Big Geodata at ITC

Status Quo and Roadmap

Revision	Date	Created By	Short Description of Changes
1.0	2020/06/10	Serkan GIRGIN	Public draft

Introduction

This document presents the current state of the big geodata-related technologies at ITC and discusses the suggested actions for more effective and efficient use of these technologies in education, research, and institutional strengthening activities.

To describe the current state, including the challenges and difficulties as well as interests and future needs, findings of the meetings with key staff and results of the surveys sent to PhD students and selected research staff are presented. Following an overall assessment of these findings, basic principles for better use of big geodata technology are identified. A high-level roadmap is provided for the development of an easily accessible, efficient, and cost effective big geodata infrastructure and for the improvement of the institutional knowledge and know-how on big geodata technology. The main goals and the corresponding short- and long-term actions are proposed under several categories based on the roadmap. Finally, some remarks on the scope of the assessment and further work are discussed.

We would like to thank staff members who shared their experience and provided helpful information through open-minded discussions and interactions, and also to Rolf A. de By and Raul Zurita-Milla who commented on earlier drafts of the document and provided valuable feedback.

In this document:

Big data technology refers to methods, tools, and services for the collection, management, analysis, and visualization of data that are too large or complex to be handled by a typical workstation.

Big geodata technology refers to big data technology involving geospatial data and related analysis methods. Whenever a topic is not only specific to the geospatial domain, but also linked to big data technology in general, the *geo* term is indicated in parentheses, i.e. big (geo)data.

Computing infrastructure refers to large-scale storage, networking, and processing hardware including specialized units (e.g. GPU/TPUs) and related software required for its operation.

Status Quo

To reveal the current state of the use of big geodata technology and relevant future needs at ITC, several meetings and surveys were conducted.

The aim of the meetings was to collect information at department level through a number of staff members who are experienced in big geodata technology. At least one representative from each department, including CRIB User Sound Board members, was interviewed. In addition to the departmental experts, information was also collected from staff working on faculty-wide services, such as the RSG-Lab and LISA. The meetings provided comprehensive information about the status quo, including the major challenges and difficulties, at organizational level.

To obtain insights into the existing degree of knowledge of and interest in big data-related technologies at individual level, surveys were performed with PhD students and selected research staff. The surveys provided useful information for the identification of the knowledge gaps.

Key Findings of the Meetings

The following points are identified based on the meetings with the selected ITC staff:

<u>Knowledge</u>

- ITC has staff with varying degrees of expertise in big geodata technology, including cloud computing, machine learning, and deep learning methods and tools.
- Large-scale distributed (i.e. out-of-core) computing for big geodata is not practiced much.
- Existing experience is not widespread and most of the staff and students have difficulties in identifying the cases where big data technology can play an important role.
- Challenges also arise in assessing which big geodata tools and methods are suitable for the research problems and how they can be utilized efficiently.
- There is a high interest in training activities on how to (better) use big data technology (e.g. methods, tools, frameworks, services etc.).
- There is also interest in learning how the technology is currently applied to solve domain-specific problems (i.e. how other people are using these methods? what kind of problems are solved?).

<u>Infrastructure</u>

- ITC does not have a common geospatial computing infrastructure. Therefore, a big geodata infrastructure also does not exist.
- Each scientific department has its own computing infrastructure solution based on its specific needs, which involves local (i.e. hosted at the department), institutional (i.e. hosted/provided by LISA), or cloud-based infrastructure (e.g. Microsoft Azure).
- UT does not provide a common computing infrastructure.
- LISA has well-established (e.g. featuring UPS, cooling, physical security, fire protection, etc.) data centres which can host hardware fulfilling certain criteria at no cost. Network and basic maintenance (e.g. OS updates, low-level cluster management) are included.
- Several ITC departments have servers hosted at LISA premises.
- LISA provides a paid large-volume storage service (i.e. NAS), which allows traditional shared file access.
- Several ITC departments and facilities (e.g. RSG-Lab) utilize the LISA storage service to store their assets (e.g. data sets).

- SURFsara also provides storage capabilities (e.g. DANS, 4TU Data Centre), which are used for research data archival as part of ITC's data policy other than a few research projects.
- UT has a special agreement with SURFcumulus for Microsoft Azure cloud services. The discount rate is around 10-15%.
- Various UT departments have cluster and high-performance computing infrastructure with GPUs (e.g. DSI clusters) which can be used for research purposes by ITC staff upon request. A contribution (e.g. additional hardware) is required in case of heavy use.
- Departments have different practices for the management of their own computational resources.
- Departmental computational resources are usually managed by staff who have other primary roles. This complicates administration and maintenance, especially due the high turn-over rate of end users (e.g. PhD students, staff).
- Departmental computational resources are not arranged or configured as clusters. This limits their use for large-scale computations and (probably) results in less effective capacity utilization because few people can use them.
- Currently, some departments are either planning for investment or are already implementing an investment plan for new computational resources.

Research

- Research projects do not fully benefit from available big geodata resources and technologies.
- At project proposal stage, there are difficulties in quantifying the needed computational resources, especially cloud services (i.e. it is not easy to answer the question "how many cloud credits?")
- Departmental computational resources are mainly allocated to staff and PhD students for research purposes.
- Computational resources available for MSc students are usually limited. They are also not indicated as big data users, but this may change from department to department.
- Access to departmental resources beyond the department staff and students is usually restricted (i.e. inter-departmental access is limited).
- Resources linked to projects (e.g. servers, GPUs) are occasionally under-utilized after the completion of the projects.
- Remotely shared assets (e.g. datasets hosted at LISA) are usually analysed locally, i.e. downloaded to the local computer and processed subsequently. For most use cases this practice is considered sufficient by the users. It is also very likely that it is simply the only known method for some users.
- Commercial cloud services offered by UT (i.e. Microsoft Azure) are very little used mainly due to high cost and limited technical support, which usually takes a considerable time even for simple support requests.
- National infrastructure (e.g. SURFsara) is usually not a preferred option for research purposes due to additional administrative overhead and cost. Experience varies on this; good cooperation is also reported.
- Occasionally national or commercial infrastructure can be used at no cost through special agreements, which may also include exclusive technical support (e.g. dedicated technical staff for a certain duration). However, these agreements are project specific and not institutional.

- Although infrastructure providers are experienced in typical data problems and can provide related training and technical support, their experience with big geodata, such as multi-temporal and multi-dimensional raster data sets, is limited.
- Software developers can help with specific computing needs. However, not all departments have such dedicated staff and if available they are usually attached to projects, i.e. serve mainly for project-specific needs. They might also have limited knowledge on certain topics, such as big geodata.

Education

- Although there are courses related or relevant to big geodata technology, courses *on* big data technology are limited (see Annex I).
- Students are heterogeneous in technical skills, especially programming skills which are required for some computation-intensive courses offered to *all* disciplines. This is highlighted as a challenge for big geodata analysis.
- It is the responsibility of the instructors to make the required infrastructure available to the students for educational purposes.
- Solutions are usually ad hoc and include local installation of software (i.e. use of personal laptops of the students), setting up a custom server for remote shell or UI access (e.g. interactive notebooks), and use of (free) cloud services.

Key Findings of the Surveys

The results of two different surveys with similar questions to assess the current level of knowledge of and interest in big geodata technology among the PhD students and the staff expressing interest in big geodata technology (i.e. participants of the Data Science and AI Workshop organized in November 2019) are given in Annex II. An example survey is available at this link.

According to the surveys, the staff and students show similar and equally high interest in different big data topics, such as big data analysis, machine learning, and deep learning (Figure A2.1). Only for big data storage and retrieval the interest of the students is found to be considerably low. Given that their studies usually focus on methods and applications at specific case-study regions, less interest in data management is understandable. Overall low interest in data analytics can be partly attributed to its confluence with data analysis.

