

PRACTICE PAPER

Earth Science Data Analytics: Definitions, Techniques and Skills

Steve Kempler¹ and Tiffany Mathews²¹NASA Goddard Space Flight Center, US²NASA Langley Research Center, US

Corresponding author: Steve Kempler (Gulliver2100@verizon.net)

The continuous evolution of data management systems affords great opportunities for the enhancement of knowledge and advancement of science research. To capitalize on these opportunities, it is essential to understand and develop methods that enable data relationships to be examined and information to be manipulated. Earth Science Data Analytics (ESDA) comprises the techniques and skills needed to holistically extract information and knowledge from all sources of available, often heterogeneous, data sets. This paper reports on the ground breaking efforts of the Earth Science Information Partners (ESIP) ESDA Cluster in defining ESDA and identifying ESDA methodologies. As a result of the void of Earth science data analytics in the literature the ESIP ESDA definition and goals serve as an initial framework for a common understanding of techniques and skills that are available, as well as those still needed to support ESDA. Through the acquisition of Earth science research use cases and categorization of ESDA result oriented research goals, ESDA techniques/skills have been assembled. The resulting ESDA techniques/skills provide the community with a definition for ESDA that is useful in articulating data management and research needs, as well as a working list of techniques and skills relevant to the different types of ESDA.

Keywords: Data Science; Analytics; Techniques; Skills; Information; Knowledge

Introduction

The continuous evolution of data management systems affords great opportunities to the enhancement of knowledge and advancement of Earth science research. With the growing need and desire to leverage information from various sources to better understand our environment, it becomes evident – through community experience and foresight – that this can be maximized by accepting new ways to cross-examine this information. As excerpts in the ‘4th Paradigm’ (Hey, Tansley & Tolle 2009) explain:

‘We have to do better at producing tools to support the whole research cycle—from data capture and data curation to data analysis and data visualization’. (xvii)

‘Clearly, data-intensive science. . . must move beyond data warehouses and closed systems, striving instead to allow access to data to those outside the main project teams, allow for greater integration of sources, and provide interfaces to those who are expert scientists but not experts in data administration and computation’. (147)

‘We are already seeing some attempts to infer knowledge based on the world’s information’. (167)

To capitalize on these opportunities, it is essential to develop a data analytics framework in which we scope the scientific, technical, and methodological components that contribute to advancing science research. Through this framework, we can categorize discussions amongst individuals of like component

interests, instead of attempting to draw specific direction from a set of starting points that may greatly vary.

Data Analytics is the process of examining large amounts of data of a variety of types to reveal hidden patterns, unknown correlations, and other useful information, key to facilitating Earth science research opportunities. Thus, the research presented here is motivated by the need to determine and categorize available data analytics techniques and skills and to identify the gaps for where they are still needed.

Today, and well into the foreseeable future, there is a rapid growth in the amount of Earth science data and value-added heterogeneous information that many Earth science researchers have not yet holistically leveraged. It is important to realize that this rate of data growth is new and challenging. It is new in that information technology is just beginning to provide the tools for advancing the analysis of heterogeneous datasets in a 'big' way to provide opportunities to discover unobvious scientific relationships, previously invisible to the science eye. The challenge is it takes individuals, or teams of individuals, with just the right combination of skills to understand the data and develop the methods to glean knowledge from data and information.

The ability to apply information technology, tools, and services necessary to facilitate the advancement of Earth science research is becoming more obvious and necessary at a rate that is accelerating. That is, if data manipulation (subsetting, data transformation, format conversion, etc.) extract information from data, then data analytics techniques and skills glean knowledge from information (Kempler, 2014).

The objectives of the Earth Sciences Information Partners (ESIP) Federation Earth Science Data Analytics (ESDA) Cluster is to understand, define, and facilitate the implementation of Earth science data analytics. As a result of cluster efforts, an ESIP adopted definition of ESDA has been generated, along with 10 ESDA goals, to set the framework for a common understanding for advancing Earth science research. In addition, the ESIP ESDA cluster performed an exhaustive search identifying ESDA types, techniques, and skills used in performing data analytics.

