Accepted Manuscript

Geological Society, London, Special Publications

A New International Initiative for Facilitating Data-Driven Earth Science Transformation

Qiuming Cheng & Molei Zhao

DOI: https://doi.org/10.1144/SP499-2019-158

Received 22 September 2019
Revised 2 February 2020
Accepted 3 February 2020

A New International Initiative for Facilitating Data-Driven Earth Science Transformation

Qiuming Cheng[1,2*] and Molei Zhao[2]

[1]State Key Lab of Geological Processes and Mineral Resources, China University of Geosciences, Beijing 100083, China

[2]School of Earth Sciences and Resources, China University of Geosciences, Beijing 100083, China

Qiuming Cheng: 0000-0003-3122-5942

** Correspondence: qiuming.cheng@iugs.org*

Abbreviated title: Big Data and Discovery for Earth Science

Highlights:

(1) Challenges of utilization of big data and data discovery in earth science.

(2) From question-driven to data-driven earth science transformation.

(3) IUGS big science program promoting international collaboration for integrative research

Abstract: Data-driven techniques including machine-learning algorithms with big data are reactivating and re-empowering research in traditional disciplines for solving new problems. For geoscientists, however, what matters is what we do with the data rather than the amount of data. Thus, how big data technologies can facilitate geoscience research is a fundamental question focusing attention from most organizations and geoscientists. A quick answer is that big data technology may fundamentally change the direction of geoscience research. In view of the challenges faced by governments and professional organizations to contribute to the transformation of earth science in the big data era, the International Union of Geological Sciences (IUGS) has initialized a new big science program – Deep-time Digital Earth (DDE). This paper elaborates on the main opportunities and benefits of utilizing data-driven approaches in geosciences and the challenges in facilitating data-driven earth science transformation. The main benefits may include transformation from human learning to integration of human learning and artificial intelligence, including machine-learning, as well as from known questions seeking unknown answers to seeking unknown questions and unknown answers. The key challenges may be associated with intelligent acquisition of massive, heterogeneous data and automated comprehensive data discovery for complex earth problem solving.

1. Introduction

Numerous scientific advances indicate that geoscience is entering a new era for real-time Earth system innovation as a result of modern space observation technology, analytical and experimental facilities, outdoor experimental infrastructure, computer simulation and data discovery technology.

Geoscientists have been continually making gratifying progress in many fields of earth sciences ranging from deep space, deep earth, deep ocean, to deep time. For example, researchers in South Africa reported that a 12,000 kilogram dinosaur named "Ledumahadi mafube" was one of the largest animals on Earth during the Mesozoic era of the planet's existence (McPhee et al. 2018). Australian scientists examined fat molecules extracted from the fossil of a mysterious creature called Dickinsonia and have confirmed that it lived 558 million years ago, making it the earliest known member of the animal kingdom (Bobrovskiy et al. 2018). British scientists reported that life, based on genetic and fossil evidence, may have begun on Earth nearly 4.5 billion years ago, much earlier than previously thought (Betts et al. 2018).

In space and planetary science, Italian scientists used radar measurements to detect a 20-km-wide lake of liquid water underneath solid ice in the Planum Australe region on Mars (Orosei et al. 2018). The presence of liquid water on Mars has long been debated and has implications for astrobiology and future human exploration. NASA has launched a new mission, "InSight", which has landed on Mars to pursue three main goals: taking the planet's temperature, measuring its size, and monitoring quakes (Voosen 2018). The new data about the thickness, size, and composition of Mars will help scientists better understand why Mars is so different from Earth. Plate tectonics has been found on Earth but not on any other planets.

In deep Earth, the ten-year Deep Carbon Observatory project (DCO) announced the discovery of considerable amounts of life forms living up to at least 4.8 km deep underground, including 2.5 km below the seabed (Schiffries et al. 2018). The high pressure form of crystalline ice, Ice VII, was identified among inclusions found in natural diamonds (Tschauner et al. 2018), presumably formed when water trapped inside the diamonds retained the high pressure of the deep mantle. This discovery might provide new insight into the deep Earth water cycle, which is a key question of Earth dynamics and plate tectonics.

In the field of critical minerals and resources, researchers in Japan reported finding centuries' worth of rare-earth metals in the deep ocean mud in the northwest Pacific (Yasukawa et al. 2016). One of the fastest-growing research fields in Earth system science is the automatic extraction and integration of information from big data in the cloud environment (Bush et al. 2017; Khozin et al. 2017; Ma et al. 2017; Gewin 2018; Savage 2018; Scheffer & Van Nes 2018; Knüsel et al. 2019; Reichstein et al. 2019).

Among the above fundamental geoscience research advances, some findings are related to questions that fill gaps in our knowledge about the Earth and its operation and other findings are directly associated with the sustainable development of our society. Some are known questions and knowledge gaps for which we look for answers in a process that can be referred as "we know what we don't know". In other cases, neither the questions nor their answers are known, i.e. "we don't know what we don't know" or "unknown unknowns", as made famous by Donald Rumsfeld in the 2000's.

Traditional research approaches are mostly related to the search for unknown answers and solutions to known questions. This type of research can be considered as problem-driven or question-driven. With the development and application of revolutionary technology, especially digital technology, earth science has just entered a new era of fast-growing innovation. The tools and technologies of the digital revolution combined synergistically with transdisciplinary and integrative efforts are boosting innovation and discovery in fundamental and applied earth science research. The digital revolution is transforming scientific research from a problem-driven paradigm into a data-driven one. In the new paradigm, through a data discovery process, scientists might be interrogated by a machine about the questions of which they might not yet be aware. From this perspective, big data discovery may boost innovation not only by finding answers to existing questions but also by revealing and defining unknown questions.

Therefore, the advantages of using big data are not only to access data volume, variety and velocity but also to implement innovative technology for enabling knowledge discovery, which may fundamentally change the methods of conducting earth science research. For example, some types of data are captured and utilized in one field but rarely known and available for uses in other fields. Big data technology enables these types of data to become accessible and interoperable to all who are interested (Halpin et al. 2010). Integrating these “new” data into the data discovery process can increase the dimensions and variety of the data, affording a better description of the complexity of the phenomena being studied and lead to the discovery of new patterns or previously unknown questions.

Using machine learning with big data is also an indispensable tool for reading and collecting data to support science, which proves to be another notable advantage. For example, Macrostrat is a platform that links the GeoDeepDive digital library with a machine reading system for rapidly aggregating geological data relevant to the spatial and temporal distribution of rocks as well as the data extracted from them. The efficiency of machine reading and extracting information from publications clearly convinces researchers that the cutting – edge technology of these approaches can not only significantly increase the big “V”s of data but also, notably, releases them from spending time and effort on data handling (Peters et al. 2018). In the remaining sections of the current paper, we will elaborate on the opportunities and challenges facilitating data-driven earth science transformation. First, we will show the advantages of big data and data-driven discovery and then provide suggestions about the challenges and solutions. Finally, we will introduce a new international initiative facilitating data-driven earth science transformation, the IUGS-led Deep time Digital Earth (DDE).

2. Data-Driven Earth Science Paradigm Transformation

2.1 From question-driven to data-driven research

From the point of view of earth system science and a habitable planet Earth, the major questions in earth science include fundamental scientific issues such as the interactions and co-evolution of multiple spheres of earth systems (Figure 1), as well as applied questions related to the occurrence and spatial-temporal distributions of natural resources (e.g., minerals, energy and water) and geohazards (e.g., earthquakes, volcanos, landslides) (Figure 2). Many global databases with either archived data accumulated in the past or monitoring data acquired from observatory networks are available for linking lithosphere, biosphere, hydrosphere and atmosphere (Figure 1). With adequate deep time scale control, these data can be georeferenced to demonstrate the variability of each relevant earth sphere through time as well as to analyze the associations and co-evolution of multiple spheres. As an example, Figure 1 shows several global datasets, from which time-series can be derived, to describe secular changes of earth biodiversity, chemical indexes of atmosphere, chemical change in seawater, and a number of mineral phases. In addition to the general secular trends of the time series, more importantly, the data show the major extreme terrestrial events that have occurred during the evolution of the Earth. These types of events include the formation of supercontinents, the mass extinction of life, iron formation, big oxidation events, glaciations and clustering occurrence of mineral deposits etc. The results plotted in Figure 1 illustrate not only the variability of the data but also complex associations of these major events.

While some studies have focused on the intersections of these events, including timing and controlling factors, new work could focus on cross-correlation and long-term interdependency (symmetrical or asymmetrical) as well as the cascading and escalating effects on each other. When time series data reach proper temporal resolution, they can be analyzed by spectral methods to reveal periodicity and internal associations. For example, five global databases of detrital zircon age data were analyzed by a new time series filtering method, Local Singularity Analysis (LSA; Cheng 2007), which revealed periodicity in the secular trends of age distributions of certain geological events (Cheng 2018). Surprisingly, the results of the LSA-filtered time series not only show the periodicities of the age peaks but also reveal systematic and simultaneous reductions of wavelength and intensity of the periodicity of peaks on the filtered time series (Figure 1c). This posed a new scientific question about why the peaks depict a linear decrease and if it can be extrapolated to predict the future of the Earth. After integrating other secular changes of Earth attributes (mantle temperature, crustal reworking rate, the ratio of collisional to collisional + accretionary orogens with time and the volume of continents; Herzberg et al. 2010; Condie et al. 2015; Spencer et al. 2014; Rino et al. 2008), two models indicate that the time required for the intensity of major magmatic activity to vanish and for the temperature of the mantle to reduce to below the mantle solidus would be 1.45 Gyr into the future (Cheng 2018).

