IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Review on Intelligent Systems for Geosciences and Big data.

R.Palson kennedy1 P.Kiran Sai2

1Prof-Peri Institute of Technology,Anna University affiliation, Chennai-48.TamilNadu, India

2Students- Geo Informatics, Institute of Remote Sensing College of Engineering Guindy,Chennai.25.

Abstract

Geoscience is a vital social area that demands solutions to numerous pressing issues confronting humanity and the earth. As the geosciences enter the era of big data, machine learning (ML)—which has been extensively successful in commercial domains—offers enormous promise to help solve geosciences challenges. This article introduces machine learning (ML) researchers to the challenges posed by geoscience problems and the potential for both machine learning and geosciences advancement. We begin by highlighting common sources of geoscience data and outlining their shared characteristics. Data science is gaining traction across a broad range of geosciences fields and applications. To meet that requirement, this article presents a review from a data life cycle viewpoint. Numerous facets of the geosciences present unique difficulties for the study of intelligent systems. Geosciences data is notoriously difficult to analyze since it is frequently unpredictable, intermittent, sparse, multiresolution, and multiscale. The spatiotemporal boundaries of geosciences processes and objects are frequently amorphous. Due to the absence of ground truth, evaluating, testing, and comparing models becomes challenging. Overcoming these obstacles will need substantial advancements in intelligent systems, which will help the geosciences tremendously in turn. Numerous successful data-driven geoscience discoveries have been reported recently, and many geoscience conferences have begun to include geoinformatics and data science sessions. Across academia, industry, and government, there is a strong desire to learn more about the current state of data science in geoscience as well as its potential. To address that need, this article provides a review from a data life cycle perspective. The data life cycle's critical steps include concept generation, data collection, preprocessing, analysis, archiving, distribution, discovery, and repurposing. The first section discusses the fundamental concepts and theoretical underpinnings of data science, while the second section summarises key points and shareable experiences from existing publications centred on each stage of the data life cycle. Finally, a future vision for data science applications in geoscience is discussed, including topics such as open science, smart data, and team science science. We hope that this review will be beneficial to data science practitioners in the geoscience community and spark additional discussion about data science best practises and future trends in geoscience.

Keyword: *Geoscience, Data Science, Intelligent system, Machine learning, Big data, data life cycle, Recent development, Trends*

1. INTRODUCTION

The goal of geosciences study is to get a better understanding of the Earth as a complex, highly interacting system of natural processes and their connections with human activities. Given the complexity of geosciences data, current methods have significant flaws. First and foremost, evidence alone is insufficient for the creation of models of the extremely complex processes under investigation; thus, preceding hypotheses must be taken into consideration. Second, data gathering can be most successful if it is guided by knowledge of current models in order to concentrate on data that will make a significant impact. Third, in order to integrate heterogeneous data and models from different disciplines, it is necessary to capture and reason about substantial qualifiers and context in order to make their integration feasible. The necessity for knowledge-rich intelligent systems that include substantial volumes of geosciences knowledge. Geosciences research seeks to comprehend the Earth as a complex, highly interactive system of natural processes and their interactions with human activities. Given the complexity of geoscience data, current approaches have fundamental flaws. To begin, using data alone is insufficient for developing models of the extremely complex phenomena under study; therefore, prior theories must be considered. Second, data collection can be most effective when guided by an understanding of existing models in order to concentrate on data that will make a difference. Third, integrating disparate data and models from disparate disciplines requires capturing and reasoning about extensive qualifications and context. These are all examples of the importance of knowledge-rich intelligent systems that incorporate a substantial amount of geoscience knowledge. Today, the speed of geosciences research is barely keeping up with the urgency created by societal requirements to manage natural resources, respond to geohazards, and comprehend the long-term implications of human actions on the globe.

Numerous aspects of geosciences pose novel problems for the study of intelligent systems. Geoscience data is notoriously difficult to analyse because it is inherently uncertain, intermittent, sparse, multiresolution, and multiscale. Processes and objects in the geosciences frequently have amorphous spatio-temporal boundaries. Due to the absence of ground truth, evaluating, testing, and comparing models becomes difficult. Overcoming these obstacles would require technological breakthroughs in intelligent systems, which would benefit the geosciences enormously. A newly formed Research Coordination Network on Intelligent Systems for Geosciences was formed in response to a workshop on this subject held at the National Science Foundation. The growing network capitalises on the momentum generated by the National Science Foundation's EarthCube initiative for geosciences and is motivated by pressing issues in Earth, ocean, atmospheric, polar, and geospace sciences.

As the deluge of big data continues to engulf virtually every commercial and scientific domain, geosciences has undergone a significant transformation from a data-poor to a data-rich field. This has been made possible by the advancement of sensing technologies (e.g., remote sensing satellites and deep sea drilling vessels), increases in computational resources for running large-scale simulations of Earth system models, and the Internet-based democratisation of data, which enables the collection, storage, and processing of data on crowd-sourced and distributed environments such as the Internet. The majority of geoscience data sets are freely accessible and do not present the privacy concerns that have stymied the adoption of data science methodologies in fields such as health care and cyber-security. The increasing availability of big geoscience data presents an enormous opportunity for machine learning (ML)—which has revolutionised almost every aspect of our lives (e.g., commerce, transportation, and entertainment)—to make a significant contribution to solving geoscience problems of significant societal importance.

2. Geoscience challenges requiring innovations in intelligent systems

Numerous recent papers have evaluated and detailed the difficulties inherent in geoscience research. Geosciences is the field of study that spans and describes the immense scales of Earth's temporal and spatial systems. These scales are accompanied by a remarkable range of data, knowledge, and scientific methodologies. Geoscience problems are rarely simple and symmetrical. The phenomena of Earth's systems are nonlinear, diverse, and highly dynamic. Extreme occurrences and long-term alterations in Earth systems will also pose challenges to geosciences study. Additionally, recent exceptional improvements in data availability, along with a greater emphasis on societal causes, underline the importance of cross-disciplinary research.

We discuss the requirements and their potential impact on a variety of scales:

2.1 Site-level requirements, for which recent research in intelligent sensors opens up new possibilities, particularly in difficult-to-reach regions. While collecting observations for all physical characteristics everywhere and at all times would be ideal, given resource and instrumentation limits, this is practically impracticable. Rather than that, the goal is to maximise the amount of science that can be accomplished within those restrictions, which requires enhancing the sophistication of existing data collection systems.

2.2 Regional-level requirements, where efficient procedures are required to integrate data from various locations, data kinds, and collection efforts spread across a large geographic area. While Earth systems are connected, geoscience data and models are not.

2.3 Global-level requirements, for which geoscience research can be both data-rich and data-deficient. That is, while it may be possible to collect enormous volumes of data about a phenomenon, the amount of information contained in the data may be trivial in comparison to the amount required to characterise the phenomenon for scientific or practical purposes. Scientists require novel ways that combine data with previously accumulated information about the underlying processes.

3. A roadmap for intelligent systems research with benefits to geosciences

Geosciences is fast transitioning from a small data to a big data age as a result of the enormous increase of observational and model data acquired about physical processes on the Earth. This has been made possible by technological breakthroughs in data collection and increased access to computing power. The increasing availability of data on the Earth system presents an enormous opportunity for intelligent systems research to speed developments in the geosciences, and vice versa.

