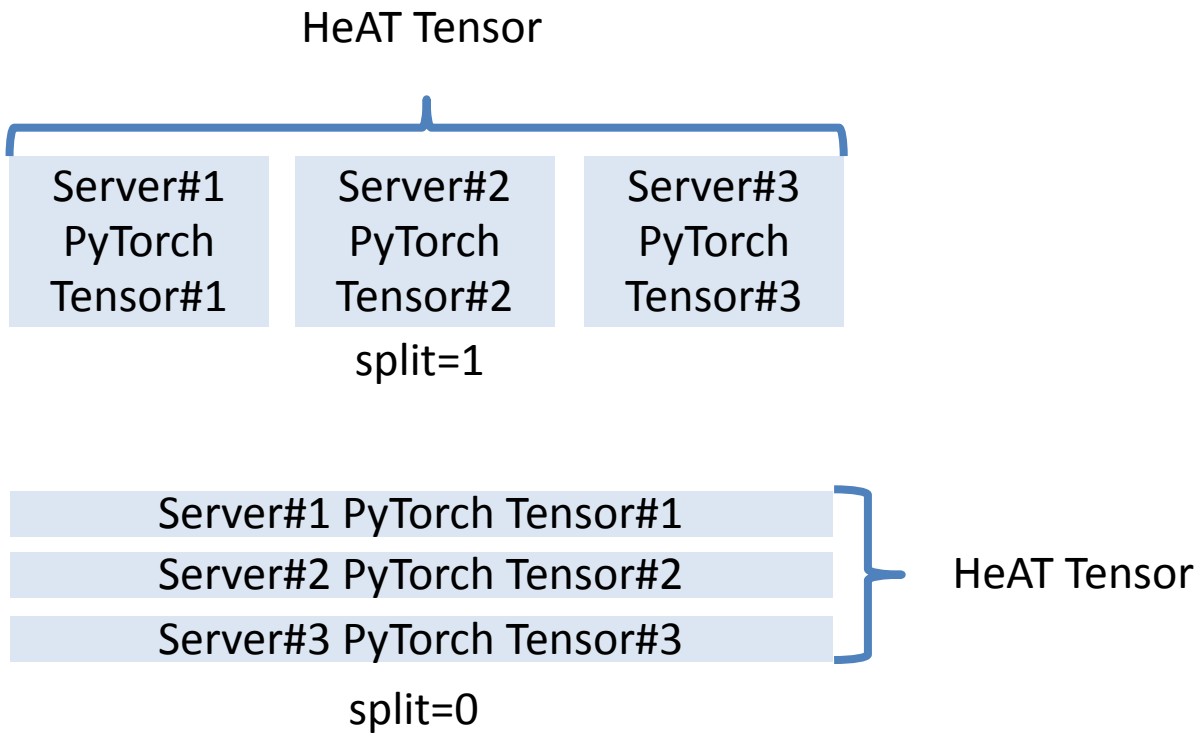


The Helmholtz Analytics Toolkit (HeAT)

Bridge data analytics and high-performance computing



Data Distribution



Example:

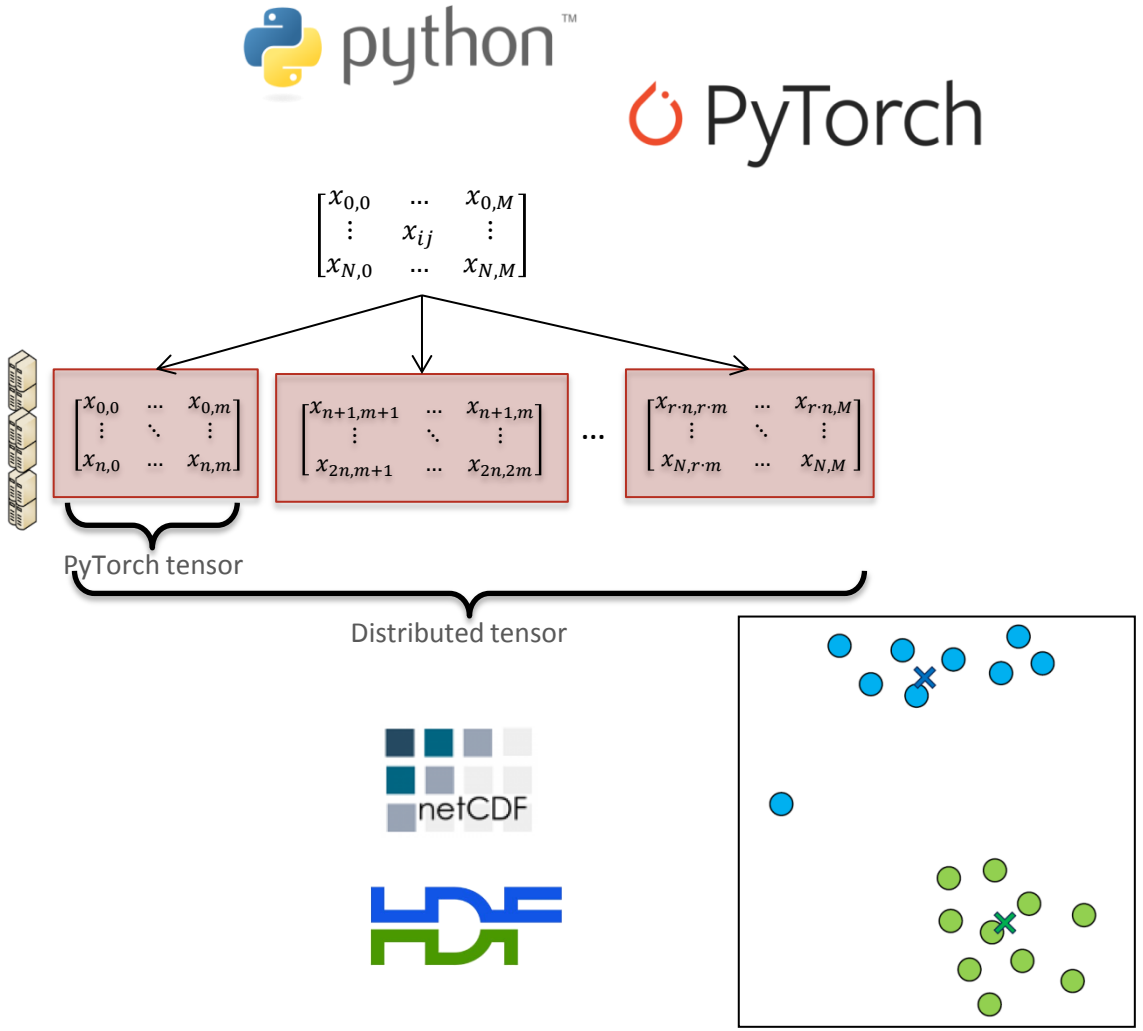
```
import heat as ht
# construct a range tensor
>>> range_data = ht.arange(6, split=1)
```

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

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

What has been done so far?

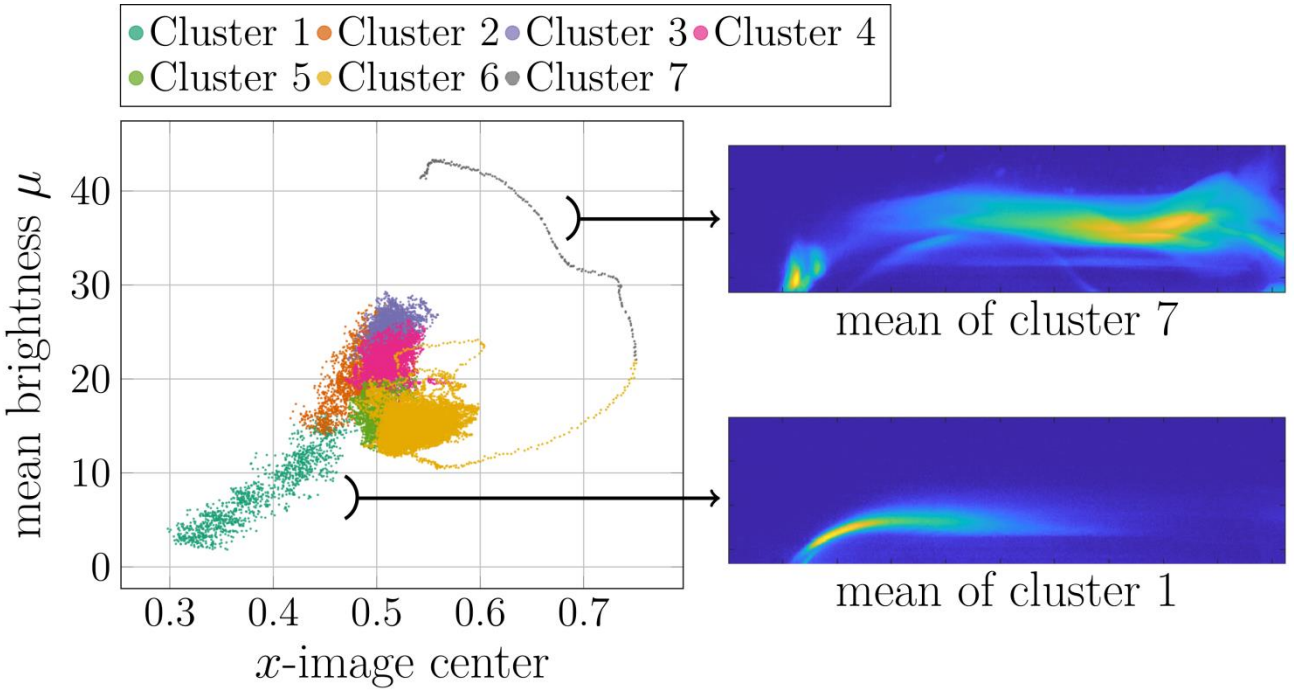
- The core technology has been identified
- Implementation of a distributed parallel tensor core framework
- NumPy-compatible core functionality
- Some linear algebra routines
- Parallel data I/O via HDF 5 and NETCDF
- K-means and spectral clustering algorithms are available



K-means clustering of combustion data

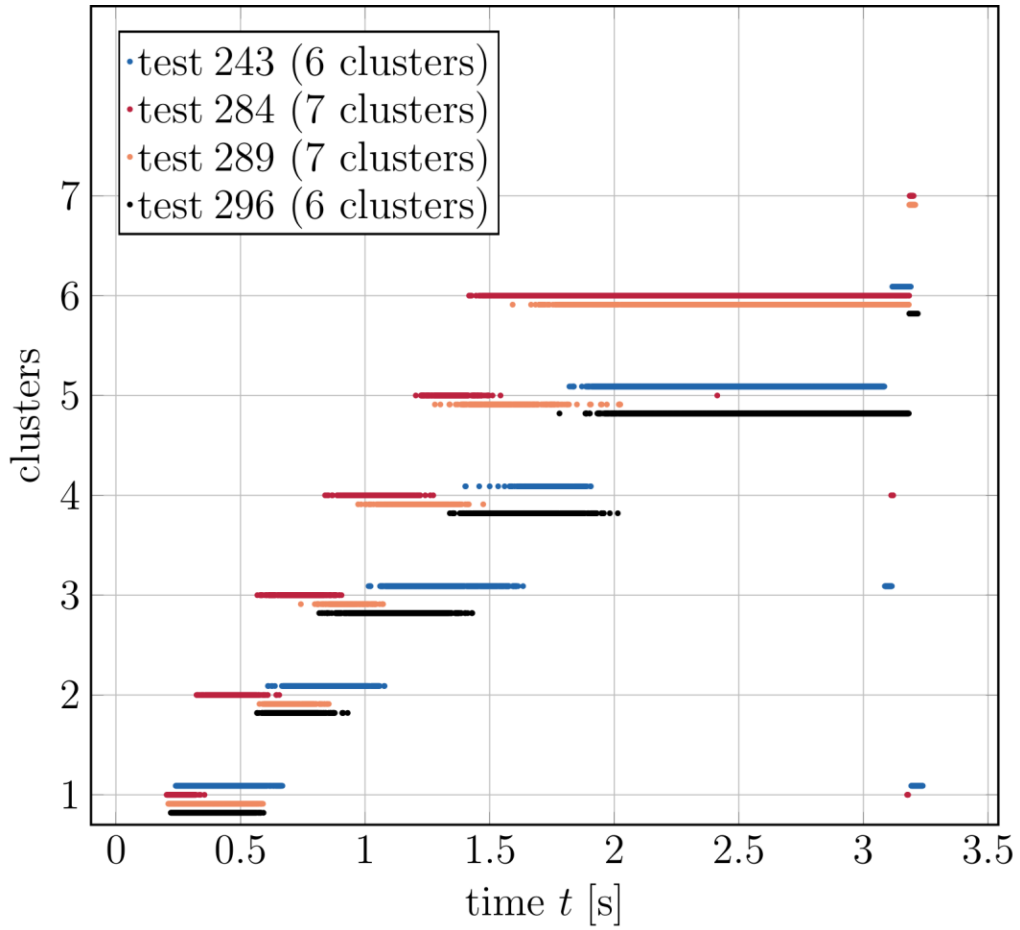
- Hyperparameter study*: K=7 optimal number of clusters for test 284

- Identification of flow phases:
  1. Ignition phase
  2. Burn phase
  3. Fuel slap burns in the middle
  4. Whole surface is burning
  5. Large side flame close to camera
  6. Constant combustion
  7. Flame extinguishing phase



*Rüttgers, A., Petrarolo, A., & Kobald, M. (2020). Clustering of paraffin-based hybrid rocket fuels combustion data. *Experiments in Fluids*, *61*(1), 4.

