DataScience
foundation

A new Data Science Framework for Analysing and Mining Geospatial Big Data

Author, Charith Silva

A Data Science Foundation White Paper

August 2018

----

www.datascience.foundation

A new Data Science Framework for Analysing and Mining Geospatial Big Data

ABSTRACT

Geospatial Big Data analytics are changing the way that businesses operate in many industries. Although a good number of research works have reported in the literature on geospatial data analytics and real-time data processing of large spatial data streams, only a few have addressed the full geospatial big data analytics project lifecycle and geospatial data science project lifecycle. Big data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processed. One of a motivation of introducing new framework is to address these big data analysis challenges. Geospatial data science projects differ from most traditional data analysis projects because they could be complex and in need of advanced technologies in comparison to the traditional data analysis projects. For this reason, it is essential to have a process to govern the project and ensure that the project participants are competent enough to carry on the process. To this end, this paper presents, new geospatial big data mining and machine learning framework for geospatial data acquisition, data fusion, data storing, managing, processing, analysing, visualising and modelling and evaluation. Having a good process for data analysis and clear guidelines for comprehensive analysis is always a plus point for any data science project. It also helps to predict required time and resources early in the process to get a clear idea of the business problem to be solved.

Keywords

Big data, Geospatial big data, Data Science, Machine learning, Data mining.

1. INTRODUCTION

“Big data” is defined as high volume, high velocity and high variety of data that cannot be stored, managed and processed by the traditional tools. It requires a new way of storing, managing and processing to enable insight discovery, decision making and process optimization [1]. Much of the industry follows Gartner's '3Vs' model to define Big Data. The 3Vs include Volume (huge amount of data is generated every second), Velocity (data are growing and changing in a rapid way), Variety (data come in multiple format). Big data' could be found in three forms: Structured, Un-structured, Semi-structured (text, sensor data, sound, video, clickstream data, log files, etc.). The structured Data are the data which could be stored in the relational database table in a row and column format and they have some specific structure and that structure defines by the data models. The Semi - Structured Data are in the form of structured data but would not fit with the data models which define the structured data. Also the semi structured data cannot be stored in the relational databases or the other form of a data tables, but they can be stored in some specific type of files which contain some tags [2]. The unstructured Data are the data which do not have any specific structure. Therefore, they cannot be stored in a row-column format of a traditional database. Ex: text files, image files, audio files, video files, web pages etc. Also, there are some examples for human-generated unstructured data as well. Eg: Social media data, Mobile data, website content data. The unstructured data are growing more rapidly than the other data types, and their exploitation could help in business decisions.

Geospatial big data refer to spatial data sets exceeding capacity of traditional computing systems. They have always been big data and size of such data is growing rapidly every year. The rapid growth of geospatial data and wide use of them emphasize the importance of the automated discovery of geospatial knowledge. A spatial referencing system consists of location coordinates of geographic space which are things and objects. These things and objects are called, geospatial data. It is hard to analyse high volumes of geospatial data with traditional data mining techniques, which are beyond the storage capacity.

Geospatial data science is the process of finding interesting and previously unknown spatial patterns. But the complexity of the spatial data sets, limit the usefulness of the traditional data mining techniques to extract spatial patterns. The geospatial data science is a combination of the spatial data mining and spatial machine learning and spatial statistics. Many advanced GIS tools are available in the market to extract useful information from geospatial data. It is important for some organizations to make decisions based on these large sets of geospatial data eg: Emergency services, Utility companies, Transport network, and research institutes, etc.…

Recently the geospatial data science has become a highly demanding field because large amounts of spatial data have been collected in various applications from remote sensors. The tremendous growth of the spatial data and the widespread use of spatial databases need a computerized discovery of spatial knowledge. In Spatial data may include two different types of features: non-spatial features and spatial features. The non-spatial features are used to characterize non-spatial characteristics of objects, such as the name, population and unemployment figures for a city. Spatial features are used to define the spatial location and size of space objects. The spatial attributes of a space object often contain information about spatial locations such as latitude, latitude [3].

2. BACKGROUND

The increasing volume and variety of data in the geospatial data collected from various sources, create new challenges of data storage, data management, data processing, verification of data quality, data analysis and visualization. Geospatial data processing techniques, machine learning techniques, spatial knowledge discovery methods and big data analytics techniques can be useful to develop new big data environment for geospatial data processing and for the improvement of geospatial data analysis. To discover knowledge from huge volumes of geospatial data, techniques like Geo-computing, data mining, simulation, statistical analysis can be applied inpidually or combining together [3].

