EDC FORUM 2017

Big Data Analytics & GIS – September 21-22, Münster

Please do not unplug any cables, the desktop computers are alarmed!

HANDS-ON WORKSHOP: **PROCESSING EARTH OBSERVATION TIME SERIES WITH SCIDB AND R**

Marius Appel
marius.appel@uni-muenster.de

ifgi
Institute for Geoinformatics
University of Münster

OBJECTIVES

- provide an overview of SciDB in practical applications and
  - how it can be used to analyze Earth observation data
  - where it has advantages / limitations

- see how accessible it is without previous experiences with the technology

- We will look at SciDB mostly from the perspektive of a „data scientist“

WHO ARE YOU?

OUTLINE (MORNING)

1. Sample dataset

2. Introduction and demonstration of basic concepts in SciDB
   - Architecture
   - Arrays in Detail
   - AFL query language

3. Introduction to the SciDB R client
   - The R package scidb
   - The R package scidbst

OUTLINE (MORNING)

4. Example time series analyses
   - Time series extraction
   - Image averaging
   - Greenest image
   - Water classification
   - Seasonality assessment

5. Ingestion of EO data

6. Hands-on

ORGANIZATION

- Material online at https://github.com/appelmar/edcforum2017

- Please
  - do not remove any data in SciDB except for arrays that you have created during the workshop
  - work with very small subsets if you try out things, all participants share the same SciDB installation!
  - ask questions, contribute ideas!

OUTLINE (AFTERNOON)

Hackthon with open topics:

- More time to experiment with SciDB and the provided Sentinel dataset
- Installation of SciDB and its extensions using Docker
- Implementation of further methods, e.g.: compute difference images from two given dates
- Ingest and analyze other data and implement (e.g., MODIS, SRTM)
- Try out SciDB-Py, e.g. by re-implementing the R examples
- your own ideas!

SENTINEL 2

- Two satellites:
  - Sentinel-2A (launched 23 June 2015)
  - Sentinel-2B (launched 07 March 2017)
- complements Landsat and SPOT missions
- part of EU Copernicus program
- 13 Spectral Bands at (10m, 20m, or 60m) spatial resolution
- covers latitudes from 56° south to 83° north
- revisit time: 5 days (with two satellites)

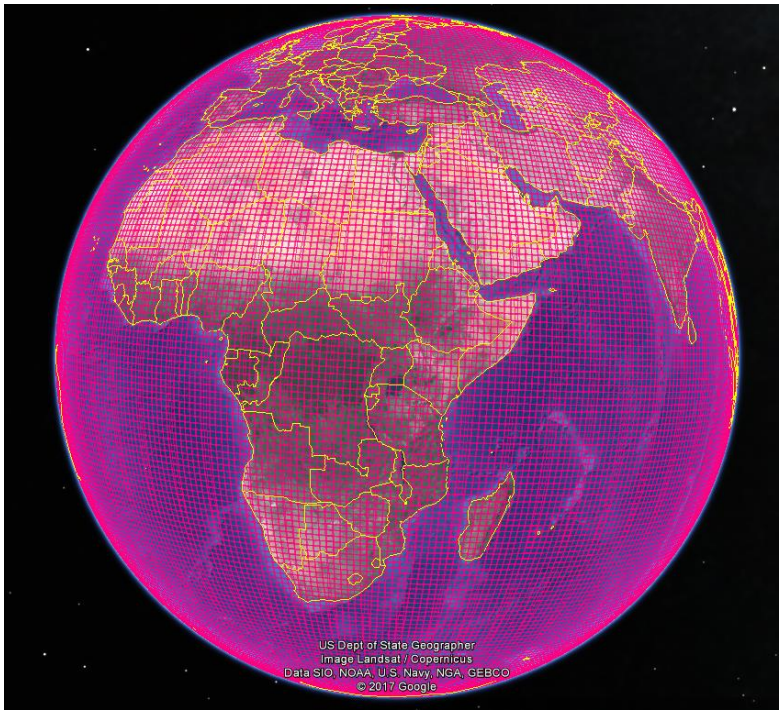
SENTINEL 2

- Applications:
  - Water monitoring
  - Forest / vegetation monitoring and change detection
  - Spatial planning
  - Crop monitoring
  - Emergency Management

- free and open access!