In this paper, we present the ESIP derived definition of ESDA and differentiate it from other publicized definitions of data analytics. We then describe different types of ESDA and their driving goals. This is followed by an exhaustive survey of current techniques and skills for performing ESDA, made available for the benefit of Data Scientists exploring new Earth science data analytics methodologies, and their potential use.

Literature Review

The advancement of information use resulting from evolving technologies, newly developed techniques, and refined skills has become the purview of the Data Scientist performing data analytics. Data analytics got its start in the business world which is why most literature and developed tools reflect back on business as the primary application. In the literature, we find that data analytics is comprised of 5 types: Descriptive, Diagnostic, Discoverative, Predictive, Prescriptive. When the ESIP ESDA Cluster attempted to categorize Earth science research use cases into these data analytics types, the use cases did not fit: categorizing was ambiguous and/or they fit in more than one type category. Where business data analytics types reflect looking for patterns, and predicting (and prescribing) actions, Earth science data analytics also include assessing, validating, calibrating, and applying techniques required to prepare raw datasets for co-use. In addition, characteristics of Earth science data introduces data analytics challenges such as dealing with differing formats, differing spatial and temporal data resolutions, inconsistent data acquisition techniques and units for the same measurement, noise, biasing, to mention a few. This led to the need for a data analytics definition directed specifically at Earth science research goals.

In addition, insights like: 'Researchers in science must work with colleagues in computer science and informatics to develop field-specific requirement' (Hey, Tansley & Tolle 2009: 151) and 'The process of combining information from existing scientific knowledge . . . including the specific methodologies that were followed to produce conclusions, should be automatic and implicitly supported' (Hey, Tansley & Tolle 2009: 170-171), are the genesis for melding the expanding manipulation of information and associated technologies, with physical science.

A significant aspect of ESDA is information literacy, the ability to "recognize when information is needed and have the ability to locate, evaluate, and use information effectively" (American Library Association, 1989). Now that we are well into the information age, ensuring the responsible and appropriate use of data and information has become extremely important and a key skill for ESDA. The Framework for Information Literacy for Higher Education (American Library Association, 2015)

provides groundwork for learners to understand and fully appreciate the value of information used when performing ESDA.

As seen in the literature, there is no shortage of data analytics definitions, and descriptions of individuals who performs data analytics, the Data Scientist. The Booz/Allen/Hamilton (B/A/H) Report, 'The Field Guide to DATA SCIENCE' (2015) reminds us that 'the term Data Science appeared in the computer science literature throughout the 1960s-1980. It was not until the late 1990s however, that the field, began to emerge from the statistics and data mining communities' (B/A/H 2015: 21). The B/A/H Report (2015) further states: 'Performing Data Science requires the extraction of timely, actionable information from diverse data sources to drive data products'.

The National Institute of Standards and Technology provides the following definitions (NIST 2015):

- Data science is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing.
- A data scientist is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes in the data life cycle.
- The analytics process is the synthesis of knowledge from information.

The article, '8 skills you need to be a Data Scientist' (Udacity 2014) defines the skills in **Table 1**.

The website, Master's in Data Science (2016) summarizes technical skills and tools needed of a Data Scientist to perform data analytics (**Table 2**).

Skills of a Data Scientist (Udacity)	
Basic Tools	Data Munging
Basic Statistics	Data Visualization & Communication
Machine Learning	Software Engineering
Multivariable Calculus and Linear Algebra	Thinking Like a Data Scientist

Table 1: Skills of a Data Scientist (Udacity).

Technical skills and tools of a Data Scientist (Master's in Data Science)

Math (e.g. linear algebra, calculus and probability)

Statistics (e.g. hypothesis testing and summary statistics)

Machine learning tools and techniques (e.g. k-nearest neighbors, random forests, ensemble methods, etc.)