Geospatial big data can't be stored, managed and processed by the traditional tools. It requires a new way of processing to enable enhanced decision making, insight discovery and process optimization [1]. Apache Hadoop and Apache Spark are the tool introduced by Apache to handle large quantities of data. They provides the platform of parallel computing for big data applications. Both Spark and Hadoop MapReduce are open source and free to use. There are two common approaches in big data processing: batch processing and stream processing (real time data processing). Batch processing is an effective way of processing high volumes of data where a group of transactions is collected over a period of time and process periodically.

3. REQUIREMENTS AND CHALLENGES

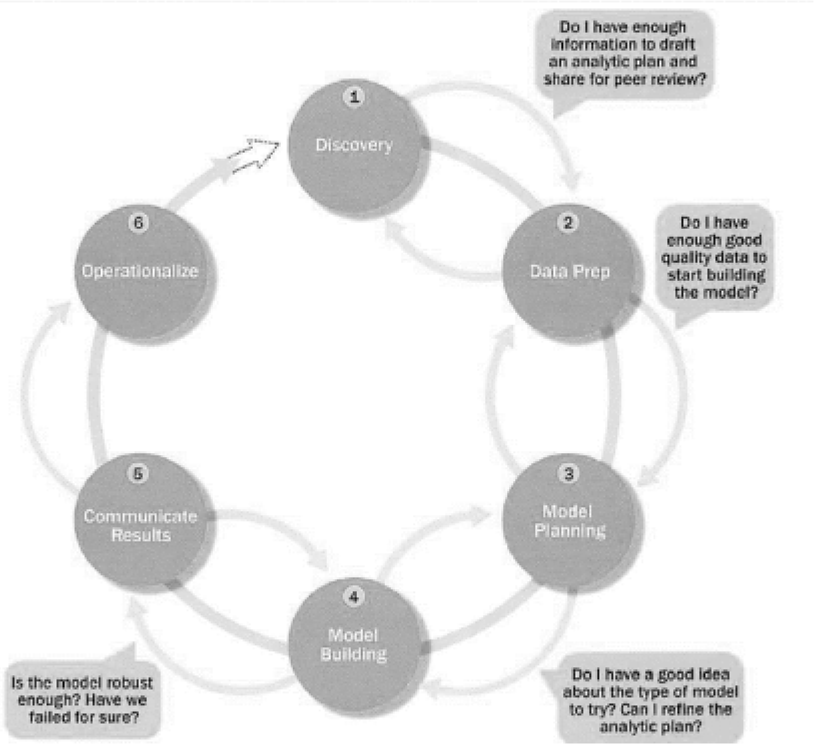
There are several challenges exist in geospatial Big Data processing, including data capturing , data transfer, data store, clean, analyse, filter, search, share, secure and visualize. Collecting and storing big data is one of the main problems in this area. Geospatial data are usually collected using ground surveying, photogrammetry and remote sensing, and more recently through laser scanning, mobile mapping, geo-located sensors, geo-tagged web contents, volunteered geographic information (VGI), global navigation satellite system (GNSS) tracking. Geospatial big data, with its defining characteristics of being large (voluminous), heterogeneous (variety), real-time processed (velocity), inconsistent (variability), and thus also of variable quality (veracity), must suffer even more from uncertainty, asynchronicity, and incompleteness [4].

Spatial statistical analysis, geo-informatics, spatial data mining methods, spatial machine learning techniques and other spatial techniques can be used to uncover knowledge from large amounts of geospatial data. An extensive analysis of spatial big data and analysis of traditional spatial data and geo-processing methods and theories can contribute to the development of large-scale analysis of geospatial data and processing framework.

Geospatial data science projects differ from most traditional data analysis projects because it can be complex and need advanced technologies comparing with traditional data analysis projects. For this reason, it is essential to have a process to govern the project and ensure that the project participants are competent enough to carry on the project. Many geospatial data science related problems that appear enormous and daunting at first can be broken down into smaller pieces or actionable phases that can be more easily addressed. Having a good data analysing and processing framework and clear guidelines ensures a comprehensive of conducting analysis. Having a good process for data analysis and clear guidelines for comprehensive analysis is always plus point for any data science project. It also helps to focus required time and resources early in the process to get a clear idea of the business problem to be solved.

