



Machine Learning for better treatment with OMICS and Clinical Data

Phani S Srinivasan Ponnappalli, Syneos Health
Subrahmanyam Rayaprolu, Syneos Health

Agenda

- Introduction
- 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
- Conclusion

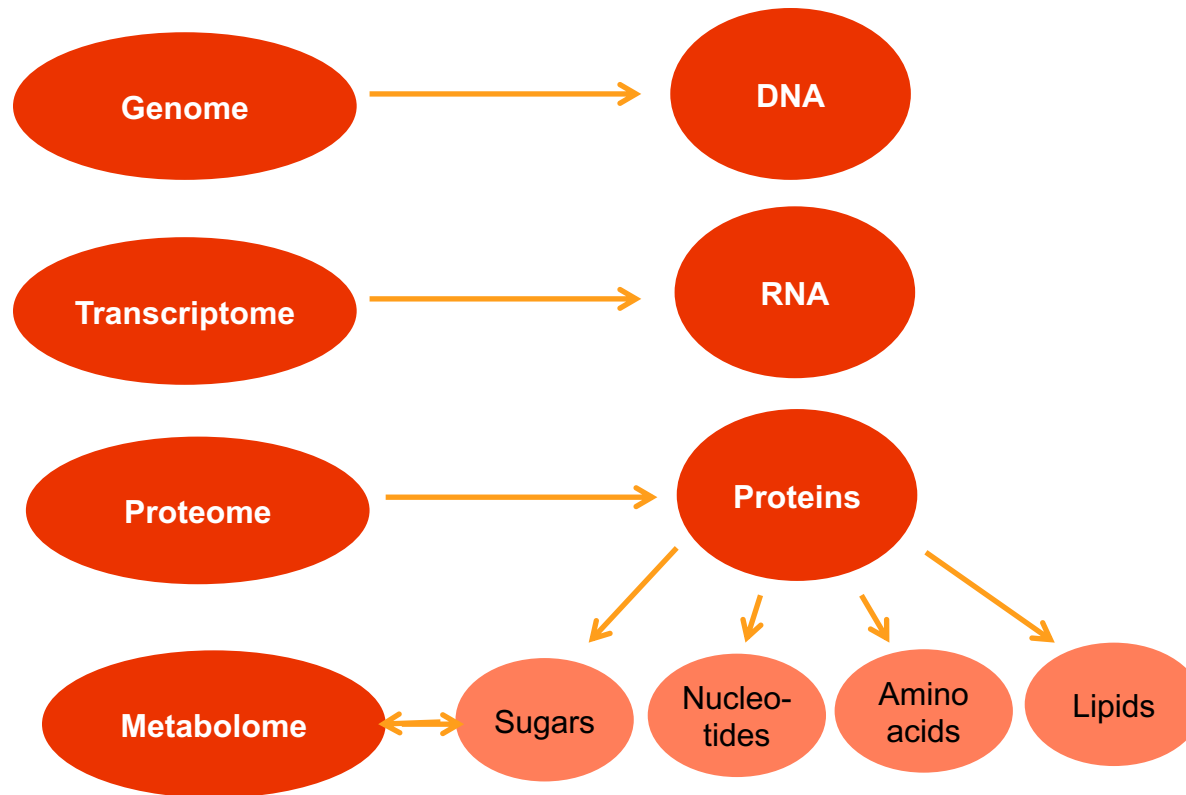
Introduction

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

OMICS Data



OMICS Data

Genomics:

Genomics is an interdisciplinary field of science focusing on the structure, function, evolution, mapping, and editing of genomes.

Transcriptomics:

The study of all coding and non-coding RNA molecules in a cell and their functions is collectively called as transcriptomics.

Proteomics:

Proteomics is the study of proteins and proteome is the entire set of proteins in a given cell

Metabolomics:

The study of the set of metabolites present within an organism, cell, or tissue is called Metabolomics.

