



UNIVERSITY OF TWENTE.

INTRODUCTION TO CLOUD COMPUTING FOR BIG GEODATA ANALYSIS

dr. ing. Serkan Girgin MSc
s.girgin@utwente.nl

Motivation

- **Not all geospatial problems require **cloud computing** technology.**
- Organizations are usually heterogeneous with respect to interests and needs, and for some people the topic **is not and will not be relevant or even interesting**.
- Even if there is *no apparent need or interest*, it is **still important** to have at least a basic understanding of the topic, because it is **becoming** a **key component** of the geospatial domain.
- This should be is an **organizational priority**.

Problem

- **Geospatial data** is getting **BIG** (e.g., satellites, drones, vehicles, social networks, mobile devices, cameras, etc.).
- **Large** and **complex** geospatial big data sets are difficult to handle using **traditional systems and methods** to analyse and extract information.
- Numerous **spatial computing** methods and systems have been developed to tackle the difficulties and enable **discovery, delivery, analysis, and visualization** of geospatial data.
- However, data processing and analysis tasks are **still** mostly performed on **local** workstations and they are **time consuming** (sometimes even ∞).

Solution?

- Recent developments in both **hardware and software** infrastructure have given big push and new direction to **geospatial data processing** capabilities.
- **Scalable and affordable** geospatial data analysis capabilities are available through:
 - **Open-source** systems that allow computing clusters on commodity hardware
 - **Proprietary** cloud-based data storage and computing services
- However, it is **challenging** to choose the **right solution(s)** depending on the nature of geospatial (big) data and the analysis needs.

Analysis Needs

- **Regional** conventional studies with medium size data
 - Analysis can be done **faster by parallel computing** on a **workstation**
- Machine learning and AI studies with medium size data
 - Analysis **requires special processing units** (e.g., **GPU/TPU**) due to computational complexity
- **National or multi-national (e.g. continental)** studies with big data
 - Analysis **requires distributed computing** on a **computing cluster** due to large volume of data or high computational complexity

All these analysis needs require specialized know-how and expertise, as well as adequate computing infrastructure...

... and transition in modus operandi!



Cloud computing is the on-demand availability of computer system resources, especially **data storage** and **computing power**, *without direct active management* by the user

Main Characteristics

- **On-demand self-service:** **provision of computing capabilities** as needed without requiring human interaction.
- **Broad network access:** availability over the Internet with **standard access mechanisms** for different client platforms (e.g., tablets, laptops, mobile phones).
- **Resource pooling:** dynamic **assignment and reassignment** of physical and virtual resources according to consumer demand.
- **Rapid elasticity:** capability to **scale rapidly** outward and inward proportionate to consumer demand.
- **Measured service:** accurate **monitoring, control, and reporting** of resource and service utilization.

They sound nice, but...

Status Quo

- Existing experience *is not widespread*, and difficulties exist in **identifying the cases** where *cloud computing* can play a role.
- Challenges exist in **proper selection and efficient use** of cloud computing *methods, tools, and services*.
- Available platforms and services are little used mainly due to **high cost and limited domain-specific technical support**.
- There is a high interest in getting **training** on *how to (better) use* cloud computing technology.
- There is also interest in **learning how** the technology is applied to solve *domain-specific problems* (e.g., what others do?)

Landscape



Source: <https://mattturck.com/data2021/>

Principle Needs

- State-of-the-art technical and scientific information should be **actively communicated** to the staff.
- **Proficiency** of the staff on cloud computing should be *improved*.
- Easy-to-use and efficient cloud computing infrastructure should be **made available** for training and work purposes.
- *Workflows* should be **enhanced and improved** with cloud computing technology where relevant.
- *Ad hoc* technical and scientific **support and advise** on cloud computing technology should be provided.
- Knowledge and good practices on better use of cloud computing technology should be **transferred** between *partner institutions*.

It is crucial to build a community that is self-motivated to learn, practice more, and share knowledge and experience!

