ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Review on Data Science in Biotechnology.

Selvarajan E[1,2*], and Punya Swaroop S[2].

[1]Department of Genetic Engineering, School of Bioengineering, SRM University, Kattankulathur, Tamil Nadu, India.
[2]School of Biotechnology, National Institute of Technology, NIT Calicut.

ABSTRACT

Data can be used in analyzing genomes, prescribing drugs in pharmaceutical, Biotechnology industries and many other areas. Now data turned into "BIG DATA" and its analysis is a whole new field of "Analytics and Data Science". With the help of analytics the cost of analysis of genome structures came down to 1000 dollars from 10 million dollars in 2007. Cancer drug was prescribed based on analysis from big data. It has whole lot of applications in many industries and businesses. There are many companies purely based on analytics and many start ups coming in the same area. Ten reasons why Biotechnology needs Big Data are clearly stated. Some of the reasons include Genomics, Drug Discovery, Electronic clinical research etc., Data mining is creating a portable business opportunities for the biotechnology industries

Keywords: Big data, Bioinformatics, Data Science, Drug Discovery

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Data Science (DS) is the practice of the repetitive, methodical analysis of huge volume of data emphasizing on Statistical analysis. DS is used to gain understandings that notify business decisions and can be used in optimization of business processes. DS tools include Statistical Analysis, data mining, Predictive modeling and multivariate testing.

When gathering huge quantities of data, significant human remedial input is necessary, this sprang up the new stream namely crowd sourcing. The suitable example is mechanical turk owned by Amazon. Possibility of modern collection of data is achieved by cloud computing and the circulation of the data over many physical resources that can be remotely accessed, rather than focused at one location [1].

While the theory of data science has been developing for many years, the belief of a data scientist has turned out to be an excellent and demanding career resulting in rise of a new era of data scientists. The core knowledge in technology development importantly unravels the astonishing growth rates of "Big Data". World Wide Web and Technology innovation cater to the development of novel data types- such as content generated by user along with tools that help interpret it [2]. Social media platforms like Facebook rely on data science to develop new, interactive features that pushers users to get involved and stay that way- all so that we know it's important.

IMPORTANCE OF BIG DATA

The usage of Big Data is becoming a vital way for leading companies to outplay their peers. In most industries, deep rooted competitors and neophytes alike will control data-driven strategies to transform, contest, and seizure value. Early examples of such use of data in every sector can be found without much effort. In healthcare, data pioneers are studying the health outcomes of pharmaceuticals when they were used prevalently and discovering aids and perils that were not conspicuous during necessarily more clinical trials. Other early adopters of Big Data are taking in data from sensors installed in products from children's toys to industrial goods to find out how these products are used in the real world. Such knowledge then informs the conception of new service offerings and the manufacturing of future products.

Big Data in instrumental in creating new growth chances and totally novel categories of companies, like those that gather and analyze industry data. Most of them will be companies that control large information flows where data is relevant to products and services, shoppers and suppliers, consumer penchants and intent can be understood and analyzed. Enterprising leaders across varied sectors could begin pass to build their organizations Big Data capabilities.

Value created by the use of Big Data

If the United States healthcare system were to use big data resourcefully and effectively to boost efficiency and quality, the sector are better off creating around 300 billion dollars in value per year. Two-thirds of that particular figure would mean that around nine percent decline in U.S. healthcare spending. In the developed economies of Europe, government bureaucrats would create more than 123 billion dollars in operational efficiency enhancements just by using Big Data and that does not include employing advanced analytic tools to lessen fraudulent and errors and increase the collection of tax revenues.

But it is not only companies and industries that strive to gain from the value that Big Data can harness. Consumers can reap highly substantial benefits. For instance, users of services allowed by personal-location data can attract 600 billion dollars in consumer surplus.

Literature Survey

The first challenge is to specify which Big Data business model (in science-based activity) will render IT services to biotechnology and life sciences firms along with research laboratories. The second challenge is to set a methodology to industry a still universally unknown service for these firms.

Since 2011 Big Data was identified as an emerging market considering the availability of huge amount of commercial and marketing data. Life sciences are known also to generate a deluge of data as an untapped source of information. The main approach was to identify what is specific for Life Sciences and what managers in Life Sciences companies and laboratories may expect from a Big Data activity-based company. Life Sciences are used to deal with large amount of data, most of them in well known structured formats. The question was to know what additional actionable information could be provided by Big Data technologies and analysis. We tried also to evaluate the needs and expectations of biotechnology and life sciences companies and research laboratories regarding data search and analysis using a survey on-line addressing life sciences companies and laboratories contacts. Most of responders require anonymised and secure data analysis and expect actionable information to launch new biotech product or to confirm a strategy.

