SCALABLE EARTH OBSERVATION ANALYTICS WITH SCIDB

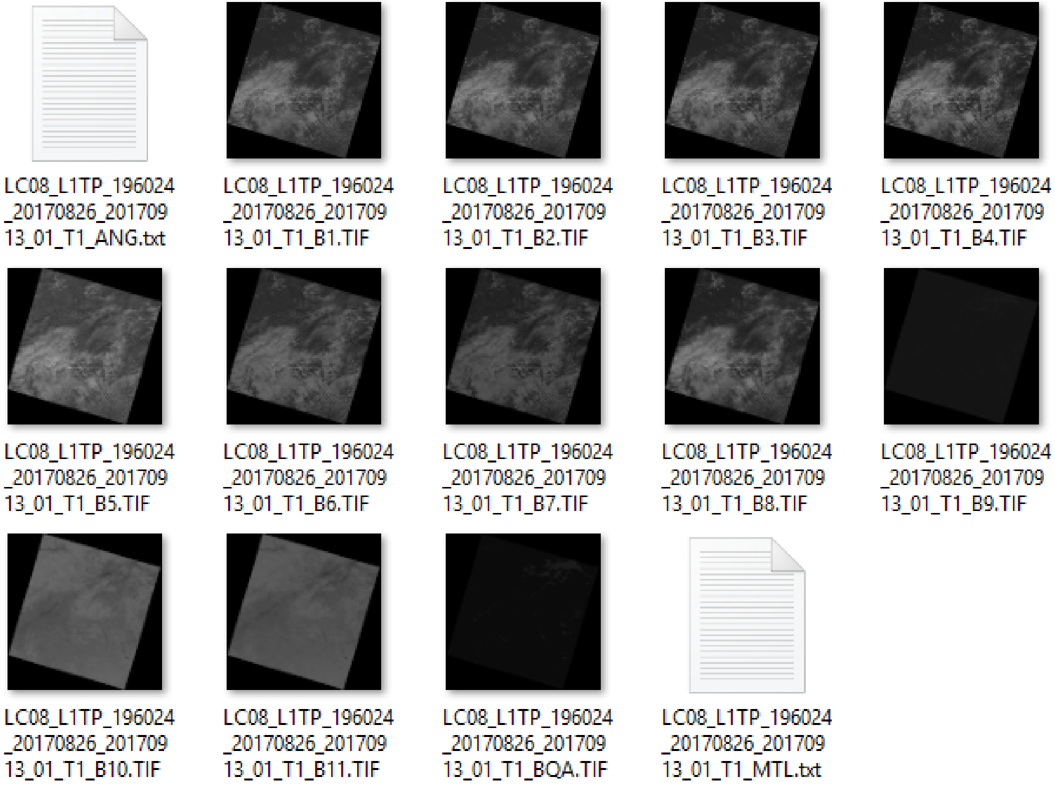
Marius Appel
marius.appel@uni-muenster.de

ifgi
Institute for Geoinformatics
University of Münster

WWU
MÜNSTER

EO DATA ORGANIZATION

LANDSAT 8



LC08_L1TP_196024_20170826_20170913_01_T1_ANG.txt

LC08_L1TP_196024_20170826_20170913_01_T1_B1.TIF

LC08_L1TP_196024_20170826_20170913_01_T1_B2.TIF

LC08_L1TP_196024_20170826_20170913_01_T1_B3.TIF

LC08_L1TP_196024_20170826_20170913_01_T1_B4.TIF

LC08_L1TP_196024_20170826_20170913_01_T1_B5.TIF

LC08_L1TP_196024_20170826_20170913_01_T1_B6.TIF

LC08_L1TP_196024_20170826_20170913_01_T1_B7.TIF

LC08_L1TP_196024_20170826_20170913_01_T1_B8.TIF

LC08_L1TP_196024_20170826_20170913_01_T1_B9.TIF

LC08_L1TP_196024_20170826_20170913_01_T1_B10.TIF

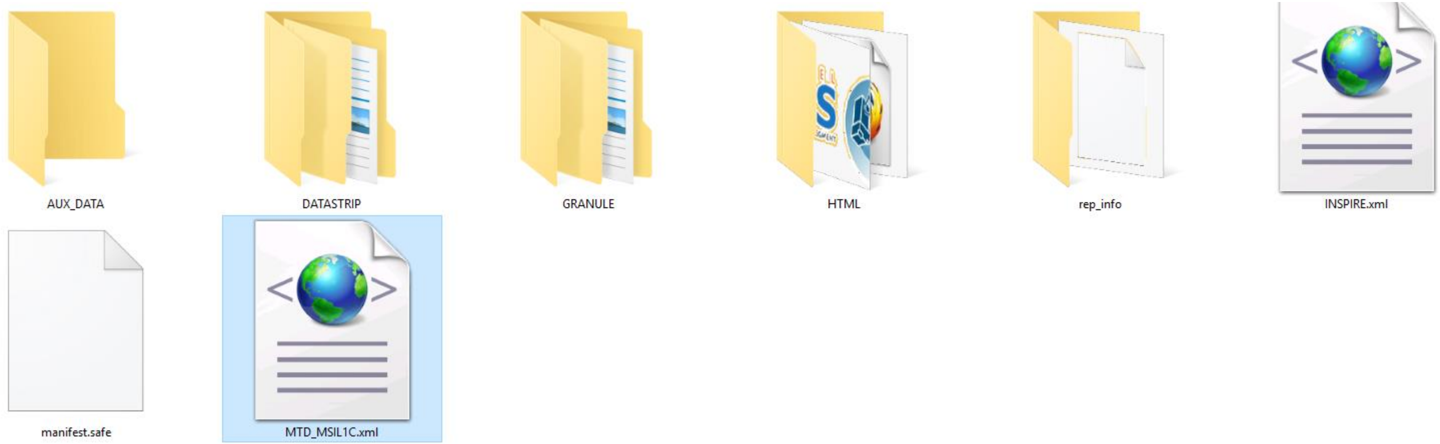
LC08_L1TP_196024_20170826_20170913_01_T1_B11.TIF

