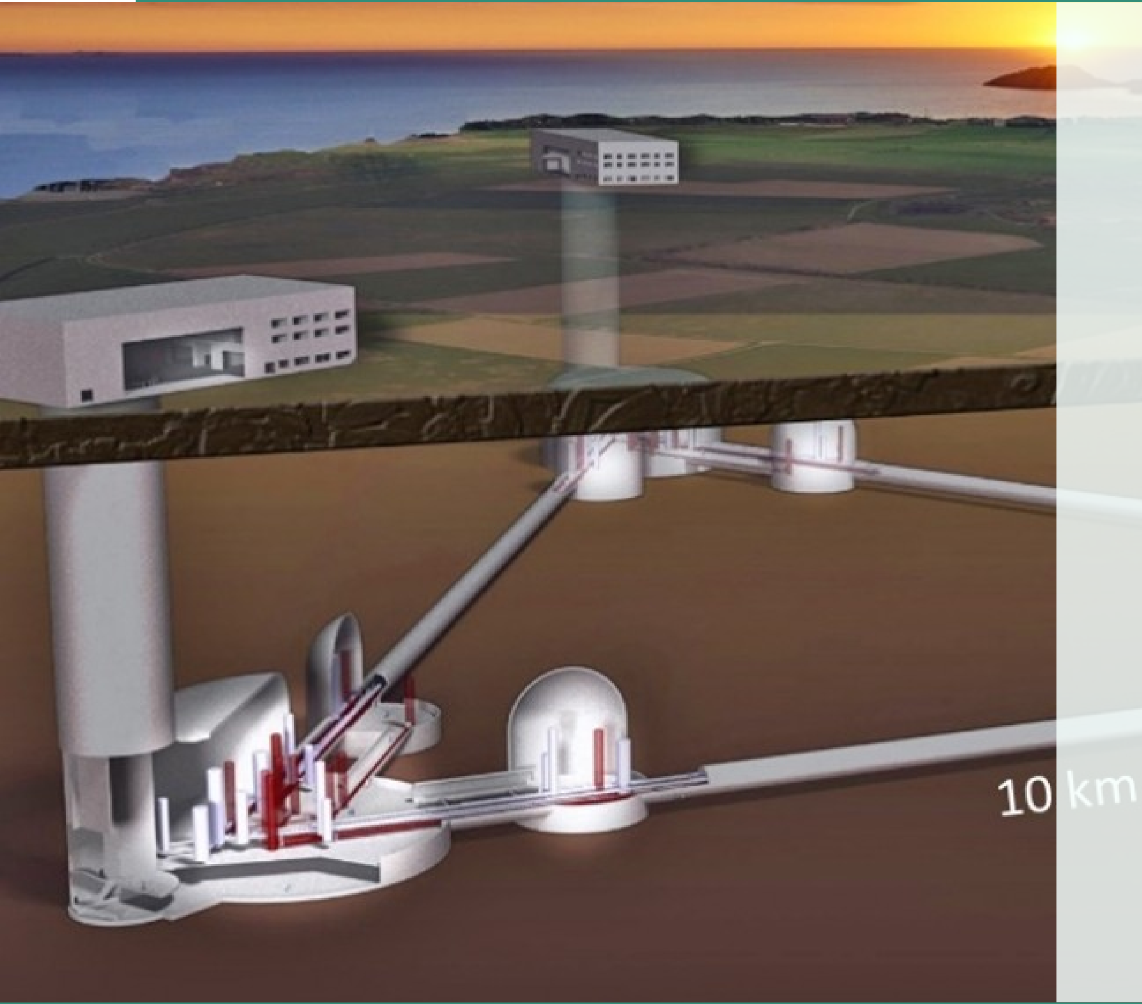


SPB workshop – ET Symposium

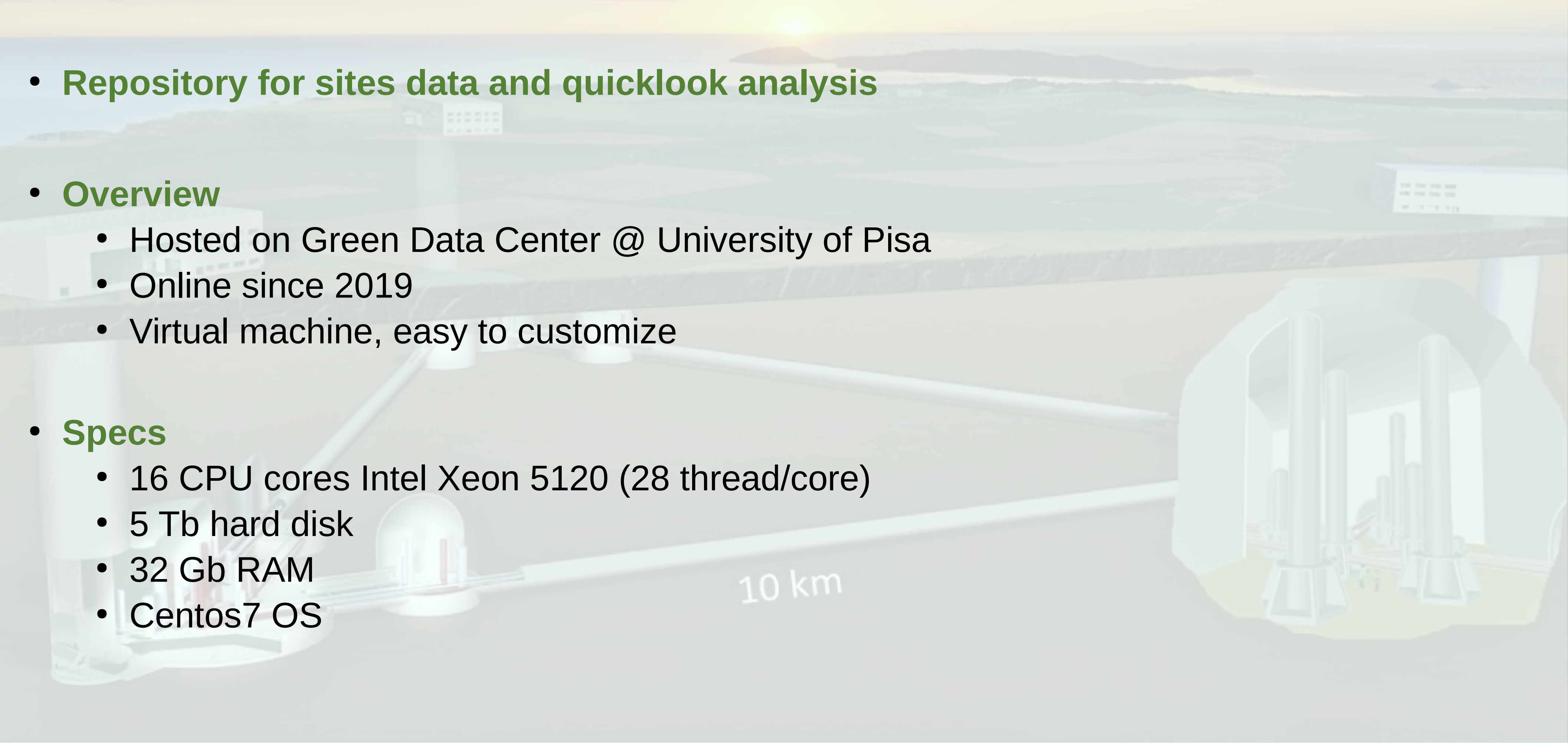
May 8, 2023

Updates on repository and data management

A. Fiori, M. Razzano and many others

University of Pisa & INFN-Pisa

ET Data Repository (etrepo)