Most of the participants indicated that they are working with or interested in big geospatial raster data updated in batches, such as satellite imagery (Figure A2.2). Interest in other big geodata products updated in batches is also high, whereas streaming big geodata (e.g. sensor data) seems to be less in use. However, there is significant interest in streaming big data that needs to be geolocated, such as news and social media feeds.

Because efficient use of big geodata technology requires acquaintance with basic spatial and big data technologies on which they are based, the surveys also measured experience with the selected tools in the following categories:

- Low-level parallel computing libraries
- High-level parallel computing frameworks
- Low-level spatial libraries
- Databases
- Spatial databases
- Machine learning frameworks

- Cloud computing services
- Large-scale storage services
- Large-scale cloud databases
- Large-scale cloud spatial databases
- Cluster-computing frameworks
- Spatial cluster-computing frameworks
- Spatial cloud computing services

The knowledge on the selected tools measured in a 5-level scale ranging from *I didn't hear* to *I'm using* are given in Figure A2.3 and Figure A2.4 for the staff and PhD students, respectively.

As anticipated, the staff is more experienced with mature tools and technologies (e.g. databases, spatial libraries). There is also a higher level of recognition of high-level computing and machine learning frameworks. However, the students are more actively using these technologies.

Among low-level computing frameworks CUDA is reported to be widely used especially by the students but given the degree of programming experience required for its direct use, this is probably limited to the use through other high-level frameworks utilizing CUDA (e.g. TensorFlow, PyTorch).

There is a high degree of use of high-level serial computing frameworks (e.g. Numpy). However, familiarity with their parallel counterparts (e.g. Dask) is very limited.

Almost no experience exists among the students for large-scale cloud (spatial) databases and (spatial) cluster-computing frameworks. The situation is also similar for the staff, which is limited to a few tries and use cases. Overall, know-how on big data tools which require low-level development (i.e. coding) or integration of various tools for problem solving is found to be quite low.

However, although it also requires a considerable amount of coding, there is a significant usage of Google Earth Engine. This indicates that both the staff and students prefer to use integrated solutions hiding the complexity of handling big data and allowing them to focus on domain-specific applications.

Although familiarity with cluster computing is low, there is a noticeable acquaintance with distributed file systems (e.g. EOS, HDFS). This is rather surprising as these technologies are closely related. One explanation can be the wrong reminiscence of the abbreviations which are rather not very specific.

Assessment

Although the analysis is aimed at assessing the current use and future needs related to big geodata technology, the results indicate that the needs are in fact broader and point out the need for support in various analysis tasks, which are related but not limited to big data. These can be roughly divided into three categories based on the analysis requirements:

1. Analyses that can be done faster by parallel computing on a single computer (e.g. continental-scale studies with medium-size data).
2. Analyses requiring special processing units (e.g. GPU) due to computational complexity (e.g. machine and deep learning).
3. Analyses requiring distributed computing on a cluster or in the cloud due to large volume of data or high computational complexity (e.g. global-scale studies with big data, e.g. petabyte scale).

There are individual experts at ITC who have experience and know-how in each category. *However, the overall institutional expertise is rather limited and more extensive use of the related tools and methods by a larger number of ITC staff and students will bring significant benefits to education, research, and institutional strengthening activities.*

Interest of the staff and students in all three categories is evident. Support in the first two categories may result in faster achievements as there are already activities that can benefit from the potential (performance) gains. Currently, a very limited number of activities have been identified that fall within the last category. *But given the interest, such big data activities can come in sight with increasing know-how and accessibility to the required infrastructure.*

Among the categories only the last one can be considered literally a part of the big data domain, hence directly related to CRIB. Although they also require special expertise, the others rather involve problems that can be (relatively) easily solved by using appropriate tools and methods without dealing with the complexities of computing infrastructure. Given the similarities in the analysis methods, these activities can also be supported by CRIB. Therefore, one strategic decision is to define the scope of the CRIB activities with respect to the listed categories.

Considering their importance, this document assumes that all three analysis needs are included in the scope and they are termed altogether as *big data technology*. However, the scope can be limited based on the final high-level decision following the discussions on the findings of this document.

As a big faculty with more than 300 staff and PhD students, ITC is quite heterogeneous with respect to interests and needs related to big data technology. Not all spatial problems involve big data or require resource-intensive computations. Therefore, for some people the topic is not and will not be interesting.

However, even if there is no apparent need or interest in using big data technology, it is still important to have at least a basic understanding of the topic since it is becoming a crucial component of the geo-information and earth observation landscape. For this reason, it is important to ensure that all ITC staff and students are familiar with big data technology and receive updates on the major developments in the field. Survey results, especially the high number of *I didn't hear* responses, highlight the deficiency (Figure A2.3 and Figure A2.4). Therefore, this should be an institutional priority.

Five user groups can be identified at ITC and these are listed in Table 1 with their specific needs and key actions required to serve these needs.

Table 1. Big geodata user groups at ITC

User Group	Needs	Actions Required
All staff and students	Basic information on big geodata technology: *capabilities, limitations, applications*	Basic training on big geodata technology: overview of the landscape Periodic updates on major developments in the field
Interested students	Detailed education in big geodata technology, with a special focus on case study applications Easy access to computing resources for education and application purposes	Courses on big geodata technology with domain-specific case studies Access to computing infrastructure
Interested staff lacking knowledge	Detailed information on big geodata technology, with a special *focus on case study applications* demonstrating solutions to domain-specific problems Guidance for the use of big data technology Easy access to computing resources for learning purposes	Technology-specific training with example case study applications Hands-on practices with tools and methods Support for system (workflow) design and implementation Access to computing infrastructure
Interested staff with limited knowledge	Detailed information on big geodata technology, with a special *focus on implementation and problem solving* Guidance for the *advanced* use of big data technology Easy access to computing resources for application purposes	Tool-specific training with hands-on practices Problem-specific expert consultation and implementation support Access to computing infrastructure
Staff actively using big geodata technology	Advanced training on selected big data technologies, with a special *focus on early adoption and efficiency* Easy access to advanced computing resources	Tool-specific advanced training including on-the-edge features Access to advanced computing infrastructure including experimental technologies Technology-specific expert consultation

Training courses preferably with hands-on practices at various levels (e.g. basic, technology-specific, tool-specific), problem-specific expert consultation and technical support, and access to proper computing infrastructure are the primary needs identified for the most of the user groups.

Basic Principles

Based on the analysis of the current state and the assessment of the needs, a set of principles is identified for better use of big data technology in the three main activity pillars of ITC: research, education, and institutional strengthening. The principles are high-level objectives that define more specific goals and detailed actions required to reach these goals, which are further discussed in the roadmap and other subsequent sections.

Research

- State-of-the-art technical and scientific information on big data technology, including good practices, should be actively communicated with the research staff.
- Proficiency of the research staff on the use of big data technology should be improved.
- Easy-to-use and efficient big geodata computing infrastructure should be made available for research purposes.
- Research projects should be enhanced and improved with big data technology where relevant.
- Ad hoc technical and scientific support and advice on big data technology should be provided to the research staff.

Education

- Students should be able to identify whether a problem can be solved by traditional methods or requires big geodata technology.
- Courses on big geodata technology should be provided, including specialized technical courses.
- Existing courses should be enhanced to include big data aspects where relevant.
- Easy-to-use and efficient big geodata computing infrastructure should be made available to the students.
- Access to the computing infrastructure for experimentation, prototyping and entrepreneurship should be encouraged (i.e. access should not be limited to the needs of specific courses).

Institutional Strengthening

- Knowledge on the use of big geodata technology and related good practices should be transferred to the partner institutions.
- Good practices in cost-effective and efficient use of big data infrastructure should be shared with the partner institutions.
- Developments in big geodata technology should be communicated with alumni to support lifelong learning.

