Analysis, Matching, and Prediction of Data with Time or Space Characteristics according to Association Rule of Data Mining

# Summary

Nowadays, tons of data are produced every second in our daily life and it could be a very tricky problem for those who have to wade through mass data. In order to solve the problem, an analysis engine of big data will be built in our keystone project. The analysis engine can provide computing service of big data, which is a solution of fast and automatic data system. Specifically, our engine will choose different algorithm against actual conditions. At the beginning, we would store data into the distributed database. After that the system will process those raw data for further analysis. As a result, those massive amount of data will be fitted into a mathematical model. At last, the analysis engine can predict the future tendency based on the mathematical model. In other words, we would use parallel computing and association rule of data mining to do comprehensive data analysis. The advantages of high concurrency and high speed give distributed database the capability of handling real-time data accurately and efficiently. In the field of big data, our analysis engine devotes to analyze, match, and predict data with time and space characteristics then put them into real case.

# Research team members

Team member: Wang Yu\*, Shao Jialin, Zhang Pengyu, Zhang Hongbin, Jiang Minyu

Mentor: Chen Juncheng (College of Computer Science, Faculty of Information Technology)

# Disciplines

Scientific fields: Computer Science

# Project outline

## Indicate the state of the art.

Because of the amount of data that we are sending and receiving every day is increasing dramatically, information explosion is becoming the core of this era. Thus we could say that we are living in the real big data era. In order to reduce the running load of data processing, we need to sort, process, and manage data, thus data mining is becoming a hot issue.

The main value of data mining is its capability of transferring huge amount of data into useful information and knowledge, which have been used in various applications. Nowadays, data mining is using ideas from the following fields: statistical sampling, estimate, hypothetical test, search algorithm and modeling in machine learning. [1]

Generally speaking, engineers around the world usually utilize the following analytical method to do data mining [2]: classification, estimation, prediction, affinity grouping or association rules, clustering, and complex type data mining. The core is the research and improvement of association rules, which could be divided into simple association, time association, and reason-result association. The objectives of association analysis is to figure out the hidden network of relationship, which is likely to be hidden label, then group or layer data. This could simplify the following search process during retrieve in order to avoid time and resources waste resulted by data traverse.

However, due to the characteristic of association rules data mining, it shows better effect in handling discrete data. Thus the rationality of dividing continuous data of raw database before data mining would influence result of association rule data mining.

Along with these ideas, database system should provide effective support in storage, index, and query. Another point is that the parallel computing (high performance computing) also plays vital role in processing massive amount of data, especially when huge amount of data are separated.

While talking about parallel computing, MapReduce [3] is an effective tool. Map operation means it could process each element separately but not generally. Aside from that, reduce operation could do recursive operation in highly parallel environment. Hadoop [4] Distributed File System (HDFS) is one of the most influential practice using MapReduce, which is a distributed file system and do well in distributed processing of massive data. Hadoop have the following advantages compared to traditional relational database: higher reliability, higher expansibility, higher efficiency, higher fault tolerance, and lower cost. The appearance of Hadoop brings a new possibility for parallel computing, which makes MapReduce not only a result in lab but also a practical and useful system.

In order to obtain the non-sampling point data, we need to fit the huge amount of data into a smooth curve. Bernstein-Bézier curve is a useful tool in data processing and fitting fields.

Bézier curves are widely used in computer graphics to model smooth curves. As the curve is completely contained in the convex hull of its control points, the points can be graphically displayed and used to manipulate the curve intuitively. Affine transformations such as translation and rotation can be applied on the curve by applying the respective transform on the control points of the curve. [5]

Quadratic and cubic Bézier curves[6] are most common. Higher degree curves are more computationally expensive to evaluate. A composite Bézier curve is commonly referred to as a "path" in vector graphics languages (like PostScript), vector graphics standards (like SVG) and vector graphics programs (like Adobe Illustrator, CorelDraw and Inkscape). To guarantee smoothness, the control point at which two curves meet must be on the line between the two control points on either side.