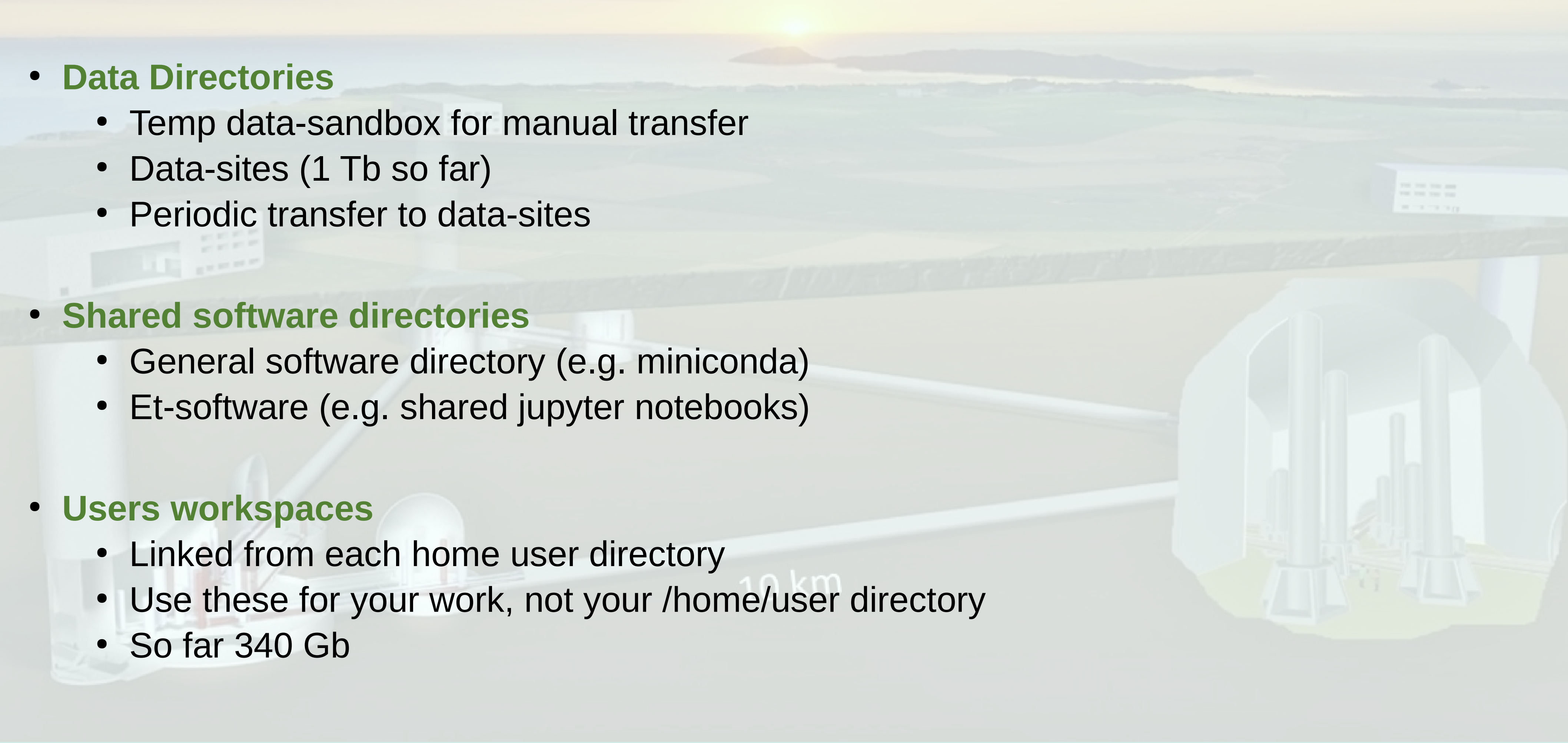


- **Repository for sites data and quicklook analysis**

- **Overview**
  - Hosted on Green Data Center @ University of Pisa
  - Online since 2019
  - Virtual machine, easy to customize

- **Specs**
  - 16 CPU cores Intel Xeon 5120 (28 thread/core)