Roadmap

Big Geodata Infrastructure

A big data infrastructure is necessary to support knowledge development and research activities on big data technology. Currently neither ITC (centrally) nor the departments (individually) have such an infrastructure. Therefore, it should be developed considering the state-of-the-art technology and good practices. As a principle, reuse or repurposing of existing in-house resources, use of infrastructure accessible through research partners, and co-use of cloud resources and services should be prioritized for cost effectiveness whenever possible.

Initially, a small-sized distributed big data infrastructure (i.e. big geodata cluster) will be developed to support crucial activities such as in-house training, exploratory research activities, and big data-related courses (e.g. Big Geodata Processing). The departments will be invited to contribute to the infrastructure by providing idle or little utilized computational resources. Existing use cases of these resources will be taken into consideration and continuity of the existing services will be ensured by dual-use mechanisms (e.g. an interactive notebook server will continue to serve notebooks besides working on big data tasks). Additional resources required for the infrastructure will be procured through the CRIB budget.

The multi-purpose big geodata infrastructure will be open to all departments and it will allow *basic use and testing of the following capabilities mainly for knowledge development purposes*:

- Multi-user access to distributed computing infrastructure
- Multi-level access to computing infrastructure for remote geospatial analysis
- Distributed storage for big data processing, e.g. by using map-reduce model
- Distributed computing by using high-level programming libraries (e.g. Dask)
- Distributed computing by using high-level frameworks (e.g. Apache Spark)
- Parallel computing by using special processing units (e.g. GPUs)
- Ready-to-use big geodata analysis tools (e.g. GeoTrellis, GeoMesa)
- Analysis-ready data (ARD) for testing big geodata analysis tools and methods

According to the evolution of the user needs, the initial infrastructure will be extended to serve a wider range of activities, including regular research projects and institutional strengthening activities. *Periodic needs assessment studies will be performed to monitor the progress.* For efficient use of the resources, similar infrastructure available at UT (e.g. DSI clusters) and beyond (e.g. SURFsara, Microsoft Azure, etc.), which can be more easily used with improved knowledge and experience of the staff and students, will be taken into consideration and unnecessary physical infrastructure investment will be avoided.

It should be noted that recent developments in computer technology allow some problems that were previously considered as big data problems to be solved by using a typical workstation. A computer with 4 core 64-bit CPU, 32 GB memory, and 1 TB SSD should be considered as a standard configuration for research purposes. For problems that can be solved by such a computer, a hardware upgrade should be considered before looking for alternative workflows or computation methods. For problems that require better resources for a limited period (e.g. several days), on-demand cloud computing instances can be considered as an alternative to the in-house big data computing infrastructure. Currently, instances with 64 virtual CPUs and 256 GB memory are available at a cost of less than 3 EUR/hour. Therefore, they provide a cost-effective solution for problems that can be solved efficiently by parallel computation methods.

Although the unit costs of hardware components (e.g. storage, memory, network) are decreasing, volume of research data and related computation needs are increasing rapidly. *Therefore, an in-house computing infrastructure serving the needs of ITC is still an expensive investment.* It also ages quickly not only in terms of computing power but also in terms of technical support (e.g. limited or no support by newer software libraries) and becomes obsolete. Considering the short-term nature (i.e. limited to project lifetime) of the infrastructure needs of most of the ITC activities, easily accessible and relatively low cost cloud infrastructure solutions can be a viable alternative not only for short-terms tasks, but also for regular computation needs. A comparison of advantages and disadvantages of in-house and cloud infrastructure is given in Table 2.

Table 2. Comparison of in-house and cloud infrastructure

In-House Infrastructure	**Cloud Infrastructure**
Cost effective for serving high number of users without financial support (e.g. students)	Provides better and easy scalability
Allows better data protection	No investment and maintenance costs
Allows in-depth knowledge acquisition on operation of big geodata infrastructure	Pay only what you use
Allows better control and fine-tuning for research and development purposes	Allows access to better infrastructure (e.g. TPU)
High investment cost	Allows further research and development on infrastructure that is difficult to access (e.g. HPC)
High maintenance cost	Facilitates individual and institution capacity building in longer term (i.e. beyond training)
Becomes obsolete quickly	*Limited data privacy*
	Restrictions may apply if provided for free

In-house infrastructure helps prototyping and experimentation by providing limited performance but continuous service at a low cost. This reduces the need to monitor the usage of the paid services (e.g. cloud credits), which creates stress on the development time that is usually idle in terms of computation, especially during the learning period. On the other hand, remote infrastructure helps analysis by providing high-performance computing resources, which are *especially useful at the later stages of the research*. Overall, efficient co-use of in-house and cloud infrastructure is necessary for cost effectiveness, better service availability, higher performance, and improved data protection. *Therefore, ITC big geodata infrastructure will also include components in the cloud.* To keep costs at a minimum, institutional collaboration agreements or research grants will be discussed with the service providers, which can be used for specialised training activities or distributed to research projects of different departments. *A part of the CRIB budget will be allocated for exploratory or experimental use of the (emerging) cloud services related to big geodata technology.*

It should be noted that know-how on the efficient use of hybrid infrastructure is valuable by itself, particularly in the current period when spatial computing is in early stages of transition from local to remote analysis. *This knowledge can also be shared with partner institutions, especially in the Global South where resources are usually scarce and limited.*

Software components of current big geodata systems are mostly open source. Therefore, there are usually no costs involved for the deployment and use of software. *But there is a high turn-over of the technology which requires continuous learning and maintenance.* Hence a stronger in-situ user support is necessary.

Various access mechanisms to the computing infrastructure will be provided to serve a wider range of users and better meet their needs. *The overall aim will be to have access mechanisms which are independent of the type of physical infrastructure*, i.e. the end-users do not need to be aware of the details of the underlying server infrastructure and will not know if they are served by an in-house or cloud infrastructure unless they select one explicitly.

A three-level access, which is a common approach for research and educational purposes, will be implemented for the big geodata infrastructure and will provide interactive notebooks for easy data visualization and analysis, remote desktop access for advanced analysis by using specific software packages, and shell access for batch processing or other complex use cases (Figure 1).

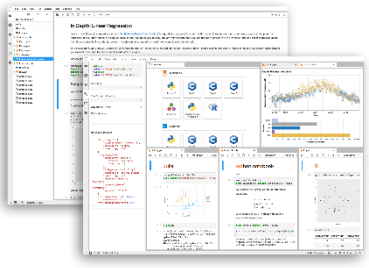
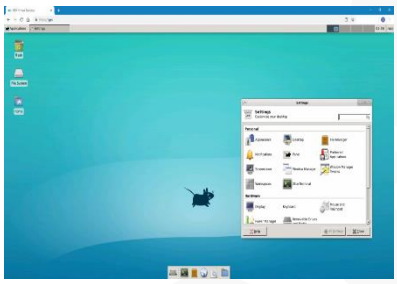
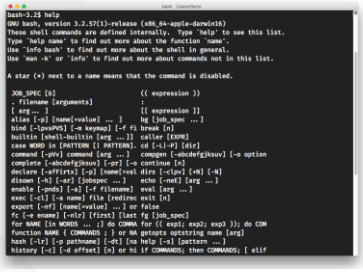
Level 1	Level 2	Level 3
		
Interactive data visualization and analysis through web-based interfaces (e.g. notebooks)	Remote desktop access for advanced analysis needs requiring specific software packages	Shell access for batch processing or other complex use cases
Computing Infrastructure		

Figure 1. Three-level access to big geodata infrastructure

Independent of the type of access, the infrastructure will provide easy and efficient access to geospatial data for analysis purposes. A data cube accessible through APIs for common programming languages (e.g. Python, R) will be part of the infrastructure. The data cube will support user authorization, i.e. private assets will be accessible to a selective user or user group. Besides locally stored private or public data, publicly available data hosted in the cloud (e.g. Amazon S3) will also be supported and indexing and caching services will be implemented to ensure a good performance.