Software engineering skills (e.g. distributed computing, algorithms and data structures)

Data mining

Data cleaning and munging

Data visualization (e.g. ggplot and d3.js) and reporting techniques

Unstructured data techniques

R and/or SAS languages

SQL databases and database querying languages

Python (most common), C/C++ Java, Perl

Big data platforms like Hadoop, Hive & Pig

Cloud tools like Amazon S3

Table 2: Technical skills and tools of a Data Scientist (Master's in Data Science).

Science Research Technologies (Sampling)		
In Atmospheric Research		In Hydrology Research
Correlation Analysis; Bias Correlation	Spectral Analysis	Linear Regression
Regression Analysis; Bivariant Regression	Temporal Trending; Trend Analysis	Monte Carlo
Decision Tree	Spatial Interpolation	Darcy Equation
Machine Learning	Revised Averaging Scheme	Poisson Regression
Data Mining	Forward Modeling; Inverse Modeling	Multi-variate time series analysis
Data Fusion	Radiative Transfer Model	BUDYKO formula
Computational Tools	Baysian Synthesis Inversion	Smoothing (Gaussian)
Constrained Variational Analysis	Temporal Stability	Filtering (Destriping)
Model Simulations	Gaussian Distribution	MESH Model
Ratios	Exponential Differentiation	
Time Series Analysis		

Table 3: Sampling of science research techniques being used.

In addition, the McKinsey Global Institute (2011), National Research Council (2013), and the ongoing on-line blog ‘What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, and Big Data?’ (Quora) provide further insights into the current thinking that defines data analytics.

‘Doing Data Science’ (O’Neil & Schutt 2014) found that most Data Scientist job descriptions ask for expertise in ‘computer science, statistics, communication, data visualization, *and* to have extensive domain expertise’.

Much information regarding data analytics techniques and skills have been found in presentations given at informatics forums such as the American Geophysical Union (AGU) Earth and Space Science Informatics (ESSI), ESIP, and SciDataCon. In addition to forum sessions being rich with experienced individuals who describe techniques and skills that facilitate data analytics, the authors of this paper have also taken the opportunity, at AGU, to visit science poster presentations to better understand research methodologies (analytics) utilized. At these meetings, research that involved co-analysis of multiple datasets were sought out. After scanning hundreds of presentations for methodologies used, 31 atmospheric science focused (study of gases) and 12 hydrologic science focused (study of liquid) presentations were targeted in which research techniques were identified (**Table 3**).

This small study opens our eyes to the data analytics techniques pertaining specifically to data analysis. Yet many of these techniques also find homes in performing data preparation and data reduction analytics.

Defining Earth Science Data Analytics

For clarity, data analytics activities fall within the scope and expertise of the Data Scientist. Data Scientists study and develop methods for analyzing, storing, and presenting data. When they practice their skills on specific problems, they are performing data analytics, applying tools and techniques to co-analyze heterogeneous data. Data Scientists, as researchers, developers, or data analytics practitioners, require similar skill sets.

Through literature research, analysis of Earth science research use cases, and integration of the science research methods, the ESIP Federation – a collaborative organization of over 170 information-centric partners – has defined and adopted the following definition of ESDA (Kempler & Mathews, 2016):

The process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data, encompassing varieties of data types. ESDA enables people to uncover patterns, correlations, and other information, to better understand our Earth.

ESDA Goals

- To calibrate data
- To validate data (note it does not have to be via data intercomparison)
- To assess data quality
- To perform coarse data preparation (e.g. subsetting data, mining data, transforming data, recovering data)
- To intercompare datasets (i.e. any data intercomparison; Could be used to better define validation/quality)
- To tease out information from data
- To glean knowledge from data and information
- To forecast/predict/model phenomena (i.e. Special kind of conclusion)
- To derive conclusions (i.e. that do not easily fall into another type)
- To derive new analytics tools

Table 4: Earth Science data analytics goals.