LC08_L1TP_196024_20170826_20170913_01_T1_BQA.TIF

LC08_L1TP_196024_20170826_20170913_01_T1_MTL.txt

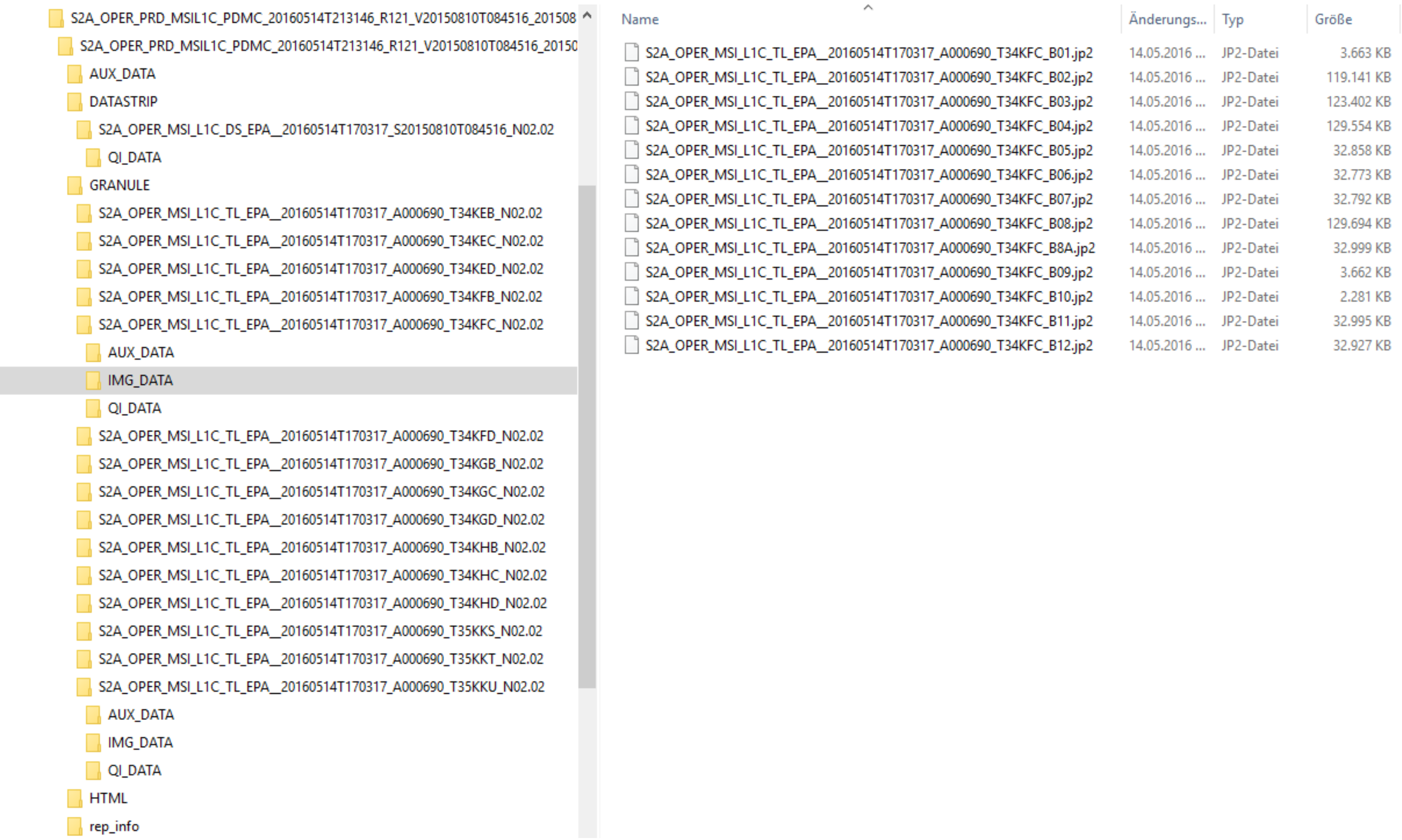
EO DATA ORGANIZATION

SENTINEL 2



EO DATA ORGANIZATION

SENTINEL 2

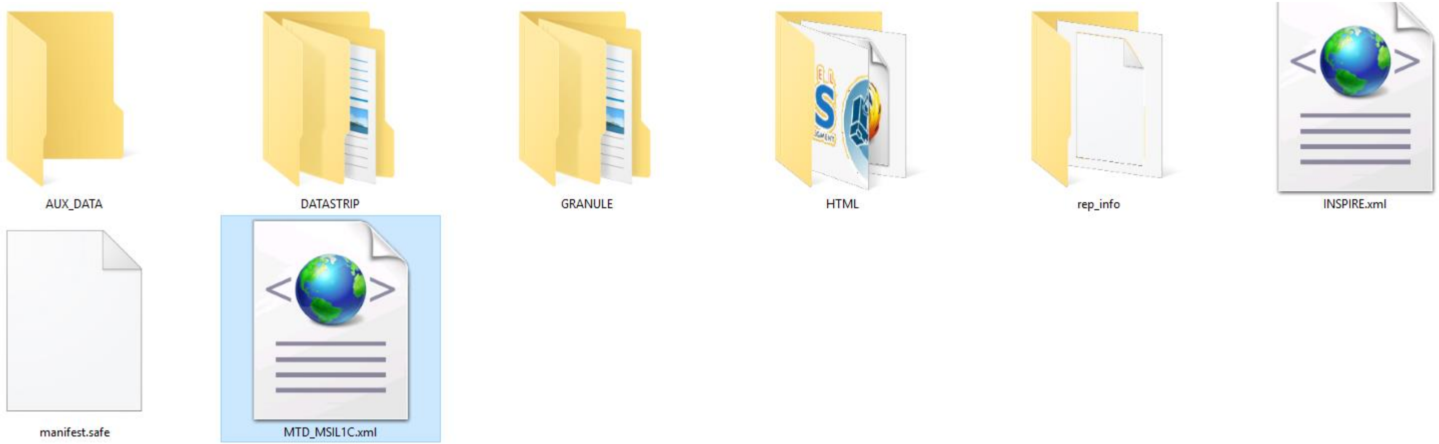


- S2A_OPER_PRD_MSIL1C_PDMC_20160514T213146_R121_V20150810T084516_201508
  - S2A_OPER_PRD_MSIL1C_PDMC_20160514T213146_R121_V20150810T084516_20150
    - AUX_DATA
    - DATASTRIP
      - S2A_OPER_MSI_L1C_DS_EPA__20160514T170317_S20150810T084516_N02.02
        - QI_DATA
    - GRANULE
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KEB_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KEC_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KED_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFB_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_N02.02
        - AUX_DATA
        - IMG_DATA
        - QI_DATA
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFD_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KGB_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KGC_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KGD_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KHB_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KHC_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KHD_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T35KKS_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T35KKT_N02.02
      - S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T35KKU_N02.02
        - AUX_DATA
        - IMG_DATA
        - QI_DATA
    - HTML
    - rep_info

