



NLP for Enhanced Treatment Decision

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

Agenda

- Introduction
- Diverse data types for clinical decision support
- Big Data and NoSQL
- Data Science Work Flow.
- Insight of NLP Algorithms.
- Python Programming Language for NLP techniques.
- Sample Python code for applying machine NLP on Integrated Clinical Data.
- Conclusion

Introduction

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

Applied Machine Learning and NLP techniques on integrated clinical information will transform the way we look at disease and enhance treatment decisions.

- 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

Different forms of Clinical Data

Genomics:

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

Electronic Health Record :

An electronic health record (EHR) is the systematized collection of patient health information and stored electronically in a digital format.

Physician Notes:

A physician note is an entry into a medical or health record made by a Physician, nurse or lab technician or any other member of a patient's healthcare team.

Clinical Trial Data:

Clinical trial data is the data generated from clinical trial research on a drug, biologic or device about their safety and efficacy.

Lab Data:

Laboratory data consists of lab test results from blood chemistry, hematology and Urinalysis. Lab data may also consists of many different collections of tests such as Microbiologic laboratory tests, ECG laboratory data etc.

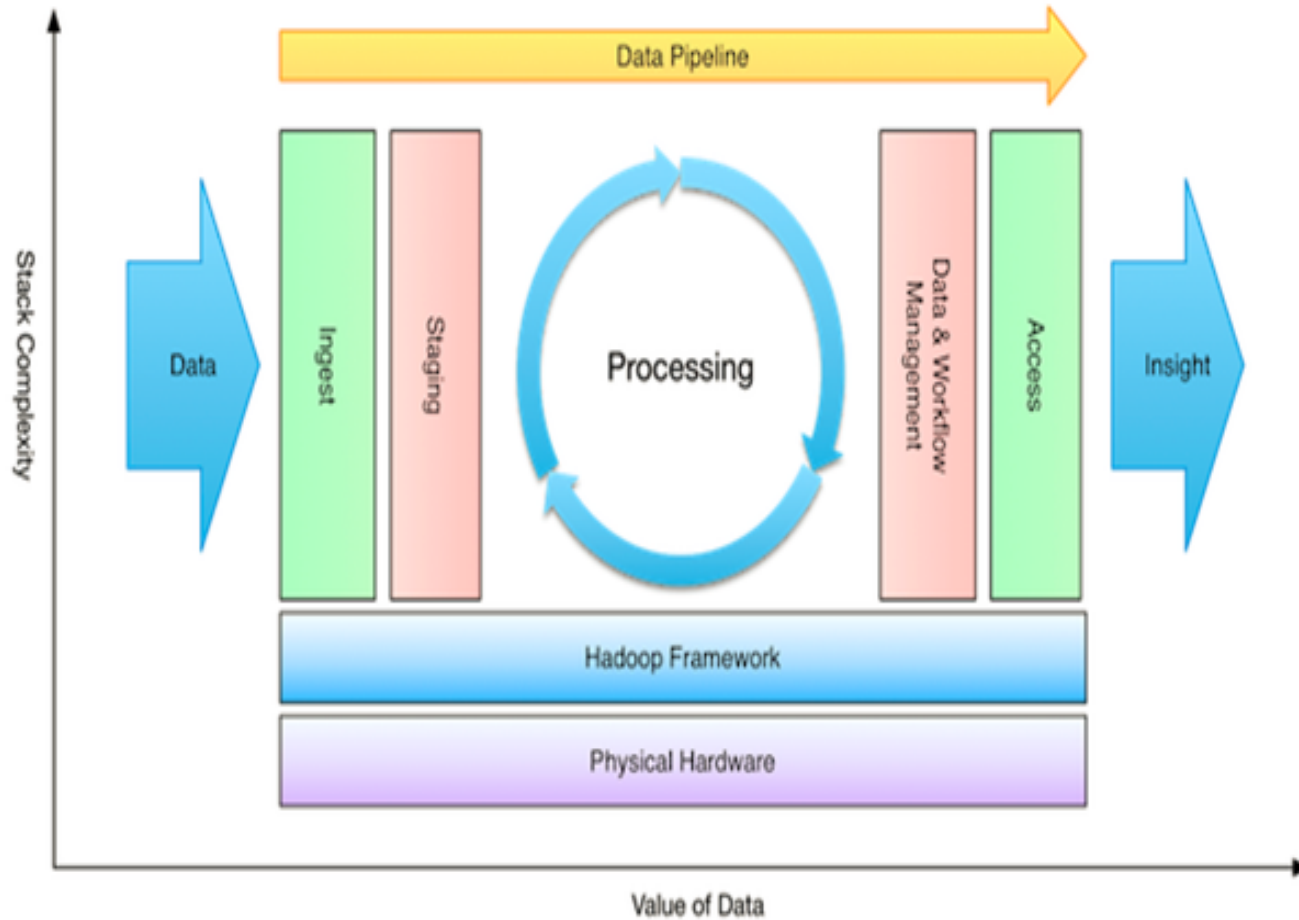
Big Data

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>
```

NATURAL LANGUAGE PROCESSING (NLP)

NLP is a way for computers to analyze, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

Popular NLP algorithms

1. **Sentiment analysis:** Sentiment analysis is a machine learning technique that automatically analyze the data and detects the Positive or Negative polarity within a text whether it is a whole document, paragraph, sentence or clause.
2. **Topic segmentation:** Topic segmentation is the process of dividing written text into meaning blocks such as words, sentences or topics. Unstructured clinical data could be segmented into meaningful blocks such as demographics, medical history, laboratory test results, hospitalizations and social history for further analysis.
