HW 1

Big Data Solution of Biomedical Image Analysis

Name: Xue, Luo
ID: 1801212899

Background and objective
In the medical field, data volume soars and the traditional methods can not handle it effectively. In biomedical computation, the continuous challenge are management, analysis and storage of biomedical data. Nowadays, big data technology attaches its importantance in data management, organization and analysis by using machine learning and artificial intelligence technology. Besides, it also allows quick access to data using a NoSQL database. Therefore, a new framework to process biomedical image data by big data technology is in need of implementation.

1. Big biological data properties explanation

Big data is often defined by three major characteristics called the **“3V”**: volume (amount of data generated), variety (data from different categories) and velocity (speed of data generation).

For this topic, another two more "V" need to be introduced: **variability** (inconsistency of data) and **veracity** (quality of captured data). Thus big data problems are now identified by the **“5V”**.

Big data in health is concerned with meaningful datasets that are too big, too fast, and too complex for healthcare providers to process and interpret with existing tools. The explanation of "5V" of biological image data is as following:

1.1 Volume & Velocity

Data are daily generated at unprecedented rates. As it can be seen in Figure 1, in the past decade alone, the number of publications incorporating digital morphology images into systematics research have more than tripled, with over half of those using CT (Fig. 1a). Considering a single dataset including image stacks, 3D models, and relevant metadata can reach over 100 GB per specimen11, management of CT and other digital files poses a serious challenge.

Another challenge is the size of the data set. Current-generation time-lapse microscopes include integrated incubation and can typically acquire 100 movies or time-lapse image sequences in a single experiment. Each movie can consist of thousands of images. In a work

analyzing stem cell image sequence data, a single data set of 200 movies requires 350 gigabytes (GB) of image data or more. This is obviously too much data to analyze by hand or by eye—we must turn to computational analysis.

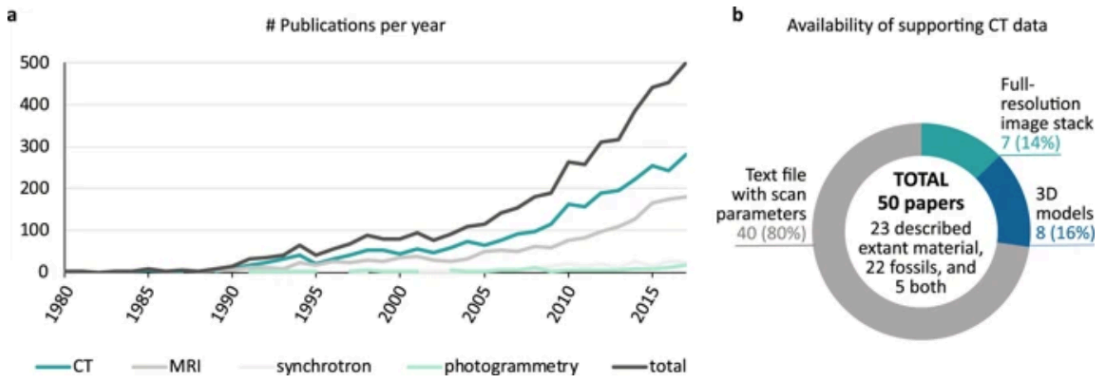

Figure 1

1.2 Variety

Data are generated from different heterogeneous sources (e.g., laboratory and clinical data, patients' symptoms uploaded from distant sensors, hospitals operations, and pharmaceutical data). In biomedical imaging, the techniques that are well established within clinical settings to capture an image are: computed tomography, magnetic resonance imaging, x-ray, molecular imaging, ultra sound, photo-acoustic imaging, fluoroscopy, and positron emission tomography - computed tomography (PET-CT).

These techniques take the medical images with high definition and large sizes. The advanced analysis of biomedical image datasets has many beneficial applications. It enables to personalize remotely radiological services (e.g., doctors can monitor online image of patients in order to provide a prescription). However, specialized doctors are very few and cannot diagnose all these millions of images generated. With this rise of biomedical image data, new demands to Artificial Intelligence for machine learning systems to learn complex models are made.