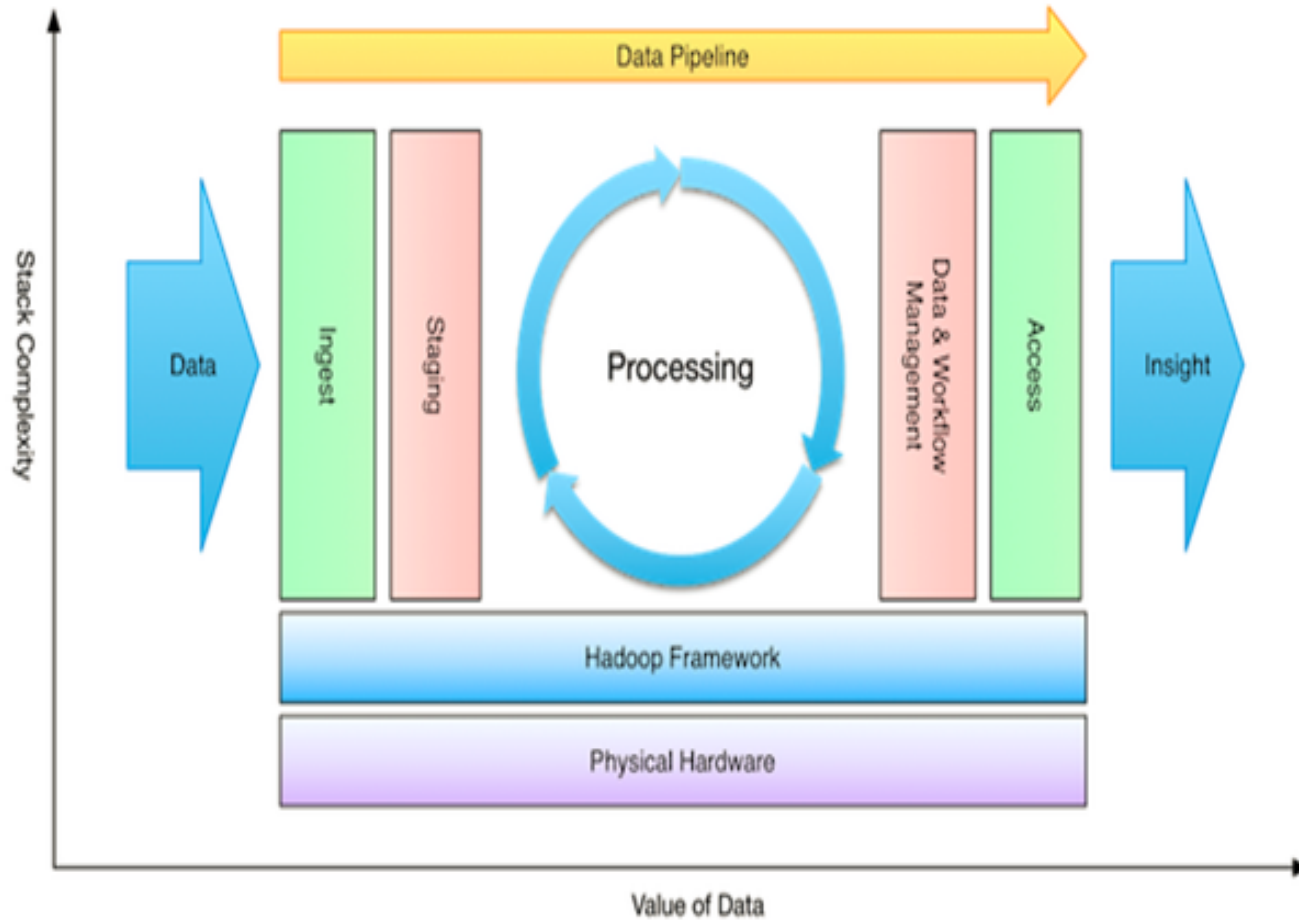
Big Data

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:**
- **Fault Tolerance**
- **Scalability**
- **Cost effectiveness**

Big Data Architecture



No SQL

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

- Key-value data stores
- Document stores:
- Wide-column stores:
- Graph stores:

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>
```

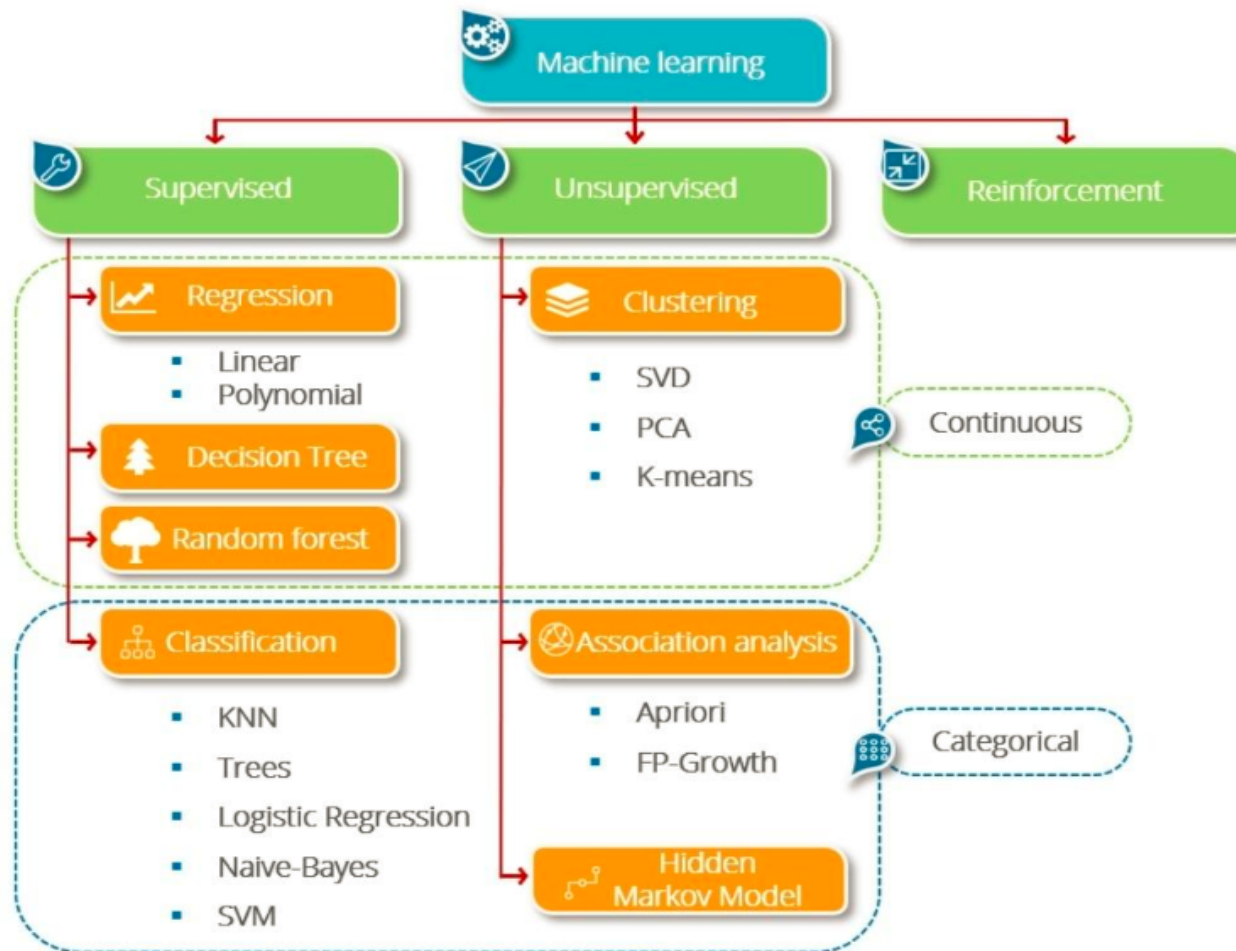
Machine Learning

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.

Machine Learning algorithm requires three main functions.

1. **Hypothesis function:** Hypothesis function is basically model for the data.
2. **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.
3. **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



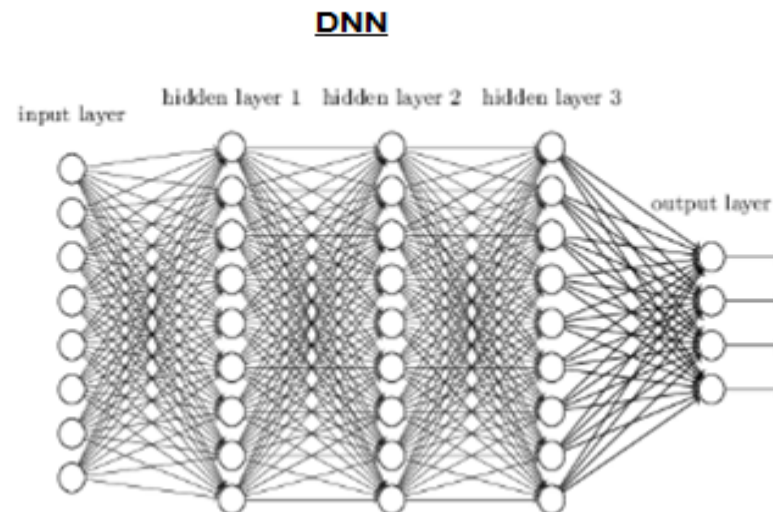
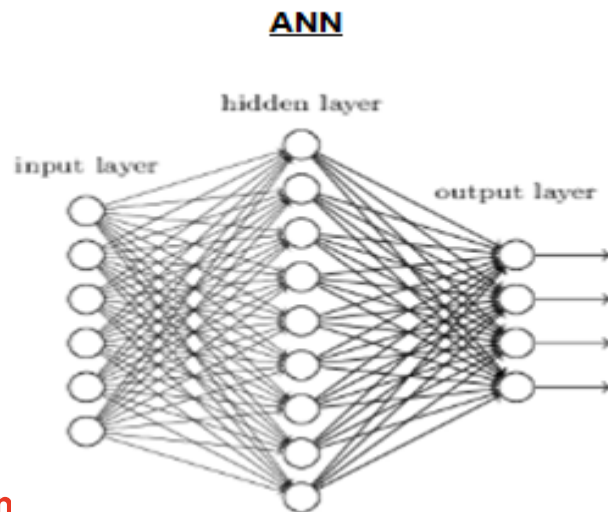
Artificial Neural Network (ANN) Vs Deep Neural Network (DNN)

Artificial Neural Network (ANN)

It is computational algorithm 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.

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.



Integrating Omics data 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

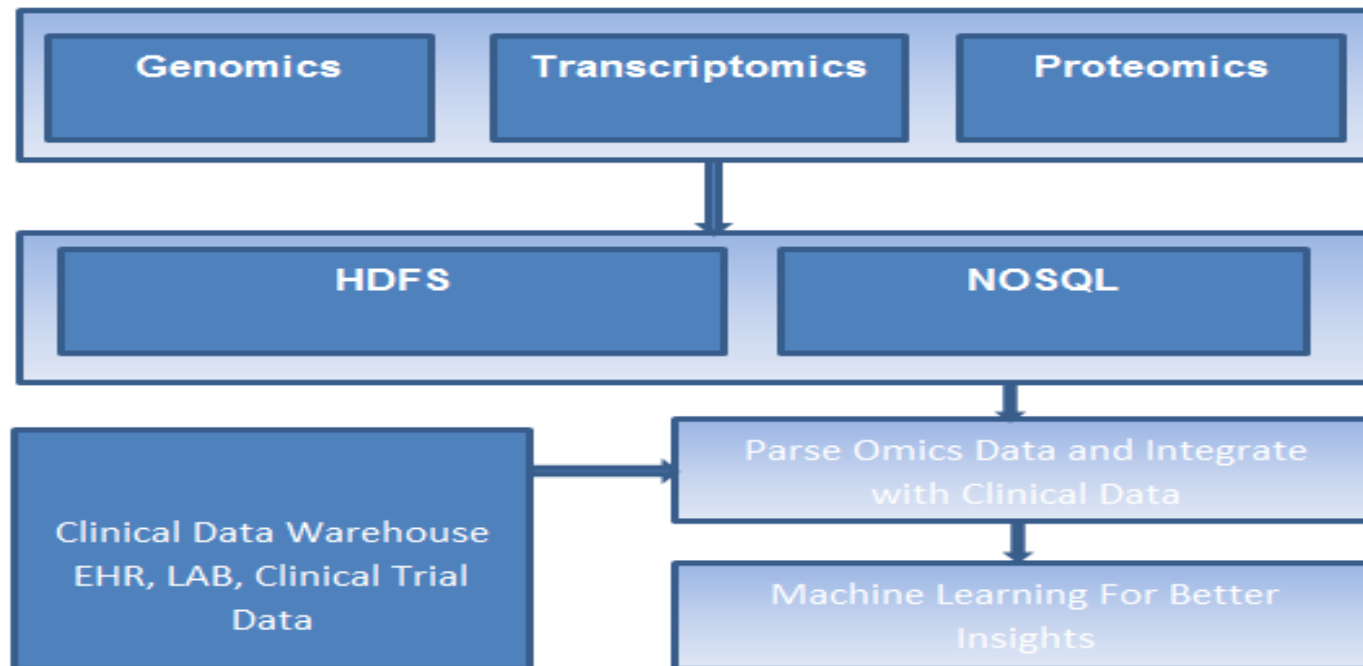
➤ Images/Videos

➤ JSON

➤ *Fastq and FASTA Format*

- 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 for integrating Omics and Clinical Data



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
```