  - 5 Tb hard disk
  - 32 Gb RAM
  - Centos7 OS

ETRepo data organization

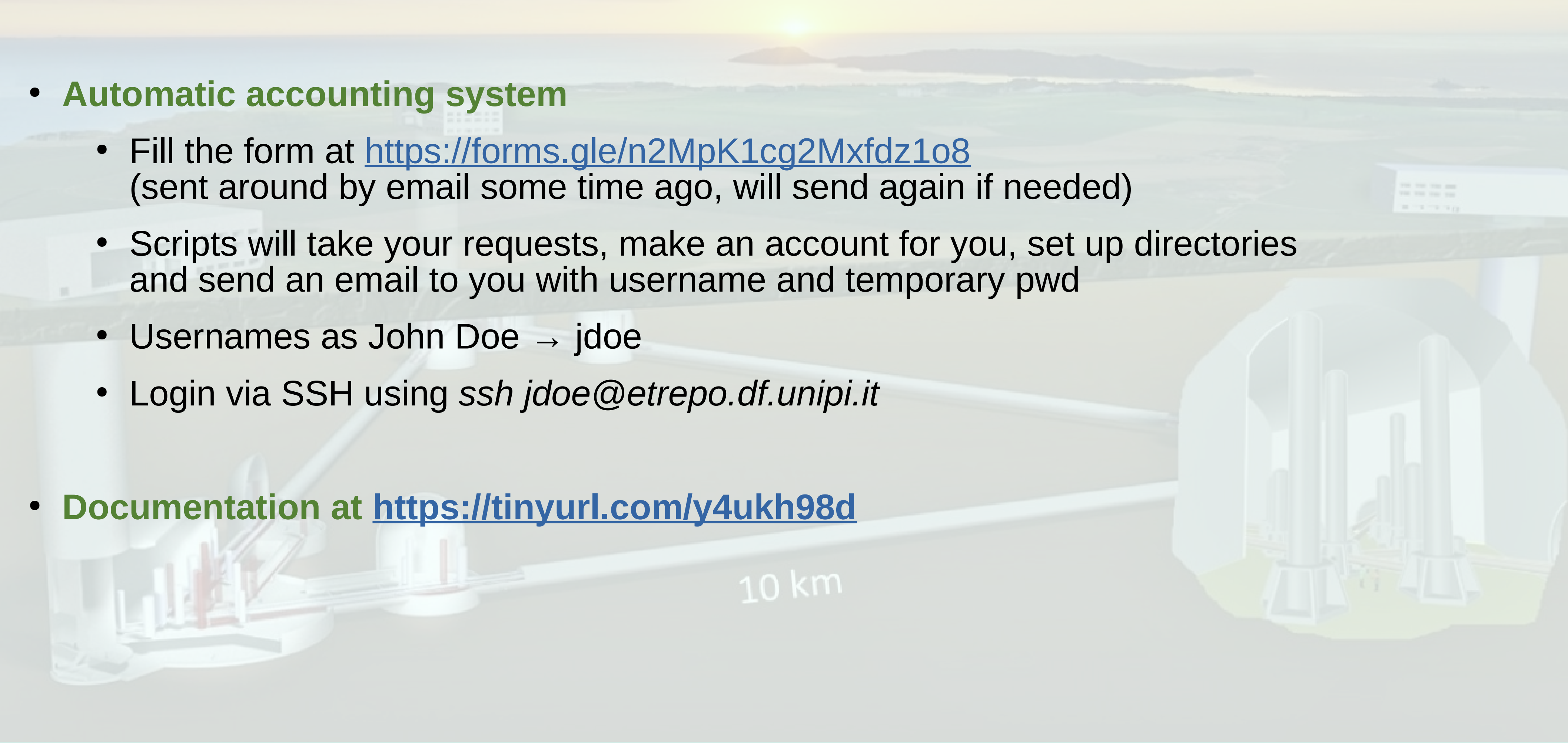


- **Data Directories**
  - Temp data-sandbox for manual transfer
  - Data-sites (1 Tb so far)
  - Periodic transfer to data-sites