Name	Änderungs...	Typ	Größe
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B01.jp2	14.05.2016 ...	JP2-Datei	3.663 KB
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B02.jp2	14.05.2016 ...	JP2-Datei	119.141 KB
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B03.jp2	14.05.2016 ...	JP2-Datei	123.402 KB
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B04.jp2	14.05.2016 ...	JP2-Datei	129.554 KB
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B05.jp2	14.05.2016 ...	JP2-Datei	32.858 KB
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B06.jp2	14.05.2016 ...	JP2-Datei	32.773 KB
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B07.jp2	14.05.2016 ...	JP2-Datei	32.792 KB
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B08.jp2	14.05.2016 ...	JP2-Datei	129.694 KB
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B8A.jp2	14.05.2016 ...	JP2-Datei	32.999 KB
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B09.jp2	14.05.2016 ...	JP2-Datei	3.662 KB
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B10.jp2	14.05.2016 ...	JP2-Datei	2.281 KB
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B11.jp2	14.05.2016 ...	JP2-Datei	32.995 KB
S2A_OPER_MSI_L1C_TL_EPA__20160514T170317_A000690_T34KFC_B12.jp2	14.05.2016 ...	JP2-Datei	32.927 KB

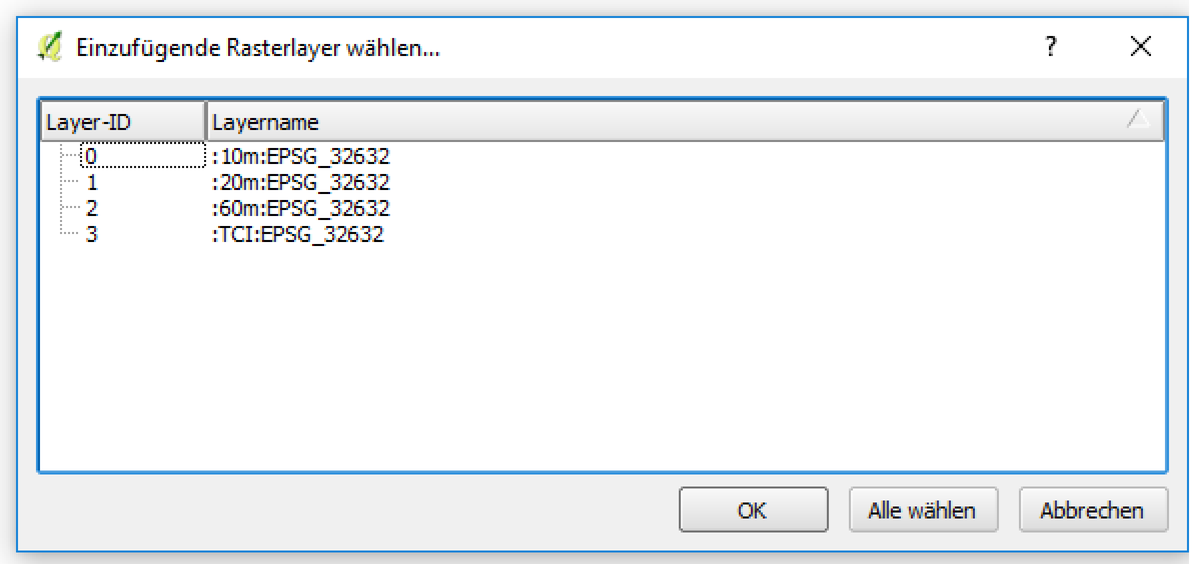
EO DATA ORGANIZATION

SENTINEL 2



EO DATA ORGANIZATION

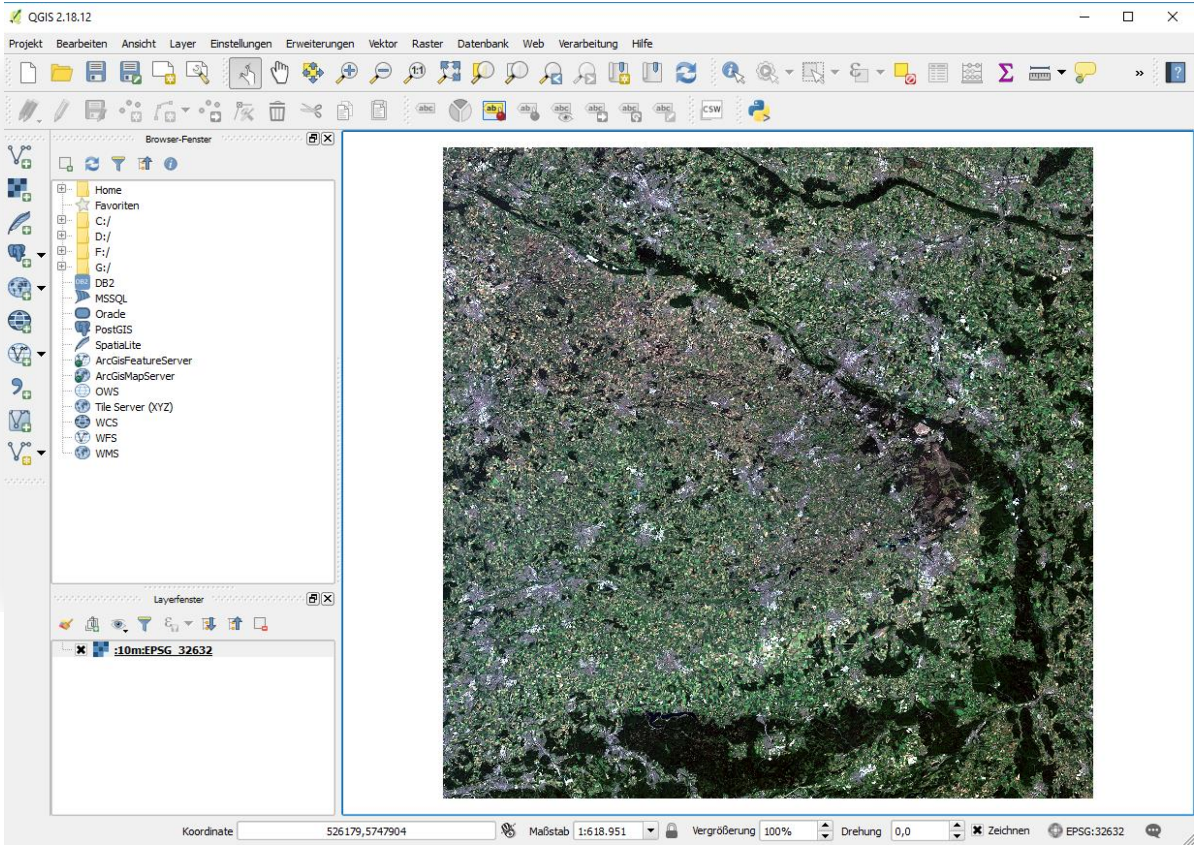
SENTINEL 2



Layer-ID	Layername
0	:10m:EPSG_32632
1	:20m:EPSG_32632
2	:60m:EPSG_32632
3	:TCI:EPSG_32632

EO DATA ORGANIZATION

SENTINEL 2



Koordinate 526179,5747904

Maßstab 1:618.951

Vergrößerung 100%

Drehung 0,0

☒ Zeichnen

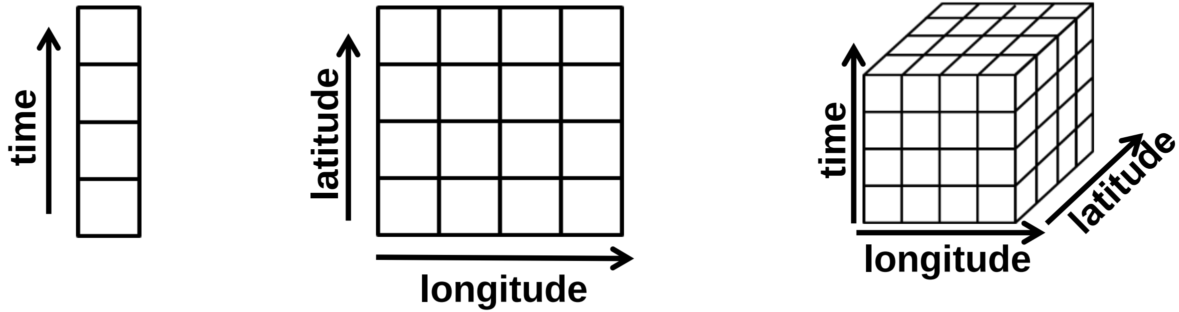
EPSG:32632

EO DATA ORGANIZATION

- EO image deployment is file-based
- GDAL interfaces EO imagery with GIS software
- Difficult to analyze large image collections due to
  - data volume
  - Irregularities
  - lack of time support in GDAL
- Higher-level data organization as an alternative to files?
  - Key requirement: scalability