The promise of intelligent systems research in the geosciences is enhanced by the recent success of classical intelligent systems methods in various commercial sectors utilizing enormous datasets, such as product recommendation and advertising. Geoscience datasets, on the other hand, exhibit a number of distinct properties that set them apart from large datasets in commercial areas. Geoscience datasets are extremely heterogeneous, are frequently spatiotemporal in nature, and the events or objects of interest lack sharp boundaries. Ocean eddies and hurricanes, for example, have amorphous spatiotemporal boundaries that manifest as patterns in continuous variables such as sea surface height. Geoscience datasets contain information on both well-known and little-understood physical processes and connections, which exhibit various features across the globe due to changes in geographies, climatic conditions, and seasonal cycles, among other factors. Even relatively uniform 'big data' from remote sensing is fraught with ambiguity, incompleteness, and a dearth of user-friendly tools.

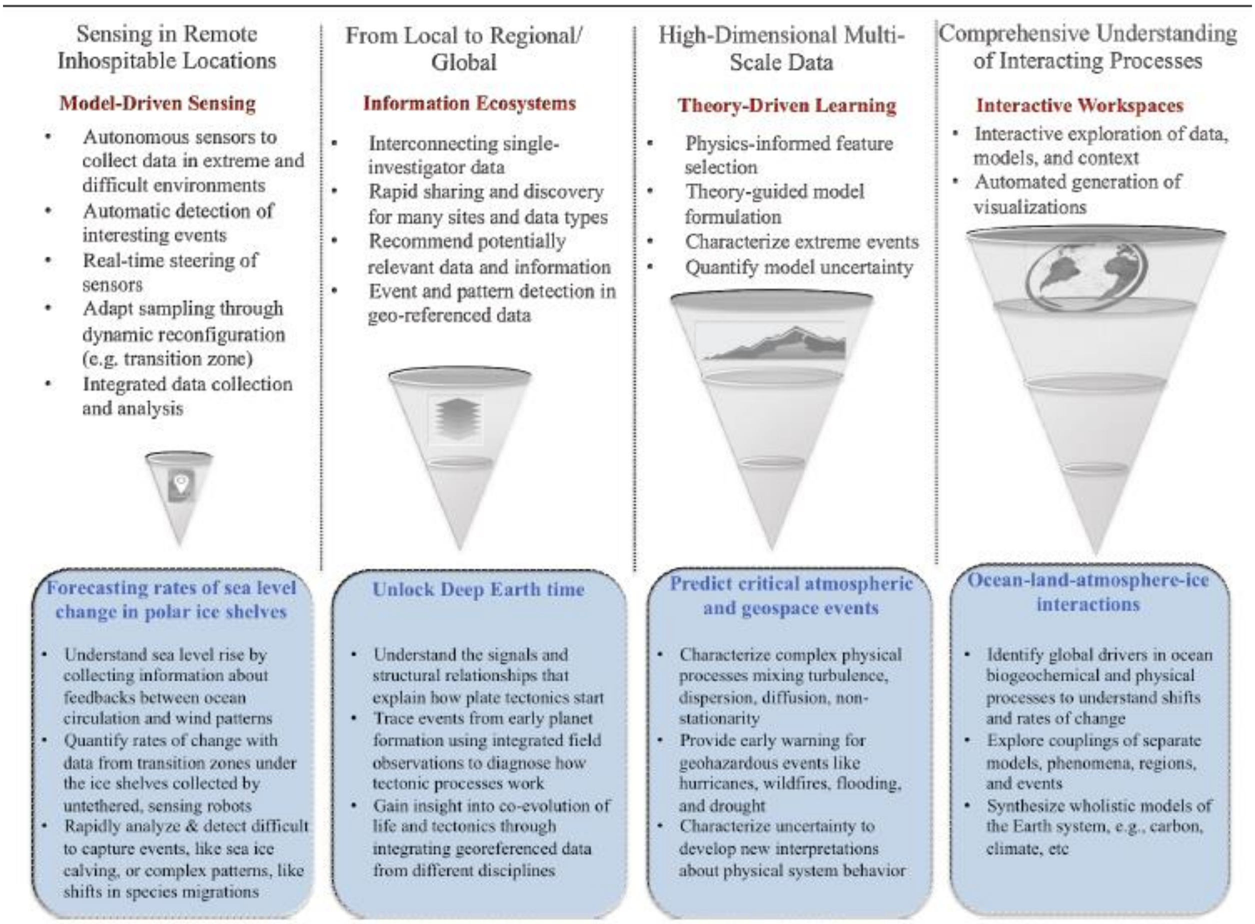
Sensing in Remote Inhospitable Locations
Model-Driven Sensing
• Autonomous sensors to collect data in extreme and difficult environments
• Automatic detection of interesting events
• Real-time steering of sensors
• Adapt sampling through dynamic reconfiguration (e.g. transition zone)
• Integrated data collection and analysis
From Local to Regional/ Global
Information Ecosystems
• Interconnecting single-investigator data
• Rapid sharing and discovery for many sites and data types
• Recommend potentially relevant data and information
• Event and pattern detection in geo-referenced data
High-Dimensional Multi-Scale Data
Theory-Driven Learning
• Physics-informed feature selection
• Theory-guided model formulation
• Characterize extreme events
• Quantify model uncertainty
Comprehensive Understanding of Interacting Processes
Interactive Workspaces
• Interactive exploration of data, models, and context
• Automated generation of visualizations
Forecasting rates of sea level change in polar ice shelves
• Understand sea level rise by collecting information about feedbacks between ocean circulation and wind patterns
• Quantify rates of change with data from transition zones under the ice shelves collected by untethered, sensing robots
• Rapidly analyze & detect difficult to capture events, like sea ice calving, or complex patterns, like shifts in species migrations
Unlock Deep Earth time
• Understand the signals and structural relationships that explain how plate tectonics start
• Trace events from early planet formation using integrated field observations to diagnose how tectonic processes work
• Gain insight into co-evolution of life and tectonics through integrating georeferenced data from different disciplines
Predict critical atmospheric and geospace events
• Characterize complex physical processes mixing turbulence, dispersion, diffusion, non-stationarity
• Provide early warning for geohazardous events like hurricanes, wildfires, flooding, and drought
• Characterize uncertainty to develop new interpretations about physical system behavior
Ocean-land-atmosphere-ice interactions
• Identify global drivers in ocean biogeochemical and physical processes to understand shifts and rates of change
• Explore couplings of separate models, phenomena, regions, and events
• Synthesize wholistic models of the Earth system, e.g., carbon, climate, etc

Figure 1. Needs and potential impact at different scales at which significant new avenues of research in geosciences would be open through advances in intelligent systems, illustrated with case examples. From left to right: 1) site-scale, 2) local scale, 3) global scale, and 4) layered wholistic scale.

To handle geosciences difficulties involving complex multi-scale, multi-process phenomena, scientists will require intelligent systems that integrate cutting-edge technology with their expertise, context, and experiences. Intelligent systems must incorporate process-centered geoscience knowledge about processes including physical, geological, chemical, biological, ecological, and human components. This will result in a new generation of intelligent systems that are rich in information and capable of unique forms of reasoning and learning from geosciences data.

3.1 Knowledge Representation and Capture

In order to create geoscience-aware intelligent systems, scientific knowledge relevant to those geoscience processes must be explicitly represented, captured, and shared.