Researchers proposed a three dimensions common framework of Big Data: Volume, Variety and Velocity known as the Three Vs [3]. Big Data in itself is worthless and requires data analysis to retrieve or acquire intelligence from the data and help in decision-making. Big Data processes or pipeline of extracting insights can be divided in two sub-processes: data management and data analytics. This process can be defined as a Business Intelligence system (BI) [4].

Most of the authors use to represent the Big Data analysis pipeline from a computer-driven process perspective. The role of data visualization as a result of data management and as a tool for data analysis is very significant. Data visualization objective is to present information clearly and efficiently to viewers using developed graphics. Data visualization is a recent field presented as one of the steps of data science [5] and a tool to communicate information from complex data sets. Big Data in Biology or in Life Sciences or in Health attributes to Big Data tuning advantageously in a specific stream or market so called Life Sciences (considering health as a Life science). Earlier, Biologists made use of the term bioinformatics to depict the procedures and techniques required for management of the generated data through biological and medicinal research. The concept of bioinformatics even though primarily devoted to genomic data has gone into disuse ever since the theory of Big Data originates as a viral buzzword in 2010.

REASONS WHY BIOTECH NEEDS DATA SCIENCE

Genomics

Numerous software companies have evolved over the past several years to handle or manage the interpretation challenges in genomics, including DNAnexus, Knome and NextBio. As per the NIH's National Human Genome Research Institute, cost of decoding an entire genome was around 10 million dollars in early 2007. From that point of time, affected largely by next-gen sequencers which came into light, that price has plummeted down faster and is now approaching 1,000 dollars per genome. Illumina and others sequence the entire genomes in hours [6].

At a startup namely NextBio, for instance, tech geniuses have made use of frameworks for easy working of massive computing tasks from Google to analyze biological data [7]. And the company recently colluded with computer chip giant Intel to improve Hadoop applications for Big Data analysis in genomics and Biotechnology. The technologies used in Data Science are represented in table 1.

Table 1: Technologies used in Data Science

Technologies	Description	Application and users
Hadoop	Java platform of Apache Software foundation for distributed applications and data management	Yahoo, Oracle, IBM, EMC, NetApp
Hive	Data analysis software using Hadoop	Facebook
NoSQL Database	Distributed NoSQL database Cassandra	Facebook, Google
Map/Reduce	Development platform for distributed data treatment	Google
Machine learning	Algorithm that can learn and make predictions on data	Amazon
Real time business intelligence	Storm, Spark	Twitter, Yahoo

ISSN: 0975-8585

Drug Discovery

Parma companies screen millions of compounds before taking the decision on competitors for testing in preclinical trials, and most number of firms has brought many different software tools into the mix required to automate discovery process. As of now, drug discovery takes a lot of time and costs huge amount of money, plus compounds in clinical trials have a low 1-in-10 shot at success. Predictive modeling which is a part and parcel of data science offers a faster way to close in on drug candidates [Figure 1].

Numerate, which has tied-up with drug makers Boehringer Ingelheim and Merck, builds virtual assays or predictive data models that are designed using the knowledge of drug target and treatment goals. Then the company releases the models to analyze huge number of virtual libraries of compounds that account to terabytes and potentially even petabytes of data. It spent months to achieve success in an HIV drug project.

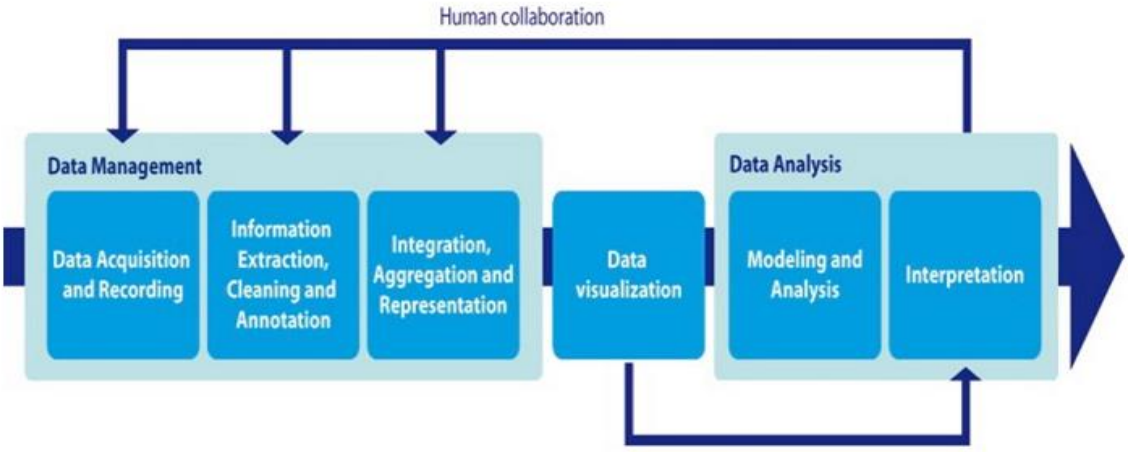

Fig. 1: Importance of Data visualization [8].

