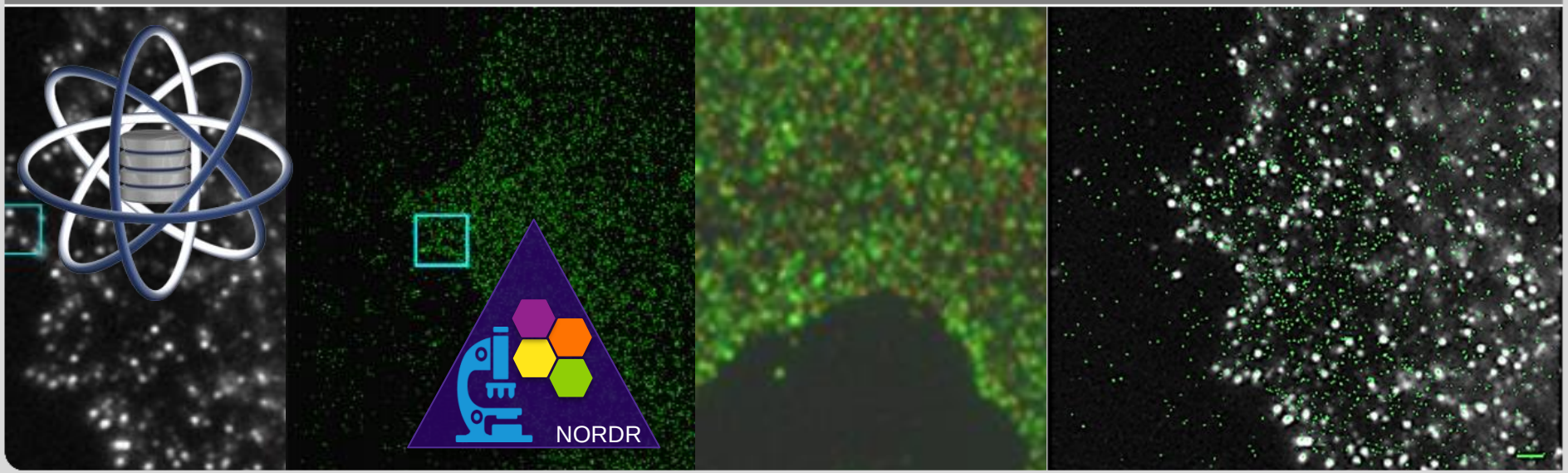
Karlsruhe Institute of Technology

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

# Enabling Large Data Processing in Nanoscopy Open Reference Data Repository using LAMBDA

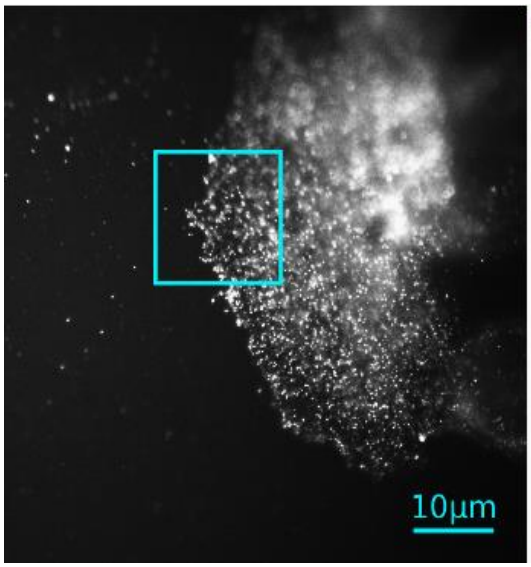
**Ajinkya Prabhune,**
**E. Schmitt, J. Hesser, M. Bach, M. Hausmann, R. Stotzka, T. Jejkal, V. Hartmann**

Institute for Data Processing and Electronics (IPE)