3.1.1 Research Directions

i. **Representing Scientific Metadata**:
Geoscientists are collecting more data than ever before, yet raw data stored on isolated servers is useless. Recent work on semantic and linked open data standards permits the publication of datasets in Web standard formats with open access licences, as well as the description of their semantics via metadata that maps the data to a domain ontology. Additionally, they enable the creation of linkages between datasets to facilitate interoperability. New ways are required for automatically integrating data from disparate sources and conducting analysis on it without requiring extensive manual effort. Additionally, new techniques for automatically inferring semantic structure from raw data are required, as well as tools for integrating, analysing, and visualising big datasets.

ii. **Capturing Scientific Knowledge**.
An even greater challenge is representing the ever-evolving, uncertain, complex, and dynamic aspects of scientific knowledge and information. While ontologies are growing in use to state basic relations between objects, existing ontologies need to be extended to represent geoscience processes with buy-in from many diverse communities and capabilities of documenting, versioning, and representing various forms, such as spatio-temporal processes interacting with each other and multi-scale phenomena. These representations can be broadly linked to existing data and ontological concepts with actionable authority. Important challenges will arise in representing mathematical concepts, dynamic processes,uncertainty, and other aspects of a constantly growing scientific knowledge base. These representations need to be expressive enough to capture complex scientific knowledge, but they also need to support scalable reasoning that integrates disparate knowledge at different scales, and scientists need to understand the representations enough to trust the outcomes.

iii. **Interoperation of Diverse Scientific Knowledge**.
Scientific knowledge comes in many forms that use different tacit and explicit representations: hypotheses, models, theories, equations, assumptions, data characterizations, etc. These representations are all interrelated, and it should be possible to translate knowledge fluidly as needed from one representation to another. A major research challenge is the seamless interoperation of alternativerepresentations of scientific knowledge, from descriptive to taxonomic to mathematical, from facts to interpretation and alternative hypotheses, from small to larger scale, and from isolated processes to complex integrated phenomena.

3.1.2 Research Vision: Knowledge Maps

We envision dense knowledge networks that comprise explicit interconnected representations of scientific information that are spatially and temporally related. These would result in five-dimensional knowledge maps (3D + time + knowledge annotations). Interpretations and assumptions shall be properly documented and corroborated by observational data and mathematical models. Today's semantic networks and knowledge graphs connect disparate facts on the Web (e.g., Wikidata), but they contain superficial facts that lack the depth and context necessary for scientific investigation. Knowledge maps will incorporate more detailed representations of spatiotemporal processes and will be physically grounded, integrating the various models of geoscience systems.

3.2 Robotics and Sensing

Collecting data is a common undertaking in the geosciences. Sensing and robotics research has the potential to have a significant impact on the geosciences through intelligent sensing and knowledge-based data collection.

3.2.1 Research Directions

i. **Data Collection Optimization:** Geoscience data are required at a variety of spatial and temporal domains. Due to the impossibility of continuously monitoring all measurements at all scales, intelligent sensing technologies are critical. Prior to sensor deployment, additional research is needed to evaluate the cost of data collecting, whether in terms of storage capacity, energy consumption, or monetary cost. A related research topic is balancing the expense of data gathering against the utility of the data that will be obtained.

ii. **Sampling in Progress:** Geoscience knowledge can be used to inform autonomous sensing systems, enabling not just long-term data collecting but also increasing sensing effectiveness by adaptive sampling, resulting in richer data sets at cheaper costs. In an adaptive sensing scheme, autonomous vehicles equipped with an embedded decision architecture assimilate data in order to produce and continually update an environmental model that is guided by geoscience knowledge and provides the sensing system with previous forecasts and estimations.

iii. **Collecting Data Through Crowdsourcing**: Citizen scientists can give valuable data (e.g., obtained via geolocated mobile devices) that would be extremely difficult to obtain otherwise. One problem in crowdsourcing data gathering is assuring the high quality of data required for geoscience research. A possible field of research is to enhance empirical methods for evaluating crowdsourced data collection and to acquire a better understanding of the biases inherent in the process.

iv. **Sensing Virtually:** Existing repositories could be enhanced through the use of virtual reality and augmented reality user interfaces to enable "virtual data collection" through navigation and selection of relevant data. A method of collecting virtual reality data would be to visualise existing datasets, utilising a highly interactive virtual reality platform to sort through accessible data.

3.2.2 Research Vision: Model-Driven Sensing

Sensor research will result in the development of a new generation of devices that will have a better understanding of the scientific context for the data being collected; they will use this understanding to maximise their performance and efficacy in modelling the phenomena being investigated. This will result in the development of new model-driven sensors with increased autonomy and exploratory capabilities.

3.3 Machine Learning

The proposed bidirectional, collaborative research program's outcome might be a scientifically correct, valuable, and trustworthy landscape of data, models, information, and knowledge. Scientific discovery generates integrated large-scale data products from raw measurements. These items are discussed in detail to illustrate the derivations and assumptions made in order to boost other scientists' comprehension and trust. These well-established scientific lines will be easily navigable, queryable, and displayed.

3.3.1 Modern machine learning tools

This decade ushers in a paradigm shift in tooling, which is directly responsible for the recent surge in use and research in both shallow and deep machine learning.

Historically, machine learning software has been dominated by proprietary applications such as MatlabTM with the Neural Networks Toolbox and Wolfram MathematicaTM, or by university-based efforts such as the Stuttgart Neural Network Simulator (SNNS). Shortly thereafter, LibSVM was released as free open-source software (FOSS), enabling the efficient implementation of support vector machines. It is still in use in a large number of other libraries, notably WEKA [Chang and Lin, 2011]. Torch, a machine learning framework with a focus on neural networks, was then released in 2002. While the original implementation in the computer language Lua has been discontinued [Collobert et al., 2002], PyTorch, the Python implementation, is one of the leading deep learning frameworks at the time of writing [Paszke et al., 2017]. Theano and scikit-learn were released as open-source Python libraries in 2007 [Theano Development Team, 2016, Pedregosa et al., 2011]. Theano is a neural network library that was developed at the Montreal Institute for Learning Algorithms (MILA) and halted development in 2017 following the availability of openly licenced deep learning frameworks by major industrial developers. Scikit-learn implements a variety of shallow machine learning algorithms, such as SVMs, Random Forests, and shallow neural networks, as well as utility functions such as cross-validation, stratification, metrics, and train-test splitting, which are required for the development and evaluation of robust machine learning models.