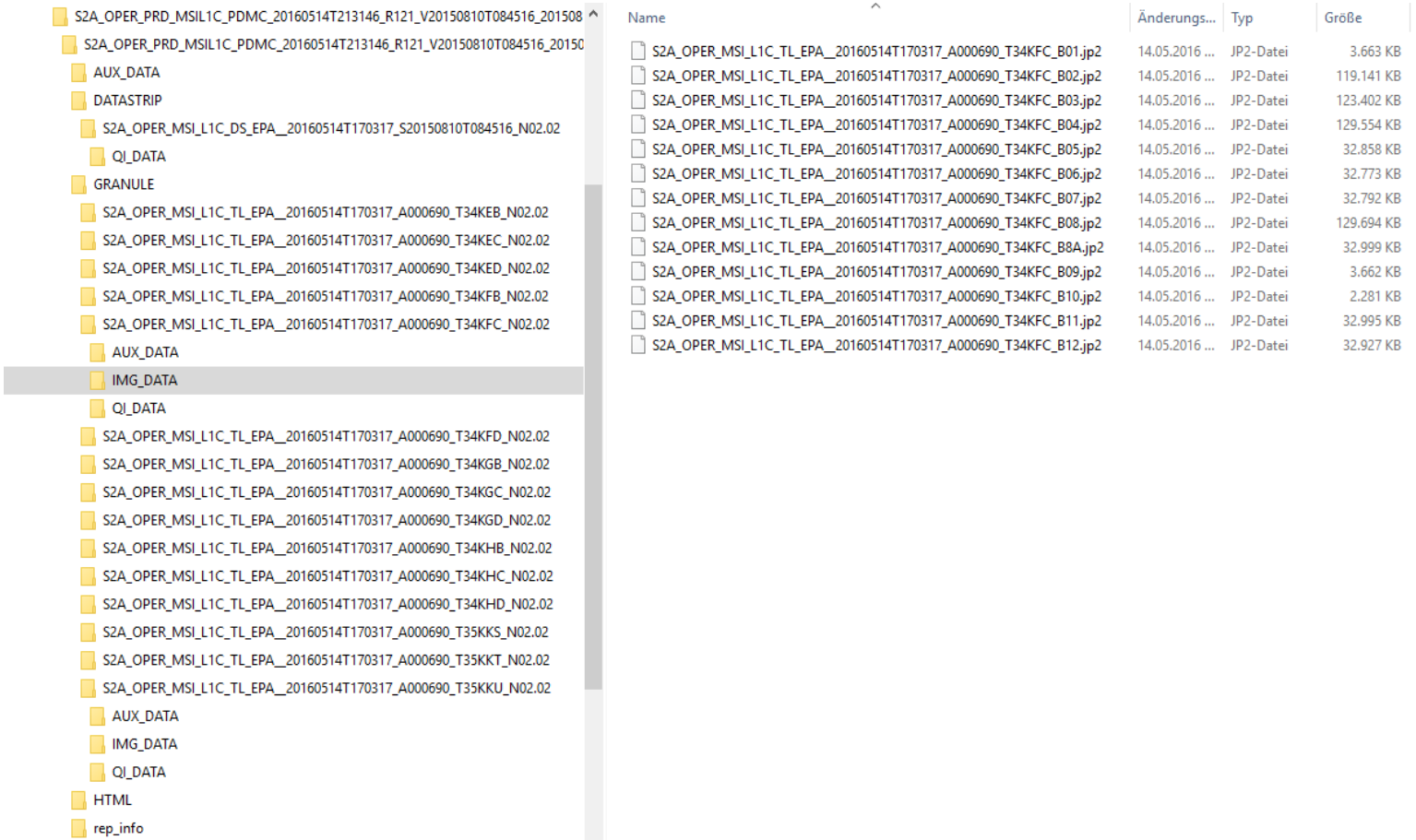
SENTINEL 2

Tiling of orthorectified products (Level 1C / 2A)