This was first attempt to predict the future dynamics of plate tectonics from the internal properties of the Earth, based on the observed geological data. The primary questions of when plate tectonics related events such as subduction, volcanism, earthquakes and rifting happen remain to be fully elucidated (https://www.nationalgeographic.com/science/2018/08/news-happens-plate-tectonics-end-earth-mountains-volcanoes-geology/). This study shows an excellent example demonstrating “unknown and unknown” scenario.

The second example to be introduced here is related to critical earth science issues threatening human life and societal development. These problems include the utilization of natural resources (e.g., minerals, energy and water) and the reduction and prevention of risks caused by geohazards (e.g., earthquakes, volcanos, landslides). Understanding the behaviors of these types of extreme events is essential for developing theory, methods and models to predict and assess their distributions and impacts. Figure 2 shows the two major tectonic converging zones and global orogenic belts: Alpine-Himalayan Orogenic Belt (Tethys Belt) and Circum-Pacific Orogenic Belt（Pacific Belt）. These two orogenic belts control most known porphyry copper mineral deposits, earthquakes, and volcanic activity of the world. They also host big cities and almost half the population of the world. In addition, much man-made infrastructure including big dams and nuclear power plants have been built along the coastal zones of these regions. As a consequence, human vulnerability is increasingly affected by earthquakes, volcanoes, landslides, floods, hurricanes, and tsunamis. While research has been focused on studying each type of event and their direct associations, few studies have emphasized the interconnection of these types of events, especially where the events are located far apart and their apparent interdependencies are weak. For example, do earthquakes occurring in the two different tectonic plate systems of the Tethys Belt and the Circum-Pacific Belt have any association? Are there indirect and weak connections that exist between these systems that could be revealed by big data mining? These are typical ‘unknown’ questions that could be revealed by a machine, data-driven research approach. Figure 2b illustrates the frequency distributions of U-Pb ages of igneous zircons from intrusions in the Tethys and Circum-Pacific belts, respectively. The results show that the two patterns of time series depict some degree of similarity between the peaks of magmatic activity (< 100Ma). A large number of porphyry copper mineral deposits (PCDs) have been found in these two global scale orogenic belts. Two fundamental questions are whether there are differences and similarities between the PCDs in these orogenic belts, and what might drive their characteristics (Yang & Hou 2009; Richards 2013). A profound property of these PCDs is that they are clustered both spatially and temporally suggesting a link between deep processes of plate subduction and mineralization processes occurring in the crust (Cheng 2019). This was demonstrated through the integration of various geodatabases with earth time and paleogeography. Temporal and spatial clusters of PCDs can be associated with localized plate motion and the geometric properties of a subducting slab. Linked databases can reveal the elements that govern cluster distributions of PCDs, a fundamental question for modelling the formation of PCDs (Cheng 2019). Comparative investigations about the causal associations between extreme events (mineral deposits and magmatic activity etc.) that occurred in the two orogenic belts is an excellent domain for calling on the assistance of big data, machine learning, artificial intelligence, complex network analysis and visualization.

2.2 From human learning, to machine learning and artificial intelligence in geosciences

Learning from experience and available knowledge, human beings can make effective judgements about new situations and in turn gain experience from the objective observation of the failure or success of these judgement calls. Machine learning (ML) has greatly contributed to this positive feedback process through the dedicated study of using machine learning from experience to improve the performance of the system itself. In short, the main purpose of ML is to find patterns hidden in data (Bishop 2006). The process of capturing models from data is termed “learning” or “training”. From this perspective, ML could be considered as a form of artificial intelligence (Provost & Kohavi 1998). ML was used initially in computer vision and natural language processing, but it has been rapidly applied in many fields of natural science, social science and engineering. The application of ML has profoundly affected most industries in the past decade, and in particular the financial and commercial sectors. ML has been introduced and utilized in geoscience for several decades for various purposes, ranging from prediction and simulation to multivariate analysis. For example, a suite of techniques including logistic regression, neural networks, weights of evidence and fuzzy logic have been developed and applied to mineral potential mapping and resource assessments (Agterberg 1989; Bonham-Carter, 1994; Cheng et al. 1994). Other applications of ML include quantitative stratigraphic comparison (Agterberg & Gradstein 1988), the classification of soil and vegetation using hyperspectral data by neural networks (Benediktsson et al. 1989) and reservoir modeling and applied geophysics by Markov models and Bayesian inference frameworks (Gavalas et al. 1976; Godfrey et al. 1980; Karamouz & Vasiliadis 1992), just to name a few. More thorough reviews of the application of ML in geosciences can be found in several recent publications (Karpatne et al. 2017; Bergen et al. 2019; Reichstein et al. 2019). The current article is not intended to detail specific technological developments, but rather to introduce a few recent examples to demonstrate how ML is advancing the development of geoscience.

ML has been successfully applied to predict and monitor extreme events such as weather, mineral deposits, volcanos and earthquakes. Since the beginning of the 20th century, scientists have been making great efforts to establish detection systems to accurately monitor these types of events. It has been demonstrated that ML has a great potential for detecting extreme events from earth monitoring data. An example is the detection of extreme weather events from climate model simulations in the field of atmospheric sciences. Boers et al. (2019) used complex networks to reveal long-distance global-scale dependencies of extreme-rainfall events, which may potentially improve the predictability of associated natural hazards. ML has also been used for automatically detecting occurrences of earthquakes (Ruano et al. 2014; Perol et al. 2018; Wu et al. 2017), distinguishing between natural earthquakes and man-made explosions (Wu et al. 2018), automatic recognition of volcanogenic seismic events (Titos et al. 2018) and classification of volcanic ash particles (Shoji et al. 2018). Other examples include establishing a global surface water monitoring system using remote sensing data (Jia et al. 2017), using radar data to predict the presence of liquid water on Mars (Orosei et al. 2018), predicting floods and tornadoes based on remote sensing and radar data (Yu et al. 2015; Zhuang et al. 2016), causality analysis of storm paths (Ebert-Uphoff & Deng 2014) and dimensionality reduction and clustering analysis of seismic attributes (Köhler et al. 2010; Zhao et al. 2017; Qian et al. 2018).

Another potential application of ML is in earth dynamic simulation and prediction. ML can provide an efficient alternative to physical models for the improvement of accuracy and efficiency in inversion modelling of earth processes. ML methods can provide elegant and accurate approximations to complex geophysical inversions such as predicting mantle flow processes by simulation of mantle convection using temperature fields as training data (Shahnas et al. 2018). Trugman & Shearer (2018) developed a set of non-parametric ground-motion prediction equations (GMPEs) by the random forest method to associate stress reduction with peak ground acceleration in northern California. Rouet-Leduc et al. (2019) used the Support Vector Machines (SVMs) method on seismic data and GPS data to predict the instantaneous velocity of a subducting plate. It is worth mentioning that these types of technologies and physical simulations based on actual observations have achieved consistent results.

3. Challenges of Earth Science in the Digital Revolution

3.1 Integrative questions require international cooperation and technology innovation

Earth scientists are still facing many profound scientific issues that require the pace of geoscience investigations to keep up with the urgent global sciences challenges. For example, since its formation 4.6 billion years ago, the Earth has undergone several major evolutionary stages, including the origin of life and the origin of plate tectonics. Solid earth operates as a complex system with interactions between and within crust, mantle and core and through the transmission of matter and energy between these layers. These internal processes do not only control the evolution of the lithosphere and the genesis of tectonism, magmatism, and subduction processes, but they also strongly influence surface systems such as mountain building, volcanism, derived sedimentation patterns and the configuration of ocean circulation through the geometry of continent and deep ocean ridges and trough. Big data show that major geological events such as the formation of supercontinents, the mass extinction of life, iron formation, big oxidation events, glaciations and clustering occurrences of mineral deposits in the earth's surface system during geological history are highly consistent with intense periods of deep subduction, plume and large-scale magma activity (Cheng 2017). Understanding how these simultaneous internal processes control the occurrences of extreme events in the lithosphere or in the crust has been a focus of much attention for scientists through the 20th century through to the present.

Another geological challenge, which requires using big data and ML for support, is the exploration for resources in frontier regions such as in the deep ocean, deep earth, deep space, polar regions, in areas covered by transported glacial or other material or obscured by vegetation, and in shallow complex environments where systematic sampling through drilling is prohibitive. The difficulty and cost of access for direct observation and mapping are problems that have invited the development and application of remote sensing technology, geophysical survey technology, indicator mineral methods, geochemical surveys and research. New equipment, such as underwater robots, the InSight Mars Lander, and a variety of new geophysical methods, have made possible the evaluation of resources in these special regions. Therefore, data processing and interpretation are key for making new discoveries pertaining to geological problem-solving (Cheng 2012). All these problems involve the unravelling of the interaction between multiple earth systems over a wide range of spatial and temporal scales. To describe these complex geo-systems and their interactions requires multi-institutional, multi-organizational international cooperation in geoscience, and technology innovation with special considerations of several challenges that will be outlined below.

3.2 Explosive growth of earth science data

Over the past few decades, with the rapid development of geophysical methods, earth observation technology, digital technology and significant expansion of the number of sensors and related monitoring systems, there has been an enormous growth in the amount of data available to geoscientists. These data span from the molecular to the global and astronomical geospatial scales, and on the time scale from near instantaneous events such as earthquakes and river discharge rates, to the time scale of mobile belts and sedimentary basins, spanning hundreds of millions of years. Among many advanced technologies, supercomputing has become a powerful tool for supporting geoscience research for simulating complex processes such as plate tectonics, mantle convection, formation of orogens and basins, the progression of tsunamis and earthquake generation through fault rupture.

