PhUSE US Connect 2019

Paper ML07

Machine learning for better treatment with Omics and Clinical Data

Phani S Srinivasan Ponnapalli, Syneos Health, Raleigh, USA
Subrahmanyam Rayaprolu, Syneos Health, Raleigh, USA

ABSTRACT

Data Driven Science is reshaping the world and is opening new opportunities across industries. In recent years, Life Science companies started embracing digital information and applying data science to optimize the time from "Lab to Life". Big Data and Machine Learning (ML) technologies provide an opportunity for Life Science firms to unlock answers to a host of high-value questions such as the true effectiveness of treatments by integrated data driven analysis of clinical trial and omics Data. This paper introduces omics data, NoSQL database, concepts of machine learning and provides insight on ML algorithms classified as Supervised, Unsupervised, Artificial Neural Network and Deep learning. Further, this paper explores Data-Driven architecture to load omics data using NoSQL and apply machine learning. Finally, this paper will show how to apply machine learning on integrated data for identifying the hidden patterns and better treatment decisions using Python libraries.

INTRODUCTION (ML AND OMICS DATA)

Advances in Digital data sciences and Bio-Informatics have resulted in the new era of precision and value based medicine.

Bringing Big Data Analytics to biological data-sets (omics) and clinical information will transform the way we look at disease and provide better health-care.

Applying AI/Machine Learning capabilities on Genomics, transcriptomics and proteomics integrated with clinical data demonstrate compelling benefits, including:

- Discovery of hidden patterns and actionable insights
- Precision Medicine
- Prediction and analysis of clinical trial outcomes
- Develop new drugs faster and reduce the costs of clinical trial

This paper would explore the possibilities of Machine Learning on "Clinical + OMICS" data by discussing:

- What is OMICS Data?
- Big Data and NoSQL
- Concepts of Machine Learning
- Insight on Machine Learning Algorithms.
- Python Programming Language for OMICS Data Parsing.
- Sample Python code for applying machine learning on OMICS data integrated with Clinical Data.

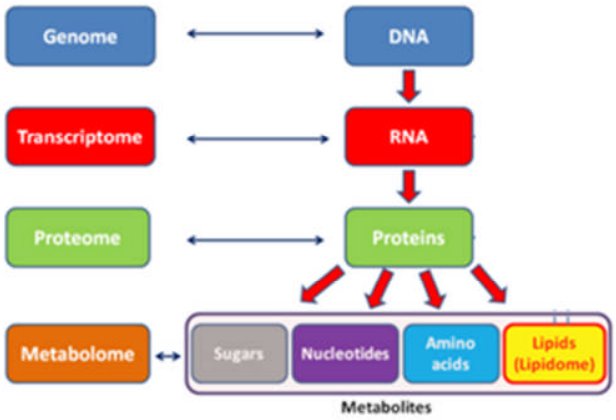
PhUSE US Connect 2019

WHAT IS OMICS DATA?

Technologies that measure some characteristic of a large family of cellular molecules, such as genes, proteins, or small metabolites, have been named by appending the suffix "-omics," as in "genomics." Omics refers to the collective technologies used to explore the roles, relationships, and actions of the various types of molecules that make up the cells of an organism.

Omics technologies provide the tools needed to look at the differences in DNA, RNA, proteins, and other cellular molecules between species and among individuals of a species. These types of molecular profiles can vary with cell or tissue exposure to chemicals or drugs and thus have potential use in toxicological assessments. Omics experiments can often be conducted in high-throughput assays that produce tremendous amounts of data on the functional and/or structural alterations within the cell.

Deoxyribonucleic acid called as DNA is the building block of the life. It contains the information the cell requires to synthesize RNA, protein and to replicate itself. DNA is the storage repository for the information that is required for any cell to function. The human genome contains approximately 3 billion base pairs of DNA, encoding approximately 20,000 genes which carry the instructions for making proteins. Advances in Bio-Informatics are providing access to the biological data sets generated by the omics technologies. There is exhaustive list of omics data types, but the following omics are mainly being analyzed for better medicine.



Genomics – A genome is an organism's complete set of DNA including all of its genes. Genomics is an interdisciplinary field of science focusing on the structure, function, evolution, mapping, and editing of genomes.

An individual's genomic information is an important determinant in influencing the state of health and disease. Examining this genetic background is of great importance for identifying mutational changes in DNA that discriminate health and disease.