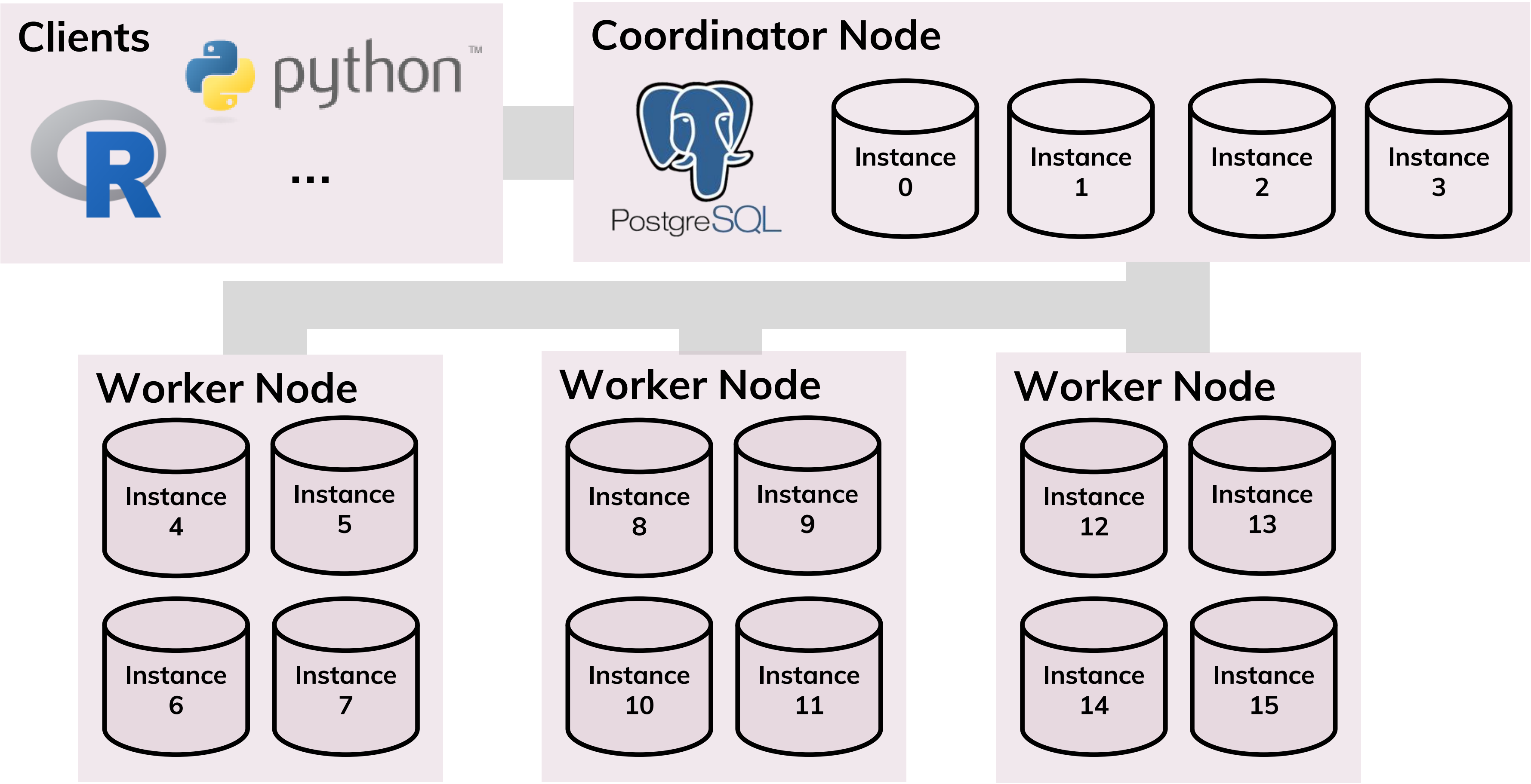
SCIDB INTRODUCTION

- Array-based data management and analytical system [1]
- Relies on shared nothing architectures
- Open-source version available, extensible by UDFs
- Basic data representation as multidimensional arrays:
  - n dimensions, m attributes with different data types



[1] Stonebraker, M., Brown, P., Zhang, D., & Becla, J. (2013). SciDB: A database management system for applications with complex analytics. *Computing in Science & Engineering*, *15*(3), 54-62.

SCIDB ARCHITECTURE



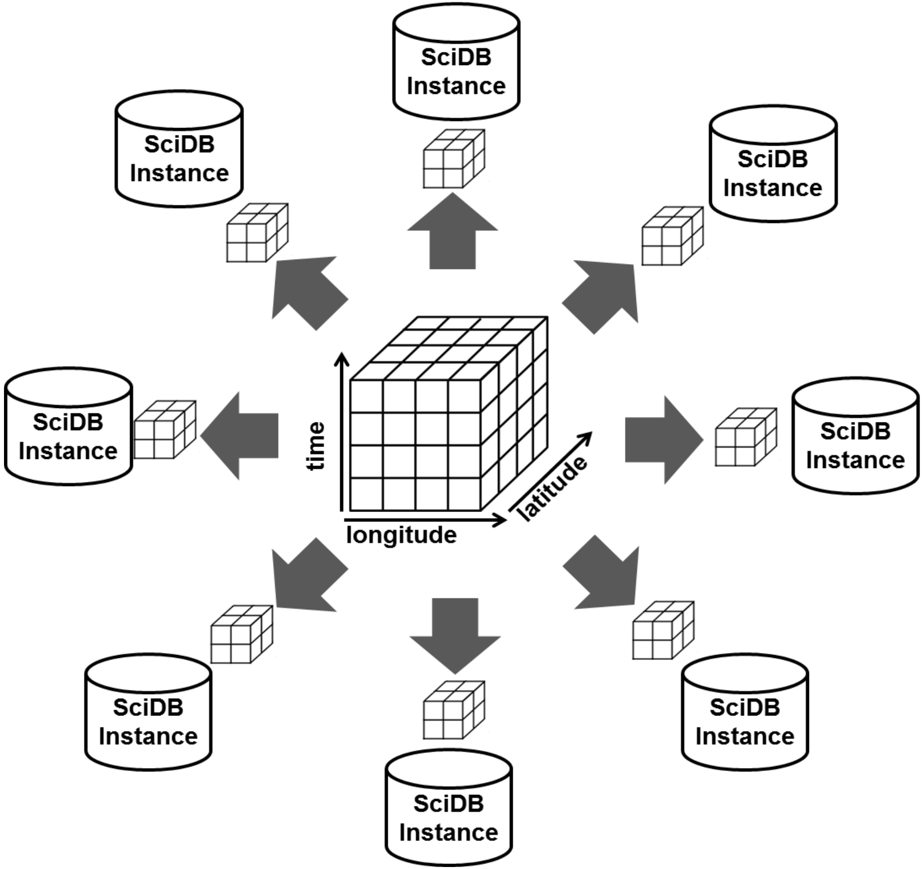
SCIDB ARCHITECTURE

- arrays are divided into equally sized chunks

- chunks are distributed over many SciDB instances

- Size and shape of chunks are defined by users per array and have strong effects on computation times

- Storage is nearly sparse



QUERY LANGUAGE AND FUNCTIONALITY

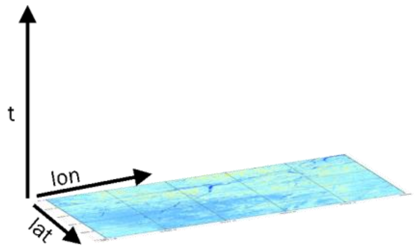
- SciDB query language: Array Functional Language (AFL)

- Built in functionality:
  - Load / write arrays from / to files
  - Arithmetic operations
  - subsetting by dimensions, attributes, or values
  - Aggregations
  - Joins
  - Changing array schemas (repartitioning, redimensioning)
  - Linear algebra routines: (GEMM, GESVD, basic statistics)
  - …

EXTENSIONS FOR EO DATA

- scidb4geo (https://github.com/appelmar/scidb4geo)
  - SciDB plugin adds metadata and simple operations on space-time referenced arrays

- scidb4gdal (https://github.com/appelmar/scidb4gdal)
  - ingest / download to / from GDAL supported files
  - spacetime mosaicing



- R package `scidbst` (https://github.com/flahn/scidbst)
  - mimics functionality of common packages on SciDB arrays