Cloud Computing Services

- **Infrastructure as a service (IaaS)**

On-demand (virtual) hardware

- Provider supplies the infrastructure
- User deploys and run arbitrary software, including operating system
- Examples
 - [Amazon AWS](#)
 - [Microsoft Azure](#)
 - [Google Cloud](#)
 - [ESA DIASs](#)
 - National Research Clouds
 - ...

Low level: Fine control on resources, custom system design, optimum performance, but difficult to manage, requires expertise!

Cloud Computing Services

- **Platform** as a service (**PaaS**)
 - Provider supplies the infrastructure, services, and tools that allow the user to deploy applications
 - User deploys applications and alters settings of the application hosting environment
 - Examples
 - [Google Earth Engine](#)
 - [Microsoft Planetary Computer](#)
 - [ITC Geospatial Computing Platform](#)
 - [Google Colab](#)
 - [Amazon SageMaker](#)
 - ...

Medium level: Limited control on resources, custom workflow design, good performance, but requires programming skills!

Cloud Computing Services

- **Software as a Service (SaaS)**

On-demand software

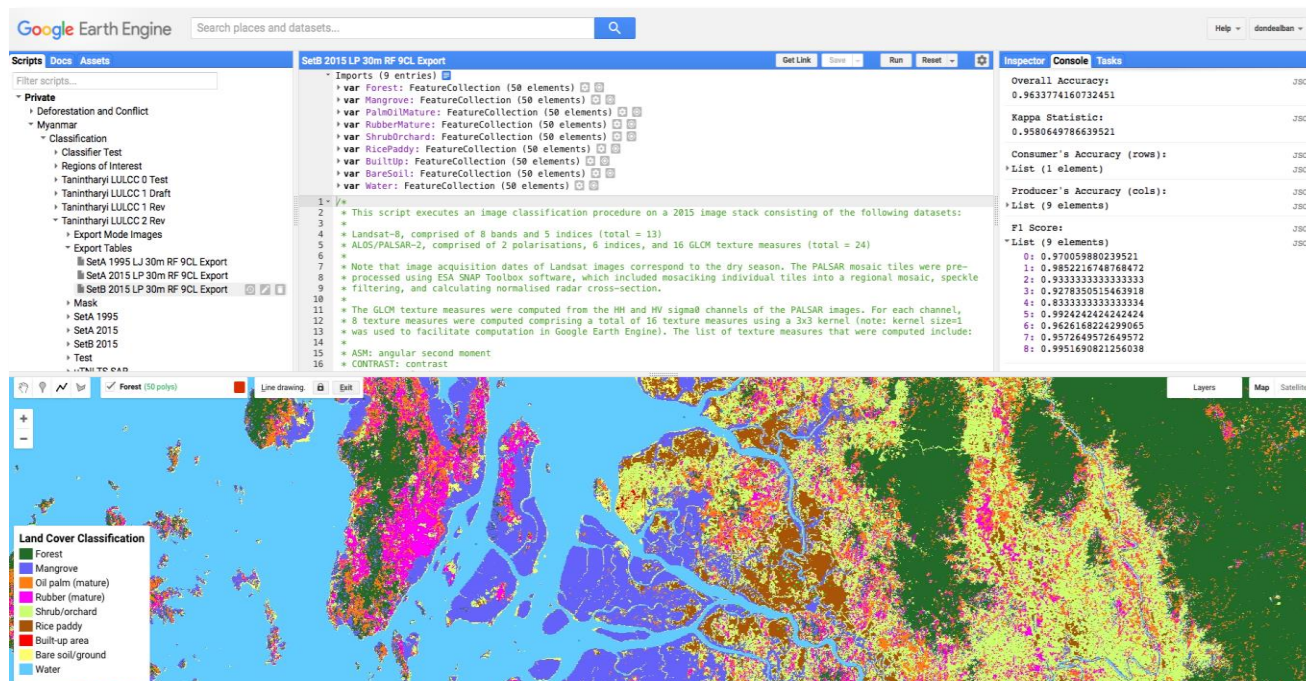
- Provider supplies the infrastructure that run the applications
- User uses provided applications through an interface
- Examples
 - [ArcGIS Online](#)
 - [CartoDB](#)
 - [Mapbox](#)
 - [R Studio Cloud](#)
 - ...

