# Chapter 1

## INTRODUCTION

According to the World Health Organisation (WHO), better health is central to human happiness and well-being. It also makes an important contribution to economic progress, as healthy populations live longer, are more productive, and save more.

The medical practice and human health are so intertwined that one cannot make mention of one without the other one being involved. Medicine is the application of scientific knowledge to human health and doctors are one important agent through which that scientific understanding is expressed. (Pubmed, 2005)

One has to know where he is coming from in order to know where he is headed. The understanding of a patient’s medical history helps to assure that the medical doctor and the individual’s health care providers provide the most appropriate and effective treatment and support for the individual’s illnesses and health conditions so that they maintain the best possible health.

A person’s medical history is made up of many different pieces of information that tell the complete story about that individual’s current and past health. A patient’s medical history is very vital as it helps the physician to know what medications a person takes, allergies, and diagnoses to treat the person in an optimal way.

### BACKGROUND OF STUDY

The impact of poor documentation of medical history or wrong reading of a patient’s medical history cannot be overemphasized. If such happens, it can lead to incorrect treatment decisions, expensive, painful, and/or unnecessary diagnostic studies, and unclear communication between consultants and referring physicians, resulting in a lack of follow through with evaluation and treatment plans. (Schaeffer, 2016)

Misdiagnoses are also the result of the fact that most doctors only spend roughly 15 minutes with each patient, which leaves little time for in-depth discussions about the patient's condition and makes giving a tentative diagnosis more difficult. (Leydon, 2017). This therefore means that inadequate time is another reason for misdiagnoses.

If there is a way to reduce the time spent reviewing case notes and also increasing the accuracy of medical diagnosis from patient’s medical history, then that indeed would be beneficial to both the doctor and the patient. One way in which this can be done is in the introduction of Natural Language Processing (NLP) to extract information from a repository containing data from case notes and then using a technique called data visualization to present the information graphically. It is true that a picture contains a thousand words.

Natural language processing is the term used to describe the process of using of computer algorithms to identify key elements in everyday language and extract meaning from unstructured spoken or written input.  NLP is a discipline of computer science that requires skills in artificial intelligence, computational linguistics, and other machine learning disciplines.

In the healthcare industry, natural language processing has many potential applications.  NLP can enhance the completeness and accuracy of electronic health records by translating free text into standardized data.  It can fill data warehouses and semantic data lakes with meaningful information accessed by free-text query interfaces.  It may be able to [make documentation easier](https://ehrintelligence.com/news/speech-recognition-supports-clinical-documentation-improvement) by allowing health care providers to dictate their notes, or generate tailored educational materials for patients ready for discharge.

But perhaps of greatest interest right now, especially to healthcare providers in desperate need of solutions for incredibly complex patient problems, is the use of NLP for clinical decision support. An NLP-powered context engine can mine unstructured documents at high speed and deliver actionable suggestions around clinical decision support.

The particular allure of NLP for healthcare is that these systems can “learn” over time and continuously deliver better insights as more interactions occur. Data processing can be also improved over time by adjusting the statistical probabilities, which account for the likelihood that the delivered bit of data matches the user request. Based on the feedback, these probabilities can be tweaked to improve the results even further (Bresnick, 2016).

Understanding a patient’s medical history is important because it helps in times of emergency to know which medications the individual is taking as well as their health conditions and any allergies that they have in order to provide safe and effective treatment. Also, it helps to know treatments and medications other doctors have provided. By knowing this information, doctors can avoid duplicating services and also avoid dangerous interactions between medications.

The most famous example of a machine learning NLP whiz-kid in the healthcare industry is IBM Watson, which has gained massive media attention due to its voracious appetite for academic literature and its growing expertise in clinical decision support (CDS) for precision medicine and cancer care (Bresnick, 2016).