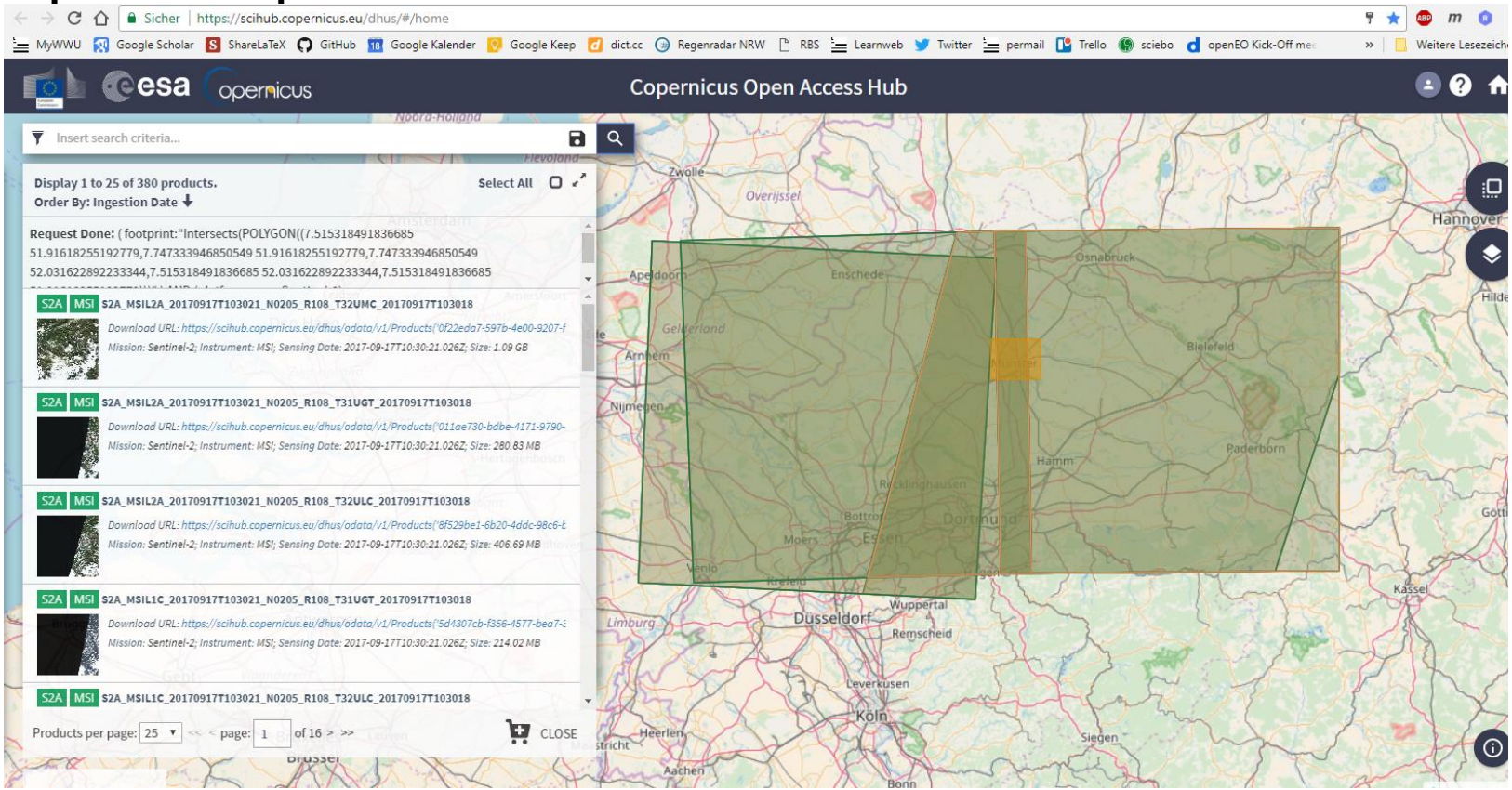
https://sentinel.esa.int/documents/247904/1955685/S2A_OPER_GIP_TILPAR_MPC__20151209T095117_V20150622T000000_21000101T000000_B00.kml