Comparison of tests – optimize configurations



Test	C_1	C_2	C_3	C_4	C_5	C_6	C_7
243	0.47	0.4	0.55	0.3	**1.21**	0.08	x
284	0.13	0.26	0.29	0.35	0.25	**1.7**	0.02
289	0.38	0.23	0.21	0.36	0.39	**1.4**	0.03
296	0.36	0.29	0.51	0.55	**1.25**	0.04	x

- Clustering allows for quantitative comparison.
- Optimization of experimental configuration is possible

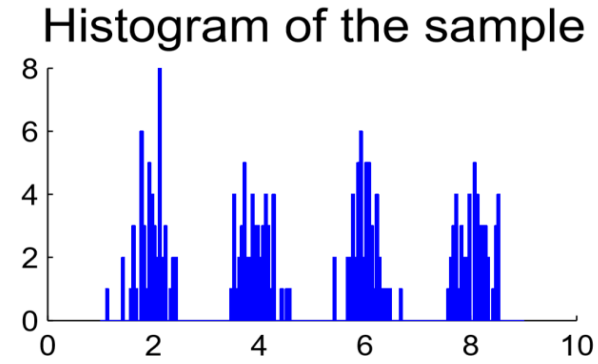
Distribution of frames to their corresponding clusters.

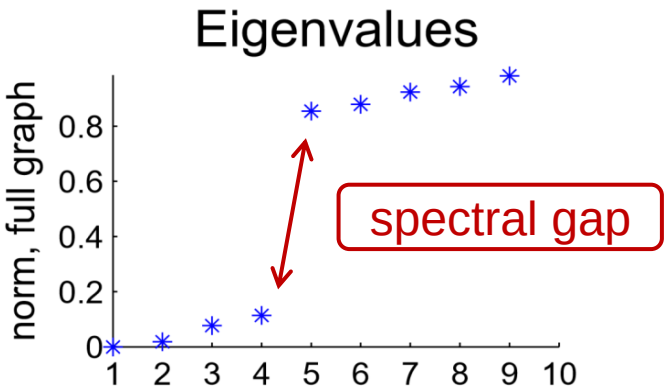
Time length of each cluster [s].

Spectral clustering: an approach to resolve anomalies?

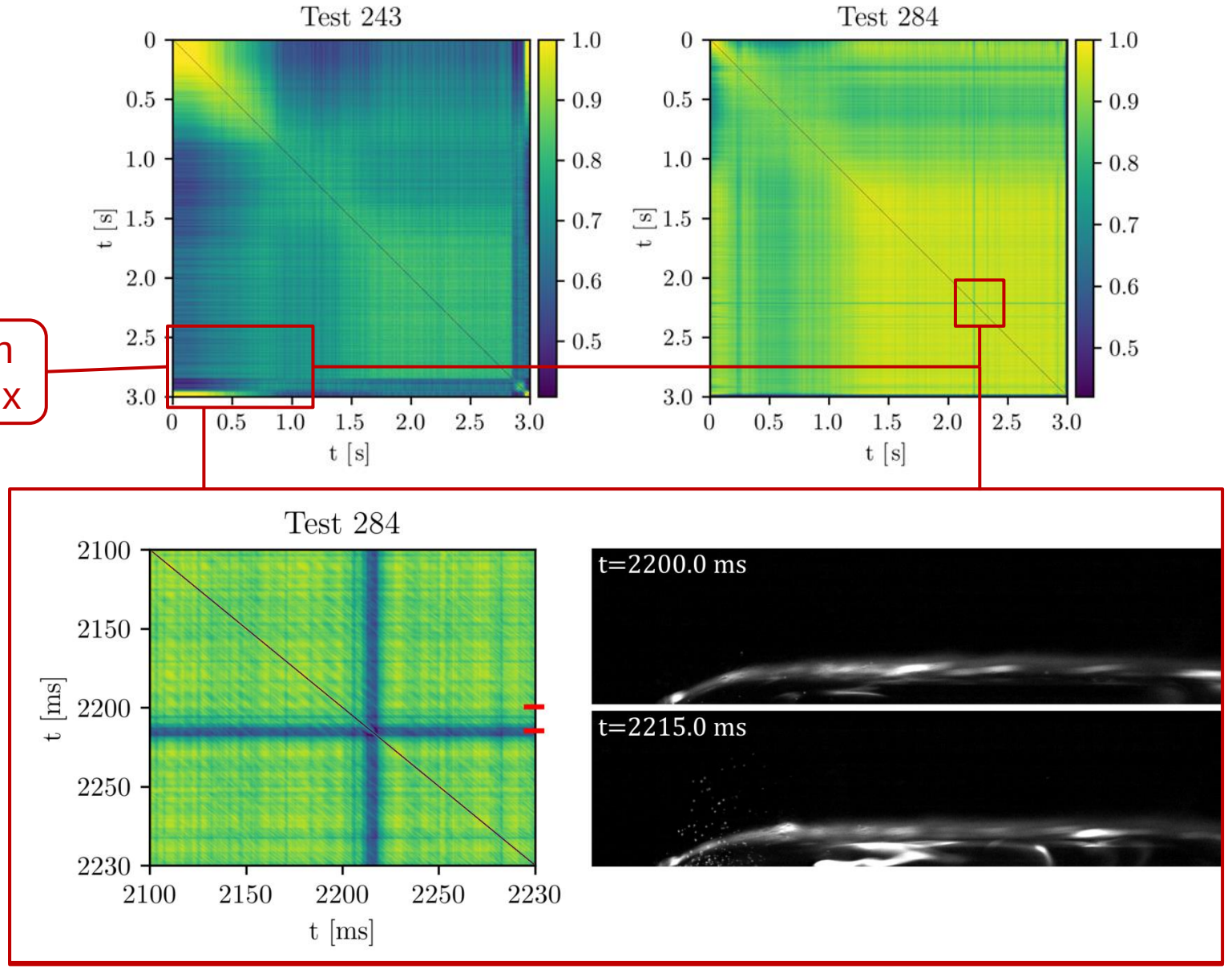
- Build graph Laplacian matrix $\mathcal{L} = D - A$
  - Similarity matrix $A = [s_{i,j}]$ of size $(30\,000 \times 30\,000)$
  - Degree matrix $D_{ii} = \sum_j A_{ij}$
  - Normalized symmetric graph Laplacian
    $\mathcal{L}^{sym} = D^{-1/2}\,\mathcal{L}\,D^{-1/2}$
- Compute eigenvalues and eigenvectors of $\mathcal{L}$

  → low-dimensional representation
- Cluster K smallest eigenvectors with e.g. k-means



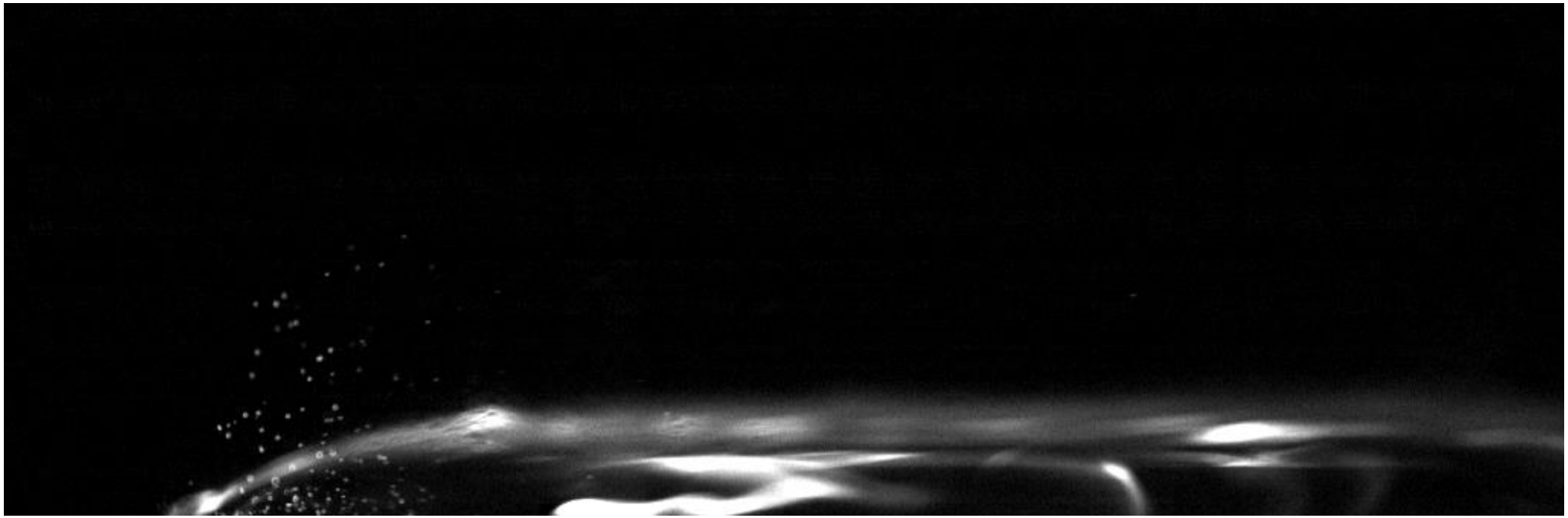


irregularities in similarity matrix

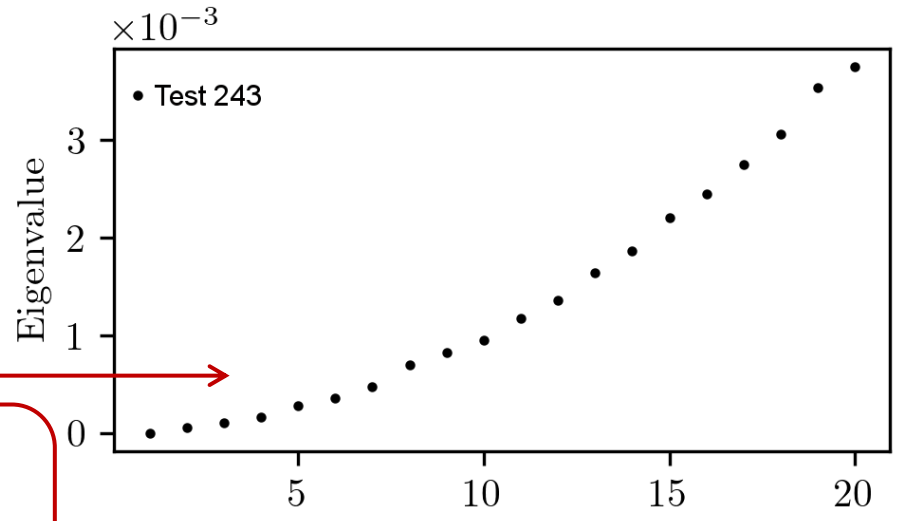


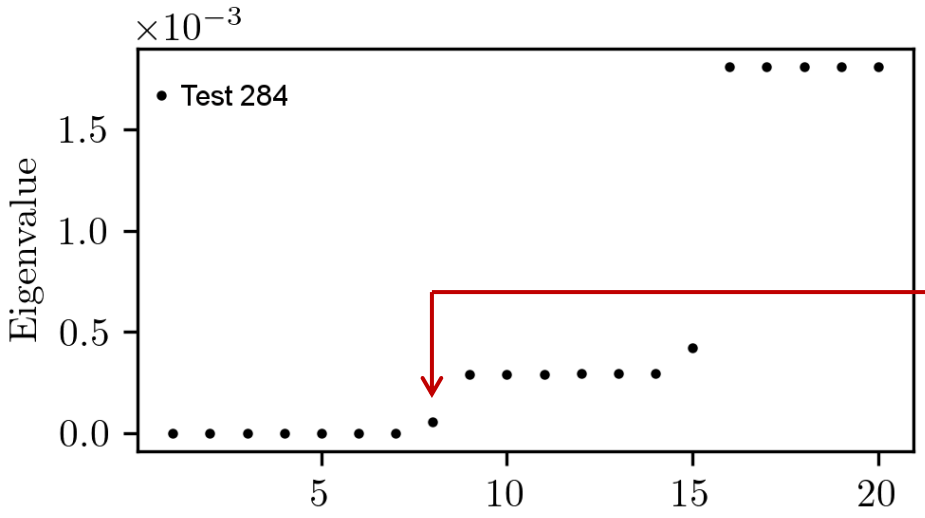
Similarity matrix with size (30 000 × 30 000) of all tests using a Gaussian kernel