- **Shared software directories**
  - General software directory (e.g. miniconda)
  - Et-software (e.g. shared jupyter notebooks)

- **Users workspaces**
  - Linked from each home user directory
  - Use these for your work, not your /home/user directory
  - So far 340 Gb

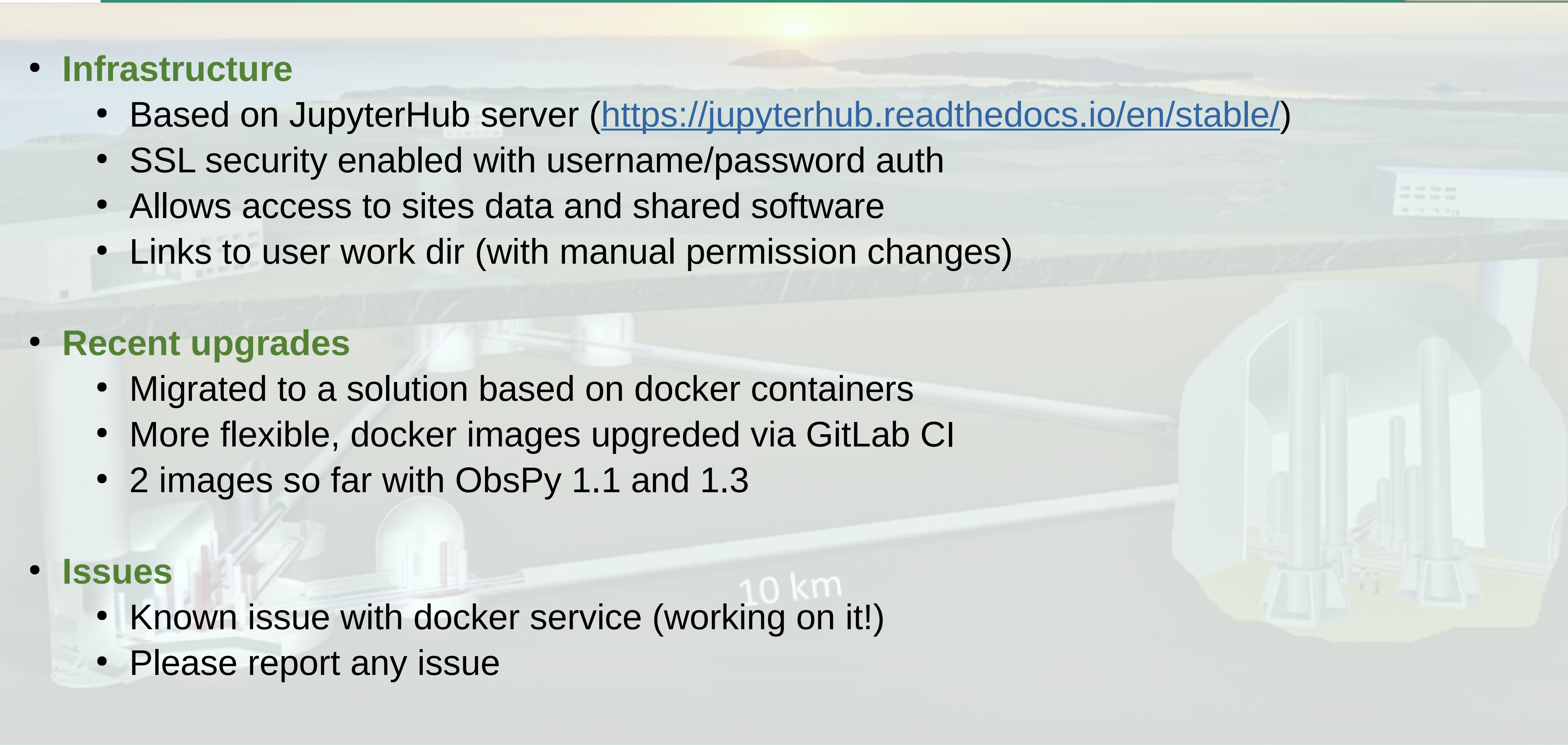
Accessing ETRepo



- **Automatic accounting system**
  - Fill the form at https://forms.gle/n2MpK1cg2Mxfdz1o8 (sent around by email some time ago, will send again if needed)
  - Scripts will take your requests, make an account for you, set up directories and send an email to you with username and temporary pwd
  - Usernames as John Doe → jdoe
  - Login via SSH using *ssh jdoe@etrepo.df.unipi.it*