Crowdsourced RD

Social networking sites gather millions of health-related comments every day, as consumers thrive on online to retrieve and spread information relevant to diseases. Parma companies rely over the possibilities to exploit on the social web trend, and a increasing number of online patient groups and social analytics tools have arose and furnish to drug makers desires. Meantime, Foldit has headed the way in bringing together its thousands of online gamers to challenge research challenges and solve protein structure puzzles associated to illnesses. 23andMe, the personal genomics company, has relied on growth of patient-driven research with its recent take-over of Cure Together (CT). CT provides online tools and surveys that allow patients to carry out their own studies and give knowledge about their conditions for the aids of others. Both Cure Together and 23andMe have made huge amounts of data from patients, and 23andMe sees an opening to control the tools and capabilities of both companies to create new genetic discoveries and attract patients to take part in research

Drug Safety

Analytics and modeling proffer the possibility to avert deaths from adversative events and other risk elements. For example, GNS Healthcare associated last year that it's working with Brigham and Women's Hospital in Boston to advance predictive models of adverse drug responses and hospital readmissions for patients who suffer with congestive heart failure, providing supercomputing-enabled tools to resist awful outcomes. In nearby Cambridge, MA, incidentally, Novartis scientists have applied computer models embedded with collaborators at the University of California-San Francisco with troves of side-effect data to forecast whether a compound will spawn bad reactions before it enters clinical trials.4 processes which comprise Data processing, Bayesian Fragment Enumeration, Parallel include sampling and Model Intervention sampling.

ISSN: 0975-8585

Drug Recycling

Drugs frequently impact genes accidentally, causing side effects in some situations but handling serious medical conditions in others. The latter could assist drug developers in reusable drugs, also known as drug reutilizing. More recently, Stanford University scientists headed by Dr.Atul Butte expounded how their algorithms related to large public databases of gene and drug information founded brand new uses for aging meds. Butte is one of the founders of a startup called NuMedii, which has used the Stanford technology and harnesses vast amounts of molecular data to find new uses in place of old drugs. National Institutes of Health chief Francis Collins has acknowledged such approaches, which offer a way to rapidly bring new treatments to patients using drugs with known safety profiles.

Business development

Every day the whole outlet of information on scientific discoveries and pharma advances from all round the words spans across many diverse sources, presenting aflood of data for biopharma out sifts to sift through to find potential licensing chances. Some Big Pharma and biotech companies have chosen analytics and data-mining technologies to scour disparate Big Data sources and deliver the correlated information they seek. Cambridge Semantics in Boston has out played some of the world's leading pharma groups with a semantic web technology, which allows business development groups to involuntarily pull together key data from a extensive variety of diverse sources such as scientific publications, websites and databases and internal troves of knowledge. And in nearby Cambridge, MA, Relay Technology Management has lately let out a software-as-a-service product that uses data-mining tech and Tibco Software's Spotfire to deliver information to pharma dealmakers, using Relay's algorithm for truly evaluating biotech assets such as compounds. Fig 2 shows comparison of Big Data and Bioinformatics.

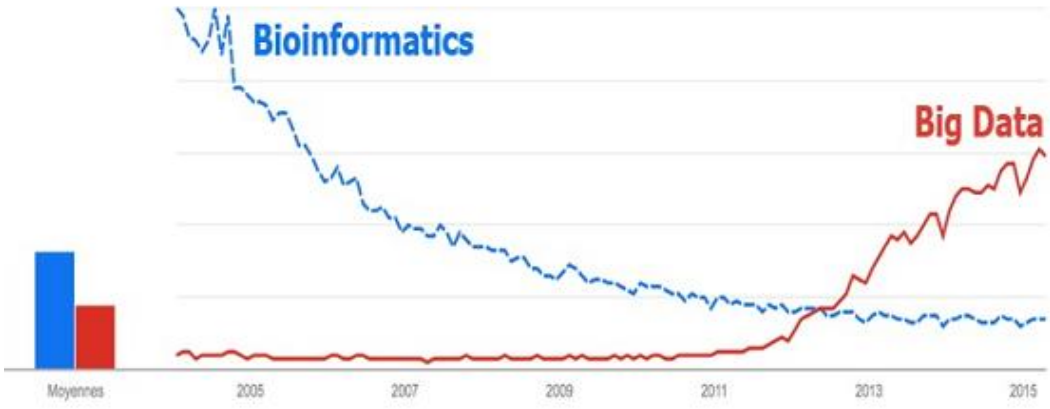

Fig. 2: Bioinformatics Vs Big Data [9]