The simplest method for scan converting (rasterizing) a Bézier curve is to evaluate it at many closely spaced points and scan convert the approximating sequence of line segments. On the contrary of common sense, it may generate too many points in areas where the curve is close to linear. A common adaptive method is recursive subdivision, in which a curve's control points are checked to see if the curve approximates a straight line to within a small tolerance. If not, the curve is subdivided parametrically into two segments, 0 ≤ t ≤ 0.5 and 0.5 ≤ t ≤ 1, and the same procedure is applied recursively to each half. There are also forward differencing methods, but great care must be taken to analyze error propagation.

During data processing, there are labels along with every data, and the most important characteristics are time and space. All time and space related data need professional database to store, analyze, process and manage. With the rapid development of database technology, spatial database and temporal database have already had some achievements and they have brilliant practical applications in various fields.

Remote sensing technology are using spatial database to manage data of animals and forestry. This could build an information system for protecting environment and endangered animals. Temporal database does well in handling data which updates fast and frequently.[7] It also show great capability in handling data which is collected from many measurement points and heavily dependent on the acquisition time. As a conclusion, there are several mature technology and achievements of spatial database and temporal database, both of them could complete data processing independently. But the problem now is how to build up connection between these two kinds of database then apply them into actual product.

After the consideration of status quo, we proposed a novel solution to solve the problem that we are facing in this big data era.

In order to analyze, compare, match, and predict data, we need to fit the data. We would use curves like Bernstein-Bézier to fit the data into a more smoothed curve. In the next stage, we would apply parallel computing in spatial database and temporal database to optimize the accuracy and efficiency of data mining process, which may figure out some hidden label of data. After we get the association functions f(x) and f(y) (x means time and y means space here), we would use the idea of pattern cognition and feature extraction to search required data fragment and predict subsequent data in the model. Nowadays there are several possible applications which has great potential, such as sharing vehicles, flow monitoring, and preference analysis to various situations.

## Describe the objectives of the research.

Low efficiency in massive data query, lack of unified data specification, and short in supporting space-time property of data are serious problems to be solved today. In this project, we need to analyze data with space or time characteristics. And use the result to extracts the needed data pieces from the huge amount of data. At last we use the obtained results for projections, comparison and draw the relevant conclusions.

Question1: How to analyze the huge amounts of data the sensors collect?

* Classifying the data according to property characteristic
* Fitting the treated data to correlation function[8]

Question2: How to use the data to fit better correlation functions?

* Using the method in data mining:

Handling the data with Bernstein- Bézier curve[9]

* Using the method in machine learning:

Combining supervise learning and unsupervised learning to handle the data

* Storing the underlying data with NoSQL database
* Classifying the data while storage, and add semantic meaning with tags [10]
* Avoiding traversal algorithm
* Using parallel computing [11], [12] to optimize the algorithm

Question3: How to search the required data fragments from the big amount of data?

* Using the idea of feature extraction [13] to find pieces of data
* Using the idea of pattern recognition [14] to find pieces of data
* Determining the trend of the data through the correlation function

Question4: Where can we use the result we obtained?

* Optimizing data query
* Helpful for the data projection
* Helpful for the data comparison

Innovation and project characteristics:

There are three aspects of innovation in our project:

1. Using parallel computing to improve optimization data query

2. Storing data with classification tag

3. Using the idea of feature extraction and pattern recognition to find pieces of data

Based on objectives of our keystone project, the complete work process is shown in Figure 1.

