Big Data Visualizations Through MongoDB For Precision Medicine In Medical Education

Altaf Siddiqui, Ph.D.
President
American Enterprises, LLC
22026 E Ridge Trail Cir
Aurora, CO 80016

Omer Siddiqui
Research Associate
American Enterprises, LLC
22026 E Ridge Trail Cir
Aurora, CO 80016

Abstract

The importance of precision medicine is increasingly being recognized in healthcare. Precision medicine is driven by patient data and physicians' diagnoses in a comprehensive manner that considers the uniqueness of individual patient where each patient is a partner in the whole process. Big data analytics can provide a means to analyze and interpret healthcare data in a manner that can be quickly implemented in patient care because the available data is not structured in the way traditional databases are. There is growing interest by physicians to take advantage of big data analytics. However, limitations in deciphering and interpreting this data by healthcare professionals has impeded implementation of this technology. Most off-the-shelf software do not provide step-by-step instructions needed for a physician to understand big data analytics. In this paper, we provide a way to create big data visualizations through MongoDB with upload & download capability on web repositories. Keywords from PubMed were integrated to provide data visualization using the MongoDB programming thereby providing a unique solution for the issues that healthcare providers face in their understanding of the big data. The web repositories with big data visualizations for precision medicine will provide healthcare professionals and specialists a readily accessible platform for efficient diagnosis and care. Recommendations are provided about patient documents and visualizations, which will provide a thorough understanding of the data, knowledge sharing, collaboration, help in medical education and efficiency to healthcare providers.

Keywords: big data visualization, mongodb, precision medicine, web repositories

Introduction

The history of precision medicine dates back to 1960 (Jane, 2002). However, the term "personalized medicine" first appeared in published works in 1999 (Managed Care, 2011). This term was revived through the Personalized Medicine Initiative (PMI) by President Obama in 2015 during his State of the Union address. The traditional one-size-fits-all approach with patients' diagnoses needs to be improved, as it lacks the inclusion of the latest technologies and collaboration. Precision medicine provides a newer approach to patients' diagnoses that is affordable and provides an opportunity to consider other important factors, such as genes, environment, and ethnicity which could provide useful information for the treatment (Dhawan, 2016). One reason why precision medicine is uncommon in routine practice is because analyzing big data is complex and beyond the scope of most physicians. In a 2020 Health Trends Report done at Stanford University with 523 physicians, 44% of them said their medical education was "not very helpful" or "not helpful at all as it applies to the emerging data related technologies (Stanford Medicine, 2020). This response came in the context of the latest data driven technologies for healthcare. The same report pointed out the fact that the majority of the physicians were open to the latest technologies dealing with data and thought it was essential for their decisions on healthcare. However, the challenge is in the understanding and use of the latest technologies that could be used in healthcare

There is enough evidence about the need and the importance of precision medicine and big data analytics (Jain, 2002 & Stanford Medicine, 2020). The problem lies when the physicians do not have sufficient skill sets to understand big data and its analysis as it is beyond the reach of most physicians. While a good percentage of physicians are going back to school for further training on the latest technologies, it might not be possible for every physician to do the same. Big data analysis through an easy-to-understand software might be a solution for those who do not have time to go back to school. This paper presents the methodology for the design of such a software through big data analysis using MongoDB visualizations. It is critical that the physicians who would utilize such a software-are able to understand the patients' data and create visualizations without learning all the technicalities of MongoDB. A concept of web repositories is also presented in this paper, demonstrating that visualizations could be stored for collaboration among other physicians, healthcare professionals, and big data experts – we call them a "team" of precision medicine in this paper. Such a "team" can benefit enormously through an easy-to-understand software which is designed for an average layman person and not for IT experts.

Methodology

The main topics of this paper include patients' documents processing, extraction of the data from patients' documents into JavaScript Object Notation (JSON) format, the storage of JSON into big data using MongoDB, the creation of visualizations using Python programming language, and the upload and download capabilities of the proposed software to an intranet web repository using Python. While JSON, MongoDB, Visualization, and Python are done by an IT expert – it does not have to be learnt by a physician. The physician is an end user of the software that used these technologies. These technologies are currently used by big data on daily basis (Agrahari1 & Rao, 2017). There are so many technologies which are out there to process medical data, however, it needs a bridge building between the IT experts and the medical community.