This includes:

- **Data Preparation** – Preparing heterogeneous data so that they can be jointly analyzed.
- **Data Reduction** – Correcting, ordering, and simplifying data in support of analytic objectives.
- **Data Analysis** – Applying techniques/methods to derive results.

Data preparation includes the methods and techniques that uncover, discover, extract data of greatest interest. This can involve filtering, mining, format conversion, smoothing, visualization, etc. Data Reduction addresses very large amounts of heterogeneous data that face Earth science research. Several methods for the purpose of data reduction can be applied, with the goal of making data transfer, computation, and analysis easier and/or more focused. Analytics to perform Data Analysis is not as clear cut. Science, by its nature, often utilizes technologies that are not decided upon until data/information is initially looked at and better understood. It is then that researchers experiment with existing and/or novel analytics techniques. In our paradigm, Data Analysis analytics includes all aspects of science research: Hypothesis and data discovery driven methods, as well as goal driven decisions, outcomes, and impacts. Data Analysis analytics aspects can be categorized separately, if desired.

ESDA is categorized by the goals of the analytics performed (**Table 4**).

With a definition of ESDA, we now have an initial framework for which to speak the same 'language', discuss ESDA scope, infrastructure, and methodologies with a common focus, and to grow upon.

Earth Science Data Analytics Techniques and Skills

ESDA techniques are considered to be computational methods. Repeatedly, individuals seeking to perform Earth science express their need to utilize mathematics, numerical modeling, statistics, software engineering and the ability to integrate data from across multiple domains. Also, there is a need for expertise in techniques, such as: rule learning, classification, cluster analysis, data fusion, machine learning, neural networks, anomaly detection, modeling, time series analysis, and visualization.

In addition, many other computational methods have been identified as potential techniques to be used in performing ESDA. **Table 5** provides a broad scope of the possibilities associated with the three types of ESIP ESDA: Data Preparation, Data Reduction, Data Analysis.

ESDA skills are considered to be the ability to apply techniques. In regards to Earth science, this refers to applying ESDA techniques to Earth science domains being studied. Thus, ESDA skills include knowledge in particular Earth science domains where data analytics can advance the understanding of science.

ESDA skills also refer to the ability to facilitate making data useful. This includes understanding the relevance of: the data lifecycle, data structures, metadata, data integration, and data interpretation.

Table 6 summarizes skills that are useful in practicing ESDA.

Earth Science Data Analytics Techniques		
Data Preparation	Data Reduction	Data Analysis
Bias Correction	Aggregation	Anomaly Detection
Coordinate Transformation	Anomaly Detection	Bayesian Techniques
Data Engineering	Cluster Analysis	Bivariate Regression
Data Mining	Data Engineering	Classification
Data Munging	Data Fusion	Correlation/Regression Analysis
Database Management	Factor Analysis	Factor Analysis
Exponential Differentiation	Filtering	Fourier Analysis
Filtering	Neural Networks	Gaussian Distribution
Format Conversion	Outlier Removal	Graphics Analysis
Imputation	Ratios	Imputation
Normalization/Transformation	Revised Averaging Scheme	Linear/Non-linear Regression
Outlier Removal	Rule Learning	Machine Learning/Decision Tree
Ratios	Time Series	Mathematics/Calculus
Rule Learning	Visualization	Modeling
Sensitivity Analysis		Monte Carlo Method
Smoothing		Multi-variate Time Series
Spatial Interpolation		Normalization
Time Series		Pattern Recognition
Visualization		Principal Component Analysis
		Revised Averaging Scheme
		Rule Learning
		Signal Processing
		Spectral Analysis
		Statistics
		Temporal Trend Analysis
		Time Series
		Visualization

Table 5: Earth science data analytics techniques (sampling).

Earth Science Data Analytics Skills

- Ability to integrate data across multiple domains
- Support domain scientists with data & computational knowledge
- Communicate across domains
- Knowledge of data cycle
- Software engineering
- Software programming
- Data Engineering
- Decision science

Table 6: Earth science data analytics skills (sampling).