Resolution of droplet entrainment recognized from affinity matrix

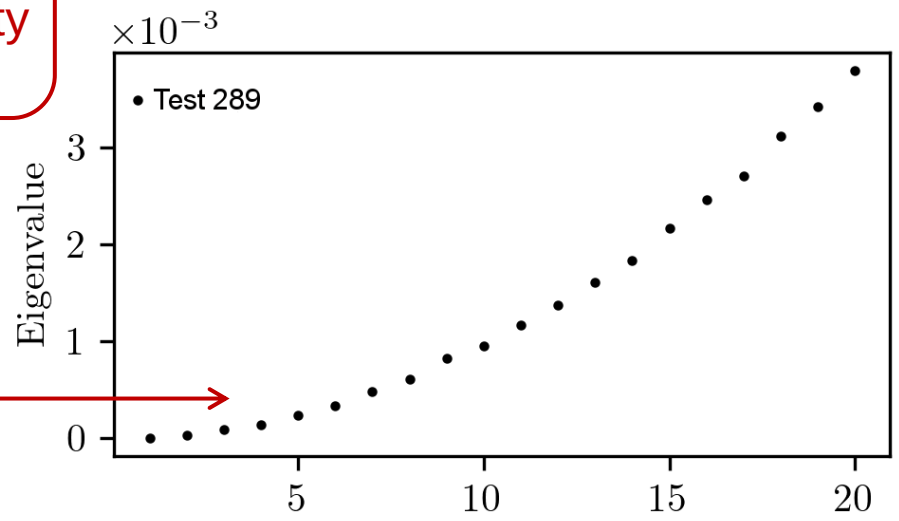


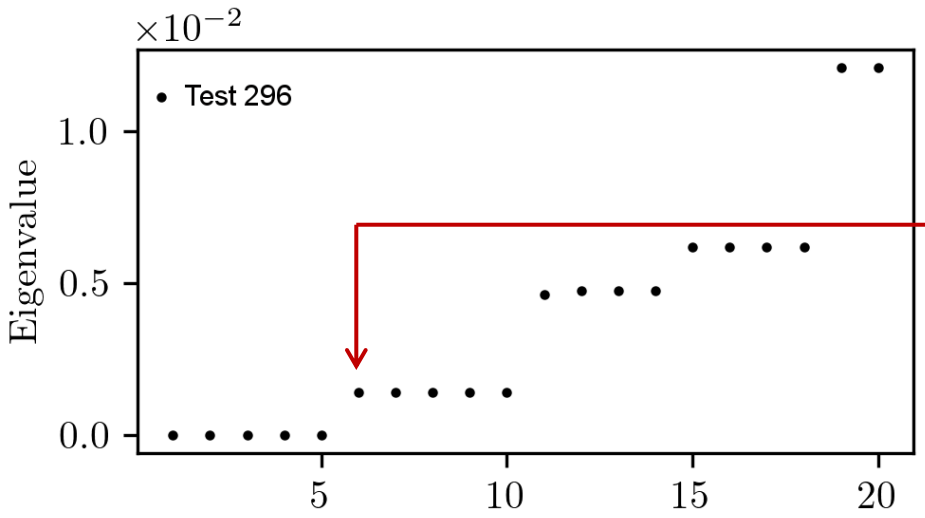
note: video has been replaced by this image





number of clusters? poor similarity measure?





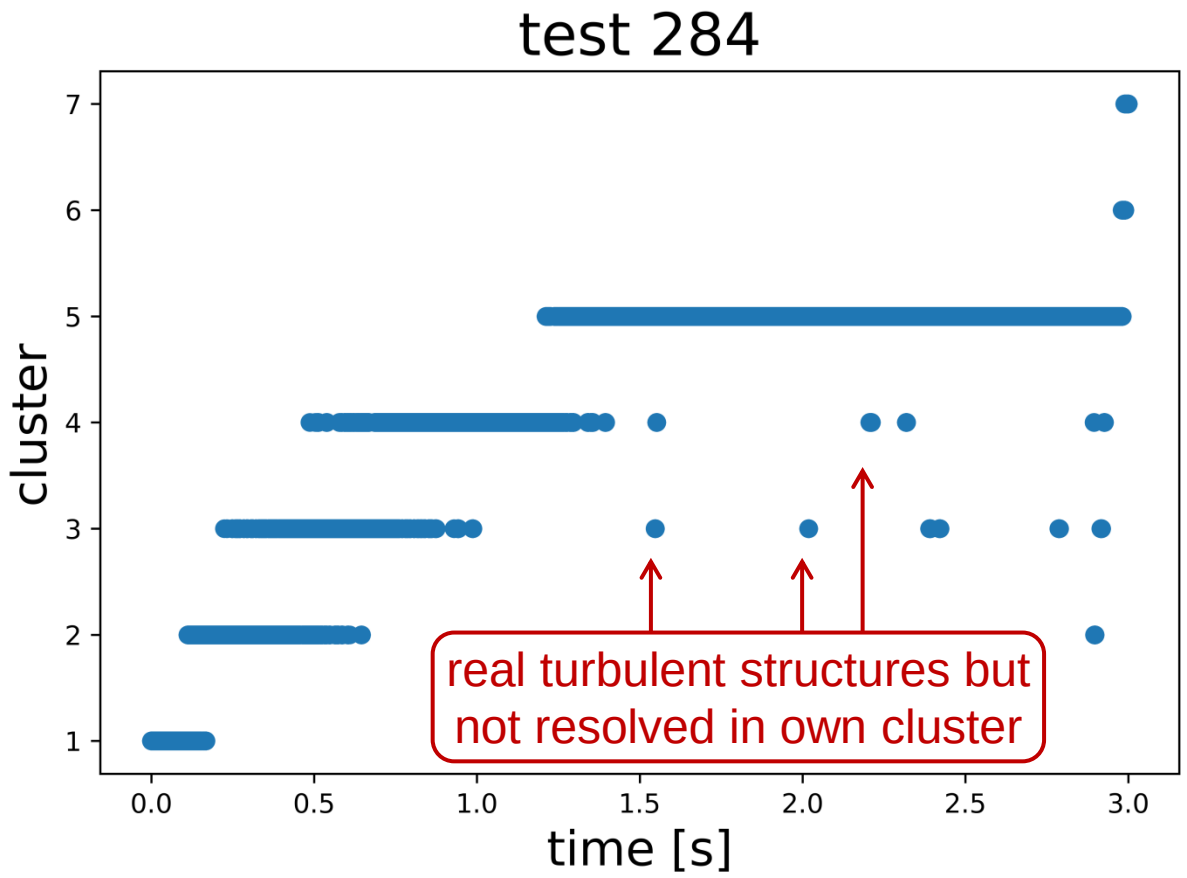
spectral gap indicates K=7 clusters and low density of the graph

spectral gap indicates K=5 clusters and higher graph density

20 smallest eigenvalues of the graph Laplacian of all four tests. The number of 0 eigenvalues of the graph Laplacian corresponds to number of connected components.*

*Von Luxburg. Statistics & Computing, 17:4 (2007)

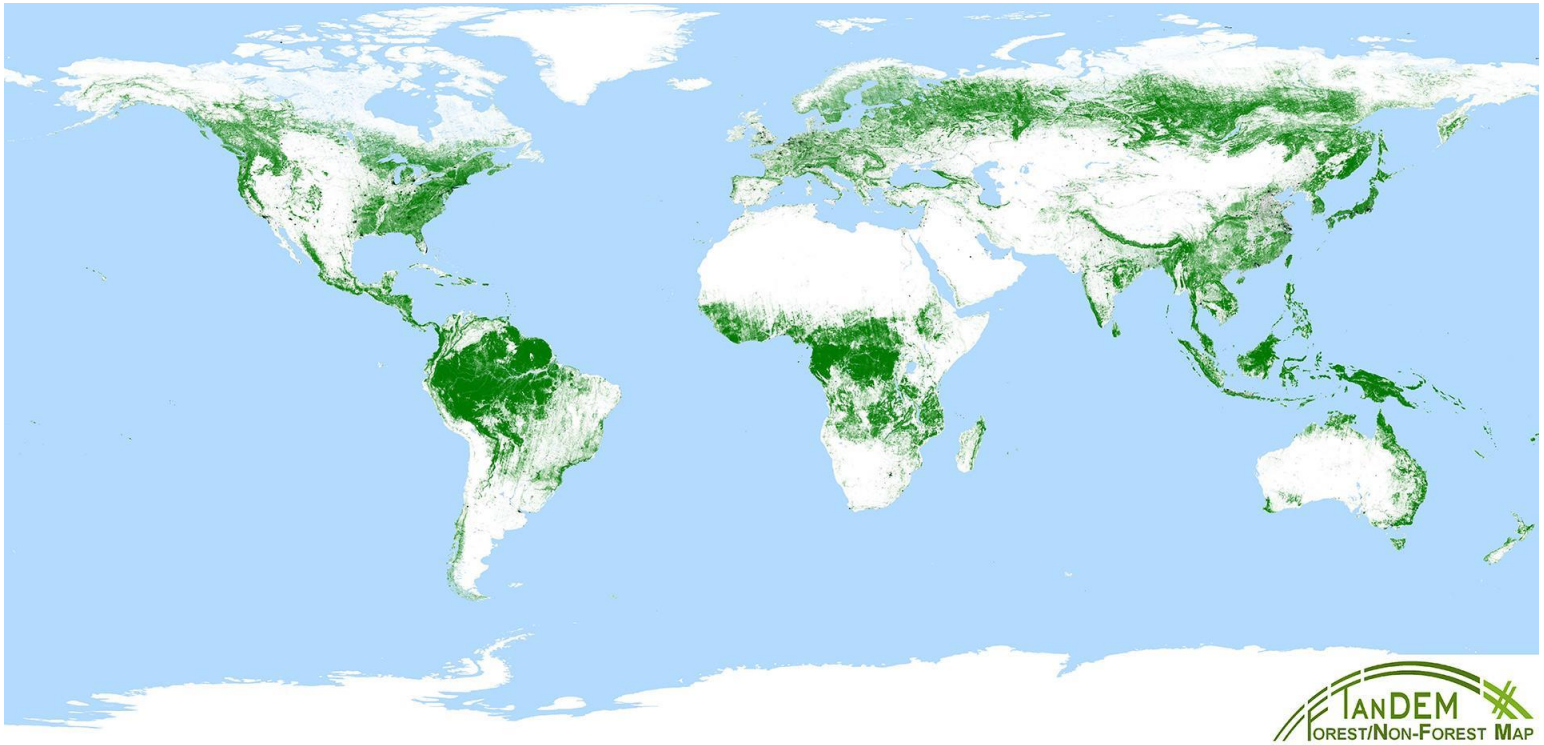
Results with spectral clustering (next step)



Spectral clustering (modified)

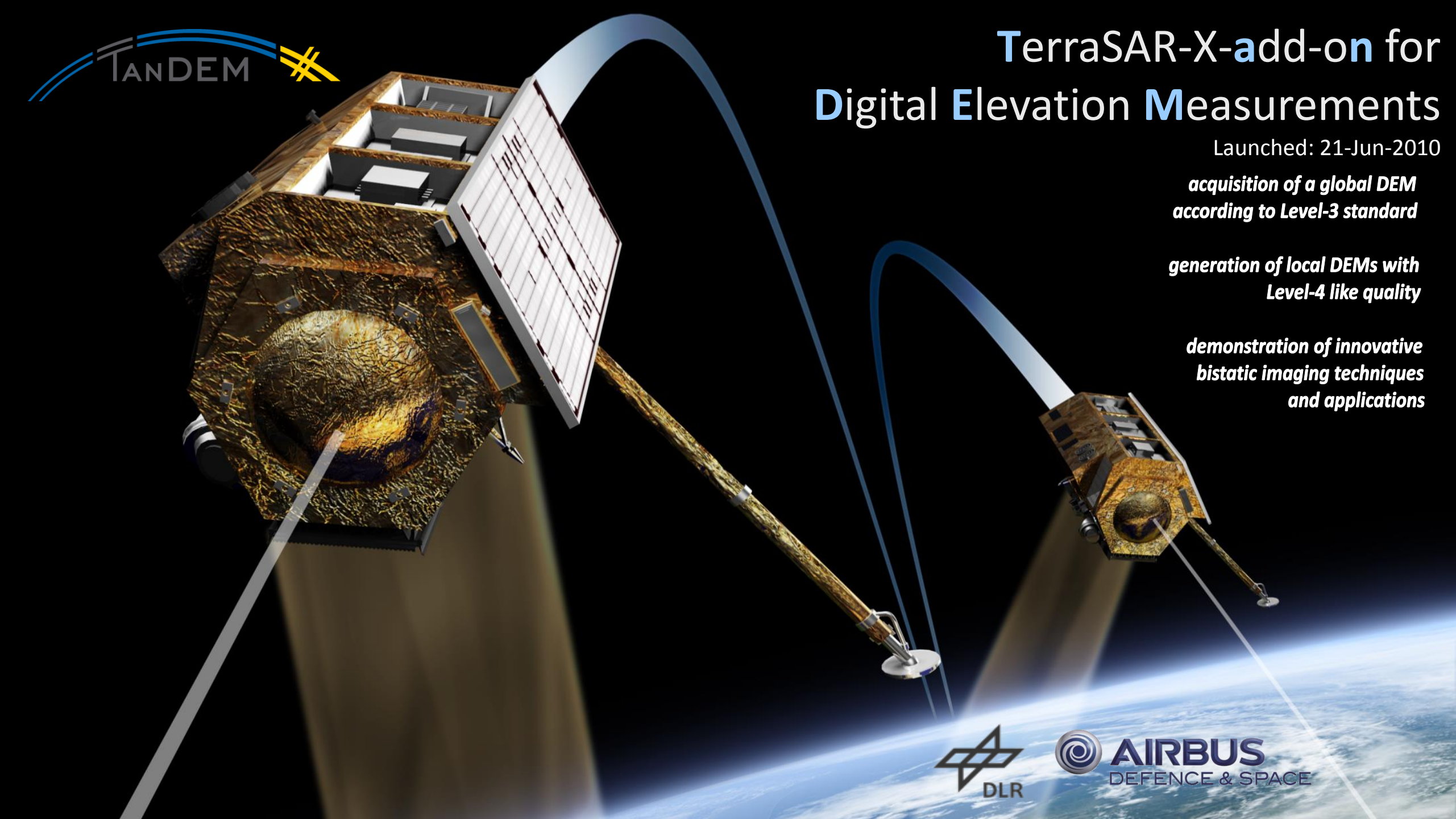
- Build graph Laplacian matrix $\mathcal{L} = D - A$
- Compute eigenvalues and eigenvectors of $\mathcal{L}$
- Cluster K smallest eigenvectors with K-means

outlier detection algorithm (DBSCAN, OPTICS-OF)



