FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION (ITC)

UNIVERSITY OF TWENTE.

INTRODUCTION TO CLOUD COMPUTING FOR BIG GEODATA ANALYSIS

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

28 February 2022

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 <u>both</u> **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 <u>commodity hardware</u>
    - **Proprietary** <u>cloud-based</u> data storage and computing services
- However, it is **challenging** to choose the **right solution(s)** depending on the <u>nature</u> of geospatial (big) data and the analysis <u>needs</u>.

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 <u>play a role.</u>
- Challenges exist in **proper selection and efficient use** of cloud computing *methods, tools, and services*.
- Available platforms and services <u>are little used</u> mainly due to **high cost and limited domain-specific technical support.**
- There is a <u>high interest</u> in getting **training** on *how to (better) use* cloud computing technology.
- There is also interest in **learning how** the technology is <u>applied to solve</u> *domain-specific problems* (e.g., what others do?)

Landscape

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

Principle Needs

- <u>State-of-the-art</u> *technical and scientific information* should be **actively communicated** to the staff.
- **Proficiency** of the staff on cloud computing should be *improved*.
- <u>Easy-to-use and efficient</u> *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 <u>provided</u>.
- <u>Knowledge and good practices</u> 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

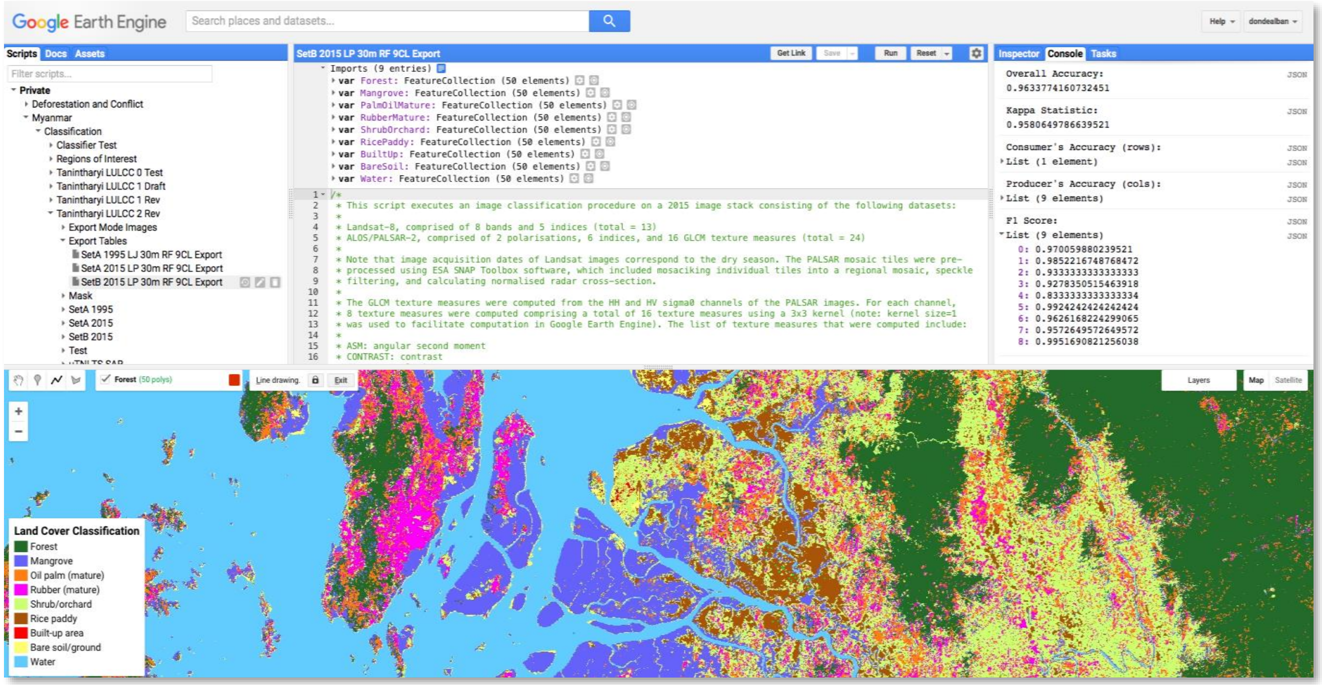
- **Software** as a Service (**SaaS**)
  On-demand software
  - Provider supplies the infrastructure that run the applications
  - User uses provided applications through an interface
  - Examples
    - ArcGIS Online
    - CartoDB
    - Mapbox
    - R Studio Cloud
    - …