By establishing an uniform application programming interface (API), scikit-learn formed the current machine learning software package [Buitinck et al., 2013]. The following code snippets demonstrate this API. To begin, we use a utility function to construct a categorization dataset. The make classification function accepts many arguments to change the desired arguments; in this case, we are creating 1000 samples (n samples) with four features (n features), two of which are genuinely significant to the classification (n informative). X contains the data, whereas y contains the labels.

```
# Generate random classification dataset for example
from sklearn.datasets import make_classification,
X, y = make_classification(n_samples=5000, n_features=5
                           n_informative=3, n_redundant=0,
                           random_state=0, shuffle=False)
```

It is recommended to divide the available labelled data into two sets: a training set and a validation or test set. This division enables models to be evaluated on previously unseen data in order to determine their generalizability to previously unseen samples. Train test split is a utility function that accepts an arbitrary number of input arrays and divides them according to provided arguments. 25% of the data is retained for the hold-out validation set and is not used in training in this circumstance. The random state variable is fixed to ensure reproducibility of these examples.

```
# Split data into train and validation set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=.25,
                                                    random_state=0)
```

Then, in light of the prior discussion of high-impact machine learning models, we need to define a machine learning model. The first example is an SVM classifier. This example uses the SVM classifier's default parameters; for optimal performance on real-world issues, these values must be modified. Machine learning training is always performed by executing classifier.fit(X, y) on the classifier object, which is the SVM object in this case.

```
# Define and train a Support Vector Machine Classifier
from sklearn.svm import SVC
svm = SVC(random_state=0)
svm.fit(X_train, y_train)

>>> SVC(C=1.0, break_ties=False, cache_size=200,
        class_weight=None, coef0=0.0, degree=3,
        decision_function_shape='ovr', gamma='scale',
        kernel='rbf', max_iter=-1, probability=False,
        random_state=0, shrinking=True, tol=0.001,
      verbose=False)
```

By using classifier.predict(data) on the learned classifier object, the trained SVM may be used to predict on new data. The new data must contain the same four characteristics as the training data. By and large, machine learning models must always be trained on the same set of input attributes as the data being predicted.

```
# Predict on new data with trained SVM
print(svm.predict([[0, 0, 0, 0, 0],
                   [-1, -1, -1, -1, -1],
                   [1, 1, 1, 1, 1]]))
>>> [1 0 1]
```

The classifier.score() function should be used to evaluate the blackbox model. Evaluating the model's performance on the training data set provides valuable insight into the model's performance.

Additionally, on the hold-out set, the trained model can be evaluated. The default score equals the accuracy, indicating that our model is around 90% accurate. Similar train and test scores show that the machine has developed a generalizable model, which enables prediction on unknown data without incurring performance degradation.

```
# Score SVM on train and test data
print(svm.score(X_train, y_train))
print(svm.score(X_test, y_test))
>>> 0.9098666666666667
>>> 0.9032
```

Support-vector machines are applicable to all categories of machine learning problems, including classification, regression, and clustering. In a two-class problem, the algorithm analyses the n-dimensional input and seeks a (n -1)-dimensional hyperplane that separates the data points in the input. If the two classes are linearly separable, commonly known as a hard margin, the task is easy. The aircraft is capable of transmitting both sorts of data without ambiguity.

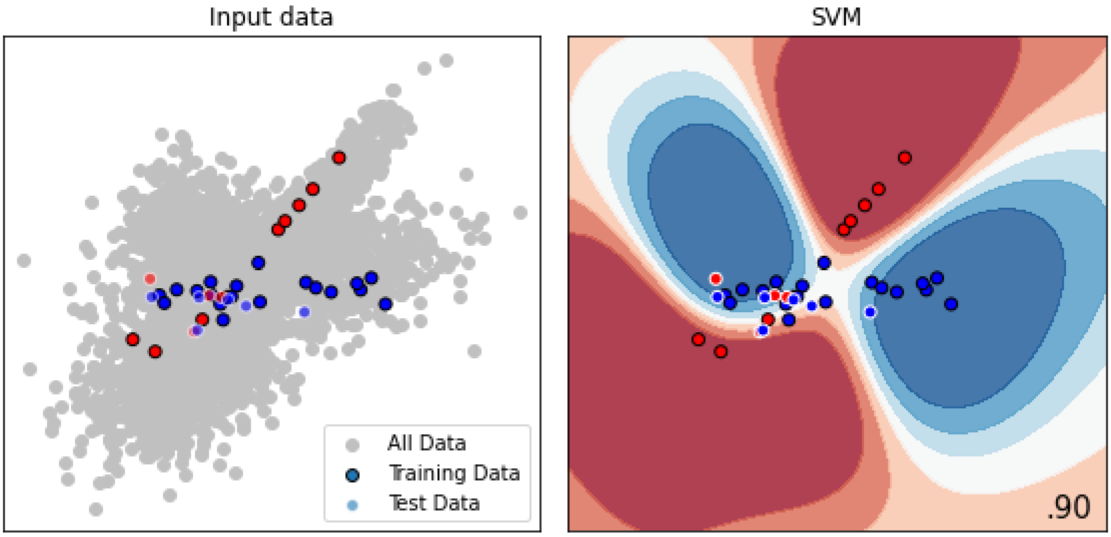

Figure 1 : Example of Support Vector Machine separating two classes of data points in 2D, showing the decision boundary learnt from the data.

Explainability is a critical concept in machine learning, as it examines the effect of input factors on the prediction. The mean values of the estimated importances indicate that three features are three orders of magnitude more significant, with the second feature providing the most information for label prediction.

```
# Calculate permutation importance of SVM model
from sklearn.inspection import permutation_importance
importances = permutation_importance(svm, X_train, y_train,
                                     n_repeats=10, random_state=0)
# Show mean value of importances and the ranking
print(importances.importances_mean)
print(importances.importances_mean.argsort())
>>> [ 2.1787e-01 2.8712e-01 1.2293e-01 -1.8667e-04 7.7333e-04]
>>> [3 4 2 0 1]
```

Support-vector machines have been used in the analysis of seismic data [Li and Castagna, 2004] and in the automatic interpretation of seismic data [Liu et al., 2015, Di et al., 2017b, Mardan et al., 2017]. These techniques typically perform worse than convolutional neural networks, because SVMs treat each sample independently. Other prominent uses of SVM in Geoscience include seismic tremor categorization [Masotti et al., 2006, 2008] and ground-penetrating radar analysis [Pasolli et al., 2009, Xie et al., 2013]. Society of Exploration Geophysicists 2016 (SEG) machine learning competition was organised with an

SVM as the baseline [Hall, 2016]. Several other authors examined well log analysis [Anifowose et al., 2017, Caté et al., 2018, Gupta et al., 2018, Saporetti et al., 2018], as well as seismology for event classification [Malfante et al., 2018] and magnitude determination [Ochoa et al., 2018]. These rely on the ability of SVMs to perform regression on time-series data. SVMs' strong mathematical foundation has enabled a wide variety of applications in geoscience, including microseismic event classification [Zhao and Gross, 2017], seismic well ties [Chaki et al., 2018], landslide susceptibility [Marjanovic et al., 2011, Ballabio and Sterlacchini, 2012], and digital rock models [Ma et al., 2012].

3.3.2 Modern Deep Learning

The decade of the 2010s saw a rebirth in deep learning, most notably convolutional neural networks. Historically, AlexNet [Krizhevsky et al., 2012] was the first convolutional neural network (CNN) architecture to enter the ImageNet challenge [Deng et al., 2009]. The ImageNet challenge is a benchmark competition and library of natural images for computer vision. This reduced the categorization error rate from 25.8% to 16.4%. (top-5 accuracy). This has sparked interest in CNN research, resulting in error rates of 2.25 percent on ImageNet's top-5 accuracy in 2017 [Russakovsky et al., 2015]. Tensorflow was introduced as an open source deep learning model library [Abadi et al., 2015], with a slightly different software design than the Theano and Torch libraries.