Global TanDEM-X forest map*

*slides provided by Dr. Paola Rizzoli and Dr. Manfred Zink


TanDEM
TerraSAR-X-add-on for
Digital Elevation Measurements
Launched: 21-Jun-2010
acquisition of a global DEM
according to Level-3 standard
generation of local DEMs with
Level-4 like quality
demonstration of innovative
bistatic imaging techniques
and applications
DLR
AIRBUS
DEFENCE & SPACE

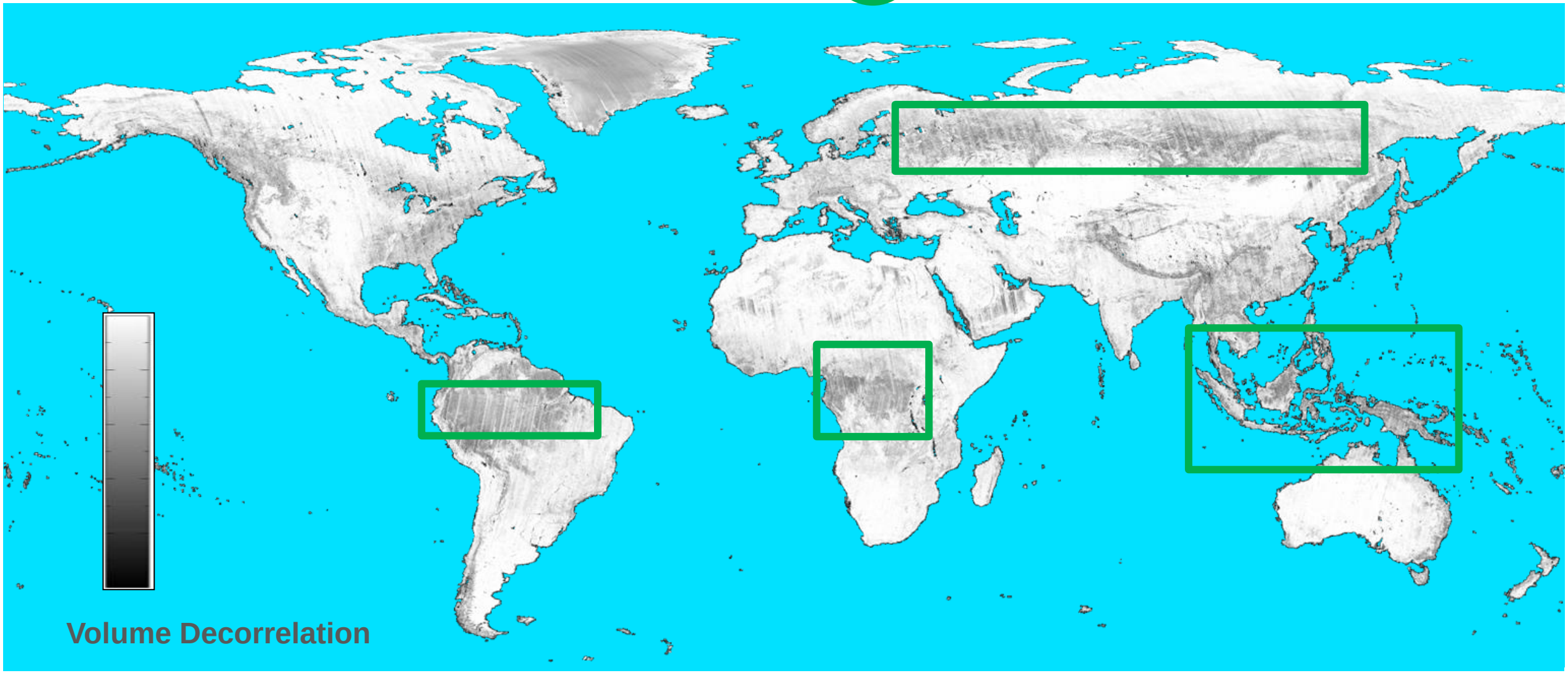
TanDEM-X Interferometric SAR Data Set

- At least 2 global bistatic coverages (global DEM)
- Acquisition time span utilized: 2011 – 2016
- About 500,000 scenes (30 km x 50 km), final posting of 12 m (**350 MB/scene)**

Quicklook Mosaics – First Global Coverage

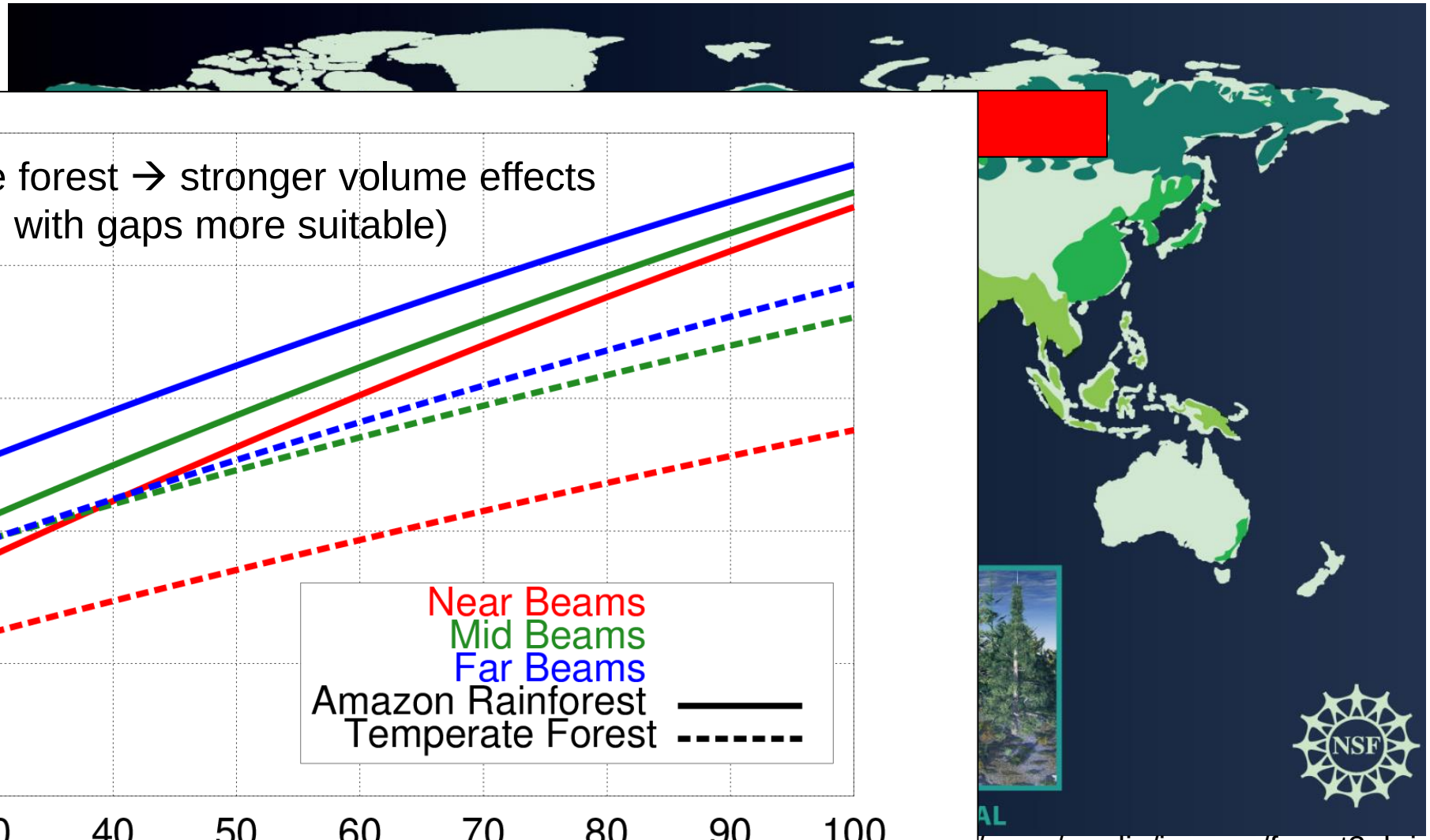
Volume Correlation Factor

$$\gamma = \gamma_{Temp} \cdot \gamma_{SNR} \cdot \gamma_{Amb} \cdot \gamma_{Rg,Az} \cdot \gamma_{Vol} \cdot \gamma_{Quant}$$



Volume Decorrelation

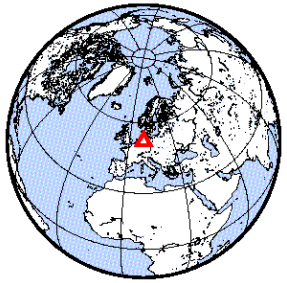
Data Training for Different Forest Types



/mmg/media/images/forest3_h.jpg

Forest/Non-Forest Classification with Fuzzy Clustering

Temperate Forest, Germany

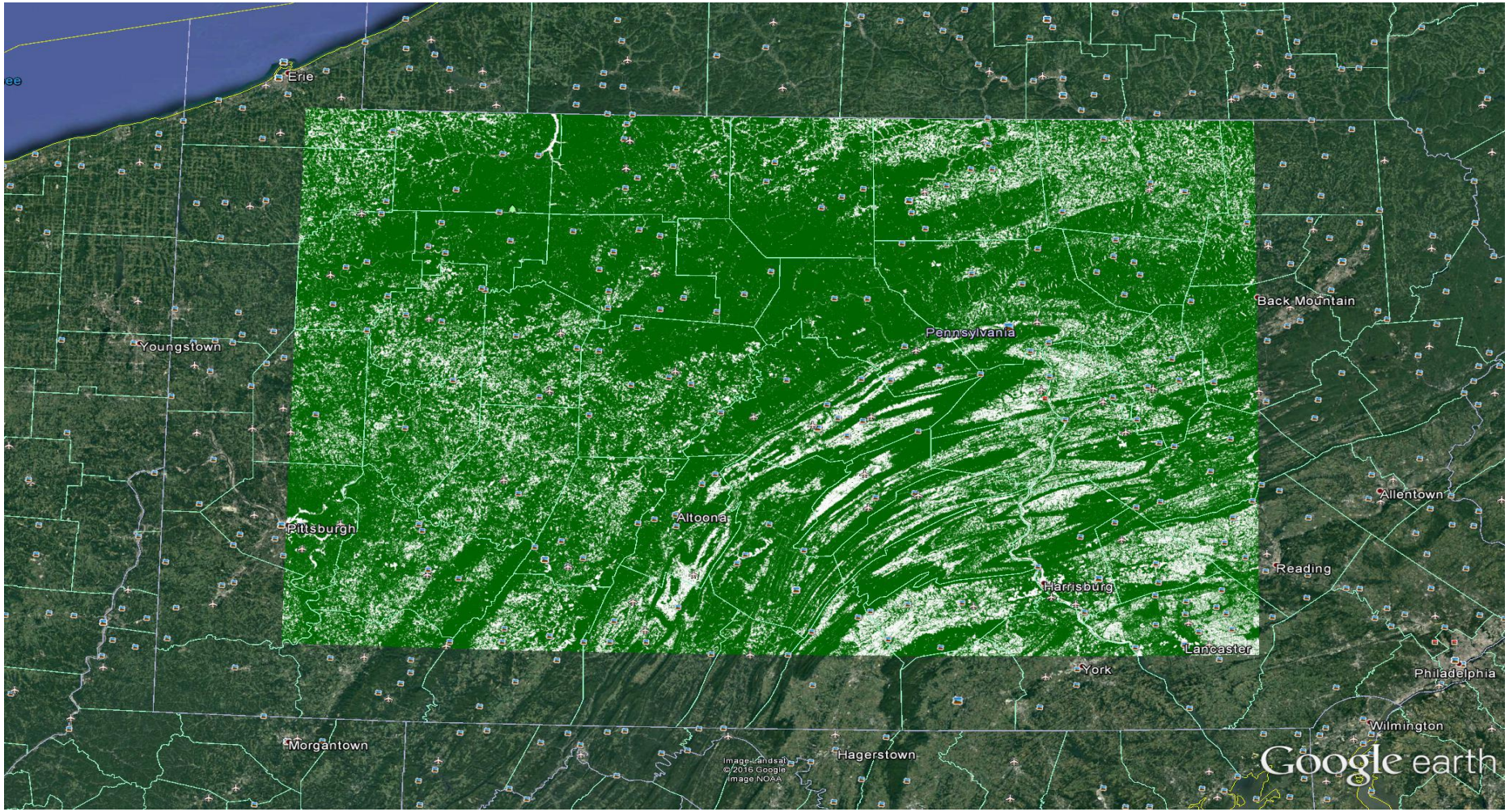


Optical Image

DLR

Mosaicking of Multiple Coverages

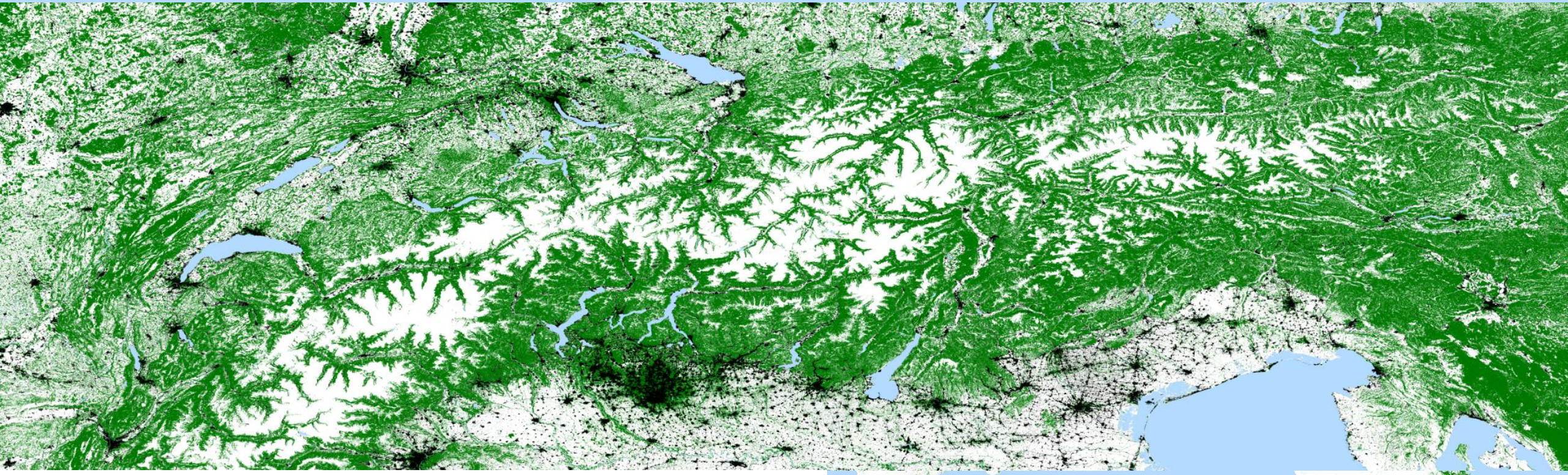
Ex: Temperate Forest, Pennsylvania



Quality-dependent optimized mosaicking logic

Global TanDEM-X Forest/Non-Forest Map

The Alps



Product resolution 50 m x 50 m
Freely available for scientific purposes
Download at …. insert link ……