3. **Text Generation :**Text generation is the task of generating text with the goal of appearing indistinguishable or missing text. A trained language model predicts the likelihood of occurrence of a word based on the previous sequence of words used in the text. Language models can be operated at character level, sentence level or even paragraph level.

INTEGRATION OF ALL FORMS OF CLINICAL DATA

Latest technological advancements in Pharma has resulted in massive amounts of life-sciences 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.

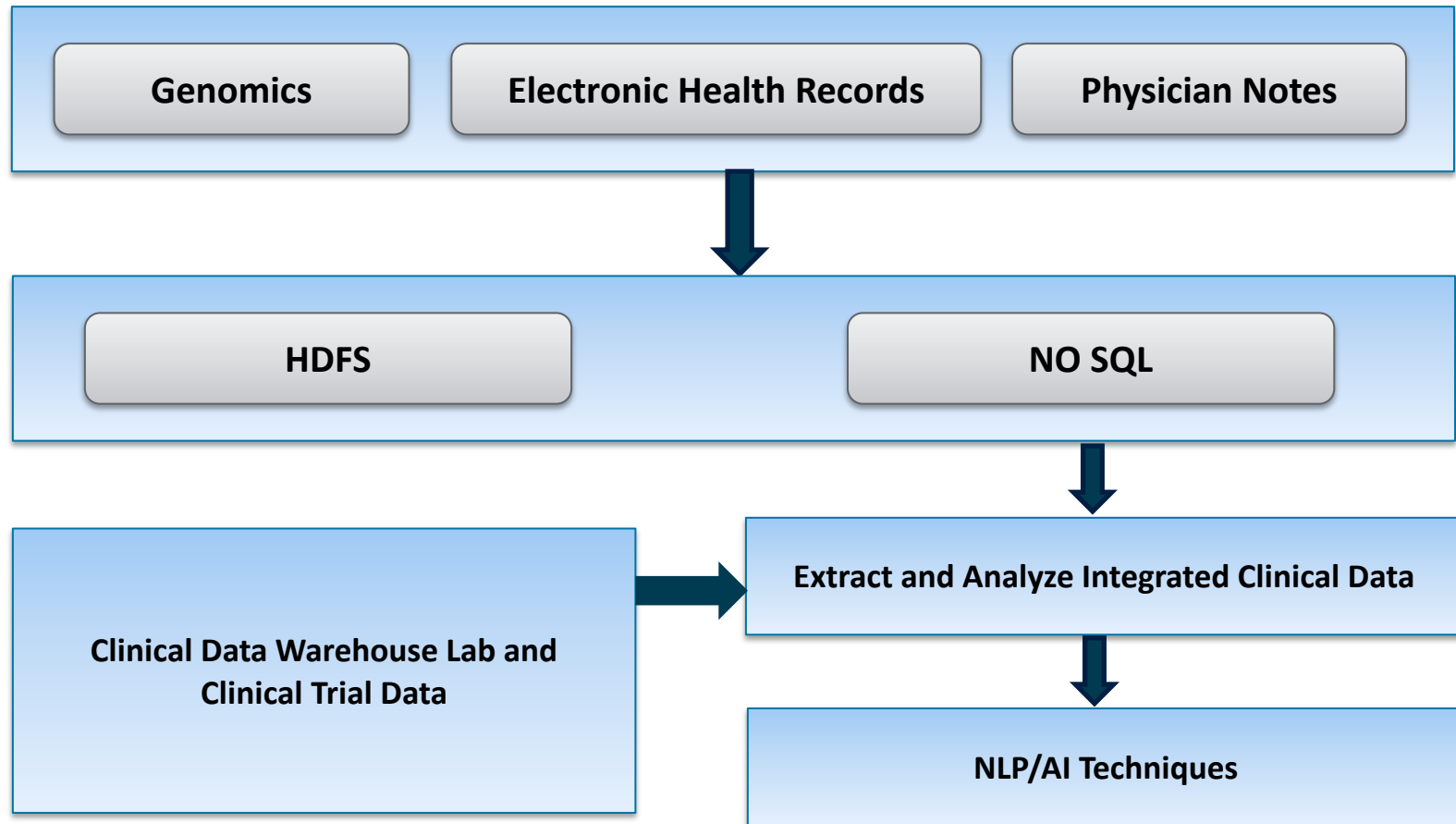
Lifesciences data is available for processing in various formats such as

- Hospital Notes
- *Pathology Reports*
- *EMR data*
- Clinical data base

Big Data and NOSQL framework is the most suitable architecture for loading this unstructured Data.

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

Architecture for integrating diverse Clinical Data



DATA SCIENCE WORK FLOW WITH PYTHON LIBRARIES

- Although data science projects can range widely in terms of their aims, scale, and technologies used, at a certain level of abstraction most of them implemented as the following workflow



Important Terminology

Document Matrix

	Terms			
Document	adverse event	visit	admit	result
Note1	20	6	45	67
Note 2	4	4	5	50
Note 3	6	8	6	66
Note 4	0	7	0	88

Corpus

Patient ID	Physician Notes
CAB101	Patient admitted with adverse reaction to Drug and Oxygen support is Given
CAB102	Patient was called on phone to find out adherence to Drug Schedule

STEP 1: IDENTIFY THE QUESTIONS TO BE ANSWERED

Analytics is about asking—and answering—smarter questions.

What questions should you be asking to drive more value in your organization? What information needed to answer these questions? Which tools can efficiently deliver the answers? Clinical Data is complex and diverse, Pharma Organizations have to identify and prioritize the questions answered.

For example:

- In a typical clinical trial, each patient responds differently to the same drug. Pharma companies challenge is to find a pattern associated with individual characteristics.
- Patient enrolment for a rare disease.
- Plan for monitoring trials in real time to identify signals requiring action to avoid significant and potentially costly issues.

STEP 2: DATA EXTRACTION AND CLEANING