- **Documentation at https://tinyurl.com/y4ukh98d**

JupyterLab



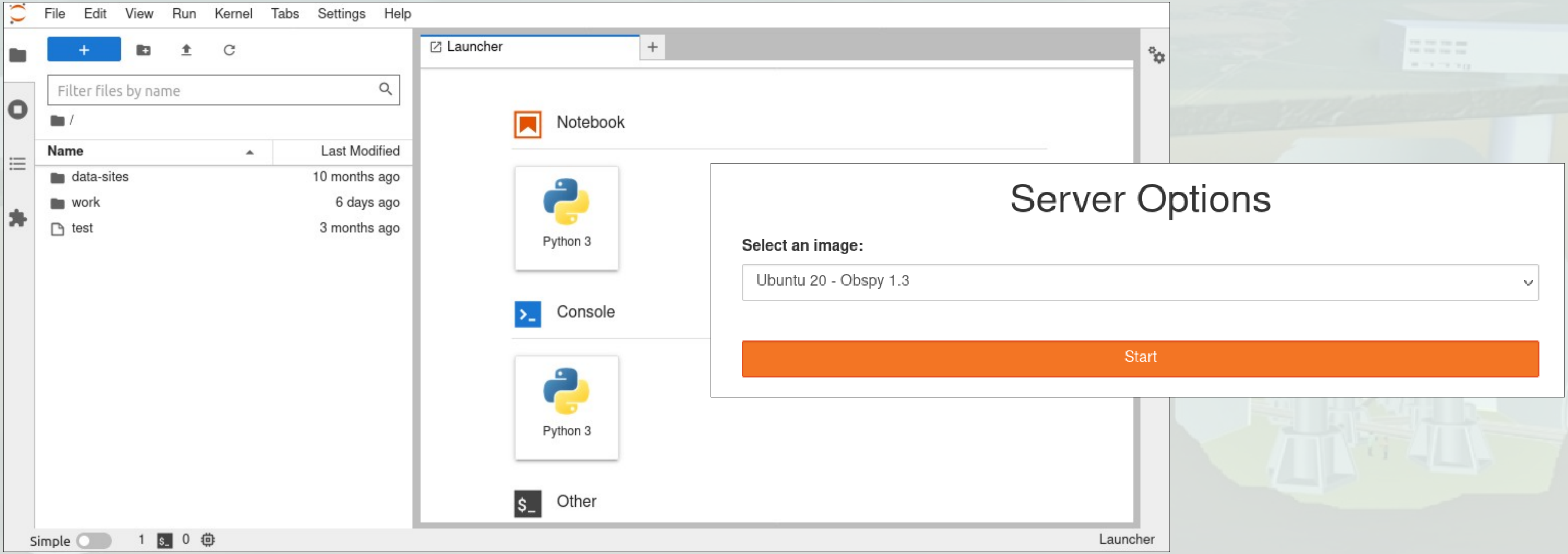
- **Infrastructure**
  - Based on JupyterHub server (https://jupyterhub.readthedocs.io/en/stable/)
  - SSL security enabled with username/password auth
  - Allows access to sites data and shared software
  - Links to user work dir (with manual permission changes)

- **Recent upgrades**
  - Migrated to a solution based on docker containers
  - More flexible, docker images upgreded via GitLab CI
  - 2 images so far with ObsPy 1.1 and 1.3

- **Issues**
  - Known issue with docker service (working on it!)
  - Please report any issue