With the rapid growth of data resources, various specialized large-scale digital databases have emerged. A related trend is that the growing scientific community is realising the benefits of sharing their data and computing services, and thereby promoting distributed data and computing community infrastructure (Ludäscher et al. 2006). Although a huge amount of data such as geological survey data or data in other established big thematic databases (e.g. global earthquake databases) are available in digital and machine-readable form, a grand challenge for utilization of these types of data is harmonization of data and interoperability of databases. Joint efforts are needed to facilitate the standardisation, harmonization and integration of these diverse data, especially in distributed databases. One excellent example of these types of collaboration efforts is OneGeology (www.onegeology.org), which is an international initiative by a number of the world's geological surveys. OneGeology makes data obtained from worldwide geological data providers accessible to those who would like to see and use them. Many of these data are portrayed in traditional geological maps. They provide an excellent geological data hub, which can be extended to include more types of data from individuals and research groups who are willing to share and allow the free use of their data. Another influential work is GroundWaterML2 (GWML2) which is an international standard for online exchange of groundwater data. GWML2 aims to overcome the problem of data heterogeneity in groundwater databases and to promote multiple forms of data exchange and information integration (Brodaric et al. 2018).

Government survey data generally has good diversity but limited geographic coverage that ends at national borders. Integrating the data compiled in academic research studies and government databases can yield a huge collection of unorganized information in the form of pictures and scanned images, tables, notes, sketches, cross-sections, videos, samples, measurements scattered in documents or even in geoscientists' notebooks. Setting free these types of "volunteer" generated information (VGI) is essential for creating big data in earth science. Proper mechanisms including incentive policy and adequate computer technology such as artificial intelligent techniques (AI) need to be found to motivate and facilitate organizations and geoscientists to share their data so that many can benefit from making their data FAIR (Findable, Accessible, Interoperable, and Re-usable) (Wilkinson et al. 2016).

There are several geodatabases operated and maintained by geological surveys, other governmental agencies, scientific organizations and industry such as the International Seismological Centre (ISC - http://www.isc.ac.uk), the National Earthquake Information Centre (NEIC - https://www.usgs.gov/natural-hazards/earthquake-hazards/connect), and the Global Centroid Moment-Tensor Project (GCMT - https://www.globalcmt.org/). In addition, there are many other databases that are developed by individual scientists, teams of scientists or consortiums through scientific programs with limited duration and specific objectives. Some established databases may lose their maintenance capacity, either technically or financially. Some of these databases even become "dead" or evolve into isolated "data islands" due to the lack of proper management. There is an urgent need to review the relevance and determine if some of these databases require maintenance or attention so that they can be linked or transformed into modern databases to broaden their accessibility and usefulness. The accessibility and quality of these earth science data are key to promoting big data utilization.

3.3 Linking knowledge systems and artificial intelligence to automate data discovery

The main interest of geoscientists in the acquisition and utilization of scientific data is to elucidate natural processes using rational, hypothesis-driven problem solving. How much information and knowledge one can extract from data is therefore the primary measure of success. Human innovation in modern civilization is closely related to the accumulation of knowledge and experience, which are disseminated through publications and other forms of media for knowledge transfer. Engineering innovation often involves systematic design and automation of the flow of high complexity processes that could be far beyond the cognitive capacity of any individual human brain. The integration of the flow of human ideas and machine learning processes can be aided and automated through the application of modern artificial intelligent technology and various advanced semantic knowledge engines. These flows can be extended to combine machine learning processes, information management and infrastructural organizations. The flow models developed by the knowledge engine can be edited, modified and reused for similar problem solving which in turn refine the models through positive feedback processes. These types of flow models can be represented visually and implemented through proprietary or native programming languages (Bennett et al. 2016; Ludäscher 2016; Goble et al. 2020). In building these types of flow models, the existing processes developed for specific tasks can be taken as primary building blocks to form more complex and sophisticated processes for tackling large-scale issues (Figure 3). Techniques and availability of big data are the key elements of the knowledge engine and models for applications. To illustrate the concept of a process model, we can use the simple Model Builder concept developed in the field of geoinformatics. Models for applications in this field are referred as workflows that string together sequences of geoprocessing tools, feeding the output of one tool into another tool as input. The models can not only ensure automation of processes built into the model but also enable smarter data processing and intelligent reasoning by iteration, optimization and conditional branching of the processes and tasks involved in the sequence of the processes (Cheng et al. 2009). The significance of these types of modeling processes is that the human intelligence including ideas and experiences, in addition to data, tools and processes that are usually shared in the science domain, can be readily shared, reused and improved by artificial intelligent technology. Newer technologies are rapidly being developed to facilitate the linking of resources to automate the processes (Figure 4). For example, Wolfram Alpha is a knowledge engine integrating expert-level knowledge and algorithms for automatically answering questions, doing analysis and generating reports. GitHub provides a platform and storehouse for project management and collaboration on code development and sharing.

There is no doubt that ML has been widely used to solve various geological problems, but it needs to be emphasized that the pursuit of learning technologies may effect a far-reaching change in the nature of research in geoscience. As an intelligent system, artificial intelligence will bring new vitality to geoscience data acquisition, applied robotics, remote and in-situ sensing, information integration and human-computer interaction (Gil et al. 2018). There is still some basic work that needs to be conducted. This includes launching global universal geoscience benchmark datasets like ImageNet (Deng et al. 2009), and expanding the use and application of new data derived from

Interferometric synthetic aperture radar (InSAR), high-resolution satellites and multispectral images (Van Rees 2016) to detect changes in the landscape that reflect deep rooted solid earth processes and development of a framework that can incorporate prior knowledge of geoscience (Hoskins 2013; Karpatne et al. 2017; Gil et al. 2019). The opportunity for further progress lies in new artificial intelligent techniques for tackling complex problems involving nonlinear earth system processes.

4. IUGS Big Science Programs for Facilitating Data-Driven Earth Science Transformation.

Social and economic development and dramatic improvements in the quality-of-life of people living in the new century have increasingly imposed heavy pressure on water, energy and mineral resources, as well as increasing risk related to earthquakes, volcanic eruptions, floods, hurricanes, water contamination, air pollution, food security, clean energy, urban space utilization, and health. These issues are closely related to the UN 2030 Sustainable Development Goals (SDGs) (Nilsson et al. 2016) and they require knowledge and solutions from geoscientists and geoengineers (Figure 5). On the other hand, with the digital revolution, the paradigm of earth science research will also undergo a tremendous transformation. The critical efforts enabling the success of a data-driven approach for knowledge discovery in earth system science must include FAIR data (Wilkinson et al. 2016), knowledge systems for semantics searching, and integration of modern algorithms for computing and physical processes-based models for knowledge discovery. These ultimately need the integration of data, models, computers and people (Figure 6). The International Union of Geological Sciences (IUGS) is keenly aware of this change and has launched new initiatives for facilitating big data and data discovery for earth science innovation. Promoting big science programs for international collaboration in facilitating data-driven and knowledge-driven earth system science has been recognized as both challenging and a strategic priority for the geoscience community (IUGS Annual Report 2017, www.iugs.org). IUGS takes the lead on this new initiative in line with its primary goals of strengthening its leading role in Earth science, increasing the level of interaction among IUGS communities as well as cooperation with other organizations, and ultimately improving its service to society and community building. In the rest of this paper the authors will elaborate on the origin of the initiative, with focus on the general challenges and opportunity to boost data-driven earth sciences.

The vision of IUGS embraces the following three aspects: (1) to promote the development of earth sciences through the support of broad-based scientific studies relevant to the entire earth system; (2) to apply the results of these and other studies to preserve Earth's natural environment, to use natural resources wisely and to improve the prosperity of nations and the quality of human life; and (3) to strengthen public awareness of geological sciences and advance geological education in the broadest sense (www.iugs.org). For the past several decades, IUGS has established or jointly initialized various international science programs such as the International Geosphere-Biosphere Programme (IGBP) (Lindesay et al. 1996), the Global Sedimentary Geology Program (GSGP), the International Geoscience Program (IGCP) (www.unesco.org/new/en/natural-sciences/environment/earth-sciences/international-geoscience-programme/), the Global

Geochemical Baseline (GGB), the International Lithosphere Program (ILP) (www.scl-ilp.org), participated in the OneGeology initiative (www.onegeology.org) and initiated "Resourcing Future Generations" (RFG) (www.iugs.org). These programs have created long-term impacts on earth science research and community development. RFG is an initiative with the mission to focus the world on the challenge of sustainable resource supply and to achieve national development and poverty reduction through a sustainable resource development framework (Nickless et al. 2014).