4. RELATED WORK

Dietrich et al. (2015) introduce a new big data Analytics Lifecycle defines analytics process best practices [5]. The lifecycle draws from well recognised methods in the realm of data analytics. This mixture of process and flow was developed after gathering input from data scientists and consulting established approaches that provided input on pieces of the process. The figure 1 presents an overview of the Data Analytics Lifecycle that includes six phases. Data science or data analytic teams commonly learn new things in a phase that cause them to go back and refine the work done in prior phases based on new insights and information that have been uncovered. This Data Analytics Lifecycle is widely used in the big data projects and it has logical and sequence and repetitive steps which is sophisticated enough for big data project. But when it comes to geospatial big data project, handling spatial data and spatial data processing and analytics is not discussed or not considered in this process. Also, methods and challenge in related to acquiring data from perse sources and data storage also not considered in this framework.



Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
5
Communicate Results
Is the model robust enough? Have we failed for sure?
6
Operationalize

Figure 1: Data Analytics Lifecycle (Source: Dietrich, 2015)

Bogorny et al. (2007) has proposed an integrated spatial data mining framework which is interoperable with any geographic information system developed under Open GIS Consortium specifications. The framework is composed of three abstraction levels: data mining, data preparation, and data repository. This work presented a solution for the problem of automatic geographic data pre-processing for data mining. Main feature of this framework is the ability to accrue data from different geographic data storage methods under the geographical database management systems such as PostGIS, Oracle,etc.

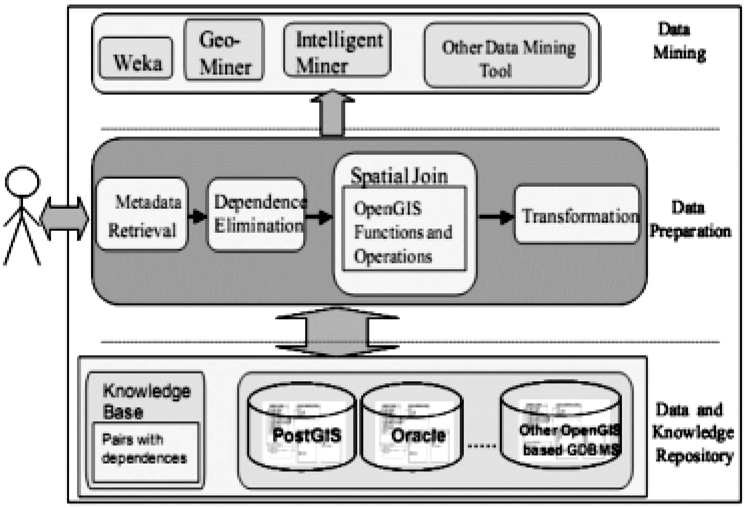


Figure 2. Integrated spatial data mining framework (Source: Bogorny, 2007)

It also contains data mining algorithms for the task of extracting the needed knowledge out of the GIS databases. Even this model has greatly helped to build the new framework,, it does not explain data accusation methods or spatial analytic techniques other than data mining. Also this framework is not addressing big data processing challenges, therefore this model is not suitable for to process highly volatile spatial data such as geospatial big data.

5. PROPOSED FRAMEWORK ARCHITECTURE

One of a motivation of building this new framework is to address geospatial big data analysis changes. Recent improvements in technology demand real-time spatial data processing and analytics and visualization to gain completive advantage of real-time decision making. After carefully examination and analysis of the above literature, there are a variety of issues in Geospatial Big Data processing and analysis. Therefore this research present new geospatial Big Data analytics and processing framework for geospatial data acquisition, data fusion, data storing, managing, processing, analysing, visualising and modelling.

Often the purpose of spatial analysis is not only to identify pattern, but to build models, if possible by gaining an understanding of process. We believe that without a proper coordination and structuring framework there is likely to be much overlap and duplication amongst project phases, and can cause confusion around the responsibilities of each project participant.

A common mistake made in geospatial big data projects is rushing into data collection and data analysis, which prevents spending adequate time to plan the amount of work involved in the project, understanding business requirements, or even defining the business problem properly. Geospatial big data has is available all around us in various formats, shapes and sizes. Understanding the relevance of each of these data sets to business problem is a key aspect to succeed with the project.