Data Source Techniques	Use Frequency	Quality
Computed tomography	✔✔✔✔	✔✔✔
Magnetic resonance imaging	✔✔✔	✔✔✔
X-ray	✔✔✔✔✔	✔✔✔✔
Molecular imaging	✔✔	✔✔✔
Ultra sound	✔✔✔✔✔	✔✔✔✔✔
Photo-acoustic imaging	✔✔✔✔	✔✔✔✔
Fluoroscopy	✔✔✔	✔✔✔✔
Positron emission tomography - computed tomography (PET-CT)	✔✔	✔✔✔✔✔

Figure 2

1.3 Variability & Veracity

In medicine, the data encountered are mainly obtained from patients. These data consist of physiological signals, images, and videos. They can be stored or transmitted using appropriate hardware and techniques. One of the services used in medicine for the storage and transmission of image data is the Picture Archiving and Communication System. It is are popular for delivering images to local display workstations. However, data exchange with this system is highly standardized, and this system relies on using structured data solely to retrieve medical images rather than leveraging the unstructured content of the biomedical images.

If we use unstructured data, it may encounter a series of problems, such as patient clinical data instability, cell variation in the process of data collection, data capture failure in the imaging process and so on.

---

2. Solution

2.1 Workflow

Medical imaging supplies important information on organ function and anatomy in order to detect the state of diseases. I propose a workflow to handle the steps of image processing.

The main goal of the workflow is to give in each step of a conceptual framework to provide a systematic method necessary for analyzing big data in biomedical imaging from patient data. The conceptual framework proposed is summarized in Fig 3. This figure shows the parts of big data processes for biomedical image processing. The whole workflow can be seen in Fig 4.

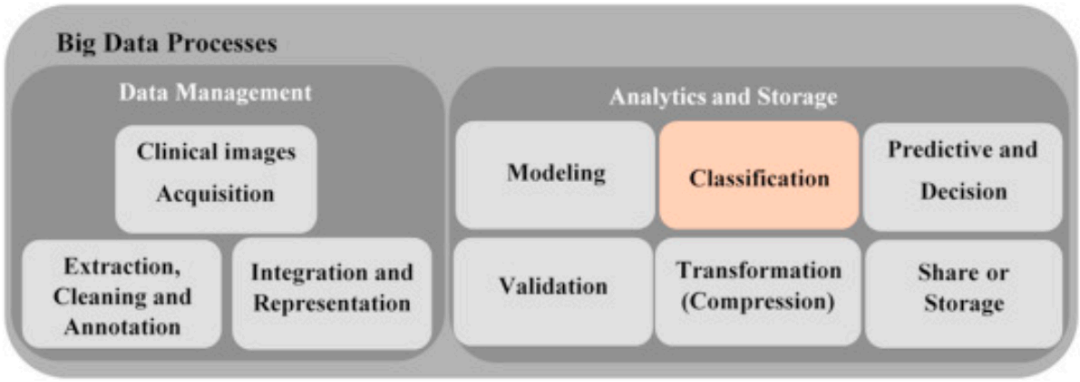

Figure 3

2.1.1 Data management

Data management is the organization, administration, and governance of large volumes of both structured and unstructured data.

- **Clinical images acquisition**

In biomedical imaging, the techniques that are well established within clinical settings to acquire an image are: computed tomography, magnetic resonance imaging, x-ray, molecular

imaging and so on.

- **Extraction, Cleaning, and Annotation**

Extraction refers to a technique that enables to obtain useful biomedical images from the raw data and, refines them so that they can be used in the following analytic steps. Cleaning is the process that eliminates noise on acquired images.

- **Integration and representation**

This is the step which involves the automatic clustering of images in the databases. Preview of images is also possible at this level before analyses.

2.1.2 Big data analysis

It is an entire program that bears the development of theoretical, mathematical, artificial intelligence, statistical methods for analysis of biomedical images, clinical diagnosis and patient monitoring.

In this process, since the classification and data store are the two issues we concern most, the explanation of them is as following:

- **Classification**

Classification is an example of pattern recognition. Classification in machine learning concerns a problem of identifying to which set of categories a new population belongs. When category membership is known, the classification is done on the basis of a training set of data containing observations. An example would be to assign a given biomedical image into "anatomic body part" or "biological systems" classes. The classification steps are processed under a supervised learning algorithm via a support vector machine (SVM). SVM is chosen from several other supervised learning algorithms because SVM and neural network are two well-known techniques used to classify biomedical image data. Indeed, in medical imaging, SVM and neural networks take up to 42% and 31% respectively of the most used algorithms. This statistic shows the efficiency of the SVM algorithm.