Undoubtedly, the aforementioned programs have greatly advanced the frontiers of earth sciences, stretched the limits of our scientific abilities in many directions, and ultimately promoted cooperation in the geoscience community. However, facilitating cross-disciplinary and convergent research and bridging natural science, social science and engineering, as well as fundamental research and solution-oriented science and technology present both challenges and new opportunities for our communities. New international programs must focus on enhancing public awareness and education, promoting international collaboration, encouraging open data sharing which is essential to facilitate “open science” in the big data world, and facilitating transdisciplinary research. During the 72nd IUGS Executive meeting held in Potsdam Germany January 2018, the initiative proposed by President Cheng for setting up new IUGS-recognized Big Science Programs and Centers of Excellence was approved, aiming to promote and support big science programs focusing on the integration of several aspects of resources and meeting the following criteria (IUGS Annual Report 2018):

- Global and major issues
- International collaboration
- Transdisciplinary
- Global – Free – Open – Sharing - Service
- Collaboration with ISC, UNESCO, etc.
- Promotion in underrepresented nations

In 2018, IUGS made significant progress by developing and approving the Deep-time Digital Earth (DDE) program as the first big science program. DDE’s mission is to harmonize global earth evolution data and share global geoscience knowledge. The program is set as a 10-year initiative and will promote international collaboration to develop open digital platforms with full FAIR data, linking the various spheres of Earth’s of geological history. The primary goal of the DDE is to facilitate efficiency and effectiveness in the use of diverse digital earth science data with proper temporal and spatial referencing, taking into account paleogeology and paleogeography rather than present day geology and geography. DDE results and outreach will be published in international peer reviewed journals and at international science meetings. DDE data will conform to FAIR, linking Earth’s various spheres in geological history. To facilitate research using new technologies of data mining, machine learning, knowledge discovery and artificial intelligence, new international programs need to increase public awareness and education, international collaboration, open data sharing in the big data world and transdisciplinary research.

4.1 Integrating FAIR data to form connected data hubs

To a large extent, understanding the long-term evolution of the earth system and its controlling factors, including anthropogenic attributes, relies on the development of new methods for integrating and querying different types of observations and models of geoscience data. In this new era of data-driven earth system science, new platforms and programs are required for facilitating efficient use and deep learning of geoscience data. For example, the construction of complete geoscientific databases would require integration of surveyed maps, which may be stored in government data warehouses, and other data collected by academia, either partially published or in their personal computers. The new initiative of IUGS also aims to provide interoperability of databases operated and maintained by national Geological Survey Organisations, other government organizations, academic organizations, and industry as well as databases developed by individual scientists, teams of scientists or consortiums through scientific programs with limit duration. From these perspectives, big science programs need to address these critical issues and deliver community-based solutions. The first-tier goal is to promote the development and adoption of specifications to achieve distributed systems of connected geoscience databases with FAIR data standards.

4.2 Integrating data science and geoscience to build transdisciplinary community

The Earth is a complex system and the Earth sciences ought to be multidisciplinary (Horton 1998). Driven by major geoscientific issues, earth science has been undergoing a significant transformation from separate disciplinary to integrated earth system science. New theoretical knowledge systems and technical and methodological platforms are required to facilitate the development of transdisciplinary and convergent research. The multiple spheres of earth dynamically interact and influence each other. The fundamental goal of contemporary earth science is to study the operational mechanisms of the various spheres of the Earth, their interrelationships and the factors that control co-evolution and regulation. Such fundamental and integrative problems of earth science require multidisciplinary efforts including new theoretical knowledge systems linking all disciplines by breaking the current discipline barriers. New platforms linking multidisciplinary knowledge systems digitally would enable integration of transdisciplinary and multidimensional earth science data, and require new data discovery techniques for handling and mining big data for knowledge discovery. The transformation from a narrow focus on separate disciplines to a comprehensive and integrative focus will rely on integration of traditional earth science disciplines and modern disciplines such as geoinformatics, geomathematics and data sciences (Figure 7).

Integrating massive big data from different disciplines for problem-solving requires algorithms that enable the rapid and efficient processing of big data and that address the challenges of rapidly growing, heterogeneous, multi-source data volumes. Advanced information technologies such as cloud computing, parallel computing, supercomputing, complex networking, knowledge graphing, machine learning and artificial intelligence can provide indispensable support to earth science. This integration will ultimately facilitate the development of complex models to enhance abductive fusion of data-driven and model-driven approaches (Duerr et al. 2015; McPhillips et al. 2015), which in turn support the use of essential computing resources and the best scientific thinking to address challenges (Bergen et al. 2019).

4.3 Promoting interactions between national Geological Survey Organisations and International Scientific Associations

The mandates of national Geological Surveys Organisations (GSOs as illustrated by the papers in this volume) are to collect, monitor, and analyze scientific information to provide knowledge about natural resources and other issues to support societal development and to improve the quality of human life. Most GSOs are responsible for database construction and maintenance. Some GSOs have broad expertise with varied sources of funds to carry out multi-scale and multidisciplinary investigations in order to provide impartial scientific information to resource managers, planners, and other clients (www.usgs.gov). International scientific associations represent academic professionals including students and typically have the objective of promoting the development of science in specific disciplines.

IUGS is supported by two types of members: Adhering Members and Affiliated Members (the statutes and bylaws of the IUGS are available on-line at https://www.iugs.org/statutes-bylaws). A geoscientific organization from a country or geographic region, supported by an appropriate authority, may become a member of the Union as an Adhering Organization. Adhering Organizations constitute the contributing and voting membership of IUGS and typically include representatives from either or an overarching committee of GSOs, Geological Societies or Academy of Sciences. The Affiliated Membership generally includes international associations and societies. One of the strategic priorities of IUGS is to facilitate interactions and collaborations of all these groups to provide integrative knowledge essential for strengthening geological science for fundamental science problem-solving and for supporting the sustainable development of societies. Given the mandates of the GSOs, typically to promote the application of geoscience for the public good, and their national roles and reach to support government decision making, enable economic development, improve public safety and protect the environment, it is important for the IUGS to act as an enabler and facilitator of international collaboration, innovation, and knowledge exchange between the GSOs and the other constituent members of the IUGS.

Many successful examples can be listed to demonstrate the essential contribution of close collaboration between governmental agencies and international associations on major science programs that tackle global-scale issues. The International Lithosphere Program (LIP), which promotes the cooperation between geology, geophysics and geotechnology is one example. IUGS, the International Union of Geodesy and Geophysics (IUGG) and national members of LIP (www.scl-ilp.org) participate in this effort and facilitate the integration of imaging, monitoring and modelling for the study of the global Lithosphere. The IUGS Big Science Programs initiative aims to focus on fundamental and integrative science questions or global-scale issues closely related to the strategic objectives of GSOs. Such an initiative may include but is not limited to the understanding and assessment of global change due to natural geological processes and anthropogenic influences, global distribution and mechanisms of extreme geological events (e.g., extreme weather, volcanoes, Earthquakes, tsunamis, and floods) that threaten human life, and global assessments of water, air, mineral and energy resources. This is intended to support the establishment of integrated decision-support systems for resource utilization with environmental stewardship, disaster risk reduction, and ecological protection, systems which are in line with the general mission of GSOs (Hill et al., this volume).

The survey data collected by GSOs and other government agencies, such as space agencies (NASA, CNSA, ESA, etc), and other types of scientific data collected by academia and associations are complementary and ought to be integrated in order to describe the whole spectra of the Earth. While the survey and monitoring data, including satellite remote sensing data, can be vast in volume, the other types of scientific data collected by scientists or scientific programs and treated as “long-tail” and “small” data (such as isotope age data and fossil samples) can provide key information about the genetic properties of geological processes and events. While most survey and monitoring data are in relatively uniform and standard formats, the data scattered in academic research studies, with small portions published and available in the repositories of publishers, are less organized with variable data standards. The collective efforts of all relevant organizations to standardize the databases and to make them FAIR will be required.

5. Conclusion

In the new era of Earth system science, driven by the digital revolution, new programs and platforms are urgently needed to facilitate efficient use of geoscience data and move from traditional research approaches to a modern approach driven by digital technologies. Many transitions towards this goal reflect the new ‘big data’ paradigm for scientific research. Since different subject areas have various scientific focuses and needs, and scientists in separate regions experience different financial conditions and scientific infrastructure, IUGS is an ideal organization to bring together their expertise through big science programs and member engagement. International collaborations are required and encouraged in establishing big science programs, especially among IUGS constituent groups including adhering national committees and affiliated international associations. Over the past several years, President Cheng and his IUGS Executive Committee team have been taking all possible opportunities of attending numerous conferences and other occasions to promote new IUGS initiatives and to convince IUGS national committees, organizations and associations to develop proposals for establishing and participating in IUGS big science programs, as well as setting up centers of excellence with the primary objectives of sharing new knowledge, facilitating cutting-edge technological innovation and tackling global issues. It is believed that organizations jointly participating in big science programs as elaborated in this paper will benefit from them as long and as much as they will invest themselves and invite their collaborators to do likewise. This investment should take the form of creating enabling projects to integrate their data and disciplinary knowledge systems, and moving toward a multidisciplinary system facilitating machine learning, artificial intelligent and data discovery, facilitating construction of knowledge engines for automated discovery flow of multiple models and techniques using big data resources, exploring fundamental and integrative questions including not only the known questions for seeking answers but also unknown question for further innovation, engaging in a dynamic and international community of transdisciplinary sciences and achieving a profound international impact on promoting the transformation of the GEO-scientific research paradigm.

Acknowledgements

The senior author thanks the IUGS EC for their comments on the development of the big science program initiative. We also thank Dr. Daniel Lebel, Director General of Geological Survey of Canada, for his invitation to give the presentation of the paper at the session of RFG2018 and to write up this paper. Dr. Phill Hill is thanked for his editing and polishing the English of the manuscript. The two reviewers are thanked for their constructive comments on reviewing the earlier version of the manuscript.

Funding

The research has been jointly supported by the National Key Technology R&D Program of China (No.2016YFC0600501) and the State Key Program of National Natural Science of China (41430320).