SENTINEL 2 – DATA ORGANIZATION



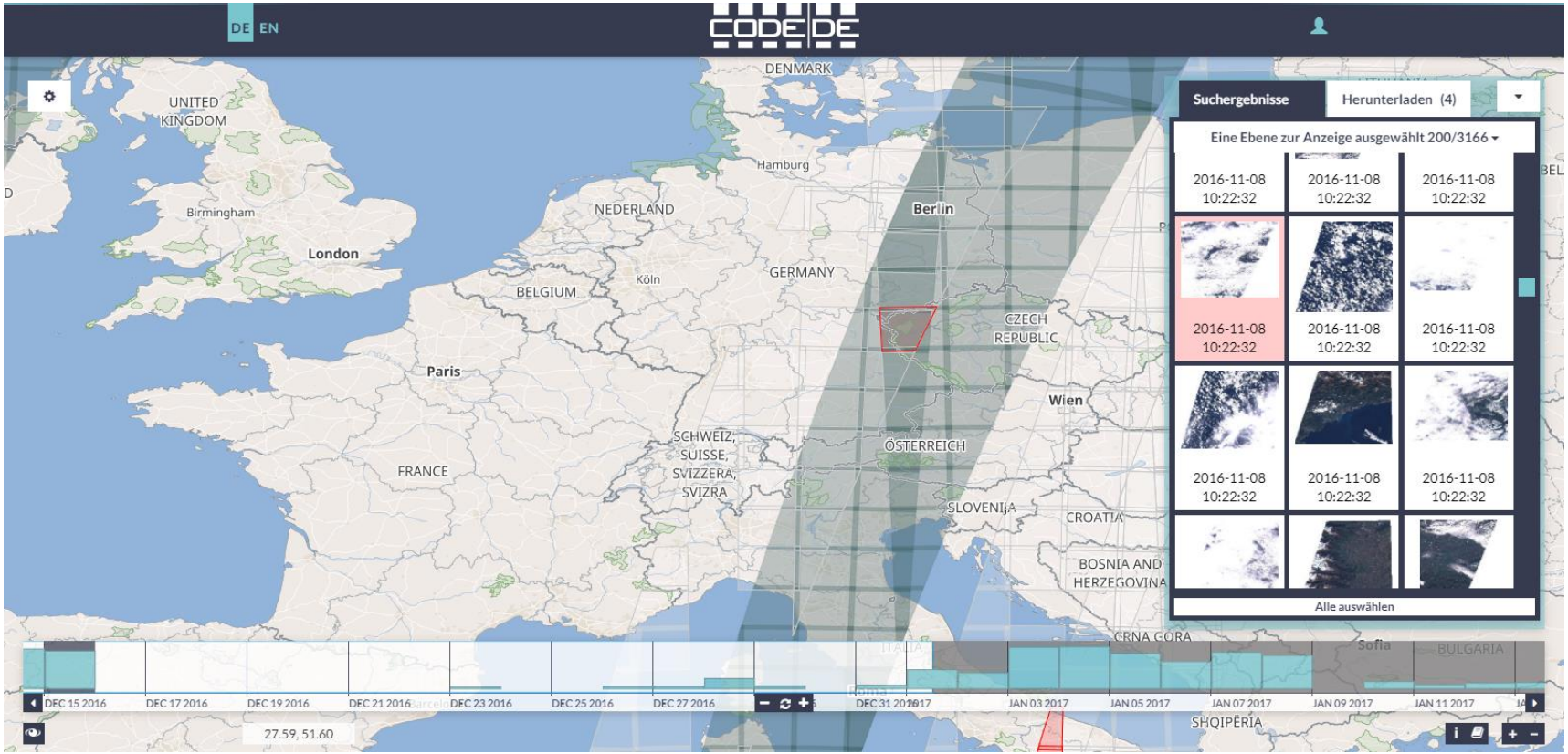
SENTINEL 2 – DATA ACCESS

Copernicus Open Access Hub