Only then, the potential of these emerging technologies could be utilized benefitting the humanity. A software approach proposed in this research is the solution.

The process of the proposed software presented in this paper starts by finding patients who would be interested in participating in a collaboration program similar to *All-of-Us* (NIH Homepage, 2020). The patients sign a consent form and provide the relevant information needed by physicians on the team. It is suggested that a big data analyst/s, such as a MongoDB expert is hired by a healthcare organization's team depending on its size. In May 2018, the *All-of-Us* Research Program was declared and funds were allocated to collect the whole-genome sequencing of 200,000 people per year (NIH Homepage, 2020) . After the patients sign consent form to become a partner in a program like *All-of Us*, the patient data is collected by his or her healthcare provider and stored in documents of history and physical exams, operative notes, discharge summaries and outpatient clinic visit notes (NIH Homepage, 2015). This document is then provided to a MongoDB expert. The MongoDB expert extracts the pertinent big data in consultation with the physicians who are on the team of precision medicine. MongoDB uses a JSON type description, which is based on key-value pairs. A sample big data obtained from a patient's document could be like figure 1 (NIH Homepage, 2015). A similar approach in storing big data as JSON format into MongoDB has been proposed previously by other authors as well (Messaoudi, Fissoune, & Badir, 2018). JSON is widely used in the IT industry for many text related data. In the past, XML was a choice for small to medium size data but JSON has more usage in the modern IT software.


```
{
    Id : 123456789
    Age: 59
    Gender: M
    Ethnicity: Black
    Marital_Status: Married
    Location: Denver
    Physical Activity: Yes
    Pain_Scale: 5
}
```


Fig. 1. A sample JSON format for the data extracted from patients' documents and stored in MongoDB

JSON is a popular format used in web and database related software technologies. MongoDB is a NoSQL (Not-only SQL) database management system used for documents. Traditional database management systems lack the capabilities to handle documents' attributes (Messaoudi, Fissoune, & Badir, 2018). Traditional databases involve relationships between

entities (tables). Traditional databases had been with the IT people for quite some time. While it will stay in many disciplines, the production of big data on daily basis demand for a new technology like big data. The above information was chosen as if a physician wanted to pick data related to Pain_Scale based on demographics. The Pain_Scale was fictitiously entered from a scale of 1 through 10 which was multiplied by 10 to create better visualization. Visualization is achieved here through python modules. Python is another popular language used in the development of software which has built in modules to help/display visualization as shown in figure 2. Visualizations are easy to understand by everyone including physicians. They do not require the technicalities involved behind the scenes. Visualizations are available in various forms of graphs, such as, line, bar, histograms, etc. Many visualizations were used in the current COVID-19 data using python modules (Ganesh, 2021). These visualizations provided the healthcare experts to see the trends of the pandemic. Similar visualizations can be used in the proposed software to understand the patients' data under examination. These visualizations are then shared by the 'team' to get the best expertise in the area. Many case studies can be looked at by various physicians for knowledge sharing and best diagnosis. It is a game changer concept in the healthcare industry. Right now, most of the case studies are done by mutual acquaintances among the physicians. However, with the software proposed by the author these visualizations can be shared through hundreds and thousands of physicians across the globe using the concept behind web repositories. With the technologies mentioned in this paper, all of this is possible and feasible.