References

Agterberg, F.P. & Gradstein, F.M. 1988. Recent developments in quantitative stratigraphy. *Earth-Science Reviews*, **25**, 1-73

Agterberg, F.P. 1989. Computer programs for mineral exploration. *Science*, **245,** 76-81.

Benediktsson, J., Swain, P. & Ersoy, O. 1989. Neural network approaches versus statistical methods in classification of multisource remote sensing data. *12th Canadian Symposium on Remote Sensing Geoscience and Remote Sensing Symposium*. IEEE, 489-492.

Bennett, K.P., Erickson, J.S., de Los Santos, H., Norris, S., Patton, E., Sheehan, J. & McGuinness, D.L. 2016. Data Analytics as Data: A Semantic Workflow Approach. *Proc. of Artificial Intelligence for Data Science Workshop at Neural Information Processing Systems (NIPS), Barcelona, Spain*.

Bergen, K.J., Johnson, P.A., de Hoop, M.V. & Beroza, G.C. 2019. Machine learning for data-driven discovery in solid Earth geoscience. *Science*, **363**.

Betts, H.C., Puttick, M.N., Clark, J.W., Williams, T.A., Donoghue, P.C. & Pisani, D. 2018. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature ecology & evolution*, **2**, 1556.

Bishop, C.M. 2006. *Pattern recognition and machine learning*. Springer.

Bobrovskiy, I., Hope, J.M., Ivantsov, A., Nettersheim, B.J., Hallmann, C. & Brocks, J.J. 2018. Ancient steroids establish the Ediacaran fossil Dickinsonia as one of the earliest animals. *Science*, **361**, 1246-1249.

Boers, N., Goswami, B., Rheinwalt, A., Bookhagen, B., Hoskins, B. & Kurths, J. 2019. Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*, **566**, 373.

Bonham-Carter, G.F. 1994. Geographic information systems for geoscientists-modeling with GIS. *Computer methods in the geoscientists*, **13**, 398.

Brodaric, B., Boisvert, E., Chery, L., Dahlhaus, P., Grellet, S., Kmoch, A., Létourneau, F., Lucido, J., Simons, B. & Wagner, B. 2018. Enabling global exchange of groundwater data: GroundWaterML2 (GWML2). *Hydrogeology Journal*, **26**, 733-741.

Bush, A., Sollmann, R., Wilting, A., Bohmann, K., Cole, B., Balzter, H., Martius, C., Zlinszky, A., Calvignac-Spencer, S. & Cobbold, C.A. 2017. Connecting Earth observation to high-throughput biodiversity data. *Nature ecology & evolution*, **1**, 0176.

Cheng, Q. 2007. Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. *Ore Geology Reviews,* **32**, 314-324.

Cheng, Q. 2012. Singularity theory and methods for mapping geochemical anomalies caused by buried sources and for predicting undiscovered mineral deposits in covered areas. *Journal of Geochemical Exploration,* **122**, 55-70.

Cheng, Q. 2017. Singularity analysis of global zircon U-Pb age series and implication of continental crust evolution. *Gondwana Research*, **51**, 51-63.

Cheng, Q. 2018. Extrapolations of secular trends in magmatic intensity and mantle cooling: Implications for future evolution of plate tectonics. *Gondwana Research*, **63**, 268-273.

Cheng, Q. 2019. Integration of Deep ‐ time Digital Data for Mapping Clusters of Porphyry Copper Mineral Deposits. *Acta Geologica Sinica-English Edition*, **93**, 8-10.

Cheng, Q., Agterberg, F.P., Ballantyne, S.B. 1994. The separation of geochemical anomalies from background by fractal methods. *Journal of Geochemical Exploration,* **51***,* 109-130.

Cheng, Q., Liu, j., Zhang, S. & Xia, Q. 2009. Application of GIS-Model Builder Technology for National Mineral Resource Assessment. *Earth Science (Journal of China University of Geosciences)*, **34**, 338-346 (in Chineses with English abstract).

Condie, K., Pisarevsky, S.A., Korenaga, J. & Gardoll, S. 2015. Is the rate of supercontinent assembly changing with time? *Precambrian Research*, **259**, 278-289.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248-255.

Duerr, R.E., McCusker, J., Parsons, M.A., Khalsa, S.S., Pulsifer, P.L., Thompson, C., Yan, R., McGuinness, D.L. & Fox, P. 2015. Formalizing the semantics of sea ice. Earth Science Informatics, 8, 51-62.

Ebert-Uphoff, I. & Deng, Y. 2014. Causal discovery from spatio-temporal data with applications to climate science. *2014 13th International Conference on Machine Learning and Applications*. IEEE, 606-613.

Gavalas, G., Shah, P. & Seinfeld, J.H. 1976. Reservoir history matching by Bayesian estimation. *Society of Petroleum Engineers Journal*, **16**, 337-350.

Gewin, V. 2018. Earth hacker. *Nature*, **560**, 273-274.

Gil, Y., Pierce, S.A., Babaie, H., Banerjee, A., Borne, K., Bust, G., Cheatham, M., Ebert-Uphoff, I., Gomes, C. & Hill, M. 2018. Intelligent systems for geosciences: an essential research agenda. *Communications of the ACM*, **62**, 76-84.

Gil, Y., Honaker, J., Gupta, S., Ma, Y., D'Orazio, V., Garijo, D., Gadewar, S., Yang, Q. & Jahanshad, N. 2019. Towards human-guided machine learning. Proceedings of the 24th International Conference on Intelligent User Interfaces. ACM, 614-624.

Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, M.R., Peters, K. & Schober, D. 2020. FAIR computational workflows. Data Intelligence, 108-121.

Godfrey, R., Muir, F. & Rocca, F. 1980. Modeling seismic impedance with Markov chains. *Geophysics*, **45**, 1351-1372.

Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L. & Thompson, H.S. 2010. When owl: sameas isn't the same: An analysis of identity in linked data. International semantic web conference. Springer, 305-320.

Herzberg, C., Condie, K. & Korenaga, J. 2010. Thermal history of the Earth and its petrological expression. *Earth and Planetary Science Letters*, **292**, 79-88.

Hill, P., Lebel, D, Hintzman, M., & Thorleifson, 2020. Introduction: the Changing Role of Geological Surveys, this volume.

Hoskins, B. 2013. The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Quartary Journal of the Royal Meteorological Soceity*. 139, 573–584.

Horton, B. 1998. Geoscientists are not just rock stars. *Nature*, **396**, 493-493.

Husson, J.M. & Peters, S.E. 2017. Atmospheric oxygenation driven by unsteady growth of the continental sedimentary reservoir. *Earth and Planetary Science Letters*, **460**, 68-75.

IUGS . 2018. Annual Report. https://iugs.org/uploads/annual%20report/IUGS_Annual_Report_for_2018-Final.pdf

Jia, X., Khandelwal, A., Nayak, G., Gerber, J., Carlson, K., West, P. & Kumar, V. 2017. Incremental dual-memory lstm in land cover prediction. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 867-876.

Karamouz, M. & Vasiliadis, H.V. 1992. Bayesian stochastic optimization of reservoir operation using uncertain forecasts. *Water Resources Research*, **28**, 1221-1232.

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N. & Kumar, V. 2017. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, **29**, 2318-2331.

Khozin, S., Kim, G. & Pazdur, R. 2017. Regulatory watch: From big data to smart data: FDA's INFORMED initiative. *Nat Rev Drug Discov*, **16**, 306.

Knüsel, B., Zumwald, M., Baumberger, C., Hadorn, G.H., Fischer, E.M., Bresch, D.N. & Knutti, R. 2019. Applying big data beyond small problems in climate research. *Nature Climate Change*, **9**, 196.

Köhler, A., Ohrnberger, M. & Scherbaum, F. 2010. Unsupervised pattern recognition in continuous seismic wavefield records using self-organizing maps. *Geophysical Journal International*, **182**, 1619-1630.

Lindesay, J., Andreae, M., Goldammer, J., Harris, G., Annegarn, H., Garstang, M., Scholes, R. & Van Wilgen, B. 1996. International geosphere‐biosphere programme/international global atmospheric chemistry SAFARI‐92 field experiment: Background and overview. *Journal of Geophysical Research: Atmospheres*, **101**, 23521-23530.

Ludäscher, B. 2016. A brief tour through provenance in scientific workflows and databases *Building Trust in Information*. Springer, 103-126.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J. & Zhao, Y. 2006. Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, **18**, 1039-1065

Ma, C., Meyers, S.R. & Sageman, B.B. 2017. Theory of chaotic orbital variations confirmed by Cretaceous geological evidence. *Nature*, **542**, 468.

McPhee, B.W., Benson, R.B., Botha-Brink, J., Bordy, E.M. & Choiniere, J.N. 2018. A giant dinosaur from the earliest Jurassic of South Africa and the transition to quadrupedality in early sauropodomorphs. *Current Biology*, **28**, 3143-3151. e3147.

McPhillips, T., Song, T., Kolisnik, T., Aulenbach, S., Belhajjame, K., Bocinsky, K., Cao, Y., Chirigati, F., Dey, S. & Freire, J. 2015. YesWorkflow: a user-oriented, language-independent tool for recovering workflow information from scripts. *arXiv preprint arXiv:1502.02403*.

Nickless, E., Bloodworth, A., Meinert, L., Giurco, D., Mohr, S. & Littleboy, A. 2014. Resourcing future generations white paper: mineral resources and future supply. *International* *union of geological sciences*.