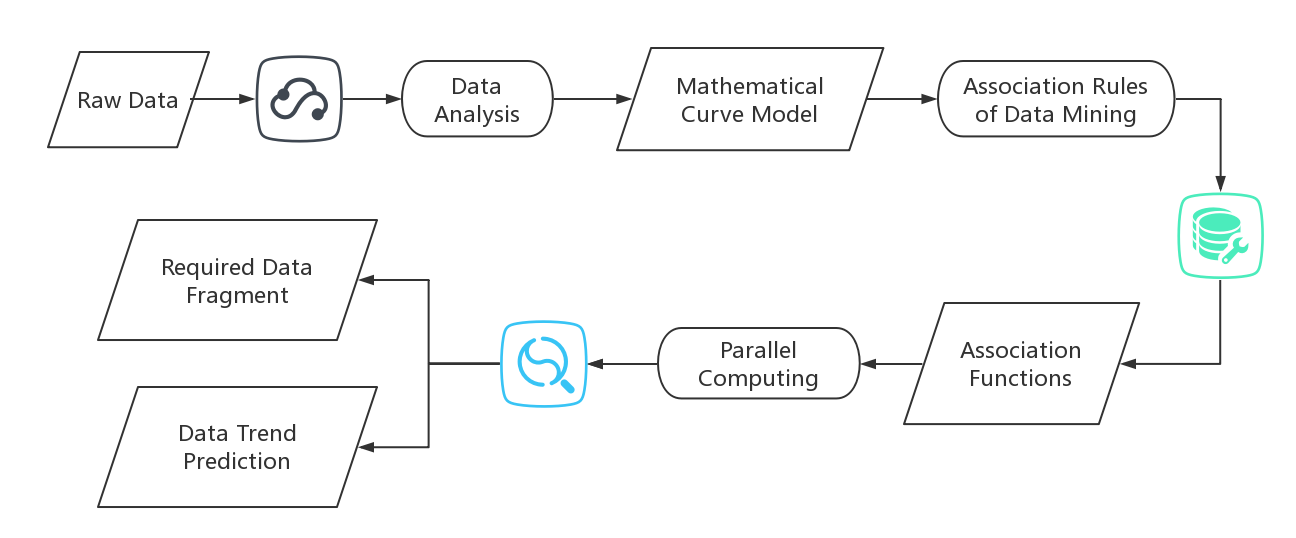


Figure Work Process

## Describe the methodology of your research.

We divide our methodology into 4 parts. With these 4 parts, we can eventually complete the overall data processing.

* **Data Collection**

Most of today's sensor data are continuous data, such as temperature sensors, wind sensors and GPS sensors. However, collecting data from multiple sensors and displaying is not an easy thing. Therefore, our solution is:

Using single-chip drive a number of sensors to collect data, and then running serial communication with the PC. After the processing of the software, we can obtain the data collected from varies of sensors.

Another way is using the DAQ device to obtain the analog or digital output of the sensor directly, then connect to the computer via USB to read the data and further process it. This can refer to a series of NI equipment and their develop environment “LabVIEW”

* **Data Storage**

After the collection of all sorts of data. It’s coming to the storage of the data, which is one of the key parts of our project. As for the storage, we need to meet the following conditions: safe, efficient, and classified.

**Safe**

In order to ensure the safety of user’s data, we will use multiple groups of servers to store data and a copy of the data. The behavior can minimize the user's data loss, conflict, error and other issues. So that it can ensure the safety of user’s data.

**Efficient**

To maximize the efficiency of data storage, we determine to use these three steps:

First, specifically defined two data tables: A used to record all the data files, B used to record large data files.

Second, for large data files, only in the A file to store the basic information, do not store the contents of the file (binary field content blank), the data content linked to the B table

Finally, B table stored in the binary field is converted to a file stream stored in the way.

**Classified**

Our goal is to create a spatially separated database. So, we need to store a separate data, including time, space and other data information.

Our initial idea is to create two sets of synchronous database servers. One for storing time information and the other for storing space information. At the same time, two database servers store a backup of another database server. So that, we can simultaneously make sure the classified and the safety of our database.

* **Data Analysis**

By using the method of Bernstein-Bézier, we can solve the problem of how to query data efficiently from massive sample points. And this could be the biggest problem for us. This could be the methodology of how to extract different methods of data from the sensor. Obviously, we will face all sorts of challenge from this methodology. Not only the very new area for us to get in, but also a high mathematical problem to deal with.

The simplest method for scan converting (rasterizing) a Bézier curve is to ：

Firstly, evaluate it at many closely spaced points and scan convert the approximating sequence of line segments. However, this does not guarantee that the rasterized output looks sufficiently smooth, because the points may be spaced too far apart. Conversely it may generate too many points in areas where the curve is close to linear.

Secondly, a common adaptive method is recursive subdivision, in which a curve's control points are checked to see if the curve approximates a straight line to within a small tolerance. If not, the curve is subdivided parametrically into two segments, 0 ≤ t ≤ 0.5 and 0.5 ≤ t ≤ 1, and the same procedure is applied recursively to each half. There are also forward differencing methods, but great care must be taken to analyze error propagation.