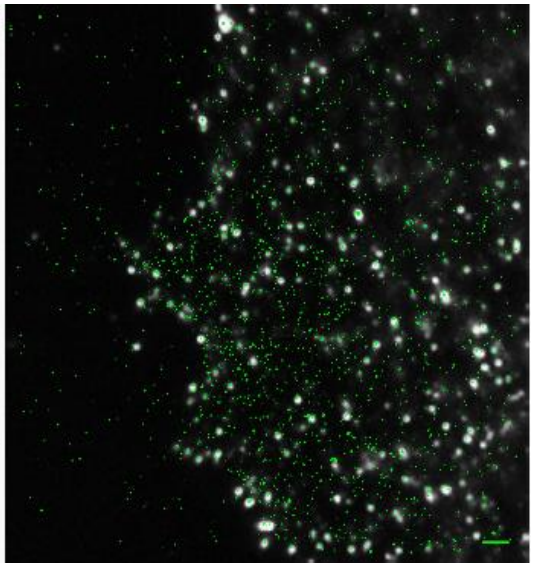


# Nanoscopy

- Novel imaging technique capable of achieving near-molecular resolution in nanometer range



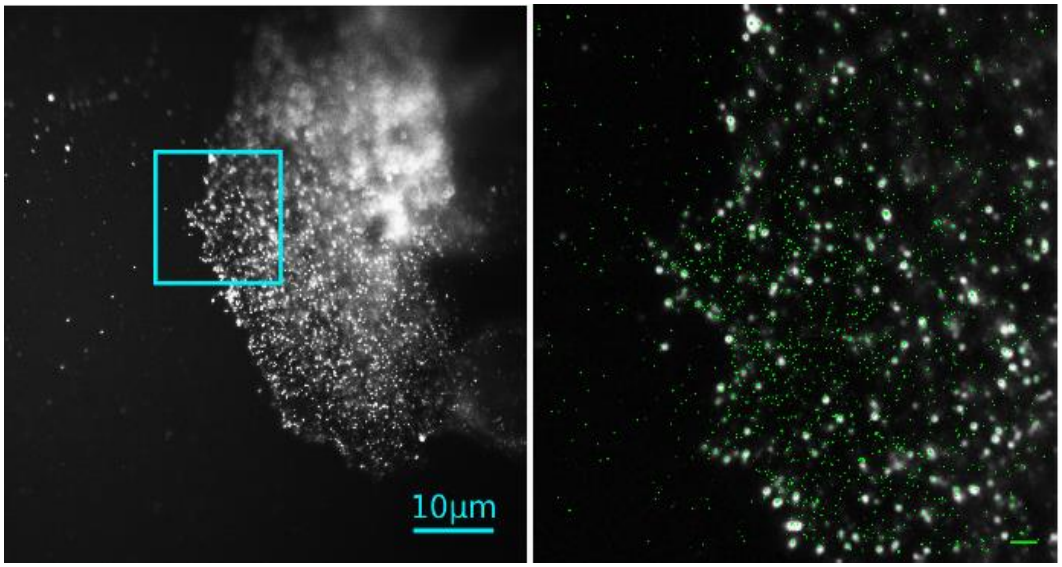

*Microscope image of the whole breast cancer cell*



*Localization image of the same breast cancer cell*

- Interpret the novel images
- Create reference datasets
- Share new scientific insights
- Analyze datasets interactively