High level: Easy to use, (usually) optimum performance, but no control on resources, usually paid!

There are also others, e.g., Function as a service (**FaaS**), Data as a service (**DaaS**), Data Processing as a service (**DPaaS**), etc.

Google Earth Engine

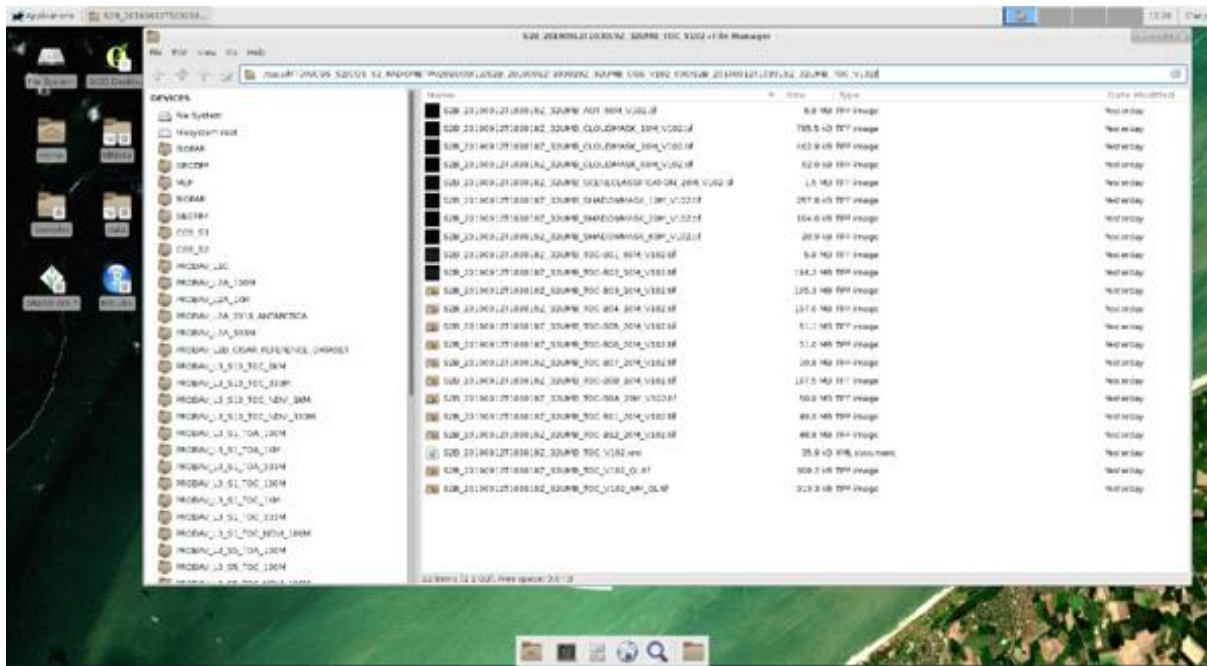
- **GEE** combines a multi-petabyte catalog of EO imagery and geospatial datasets with planetary-scale analysis capabilities available for free*.



<https://earthengine.google.com/>

Terrascope

- **Terrascope** provides VMs (4 vCPU, 8 GB RAM, 80 GB storage) with ready to use datasets (e.g., Sentinel-1, Sentinel-2, SPOT-VEGETATION, PROBA-V) and customizable pre-installed environment (e.g., QGIS, SNAP) for free*.