A group of [French researchers](http://www.aclweb.org/anthology/W09-4506) has developed another NLP-powered algorithm aimed at monitoring, detecting, and preventing hospital-acquired infections (HAI) among patients. Performing hospital-wide surveillance on this is a rather complex and time-consuming task. However, computer science enables parsing all the patient records, while NLP can help render unstructured data such as speech accounts that are used to detect early signs of a possible infection and notify the clinicians. (Proux, Pierre, Segond, Kergourlay, Darmoni, Pereira, Gicquel, and Metzger, 2009)

Another successful experiment was [conducted to automate the identification and predict risk for hospitalized heart failure patients](https://www.ncbi.nlm.nih.gov/pubmed/26911827). Scientists analysed free text reports from the previous 24-hours each day with the help of NLP to determine at-risk hospitalized patients and predict the patient’s risk of 30-day hospital readmission and 30-day mortality.  The algorithm performed really well and increased the identification score’s sensitivity from 82.6% to 95.3% and its specificity from 82.7% to 97.5%, and the model’s positive predictive value is 97.45%. (Evans, Benuzillo, Horne, Lloyd, Bradshaw, Budge, Rasmusson, Roberts, Buckway, Geer, Garrett and Lappé, 2016)

In 2013, the Department of Veterans Affairs used NLP techniques to review more than 2 billion EHR documents for indications of PTSD, depression, and potential self-harm in veteran patients.  The pilot was 80 percent accurate at identifying the difference between records of screenings for suicide and mentions of actual past suicide attempts.

Researchers at MIT in 2012 were able to attain a [75 percent accuracy rate](https://ehrintelligence.com/news/mit-develops-new-nlp-system-for-freehand-ehr-medical-notes) for deciphering the semantic meaning of specific clinical terms contained in free-text clinical notes, using a statistical probability model to assess surrounding terms and put ambiguous terms into context (Bresnick, 2012).

All the above mentioned systems, in one way or the other, have recorded success when it comes to parsing clinical notes and using the data to produce meaningful insights to existing problems. However, there is still some missing link. The addition to this system being built which is going to differentiate it from other systems would be the visualization of the data. After parsing the records and extracting relevant information, data visualization tools would be used to provide infographics for better understanding of the data.

### THEORECTICAL BACKGROUND

This NLP-based visualization system has its roots in the computer science fields as it is related to data science, artificial intelligence and machine learning which are new and emerging fields in computer science.

**Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract [knowledge](https://en.wikipedia.org/wiki/Knowledge) and insights from [data](https://en.wikipedia.org/wiki/Data) in both structured and unstructured forms. It unifies statistics, data analysis, machine learning and their related methods in order to "understand and analyse actual phenomena" with data.

**Data mining** is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. The overall goal of which is to extract information from a data set, and transform it into an understandable structure for further use through data analysis. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

**Artificial intelligence (AI)** is an area of computer science that emphasizes the creation of intelligent machines that work and react like humans. Some of the activities computers with artificial intelligence are designed for include:

* Problem solving
* Speech recognition
* Learning
* Planning

**Machine learning** is an application of artificial intelligence (AI) that provides systems the ability to automatically learn without human intervention, 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.

**Natural Language Processing (NLP)** is the process of using of computer algorithms to identify key elements in everyday language and extract meaning from unstructured spoken or written input.  NLP is a discipline of computer science that requires skills in artificial intelligence, computational linguistics, and other machine learning disciplines.

Some NLP efforts are focused on mimicking human-like responses to queries or conversations.  Others try to understand human speech through voice recognition technology, such as the automated customer service applications used by many large companies.

Many NLP systems “learn” over time, reabsorbing the results of previous interactions as feedback about which results were accurate and which did not meet expectations.  These machine learning programs can operate based on statistical probabilities, which weigh the likelihood that a given piece of data is actually what the user has requested.  Based on whether or not that answer meets approval, the probabilities can be adjusted in the future to meet the evolving needs of the end-user (Bresnick, 2016).

**Data Visualization** is a general term that describes any effort to help people understand the significance of **data** by placing it in a visual context (pictorial or graphical format). Patterns, trends and correlations that might go undetected in text-based **data** can be exposed and recognized easier with **data visualization** software (Rouse, 2012).

**Clinical decision support system (CDSS)** is an application that analyses data to help healthcare providers make clinical decisions. Physicians, nurses and other healthcare professionals use a CDSS to prepare a diagnosis and to review the diagnosis as a means of improving the final result. [Data mining](https://searchsqlserver.techtarget.com/definition/data-mining) may be conducted to examine the patient's medical history in conjunction with relevant clinical research. Such analysis can help predict potential events, which can range from drug interactions to disease symptoms (Rouse, 2014).