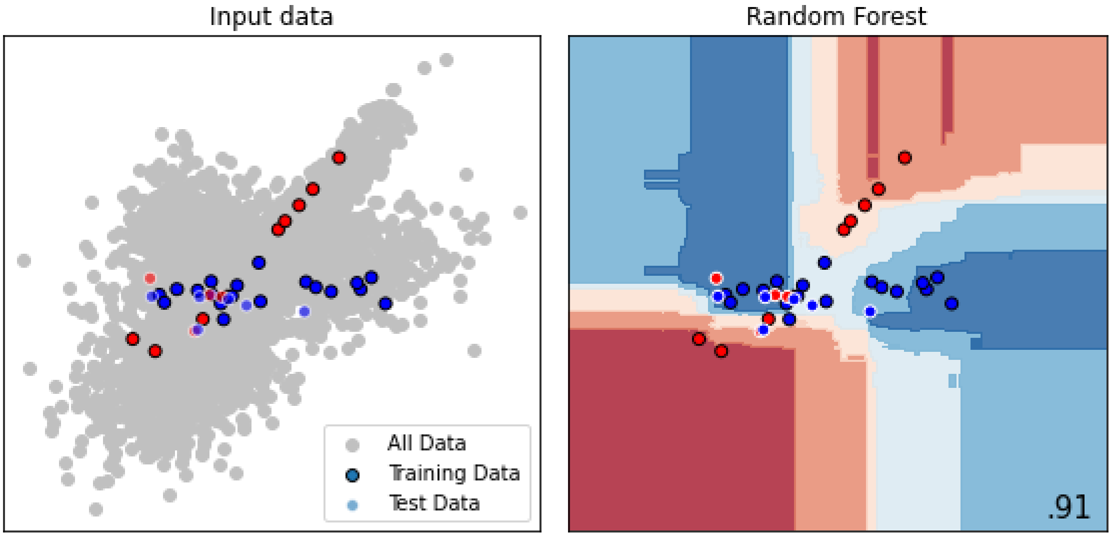

Figure 2: Binary Decision Boundary for Random Forest in 2D

The following example illustrates how deep learning is applied to the data supplied in the preceding examples. We employ independent samples in the categorization data set, which necessitates the usage of basic densely connected feed-forward networks. While it is great to feed image data or spatially linked datasets to a convolutional neural network (CNN), time series are frequently better tackled using recurrent neural networks (RNN). This example is created in Python and makes use of the Tensorflow package. While PyTorch is an excellent tool to use, the author prefers to write a succinct example using the Tensorflow API.

The sample model is composed of Dense layers and a Dropout layer that are sequentially assembled. Densely linked layers contain a predetermined number of neurons with a predetermined activation function, as illustrated in the example below. Each neuron executes the calculation described in Equation 1, with the activation defined. Nowadays, modern neural networks rarely employ sigmoid and tanh activations. Their activation property causes them to lose information at extreme positive and negative input values, which is referred to as saturation. This saturation of neurons hampered the performance of deep neural networks until new non-linear activation functions were introduced. The activation mechanism Due to their non-saturating qualities, the rectified linear unit (ReLU) is widely credited with aiding the creation of very deep neural networks [Hahnloser et al., 2000]. As seen in equation 6, it zeroes out all negative values and delivers a linear response for positive values. Numerous other rectifiers with varying qualities have been introduced since its start.

$$\_(a) = \max(0; a)$$

The other activation function used in the example is the "softmax" function on the output layer. This activation is commonly used for classification tasks, as it normalizes all activations at all outputs to one. It achieves this by applying the exponential function to each of the outputs in ~a for class C and dividing that value by the sum of all exponentials:

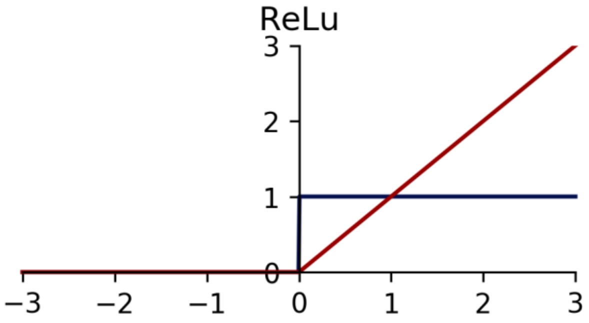

Figure 3: ReLU activation (red) and derivative (blue) for efficient gradient computation.

$$\sigma(\vec{a})_j = \sum_{p}^{e^{a_j}} e^{a}\ p$$

Additionally, the example employs a Dropout layer, which is a widely used technique for regularising networks by randomly changing a preset percentage of nodes to zero for each iteration. Neural networks are particularly prone to overfitting, which can be mitigated using a variety of regularisation strategies, including input data augmentation, noise injection, L1 and L2 limitations, and early training loop termination [Goodfellow et al., 2016]. For regularisation, modern deep learning systems may even employ noisy student-teacher networks [Xie et al., 2019b].

```
import tensorflow as tf
model = tf.keras.models.Sequential([
tf.keras.layers.Dense(32, activation='relu'),
tf.keras.layers.Dropout(.3),
tf.keras.layers.Dense(16, activation='relu'),
tf.keras.layers.Dense(2, activation='softmax')])
```

3.3.3 The State of ML on Geoscience

Geoscience, particularly geophysics, has closely followed breakthroughs in machine learning. Machine learning techniques have been applied across fields to a variety of challenges that may be broadly classified into three categories:

1. Create a fictitious machine learning model of a well-understood process. This paradigm typically has a cost advantage in terms of computation.

2. Create a machine learning model for a task that could previously only be accomplished through human contact, interpretation, or knowledge and experience.

3. Create a fresh machine learning model capable of performing a previously impossible task.

4. Data Science for Geosciences:-

The last decade has seen a surge in interest in data-driven discovery in geoscience research, as seen by the increasing number of financed initiatives, new facilities, shared datasets, and published scientific findings. Cyberinfrastructure, data portals, databases, workflow platforms, statistical models, machine learning algorithms, data management, and data sharing are all becoming increasingly common in the daily work of many geoscientists. Numerous successful instances of data-driven geoscience discovery over the last few years have proven the data revolution's great potential. It is self-evident that data science will play a critical role in the coming decades in order to scale up innovation and accelerate new discoveries in geoscience. Nonetheless, because data science's theoretical foundations are still being developed, there is little debate and review of data science in geoscience. By contrast, geoscientists are currently in high demand for data science methodologies and tools. To meet that requirement, the objective of this work is to synthesise recent advances in both data science and data-driven geoscience in order to give a review and anticipate future developments.

4.1 Trends in data science

To gain a better grasp of data science workflows, it is vital to comprehend a few key ideas. In recent years, the author has taught database and data science classes to senior undergraduate and graduate students. Even students majoring in computer science may become perplexed by the definitions of data, metadata, information, and knowledge, as experience has demonstrated. The term "data" refers to the documented representation of facts. Nowadays, in the digital era, records are typically stored digitally in formats such as plain text, spreadsheet, relational database, or graph database. The meaning or message extracted from data is referred to as information. The process of extracting information is frequently determined by the objective of the data analysis, the methodologies and instruments utilised, and the interpretation of the data analysis results.