Transcriptomics – The transcriptome is the set of all ribonucleic acid (RNA) transcripts in a cell. The entire DNA sequence provides a template for the synthesis of a variety of RNA molecules like, mRNA, tRNA and rRNA. The study of all coding and non-coding RNA molecules in a cell and their functions is collectively called as transcriptomics.

The analysis of mRNAs provides direct insight into cell and tissue specific gene expression. This information is fundamental for a better understanding of the dynamics of cellular and tissue metabolism, and to understand how changes in the transcriptome profiles affect health and disease.

Proteomics – Proteomics is the study of proteins and proteome is the entire set of proteins in a given cell. RNA acts as a messenger carrying instructions from DNA for controlling the synthesis of proteins.

Studying the human proteins helps in the identification of potential new drugs for the treatment of disease. This relies on proteome information to identify proteins associated with a disease, For example, if a certain protein is implicated in a disease, its 3D structure and protein-protein interaction analysis provides the information to design drugs to interfere with the action of the protein.

Metabolomics - Proteins, carbohydrates, and fats in the food we eat are metabolized into amino acids, sugars, and fatty acids, providing the building blocks of cells and tissues. The study of the set of metabolites present within an organism, cell, or tissue is called Metabolomics. The study of these compounds and their relationships in a living organism provides a new avenue for exploration in clinical research and to detect cancer in advance.

BIG DATA AND NOSQL:

Big Data

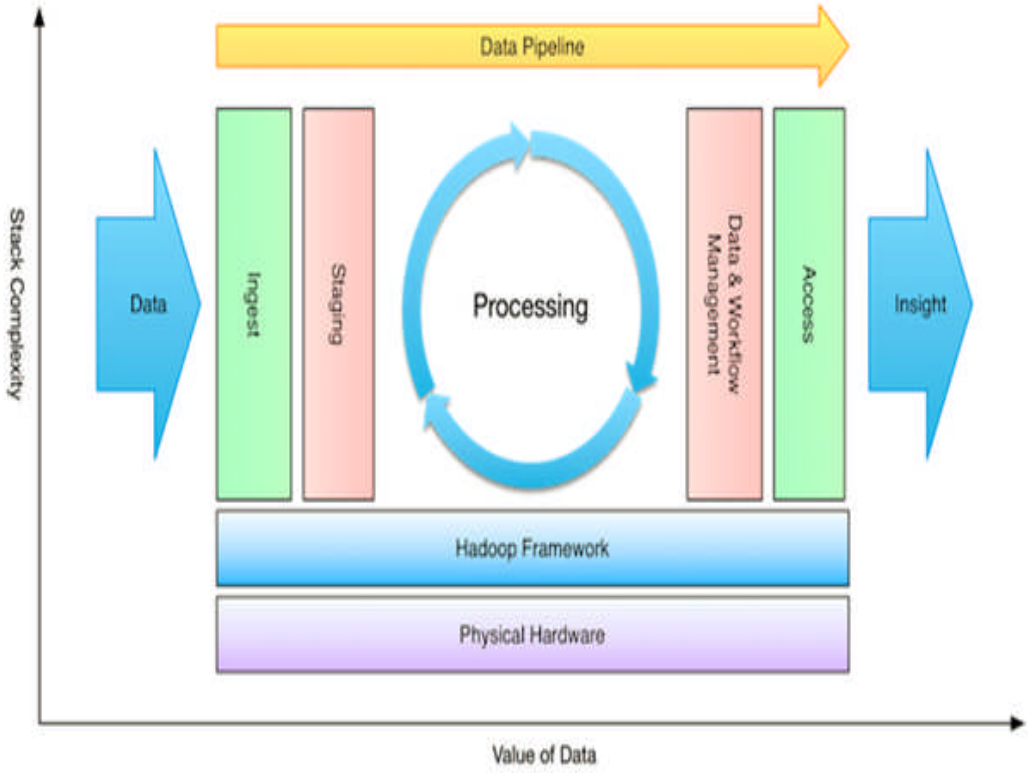
Big Data architecture is designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database systems. Big data solutions work on a fundamentally different principle to handle device data and streaming data. Big Data architecture has the following characteristics.

- Distributed Data and parallel processing: Big Data solutions store huge data in a distributed manner in a file system. Process the data in parallel on a cluster of nodes. In simple words, Data is broken in to small pieces of information and stored in multiple blocks called HDFS. Tasks are executed in parallel by processing these blocks in parallel and results are merged back, is called map reduce.