Also geospatial big data has multiple layers of hidden complexity that are not visible by simply inspecting. Poorly planned project can ruin entire project and the finding of the project in any organization. If the project does not clearly identify the appropriate level of complexity and the granularity, then the chances are high an erroneous result set will occur that twists the expected analytical outputs. In this research, we concentrate only on developing a big data environment for geospatial data mining and machine learning . In which the data can be managed in the distributed environment to store huge data. Big data environment for analyzing geospatial data provides the ability to deal with geospatial data on a large scale.

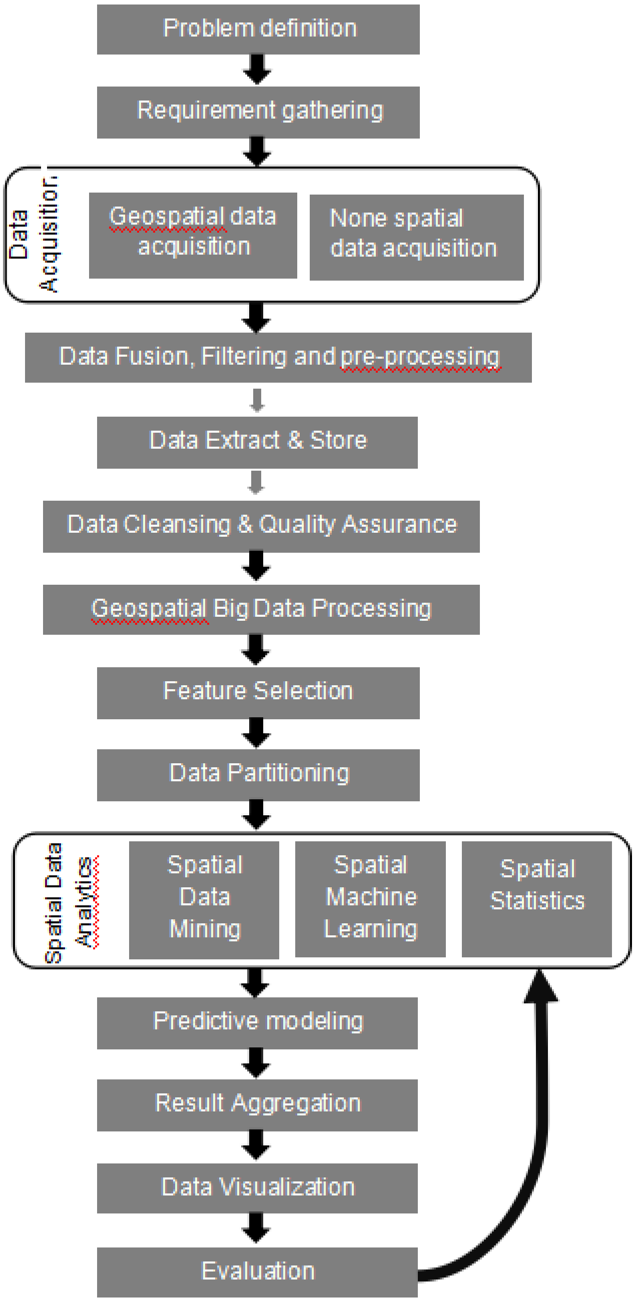


Figure 3. Proposed geospatial big data mining and machine learning framework for Data Science projects

1. *Problem definition*
   A problem definition statement is a brief description of the issues that need to be addressed by the project. It is used to focus the team from the project start and keep the team on track during the project life time, and also it can be used to validate the result. The (5W+H) method can be used to define the problem, because it is simple and easy to understand. The

five W's and the H are acronyms forWho? What? Where? When? Why? And How? It is a good tool for gathering information methodically in a problematic situation.

2. *Requirement gathering*

   Gathering business requirements is a critical initial step for any kind of project. It is an essential part of any project and project management. Poor requirements gathering techniques are the cause of many project failures. Creating a comprehensive set of requirements in beginning of the project will enable accurate estimates of costs, shorter delivery times, increase customer satisfaction, and increase the accuracy of the of the final product or solution. It is always better to avoid talking technology or solutions until the requirements are fully gathered and understood by the participants. It is important to create a clear, summarised and thorough requirements document and share it with the project participant.
3. *Data acquisition*

   This framework suggests to use both spatial as well as non-spatial data, which makes it more beneficial over the traditional data analysis.
   1. *Geospatial data acquisition*