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

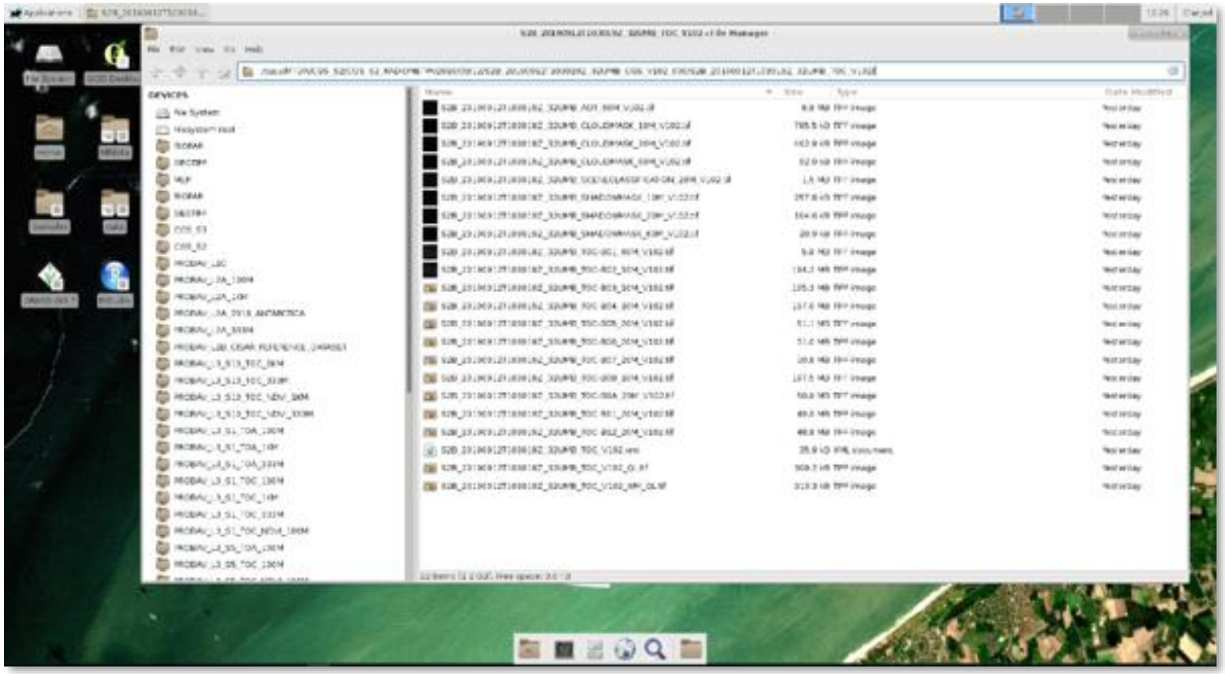
- **GEE** combines a multi-petabyte catalog of EO imagery and geospatial datasets with planetary-scale analysis capabilities available for free*.



https://earthengine.google.com/

Terrascope

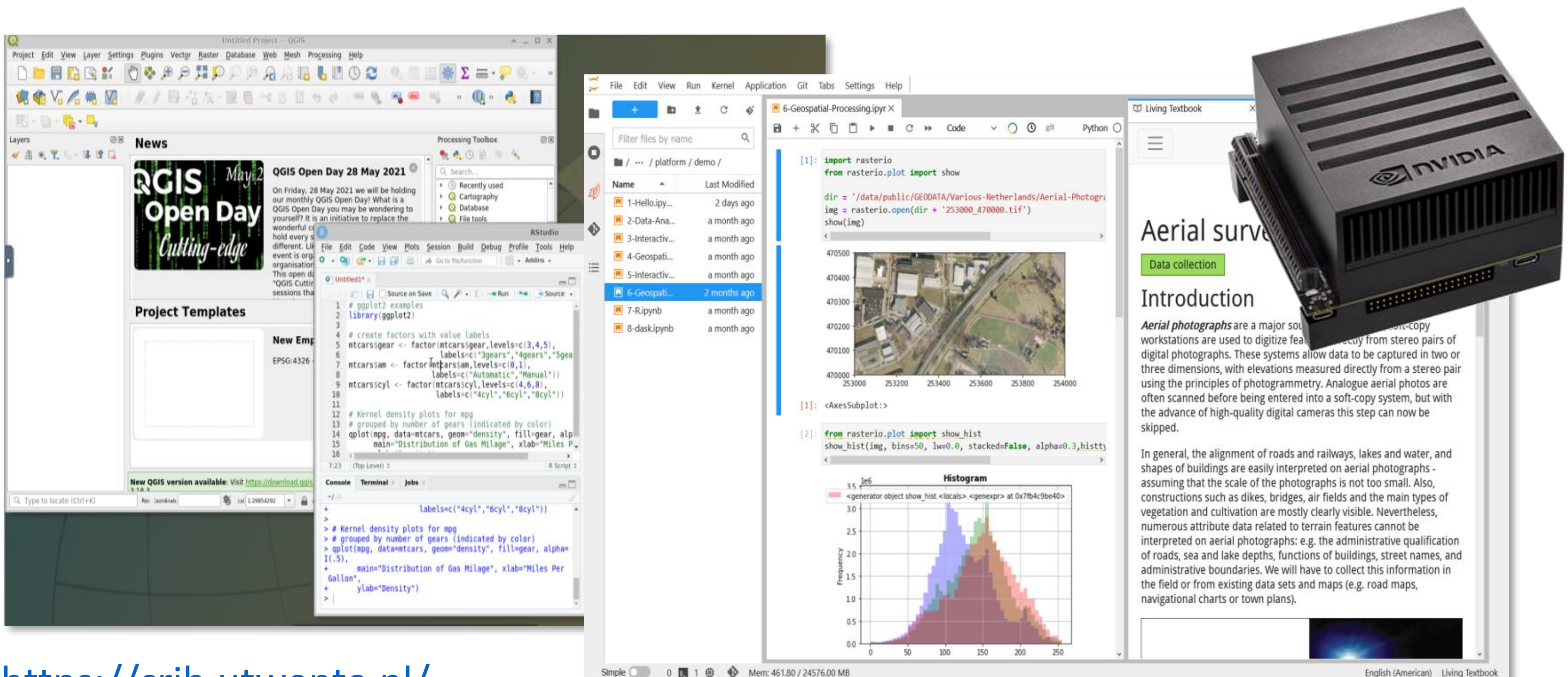
- **Terrascope** provides VMs (4 vCPU, 8 GB RAM, 80 GB storage) with ready to use datasets (e.g., Sentinel-1, Sentinel-2, SPOT-VEGETATION, PROBA-V) and customizable pre-installed environment (e.g., QGIS, SNAP) for free*.