<https://terrascope.be/>

Geospatial Computing Platform

- **GCP** provides **GPU-enabled** (8 vCPU, 32 GB RAM, unlimited storage) and **Big Data** (72 vCPU, 768 GB RAM, unlimited storage) VM **clusters** with ready to use datasets (e.g., OSM), customizable pre-installed **interactive** and **desktop** environments, and **shared workspaces**.

The collage illustrates the Geospatial Computing Platform's capabilities. It features a QGIS desktop environment with a map of the Netherlands, a Jupyter Notebook with Python code for rasterio and ggplot2, a histogram plot, and a document titled 'Aerial survey' with text about digital photography. An NVIDIA GPU card is also shown.

QGIS Desktop Environment: The interface shows a map of the Netherlands with a layer named 'QGIS Open Day 28 May 2021'. The map is displayed in a web browser window.

Jupyter Notebook: The notebook displays Python code for rasterio and ggplot2. The code includes a list of files in the '/ platform / demo /' directory, a rasterio plot, and a histogram plot.

Aerial survey Document: The document titled 'Aerial survey' contains text about digital photography and the use of aerial photographs for data collection. It mentions that aerial photographs are a major source of data for geospatial analysis and that they can be used to capture data in two or three dimensions.

NVIDIA GPU Card: An NVIDIA GPU card is shown, indicating the platform's GPU-enabled capabilities.

<https://crib.utwente.nl/>

Available Software



... and many more: **800+ Python** and **400+ R** packages!

Available Services



GeoServer

Open source server for sharing
geospatial data



MapServer

Open source platform for
publishing spatial data



PostgreSQL

Open source relational database



MariaDB

Open source relational database



GeoNode

Open source geospatial content
management system



Dataverse

Open source research data
repository software



Gitea

Open source lightweight code
hosting solution



Open Data Kit

Open source platform to collect
data quickly, accurately, offline, and
at scale

Support

Welcome to the CRIB Support Center!

In order to streamline support requests and better serve you, we utilize a support ticket system. Every support request is assigned a unique ticket number which you can use to track the progress and responses online. For your reference we provide complete archives and history of all your support requests.

Quick Access

- [Report a Problem](#)
- [Shared Workspace Request](#)
- [Course Workspace Registration with Canvas Integration](#)
- [External Account Request](#)
- [Account Removal Request](#)
- [Account Transfer Request](#)
- [Software Request](#)
- [Dataset Request](#)
- [Database Request](#)

Featured Questions

[How can I access to the platform?](#)

[Is it secure?](#)

[How can I use the platform?](#)

[Which programming languages are supported on the platform?](#)

[Which libraries and packages are supported by the platform?](#)