# Challenges: Processing extremely large datasets



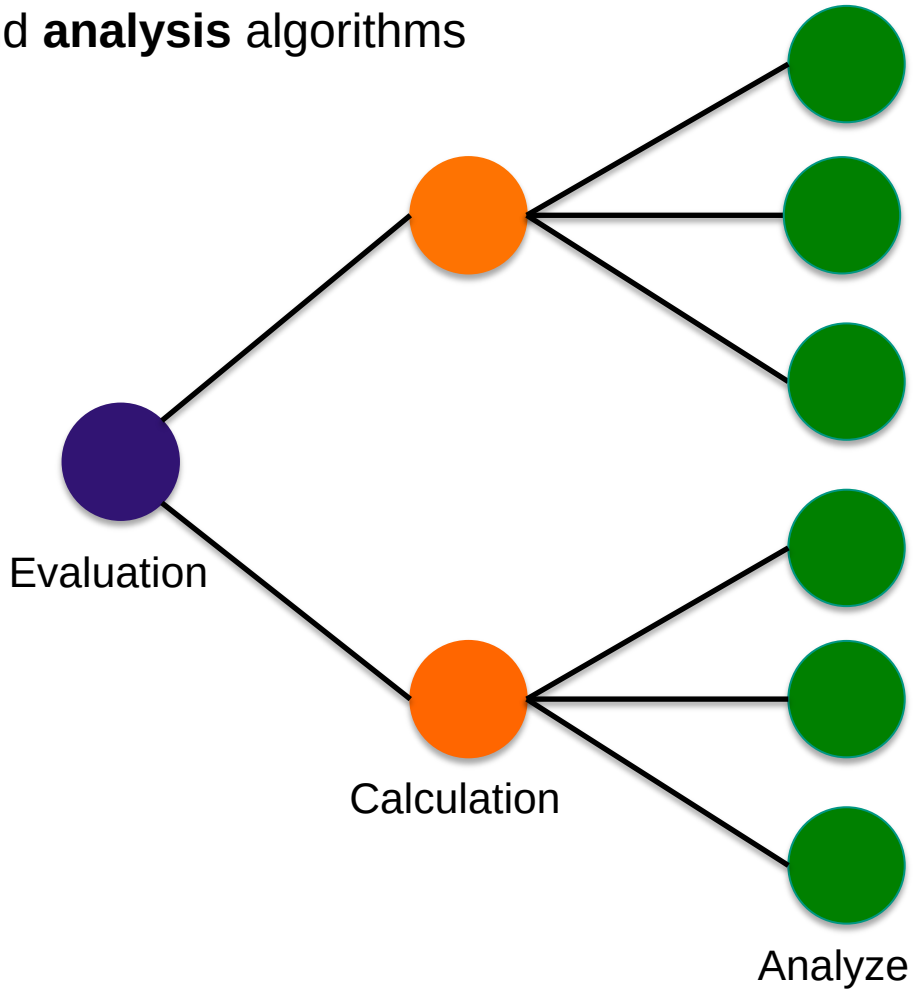

*Microscope image of the whole breast cancer cell*

*Localization image of the same breast cancer cell*

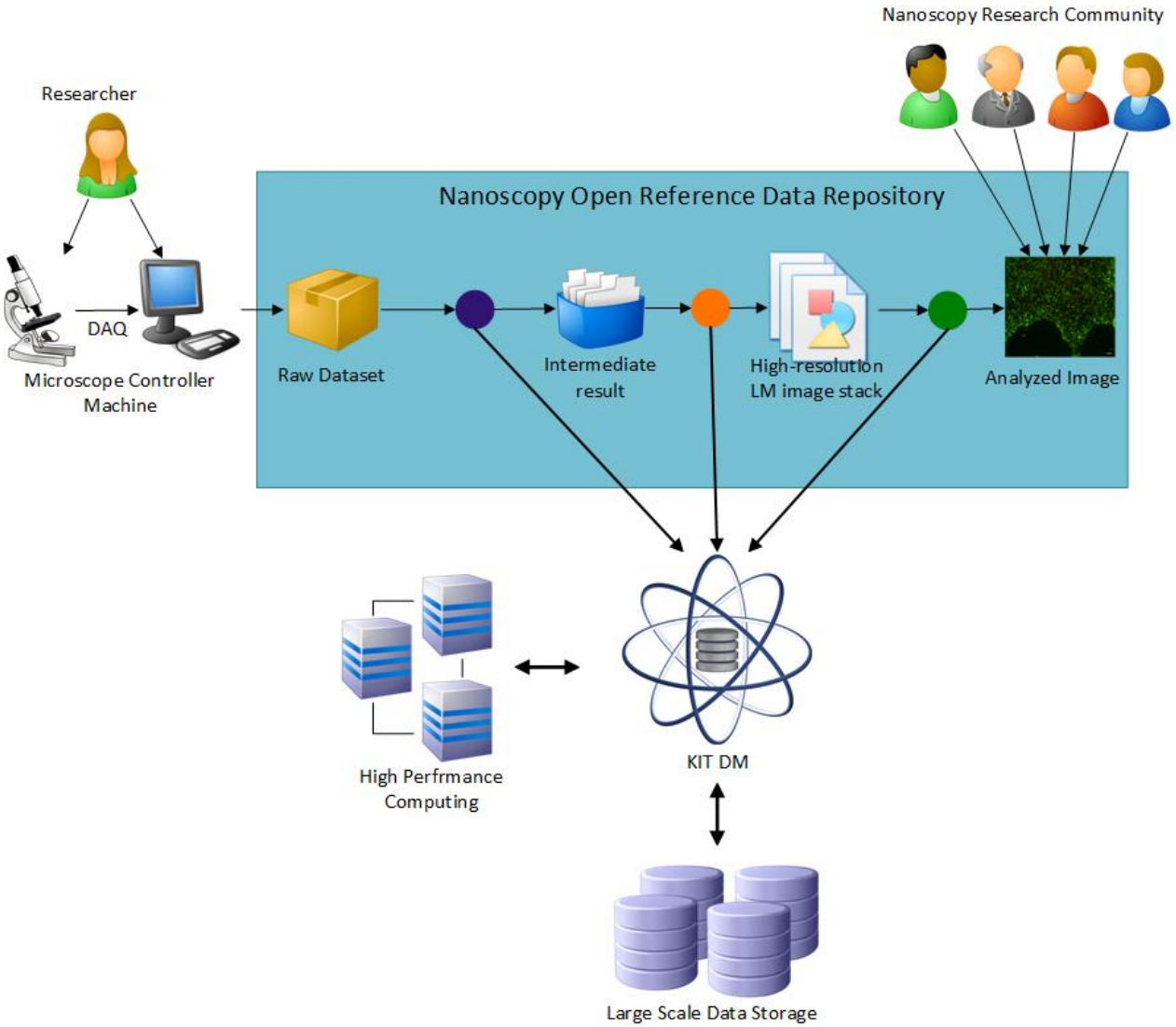
- Each dataset size ~50-80 GB, complete investigation dataset size ~150-200 TB
- Processing of large datasets
- Ability to manage complex scientific workflows
- Provenance information for comparing, analysing and sharing various workflows
- Enable data reproducibility

# Goals

- Define and manage **systematic workflows** of Nanoscopy experiments
- Integrate **evaluation**, **calculation** and **analysis** algorithms with KIT DM LAMBDA
  - **Evaluation**
    - fastSPDM
  - **Calculation**
    - Orte2NNBild
    - Orte2StdBild
  - **Analyse**
    - AutoCluster
    - AutoDistribution
    - AutoNNDistribution