- **OMICS data and 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 be parsed through libraries.

```
TCGCTGCGTCATCCGCCAGCAGGAGCTGGACTTTACTGATGCCCGTTATATCTGCCAAAAAGACCGGGATC
TGGACCCCGTGATGGCATTCTCTGGTTTTCGTCATCCGGTGAAGAGATTGAGCCACCTGACAGTGTGACCT
TTCACATCTGGACAGCGTACAGCCCGTTCACCACCTGGGTGCAGATTGTCAAAGACTGGATGAAAAACGAA
AGGGGATACGGGAAAAACGTAAAAACCTTCGTAACACACCACGCTCGGTGAGACGTGGGAGGCGAAAAATTGGC
GAACGTCGGGATGCTGAAGTGATGGCAGAGCGGAAAGAGCATTATTCAGCGCCCGTTCTTGACCGTGTGG
CTTACCTGACCGCCCGGTATCGACTCCAGCTGGACCGCTACGAAATGCGCGTATGGGGATGGGGGCGCGGG
TGAGGAAAGCTGGCTGATTGACCGCGCAGATTATTATGGGGCCGCCAGCAGATGAACAGACGCTGCTGCGT
GTGGATGAGGCGATCAATAAACCTATACCGCGCGGAATGGTGCAGAAATGTCCGATATCCCGTATCTGCT
GGGATACTGGCGGGATTGACCCGACCATTTGTGTATGAACGCTCGAAAAAACATGGGCTGTTCCGGGTGAT
CCCCATTAAAGGGGATCCGCTCTACGGAAAGCGGTGGCCAGCATGCCACGTAAGCGAAACAAAAACGGG
GTTTACCTTACCGAAATCGGTACGGATACCGCGAAAGAGCAGATTATAACCGCTTCACACTGACGCGCG
AAGGGGATGAACCGCTTCCCGGTGCCGTTCACTTCCCGAATAACCGGATATTTTGTATCTGACCGAAGC
GCAGCAGCTGACTGCTGAAGAGCAGGTGCAAAAAATGGGTGGATGCCAGGAAAAAAATACTGTGGGACAGC
AAAAAGCGACGCAATGAGGCACCTCGACTGCTTCGTTTATGCGCTGGCGGCGCTGCGCATCAGTATTTCCC
"omics_sample.fa" 695 lines, 49270 characters
```