- Fault Tolerance: Big Data solutions work on failure to tolerance by redundancy. The same information is stored in multiple places called racks. There are multiple machines available as cluster for processing. If any machine in the cluster goes down still system works due to data redundancy and multiple machines.

- Scalability: Big Data systems are very flexible in scaling storage space and computing power on fly whenever required.

- Cost effectiveness: Big Data systems like Hadoop are open source and uses commodity hardware. They do not require a very high-end server with large memory and processing power. This makes the system very cost effective.



PhUSE US Connect 2019

NoSQL

NoSQL is an approach to databases that represents a shift away from traditional relational database management systems (RDBMS).NoSQL databases do not rely on predefined schema structures and use more flexible data models. NoSQL can mean “not SQL” or “not only SQL.” NoSQL is particularly useful for storing unstructured/semi-structured data.

Features of NoSQL:

- Minimal data modeling
- Minimal/no ETL
- No pre-defined schema necessary

Types of NoSQL Databases

Several different varieties of NoSQL databases have been created to support specific needs and use cases. These fall into four main categories:

Key-value data stores: Key-value NoSQL databases emphasize simplicity and are very useful in accelerating an application to support high-speed read and write processing of non-transactional data. Stored values can be any type of binary object (text, video, JSON document, etc.) and are accessed via a key. The application has complete control over what is stored in the value, making this the most flexible NoSQL model. Data is partitioned and replicated across a cluster to get scalability and availability. For this reason, key value stores often do not support transactions. However, they are highly effective at scaling applications that deal with high-velocity, non-transactional data.

Document stores: Document databases typically store self-describing JSON, XML, and BSON documents. They are similar to key-value stores, but in this case, a value is a single document that stores all data related to a specific key. Popular fields in the document can be indexed to provide fast retrieval without knowing the key. Each document can have the same or a different structure.

Wide-column stores: Wide-column NoSQL databases store data in tables with rows and columns similar to RDBMS, but names and formats of columns can vary from row to row across the table. Wide-column databases group columns of related data together. A query can retrieve related data in a single operation because only the columns associated with the query are retrieved. In an RDBMS, the data would be in different rows stored in different places on disk, requiring multiple disk operations for retrieval.

Graph stores: A graph database uses graph structures to store, map, and query relationships. They provide index-free adjacency, so that adjacent elements are linked together without using an index.

Multi-modal databases leverage some combination of the four types described above and therefore can support a wider range of applications.

Relational Table Vs NoSQL

Relational Table

Product_id	Product_name	Product_price
2	Product 2	3000

NO SQL

```
<Table >
<product >
<product_id>2</product_id>
<product_name>product 2</product_name>
<product_price>3000</product_price>
</product>
```

PhUSE US Connect 2019

CONCEPTS OF MACHINE LEARNING

Machine Learning is an application of artificial intelligence (AI) that provides systems or machines the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Insight on Machine Learning Algorithms

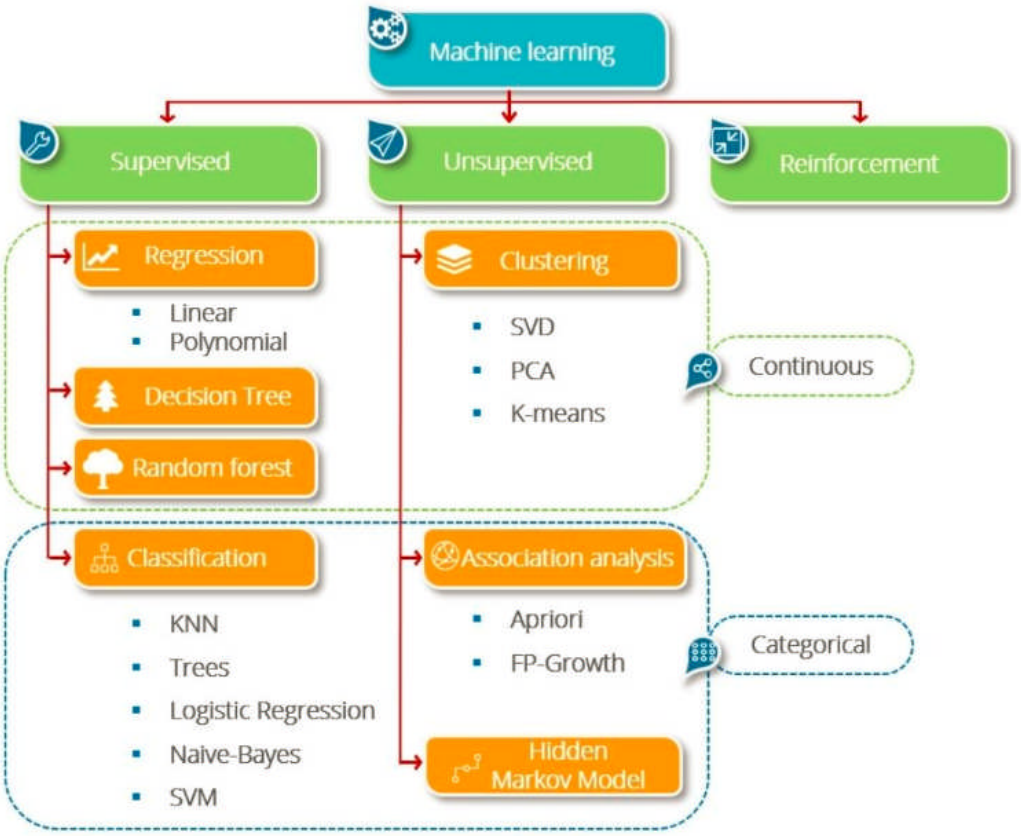
Machine Learning algorithm requires three main functions.