* **Data Mining**

We need a method to process the relationship among complex data label, it is, a algorithm to deal with association rule data mining.

The intensity of association rule can be describe as support and confidence.

The definition of support is that the number of time when a element of assemblage X and assemblage Y that simultaneously appears in one record/ the number of recording data.

Definition of confidence is the number of time when an element of assemblage X and Y that simultaneously appears in one record/ the appearing number of assemblage X.

The higher support and confidence is, the stronger intensity is. The association rule mining is to mine a rule that satisfy a certain intensity.

So we can divide it into two part:

Generate frequent item set.

The first step is to find a assemblage that satisfy the minimal support. The finding item called frequent item set.

The second step is to generated in the previous step of frequent item sets based on rules to produce a minimum confidence level, the resulting rules are called strong rule.

In order to reduce the generation time of frequent item sets, we should eliminate some of the impossible that is certainly not a collection of frequent item sets as soon as possible.

In this way, Apriori algorithm can meet the complex calculations. Apriori algorithm is for mining frequent item sets of association rules. The core idea is to mine frequent item sets through candidate generation and closed down detection plot in two phases.

## Provide a work plan, i.e. the different work packages and a detailed timetable.

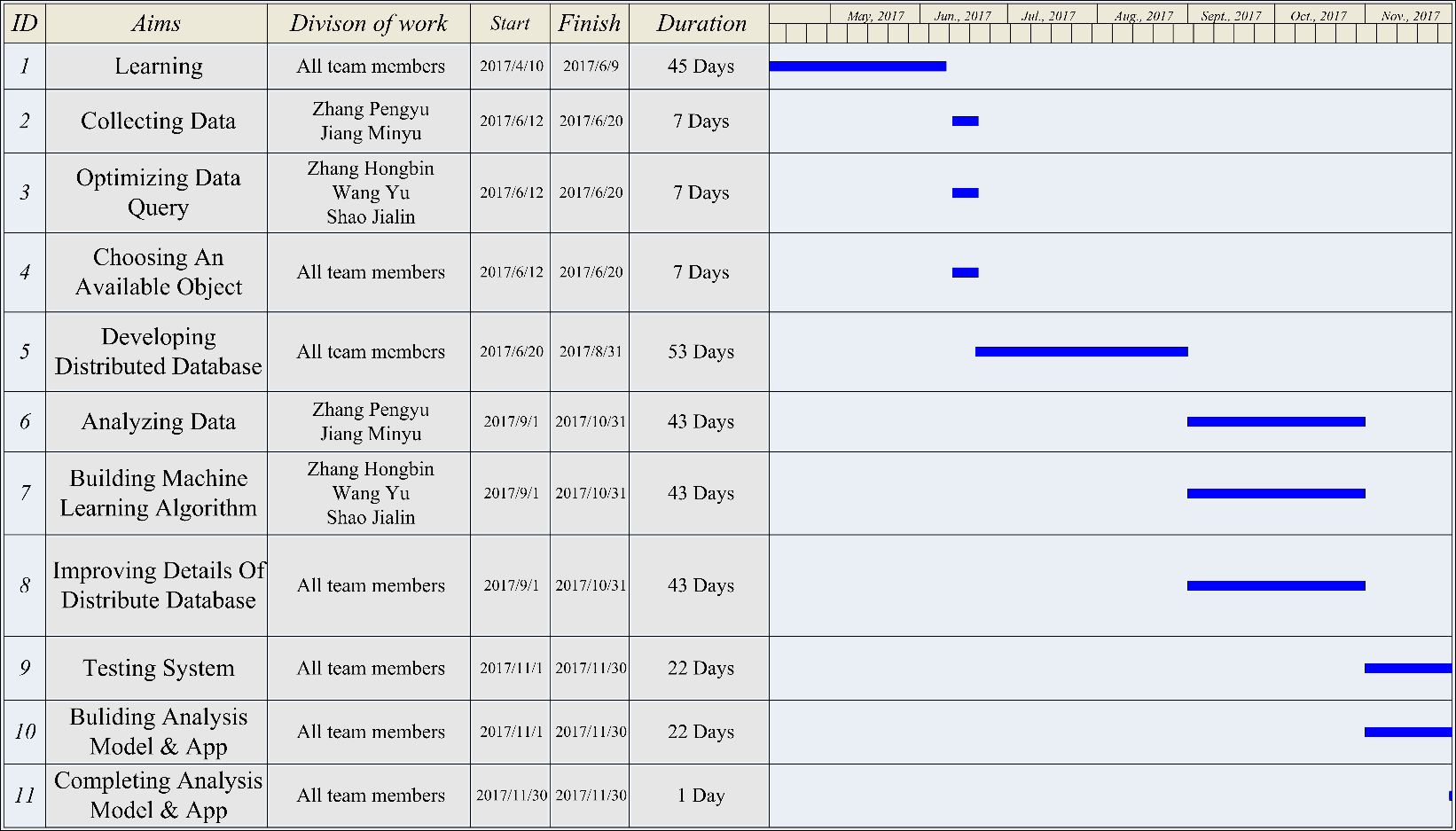


Table Detailed Timetable

Outcome: An App with the analysis engine and model dealing with big data in analyzing data of specific object.

## Enumerate the bibliographical references that are relevant for your research proposal.

[1] S.N. John, C. Anele, O. Okokpujie Kennedy, F. Olajide, Chinyere Grace Kennedy. Realtime Fraud Detection in the Banking Sector Using Data Mining Techniques/Algorithm. Computational Science and Computational Intelligence (CSCI), 2016 International Conference on (2016).

[2] Leslie Mertz. What Can Big Data Tell Us About Health?: Finding Gold Through Data Mining. IEEE Pulse. (2016) 7(5): 40-44.