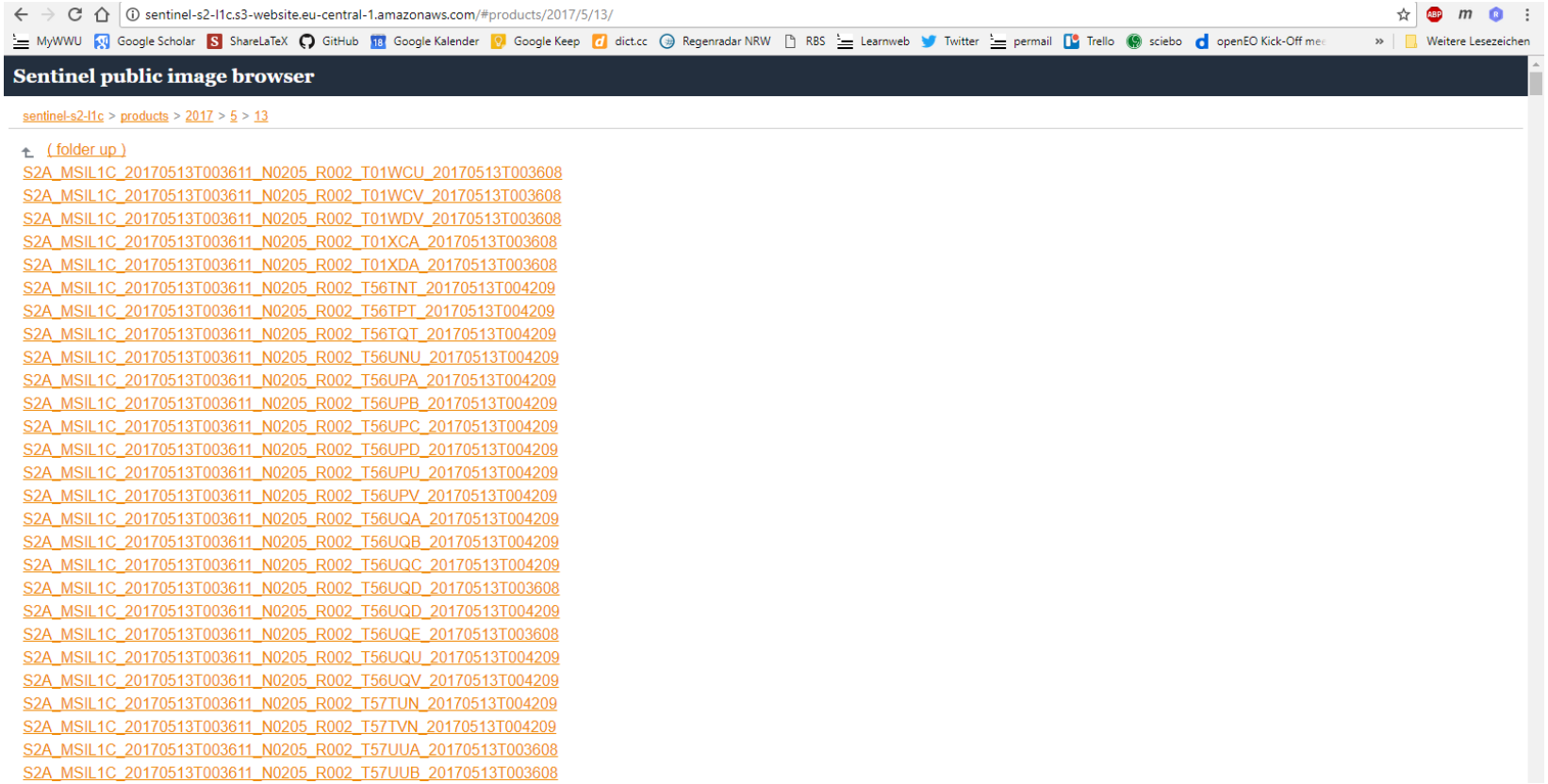
SENTINEL 2 – DATA ACCESS

Copernicus Data and Exploitation Platform – Deutschland (CODE-DE)