Big-Data frame works like apache Hadoop and Ingestions tools are used to acquire the data and load in to enterprise Data Lake. Big Data Ecosystems contain tons of data in raw format.

Data extraction and cleaning processing step is to turn the raw source data into a system analyzable format. Different types of analysis require type of data formats.

- Make text all lower case
- Remove punctuation, numerical values and stop words.
- Tokenize text

Python Programing Paradigm:

Input: Raw Data

Python Libraries:

- Pandas – is a data analysis library that includes dataframes
- Pickle - The pickle module implements binary protocols for serializing and de-serializing a Python object structure.
- NLTK (Natural Language Toolkit) is used for such tasks as tokenization, lemmatization, stemming, parsing, POS tagging, etc. This library has tools for almost all NLP tasks.

STEP 3: EXPLORATORY ANALYSIS

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations

Python Programing Paradigm:

Input: Document Matrix, Pickled file

Python Libraries:

- Numpy - is used for its N-dimensional array objects
- Matplotlib – is 2D plotting library for creating graphs and plots

Output:

- Distribution plots
- Simple Graphs and Charts

STEP 4: APPLY AI ALGORITHMS

At this point data is integrated and converted to format required by particular NLP/AI techniques.

Common Tasks solved by NLP algorithm include:

- Text classification
- Topic modeling
- Text generation

Python Programing Paradigm:

Input: Corpus, Document Matrix, Pickled file

Python Libraries:

- Gensim - is the package for topic and vector space modeling, document similarity
- TextBlob- is a package for part-of-speech tagging, noun phrase extraction, sentiment analysis, classification

STEP 5: VISUALIZATION

Final step in data science workflow include interpretation and presentation of the results in form of graphs, reports. It provides easy-to-understand format and helps in gaining insights of hidden patterns in data.

Python Programing Paradigm:

Input: Corpus, Document Matrix, Pickled file

Python Libraries:

- seaborn – a data visualization library based on matplotlib
- scikit-learn - the algorithms used for data analysis and data mining tasks
- seaborn – a data visualization library based on matplotlib

Output:

- Graphs and Charts
- Visualizations.

PYTHON NLP SAMPLE CODE

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
```