[3] Seema Maitrey, C.K. Jha. Handling Big Data Efficiently by Using Map Reduce Technique, Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on (2015).

[4] Trupti V. Kenekar, A.R. Dani. An efficient private FIM on hadoop MapReduce, Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on (2016).

[5] Gerald E. Farin, Josef Hoschek, Myung-Soo Kim. Handbook of Computer Aided Geometric Design. Elsevier. (2002): 4–6.

[6] Changxin Song. Research of Association Rule Algorithm Based on Data Mining. IEEE International Conference on Big Data Analysis (2016).

[7] Li Jun, Yang Rui, Zhou Ru-Lian. Comparison of Data Forecasting Methods in large-scale Small Samples. Forest Inventory and Planning. (2010) 35(5): 7-11.

[8] Zvonarev MB, Cheianov VV, Giamarchi T. The time-dependent correlation function of the Jordan-Wigner operator as a Fredholm determinant. Journal of Statistical Mechanics-theory and Experiment. (2009).

[9] Goldman R, Simeonov, P. Two essential properties of (q, h)-Bernstein-Bezier curves. (2015) 96: 82-93.

[10] Buche P, Dibie-Barthelemy J, Haemmerle O, Hignette G. Fuzzy semantic tagging and flexible querying of XML documents extracted from the Web. (2006) 26(1): 25-40.

[11] Muchnick VB, Shafarenko AV. Dynamic evaluation strategy for fine-grain data-parallel computing. IEEE Proceedings-Computers and Digital Technoques. (1996) 143(3): 181-188.

[12] Ma Hui, Li Yongqi. Effective Data Exchange in Parallel Computing. 2013 International Conference on Information Science and Cloud Computing (ISCC). (2014): 106-112.

[13] Fields MC, Anderson DC. Fast Feature-extraction for Machining Applications. Computer-aided Design. (1994) 26(11): 803-813.

[14] Liu Haisong, Wu Minxian, Jin Guofan, Yan Yingbai. A postprocessing algorithm for the optical recognition of degraded characters. Proceedings of the Society of Photo-optical Instrumentation Engineers (SPIE). (1999) 3651:41-48

[15] Dongyu Feng, Ligu Zhu, Lei Zhang. Research on improved Apriori algorithm based on MapReduce and HBase. Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). (2016).

[16] Hu Shimin, Sun Jiaguang, Jin Tongguang, Wang Guozhao. Approximate degree reduction of bezier curves. Tsinghua Science and Technology. (2012) 3(2): 997-1000.