- **Data storage**

Big data applications commonly use Not Only SQL (NoSQL) technologies as a database. NoSQL refers to a database category that appeared in 2009 which differs from the relational databases. Indeed, one of the recurring problems of relational databases is the loss of performance when one should process a very large volume of data. Moreover, distributed architectures provide the need to adapt solutions natively to replication mechanisms of data and load management. Thus, I will use NoSQL database for this problem, the detailed illustration can be seen in *2.2 Database Choice*

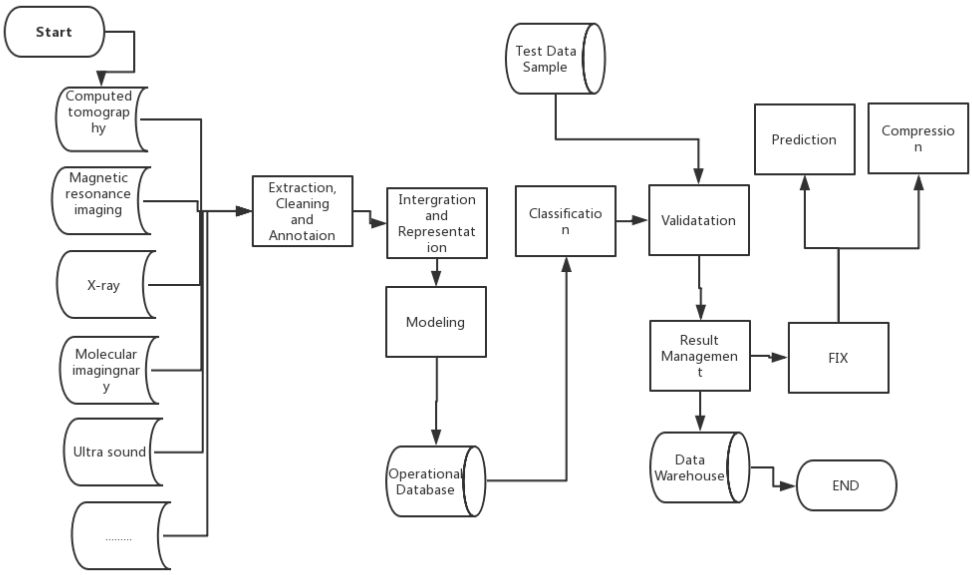

Figure 4

2.2 Database Choice

I will use Hadoop to solve this problem. Since the classification is the main process of biological image processing, MapReduce can naturally and easily to classify all of the images resulting from the modeling step in each defined category.

MapReduce is a processing technique and programming model done in a lateral and scattered manner. MapReduce programming is a special form of a directed acyclic graph (DAG) which is applicable to a wide range of used cases. It is organized in two functions. The first one is a Map function, which transforms an element of data into some number of key/value pairs. The second is the Reduce function, which is used to merge the values (of the same key) into a single result.

Therefore, the simplicity of the implementation of MapReduce programming becomes the reason why I chose Hadoop. All of the images resulting from the modeling step will be automatically classified in each defined category. That will optimize the prediction and decision methods to be applied to the images. Thus, we can use Hadoop and apply a deep learning algorithm in each category resulting from the classification step, in order to predict and make decisions automatically on each image.

References

[1]Andreu-Perez, C.C.Y. Poon, R.D. Merrifield, S.T.C. Wong, G.-Z. YangBig data for health
Journal of Biomedical and Health Informatics, 19 (4) (2015), pp. 1193-1208
[2]J. Luo, M. Wu, D. Gopukumar, Y. Zhao Big data application in biomedical research and health

care: a literature review Biomed Inf Insights, 8 (2016), pp. 1-10
[3]A. Belle, R. Thiagarajan, S.M.R. Soroushmehr, F. Navidi, D.A. Beard, K. Najarian1 Big data analytics in healthcare BioMed Res Int (2015), p. 16
[4]M. Viceconti, P. Hunter, R. Hose Big data, big knowledge: big data for personalized healthcare Journal of Biomedical and Health Informatics, 19 (4) (2015), pp. 1209-1215
[5] A. Yang, M. Troup, J.W.K. Ho Scalability and validation of big data bioinformatics software Comput Struct Biotechnol J (2017), p. 8 Article in press
[6] S. Istephan, M.-R. Siadat Unstructured medical image query using big data – an epilepsy case study J Biomed Inf, 59 (2016), pp. 218-226