SCIDB CLIENTS

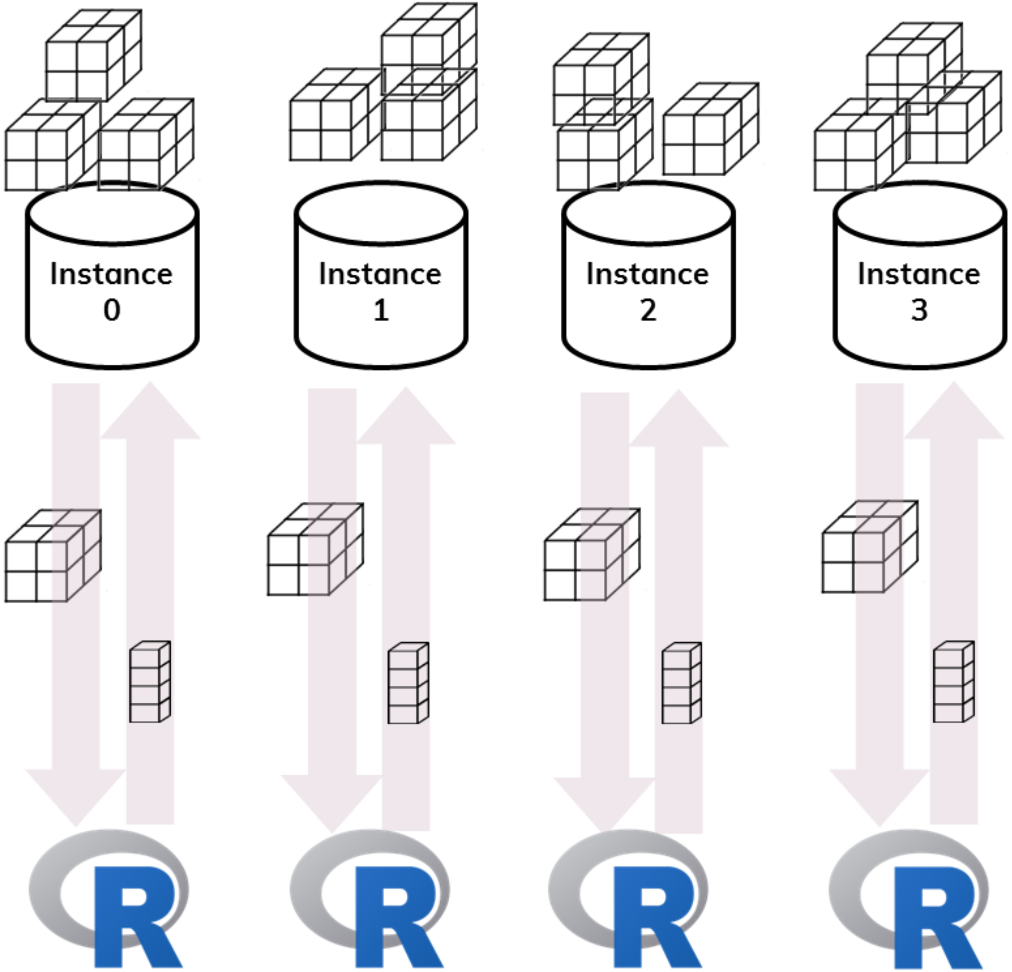
- Low-level clients: iquery, Shim

- High-level R client (similar for Python)
  - overrides standard methods, e.g. %*%
  - make extensive use of proxy objects
  - lazy evaluation:
    - compute things when result is being read
    - ignore computations for unread parts of the results

SCIDB STREAMING

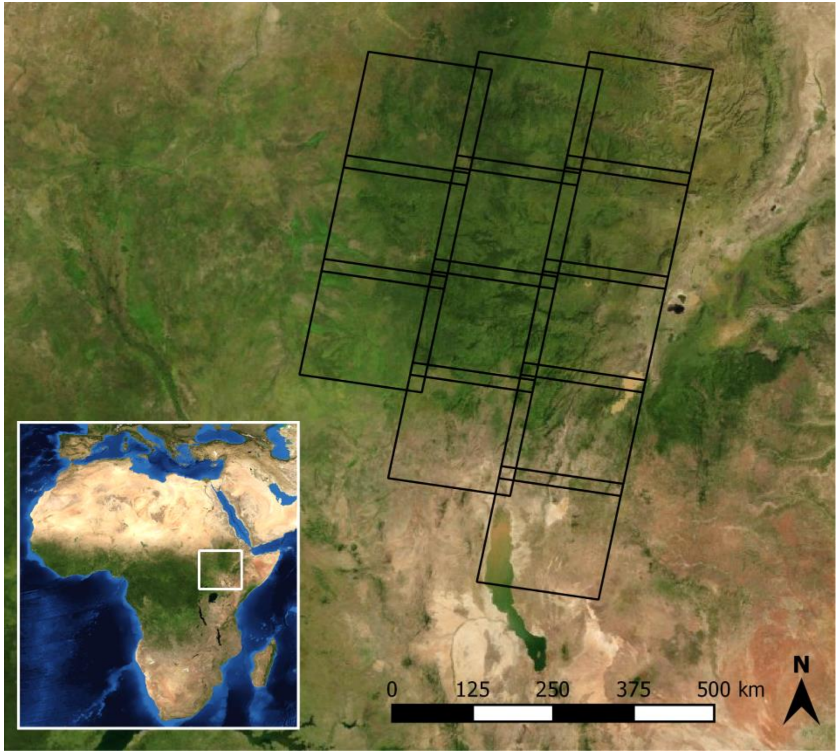
- Run external programs (e.g., R, python) within SciDB at chunk level parallelism