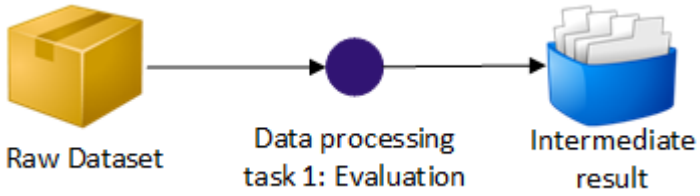


# Nanoscopy scientific workflow



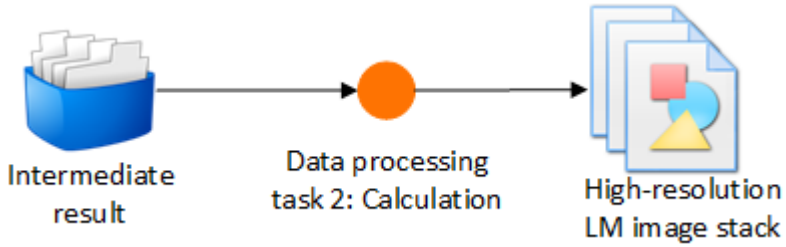
# Data processing task 1: Evaluation

| | |
|---|---|
| Input data | Raw dataset |
| Data size | ~50-80 GB(few sections) → 150-200 TB (total) |
| Processing time | 6-8hrs (few sections) |
| # of algorithms | 1 |
| Algorithm implementation | Matlab |
| Provenance information | Required |
| Automated data ingest | Required |
| Output data | Intermediate result |



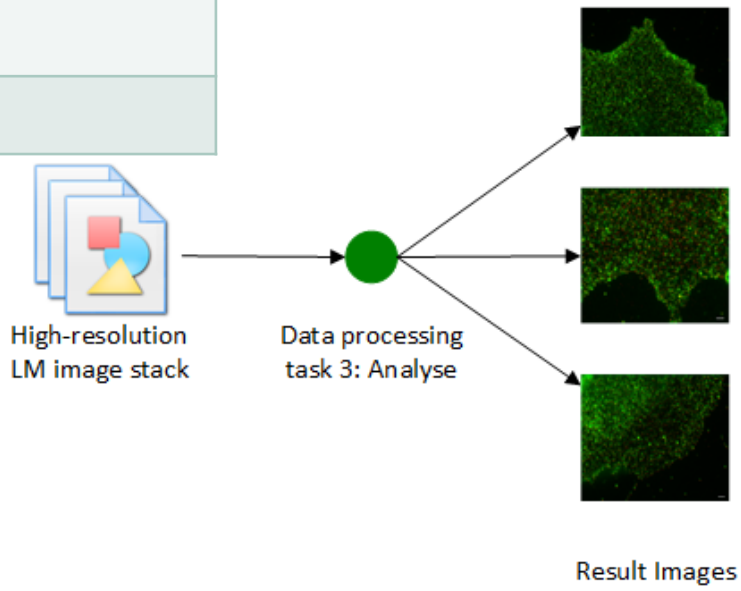
# Data processing task 2: Calculation

| | |
|---|---|
| Input data | Intermediate dataset |
| Data size | ~50-80 MB(few sections) → 150-200 GB (total) |
| Processing time | ~30-45 min (few sections) |
| # of algorithms | 3 |
| Algorithm implementation | Matlab |
| Provenance information | Required |
| Automated data ingest | Required |
| Output data | High resolution LM image stack |



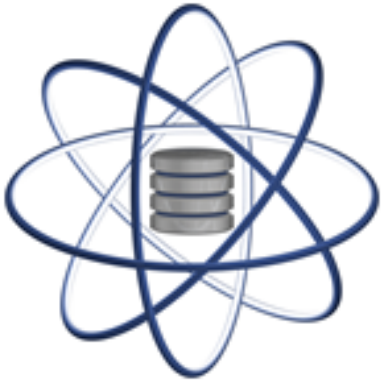
# Data processing task 3: Analyse

| | |
|---|---|
| Input data | High-resolution LM image stack |
| Data size | Few MBs (section) → 10GB (total) |
| Processing time | 5-10 min(few sections) |
| # of algorithms | 4-5 |
| Algorithm implementation | Matlab |
| Provenance information | Required |
| Automated data ingest | Required |
| Output data | Result Images |



# LAMBDA computing service

- Enable processing of large datasets
- Define and execute complex scientific workflows
- Capturing the provenance information of the processing task
  - Interoperability of provenance metadata in OPM and PROV model
- Allow reproducibility of data
- Compare, analyse and improve complex scientific workflows



# Conclusion

- Extend the NORDR with KIT DM LAMBDA computing service
- Possible to define and execute complex scientific workflows
- Provenance information in OPM or PROV standards needed
- Reusable LAMBDA tasks enable reproducibility of scientific results