Parsing Genome Data

- 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:

```
Counter({'G': 12820, 'A': 13564, 'T': 11867, 'C': 124574})
```

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.

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

Sample Input to ML Algorithm:

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

General packages and libraries for ML 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 with Python

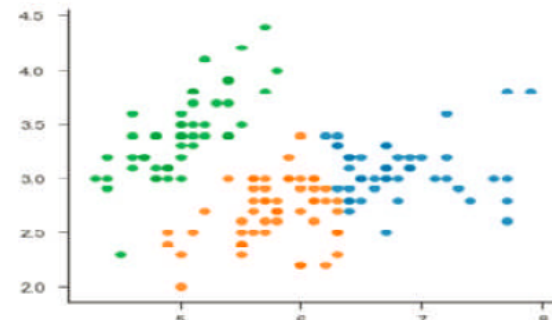
K-means is type of unsupervised algorithm which deals with the clustering problems. Its procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). Data points inside a cluster are homogeneous and are heterogeneous to peer groups.

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

- The vast volume of digital data generated by health devices and sensors will be transformative across the entire health care spectrum, from wellness and prevention to treatment and research.
- Machine learning offers tremendous opportunities to more efficiently access and understand vast amounts of data produced by bio-informatics Industry.
- During the next five years, Big Data and AI technologies will mature to enable advanced research models, including cloud-based health databases of continuously uploaded patient data and Internet-based trials conducted remotely.
- 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

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

Thank you!