<https://crib.utwente.nl/support/>

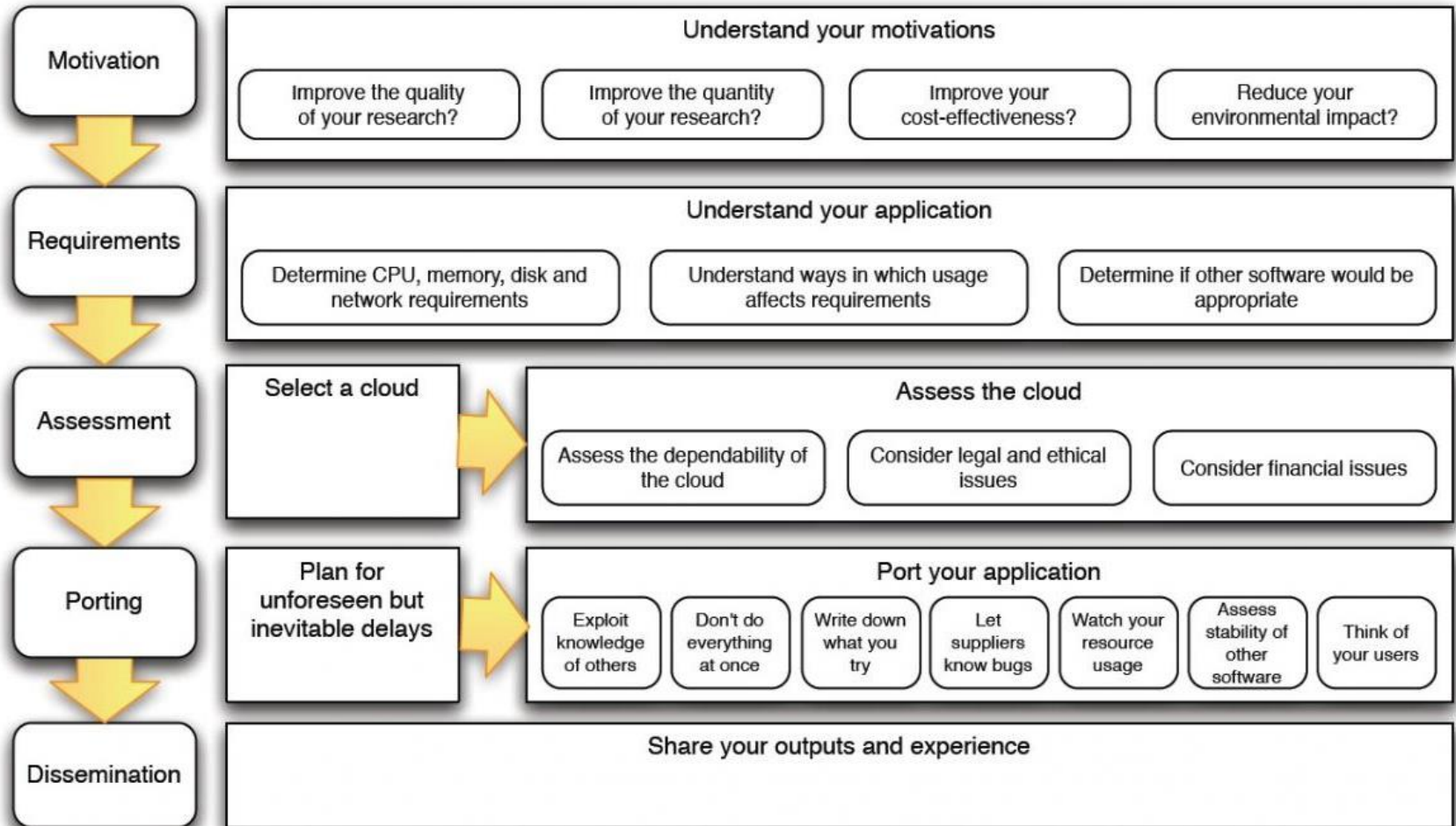
Potential Benefits

- Better **computing infrastructure** (e.g., more CPUs, GPUs, RAM)
- Better **storage** (e.g., replicated, backup)
- Better **scalability** (e.g., more resources on-demand)
- Improved workflow **performance** due to *colocation of data and computing resources* (i.e., no download)
- Improved **collaboration** (e.g., direct access to same assets)
- Improved **resource utilization** (e.g., less idle time)
- No cost for **investment and maintenance** (*if remote cloud*)
- Low cost for **extensive use** (*if local cloud*)

Suggestions

- **Ensure familiarity** with the cloud computing technology through short talks and lectures
- **Improve know-how** by participating tool- and technology-specific training
- **Try and use** the infrastructure and platforms available for free or through partner organizations
- **Follow** a hybrid approach (i.e., local + cloud) to maximize the benefits
- **Ask for technical and scientific support** for better implementation and integration of the technology.
- **Ask for guidance** for the planning of future activities.
- **Share your knowledge** and good practices (e.g., for cost-effective and efficient use of the technology) with your colleagues and partners.

Moving to the Cloud



Source: [Best practice for using cloud in research \(Hong et al., 2018\)](#)

Big Geodata Newsletter

<https://itc.nl/big-geodata/newsletter/>

Subscribe Now

Cloud credits for Earth observation projects



Do you know that most of the academic research? **RESEARCH GRANTS**

Amazon has a **RESEARCH GRANTS** program that provides an **EC2** credits up to 100,000 USD. There are also Azure training materials, and education opportunities. **EC2** is the next deadline being.