Hypothesis Function:
Hypothesis function is basically model for the data.

Cost Function:
The cost function measures how well hypothesis function fits into data. It calculates the difference between actual data point and hypothesis data point.

Optimization:
Machine Learning uses gradient descent to find the best model of the data. Gradient Descent will minimize the cost function and eventually find the best model for the data.

Machine Learning Types



Supervised Learning

In supervised learning the data is labeled in order to assist the machine to understand whether they are making the right kind of progress or not. Then the machine applies the learning from the past data that has been labeled to predict the events of the future. Upon the analysis of the training dataset, the learning algorithm can help to infer the functions by making the predictions form the output values. Once sufficient amount of training is done the system is able to take new inputs and give new targets. The machine is also able to rectify itself, make course correction to be in line with intended output and find the mistakes in the previous output and modify the entire model of working in the real world with utmost precision.

Supervised learning can be further grouped into regression and classification.

Classification: Classification algorithm is where the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.

Regression: Regression algorithm is when the output variable is a real value, such as “dollars” or “weight”.

Unsupervised Learning

Unsupervised learning on the other hand is one wherein the machine is given any labeled data. This means there is no classification of the data. The unsupervised learning algorithms are based on the premise of how the machines will be able to infer from a certain the function the hidden structures in unlabeled data. This way the system is not able to figure out the right output from the wrong output but upon large amounts of data being fed into the system the machine is able to find out the hidden patterns and draw conclusions about what might be the right output. The hidden structure in the unlabeled data is explored.

Unsupervised learning can be further grouped into Clustering and Association.

Clustering: Clustering problem is where we can discover the inherent groupings in the data.

Association: Association rule learning is where we can discover rules that describe large portions of your data.

Reinforcement Learning

Reinforcement Learning is the type of learning wherein it is not possible to write all the rules or label all the data as right or wrong due to the sheer amount of data or the fuzzy nature of the data. The machine is exposed to an environment where it trains itself continually using trial and error.

Artificial Neural Network (ANN)

Artificial Neural networks (ANN) or neural networks are computational algorithms.
It intended to simulate the behavior of biological systems composed of “neurons”. ANNs are computational models inspired by an animal’s central nervous systems. It is capable of machine learning as well as pattern recognition. These presented as systems of interconnected “neurons” which can compute values from inputs.

A neural network contains the following 3 layers:

Input layer: The activity of the input units represents the raw information that can feed into the network.

Hidden layer: To determine the activity of each hidden unit, activities of the input units and the weights on the connections between the input and the hidden units. There may be one or more hidden layers.

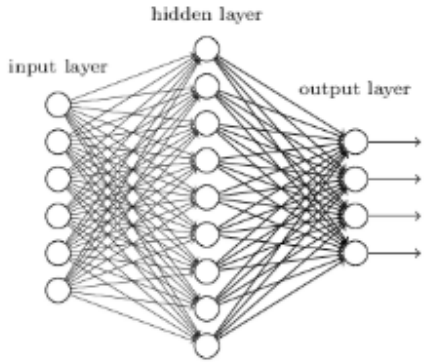
Output layer: The behavior of the output units depends on the activity of the hidden units and the weights between the hidden and output units.