      The acquisition of sophisticated and feature rich geospatial dataset is essential to modern day geospatial data science projects and applications. Data collection and the maintenance of spatial databases is the most expensive and time-consuming aspect of geospatial projects. Geographic data may be available in a variety of file formats and it might be very challenging to acquire correct type of data for the project.
   2. *Non-Geospatial data acquisition*

      Integrated analysis of spatial data and none spatial data from multiple data sources improve capability to identify hidden spatial patterns, trends, and relationships. Therefore, it’s always good idea to acquire relevant none spatial data for further analysis.
4. *Data Fusion, Filtering and pre-processing*
   1. *Data Fusion*

      The aim of a data fusion process is to maximize the useful information content acquired by heterogeneous sources in order to infer relevant situations and events related to the observed environment[6]. Spatial data fusion refers to the combining spatial data from multiple sources to extract meaningful information with respect to a specific application context. Spatial data fusion help to improve capability to identify spatial patterns, trends, and relationships. Also, it provides a platform to rapidly integrate new/additional none spatial information to original spatial source of data.
   2. *Filtering and pre-processing*

      The phrase "garbage in, garbage out" can be applicable to any data analytic project. Data pre-processing transforms the data into a format that will be more easily and effectively processed. Data filtering is the steps to explore, filter and condition data prior to pass in to ETL process. Data filtering and pre-processing steps can take considerable amount of processing time. Having good understanding of the content of the data is essential for determining the best approach for filtering and pre-processing, eg: an unhandled NULL value can destroy any ETL process in the future steps.
5. *Data Extract & Store*

   Data storage might be a vital part of any data analytic projects, it will determine the data security, ETL process speed and performance. Data Extract & Store is often a complex combination of process and technology that consumes a significant portion human resources and cost. Each data source has its distinct set of characteristics that need to be conceded in order to effectively extract and store data. Understanding the content of the data is crucial for determining the best approach for data extract and store.

6. *Data Cleansing& Quality Assurance*
   Key criterion for the success of the data science project is the cleanliness and cohesiveness of the data. Real world data are generally incomplete, inconsistent and noisy. Therefore, data cleansing is paramount to the data science project, because in this stage project participants can work on missing values, smooth noisy data, identify outliers and resolve inconsistencies. Machine learning algorithms can be used to handle missing data. Domain knowledge is vital for data quality assurance.
7. *Geospatial Big Data Processing*
   There are two common approaches in processing the big data: batch processing and stream processing. Batch processing means running the data processing queries in a scheduled way. Also, batch processing can be defined as a effective way of processing high volumes of data where a group of transactions is collected over a period of time and process periodically. Stream processing means processing data as it comes in. The demand for stream processing is increasing. Stream processing analyses and performs actions on real-time data. In real-time or near real-time data processing is when speed is important, but processing time in seconds is acceptable. Very interesting paper (The 8 Requirements of Real-Time Stream Processing [7]) writing by Atta at al. (2016) outlines eight requirements that a system should meet to any real-time system processing applications. The main goal of paper is to provide high-level guidance what to look for when evaluating stream processing solutions.

   Rule 1: Keep the Data Moving
   Rule 2: Query using SQL on Streams (StreamSQL)
   Rule 3: Handle Stream Imperfections (Delayed, Missing and Out-of-Order Data)
   Rule 4: Generate Predictable Outcomes
   Rule 5: Integrate Stored and Streaming Data
   Rule 6: Guarantee Data Safety and Availability
   Rule 7: Partition and Scale Applications Automatically
   Rule 8: Process and Respond Instantaneously

   This 8 rules framework will be used in his section to process streaming data.
8. *Feature Selection*
   Feature selection is a crucial step usually mandatory in data science projects. Feature selection is also called variable selection or attribute selection. Its aim is to reduce data dimensionality by removing irrelevant and redundant features from a dataset. The irrelevant input features will induce greater computational cost therefore feature selection step increase the performance of the process.
9. *Data Partitioning*
   Separating data into training and testing sets is an important part of evaluating data mining and machine learning models. The training set is used to train or build a model. Once a model is built using the training Set, the performance of the model must be validated using new data, this data set is known as the testing set
10. *Spatial Data Analytics*

Spatial data analytics is concerned with analysis of data describing geographic phenomena. Systematically analyzing relationships between the spatial environment and relevant none spatial data offers a wealth of hidden information. Spatial analytics are how we understand our world—mapping where things are, how they relate, what it all means, and what actions to take. The spatial data science provided a better knowledge and understanding on spatial relationships among spatial and none spatial variables in the data source. Spatial data analytics can be pided in three sub categories such as Spatial Data Mining, Spatial Machine Learning and Spatial Statists.