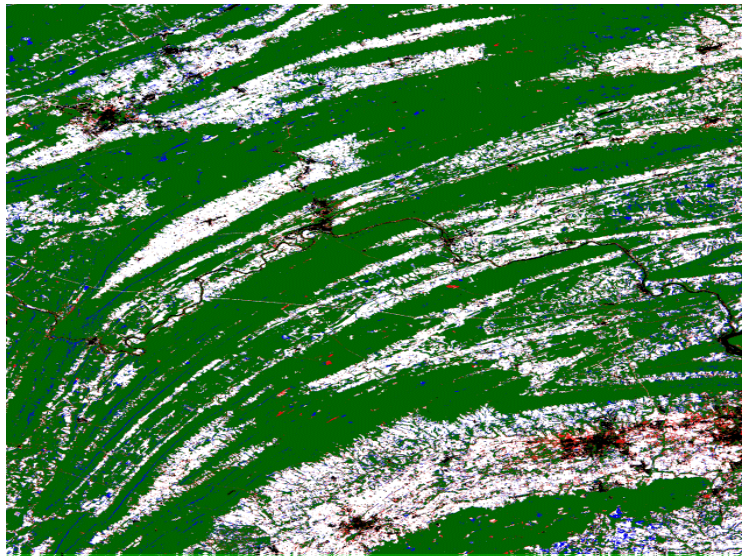
Microwaves and Radar Institute - DLR

TanDEM
Forest/Non-Forest Map

Validation Example: Pennsylvania

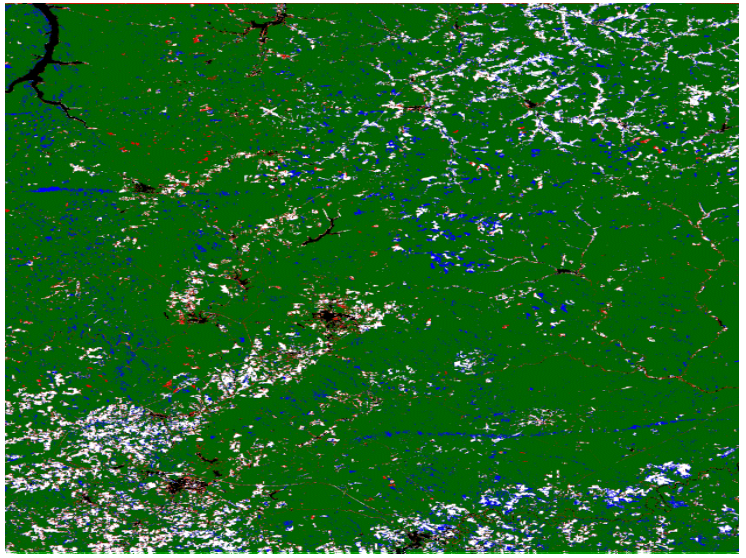
Comparison against Lidar/Optic Validation Reference (LO) (1m resolution)

N40W078



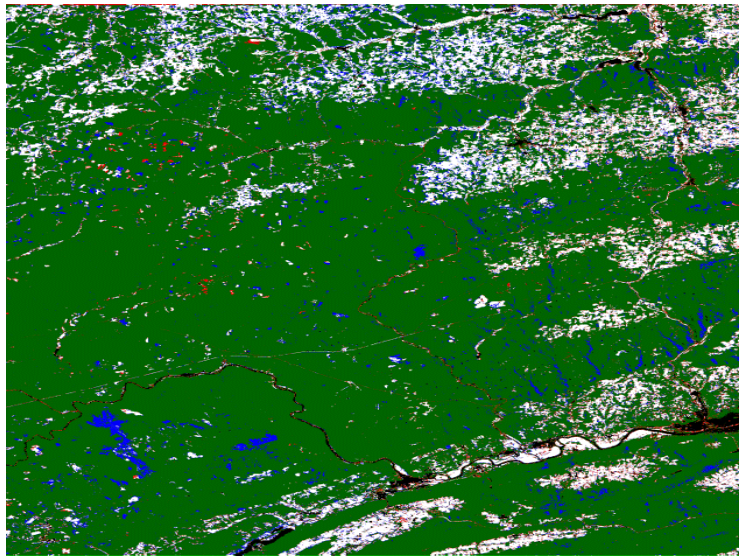
91% Accuracy

N41W079



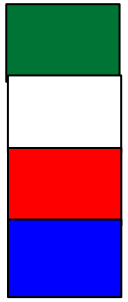
92% Accuracy

N41W078



92% Accuracy

Confusion Matrix:



TDX forest / Reference forest

TDX non-forest / Reference non-forest

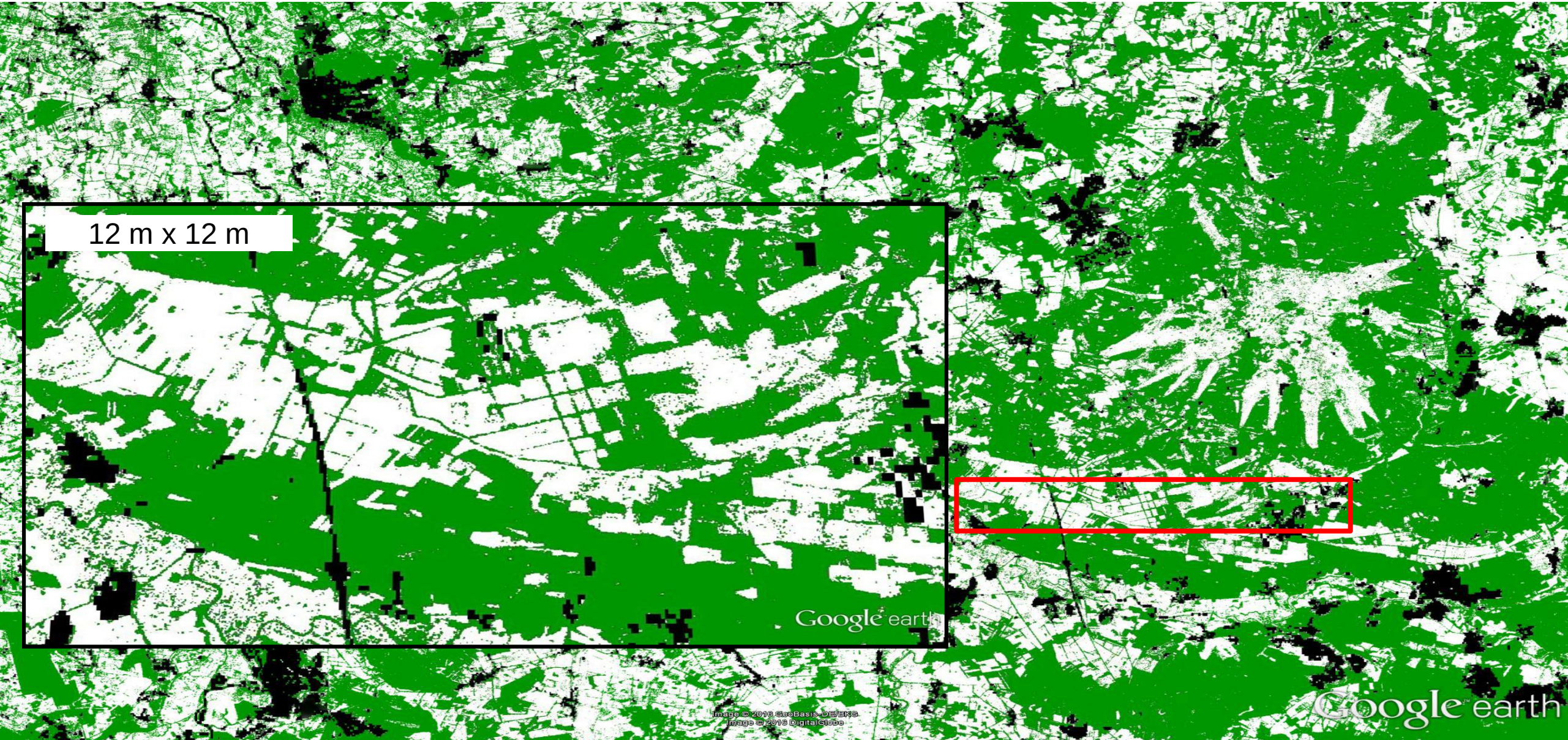
TDX forest / Reference non-forest

TDX non-forest / Reference forest

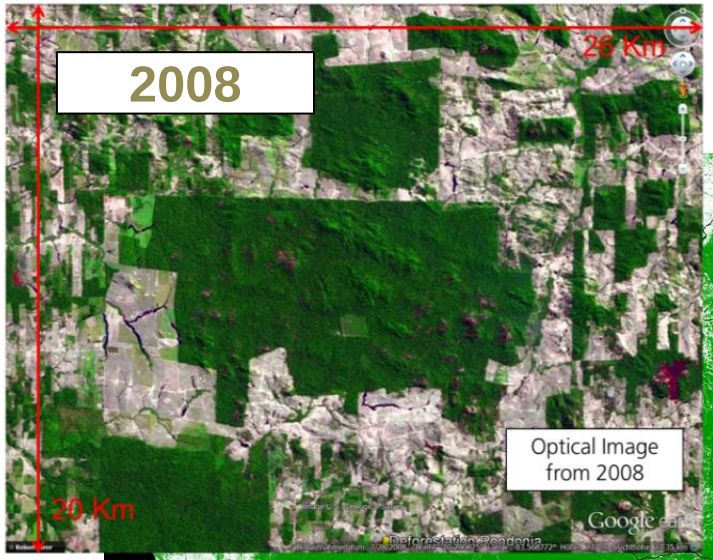
$$\text{Accuracy} \triangleq \frac{TP+TN}{TP+TN+FP+FN}$$