PhUSE US Connect 2019

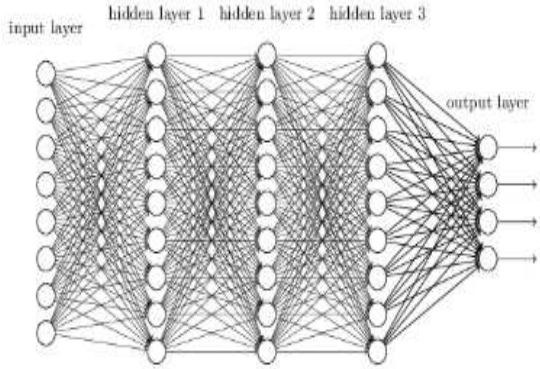
Deep Neural Network (DNN)

A deep neural network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers. The DNN finds the correct mathematical manipulation to turn the input into the output, whether it be a linear relationship or a non-linear relationship. The network moves through the layers calculating the probability of each output.

ANN



DNN



LOADING OMICS DATA AND INTEGRATION WITH CLINICAL DATA

Next Generation Sequencing (NGS) technology has resulted in massive amounts of OMICS data consisting of characteristics like Variety, Velocity, Veracity, Volume and Potential Value. In clinical domain, data of many patients is collected for intelligent decision making. This data needs to be stored and analyzed in an efficient way that would be helpful for making best decisions. Different data mining and machine learning algorithms require input data in specific types and formats.

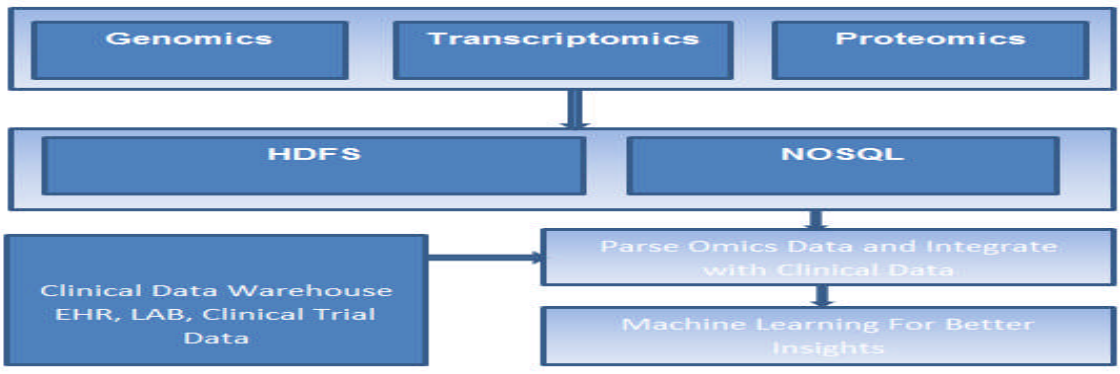
OMICS data is available for processing in various formats such as

1. Images/Videos
2. *JSON*
3. *Fastq and FASTA Format*

Above mentioned formats are mostly unstructured / semi-structured do not fit the enterprise relational data model. Big Data and NOSQL frame work is the most suitable architecture for loading this Data.

OMICS data is to be parsed and integrated with clinical trial data to obtain value out of it. Applying Machine Learning and AI algorithms on the integrated information will open the door of hidden insights and better treatment decisions.

Architecture



PhUSE US Connect 2019

MACHINE LEARNING ON INTEGRATED DATA WITH PYTHON LIBRARIES

Python is a popular open source programming language and it is one of the most-used languages in artificial intelligence and Machine Learning fields.

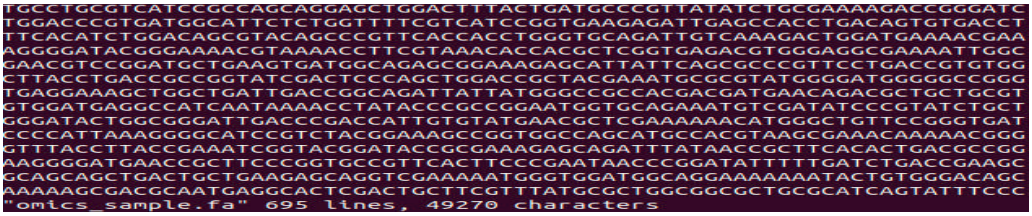
Anatomy of Basic Python Program

```
# Import Statements of Libraries
# Variable Definitions
# Classes Definitions
# Function Definitions
# Processing
```

Parsing OMICS data with Python:

Omics data is semi structured data which can be parsed through python using text processing operations.