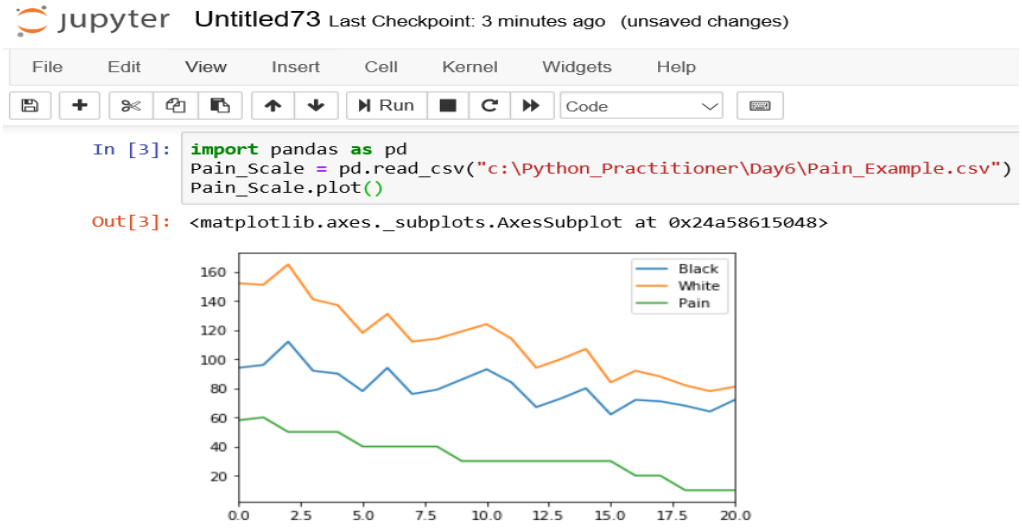

Fig. 2. A plot showing visualization of patient data. The MongoDB key-value pairs are converted into csv format and then python programming is used to draw the visualization using Jupyter.

In the above sample data, 8 key-value pairs were used. However, this data could be structured involving more key-value pairs as needed. The JSON file that stored the key-value pairs of data was then converted to a comma separated value(csv) file by using Python

programming language. The Python programming code used is shown in figure 2 to create the visualization from the data originally stored in MongoDB.

The pain_scale is between 9 through 58 which were rounded to 10 and 60 respectively. The different frequency for Black and White were entered for a certain pain_scale. All of these numbers were fictitious to demonstrate that a visualization is possible for big data (MongoDB) using Python. The Python code in Figure 2 imports "pandas" which is open source, meaning it is free of charge to its users. Jupyter is used as an Integrated Development Environment(IDE) to run the Python code. Python has rich libraries and modules like pandas, which have built-in methods to help plot the visualization as well. A physician, after consultation and training on big data should be able to create a similar visualization about "Pain_Scale" as it relates to the demographic data. MongoDB charts and many third-party software tools also allow MongoDB data to be used for visualizations. The visualizations illustrate the level of Pain_Scale among two key-value pairs based on their ethnicity. These two pairs of data are a sample for demonstration purposes and can be changed to other key-value pairs on a as needed basis. The whole process from the collection of patients' data to the creation of visualization is shown in Figure 3.

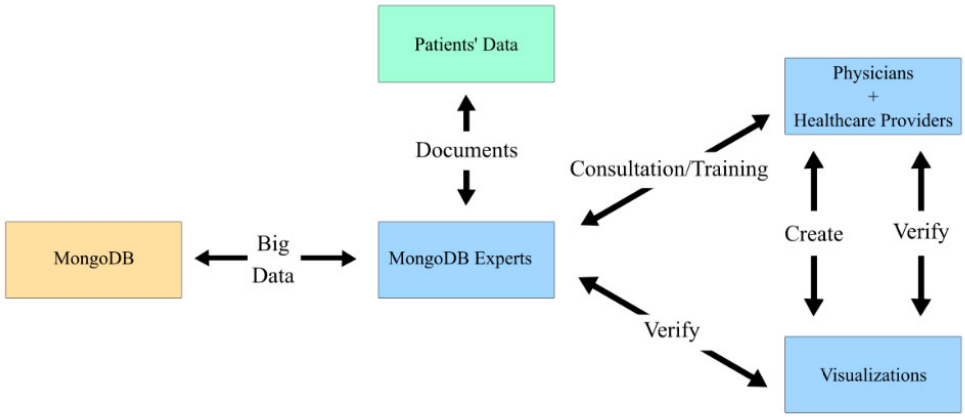

Fig. 3. Schematic outlining the whole process from the collection of big data from patients' documents until the creation of visualization is shown.