### THE PROBLEM STATEMENT

The amount of information the doctor has to go through before starting the consultation process can sometimes be overwhelming and thereby leading to wrong inferences. Also, with the lack of access to coherent data, it may be hard for the doctor to extract actionable insights from all the unstructured data available within the given time frame, especially in case of emergency. As a result, the doctor has a fragmented view of the patient’s medical history.

There is a need to turn the vast amount of case notes data into more intelligent, intuitive and streamlined reports that is beneficial for both the doctor and the patient. The introduction of Natural Language Processing to medical records alongside the visualization of the medical history is the proposed solution.

### AIM AND OBJECTIVES

The aim of this project is to design and implement an NLP-based visualization model that would improve the medical consultation process.

The objectives are:

1. To capture information relating to the documentation of case notes.
2. To design a repository that aggregates case notes’ data.
3. To design an NLP-based system that mines the data and then provides a complete, easy-to-review summary of patient’s information, available to the doctor.
4. To test the system for the visualization of a patient’s medical history.

### METHODOLOGY RESEARCH DESIGN

Building a fully functional NLP-based visualization system would involve about four major steps.

1. Data / information collection: This stage involves capturing the data (patients’ case notes) that would be used for the project. The major stakeholders (doctors) would be interviewed to properly understand what kind of information exists in a case note. Sample case notes would be collated for use as input to the system.
2. Designing a classifier which would group the different aspects of a case note into their different entities or part of speech. This would form the knowledge-base or the knowledge repository for this project.
3. Designing and implementing an NLP-based system which would use text summarization techniques and algorithms to extract the main content from the knowledge repository.
4. Data Visualization: The summarized text from the last stage would then serve as input to the visualization tool. The output from this stage would be an infographic showing a detailed but summarized version of the patient’s case note.

Text mining techniques would be used to process and clean the unstructured data, making it a standard input for the machine learning model. The model then trains from this input data and is able to classify the text and extracts keywords, presenting a summary of the entire note.

**Requirements Elicitation**

The information gathering techniques that would be used include the following:

* A feasibility study would be carried out to determine the economic and organisational impact of the system.
* Interviews with necessary stakeholders such as doctors and data entry operators to understand the workings of the existing system.

**Requirements Analysis**

In this task, both the functional (what the system is expected to do) and the non-functional requirements (how the system is expected to behave) would be stated.

**Requirements Modelling**

For process modelling, an activity diagram would be used to show how one activity connects with the next within a system.

Use Case model would be used to show the interaction between the different actors in the system and how they operate in the system.

For data modelling, a class diagram would be used to properly define the relationship between the different data objects in the system.

For Logic modelling, structured English would be used as a tool for representing the logical processes that takes place within the system.

**System Design**

Respective tasks and tools to be used in specifying the system design are stated below:

* The general system architecture modelling would be designed using Subsystem Hierarchy Relationship Architecture.
* The System Logical Design showing the transformation of inputs to outputs would be clearly stated using algorithms (and flowchart or pseudocode if needed).

**Implementation**

The proposed system would be implemented using Python programming language.

### JUSTIFICATION OF STUDY

Whether manually or digitally, there is some form of medical documentation in every hospital. Each patient has a ‘history’ which shows the diagnosis, known allergies, current medications, past and present illnesses, medication history, current doctors, emergency contact information, previous surgeries, family medical history, immunization records, insurance information and many other things. This is actually a whole lot of information. A new doctor or medical personnel has to go through all these before diagnosing a patient as it is essential for standards of care to be met and to facilitate continuum of care (Kanaan, 2017).

Sometimes, as a result of an emergency or need to provide first aid, the process of actually going through one’s case note is hastened, which can lead to misdiagnoses (Leydon, 2017). Other times, it can be as a result of poor documentation which could result in incorrect treatment decisions or expensive, painful, and/or unnecessary diagnostic studies (Schaeffer, 2016).