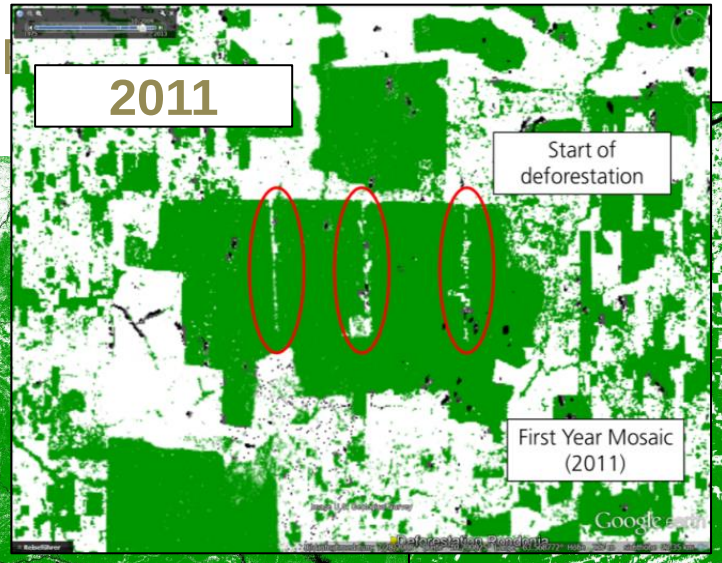
High-Resolution Maps

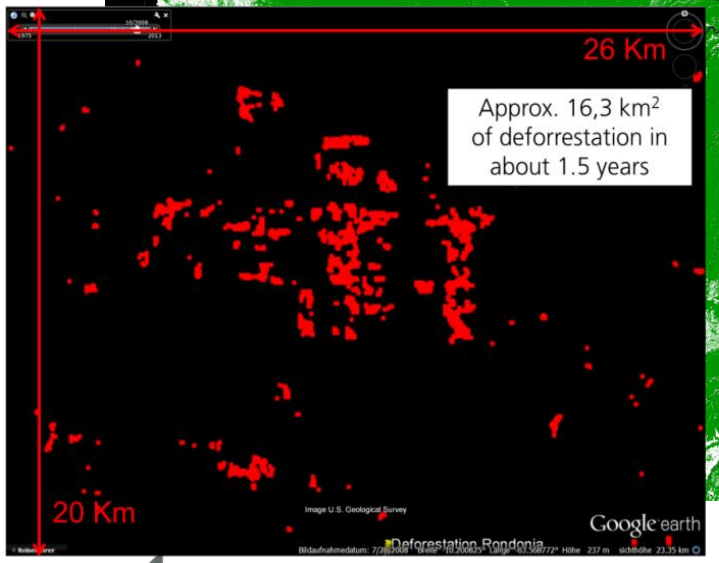
12 m x 12 m

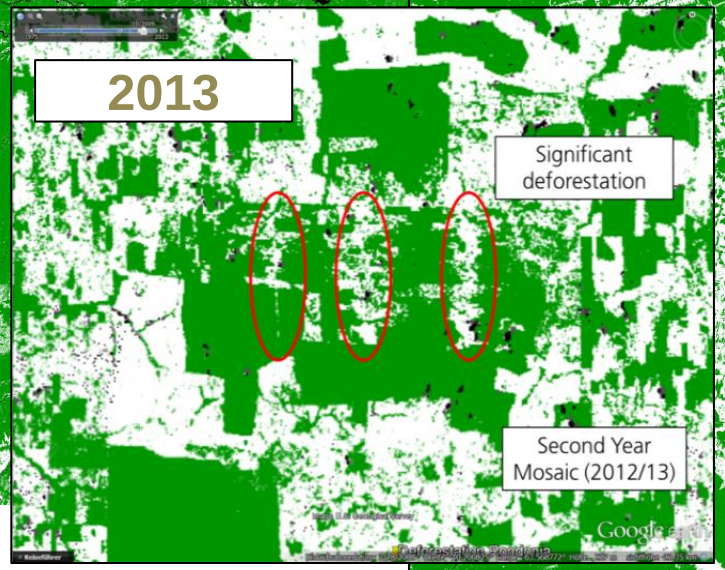


Potentials for change detection

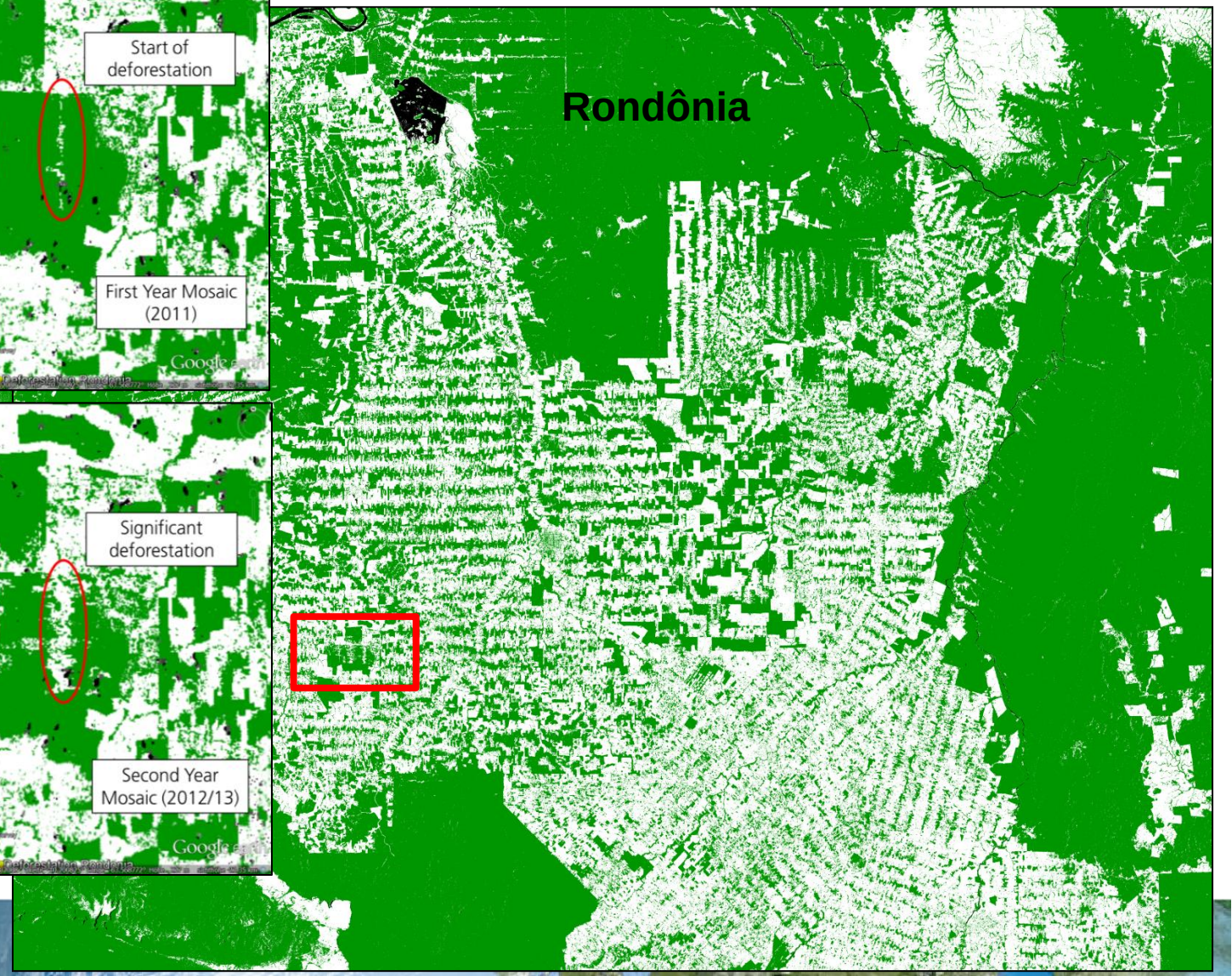








Changes between 2011 and 2013



DLR

High-resolution time-series for change monitoring

9th Sept. 2016 (reference)

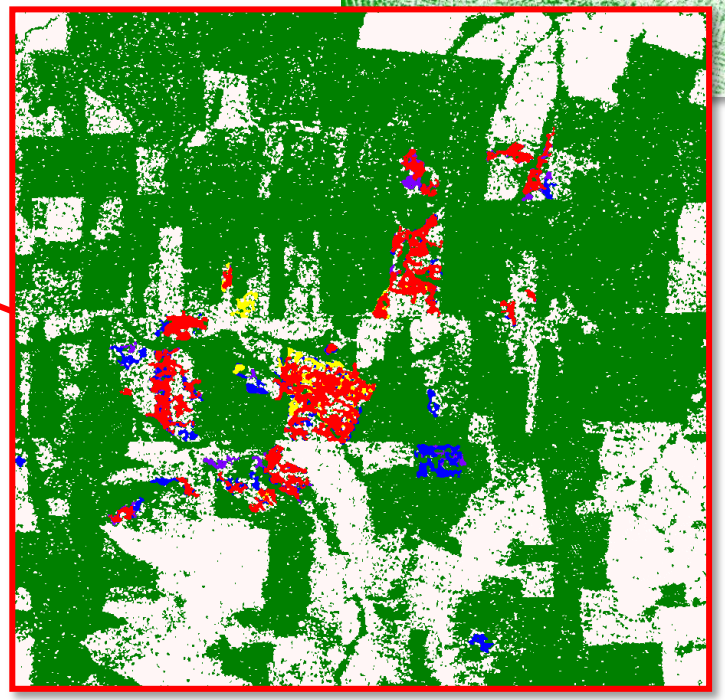
19th Nov. 2016

15th Feb. 2017

31st Mar. 2017

Overall

Rondônia



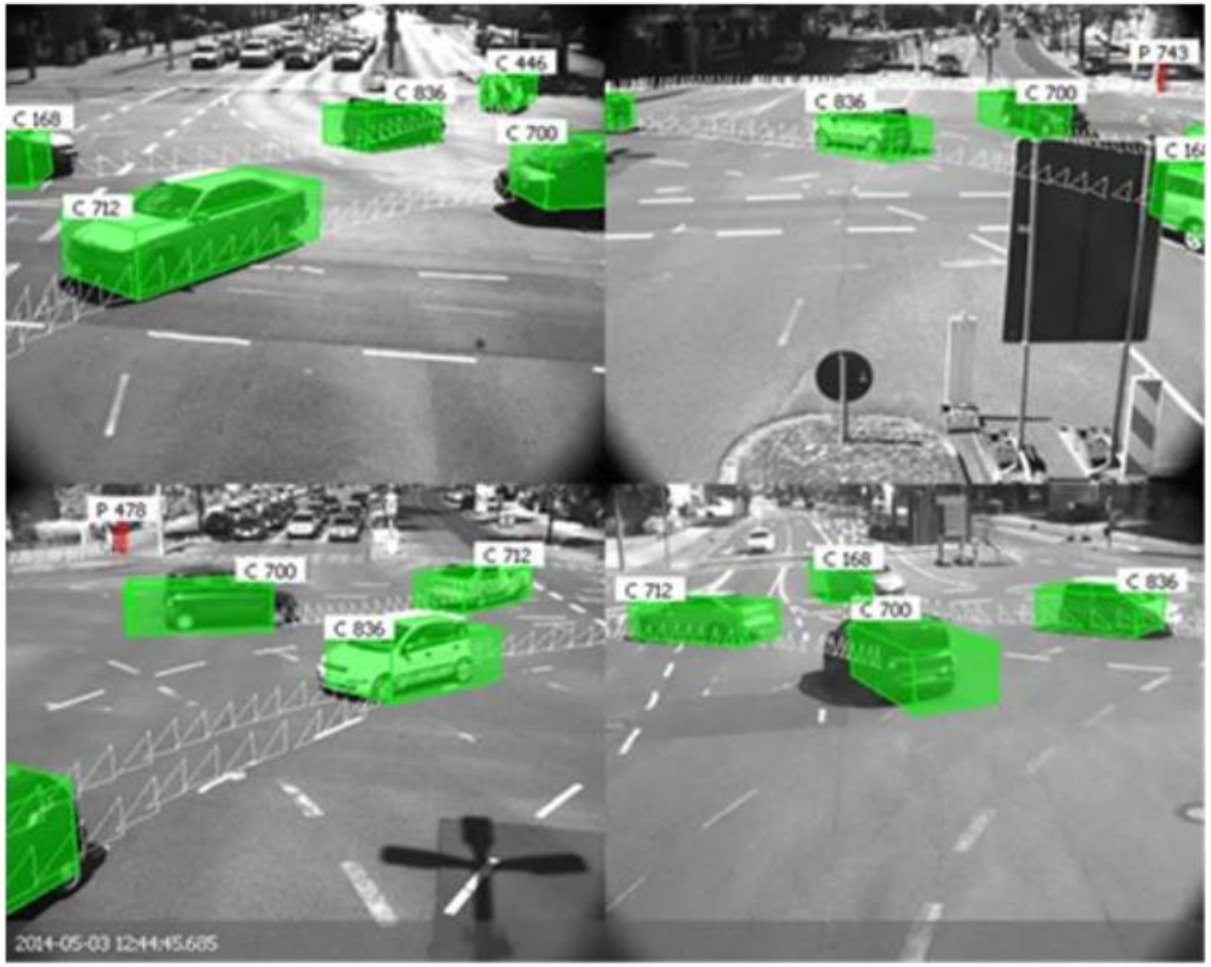
Sep.-Nov. 2016
Nov. 16-Feb. 17
Feb.-Mar. 2017