In Figure 3, the MongoDB experts receive the patients' documents in electronic format and convert into JSON format. The JSON format is ready to be stored as big data into MongoDB. During this process, the MongoDB expert is in constant consultation with the physicians, providing consultation and training on what data is being stored that is relevant and important. Once trained, the physicians can create visualizations themselves that are ready to be uploaded. During this process, the physicians verify the accuracy of visualizations among their peers and MongoDB experts.

The process of collaboration among physicians begins when a physician starts uploading and downloading such visualizations and shares with other physicians and healthcare professionals. In the proposed software, a physician would have to provide a password for authentication before he or she can upload or download a visualization on a web repository. Web

repositories provide a great way for sharing knowledge and collaboration in many areas (Siddiqui, 2015). The visualizations will be stored on an intranet web repository based on its classification. Classification plays an important role in searching and saves time (NIH Homepage, 2015). The web repository is a central location like a Microsoft SharePoint intranet site (Williams, 2011). The classification could correspond to folders. A sample web repository would look like figure 4 where a physician will click on the visualization of his/her choice after logging in.

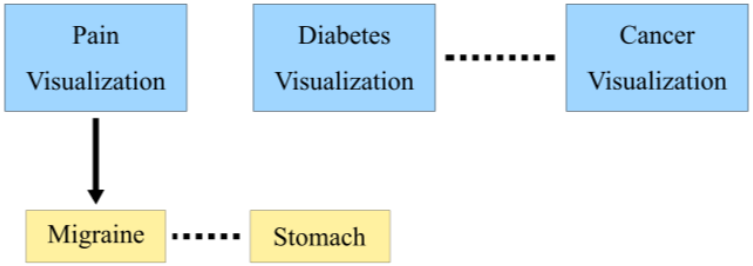

Fig. 4. The web repository classified into different categories of diseases.

Once a visualization is completed by a physician, they can authenticate their identity though a username and password for the intranet that they are part of. This is essential for the security of the data of their patients. The visualizations are classified based on the type of a disease. The classification is critical for an efficient use of information. With the overwhelming amount of data that is available, it is difficult to filter the data that is relevant. Therefore, the diseases are put into their corresponding folder shown in Figure 4. This also gives an opportunity for the physicians to click on the folder of their specialty. This classification is done by the MongoDB experts after consultation with the team. It is possible to have further classification within a disease category as well.

The whole process after the creation of a visualization to uploading/downloading on a web repository is shown in Figure 5. The visualizations will help physicians to understand the disease with more data and help in the diagnosis of a patient in a personalized manner.

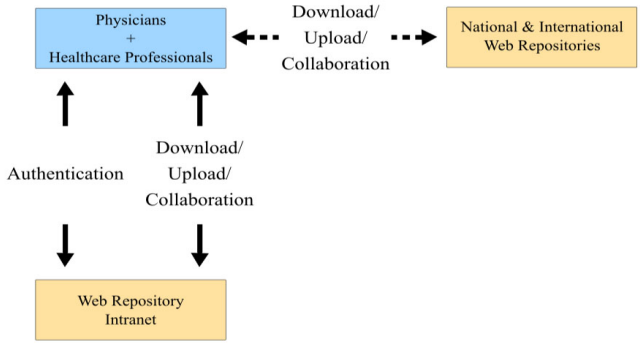

Fig. 5. After the visualization is created, a physician is authenticated through username/password and then given the option to upload or download on an intranet for collaboration with other physicians. This concept could be extended to the national and international level web repositories.

As shown in Figure 5, once a visualization is uploaded, it can be shared with other physicians and healthcare providers. The power of collaboration through web repositories should be limited to the local doctors' offices and healthcare providers. It should be expanded to the national and international level web repositories in situations where the cure is not easily found as in the case of the COVID-19 outbreak.

Results and Discussion

While all of the data related technologies discussed above already exist, and many works have been published on each of the above topics, it lacks the comprehensive approach that is needed to benefit from each of the above technologies at smaller doctors' offices and healthcare organizations. Moreover, the element of collaboration is also lacking at smaller medical offices in the whole diagnostic process of a patient. The objective of this paper is to synthesize all the emerging technologies into one easy-to-use software that will benefit the physicians who would otherwise not pursue a data emerging degree program and would be excluded from the power of collaboration and modern technologies. The key thing is to develop a sophisticated software using the emerging technologies.

Precision medicine is becoming the future of treatment for the patients in developed countries where the latest data related technologies have progressed tremendously in the past decade. Many physicians are realizing the power of big data and its application in the field of precision medicine. While a good percentage of the physicians are going back to school for data technologies related degrees, not all of them have this flexibility. A user-friendly software with little training for physicians is proposed in this paper to better equip physicians who are unable to pursue another big data technology degree. The software proposed could be written with the help of IT experts, physicians, and healthcare professionals who are part of the same team. The technologies involved in the writing of the software include, python programming with visualizations and big data using MongoDB. These technologies are proven in the industry and are widely used for the software that deal with big data.

Patients should be notified requesting their consent to be included into a research study about a certain disease before their documents are accessed for the proposed software. The key-value pairs in the MongoDB are selected from the patients' documents, which are those fields that are useful to physicians' diagnoses process. This paper picked one variable of pain-scale randomly, and fictitious responses from two ethnicities were stored in a csv format. Python was used to visualize the csv file. Python scripts can be used to create a button in a software's user interface for upload/download capabilities that could be merged with the proposed software (More complex APIs, 2020). The physicians are authenticated with a username/password before they can access the web repository in an intranet. The participating physicians could collaborate in a similar fashion through web repositories on the internet at national and international level for

a particular disease. Collaboration is an important factor to learn about a disease and diagnoses (NIH homepage, 2020). The collaboration will help physicians to understand the disease with evidence from more data, help gain expertise from their colleagues, and allow faster and more accurate diagnoses of their patients' illnesses. All of the technologies discussed in this paper are available and tested by the industry. The thing which is missing is the connection of these technologies with the physicians. There is no better way to communicate and collaborate than web repositories in the modern age when we are all connected through today's cutting-edge internet technology.

References

1. Agrahari1, A. Rao, D. : A Review paper on Big Data: Technologies, Tools and Trends. International Research Journal of Engineering and Technology (IRJET), 4(10), 640-649. (2017).
2. Dhawan, A.: Collaborative Paradigm of Preventive, Personalized, and Precision Medicine with Point-of-Care Technologies. IEEE Journal of Translational Engineering in Health and Medicine, 4,1-8. (2016).
3. Ganesh, S.: Impact of COVID-19 – Data Visualization using Python. https://towardsdatascience.com/impact-of-covid-19-data-visualization-using-python-6f8e3bdc860b, last accessed 2021/10/2021.
4. Jain, K.: Personalized medicine. Curr Op Mol Ther 4(6), 548-558. (2002).
5. Managed Care, https://www.managedcaremag.com/archives/2011/8/history-and-future-personalized-medicine, last accessed 2020/04/28.
6. Messaoudi, C., Fissoune, R., Badir, H.: A performance evaluation of NoSQL databases to manage proteomics data. Int. J. Data Mining and Bioinformatics, 21(1), 70-89 (2018).
7. More complex APIs: Upload and Download Files with Flask, https://docs.faculty.ai/user-guide/apis/flask_apis/flask_file_upload_download.html, last accessed 2020/04/28.
8. NIH Homepage, https://www.nih.gov/sites/default/files/research-training/initiatives/pmi/pmi-working-group-report-20150917-2.pdf, last accessed 2020/04/27.
9. NIH Homepage, https://allofus.nih.gov/about/protocol/all-us-consent-process, last accessed 2020/04/28.
10. NIH Homepage, https://www.ncbi.nlm.nih.gov/pubmed/28122115, last access 2020/04/27.
11. Siddiqui, A.: Design of Instructional Modeling Language and Learning Objects Repository. In: AECT CONFERENCE 2015, pp. 1-9. Springer, Indianapolis(2015).
12. Stanford Medicine, https://med.stanford.edu/news/all-news/2020/01/health-trends-report-spotlights-rise-of-data-driven-physician.html, last accessed 2020/04/28.
13. Williams, J.: SharePoint Site Eases Information Flow for Clinical Engineering Team. Biomedical Instrumentation & Technology, 45(1), 49-52. (2011).