Example:
Genome file is fasta (".fa") file which can be loaded as text file to python and can parsed trough libraries.



Reading genome and parsing it with string functions:

```
def readgenome(filename):
genome=' '
with open(filename,'r') as f:
for line in f:
if not line[0] == '>':
genome += line.rstrip()
return genome
genome = readgenome('omics_sample.fa')
```

Once file loaded to Python environment we can parse it by libraries like collections. We can get count of each basis by using counter module.

```
Import Collections
Collections.Counter(genome)
```

Sample Output of counter object:

PhUSE US Connect 2019

PYTHON MACHINE LEARNING LIBRARIES AND SAMPLE CODE

Python has machine learning libraries that enable developers to apply data science algorithms. It implements popular machine learning techniques such as recommendation, classification, and clustering. Therefore, it is necessary to have a brief introduction to machine learning before we move further.

Machine learning algorithms require input data in a particular format. We could use parsing and ETL techniques to create proper data.

Sample Input Data with various features calculated multiple subjects

Subject_id	Feature_1	Feature_2	Feature_3
U101	1.5	30	100
U102	2.6	20	100
U103	4	30	100
U104	5.6	20	100

Following are general packages and libraries used to apply machine learning Algorithms with Python

- numpy - is used for its N-dimensional array objects
- pandas – is a data analysis library that includes dataframes
- matplotlib – is 2D plotting library for creating graphs and plots
- scikit-learn - the algorithms used for data analysis and data mining tasks
- seaborn – a data visualization library based on matplotlib

K-means Clustering:

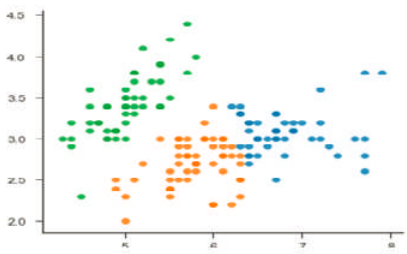
It is a type of unsupervised algorithm which deals with the clustering problems. Its procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). Data points inside a cluster are homogeneous and are heterogeneous to peer groups.

Patients can be stratified by applying Clustering algorithm (OMICS data + Clinical Data). We can understand how a patient's unique molecular and genetic profile makes them responsive for certain medications. Better treatment decision could be made by tailoring the medical treatment to the individual characteristics of each patient. This will enhance the ability of a clinical trial to predict which medical treatments will be safe and effective for each patient, and which ones will not be.

Code Snippet

```
from sklearn.cluster import KMeans
df = pd.read_csv('sample_clin_omics_summ.csv')
df.columns = ['X1', 'X2', 'X3', 'X4', 'Y']
df= df.drop(['X4', 'X3'], 1)
X = df.values[:, 0:2]
kmeans = KMeans(n_clusters=3)
kmeans.fit(X)
plt.scatter(X[:,0], X[:,1], c=kmeans.labels_, cmap='rainbow')
```

Sample Output



CONCLUSION

Machine learning offers tremendous opportunities to more efficiently access and understand vast amounts of data produced by bio-informatics Industry. Advances in technology and reducing costs of Next Generation Sequencing are unlocking the doors for new era of clinical research and opportunities for better treatment.

REFERENCES

1. http://alttox.org/mapp/emerging-technologies/omics-bioinformatics-computational-biology/
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6018996/
3. https://en.wikipedia.org/wiki/Transcriptome
4. https://www.lexjansen.com/pharmasug/2018/SI/PharmaSUG-2018-SI06.pdf
5. https://www.sciencedirect.com/science/article/pii/S2001037014000464
6. http://blog.cloudera.com/blog/2014/09/getting-started-with-big-data-architecture/
7. https://www.analyticsvidhya.com/blog/2013/07/big-data/

ACKNOWLEDGMENTS

The authors would like to thank Carol Frazee, Associate Director, Syneos Health, for her valuable comments and feedback.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Phani Ponnapalli
Enterprise: Syneos Health
Work Phone: +1 732.666.8572
E-mail: phani.ponnapalli@syneoshealth.com

Name: Subrahmanyam Rayaprolu
Enterprise: Syneos Health
Work Phone: +1 972.370.3552
E-mail: subrahmanyam.rayaprolu@syneoshealth.com