1. Data science challenges in developing intelligent and information systems for geoscience

a) Presence of heterogeneity: Due to differences in geographies, topographies, climatic conditions, and vegetation types, different locations on the Earth show varying characteristics in geoscience datasets. Furthermore, the same locations can show varying characteristics at different times, due to the presence of Earth's seasonal cycles and inter-annual patterns. This heterogeneity in geoscience datasets restricts the global applicability of traditional big data analytics approaches that are designed to work with i.i.d data. Thus, there is a need to advance the state-of-the-art techniques in big data analytics for handling heterogeneity in geoscience datasets, as explored by recent advances in heterogeneous machine learning approaches, e.g. multi-task learning and multi-instance learning.

b) Uncertainty and incompleteness: Geoscience datasets are frequently plagued with noise/uncertainty and incompleteness due to sensor interference and instrument malfunctions. This issue is particularly acute in the case of remotely sensed land surface data, where atmospheric (clouds and other aerosols) and surface (snow and ice) interference are constantly encountered. The presence of a high degree of noise and missing values significantly impacts the performance of traditional big data analytics approaches that are designed to work with relatively noise-free datasets, e.g. advertising and social network datasets. This motivates the need for developing algorithms that are robust to presence of uncertainty and incompleteness in geoscience data.

c) Lack of representative ground truth: Obtaining gold-standard ground truth is often expensive in Earth science applications. Additionally, it is difficult to obtain representative training samples when the phenomena of interest shows class imbalance, e.g. detection of ultra-rare land cover changes such as forest fires. The paucity of representative training samples can significantly impact the performance of traditional data-centric approaches. Techniques are needed that can work in the absence of gold-standard ground truth, by making use of imperfect but abundant weak labels that are more readily available in a number of Earth science applications.

d) Multi-scale and multi-resolution: Naturally occurring geoscience phenomena occur at different scales and are observed in a variety of Earth science datasets that are available at varying spatial and temporal resolutions. There is a need for developing big data analytics approaches that can effectively exploit the complementary nature of multiple datasets at varying resolutions, instead of relying on a single dataset at a particular resolution.

2. Emerging research trend: Need for theory-guided data science

Advances in data science, the growth of datasets, and access to abundant computing facilities have given rise to the notion that any problem with a clear objective function can be solved given sufficient data. However, traditional big data analytics approaches that have been tremendously successful in the commercial domain (e.g. in finance, advertising, and social network analysis) are not directly suited for geoscience applications, due to their disregard to common domain knowledge. What is needed is an approach that leverages the advances in data-driven research yet constrains both the methods and the interpretation of the results through scientific principles that govern the domain. Thus, to make significant contributions to geoscience, new data science methods must encapsulate domain knowledge to produce theoretically consistent results. As an illustration, consider the problem of mapping the dynamics of water bodies at a global scale using remote sensing datasets. Given the heterogeneity in the land and water bodies globally and the poor quality of remote sensing datasets, even the state-of-the-art machine learning algorithms cannot provide adequate accuracy to map the dynamics of water bodies satisfactorily. However, the waterbeds of naturally occurring water bodies are generally concave in shape, which implies that locations at a lower elevation are more likely to be filled with water at a given time-step as compared to locations at a higher elevation. Elevation information, if available, can then be incorporated in the predictive model to account for the inadequacies of the state-of-the-art machine learning methods. However, obtaining accurate and precise elevation information is expensive, and existing global elevation datasets are of too coarse granularity to be of use in solving the problem. Hence, there is a need to develop novel predictive modeling techniques that can adhere to the physical constraints of water bodies even in the absence of elevation data.