→ chunk size selection must be adapted to the analysis



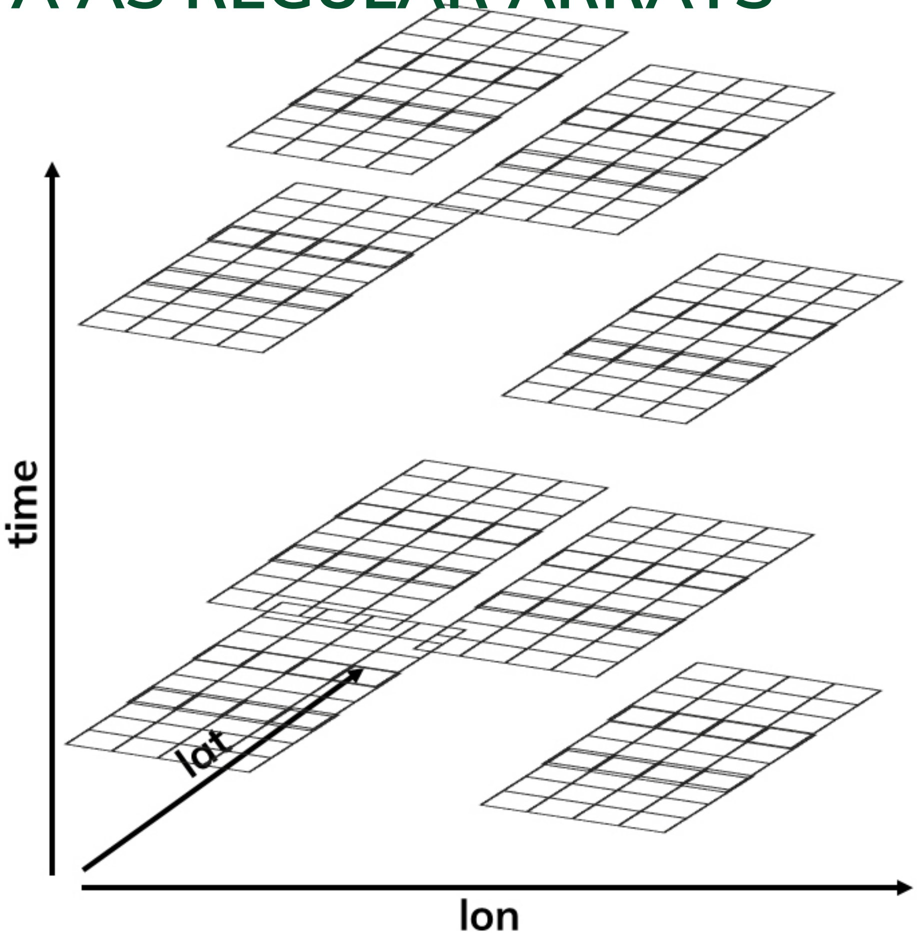
STUDY CASE: LAND USE CHANGE MONITORING IN SOUTH WEST ETHIOPIA FROM LANDSAT 7 IMAGERY

- Landsat 7 data from 12 tiles captured between 2003-07-21 and 2014-12-27 → 1975 scenes
- approx. 325,000 km^2
- monitor changes starting with 2010-01-01
- using R and Breaks For Additive Season and Trend and its R implementation [1]



[1] Verbesselt, J., Hyndman, R., Newnham, G., & Culvenor, D. (2010). Detecting trend and seasonal changes in satellite image time series. Remote Sensing of Environment, 114, 106-115. DOI: 10.1016/j.rse.2009.08.014.

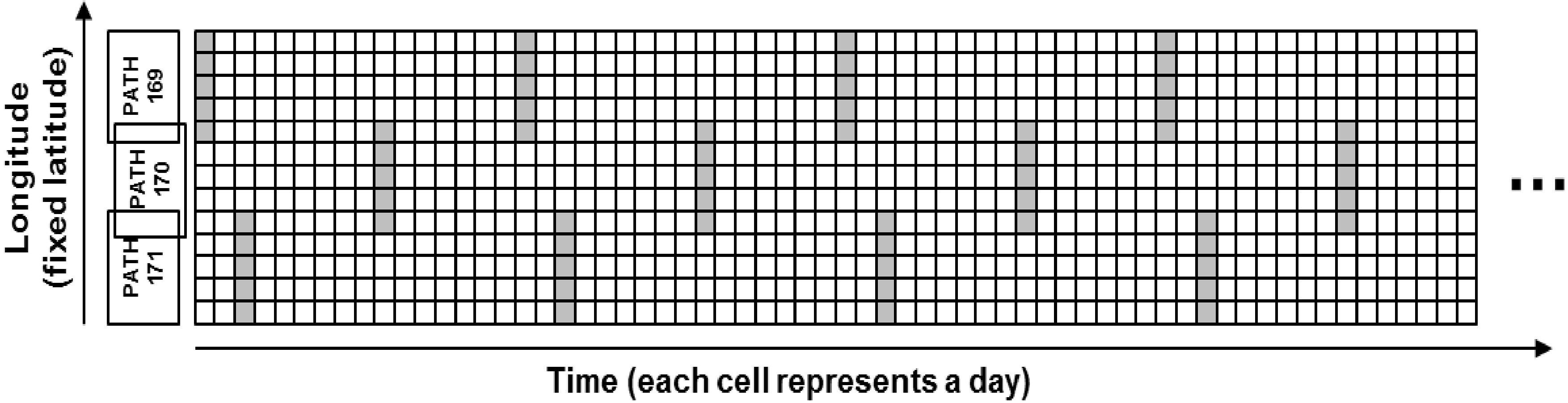
EO DATA AS REGULAR ARRAYS



LANDSAT 7 IN SCIDB

Images form a single three-dimensional array with **daily temporal resolution** and

- 49548 x 47713 x 4177 cells in total
- Only 0.5% ($54 \cdot 10^9$) of the cells contain data → sparse storage