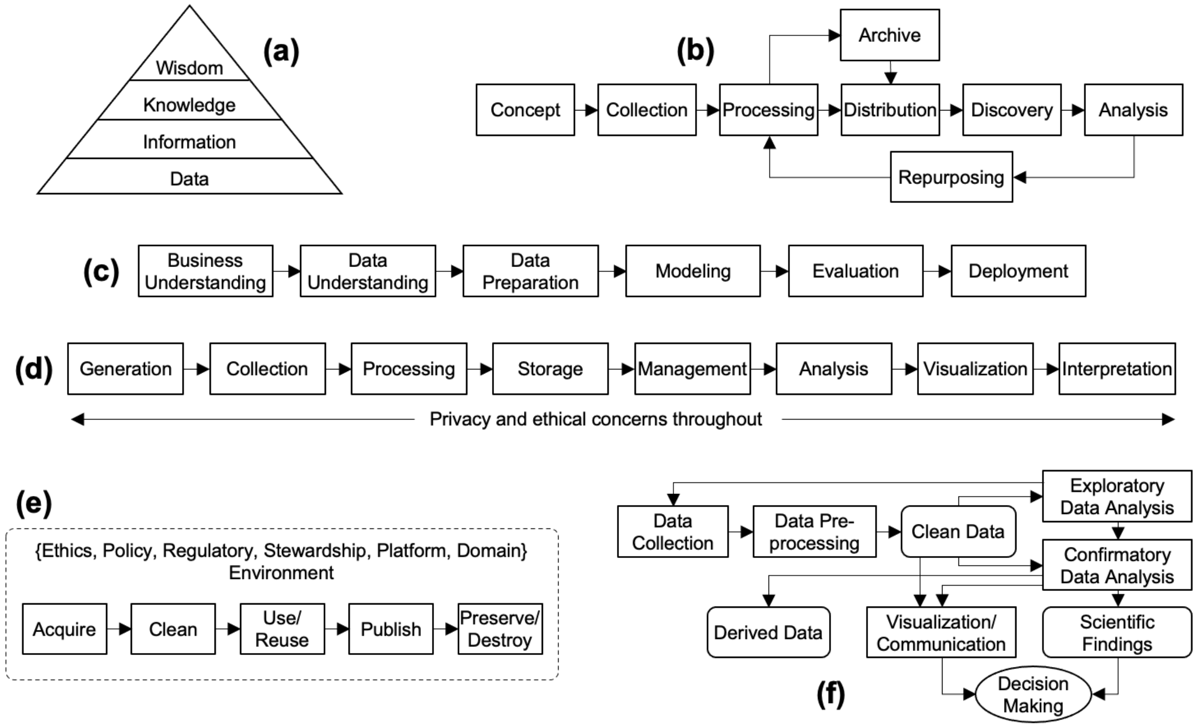

Figure 3. Different depictions of the data life cycle and the data science process. (a) the DIKW model; (b) the Data Documentation Initiative (DDI) data life cycle (DDI Alliance, 2021); (c) the cross-industry standard process for data mining (CRISP-DM) (Chapman et al., 2000); (d) the data life cycle in data science (Wing, 2020) (e) the data life cycle and surrounding data ecosystem (Berman et al., 2018); and (f) the data science process (Schutt and O'Neil, 2013).

Data science emerged and evolved as a result of multidisciplinary collaboration. Donoho (2017) provided a comprehensive overview of the evolution of data science over the last three decades. He highlighted numerous statisticians' viewpoints on the importance of broadening the scope of classical statistics to include data preparation, presentation, and prediction. According to a recent report from the National Academies of Sciences, Engineering, and Medicine (NASEM, 2018a), a critical task of data science education is to develop data acumen, which encompasses the following key concepts: mathematical foundations, computational foundations, statistical foundations, data management and curation, data description and visualisation, data modelling and assessment, and workflow and analysis. These data literacy issues are mirrored in the data life cycle and data science methodology (Figure 1), which are designed to fulfil the real-world requirements of data science applications. Numerous colleges have already begun to offer courses in data science. For instance, the University of California, Berkeley's Data 8: Foundations of Data Science course is designed for freshmen in any major (Adhikari and DeNero, 2017). Its curriculum encompasses the majority of the courses mentioned in the preceding list of data acuity.

4.1 A reflection on the key steps of a data life cycle

Focusing on the theme of data science for geoscience, the following sub-sections will review a list of recent publications for each key step in the data life cycle, and summarize the shareable experience from them.

4.1.1 Business understanding and concept

The steps labelled "concept" in Figure 1b and "business knowledge" in Figure 1c are meant to help define the data science project's objectives and estimate data requirements (Chapman et al., 2000; DDI Alliance, 2021). They are concerned with translating business objectives into data science plans. If database development is part of the intended activities, this step will also include work on data structures such as conceptual models, logical models, physical models, and controlled vocabularies for data standards. Cyberinfrastructure researchers have realised that early consideration and action on data semantics can aid in improving data interoperability when data is generated, gathered, integrated, and shared (Reitsma et al., 2009; Narock and Shepherd, 2017).

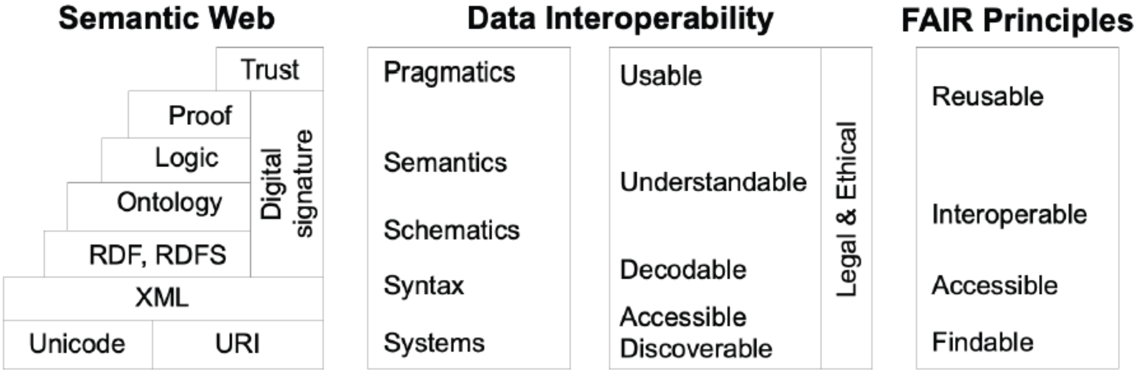

Figure 4. Comparing the layered structure of data interoperability with the Semantic Web architecture and the FAIR data principles

Numerous academics have described the layered structure of data interoperability, which encompasses systems, syntax, schematics, semantics, and pragmatics (Bishr, 1998; Sheth, 1999; Ludäscher et al., 2003; Brodaric, 2007, 2018). Several further studies defined these levels in layman's words, such as discoverable, accessible, decodable, intelligible, and useable (Wood et al., 2010; Ma et al., 2011). The layered structures of data interoperability and the FAIR principle are similarly comparable to the Semantic Web's technological design (Berners-Lee, 2000). Numerous examples of data interoperability best practises may be found in the domain of geoscience.

The United States Geological Survey's (USGS) National Geologic Map Database has adopted the North American Geologic Map Data Model (NADM) (NADM Steering Committee, 2004) as a standardised schema for organising state-level geologic map databases. At the USGS, such efforts on standards are ongoing, as evidenced by the recently released Geologic Map Schema (GeMS) (USGS NCGMP, 2020). Similarly, NASA uses the Global Change Master Directory (GCMD) Keywords as a hierarchical collection of controlled vocabularies to ensure the interoperability of its data and services (GCMD, 2020). In Europe, the INSPIRE Directive intends to establish a spatial data infrastructure for the European Union (Bartha and Kocsis, 2011; Ma and Fox, 2014). Its data and metadata guidelines span 34 data topics in Earth and environmental sciences, and complete implementation across all participating European nations is required by 2021.

