Introduction to Analysis Methods for Big Earth Data

Christopher Lynnes[1]

[1]NASA/Goddard Space Flight Center (U.S. Government Civil Servant).

Corresponding author: Christopher Lynnes (christopher.s.lynnes@nasa.gov)

Key Points:

- Big Earth Data are too big to be tractable to simple data inspection and require models to make sense of all the data.
- Useful models for Big Earth Data may be physical, statistical, or machine learning based. In many cases, hybrid models combine attributes of two or more of these types.

The National Institute for Standards and Technology (NIST) defines Big Data as "extensive datasets--primarily in the characteristics of volume, velocity, variety, and/or variability--that require a scalable architecture for efficient storage, manipulation, and analysis" (Chang et al., 2018). The Volume aspect of this definition has so far dominated the work in Earth Science Informatics area (though as time goes on and data sources multiply, variety and variability may well surpass the volume challenge). What does this mean for data analysis of Big Earth Observations?

Before the advent of computers, the only practical way to analyze data was through visual inspection, either of the individual data points or in some cases relatively simple statistical models. However, human vision can only scale so far, particularly when limited to static, two-dimensional representations of data, such as science papers. The most common approach to addressing this limitation is to employ some kind of model-based analysis. Model-based analysis can be viewed as three different classes of model: physical, statistical, and machine-learning (Fig. 1), but as we will see, these are not necessarily disjoint.

Physical models are developed based on our understanding of the physical world and its workings. This includes geophysical, biological and ecological models. Observed data may be used in a number of different ways in this type of model: as initial or boundary conditions for the model; directly assimilated into the model itself; or as validation or falsification of the model. As such, particularly in the latter case, analysis based on physical models can be a particularly stringent test of our knowledge of the physical aspects being investigated. However, physical models require a thorough knowledge of underlying principles, which may not be available for studies in new or poorly understood areas. This is particularly the case in Earth Observation as instruments are able to produce data on finer spatial and temporal scales, where the physics may not be well enough known to incorporate into a model. Physical models are also typically complex computer programs which are susceptible to bugs and can be difficult to reproduce by others, unless the model code is both published and portable.

Statistical techniques range from computing simple averages and standard deviations to sophisticated, compute-intensive techniques such as kriging and wavelet analysis. Although not always obvious, these techniques usually rest on underlying assumptions of how the data are likely to be distributed, such as the Gaussian distribution assumed to be the premise for simple averaging. Statistical models can also be powerful tools for exploring the important characteristics of datasets, particularly when we move beyond simple averaging to look at other aspects of the data, such as extreme values, as discussed in the following chapter on "An Unsupervised Anomalous Event Detection and Interactive Analysis Framework for Large-scale Satellite Data". However, if the distribution of the data deviates significantly from the assumed statistical model, statistical analysis can produce misleading results. Also, since most statistical methods compute aggregate quantities to reduce the data volume, the aggregate estimates may fail to find important features in the data. The geospatial nature of Earth Observation statistics represent yet another complication, which is treated in the following chapter on "Spatial statistics for big data analytics in the ocean and atmospheric realm: techniques, examples, and challenges".

Machine learning models are in some ways an extension of statistical models. Indeed, some of the earliest forms of machine learning, such as logistic regression and Naive Bayesian classification, have an obvious basis in statistics. While machine learning algorithms can in theory consume a virtually unlimited amount of data, in reality adding more data to a machine learning model with high bias (underfitting) does little to improve the model. Another key

challenge with the use of Machine Learning Models in data analysis is that many of the models are purely mathematical, with no obvious connection to the underlying physics. As a result, efforts toward "interpretable machine learning" have sprung up in a number of fields.

Perhaps the most knowledge can be gleaned from Big Earth Data via hybrid models. For instance the following chapter "Mapping Surface Water Dynamics on a Global Scale using Data from Earth Observing Satellites and Machine Learning" uses physical constraints to detect and correct for errors in the machine learning results. The chapter on "An Unsupervised Anomalous Event Detection and Interactive Analysis Framework for Large-scale Satellite Data" uses a combination of statistical and machine-learning-based techniques to identify anomalies that are likely to represent interesting physical phenomena.

References

Chang, W.L., 2015. NIST Big Data Interoperability Framework: Volume 1, Definitions (No. Special Publication (NIST SP)-1500-1).

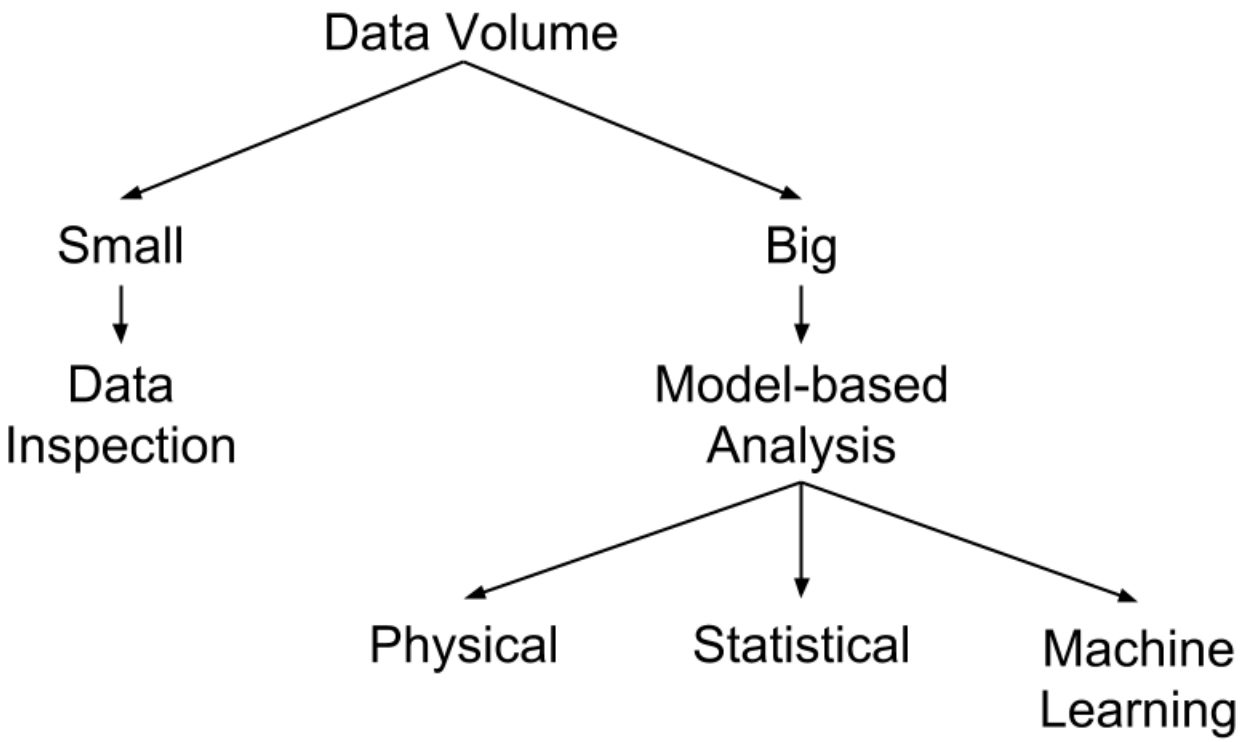

Fig. 1. Big Data volumes are too large to be tractable purely to inspection of the data. Instead, analysis must incorporate some kind of model--physical, statistical or machine-learning--in order to make most sense of the data.