Patient Response trend:

Sentiment Analysis and Segmentation on clinical trial site notes and Investigator notes, which are sometimes available in comments and not loaded to clinical database could be used to gain more insight and decision making during a trail.

- Pattern of Patients response positive or Negative to a drug across the visit?
- Any additional reasons of patient's response like an event in family or individual Genome?

Sentiment Analysis with Python

Sentiment analysis is a machine learning technique that automatically analyze the data and detects the Positive or Negative polarity within a text whether it is a whole document, paragraph, sentence or clause. Much of the information of clinical data i.e. Genomics, Health care records and Physician notes lies buried in unstructured text and it is very hard to analyze this data. Sentiment analysis helps in making sense of the unstructured clinical data by automatically tagging and detecting the polarity.

- Each word in a corpus is labeled in terms of polarity and subjectivity. A corpus sentiment is the average of these.
- Polarity: How positive or negative a word is. -1 is very negative. +1 is very positive.
- Subjectivity: How subjective or opinionated a word is. 0 is fact. +1 is very much an opinion.

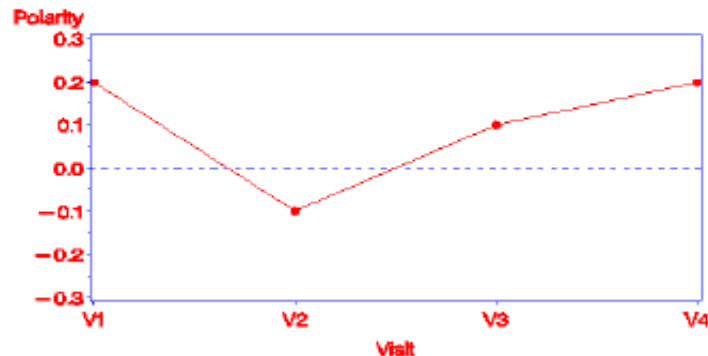
Code Snippet

```
# Calculate the average polarity of each patient
polarity_patient = []
for lg in notes_visit:
    polarity_visit = []
    for p in lg:
        polarity_visit.append(TextBlob(p).sentiment.polarity)
    polarity_patient.append(polarity_visit)

polarity_patient

# Show the plot for one patient
plt.plot(polarity_patient[0])
plt.title(data['Patient_name'].index[0])
plt.show()
```

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.
- The combination of NLP and ML provides a powerful paradigm for developing predictive models in healthcare and life science applications.
- 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.
- NLP can provide access to most of the data otherwise trapped in unstructured text, turning it at large scale into features that can drive predictive models

References

- <https://github.com/adashofdata/nlp-in-python-tutorial/>
- <http://alttox.org/mapp/emerging-technologies/omics-bioinformatics-computational-biology/>
- <https://www.sciencedirect.com/science/article/pii/S2001037014000464>
- <http://blog.cloudera.com/blog/2014/09/getting-started-with-big-data-architecture/>
- <https://monkeylearn.com/sentiment-analysis/>
- https://en.wikipedia.org/wiki/Electronic_health_record
- <https://practicefusion.com/medical-notes/>
- <https://www.analyticsvidhya.com/blog/2013/07/big-data/>
- <https://emerj.com/ai-sector-overviews/natural-language-processing-in-pharma-current-applications/>
- <https://www.sciencedirect.com/science/article/pii/S1532046418302016>

Thank you!