JupyterLab

- **Reachable via browser at https://etrepo.df.unipi.it:8000**



Name	Last Modified
data-sites	10 months ago
work	6 days ago
test	3 months ago

Server Options

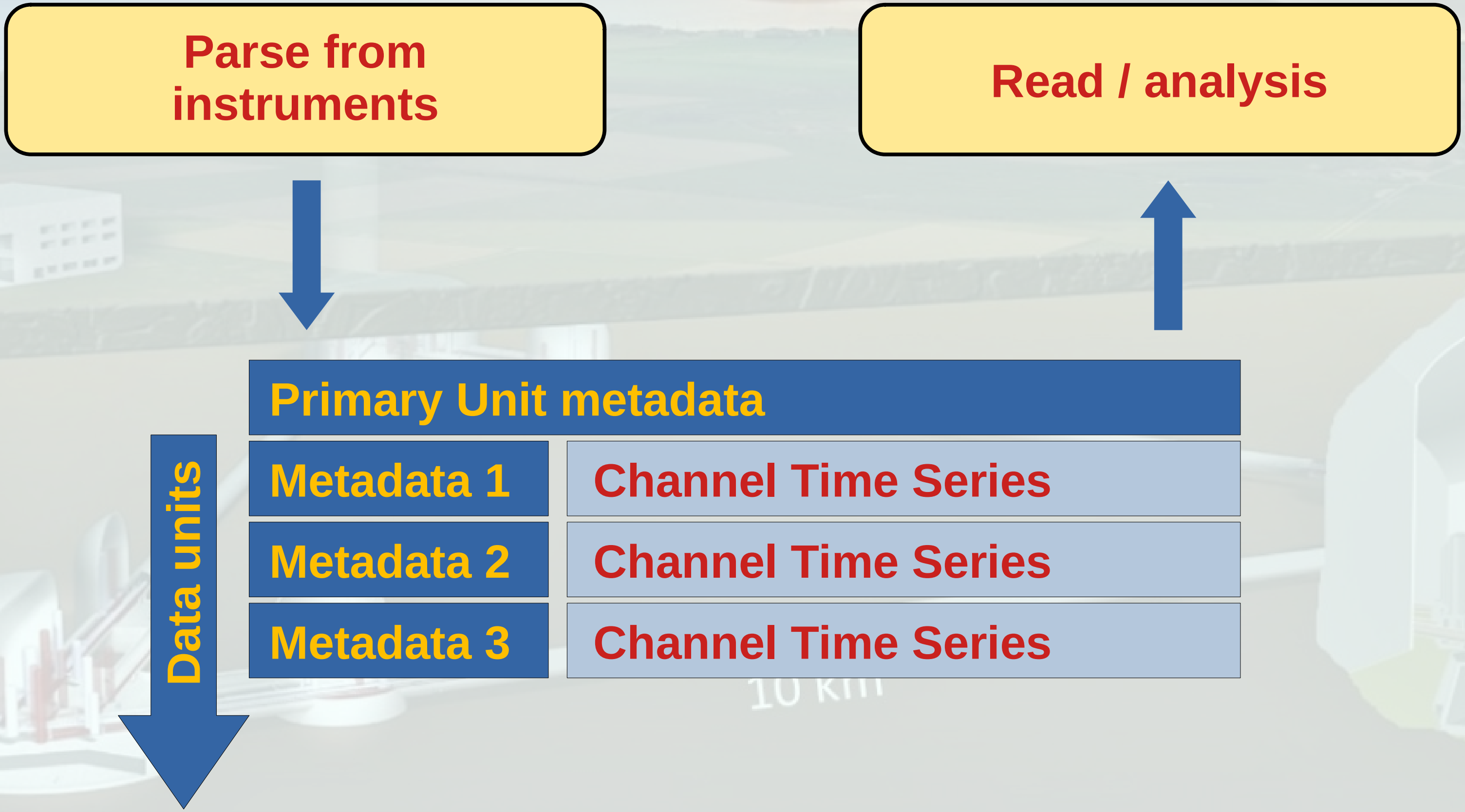
Select an image: Ubuntu 20 - Obspy 1.3

Data format: motivation

- **Requirements**
  - Manage data from different sources/instruments/format, both existing and future
  - Act as intermediate layer for existing formats (e.g. miniseed)
  - Collect and manage metadata from different instruments
  - Hierarchical, multichannel structure, similar to existing aux channels in GW detectors
  - Easy and fast I/O

- **Hierarchical Multichannel Data Format**
  - Concept as data formats like FITS, GWframes, mseed,etc...
  - Data streams in Data Units, containing metadata+channel data
  - Possibility to group channels (e.g. same sensor) and add Data Quality flags
  - Rely on HDF format as container easy to read in Python and other languages
  - HDF files also used in Adaptable Seismic Data Format (ASDF) for seismic data
  - Needed a full definition and a package to manipulate this format