STUDY CASE IMPLEMENTATION

1. Ingestion using GDAL

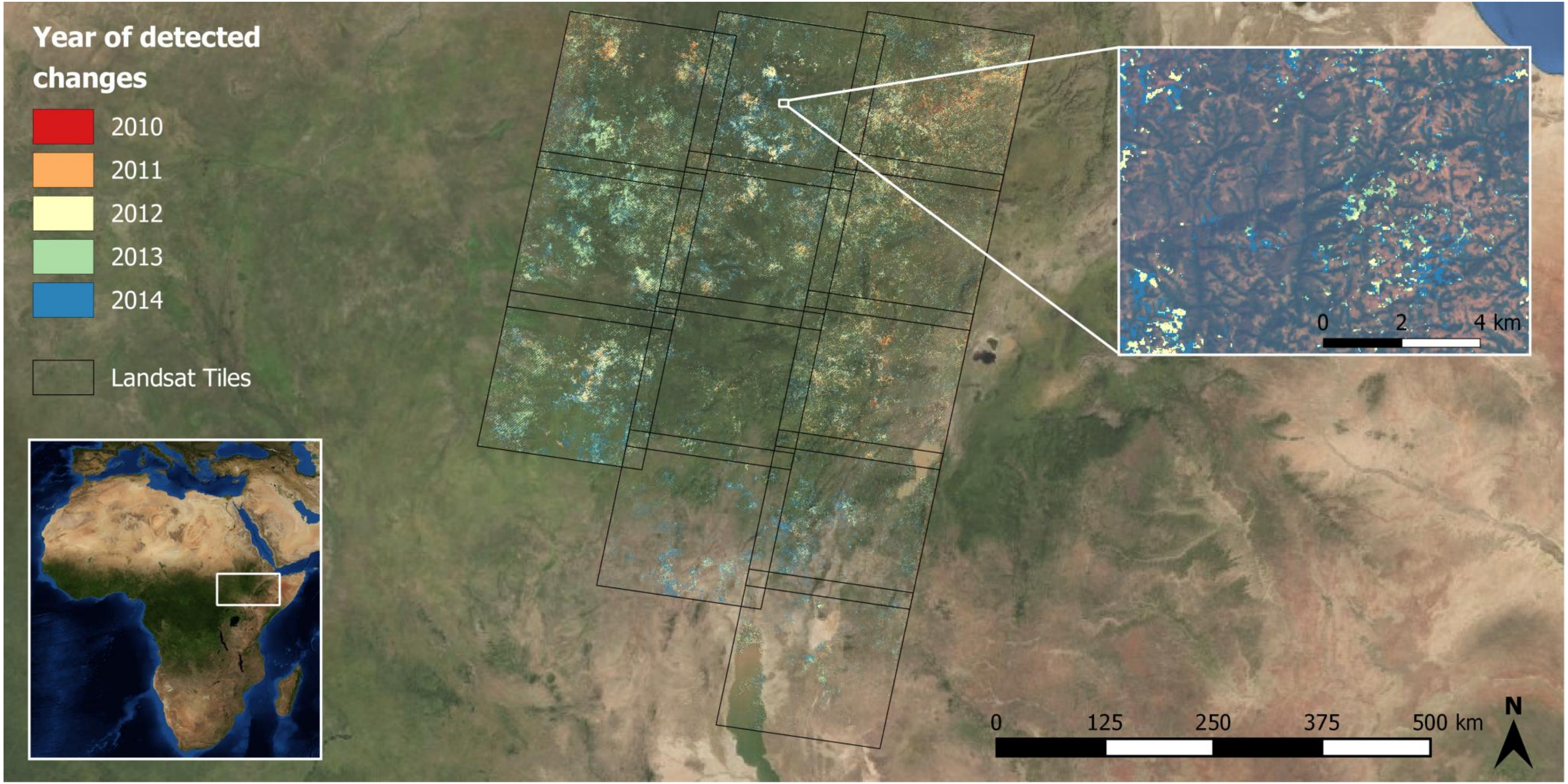
2. Preprocessing (with built-in SciDB functionality)
   - remove any values <= -9999 or >10000
   - compute NDVI vegetation index
   - Reorganize chunks such that one chunk stores complete time series of 64 x 64 pixels

3. Run R scripts on all chunks using streaming

4. Postprocessing (with built-in SciDB functionality)
   - Reshape one-dimensional result array to form a two-dimensional map

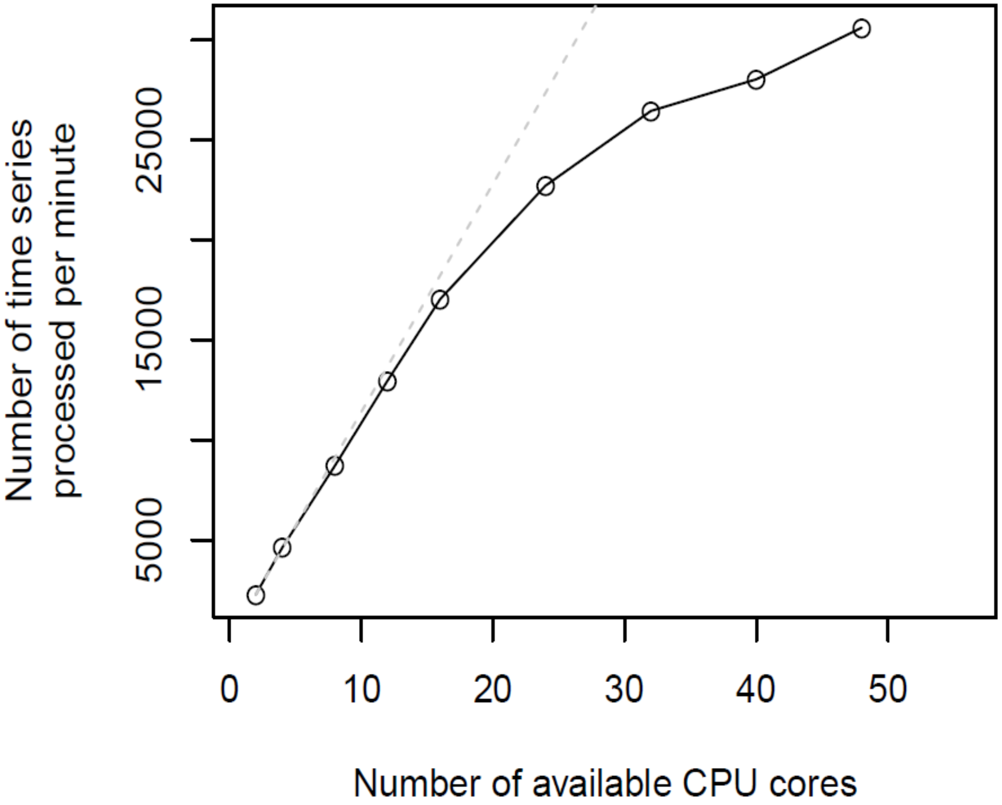
5. Export results using GDAL

STUDY CASE: RESULTS



STUDY CASE SCALABILITY

- 16 SciDB instances
- running change analysis repeatedly with different number of available CPU cores



CONCLUSIONS

- The array model with chunking and sparse storage seems well-suited to represent large EO datasets from many scenes at a higher level than files

- Analyses scale well with available hardware

- Little reimplementation needed to scale complex time-series processing through streaming (and no need to care about parallelization / external memory)

- Installation and data ingestion not straightforward and time-consuming

- Mostly useful for re-analysis but not real-time processing

- Missing interactive(!) user interfaces (á la Google Earth Engine) to make the technology more accessible to end users?

THANK YOU

- Questions?
- Hands-on with SciDB tomorrow!
- Slides available at GitHub: https://github.com/appelmar/edcforum2017
- Contact **marius.appel@uni-muenster.de**