Forest Loss
~ -25 ha
~ -2.6 ha
~ -6.7 ha

DLR

Traffic monitoring and analysis*

*slides provided by Christian Linder and Kay Gimm

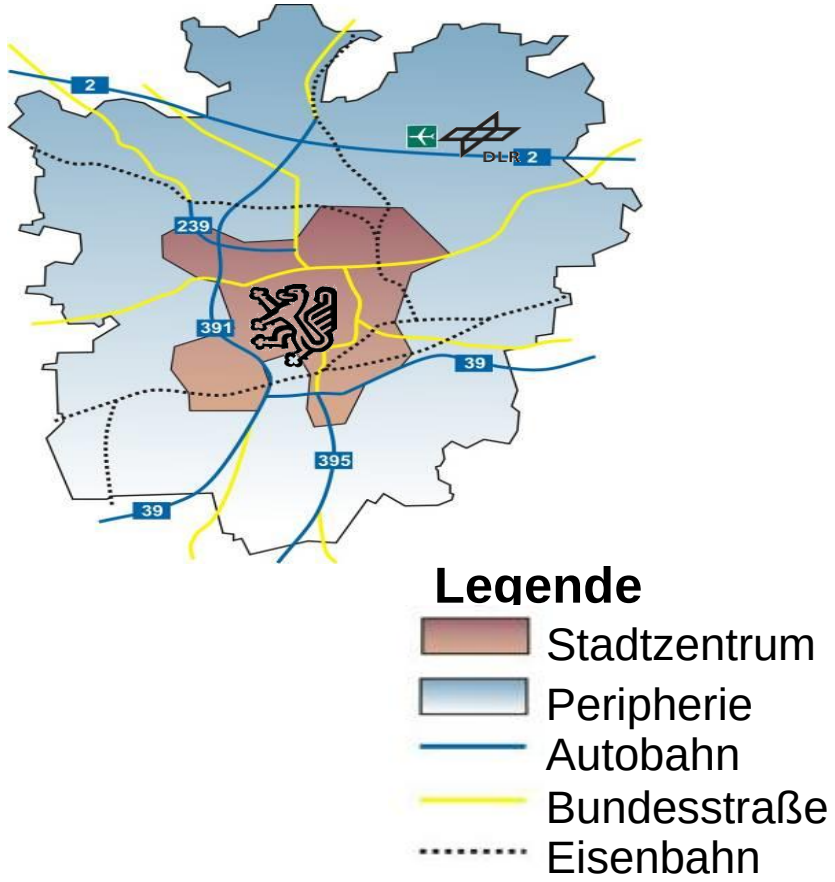


note: video has been replaced by this image

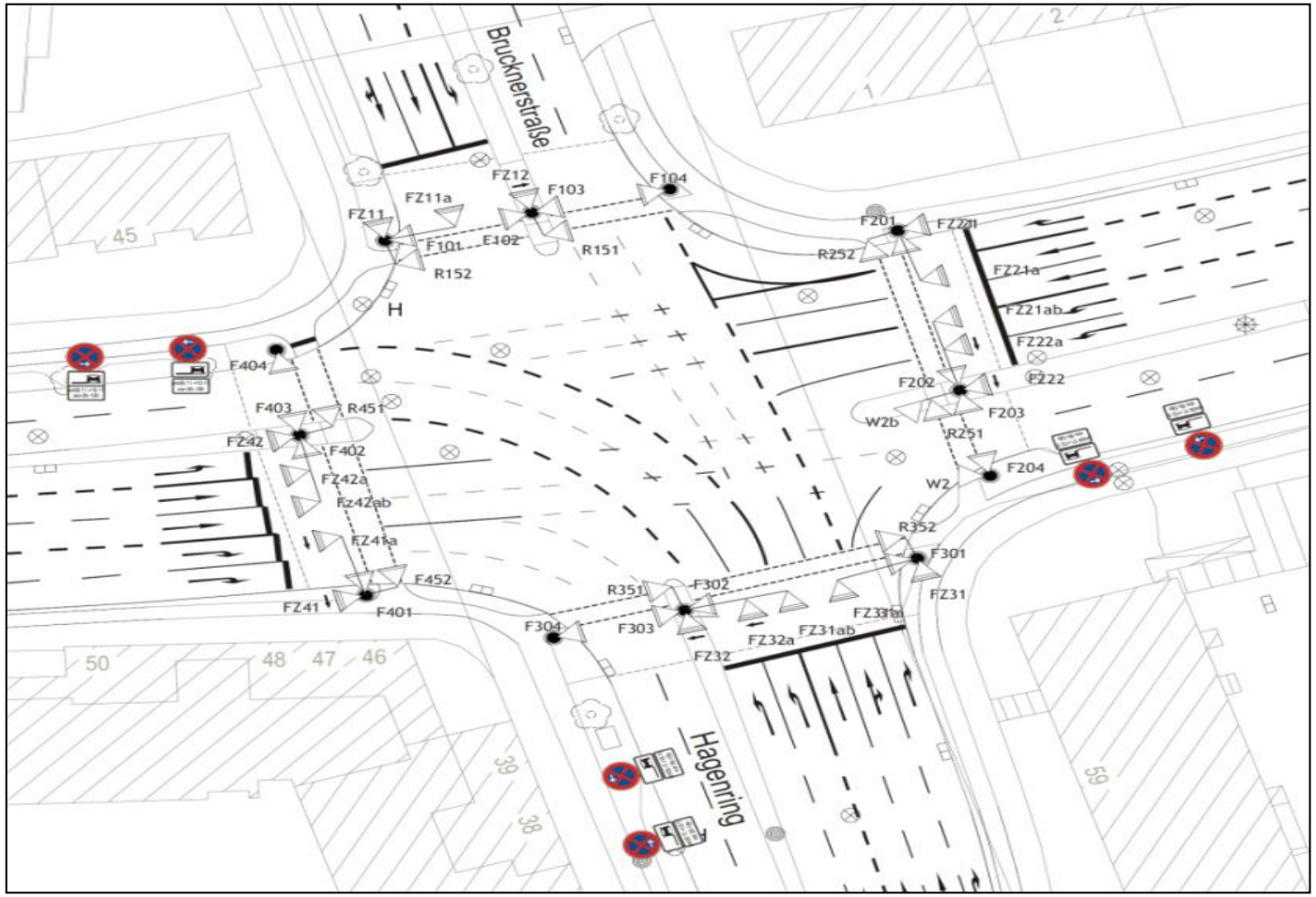
Anwendungsplattform intelligente Mobilität (AIM)

Als Plattform für anwendungsorientierte Wissenschaft, Forschung und Entwicklung in der Dimension einer Stadt bietet AIM eine große Bandbreite an Technologiebausteinen:

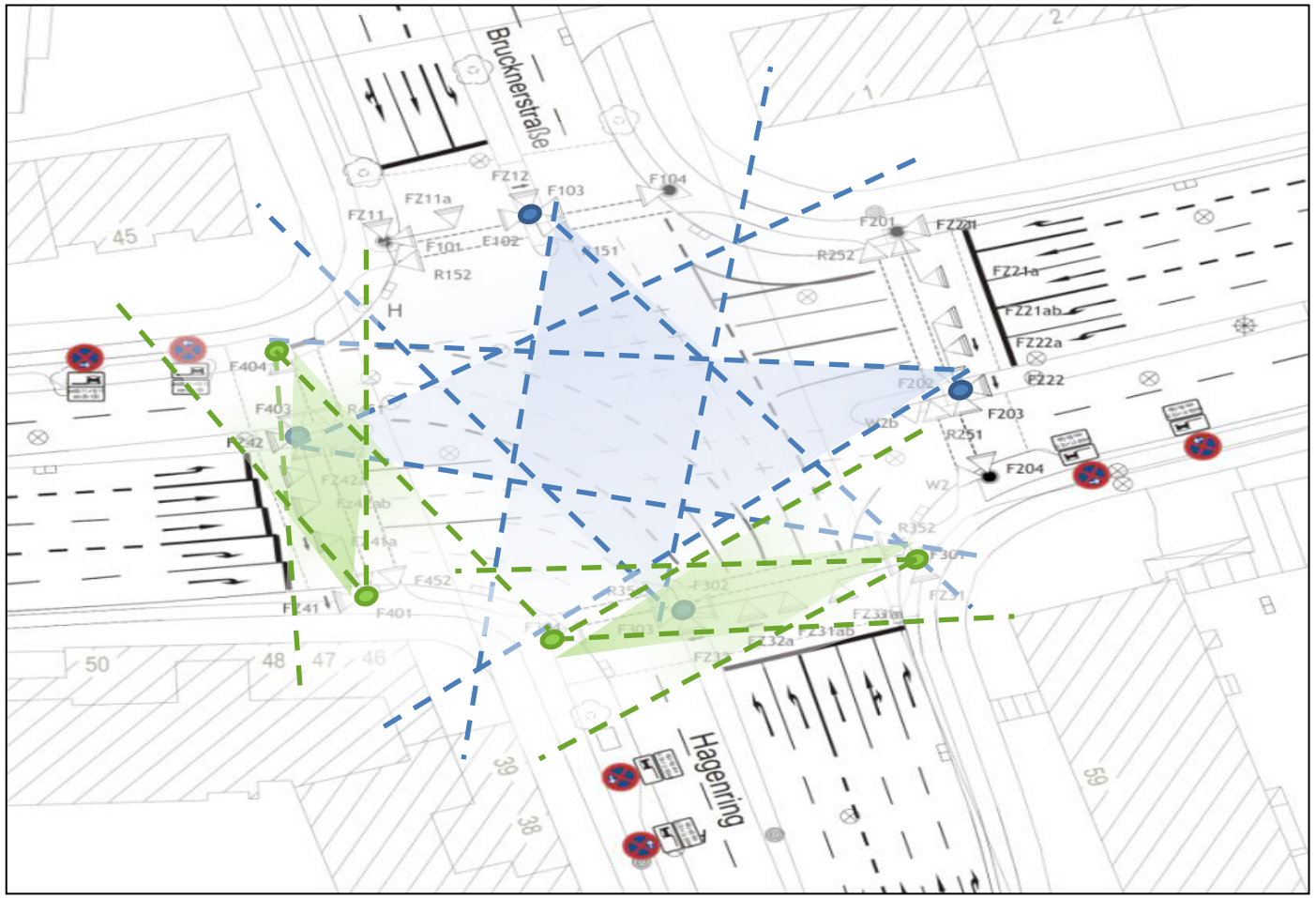
- Sensorische Erfassung der Realität des städtischen Verkehrsumfelds
- Anwendung von Simulationen zur Ableitung tragfähiger Erkenntnisse für den realen Verkehrsablauf
  - großräumige (makroskopische) Aspekte
  - kleinräumige (mikroskopische) Aspekte
- Gezielte Beeinflussung des Verkehrsgeschehens
  - Datenübertragung zwischen Infrastruktur und Verkehrsteilnehmern
  - Einbettung in vorhandene Teilsysteme des städtischen Verkehrsmanagements

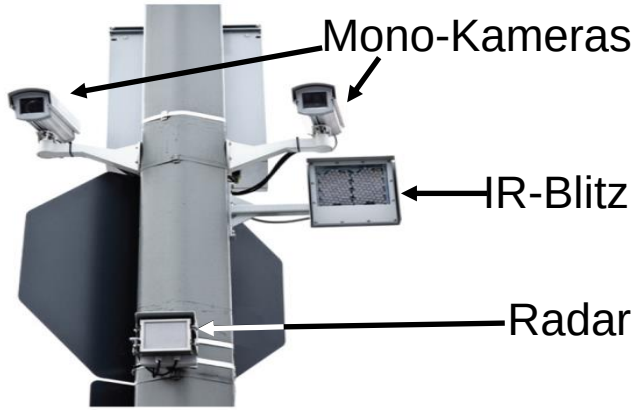


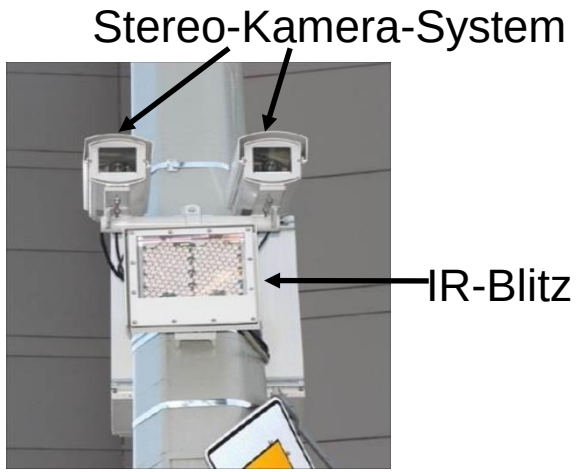
AIM Forschungskreuzung – Sensorischer Aufbau