Data format: proposed solution



Parse from instruments

Read / analysis

Primary Unit metadata

Metadata 1

Channel Time Series

Metadata 2

Channel Time Series

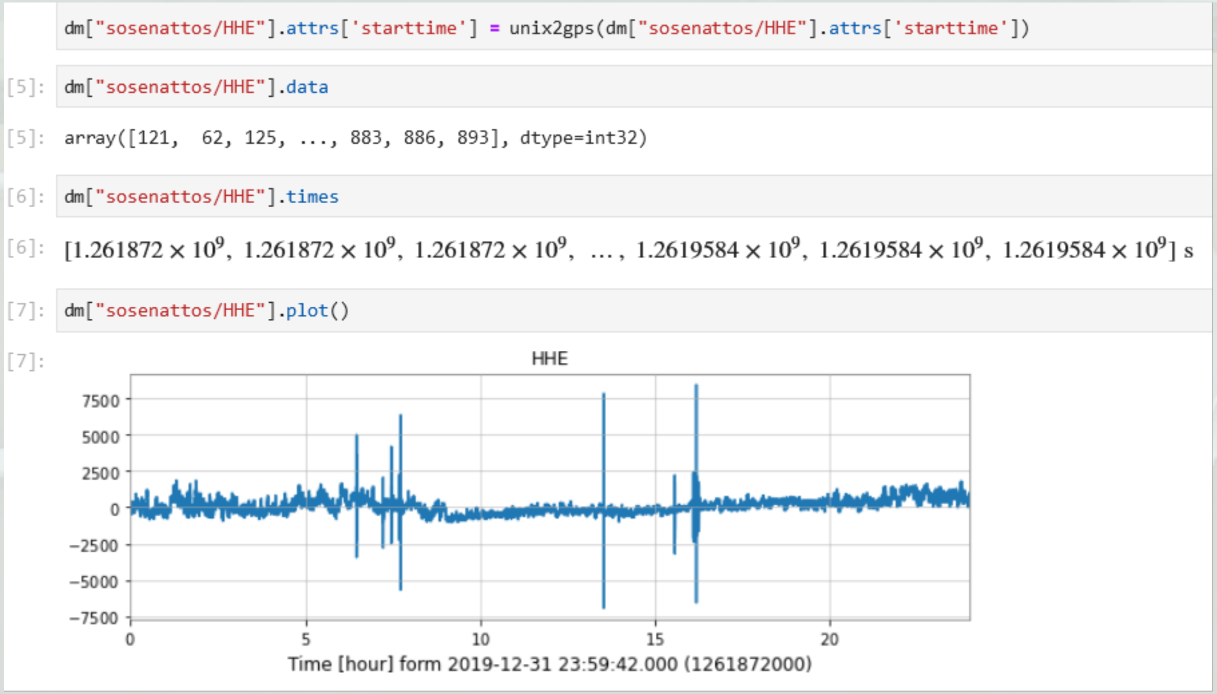
Metadata 3

Channel Time Series

Data units

GW Data Manager (GWDama)

- **Status:**
  - Based on library developed @unipi (M. Razzano, F. Di Renzo, N. Sorrentino, et al.)
  - Open-source, compatible with the main data analysis packages: **GWpy**, **Obspy** and **Pandas**
  - Multi-format/channel I/O:
    - GW data (gwf)
    - Open data center (GWOSC)
    - Geophysics (mseed, csv)
  - Hierarchical structure
  - Parse and manage from various sources into this format



```
      dm["sosenattos/HHE"].attrs['starttime'] = unix2gps(dm["sosenattos/HHE"].attrs['starttime'])

[5]:  dm["sosenattos/HHE"].data

[5]:  array([121,  62, 125, ..., 883, 886, 893], dtype=int32)

[6]:  dm["sosenattos/HHE"].times
```

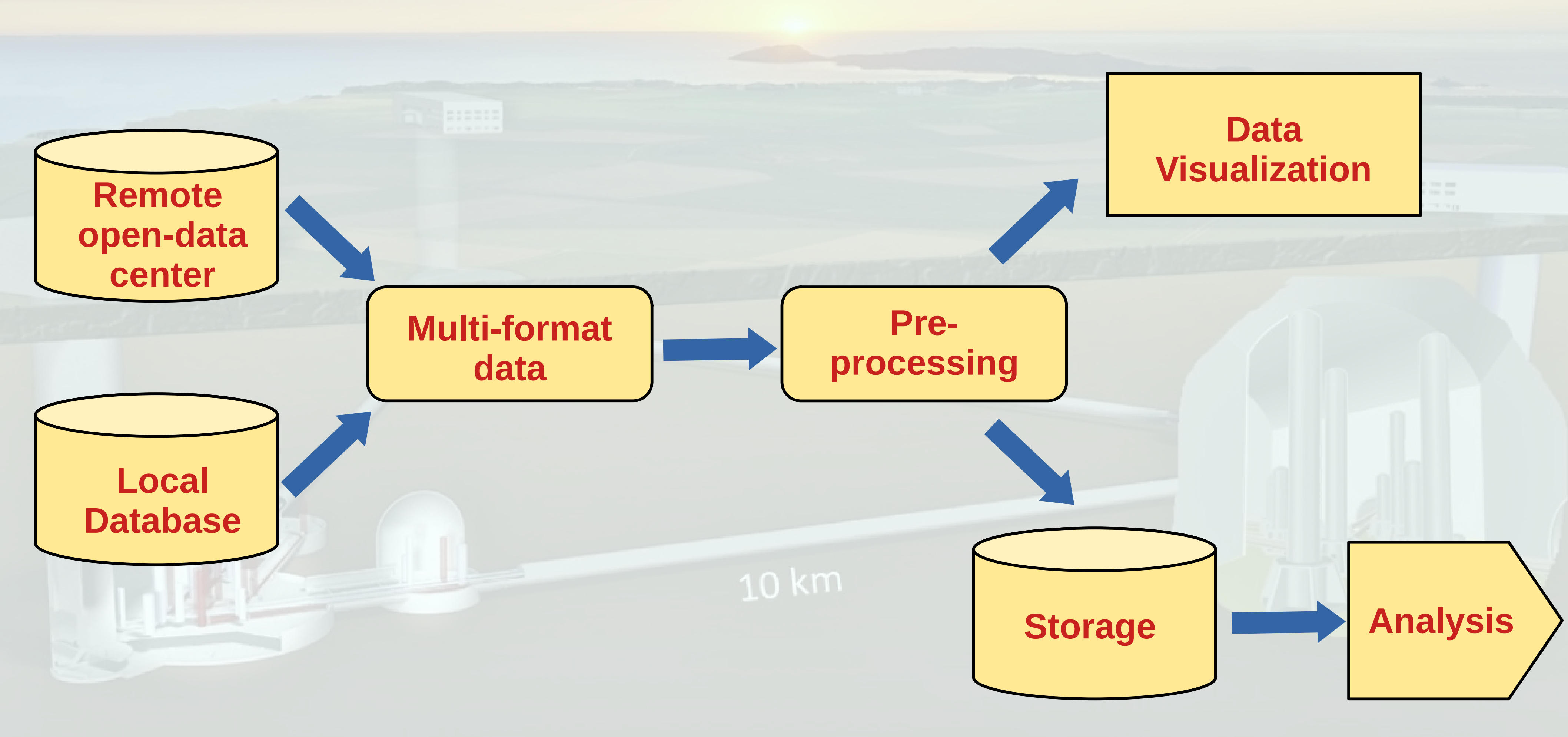
[6]: $[1.261872 \times 10^9,\ 1.261872 \times 10^9,\ 1.261872 \times 10^9,\ \ldots,\ 1.2619584 \times 10^9,\ 1.2619584 \times 10^9,\ 1.2619584 \times 10^9]\ \mathrm{s}$

```
[7]:  dm["sosenattos/HHE"].plot()

[7]:
```