Nilsson, M., Griggs, D. & Visbeck, M. 2016. Policy: map the interactions between Sustainable Development Goals. *Nature News*, **534**, 320.

O'Neill, C., Lenardic, A. & Condie, K.C. 2015. Earth's punctuated tectonic evolution: cause and effect. *Geological Society, London, Special Publications*, **389**, 17-40.

Orosei, R., Lauro, S.E., Pettinelli, E., Cicchetti, A., Coradini, M., Cosciotti, B., Di Paolo, F., Flamini, E., Mattei, E., Pajola, M., Soldovieri, F., Cartacci, M., Cassenti, F., Frigeri, A., Giuppi, S., Martufi, R., Masdea, A., Mitri, G., Nenna, C., Noschese, R., Restano, M. & Seu, R. 2018. Radar evidence of subglacial liquid water on Mars. *Science*, **361**, 490-493.

Peters, Shanan E., et al. 2018. Macrostrat: A Platform for Geological Data Integration and Deep-time Earth Crust Research. EarthArXiv, 27 Jan. 2018. Web.

Perol, T., Gharbi, M. & Denolle, M. 2018. Convolutional neural network for earthquake detection and location. *Science Advances*, **4**, e1700578.

Provost, F. & Kohavi, R. 1998. Glossary of terms. *Journal of Machine Learning*, **30**, 271-274.

Puetz, S.J. 2018. A relational database of global U–Pb ages. *Geoscience Frontiers*, **9**, 877-891.

Qian, F., Yin, M., Liu, X.-Y., Wang, Y.-J., Lu, C. & Hu, G.-M. 2018. Unsupervised seismic facies analysis via deep convolutional autoencoders. *Geophysics*, **83**, A39-A43.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J. & Carvalhais, N. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, **566**, 195.

Reyners, M., Eberhart-Phillips, D. & Stuart, G. 2007. The role of fluids in lower-crustal earthquakes near continental rifts. *Nature*, **446**, 1075-1078.

Rino, S., Kon, Y., Sato, W., Maruyama, S., Santosh, M. & Zhao, D. 2008. The Grenvillian and Pan-African orogens: world's largest orogenies through geologic time, and their implications on the origin of superplume. *Gondwana Research*, **14**, 51-72.

Richards, J.P. 2013. Giant ore deposits formed by optimal alignments and combinations of geological processes. *Nature geoscience*, **6**, 911-916.

Ruano, A.E., Madureira, G., Barros, O., Khosravani, H.R., Ruano, M.G. & Ferreira, P.M. 2014. Seismic detection using support vector machines. *Neurocomputing*, **135**, 273-283.

Savage, N. 2018. Big data goes green. *Nature*, **558**, S19-S19.

Scheffer, M. & Van Nes, E.H. 2018. Seeing a global web of connected systems. *Science*, **362**, 1357-1357.

Schiffries, C.M., Mangum, A., Mays, J., Hoon-Starr, M. & Hazen, R. 2018. The Deep Carbon Observatory: The Carnegie Institution for Science as a Global Center for Collaborative and Interdisciplinary Research. *AGU Fall Meeting Abstracts*.

Shahnas, M., Yuen, D. & Pysklywec, R. 2018. Inverse Problems in Geodynamics Using Machine Learning Algorithms. *Journal of Geophysical Research: Solid Earth*, **123**, 296-310.

Shoji, D., Noguchi, R., Otsuki, S. & Hino, H. 2018. Classification of volcanic ash particles using a convolutional neural network and probability. *Nature, Scientific reports*, **8**, 8111.

Spencer, C.J., Cawood, P.A., Hawkesworth, C.J., Raub, T.D., Prave, A.R. & Roberts, N.M. 2014. Proterozoic onset of crustal reworking and collisional tectonics: Reappraisal of the zircon oxygen isotope record. *Geology*, **42**, 451-454.

Titos, M., Bueno, A., Garcia, L. & Benitez, C. 2018. A deep neural networks approach to automatic recognition systems for volcano-seismic events. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **11**, 1533-1544.

Trugman, D.T. & Shearer, P.M. 2018. Strong correlation between stress drop and peak ground acceleration for recent M 1–4 earthquakes in the San Francisco Bay area. *Bulletin of the Seismological Society of America*, **108**, 929-945.

Tschauner, O., Huang, S., Greenberg, E., Prakapenka, V.B., Ma, C., Rossman, G.R., Shen, A.H., Zhang, D., Newville, M., Lanzirotti, A. & Tait, K. 2018. Ice-VII inclusions in diamonds: Evidence for aqueous fluid in Earth's deep mantle. *Science*, **359**, 1136-1139.

UNESCO 2019. International Geoscience Programme. http://www.unesco.org/new/en/natural-sciences/environment/earth-sciences/international-geoscience-programme/.

Van Rees, E. 2016. DigitalGlobe and big data. *GeoInformatics*, **19**, 6.

Voosen, P. 2018. NASA lander to probe interior of Mars. *Science*, **360**, 247-248.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B. & Bourne, P.E. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, **3**.

Wu, Y., Lin, Y., Zhou, Z. & Delorey, A. 2018. Seismic-Net: A deep densely connected neural network to detect seismic events. *arXiv preprint arXiv:1802.02241*.

Wu, Y., Lin, Y., Zhou, Z., Bolton, D.C., Liu, J. & Johnson, P. 2017. Cascaded region-based densely connected network for event detection: A seismic application. *arXiv preprint arXiv:1709.07943*.

Yang, Z., Hou, Z. 2009. Porphyry Cu deposits in collisional orogen setting: A preliminary genetic model. *Mineral Deposits,* **28**, 515-538.

Yasukawa, K., Nakamura, K., Fujinaga, K., Iwamori, H. & Kato, Y. 2016. Tracking the spatiotemporal variations of statistically independent components involving enrichment of rare-earth elements in deep-sea sediments. *Scientific reports*, **6**, 29603.

Yu, K., Wang, D., Ding, W., Pei, J., Small, D.L., Islam, S. & Wu, X. 2015. Tornado forecasting with multiple markov boundaries. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2237-2246.

Zhao, T., Li, F. & Marfurt, K.J. 2017. Constraining self-organizing map facies analysis with stratigraphy: An approach to increase the credibility in automatic seismic facies classification. *Interpretation*, **5**, T163-T171.

Zhuang, Y., Yu, K., Wang, D. & Ding, W. 2016. An evaluation of big data analytics in feature selection for long-lead extreme floods forecasting. *2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC)*. IEEE, 1-6.

Figure captions

Fig. 1. (a) Data linking various earth spheres demonstrating co-evolution; **(b)** Major biological, environmental, resources and atmospheric impacts of Precambrian tectonic episodicity, modified from O'Neill et al. (2015) and Husson & Peters (2017); **(c)** Future evolution of magmatism, modified from Cheng (2018). The singularity is an index with values calculated by LSA filter which indicates intensity of magmatic activities.

Fig. 2. (a) A map showing two global tectonic converging zones: Alpine-Himalayan Orogenic Belt (Tethys) and Circum-Pacific Orogenic Belt, and **(b)** Frequency distributions created using U-Pb ages of the detrital zircons from Pacific and Tethys metallogenic belts. Age range is 0 to 100Ma, data from Puetz (2018).

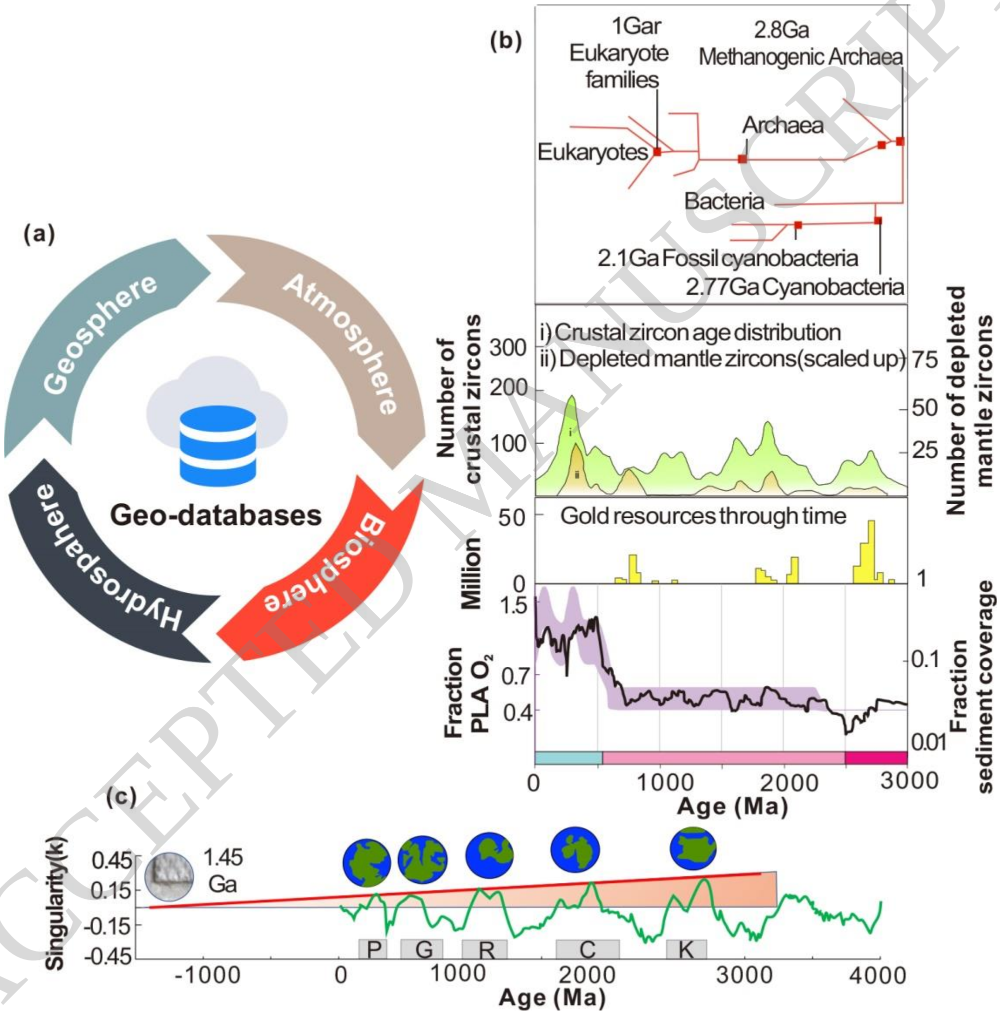
Fig. 3. Linking data, algorithms and projects for model automation and researcher collaboration.