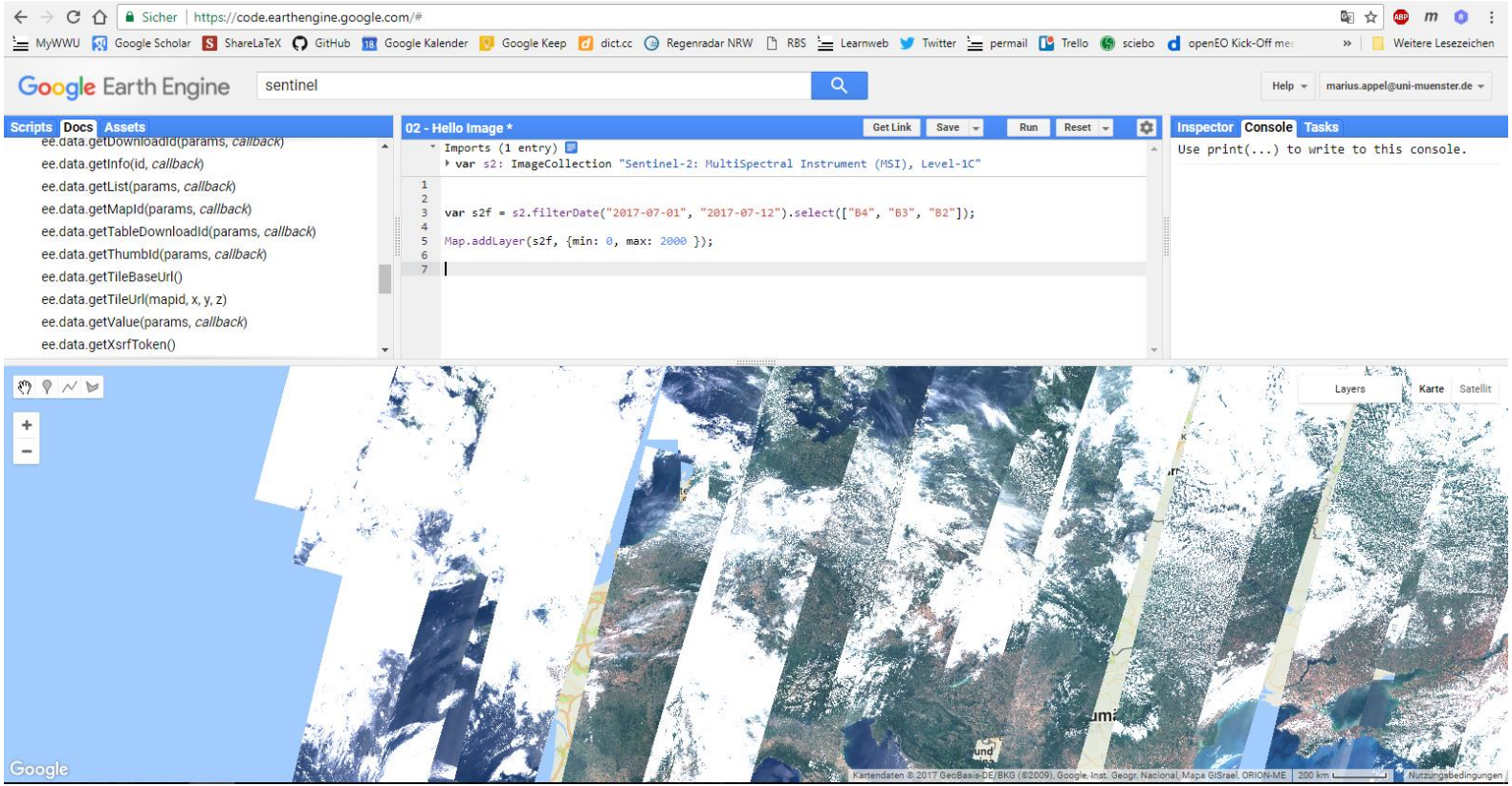
SENTINEL 2 – DATA ACCESS

Amazon Web Services



SENTINEL 2 – DATA ACCESS

Google Earth Engine



(FUTURE) ACCESS TO EO DATA?

Alternatives to downloading scenes from space agencies?

- dedicated EO data centers / cloud services that provide access at a
  - low-level (file-based): AWS, EODC, ….
  - high-level: Google Earth Engine, SciDB,