In short, ESDA techniques and skills need to be interdisciplinary from the start. One needs to know what domain specific information is available, where to get it, how it is generated, as well as statistical, mathematical, and computational methods to manipulate it.

Conclusions

Although data analytics definitions and types that are oriented at business are well documented, data analytics to facilitate the inter-analysis of large heterogeneous Earth science datasets has only begun to be addressed methodically. The significance of the development of a set of definitions, types, goals, techniques, and skills that target ESDA specifically provides Earth scientists the opportunity to better articulate the techniques and skills they employ in furthering their science research. In particular, now communications can be performed in terms that engage information technologists who can provide support by implementing responsive tools. With a categorization of known techniques and skills associated with data preparation, data reduction, and data analysis analytics, we know what techniques and skills are presently available, what tools have been implemented that perform these techniques, and what tool gaps need to be filled as science research methodologies evolve.

Next steps include: Engaging the Earth science research community, to better understand their research methodologies and share information technologies that may be useful; Engaging scientists to acquire additional use cases to further validate and update our knowledge of known ESDA techniques and skills; Continuing to refine our understanding of the skills needed to perform ESDA; Promoting the development of ESDA techniques and skills through university curriculums, and; Addressing the 'moving' gap analysis.

Acknowledgements

The authors would like to thank the ESIP Federation and, in particular, the ESDA Cluster members for their support and insights, as well as Peter Fox and Erin Robinson for their wisdom and encouragement.

Competing Interests

The authors have no competing interests to declare.

References

- American Library Association** 1989 Presidential Committee on Information Literacy: Final Report. Available at: <http://www.ala.org/acrl/publications/whitepapers/presidential>.
- American Library Association** 2015 Framework for Information Literacy for Higher Education. Available at: <http://www.ala.org/acrl/standards/ilframework>.
- ESIP** Earth Science Data Analytics ESDA – Earth Sciences Information Partners (ESIP). Available at: http://wiki.esipfed.org/index.php/Earth_Science_Data_Analytics.
- Hey, T, Tansley, S and Tolle, K** (eds.) 2009 *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.
- Kempler, S** 2014 How Do We Facilitate the Use of Large Amounts of Heterogeneous Data, In: NASA/GSFC Information Science and Technology Colloquium Series.
- Kempler, S and Mathews, T** 2016 In: *Earth Science Data Analytics (ESDA) Telecon XX*. Earth Science Data Analytics Cluster.
- Master's in Data Science** 2016 Complete Directory of Data Science Graduate Degrees. Available at: <http://www.mastersindatascience.org/schools/>.
- McKinsey Global Institute** 2011 *Big data: The next frontier for innovation, competition, and productivity*. Available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.
- National Research Council** 2013 *Frontiers in Massive Data Analysis*. The National Academies Press, Washington, D.C.
- NIST Big Data Interoperability** 2015 NIST Big Data Public Working Group Definitions and Taxonomies Subgroup. *Framework: Volume 1, Definitions, NIST Special Publication 1500-1*.
- O'Neill, C and Schutt, R** 2014 *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media, ISBN: 978-1-449-35865-5.

Quora What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, and Big Data? (ongoing blog). Available at: <https://www.quora.com/What-is-the-difference-between-Data-Analytics-Data-Analysis-Data-Mining-Data-Science-Machine-Learning-and-Big-Data-1>.


The Booz/Allen/Hamilton (B/A/H) Report 2015 *The Field Guide to DATA SCIENCE, second edition*.

Udacity 2014 8 skills you need to be a Data Scientist. Available at: <http://blog.udacity.com/2014/11/data-science-job-skills.html>.

How to cite this article: Kempler, S and Mathews, T 2017 Earth Science Data Analytics: Definitions, Techniques and Skills. *Data Science Journal*, 16: 6, pp. 1–8, DOI: <https://doi.org/10.5334/dsj-2017-006>

Submitted: 27 October 2016 **Accepted:** 16 January 2017 **Published:** 24 February 2017

Copyright: © 2017 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 