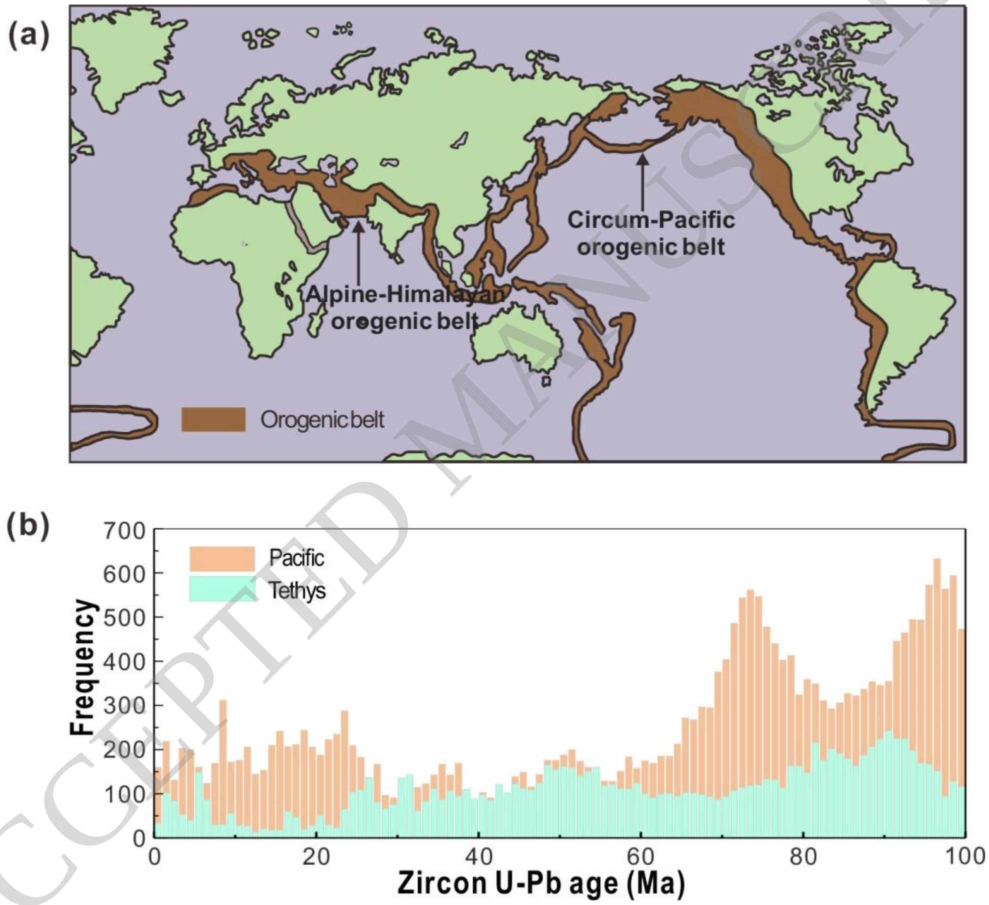
Fig. 4. Flow of discovery processes in linking data and processes for applications.

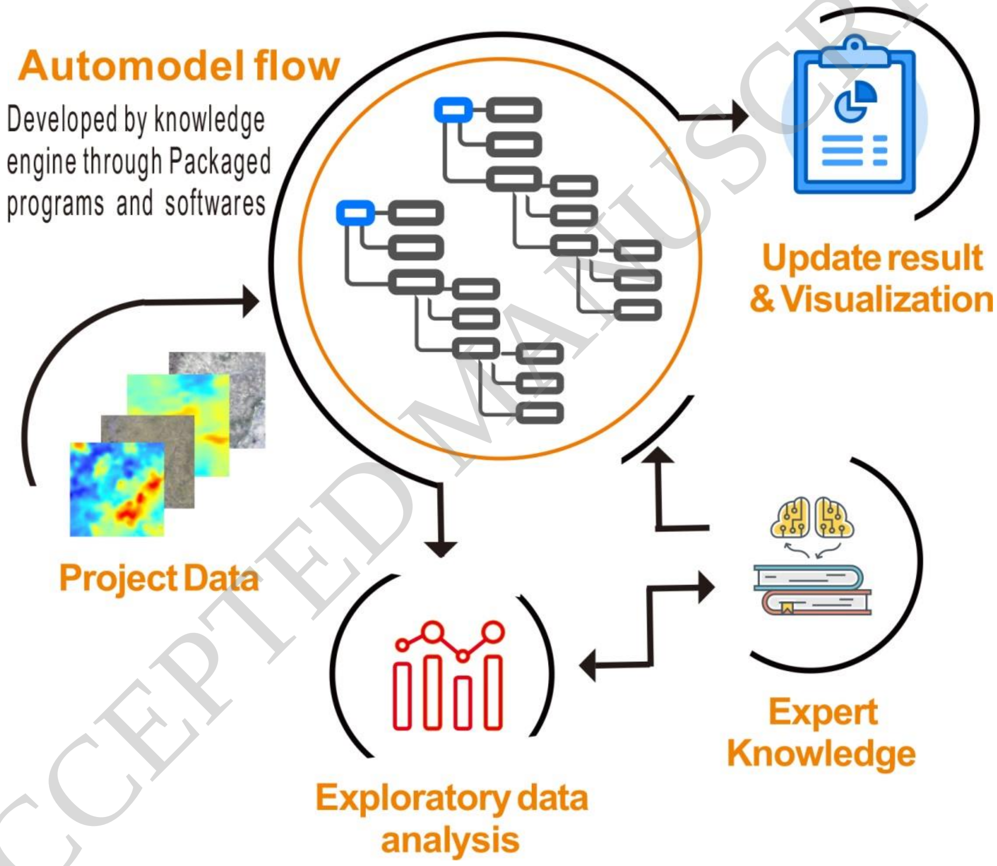
Fig. 5. The issues addressed by geoscientists are closely related to the UN 2030 Sustainable Development Goals (SDGs).

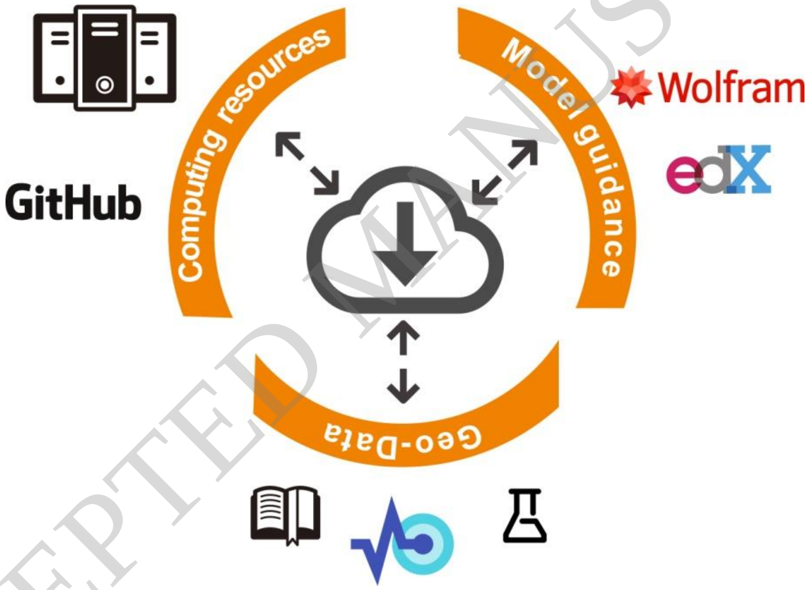
Fig. 6. Integration of main components of data, model, algorithms, data mining processes, computer network and people in supporting data-driven knowledge discovery.