In addition to the common tools and libraries available for big geodata analysis, including machine learning and AI frameworks, the most recent versions of the software applications, extensions, and libraries developed by ITC (e.g. ILWIS) will be available through the infrastructure. Software components of the infrastructure will be offered for custom installation, e.g. for projects and partner organizations. This will allow easy and quick set-up of private infrastructure if hardware resources are available either locally or in the cloud. A repository of automated installation tools (e.g. scripts) and containers will be maintained and up-to-date how-to documents will be provided for service needs and custom installations.

Given their broad use in various departments, both R and Python will be first class languages of the computing infrastructure and periodic updates will be ensured, including common numerical, statistical, and spatial analysis packages. Other languages will also be supported depending on the user needs.

It should be noted that the analysis results indicate the need for an institutional geospatial computing infrastructure which is not limited to big data. The need is evident not only for research purposes, but also for educational activities. Given that several departments are already investing or finalizing investment plants for additional computational resources, it is not realistic to think that a faculty-wide common infrastructure can be achieved soon. Clearly, there were reasons in the past for the departments to choose to have their own dedicated resources and some of the reasons may still be valid. However, given the similarities in the existing and future needs, difficulties in managing and maintaining the computational resources, and the prevalent shift to remote data analysis, there are also clear benefits in having a common infrastructure. If fair usage is guaranteed, such a system will not only ensure efficient use of the resources but will also support other horizontal initiatives that require an infrastructure in place to work effectively. One example is the ITC Open Science initiative, which requires technical components such as code repository, data repository, and collaborative development environment (e.g. notebooks) to support an Open Knowledge Hub.

Development of a common general-purpose computing infrastructure is beyond the scope of CRIB, but still worth to be discussed. *The following steps can be suggested for a stage-wise transition to a common geospatial computing infrastructure*:

- *Supporting management and maintenance of the departmental computational resources:* This will reduce the administrative burden on the departments. Standard management practices will allow similar end-user experience between the departments and facilitate future integration.
- *Supporting formation of computing clusters:* Computational resources, which are currently available individually, can be arranged in a cluster formation. Besides allowing more efficient use of the resources (e.g. serving a greater number of users), this will also provide an infrastructure supporting distributed computing.
- *Facilitating sharing of idle resources:* Good management practices together with standard end-user experience may facilitate inter-departmental sharing of the resources. This can be supported by good practices in usage analytics and resource allocation / scheduling, so that departmental priorities are respected. In practice, this will result in a virtual common infrastructure.
- *Centralization of the resources:* Once common management practices and shared end-user experience become the norm, computation resources can be located at a central facility (e.g. LISA data centre) resulting in a physical common infrastructure.
- *Improving the common infrastructure:* Centralised infrastructure can be improved with common investment plans supported by the faculty and departmental budgets.

Independent from the overall aim of developing a common infrastructure, the initial steps are still useful to have a common end-user experience at ITC. Given the similarities with the work required for the development of the big geodata infrastructure, some of these steps can be supported by CRIB.

Knowledge Development

ITC is a world-renowned organisation for achievements in education, research, and capacity development in the field of geo-information science and earth observation. *To strengthen its recognition also in the big geodata domain, not only the staff members who excel in big data-related topics should be further supported, but also interested staff should be given the opportunity to learn and practice big data technology, and the overall knowledge level of all staff and students should be improved.*

Familiarity with the technology with a special focus on the capabilities and limitations of the available tools and case-studies in various ITC domains (e.g. natural resources, urban planning, etc.) should be the first objective. Short special lecture(s) as part of the core (common) course will be used to increase the awareness of the students early in their education programme. Similar short overview presentations, preferably at departmental level, will be provided to the research and support staff to refresh their knowledge on rapidly evolving big geodata landscape.

To improve technology-specific practical know-how, *hands-on training courses will be organized for selected big geodata tools and services, which have the potential to be used by many people from different departments.* Some of these trainings can be provided by the core CRIB team or expert ITC staff, whereas for some trainings external professional services might be preferential or necessary. *A training budget will be allocated for this purpose.*

Trainings will be followed by private consultation sessions to discuss application-oriented questions or challenges of the staff and students. Basic support for analysis workflow design and advice on implementation can be part of these sessions. The use cases will be monitored and if necessary, consultation meetings will be continued on a regular basis depending on the needs and available resources. Tool-specific detailed training and technical expert consultation can also be possible to support experienced staff in long term.

It is challenging to identify whether a problem requires big data technology or not and to assess the proper and efficient approaches for the solution. *To improve required know-how, a stage-wise approach is suggested* which involves training activities on the following topics ranging from better use of traditional methods to use of computer clusters for distributed computing:

1. How to use existing methods and tools more efficiently to avoid distributed computing (i.e. avoid distributed computing if you can) (e.g. data engineering, proper sampling strategy, indexing, workflow optimization)
2. How to use parallel computing on a single workstation (i.e. parallelism with multi-cores) (e.g. multi-threading, multi-processing, data serialization)
3. How to use parallel computing with specialized processing units (i.e. GPU/TPUs)
4. How to use out-of-core computing on multiple computers (i.e. computing clusters) (e.g. distributed data management, distributed computing)
5. How to use large-scale cloud computing (i.e. cloud infrastructure, scaling, cost effectiveness, system customization)
6. How to move from prototyping to production (i.e. system optimization)

Some of these topics can also be offered as part of the existing courses or as new specialized courses in the MSc and PhD programmes.

Big geodata technology not only requires technical and scientific know-how, but also to a large extent *demands a transition in modus operandi from local desktop-based computing to remote cloud-based computing.* A recent user requirements survey on cloud-based systems for big geospatial computing has the following conclusions that are mostly also valid for ITC (Wagemann et al., 2020):

- Despite the high interest in using cloud-based services, many users face technical hurdles in using them.
- Users are not familiar with working with cloud-based systems. A shift will require a change in mindset and time.
- Combining different types of data is one of the most important tasks users do. Current cloud solutions do not facilitate this need and face the problem of data interoperability.

Given the difficulties known to the end-users, attention will be paid to facilitate this transition by providing initial support and help for the use of provided services. It should be noted that this is rather a generic requirement not limited to big geodata; therefore, can also be supported by other initiatives and teams at UT/ITC.

The current trend with major big data and machine learning frameworks is to support multiple languages for the end-user needs. *Support for each language varies from framework to framework, and not all languages are first-class citizens, resulting in significant underperformance.* This can still be acceptable for some use cases, but limitations and performance penalties should be known at the beginning so that decisions, including not using a specific tool or method, can be taken early. Once a framework is chosen, switching to another one later is usually non-trivial and time consuming. The initial training will focus on the overall design principles, capabilities, and limitations of the frameworks, aiming to provide basic knowledge without going into details of language-specific implementation. Such details will be considered in the follow-up training for selected user groups when there is enough interest, or they will be discussed in problem-specific consultation meetings.

For improving institutional know-how in a consistent manner, it is crucial to develop a community which is self-motivated to learn and practice more, and share knowledge, experience, and good practices. Besides establishing an encouraging environment, it is also essential to assist the community by providing information on latest developments in the field, including recent big data resources, analysis platforms, and case studies. Some of the activities identified for this purpose are listed below. *In addition facilitating knowledge development and sharing, these activities will also help to increase the visibility of CRIB among ITC staff and students.*

- Big Geodata Newsletter: A non-technical, periodic newsletter on recent developments in the big geodata landscape, including case-study applications (see the first edition here).
- Big Geodata Talks: A technical series of talks on big geodata technology, especially tools and platforms. Talks will be organized in the following main categories:
    - *Big geodata centres*: First-hand experience from initiatives similar to CRIB with a focus on structure, provided services, lessons learned, and user stories
    - *Big geodata case-studies*: Case-studies from research groups showing good and state-of-the-art practices of big geodata technology
    - *Big geodata tools*: First-hand information from the developers of the tools and frameworks with a focus on existing and planned features, and what kind of problems they help to solve
    - *Big geodata platforms*: First-hand information from the operators of big geodata platforms with a focus on existing and planned capabilities, and user stories
- Big Geodata Meetings: Periodic meetings of the staff and students interested in big geodata technology to discuss practical problems that they experience. Aims to facilitate knowledge transfer and sharing of
- good practices.