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), <u>customizable</u> pre-installed **interactive** and **desktop** environments, and **shared workspaces**.



https://crib.utwente.nl/

Available Software

Coming Soon!

... 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

UNIVERSITY OF TWENTE. | CRIB

Support Center Home | Knowledgebase | Open a New Ticket | Check Ticket Status

Search our knowledge base [Search]

[Open a New Ticket]

[Check Ticket Status]

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/

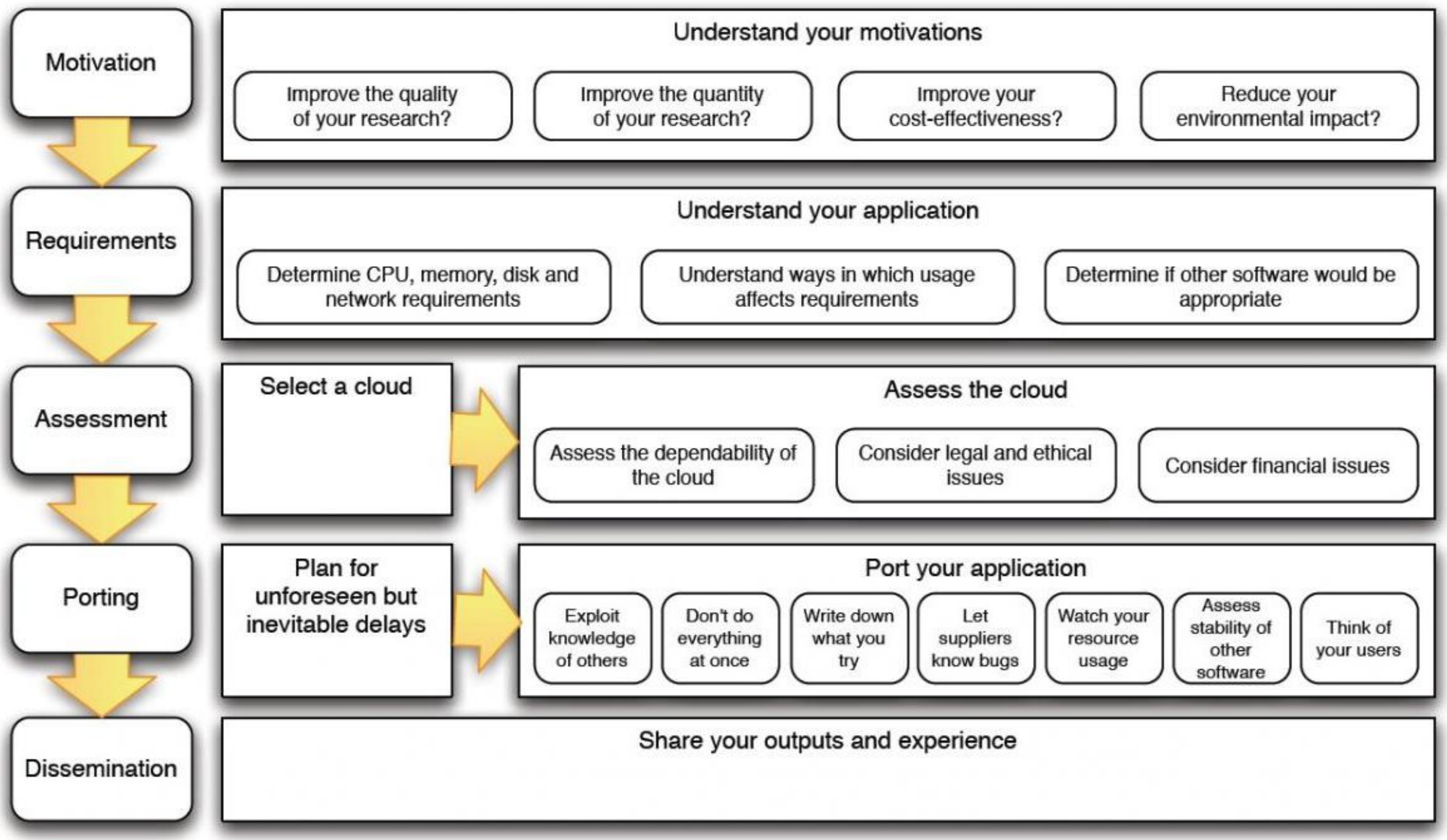
Potential Benefits

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

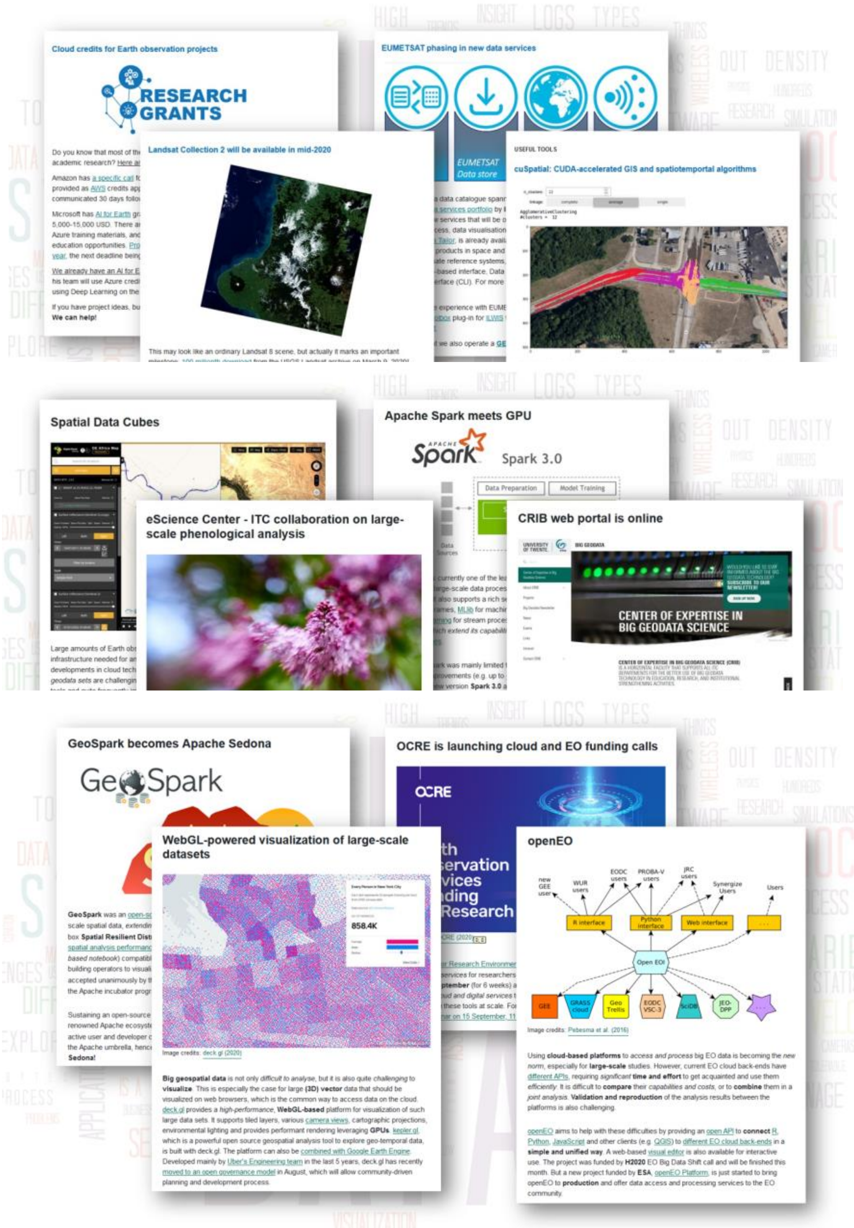
Suggestions

- **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

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