4.1.2 Data understanding, generation and collection

NASA manages about 100 missions and hundreds of platforms, equipment, and sensors orbiting the Earth and nearby space, and is one of the world's largest producers of geoscience data. According to Shannon (2019), NASA generated 12.1TB of data per day in 2016. Additionally, the same storey said that NASA was installing new sensors capable of generating 24 terabytes of data every day. The same advancements in instrumentation and data generation, transmission, and management were observed in field-based geological survey (Mookerjee et al., 2015). Wing (2019) distinguished data generation from data collection, noting that not all data generated is captured (Figure 1d). This could be because we just want to capture a subset of the data, or because the velocity of data streams is too high for present technologies to process.

4.1.3 Data preprocessing and preparation

Preprocessing data is becoming an increasingly critical stage in data science. Additionally, it is referred to by various alternative terms, including data cleansing, data wrangling, and data munging. The goal of data preprocessing is to assure the quality of data prior to conducting any data analysis. It may include tasks such as clearing out noisy and unreliable records, lowering data dimensionality, changing data formats, choosing records of interest, enriching existing data with extra properties, and combining data from many sources to create a new piece of data (Wang, et al., 2018). Numerous new research discoveries have been made as a result of the upgraded database, including mineral evolution and ecology (Morrison et al., 2019, 2020) and the co-evolution of the geosphere and the biosphere (Spielman and Moore, 2020). Additionally, the database resulted in new designs for mineral species databases and talks about improved data curation and sharing methods (Prabhu et al., 2021).

4.1.4 Data archive, distribution, and discovery

Funding agencies increasingly demand researchers to provide a data management plan with their grant submissions (Dietrich et al., 2012; NSF, 2015). Data are increasingly being viewed as a formal research output on par with paper papers and receiving the same level of attention. The ideas of FAIR data (Wilkinson et al., 2016) are now widely accepted across practically all scientific disciplines, including geoscience (Stall et al., 2019; Lannom et al., 2020). The FAIR data principles build on a long history of data management and stewardship activities and provide a systematic way to sharing and reusing scientific data in open science. NASA, the US Geological Survey, the National Oceanic and Atmospheric Administration, and the United States Department of Agriculture all have their own data archives and portals that enable users to search for and retrieve relevant data. For example, through a central interface, the USGS supports federated querying of a large number of spatial datasets devoted to mineral resources (USGS MRDATA, 2021). With the increased use of workflow platforms such as Jupyter Notebook and R Markdown, many data portals have developed packages to facilitate data access from workflow platforms, such as the paleobioDB R package for the Paleobiology Database (Varela et al., 2015) and the neotoma R package for the Neotoma Paleoecology Database (Varela et al., 2015). (Goring et al., 2015).

4.1.5 AI and Small Data Scalable

Covid-19 severely disturbed the sorts of data accessible for analysis and, as a result, the utilization of that data. More individuals are accessible online to study a wider range of data, yet these data are quite different from past sets of big data. That is why the AI 'small data' approaches take primacy, based on fewer consumer behavior occurrences. Therefore, **artificial intelligence** (AI) must be scalable to respond, despite the knowledge that huge amounts of data are historically better at predicting accurately. **Machine learning** must also adapt to the new analytical limitations arising from increased internet activity. New privacy laws such as the California Consumer Privacy Act of 2020 will make it more difficult to focus on 'little data' and allow more past data to be accessible.

4.1.6 Cloud Computing

The transition to cloud-based data storage has made a difference for many companies that prefer the safety of local servers and simply see the cloud as a transaction tool, as its initial function was. However, as cloud technology develops fast, new data science trends have enticed many companies to replenish their data storage. Providers like Amazon, Microsoft and Google are now the main method organisations may store their data and offer built-in analytics to ease the process of data management. By the end of 2022, Gartner says that 90% of innovation in data and analytics would need public cloud services, with a cloud-based AI five times as important as it was in 2019 within a year.

4.1.7 Real- time data

Real-time automated testing is one of the largest new data analysis capabilities in 2021. This signifies a trend away from historical data that is out of date by definition. Companies may now connect more effectively with their product or service consumers, responding to customer behaviors, instead of analyzing their data at a later period. According to Seagate, 75% of the world's population will interact every 18 seconds with data by 2025, making it vital to speed up the data analysis and the following reaction.

4.1.8 Progress in Data science

The rapid growth of Big Data and Data Science has spurred greater ideas and goals for data-driven geosciences study. The Carnegie Institution for Science launched the "4D" programme in 2018. (4D Initiative, 2018). In 2019, the International Union of Geological Sciences started the major research initiative Deep-time Digital Earth (DDE) (Cheng et al., 2020). Open data and community of practices on cyber infrastructure requirements and progress were made as part of the major recommendation in the vision (NASEM, 2020) for the next ten earth-science goals for the U.S. National Science Foundation (NSF). We are at a major turning point in science—a moment in which the way geoscientists do research will be altered by open data resources, cyber-infrastructure facility and new data science methods of analysis and visualization. Caps to uncover are the ongoing creation, integration and exploitation of facilities, data and knowledge to create and explore methods to understand the Earth more deeply (Hazen et al., 2019).

5. Conclusion

In the world of data science, it is new, and we are still figuring out what it is. For the time being, the term is best defined by the work of a data scientist. A data scientist is someone who utilizes programming as the foundation for a more in-depth and flexible approach to data analysis. Researchers in intelligent systems and geosciences collaborate to develop knowledge-rich frameworks, algorithms, and user interfaces that are easy to use and understand. Understanding that these connections are unlikely to occur without considerable assistance, a new Research Coordination Network on Intelligent Systems for Geosciences has been established to facilitate sustained communication across various areas that do not normally cross paths with one another. Enabling these advancements will require collaboration between academics in intelligent systems and geosciences to develop knowledge-rich frameworks, algorithms, and user interfaces. Recognizing that these linkages are unlikely to occur without major facilitation, a new Research Coordination Network on Intelligent Systems for Geosciences has been established to facilitate sustained communication across these domains that rarely intersect. This network is focused on three primary objectives. To begin, collaborative workshops and other platforms will facilitate synergistic talks and reveal shared interests. Second, repositories of challenge issues and datasets with succinct challenge statements are intended to minimise the entry barriers. Third, a curated archive of educational materials will be established to assist researchers and students in overcoming the steep learning curve associated with advanced topics in the other discipline. In geoscience, machine learning has a lengthy history. Kriging has evolved into more generic machine learning techniques, and geology has made tremendous strides with the application of deep learning. Applying deep convolutional networks to autonomous seismic interpretation has advanced these systems beyond what was previously conceivable, however this area of research remains active. Developing custom neural networks and shallow machine learning pipelines has become increasingly simple with new tools, enabling widespread applications in every subject of geoscience. Nonetheless, it is critical to recognise machine learning's limitations in geoscience. These are cutting-edge technologies, but developing fully tested models takes time, which might be disadvantageous when working in a hot scientific subject. While none of these applications are totally automated, as the allure of artificial intelligence would suggest, substantial new discoveries have been provided in applied geoscience. These applications make use of machine learning as a pre-processing tool for data, extending previous insights beyond theory and synthetic instances, or the model itself enabling previously unimaginable applications in geoscience. In general, applied machine learning has developed into a well-established tool in computational geoscience and has the ability to throw new light on geoscience theory.