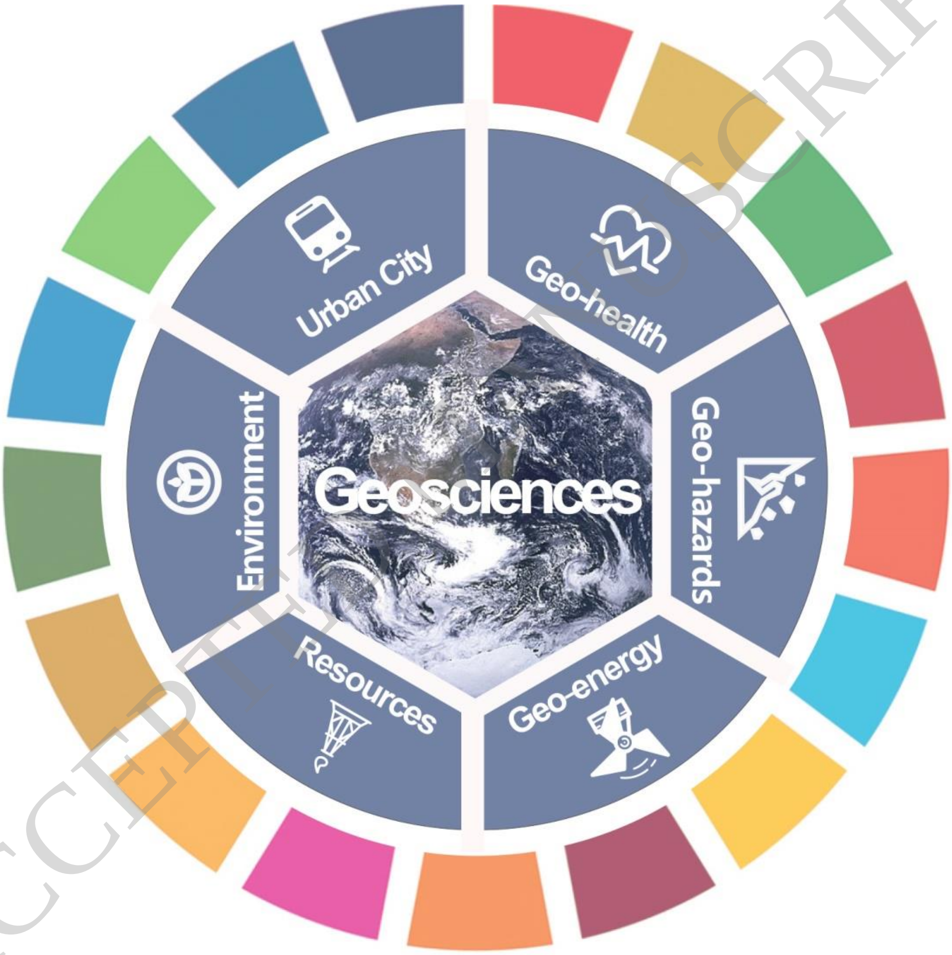
Fig. 7. Integration of earth science, computer science and mathematics.

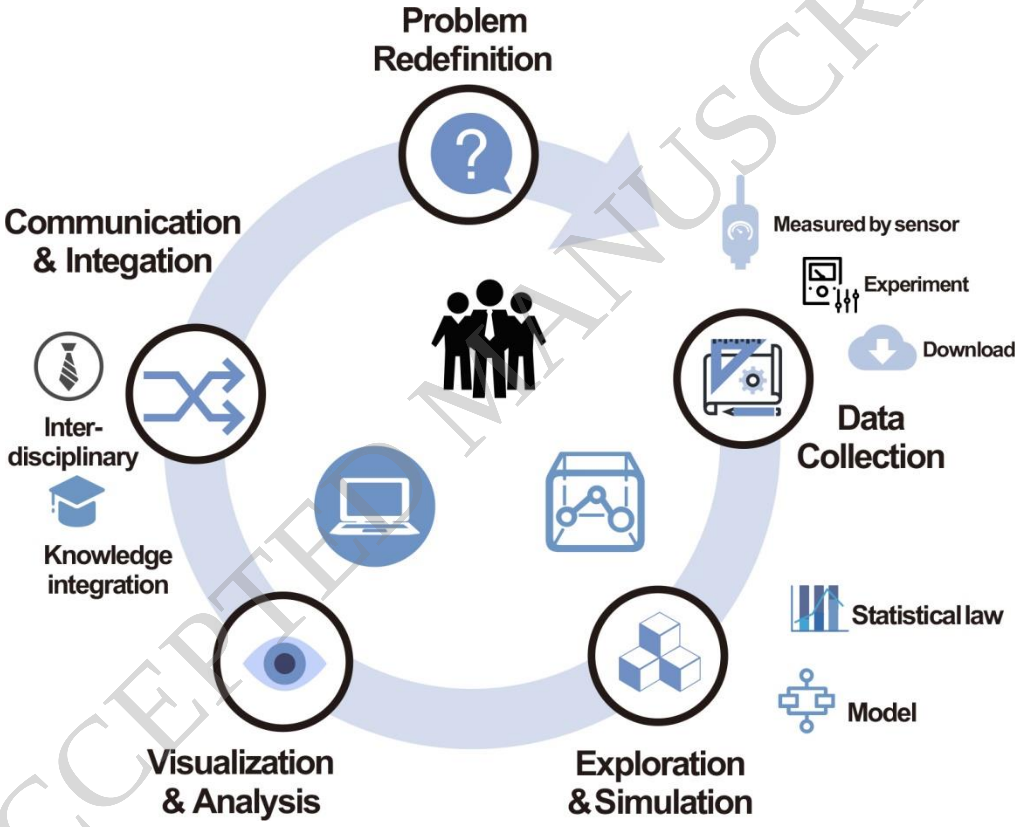

(a)
Geosphere
Atmosphere
Biosphere
Hydrospahere
Geo-databases
(b)
1Gar
Eukaryote
families
2.8Ga
Methanogenic Archaea
Archaea
Eukaryotes
Bacteria
2.1Ga Fossil cyanobacteria
2.77Ga Cyanobacteria
Number of crustal zircons
300
200
100
i) Crustal zircon age distribution
ii) Depleted mantle zircons(scaled up)
75
50
25
Number of depleted mantle zircons
Gold resources through time
50
0
Million
1.5
0.7
0.4
Fraction PLA O_2
0.1
0.01
1
Fraction sediment coverage
0
1000
2000
3000
Age (Ma)
(c)
Singularity(k)
0.45
0.15
-0.15
-0.45
1.45
Ga
P
G
R
C
K
-1000
0
1000
2000
3000
4000
Age (Ma)

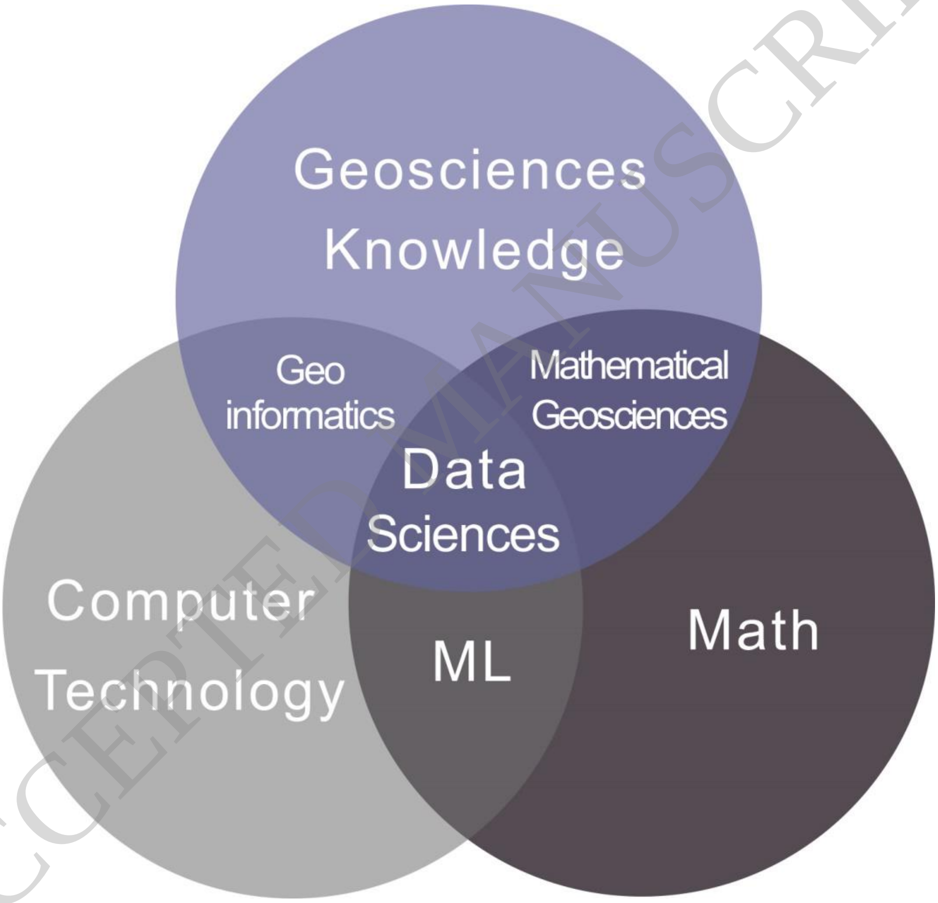

(a)
Circum-Pacific
orogenic belt
Alpine-Himalayan
orogenic belt
Orogenic belt
(b)
Pacific
Tethys
Frequency
0
100
200
300
400
500
600
700
0
20
40
60
80
100
Zircon U-Pb age (Ma)


Automodel flow
Developed by knowledge engine through Packaged programs and softwares
Update result & Visualization
Project Data
Expert Knowledge
Exploratory data analysis


Computing resources
Model guidance
Geo-Data
GitHub
Wolfram
edX


Geosciences
Urban City
Geo-health
Geo-hazards
Geo-energy
Resources
Environment


Problem
Redefinition
Measured by sensor
Experiment
Download
Data
Collection
Statistical law
Model
Exploration
& Simulation
Visualization
& Analysis
Knowledge
integration
Inter-
disciplinary
Communication
& Integration


Geosciences Knowledge
Geo informatics
Mathematical Geosciences
Data Sciences
Computer Technology
ML
Math