We already have an **AI** for **E** team will use Azure credit using Deep Learning on the **E** if you have project ideas, but we can help!

Landsat Collection 2 will be available in mid-2020



This may look like an ordinary Landsat 8 scene, but actually it marks an important

EUMETSAT phasing in new data services



EUMETSAT data store

A data catalogue spans a wide range of data services that will be a core, data visualisation **EC2** is already available in space and also reference systems, latest interface, Data interface (CLI). For more

experience with EUMETSAT data store in for **EC2** we also operate a **EC2**

cuSpatial: CUDA-accelerated GIS and spatiotemporal algorithms



Spatial Data Cubes



eScience Center - ITC collaboration on large-scale phenological analysis



Large amounts of Earth observation data are needed for developments in cloud-based geodata sets are challenging

Apache Spark meets GPU



Spark 3.0

CRIB web portal is online



Currently one of the big data-scale data processing also supports a rich set of services, **EC2** for machine learning for stream processing, extended its capabilities

work was mainly limited to environments (e.g. up to 1000000 Spark 3.0.0)

GeoSpark becomes Apache Sedona



WebGL-powered visualization of large-scale datasets



GeoSpark was an open-source spatial data engine. **SPATIAL RESILIENT DISK** spatial analysis performance based on a distributed computing building operators to visualise accepted unanimously by the Apache incubator group

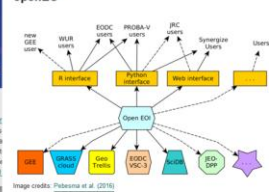
Sustaining an open-source ecosystem Apache ecosystem active user and developer of the Apache umbrella, hence **Sedona**

Big geospatial data is not only difficult to analyse, but it is also quite challenging to visualise. This is especially the case for large **EC2** vector data that should be visualised on web browsers, which is the common way to access data on the cloud. **EC2** provides a high performance, **WebGL**-based platform for visualization of such large data sets. It supports tiled layers, various **EC2** styles, cartographic projections, environmental lighting and provides performance rendering leveraging **GPU**. **EC2** is a powerful open source geospatial analysis tool to explore geo-temporal data, is built with **EC2**. The platform can also be combined with **EC2** Earth Engine. Developed mainly by **EC2** (Community team) in the last 5 years, **EC2** has recently moved to an open governance model in August, which will allow community-driven planning and development process.

OCRE is launching cloud and EO funding calls



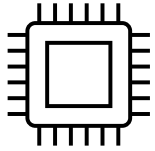
openEO



Using cloud-based platforms to access and process big EO data is becoming the new norm, especially for large-scale studies. However, current EO cloud back-ends have different APIs, requiring significant time and effort to get acquainted and use them efficiently. It is difficult to compare their capabilities and costs, or to combine them in a joint analysis. Validation and reproduction of the analysis results between the platforms is also challenging.

openEO aims to help with these difficulties by providing an **open API** to connect to Python, JavaScript and other clients (e.g. **QGIS**) to different EO cloud back-ends in a simple and unified way. A web-based **openEO** editor is also available for interactive use. The project was funded by **ESA** (ESA Big Data Shift call) and will be refreshed this month. But a new project funded by **ESA**, **openEO Platform**, is just started to bring **openEO** to production and offer data access and processing services to the EO community.

Contact



<https://crib.utwente.nl>



<https://itc.nl/big-geodata>



crib-itc@utwente.nl



[@BigGeodata](https://twitter.com/BigGeodata)



dr.ing. Serkan Girgin MSc

Head of the Center of Expertise in Big Geodata Science

Assistant Professor

Faculty of Geo-information Science and Earth Observation

University of Twente

s.girgin@utwente.nl

+31 53 489 55 78