11. *Predictive modeling*
    Predictive modelling is the process of building, testing and validating a model to predict the probability of an outcome. This is an iterative process and often involve training the model and testing model on same data set and finally find best fit model based on the business requirement.
12. *Result Aggregation*
    In this section, various analytical results will be analysed, evaluate and aggregate by project participant who has good domain knowledge. The systematic aggregation of the results from the multiple inpidual data analytics methods helped to provide comprehensive results set. Also, it might help to improve ability to identify hidden patterns, trends, and relationships.
13. *Data Visualization*
    It is important to note that visualizations have become more common conceptualizing means of communication. Data visualisation stage is an effective way of communicating complex and big data in a simple and comprehensible form data. In this stage, we visualize the data which has been analyzed and design it in a way that other people can take advantage of the analysis without digging deeper into data. Ideally, visualization or knowledge discovery in general should not be limited by data available or by tools used, data can be modified or extended and tools can be replaced or supplemented but visualizations should not be constrained or compromised. Data visualization aids people to understand the data by engaging it in a visual context. Patterns, trends and correlations which might be undetected in text based data are exposed and recognized much easier if users can use proper data visualisation techniques.
14. *Evaluation*
    Evaluation is systematic and objective assessment of ongoing projects or completed program's design, implementation and results. The main goal of the evaluation is to determine the relevance and performance of objectives, efficiency, effectiveness, impact.

6. CONCLUSION & FUTURE WORK

With the development of advanced remote sensing and communication technology, new sources of geospatial data began to develop in the lots of industries such as transport, utility, etc.. These new types of geospatial data are being received continuously at a very high speed. Researchers in academia and industry have made many efforts to improve the value of Geospatial big data and significant use of its value using data science. Having a good process for data mining and machine learning and clear guidelines is always plus point for any geospatial data science project. It also helps to focus required time and resources early in the process to get a clear idea of the business problem to be solved. Hence, the framework is proposed to aid geospatial data science project lifecycle and bridge the gap with business needs and technical realities.

Downside of existing frameworks are, they mainly focus on the data visualisation rather than the advance data analytics such as data mining, machine learning and statistics. Therefore existing framework are might not be a great choice for geospatial big data analytic and data science projects.

As an additional enhancement to this framework, performance evaluation of the proposed system can be conducted by using different datasets as an input. In addition, a simulation model for the analysis can be built for better understanding.

REFERENCES

1. BEYER, M. A. & LANEY, D. 2012. The Importance of “Big Data”: A Definition. https://www.gartner.com/doc/2057415/importance-big-data-definition.
2. DAS, A. C., MOHANTY, S. N., PRASAD, A. G. & SWAIN, A. A model for detecting and managing unrecognized data in a big data framework. 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 3-5 March 2016 2016. 3517-3522.
3. SUMATHI, N., S GEETHA, R., SATHIYA, B. & S RANGASAMY, K. 2018. Spatial Data Mining -Techniques Trends and its Applications.
4. LI, S., DRAGICEVIC, S., CASTRO, F. A., SESTER, M., WINTER, S., COLTEKIN, A., PETTIT, C., JIANG, B., HAWORTH, J., STEIN, A. & CHENG, T. 2016. Geospatial big data handling theory and methods: A review and research challenges. ISPRS Journal of Photogrammetry and Remote Sensing, 115, 119-133.
5. DIETRICH, D. 2015. Data science and big data analytics : discovering, analyzing, visualizing and presenting data.
6. MASTROGIOVANNI, F., SGORBISSA, A. & ZACCARIA, R. A Distributed Architecture for Symbolic Data Fusion. IJCAI, 2007.
7. ATTA, S., SADIQ, B., AHMAD, A., SAEED, S. N. & FELEMBAN, E. Spatial-crowd: A big data framework for efficient data visualization. 2016 IEEE International Conference on Big Data (Big Data), 5-8 Dec. 2016 2016. 2130-2138.
8. Vania Bogorny, Bart Kuijpers,, Andrey Tietbohl,, Luis Otavio Alvares, 2007. Spatial Data Mining: From Theory to Practice with Free Software. https://pdfs.semanticscholar.org/61a1/175964cf6366a9c0b97434d402bc84519f8c.pdf

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:contact@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation