Data Science and Cancer: An Approach to the Challenges Involved.

Deekshita S[1]

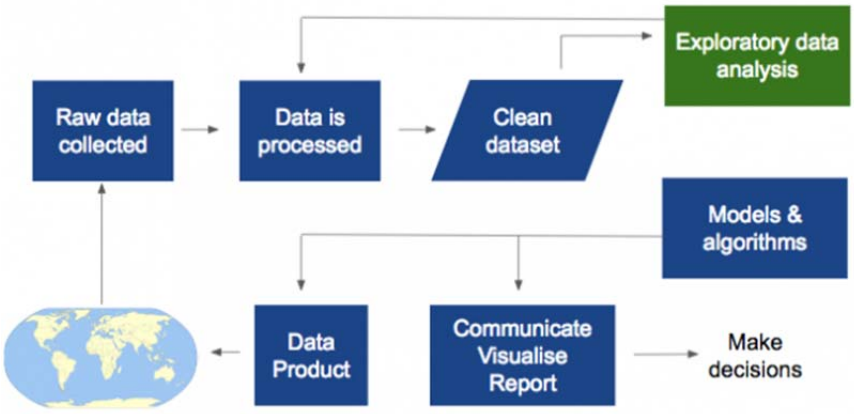
Third Year Engineering Student, New Horizon College of Engineering, Bangalore, India

Abstract: Cancer is one of the fastest evolving, ever changing and an adapting complex disease. In order to understand this complexity, there lies a need to study the snapshots of the tumor's genetic make up. However, these snapshots need to be acquired frequently to understand the process behind the evolution of the tumor. Humongous amounts of data is generated as a result of the measurements underlying these snapshots. The important patterns identified in the specific cancers can be leveraged to develop models that diagnose and treat the disease. The complexity of the disease can be solved by data science, involving sequencing methods. This can be done with intelligent machines using the genomic data sets and artificial intelligence as a tool.

Keyword: Cancer, Data Science, High performance computing and Algorithms, Sequencing**.**

I. Introduction

Data Science is an interdisciplinary field that involves statistics, scientific methods, processes and systems to extract knowledge from data. It expands over several domains such as Mathematics, Information Science, Statistics and Computer Science. Jim Gray, imagined Data Science as "fourth paradigm" of Science. Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data. Data is increasingly cheap and ever available. The digitizing of analog content that dates back to several centuries and collection of a plethora of data from various platform is linked to the usage that the data generated and collected can be put to with emerging and existing technologies. The high level overview of the data science process is data collection, data modeling and analysis followed by problem solving and decision support.



Source: David Wilks, Data at Government Digital Service, UK

Data Science has the potential to revolutionize health care. The field of medicine has been generating and storing data for several decades in the form of clinical reports, case studies, insurance data and hospital records. Specifically, biological data; gene expression, next-generation DNA sequence data, proteomics, and metabolomics, electronic health records (EHRs) and longitudinal drug and medical claims. The combination of data and resources can tackle the challenges of the problem set intended to be solved. With leveraging of data accumulated over the years, the field of medicine, benefits to a large extent with respect to the diagnosis of an individual's disease.

As compared to the traditional method of treatment wherein what worked for most patients is used, the treatments specific to individual solves the problem of complexities and this is possible using data science. The technology for storing, analyzing and ubiquitous data networking through mobile network has branched ways to possibly cure cancer. The only shortcoming being the challenges involved.

II.Challenges of data science

The challenges of data science are several with respect to its specific application. The thesis focuses on the challenges that data science holds to treat cancer. They can be broadly categorized as below-

I) Demand for reliable software.

II) Lack of available data.

III) Lack of volume and ethnic diversity in data sets.

IV) Data processing and recognition.

V) Data Privacy.

Demand for reliable software:

There exists a huge amount of contribution from a reliable software for treating cancer with the help of data science. Software that will effectively identify important mutations in the cancer genome, help with understanding how the cancer genome changes in time and discovering new potential cancer drugs is necessary. As much as the software that does the work is needed, its reliability is a cause of concern. Statistical machine learning approaches are needed that can automatically detect and extract deeply embedded patterns hidden in massive datasets and correlate them to outcomes of interest to the analyst. Machine learning methods such as dictionary learning, reinforcement learning, similarity learning, and transfer learning must be scalable to massive data scales.The software should be consistent in its functioning, considering the fact that it deals with crucial health related conditions and emergencies. The scope of error or benefit of doubt must be infinitesimally small, or rather, close to none. Since, studying the genomic patters and every minute change is the foundation of the end result, reliability at its epitome is what is expected out of the software. In order to have reliable software, high performing computational algorithms will need to be modeled and defined.

Lack of available data:

The biggest challenge in leveraging talent in the data science community to help address medicine's biggest problems. HIPAA restrict the storage, dissemination, and use of patient data to a large extent. Data is the fuel of any analysis and, without it, further developments is strenuous. Researchers face restrictive legal terms and reluctant sharing partnerships. Although some organizations share genomic data sets, the agreements are typically between individual institutions for individual data sets. There exists a need for standardized terms and platforms to accelerate access. For this problem to be tackled, national databases should be established. These databases should provide storage and dissemination of medical data such as charts, results from laboratory tests and medical images. An Opt-In clause needs to be provided for the patients, through which they could volunteer to anonymize their data for longitudinal studies.

Lack of volume and ethnic diversity in the data sets:

An evenly distributed dataset serves as a resource pool for diverse datasets. This not only helps in analyzing and working on several different kinds of sequencing patterns but also leads to advances in treatment for all ethnicities of people ensuring cancer research that is accessible to all. New analytical tools to facilitate networked single-user and collaborative visualization of massive multi-modality datasets must be developed.

Data Processing and Recognition:

Electronic health records contain patient data in a more-or-less standard form that can be shared efficiently, data that can be moved from one location to another at the speed of the Internet. Not all data formats are created in this manner, and some are certainly better than others.

However, any machine-readable format, even simple text files, is better than nothing. While there are currently hundreds of different formats for electronic health records, the fact that they're electronic means that they can be converted from one form into another. Standardizing on a single format would make things much easier, but just getting the data into some electronic form, any, is the first step. There exists a category of response identified as "Don't Know Response." If the group of DK response is small, it is of little significant. But if it is relatively big, it is a matter of major concern. Thus, data processing and recognition had to be dealt with utmost precision.

Data Privacy:

Data Privacy is the foremost challenge as well as concern with respect to advancements in data science. An alarming number of companies are using the technology to store and analyse petabytes of data including web logs, click stream data and social media content to gain better insights about their customers and their business. As a result, information classification becomes even more critical; and information ownership must be addressed to facilitate any reasonable classification. The deployment of Data Science for fraud detection, and in place of security incident and event management (SIEM) systems, is attractive to many organizations. The overheads of managing the output of traditional SIEM and logging systems are proving too much for most IT departments and Data Science is seen as a potential saviour. There are commercial replacements available for existing log management systems, or the technology can be deployed to provide a single data store for security event management and enrichment.

III.Recent work in this regard

I) Olivier Elemento, Cornell Research

Elemento's lab focuses on identifying important mutations in the cancer genome, understanding how the cancer genome changes in time, and discovering new potential cancer drugs. A major thrust of the Elemento lab's research is all about sequencing cancer genomes therein to guide patient treatment and diagnoses. These efforts produced huge amounts of data due to the sheer amount of sequenced DNA. The researchers break up a cancer genome into 100 base-pair long fragments and sequence hundreds of millions of these pieces. Custom software and supercomputers then piece all of the data back together. But sequencing a genome doesn't provide any information on its own. The challenge is in identifying the critical mutations in a genome.

That's where additional measurements on patient samples, big data analytics, and machine learning come in. The researchers perform assays that measure the effect of mutations in the genome. One method is to examine changes in the transcriptome—the entire set of genes that are expressed. These assays create enormous amounts of additional information, which are then integrated with the DNA sequencing data.

Diagonostic model for Thyroid Cancer:

Elemento's lab has built a machine-learning model that predicts whether a patient has thyroid cancer. This is done by analyzing expression levels of specific genes. Thyroid cancer usually presents in the form of a thyroid nodule, a lump that forms at the base of the neck, and around 5 to 15 percent of these nodules are malignant. Using gene measurements of the nodule, the model is able to predict with greater than 90 percent accuracy—higher than standard diagnostic tools—whether a nodule is malignant or benign. The work was published in Clinical Cancer Research in 2012.

II) Data Science Bowl

Data scientists are using machine learning to tackle lung cancer detection. Starting from January, nearly 10,000 data scientists around the world competed in the Data Science Bowl to develop the most effective algorithm to help medical professionals detect lung cancer earlier and with utmost accuracy. Through the large scale competition, several perspectives to a common goal were driven in diagnosing cancer through data science.

National Cancer Institute, USA supplied the Data Science Bowl with 2,000 anonymized, high-resolution CT scans, each image containing gigabytes of data. Sullivan says 1,500 of the images were the training set, accompanied by the final diagnosis. The remaining 500 images were the problem set. Using the training set, competitors' machine learning algorithms had to learn how to correctly determine whether lesions in the lungs were cancerous in the remaining 500 images. The algorithms were scored based on the percentage of correct diagnoses.

III) Dr. Lucas Wartman, Washington University School of Medicine

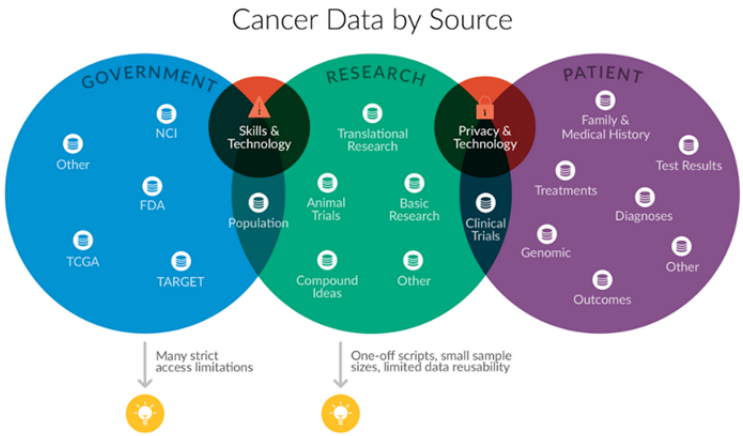
The team involved with Leukemia Research used a supercomputer and one of the university's 26 gene sequencing machines, spent a month of working frantically to beat the cancer clock.

The results exhibited that the cancer DNA had many mutations, but unfortunately, no drugs existed to attack any of them. However, the RNA results had one lead: One normal gene in the leukemia cells was in overdrive, churning out huge amounts of a protein that appeared to be spurring the cancer's growth. There was a new drug, Sutent, that could work on that malfunctioning gene. Though it had only been tested and approved for kidney cancer, the main obstacle to Dr. Wartman's receiving it was actually cost: It was $330 a day.

IV) Data Products Team, Memorial Sloan Kettering Cancer Centre

The team works on projects related to clinical research and the 'machine learning', "teaching" its software how to understand the clinical essays with an unprecedented detail level. They have already been able to develop a specialized programme that was implemented in the center.

The department is in charge of training Watson, IBM's supercomputer, so that it would be able to understand a cancer case and would try to replicate the decision process used by the doctors of the centre, with the objective to develop, in the future, tools that would contribute to the diagnosis. The research team is integrated by eight people: three data scientists, three software engineers, a designer and Paul, the team chief. The team intend to obtain effective cure with the help of skilled data scientists and reliable data from various sources.



Source: Dan Wanger, Chief Analytics Officer, Cancer Moonshot Blog

IV. FUTURE SCOPE

Data science has already produced research that has helped patients. For instance, a data network for children with Crohn's disease and ulcerative colitis called ImproveCareNow helped increase remission rates for sick children. The scope of data science is thus large in the field of medicine.

Therefore, a need exists for skilled data scientists and computational engineers who could put the data to good use, such that it is more useful, fluent and visual. This process can accelerate when data is available in large amounts. Researchers are slicing and dicing digital data to reveal similarities between tumors in different parts of the body that were not apparent before. The majority of medical data is considered confidential, and therefore cannot currently be used without the consent of the individual concerned. This has presented huge difficulties for companies using Machine Learning, which requires a large pool of data from which it can learn to detect conditions as efficiently as a human. Although these companies are fighting hard to be allowed access to the necessary data, it is the decision of the patients themselves whether they wish for their medical information to be used.

Data Science offers enormous opportunity for healthcare around the world to tackle some of the deadliest conditions head on. It does however, require a compromise from the patients and medical professionals. There is a need to pool medical records in order to provide doctors and data scientists a larger spectrum of data with which to research. Resulting in the best chance of diagnosing, treating or even stopping these conditions before they can take any more lives.

The future of cancer treatment is whole genome sequencing. Knowing the genetic makeup of the tumor will help doctors choose drugs to attack the problem genes and stop the cancer. However, the method is now in an infantry stage and is quite complex. The method is too expensive to be widely available. On a positive note, the cost of sequencing is dropping steeply, and research on genes is exploding, so medical experts believe that soon, whether it's a few years or a decade from now, genetic analyses of cancer will become routine. To get ahead of the curve, drug companies and small biotech firms are working on drugs that attack specific genes rather than types of tumors. Similarly, cancer researchers are finding companies to track down genes that cause cancer to grow, and to find and test drugs against these genetic culprits. Venture capital firms are also getting into the field.

III. Conclusion

In order to make data available for effective cancer research, patients should be willing to contribute data easily, companies and organizations need to come up with legal consent and funding to carry out such high performing algorithms and computational methods and generate datasets that are diverse.

The whole genome sequencing method will be a success and widely accessible only when its need of raw materials, that is, data sets is obtainable and cost effective. Intervention from the government and health agencies, with respect to the cost and procedural skill of data scientists involved will pave a new high for the betterment of the technology in itself to cure cancer. Technical improvements have decreased sequencing costs and, as a result, the size and number of genomic datasets have increased rapidly. Because of the lower cost, large amounts of sequence data are now being produced by small to midsize research groups.

REFERENCES

[1] International Council for Science : Committee on Data for Science and Technology. (2012, April). CODATA, The Committee on Data for Science and Technology. Retrieved from International Council for Science : Committee on Data for Science and Technology: http://www.codata.org/

[2] ."The Cost of Sequencing a Human Genome". www.genome.gov.

[3] . McGuire, Amy, L; Caulfield, Timothy (2008). "Science and Society: Research ethics and the challenge of whole-genome sequencing". Nature Reviews Genetics.

[4] . "Complete Human Genome Sequencing Datasets to its Public Genomic Repository". Archived from the original on June 10, 2012

[5] Article: “How datascience is transforming health” By Tim O'ReillyMike LoukidesColin Hill May 4, 2015. www.oreilly.com/ideas/how-data-science-is-transforming-health-care

[6] Borry P; Fryns JP; Schotsmans P; Dierickx K (February 2006). "Carrier testing in minors: a systematic review of guidelines and position papers". Eur. J. Hum. Genet.

[7] . Article: “Cancer and Big Data Analytics” By Alexandra Chang, Cornell Research, 2015.