# Budget plan

1. **Consumables: total amount of 6000:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| Consumable type | Data | Database | E-MapReduce | Paper |
| Detailed description | Purchasing data from aliyun | MongoDB | Hadoop, Spark, HBase | Publishing paper in journal |
| Motivation | Analyzing & Building algorithm | Learning & Reference | Analyzing data | Publishing paper in journal |
| Approximate amount | 1000 | 2000 | 1000 | 2000 |

1. **Equipment: total amount of 4000:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Equipment type | Books | MaxCompute | Data IDE |
| Detailed description | Ceph, golang, data analysis, statistics | Purchasing from aliyun | Visual development interface |
| Motivation | Learning | Learning & Reference | Development |
| Approximate amount | 1000 | 2000 | 1000 |

# Ethics

If you have indicated “Yes” for at least one of the items below, you must submit your proposal to the research ethics committee of your host institution for ethical clearance, as soon as your application has been approved for funding. Your project can only start when this clearance has been formally given.

☒ I confirm that none of the issues below apply to my proposal.

☒ I hereby confirm having taken note that an ethical clearance is needed for the start of my project. I will thus ensure submission of my proposal to the research ethics committee of my host institution.

☒ I hereby confirm that I’ve already submitted an application for an ethical clearance at the local ethical committee of my host institution.

A full list of questions is provided as below.

1. Does your research involve the use of human embryos?

☐ Yes ☒ No

1. Does your research involve the use of human foetal tissues/cells?

☐ Yes ☒ No

1. Does your research involve human participants?

☐ Yes ☒ No

1. Does your research involve volunteers for social or human sciences research?

☐ Yes ☒ No

1. Does your research involve persons unable to give informed consent?

☐ Yes ☒ No

1. Are they vulnerable individuals or groups?

☐ Yes ☒ No

1. Does your research involve children/minors?

☐ Yes ☒ No

1. Does your research involve patients?

☐ Yes ☒ No

1. Does your research involve healthy volunteers for medical studies?

☐ Yes ☒ No

1. Does your research involve physical interventions on human study participants?

☐ Yes ☒ No

1. Does your research involve invasive techniques on human study participants?

☐ Yes ☒ No

1. Does your research involve human cells or tissues (other than from Human Embryos/Foetuses, i.e. section 1)?

☐ Yes ☒ No

1. Does your research involve personal data collection and/or processing?

☐ Yes ☒ No

1. Does it involve the collection and/or processing of sensitive personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?

☐ Yes ☒ No

1. Does it involve collecting/processing of genetic information/data?

☐ Yes ☒ No

1. Does it involve tracking or observation of participants?

☐ Yes ☒ No

1. Does your research involve further processing or previously collected personal data (secondary use)?

☐ Yes ☒ No

1. Does your research involve animals?

☐ Yes ☒ No

1. Does your research involve vertebrates?

☐ Yes ☒ No

1. Does your research involve non-human primates?

☐ Yes ☒ No

1. Does your research involve genetically modified animals?

☐ Yes ☒ No

1. Does your research involve cloned farm animals?

☐ Yes ☒ No

1. Does your research involve endangered species?

☐ Yes ☒ No

1. Do you plan to use local resources (e.g. animal and/or human tissue samples, genetic material, live animals, human remains, materials or historical value, endangered fauna or flora samples, etc.)?

☐ Yes ☒ No

1. If your research involves low and/or lower middle income countries, are benefits-sharing measures foreseen?

☐ Yes ☒ No

1. Could the situation in the country put the individuals taking part in the research at risk?

☐ Yes ☒ No

1. Does your research involve the use of elements that may cause harm to the environment, to animals or plants?

☐ Yes ☒ No

1. Does your research deal with endangered fauna and/or flora and/or protected areas?

☐ Yes ☒ No

1. Does your research involve the use of elements that may cause harm to humans, including research staff?

☐ Yes ☒ No

1. Does your research have the potential for military applications?

☐ Yes ☒ No

1. Does your research have the potential for malevolent/criminal/terrorist abuse?

☐ Yes ☒ No

1. Are there any other ethics issues that should be taken into consideration? Please specify.

No any other ethics issue.