Some of these activities, especially the newsletter and talks, can reach a wider audience beyond ITC and can help to increase the visibility of UT/ITC in the big geodata domain. The talks will also facilitate networking with prominent big data initiatives and organisations, which can *facilitate early adoption of the technology by providing access to new features or capabilities* before they become available to the ordinary users.

Goals and Actions

The principles, basic needs and possible solutions identified in the previous sections are used to define specific goals and actions, which are grouped under the following categories:

- Infrastructure Development
- Knowledge Development
- Project Services
- Monitoring
- Visibility
- Special Tasks

Please note that the goals and actions naturally overlap with the content discussed in the previous sections, in particular the roadmap section, as they aim to be concise and quantifiable statements of the proposed activities. *Currently, they are rather high level and do not provide a clear timeline except expressing the intention to reach selected goals sooner than the others.*

After consulting the stakeholders and discussing the findings of this report, precise formulations of goals and actions will be provided together with an implementation plan. *The plan will include a well-defined schedule, quantitative performance measures, necessary resources, and corresponding budget requirements.*

Infrastructure Development

Better use of big data technology at ITC requires an infrastructure, which is not limited to research and education needs, but also available for personal capacity development activities. It should fit the purpose, be efficient, easily accessible, user friendly, and cost effective. The infrastructure should utilize already existing assets and employ a hybrid approach benefitting both from local and remote resources. See Big Geodata Infrastructure section for more details.

Short-term Goals

1. Development of a common big geodata infrastructure
2. Improved utilization of the computing resources available at ITC departments
3. Improved utilization of the infrastructure provided by partner UT departments and institutes
4. Better use of cloud infrastructure and services

Long-term Goals

1. Better use of external large-scale infrastructure (e.g. national, European, international)
2. Development of analysis ready data and related services

Actions

1. Identify infrastructure needs based on existing, planned, and potential education, research, and institutional strengthening activities.
2. Create and maintain an inventory of available and planned computing infrastructure at different organizational levels (i.e. ITC, UT, national, European, international, commercial cloud) that can be utilized for big geodata needs.
3. Develop and maintain a big geodata infrastructure to enable critical activities, such as training, courses, and exploratory research.
4. Implement easy-to-use, efficient, and standard access mechanisms to the infrastructure allowing hybrid use of local and remote resources.

5. Keep ITC staff informed about available infrastructure and related developments.
6. Facilitate sharing of available resources among ITC departments.
7. Facilitate centralization of available resources at ITC departments, if feasible.
8. Collaborate with other infrastructure owners at various levels (e.g. UT, national, European, international) to make additional resources available at minimum or zero cost.
9. Facilitate two-way communication and cooperation between the staff and infrastructure providers, especially for technical support.
10. *(Depending on the interest)* Upgrade existing file repositories into data cubes, which can be easily accessed, queries, and analysed through API (e.g. Python, R) or Desktop GIS.
11. *(Depending on the interest)* Produce Analysis Ready Data sets for selected project areas based on publicly available EO data (e.g. Landsat, Sentinel, etc.)

Knowledge Development

Better and wide-spread use of big geodata technology at ITC requires improving the existing expert knowledge, providing theoretical and hands-on training to the interested staff and students, and rising the overall familiarity with the technology. See Knowledge Development section for more details.

Short-term Goals

1. Well-informed staff and students on advances in big geodata technology and infrastructure.
2. Up-to-date catalogue of staff know-how and interest on big geodata technology
3. Better knowledge sharing among the staff on big geodata technology and applications.

Long-term Goal

1. Improved knowledge and hands-on experience of staff and students in using big geodata technology

Actions

1. Actively communicate with the staff and students to inform about recent developments in big (geo)data domain (e.g. Big Geodata Newsletter).
2. Provide basic overview of the big geodata technology to the staff and students
3. Organize crash training courses on:
    a. Cloud services
    b. Parallel computing frameworks
    c. GPU computing
    d. Cluster computing frameworks
    e. Machine learning frameworks
4. Organize invited talks domain-specific case-study applications of the big geodata technology.
5. Invite managers of other big geodata centres and organisations to present their activities, including user stories.
6. Invite developers of big geodata tools to present their tools, including future development plans and user stories.
7. Facilitate periodic inter-departmental big data technology meetings to discuss problems, practical solutions, and recent development (like DSI AI meetings).
8. Organize public workshops, hackathons, and similar events on big geodata technology.

Project Services

Research projects should include the big data technology as needed, from the proposal stage onwards, budget appropriately, and ensure that the infrastructure available will be optimally utilized. Providing consultancy and advisory services for the integration and better use of the technology is therefore crucial. To set the potential in motion, it is important to demonstrate possible support by CRIB as early as possible. For this purpose, case-study projects from each department can be selected in close cooperation with interested staff, which can be enhanced or scaled-up by using big geodata technology.

Short-term Goals

1. Improved project proposals that leverage big geodata technology and infrastructure
2. Better implementation and utilization of big geodata technology in the projects
3. Better planning of future activities considering big geodata aspects.

Long-term Goals

1. Periodically reviewed needs assessment on big geodata technology
2. Early adoption of emerging big (geo)data technology.

Actions

1. Support project proposals by providing technical and scientific guidance and advice on big geodata technology.
2. Facilitate early adoption of emerging big (geo)data technology by providing technical and scientific support.
3. Provide ad hoc technical and scientific support for better implementation and integration of big data technology.
4. Support scientific projects by active participation in big geodata-related tasks
5. Provide guidance for the planning of future activities in the context of big geodata.
6. Carry out surveys to identify needs and follow progress in time.

Monitoring

Big geodata domain is evolving rapidly. To keep up with the progress, it is important to closely monitor the developments, especially related to data resources, analysis methods, tools, and platforms. Following closely the data providers, service providers, research institutions, donor organisations, and international bodies is beneficial.

Short-term Goal

1. Up-to-date information on major components and actors of the big data technology

Actions

1. Monitor developments in big (geo)data technology.
2. Monitor big data and computing infrastructures.
3. Monitor big geodata sources with a special emphasis on future availability and early access.
4. Monitor grants offered for big geodata projects and applications, especially by service providers and support funds
5. Monitor the activities of other big geodata centres and organisations

Visibility

To establish ITC as a well-known player in the big geodata domain, it is important to ensure high visibility of the related activities. Other than the personal communication channels and networks of the staff, there are also institutional mechanisms in place to increase the visibility of the ITC activities, such as the services provided by the UT/ITC communication teams. There is also the new ITC Open Science initiative, which will provide a dedicated support for sharing research data and deliverables. *The activities listed under this category mainly aim to support these services by ensuring effective communication and collaboration between the related parties.* They will also provide further visibility through additional channels linked to big (geo)data communities.

It should be noted that initially there is also a need to ensure internal visibility of CRIB at ITC so that its activities are welcomed and supported by the staff, students, and alumni.

Short-term Goals

1. Recognition of CRIB at ITC as a competence centre facilitating the use of big data technology to enhance and scale-up education, research, and institutional strengthening activities.
2. Recognition of ITC at UT as an expert institution on big geodata technology.

Long-term Goals

1. Recognition of ITC at UT as a role model for good practices in big data technology.
2. Recognition of UT/ITC in the Netherlands as an expert institution on big geodata.
3. Recognition of UT/ITC as a global player in big geodata domain.

Actions

1. Build a web portal.
2. Activate social media (e.g. Twitter, LinkedIn) accounts to faciliate communication.
3. Share and promote big geodata related activities at ITC.
4. Share and promote big datasets and other big data products (e.g. DNN) produced by ITC and partner organizations in cooperation with the ITC Open Science and Data Policy teams.
5. Share big data-related software and publications in cooperation with the ITC Open Science team.
6. Ensure good communication and cooperation with other faculties (e.g. ET, EEMCS, TNW, BMS) and programmes (e.g. Computer Vision and Biometrics, Data Science & Business, Data Science & Technology, Mathematics of Data Science) related to big data technology at UT.
7. Ensure good communication and cooperation with the Dutch big (geo)data initiatives and organizations (e.g. eScience).
8. Ensure good communication and cooperation with international big (geo)data initiatives and organizations (e.g. JEODPP).

Special Task: Reference Big Geodata Curriculum

In 2019, ITC has been asked by GEO to develop a reference curriculum on big geodata technology. The primary motivation of the request was the Digital Earth Africa, which is a community activity under GEO aiming to provide an operational big earth observation data service to address challenges of Africa. The longstanding relationship between GEO and ITC, ITC's renowned expertise in capacity development, and its partner and alumni network in Africa stimulated a partnership for this purpose.

One of the special tasks of CRIB will be the development of this curriculum for graduate and professional education, which may also serve a broader audience, including but not limited to partner academic institutes world-wide.

Short-term Goals

- Development of a modular and scalable reference curriculum on big geodata for professional and graduate education, which is in line with existing reference geospatial curriculum initiatives.

Long-term Goals

- Promoting the reference big geodata curriculum for implementation purposes and supporting related initiatives.

Actions

- Determine knowledge and skills needed for state-of-the-art big geodata analysis.
- Collect and maintain information on online and in-class big (geo)data courses and training provided by academic and educational organizations.
- Network with existing geospatial curriculum initiatives (e.g. GIS&T BoK, EO4GEO) to collaborate on big geodata topics.
- In collaboration with ITC colleagues, prepare and maintain big geodata reference curriculum for professional and graduate education.
- Promote big geodata reference curriculum through various communication channels.
- Support the implementation of the curriculum at academic and educational institutions, especially in the Global South.

CRIB Activities

The Centre of Expertise in Big Geodata Science (CRIB) has been initiated in March 2020. Initially the core team will be composed of two dedicated staff who are expert in big geodata technology. Dr. Serkan Girgin has been recruited as Assistant Professor and leads the activities. Recruitment for the second position (Expertise Development Coordinator) is currently on-going. As part of the governance structure a Steering Committee consisting of ITC's senior scientific staff with one external member and a User Sound Board with members from scientific departments who are expert in geodata science are established. The current members of the Steering Committee and the User Sound Board are listed in Annex III.

The core team will manage, coordinate, and perform the majority of the tasks required for the actions identified in the preceding sections. The following actions are also identified for the effective and efficient operation of CRIB.

Actions

- Put the proposed governance structure into practice.
- Organise quarterly consultation meetings with the User Sound Board.
- Organise biannual meetings with the Steering Committee
- Prepare an annual activity report.
- Prepare an annual budget.

Needs

To be able to perform designated tasks more effectively and efficiently, the core CRIB team also has some needs, including means of professional development during their career. The following activities are identified for this purpose:

- Participation in specialized workshops and training courses on big (geo)data technology to improve and extend existing knowledge.
- Participation (with or without presentation) to conferences and meetings on big (geo)data technology to share experience, keep informed on recent developments, and improve networking.
- Co-authoring of (peer-reviewed) technical and scientific articles on big geodata technology and related applications.
- Acting as supervisor of MSc studies on big geodata technology.
- Acting as formal co-advisor of PhD studies that involves big geodata technology.

Remarks and Further Work

The analysis part of this document is mainly based on information collected through teleconference meetings in a period during which access to ITC was not possible. Hence, communication opportunities were limited. *Reaching a larger number of stakeholders (e.g. academic staff, students) through normal discussion meetings can improve the completeness and provide a better assessment.*

Although sent to all PhD students (n = 114), participation to the big data surveys were limited (n = 21) probably due to unusual conditions (i.e. COVID-19) and low level of recognition of CRIB among the students. It is a reasonable assumption that mostly the students interested in the topic participated in the survey; hence, the results can be more inclined to be positive (i.e. more familiarity with the technology) rather than negative. The staff survey was limited to the participants of the ITC Big Data and AI workshop (n = 16). The participants cover a significant portion of the staff already experienced with or interested in this technology and thus with a higher knowledge level compared to the rest. *Still it can be useful to reach all staff members for more representative results.* Recently all ITC staff and students, including MSc students, were asked to participate in the survey through an announcement in the first edition of the Big Geodata Newsletter sent on May 27th, 2020. Results are pending.

Review of big geodata initiatives of other national or international academic and research organizations can provide additional information, especially in terms of implementation details, good practices, solutions to common problems, and lessons learned. Currently this is not included in the analysis. But a survey will be performed before the implementation of the actions identified to benefit from the existing experience.

Although the actions identified for the better use of big data technology are grouped as short-term and long-term actions, *a definite timeline of the actions is not provided.* This was a deliberate choice and a timeline will be prepared based on the discussion of the findings and proposed actions with the CRIB User Sound Board and other stakeholders. It should be noted that some actions (e.g. training) may require normalisation of the extraordinary conditions due to the COVID-19 pandemic so that they can be performed effectively. Therefore, a delay is expected and will be taken into account. We will ensure that this period is efficiently used by other activities.

For the time being, no specific needs for institutional strengthening activities are identified. However, in the future needs may arise in this pillar as well. Due to their highly variable nature, the possible inclusion of big geodata-related activities in institutional strengthening activities needs to be carefully weighed based on utility, institutional absorption capacity, and systemic potential to sustain its use. Further in-depth analysis is required for this purpose.

Annex I

The courses available at ITC which are directly linked, related, or relevant (i.e. covers similar topics) to big data technology are listed in this annex.

Courses

- Big Geodata Processing:

  Introduction to big geodata; big geodata management; big geodata modelling and analysis; solution setup; types of solutions; available solutions; building scalable workflows; versioning; scientific reproductivity.
  Zurita Milla, R. (Raul)

Related Courses

- Scientific Geocomputing:

  Fundamental mathematics; algorithmics; literate programming; data types; input-output; control flow; file operations; array-matrix algebra; functions; spatial data types; spatial data-handling libraries; review of programming languages for geospatial applications.
  By, R. A. de (Rolf)

- Programming Solutions:

  Object-oriented programming in Python; scientific computing; building graphical user-interfaces; add-on mathematical, scientific, and engineering packages; case study on a topic related to geoinformatics.
  Bakker, W. H. (Wim)

- Advanced Image Analysis:

  Support vector machines; random forest; deep learning with convolutional neural networks; markov random fields.
  Tolpekin, V. A. (Valentyn) [Left]

Relevant Courses

- Spatio-temporal Analytics and Modelling:

  … Data mining and ML methods, including an introduction to cloud computing …
  Augustijn, P. W. M.

- Earth Observation for Wetland Monitoring and Management:

  … Cloud processing; practical use of time series analysis and big data processing …
  Vekerdy, Z. (Zoltan)

- Spatio-temporal Analysis of Remote Sensing Data for Food and Water Security:

  … Platforms (cloud-based) and tools/algorithms to (pre-)process space-time cube data …
  Bie, C. A. J. M. de

- Unmanned Aerial Vehicles for Scene Understanding

  … State-of-the-art deep learning algorithms …
  Yang, Y

- Extraction, Analysis and Dissemination of Geospatial Information

  … Principles of web architectures and web services requirements for web/cloud applications …
  Gevaert, C. M. (Caroline)

- Integrated Geospatial Workflows

  … Infrastructural system design: deployment (e.g. cloud services) …
  Lemmens, R. L. G. (Rob)

- Statistics for Spatial and Spatio-temporal Data

  … Applied spatial data analysis with R …
  Osei, F. B. (Frank)

- Water, Carbon and Ecosystem Dynamics

  … Geospatial programming of satellite data using OS software (GDAL, Ilwis, QGIS, R) …
  Mannaerts, C. M. M. (Chris)

- Cadastral Data Acquisition Technologies and Dissemination Methods

  … Web architectures, web services, open systems …
  Koeva, M. N. (Mila)

- Mapping and Monitoring for Natural Resources Management

  … Mapping in agroecosystems using decision-tree classification and machine learning …
  Marshall, M. T. (Michael)

- From Data to Geo-information for Natural Resources Management

  … Searching data and storing data and data products in the cloud. …
  Leeuwen-de Leeuw, L. M. van (Louise)

Annex II

The results of the big geodata user surveys are summarised in this annex.

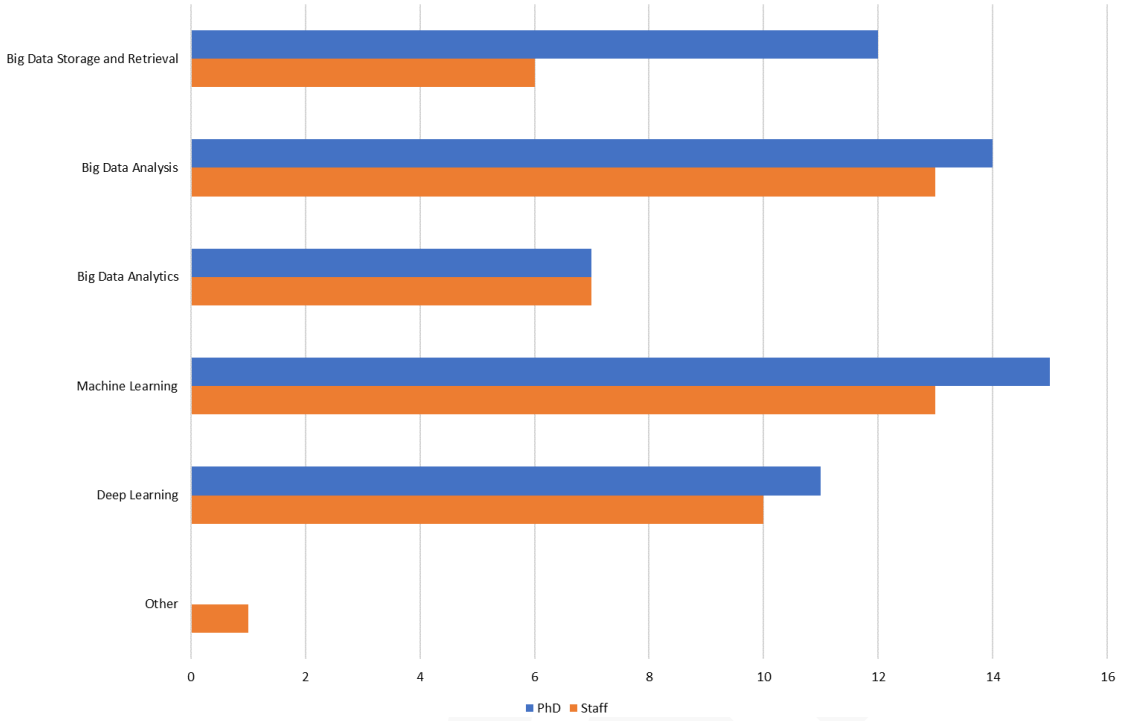


Figure A2.1. Big data topics the participants interested in.

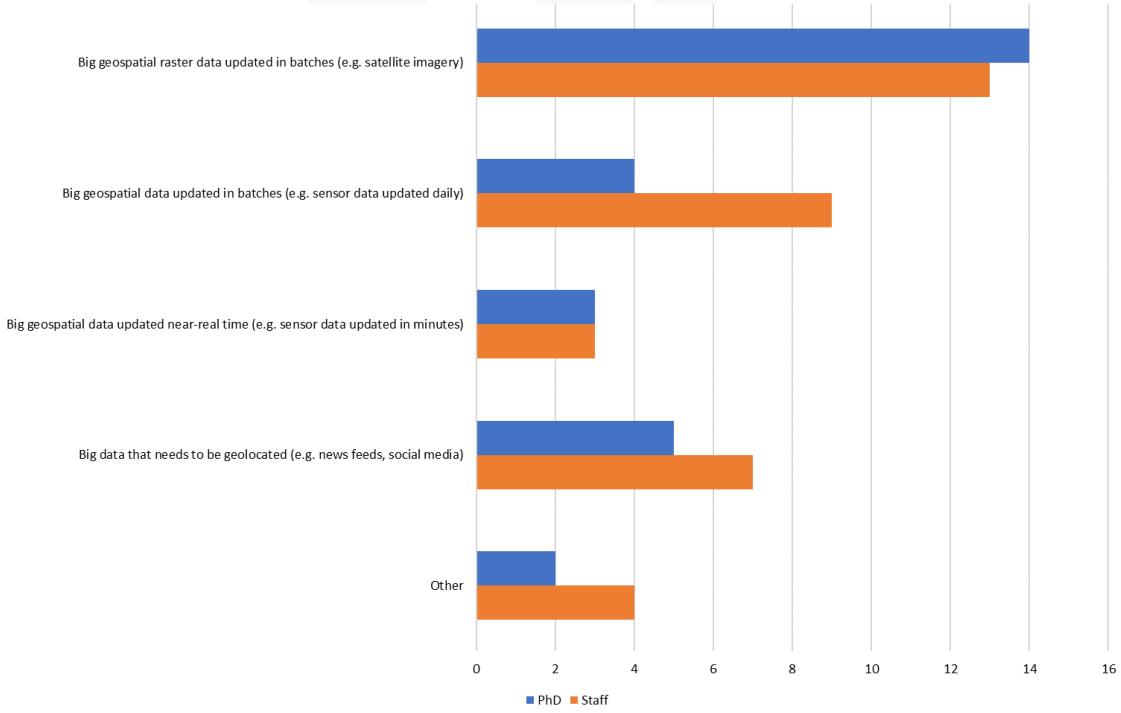


Figure A2.2. Big data types the participants working with or interested in.