References

1. Bergen Karianne J, Johnson Paul A, de Hoop Maarten V, Beroza Gregory C (2019) Machine learning for data-driven discovery in solid Earth geoscience. Science. https://doi.org/10.1126/science.aau0323
2. Bergen, K.J., Johnson, P.A., Maarten, V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science, 363(6433), eaau0323. doi:10.1126/science.aau0323
3. Fox, P., Hendler, J., 2011. Changing the equation on scientific data visualization. Science, 331(6018), 705-708.
4. Gil, Y. and S. Pierce (Eds). "Final Report of the 2015 NSF Workshop on Information and Intelligent Systems for Geosciences." National Science Foundation Workshop Report, October 2015. Available from the NSF IIS collection at the ACM Digital Library: at http://dl.acm.org/collection.cfm?id=C13 and from http://is-geo.org/
5. "Dynamic Earth: GEO Imperatives and Frontiers 2015-2020." National Science Foundation, Advisory Committee for Geosciences, 2014.
6. Kawale, J., S. Liess, A. Kumar, M. Steinbach, P. Snyder, V. Kumar, A. R. Ganguly, N. F. Samatova, and F. Semazzi. "A graph-based approach to find teleconnections in climate data." Stat. Anal. Data Mining, 6, 158-179, 2013.
7. Peters SE, Zhang C, Livny M, Ré C. "A Machine Reading System for Assembling Synthetic Paleontological Databases." PLoS ONE 9(12), 2014.
8. Karpatne, A. et al. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Transactions on Knowledge and Data Engineering 29, 10 (2017) 2318--2331.
9. Wang, C., Ma, X., Chen, J., Chen, J., 2018. Information extraction and knowledge graph construction from geoscience literature. Computers & Geosciences, 112, 112-120.
10. Donoho, D. (2015). 50 years of Data Science. In Princeton NJ, Tukey Centennial Workshop. Retrieved from http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf
11. Tukey, J. W. (1962). The future of data analysis. The Annals of Mathematical Statistics, 33(1), 1–67. Retrieved from http://projecteuclid.org/euclid.aoms/1177704711
12. National Academies of Sciences, E., & Medicine. (2018). Data Science for Undergraduates: Opportunities and Options. Washington, DC: The National Academies Press. https://doi.org/10.17226/25104
13. Press, G. (2013). A Very Short History Of Data Science. Forbes. Retrieved from https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science
14. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin,S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow. org/. Software available from tensorflow.org.
15. F. Agterberg. Markov schemes for multivariate well data. In Proceedings, symposium on applications of computers and operations research in the mineral industries, Pennsylvania State University, State College, Pennsylvania, volume 2, pages X1–X18, 1966.
16. M. A. Aizerman. Theoretical foundations of the potential function method in pattern recognition learning. Automation and remote control, 25:821–837, 1964.
17. M. Alfarraj and G. AlRegib. Petrophysical property estimation from seismic data using recurrent neural networks. arXiv preprint arXiv:1901.08623, 2019.
18. F. Anifowose, C. Ayadiuno, and F. Rashedian. Carbonate reservoir cementation factor modeling using wireline logs and artificial intelligence methodology. In 79th EAGE Conference and Exhibition 2017-Workshops, 2017.
19. M. Araya-Polo, T. Dahlke, C. Frogner, C. Zhang, T. Poggio, and D. Hohl. Automated fault detection without seismic processing. The Leading Edge, 36(3):208–214, 2017.
20. S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27):7353–7360, 2016.
21. Y. Babakhin, A. Sanakoyeu, and H. Kitamura. Semi-supervised segmentation of salt bodies in seismic images using an ensemble of convolutional neural networks. German Conference on Pattern Recognition (GCPR), 2019.

22. C. Ballabio and S. Sterlacchini. Support vector machines for landslide susceptibility mapping: The staffora river basin case study, italy. Math. Geosci., 44(1):47–70, Jan. 2012. ISSN 1874-8961, 1874-8953. doi: 10.1007 s11004-011-9379-9. URL https://doi.org/10.1007/s11004-011-9379-9.
23. T. Bayes. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. Philosophical transactions of the Royal Society of London, (53): 370–418, 1763.
24. K. J. Bergen, P. A. Johnson, V. Maarten, and G. C. Beroza. Machine learning for data-driven discovery in solid earth geoscience. Science, 363(6433):eaau0323, 2019.
25. P. Bestagini, V. Lipari, and S. Tubaro. A machine learning approach to facies classification using well logs. In SEG Technical Program Expanded Abstracts 2017, SEG Technical Program Expanded Abstracts, pages 2137– 2142. Society of Exploration Geophysicists, Aug. 2017. doi: 10.1190/segam2017-17729805.1. URL https://doi.org/10.1190/segam2017-17729805.1.
26. M. Beyreuther and J. Wassermann. Continuous earthquake detection and classification using d iscrete hidden markov models. Geophys. J. Int., 175(3):1055–1066, Dec. 2008. ISSN 0956-540X. doi: 10.1111/j.1365-246X.2008.03921.x. URL https://academic.oup.com/gji/article-abstract/175/3/1055/634811.
27. M. Bicego, C. Acosta-Muñoz, and M. Orozco-Alzate. Classification of seismic volcanic signals using Hidden-Markov-Model-Based generative embeddings. IEEE Trans. Geosci. Remote Sens., 51(6):3400–3409, June 2013. ISSN 0196-2892. doi: 10.1109/TGRS.2012.2220370. URL http://dx.doi.org/10.1109/TGRS.2012.2220370.
28. H. Blondelle, A. Juneja, J. Micaelli, and P. Neri. Machine learning can extract the information needed for modelling and data analysing from unstructured documents. In 79th EAGE Conference and Exhibition 2017-Workshops. earthdoc.org, 2017. URL http://www.earthdoc.org/publication/publicationdetails/?publication=89273.
29. M. Blouin, A. Caté, L. Perozzi, and E. Gloaguen. Automated facies prediction in drillholes using machine learning. In 79th EAGE Conference and Exhibition 2017-Workshops. earthdoc.org, 2017. URL http://www.earthdoc.org/ publication/publicationdetails/?publication=89276.
30. Prabhu, A., Morrison, S.M., Eleish, A., Zhong, H., Huang, F., Golden, J.J., Perry, S.N., Hummer, D.R., Ralph, J., Runyon, S.E., Fontaine, K., 2021. Global earth mineral inventory: A data legacy. Geoscience Data Journal. In Press. doi:10.1002/gdj3.106.
31. Qiu, Q., Xie, Z., Wu, L., Li, W., 2019. Geoscience keyphrase extraction algorithm using enhanced word embedding. Expert Systems with Applications, 125, 157-169.
32. 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(7743), 195-204.
33. Spielman, S.J., Moore, E.K., 2020. dragon: A New Tool for Exploring Redox Evolution Preserved in the Mineral Record. Frontiers in Earth Science, 8, 585087. doi:10.3389/feart.2020.585087.
34. Stephenson, M.H., Cheng, Q., Wang, C., Fan, J., Oberhansli, R., 2020. Progress towards the establishment of the IUGS Deeptime Digital Earth (DDE) programme. Episodes Journal of International Geoscience, 43(4), 1057-1062.
35. Valentine, D., Zaslavsky, I., Richard, S., Meier, O., Hudman, G., Peucker-Ehrenbrink, B., Stocks, K., 2020. EarthCube Data Discovery Studio: A gateway into geoscience data discovery and exploration with Jupyter notebooks. Concurrency and Computation: Practice and Experience, In Press. doi:10.1002/cpe.6086.