https://gwnoisehunt.gitlab.io/gwdama/

Data organization proposal



Remote open-data center

Local Database

Multi-format data

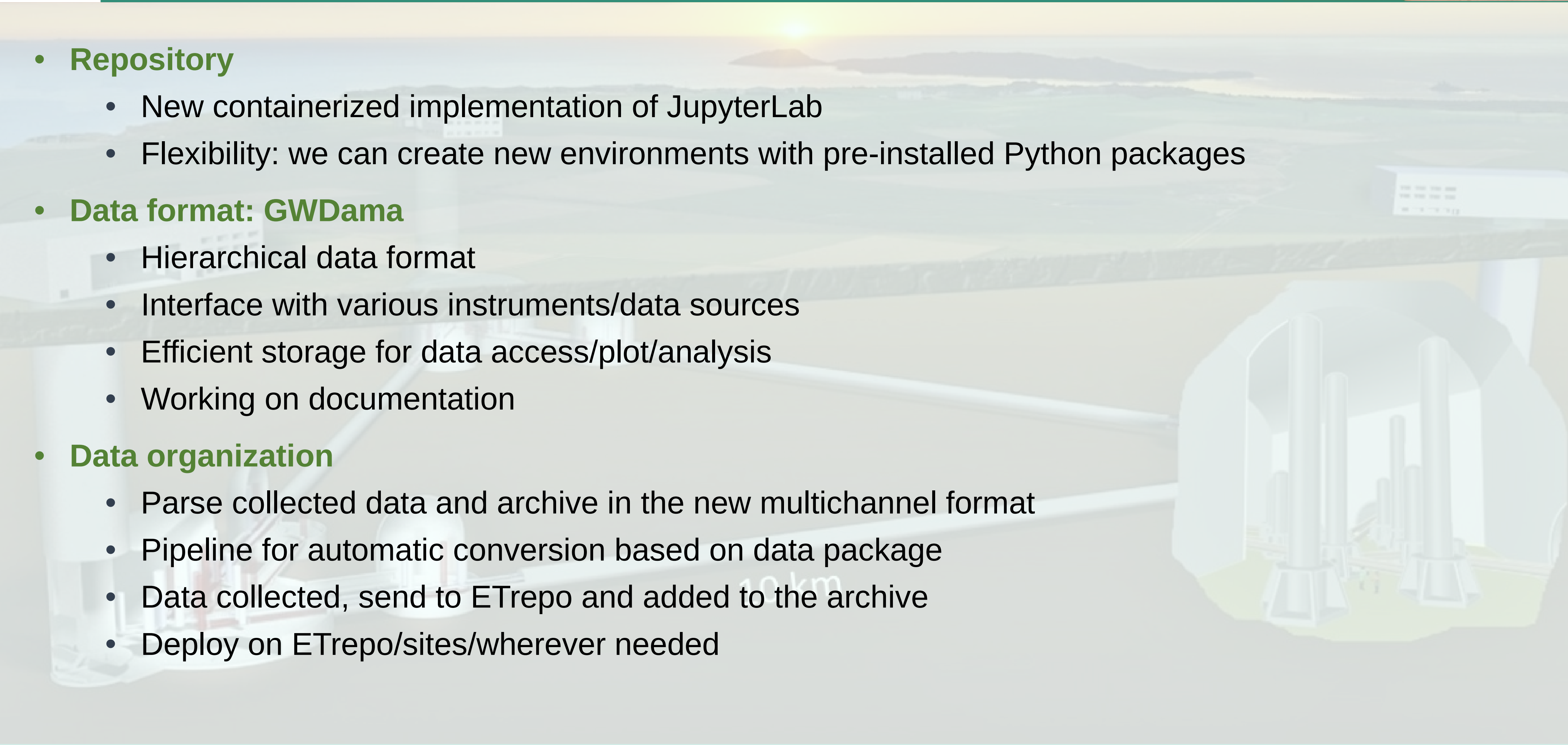
Pre-processing

Data Visualization

Storage

Analysis

Summary



- **Repository**
  - New containerized implementation of JupyterLab
  - Flexibility: we can create new environments with pre-installed Python packages
- **Data format: GWDama**
  - Hierarchical data format
  - Interface with various instruments/data sources
  - Efficient storage for data access/plot/analysis
  - Working on documentation
- **Data organization**
  - Parse collected data and archive in the new multichannel format
  - Pipeline for automatic conversion based on data package
  - Data collected, send to ETrepo and added to the archive
  - Deploy on ETrepo/sites/wherever needed