- Processing where the data is stored

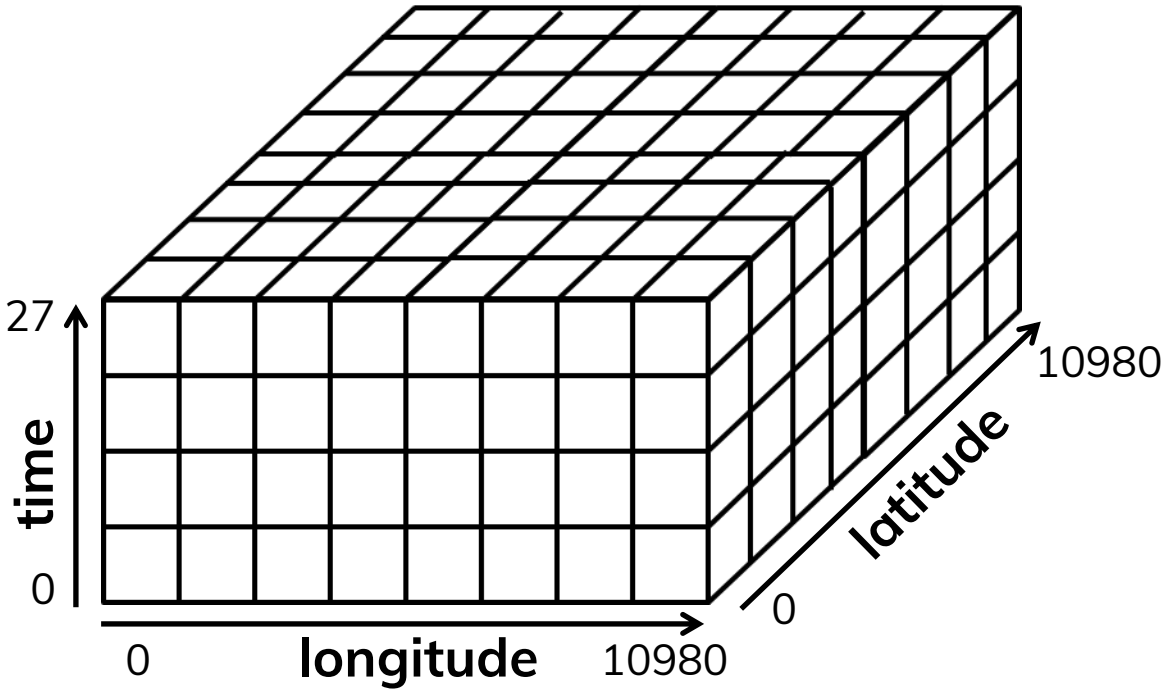
SAMPLE DATASET

- 26 scenes from Sentinel 2A Level 1C
- Tile 34KGD (Okavango Delta region)
- Captured between 2016-12-12 and 2017-08-29
- approx. 12e09 measurements, approx 22 GB

SAMPLE DATASET

SAMPLE DATASET

- SciDB representation has three dimensions (10980 x 10980 x 27 cells)
- Each cell has 4 attributes (bands)
- Cells cover 10m x 10m x 10d
- Two arrays with the same data but different chunk organization:
  - **S2_OKAVANGO_S**
  - **S2_OKAVANGO_T**



BASIC SCIDB CONCEPTS

Multidimensional Arrays (n dimensions, m attributes)

$$A: D \rightarrow V$$

$$D \subset D_1 \times D_2 \times \cdots \times D_n$$

$$V \subset V_1 \times V_2 \times \cdots \times V_m$$

- Dimensions are ordered and finite
- SciDB dimensions are always 64bit integers → space or time must be transformed
- SciDB attributes can be of any type, and nullable if desired

BASIC SCIDB CONCEPTS

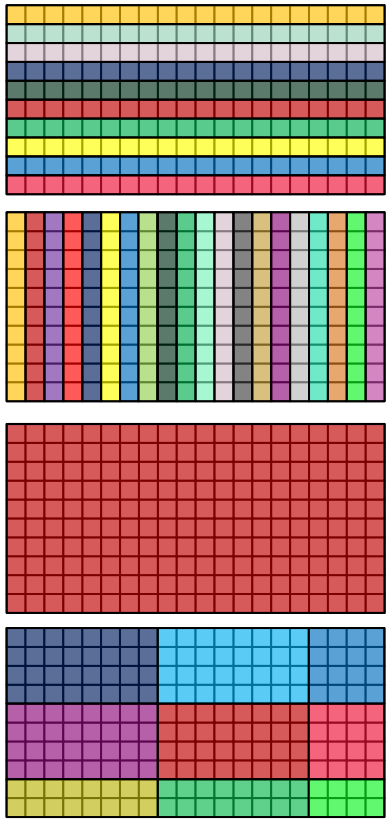
Examples:

- RGB image: two dimensions, three attributes
- multivariate time series: one dimension, multiple attributes
- hyperspectral EO image time series: four dimensions, one attribute
- DEM: two-dimensions, one attribute

BASIC SCIDB CONCEPTS

Chunking

- equally sized rectangular subsets of the data within an array
- specified by number of cells per dimension
- defined during array creation
- optional chunk overlap
- strongly affects scalability and performance



BASIC SCIDB CONCEPTS

Nearly Sparse Chunk Storage

- **Empty cells**: no value or any of the attributes
- **Empty vs Null**: cell value of a single attribute can be null but still have data for other attributes, null also adds a missing reason code and is stored

BASIC SCIDB CONCEPTS

Nearly Sparse Chunk Storage

- run-length encoded dense bitmaps and
- actual values of attributes for nonempty cells

SCIDB QUERY LANGUAGE AFL

→ see `scripts/1_AFL_basics.md`

SCIDB R CLIENT

→ see `scripts/2_SciDB_R.R`

SCIDBST R PACKAGE

→ see `scripts/3_scidbst.R`

EXAMPLE TASKS

1. Time series extraction
2. Image averaging
3. Greenest image
4. Water classification
5. Seasonality assessment