AIM Forschungskreuzung – Sensorischer Aufbau
Gemeinsamer Erfassungsbereich

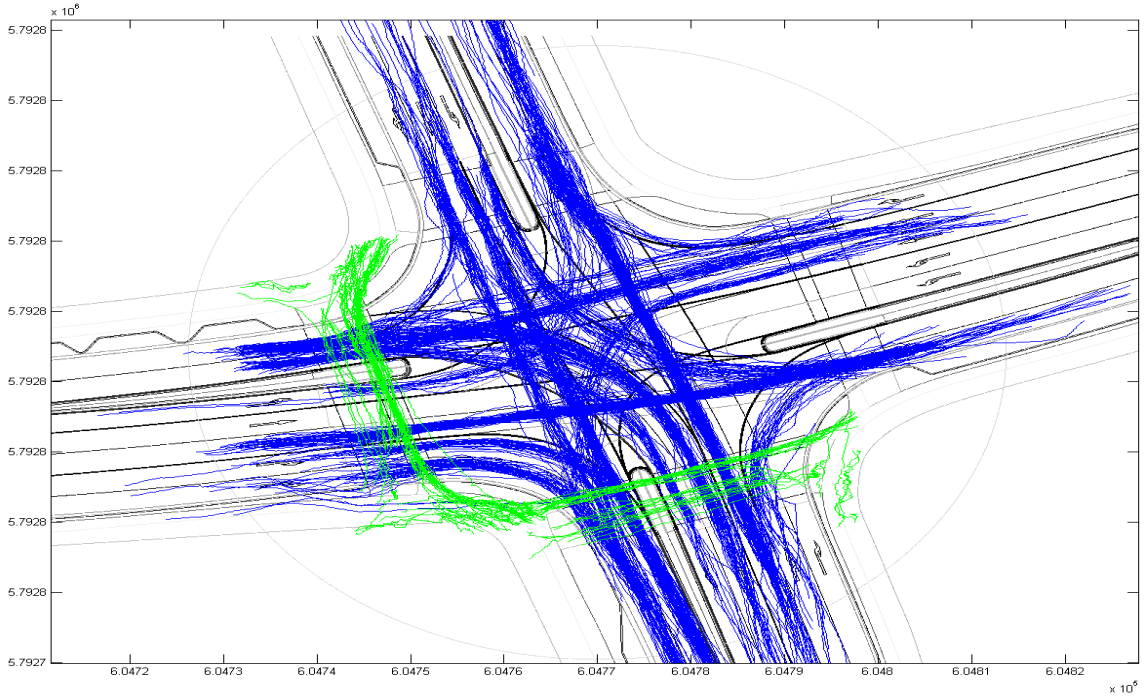


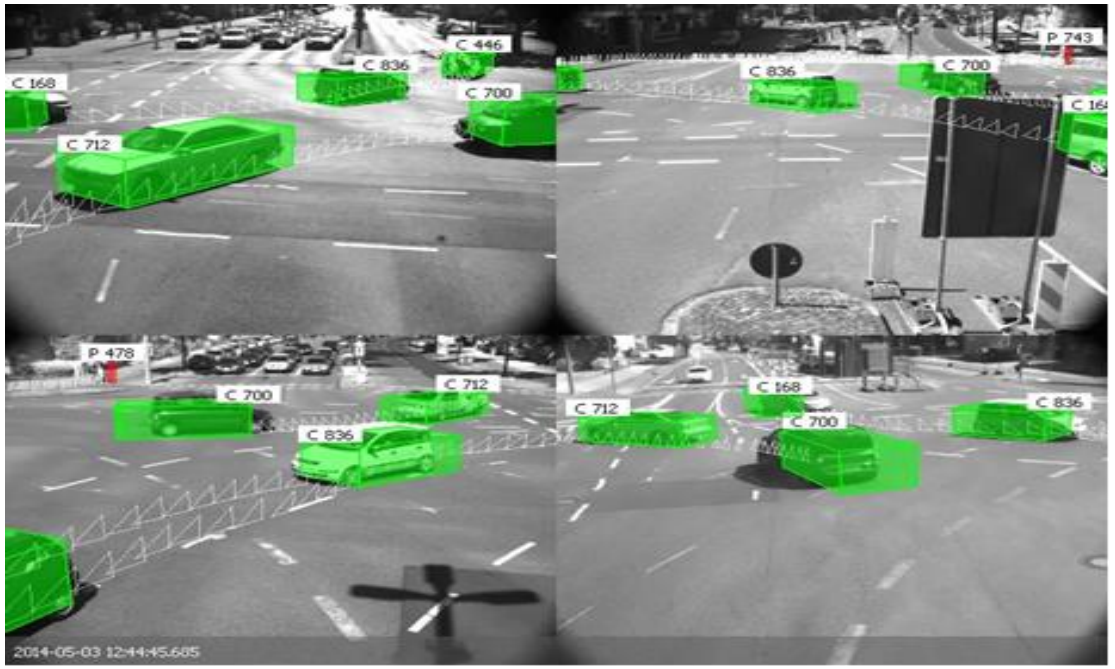




Ergebnisdaten Trajektorien und Videodaten

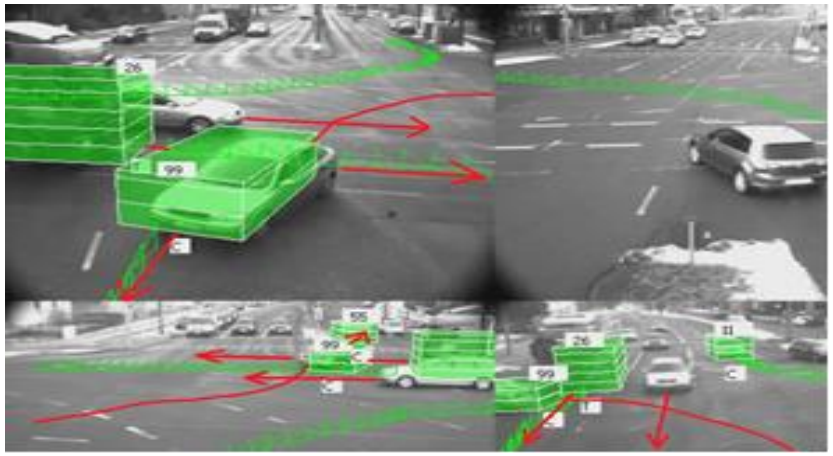
- Trajektorien mit 25Hz Datenrate
- Zustandsdaten umfassen neben dem Zeitstempel:
  Position, Geschwindigkeit, Beschleunigung, Abmessungen, Klassifikation

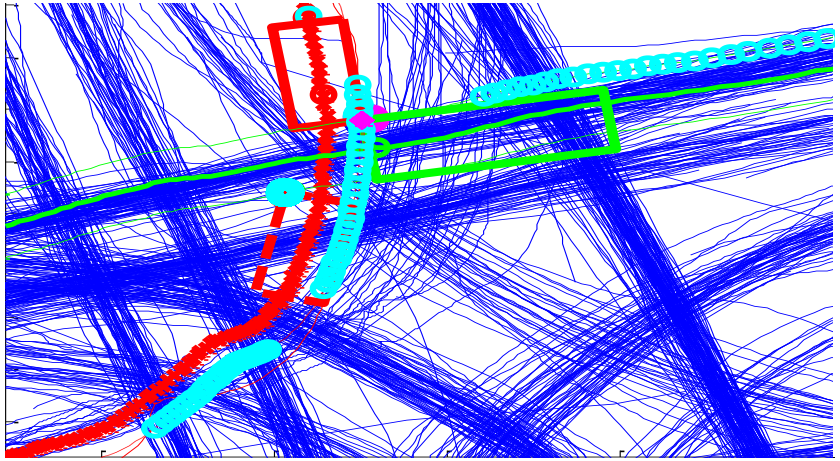




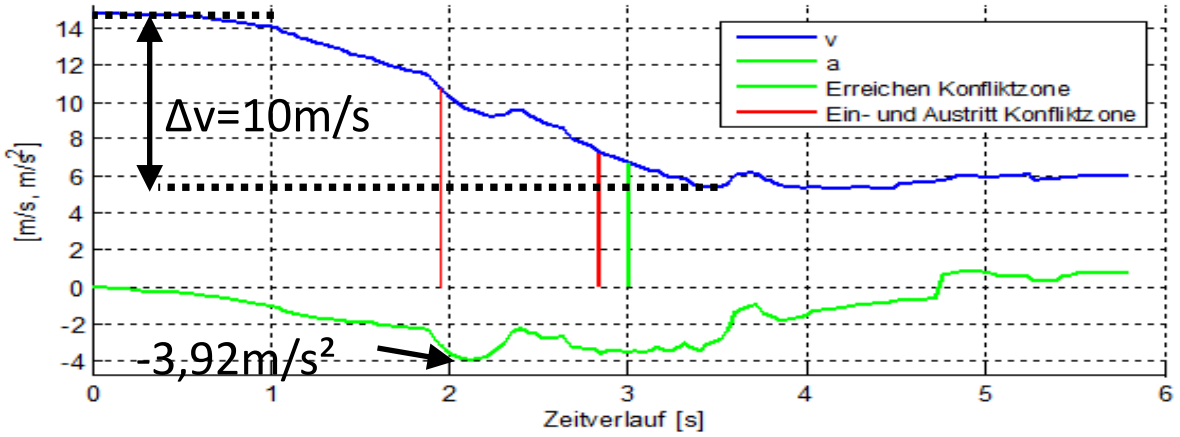
Situationsanalyse – Linksabbieger mit Gegenverkehr
Einzelfallanalyse: Beschreibung

PKW (ID99) vs.
LKW (ID26)

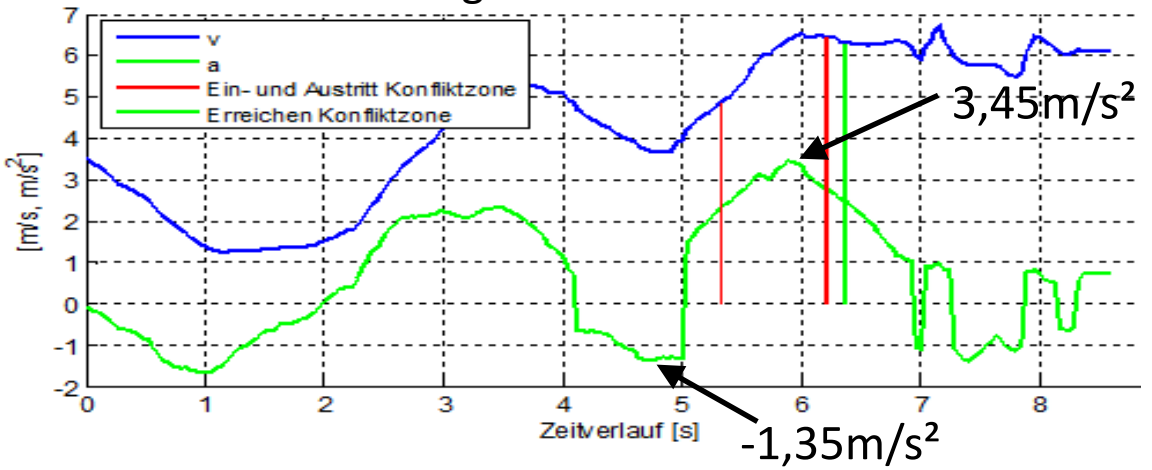




Längsverkehr: LKW



Abbiegender Verkehr: PKW



Großinvestitionen AIM Cluster 3a (im Zeitraum 2018 -2020)

- **AIM-Backend (Hintergrundsystem)**
  Erneuerung Oracle Datenbank (aktuell Exadata X3 -> Exadata X7)
  Erneuerung / Erweiterung Dateispeicher (NetApp FAS3200)
  Erweiterung Datenbanksystem auf Standort DLR-BA
  ca. 60% der Mittel

- **AIM-Referenzstrecke**
  Austausch der LSA Gateways (erlaubt externen Zugriff auf LSA Steuerung und Integration mit V2X Technik)
  Austausch alte V2X Kommunikationseinheiten (Vorhandene HW beherrscht aktuelle Protokolle nicht, Beibehaltung State-of-the-art)
  ca. 25% der Mittel

- **AIM-Forschungskreuzung**
  Ersatz alte Rechenknoten zur Bildverarbeitung (neue Server Gener.)
  Tausch der Kamerasysteme (Höhere Auflösung, mehr Datenmenge)
  ggf. Verlegung zusätzliche Glasfaser für höchste Datenraten
  ca. 9% der Mittel

- **AIM-Mobile Aufbauten**
  Ersatz alte Rechenknoten zur Bildverarbeitung (neue Server Gener.)
  ca. 6% der Mittel

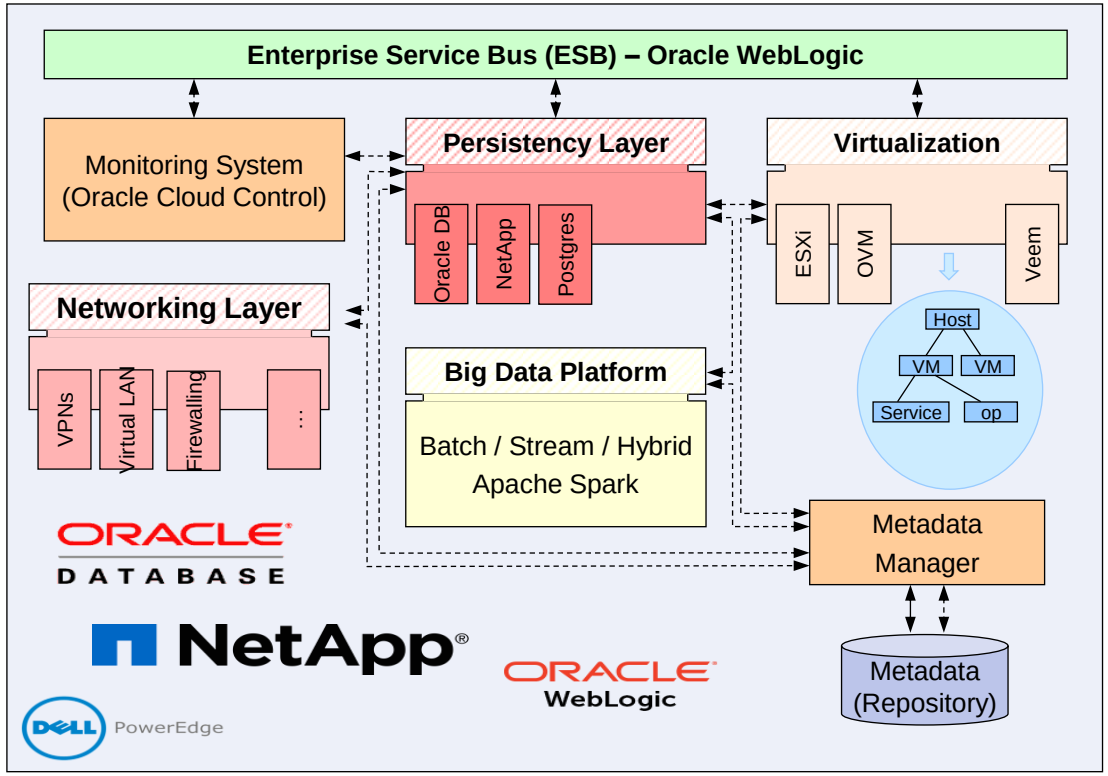


Figure: Backend IT-Architecture

Conclusion and outlook

- **Research topics such as Big Data and AI are of central importance at DLR.**

- **A lot of knowledge exists, but there is the risk of two DLR institutes developing similar methods.**

- **DLR's project Big-Data-Plattform shall establish common standards at DLR:**
  - **usage of standardized interfaces** → white paper
  - **develop databases for distributed and heterogeneous data** → recent milestone
  - **if possible, use established open-source software solutions** → HeAT, HDF5
  - **improve (parallel) scaling of applications** → HPC, HPDA clusters at DLR, cloud services
  - **knowledge database** → work in progress
  - **….**

And there is still more to come …