Electronic clinical research

The U.S. government strategies to issue 19 billion dollars into enhancing espousal of electronic health records, an investment that intent to digitize the medical histories of crores of patients. While countless of the digital records systems are deprived interoperability, pharma companies and other stakeholders have begun to find ways of investing vast amounts of electronic patient information to support and propagate clinical research. For example, Partnership to enhance Clinical Electronic Research (PACeR) in New York proposes to scheme a system that would allow investigators and sponsors to procure from the application of electronic patient records to realize patients for clinical trials and some more studies. While some sticky legal, technical and privacy issues are hindering the progress, the associates of the nonprofit organization include a different kind of drug makers and developers, hospitals, health systems and software companies with a bestowed interest and recognizing this vision.

Catching Drug Fraud

Fraudulent drugs kill people, majorly in developing countries, who believe they are taking the real ones to battle their illnesses. In this pathetic scenario, patients and their families lose and, from a business position, creators of the real drugs miss sales. The World Health Organization approximates that 700,000 patients in Africa die as a result of fake versions of anti-malaria and tuberculosis meds and the difficulty costs drug makers 75 billion dollars annually. This problem has instigated the startup Sproxil to tech giant IBM to allow drug companies to evaluate Big Data sources to spot patterns of phony drug activity. Sproxil aims to collect huge amounts of transactional data with a system that aids patients to text-message codes from medicine bottles to study whether the meds are authentic or not. With IBM's conception tech and other analytics, drug makers can knock petabytes of data on the drug transactions in actual time, according to Big Blue. Presumably, prescription drug swindles can be spotted with in short time with those capabilities. And Sproxil has been working with Merck KGaA and GlaxoSmithKline to certify that patients in Africa are getting the manufacturers' drugs and not fakes.

CONCLUSION

We have come across the importance of Big Data in 2016, where Bioinformatics went in disuse to give the Big Data a role to become one significant tool to analyze the gene sequences. As of now, there is a global shortage of Data Scientists. This above statement can be illustrated by this: A new species of techie is in demand these days not only in Silicon Valley, but also in company headquarters around the world. Data scientists are the new superheroes, says Pascal Clement, the head of Amadeus Travel Intelligence in Madrid. A study by McKinsey projects that by 2018, the U.S. alone may face a 50 percent to 60 percent gap between supply and requisite demand of deep analytic talent. The shortage is already being felt across a broad spectrum of industries, including aerospace, insurance, pharmaceuticals, and finance.

ACKNOWLEDGEMENT

The authors are grateful to National Institute of Technology Calicut, Kerala, India for providing tremendous facilities and support to carry out the research work.

REFERENCES

[1]. Diebold, F.X. (2003) Big Data Dynamic Factor Models for Macroeconomic Measurementand Forecasting: A Discussion of the Papers by Reichlin and Watson," In M. Dewa-tripont, L.P. Hansen and S. Turnovsky (eds.), Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society, Cambridge University Press, 115-122

[2]. Wang R (2012) What a Big-data Business model looks like?. Havard Business review.

[3]. Laney D (2001) 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6.

[4]. Wang Y, Liu Z. Study on Port Business Intelligence System Combined with Business Performance Management. Proceedings of the 2009 Second International Conference on Future Information Technology and Management Engineering, Washington, DC, USA: IEEE Computer Society; 2009, p. 258–260.

[5]. Friedman V (2008) Data Visualization and Infographics in: Graphics, Monday Inspiration, January 14th, 2008

[6]. Dulbecco R (1986) Turning Point in Cancer Research, Sequencing the Human Genome. Science 231 (4742): 1055-1056.

[7]. Groves P, Kayyali B, Knott D, Van Kuiken (2103) The big-data revolution in US health care: Accelerating value and innovation. McKinsey & Company report

[8]. Gandomi A, Haider (2015) Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35: 137-144.

[9]. Mordret G (2015) Big data in science: which business model is suitable?. Journal of Antibody Drug Cojucates.