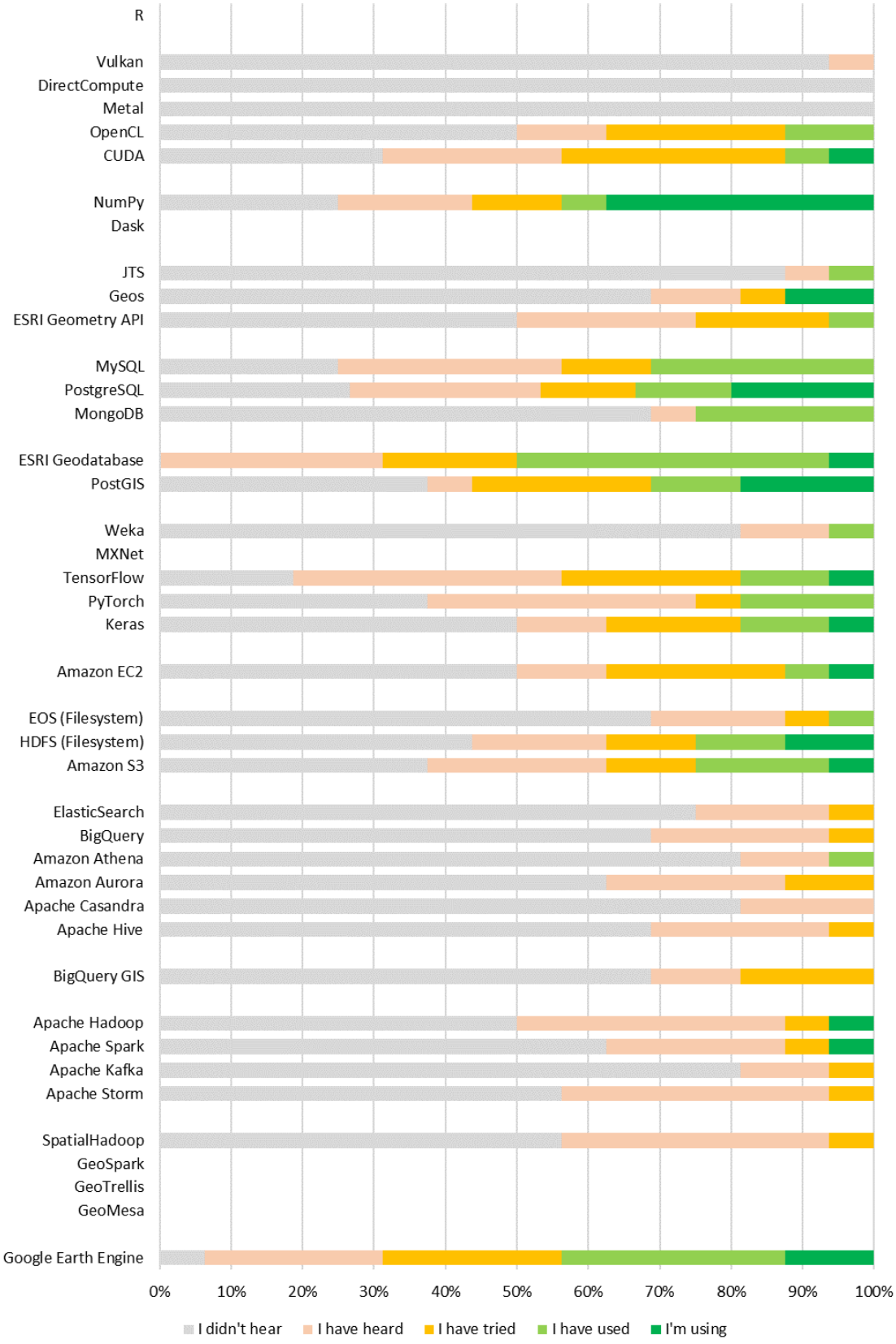


Figure A2.3. Knowledge on selected big data-related tools (Staff, n = 16)

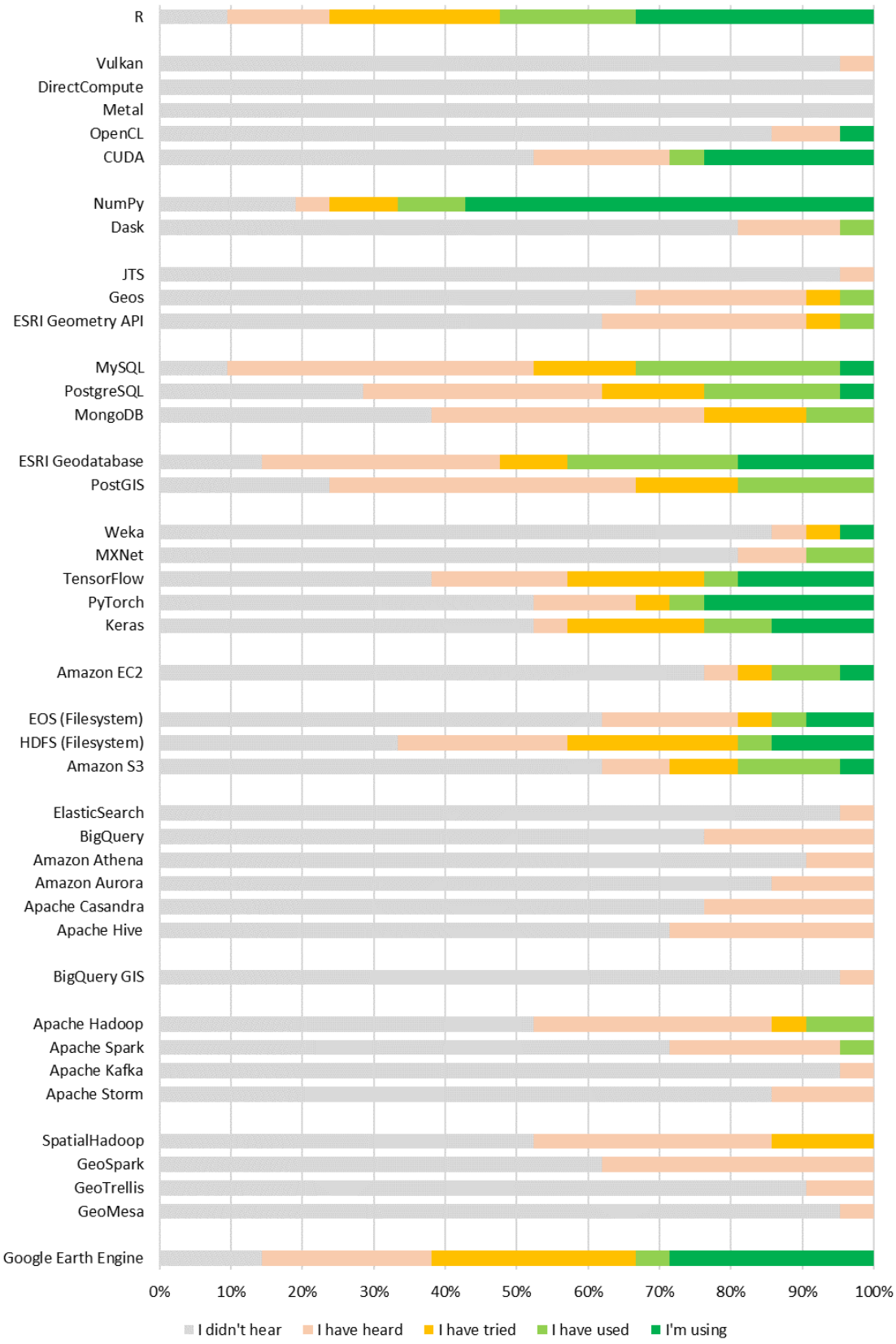


Figure A2.4. Knowledge on selected big data-related tools (PhD students, n = 21)

Annex III

The current structure of the Centre of Expertise in Big Geodata Science (CRIB) is given in this annex.

Core Team

Members

- Serkan Girgin (ITC FB)

User Sound Board

Tasks*

- Communication channel into departments and their big data challenges
- Discuss needs and wishes and aim to schedule developments in CRIB's activities

Members

- Ben Maathuis (ITC WRS)
- Claudio Persello (ITC EOS)
- Harald M. A. van der Werff (ITC ESA)
- Michael T. Marshall (ITC NRS)
- Nina Schwarz (ITC PGM)
- Raul Zurita-Milla (ITC GIP)

Steering Group

Tasks*

- Review proposed CRIB strategy and reflect on CRIB's functioning.
- Review CRIB annual report and ensure it is entered on the agenda of ITC's Academic Board

Members

- Caroline Gevaert (ITC EOS)
- Raymond Veldhuis (UT DSI)
- Rolf A. de By (ITC GIP)

* According to the "Formation and Organizational Embedding of Expertise Team Big Geodata" document.