There is a need to implement a solution which addresses this problem, which reduces human error on the part of the doctors. The digital age comes with many advantages, one of which is the availability of large amounts of datasets. The data can then be mined to extract meaningful information and to discover patterns.

One of such data mining techniques is Natural Language Processing which uses computer algorithms to identify key elements in everyday language and extract meaning from unstructured spoken or written input. Also, there is a concept known as data visualization which help people understand the significance of data by placing it in a visual context (pictorial or graphical format).

Applied to the process of medical consultation, these techniques can be used to design a system that extracts meaning from patients’ case notes or medical history and then visualizes the information extracted so that the doctor sees at a glance, the history of that particular patient. Hence, the need for an NLP-based visualization system which improves the process of medical consultation.

**REFERENCES**

Bresnick J. (2012), MIT develops new NLP system for freehand EHR medical notes. *HealthITAnalytics, https://ehrintelligence.com/news/mit-develops-new-nlp-system-for-freehand-ehr-medical-notes*

Bresnick J. (2016),What is the Role of Natural Language Processing in Healthcare? *HealthITAnalytics*

Evans R.S, [Benuzillo J](https://www.ncbi.nlm.nih.gov/pubmed/?term=Benuzillo%20J%5BAuthor%5D&cauthor=true&cauthor_uid=26911827). , [Horne B.D](https://www.ncbi.nlm.nih.gov/pubmed/?term=Horne%20BD%5BAuthor%5D&cauthor=true&cauthor_uid=26911827). , [Lloyd J.F](https://www.ncbi.nlm.nih.gov/pubmed/?term=Lloyd%20JF%5BAuthor%5D&cauthor=true&cauthor_uid=26911827). , [Bradshaw A](https://www.ncbi.nlm.nih.gov/pubmed/?term=Bradshaw%20A%5BAuthor%5D&cauthor=true&cauthor_uid=26911827). , [Budge D](https://www.ncbi.nlm.nih.gov/pubmed/?term=Budge%20D%5BAuthor%5D&cauthor=true&cauthor_uid=26911827). , [Rasmusson K.D](https://www.ncbi.nlm.nih.gov/pubmed/?term=Rasmusson%20KD%5BAuthor%5D&cauthor=true&cauthor_uid=26911827). , [Roberts C](https://www.ncbi.nlm.nih.gov/pubmed/?term=Roberts%20C%5BAuthor%5D&cauthor=true&cauthor_uid=26911827). , [Buckway J](https://www.ncbi.nlm.nih.gov/pubmed/?term=Buckway%20J%5BAuthor%5D&cauthor=true&cauthor_uid=26911827). , [Geer N](https://www.ncbi.nlm.nih.gov/pubmed/?term=Geer%20N%5BAuthor%5D&cauthor=true&cauthor_uid=26911827). , [Garrett T](https://www.ncbi.nlm.nih.gov/pubmed/?term=Garrett%20T%5BAuthor%5D&cauthor=true&cauthor_uid=26911827). and [Lappé D.L](https://www.ncbi.nlm.nih.gov/pubmed/?term=Lapp%C3%A9%20DL%5BAuthor%5D&cauthor=true&cauthor_uid=26911827). (2016), Automated Identification and Predictive Tools to help identify high-risk heart failure patients: pilot evaluation*, Pubmed, https://www.ncbi.nlm.nih.gov/pubmed/26911827*

Halyna (2017), Natural Language Processing in Healthcare: Is It Worth Implementing? *Romexsoft*

**Juliann S. (2016) Poor Documentation: Why It Happens and How to Fix It*,*** For The Record**, Vol. 28 No. 5 P. 12**

Kanaan S. (2017), The Importance of Documentation: The Medical and Legal Issues

[MarksMan Healthcare](https://marksmanhealthcare.com/author/amit/) (2017), How Data Visualization is Helping in Healthcare Decision Making? *MarksMan Healthcare club*

Melanie M. (2017), The Rise of Healthcare Data Visualization*, A journal of Ahima blog*

Noy F. N and McGuinness L. D. (2001), Ontology Development 101: A Guide to Creating Your First Ontology, *Stanford University*, pp: 1. *https://protege.stanford.edu/publications/ontology\_development/ontology101-noy-mcguinness.html*

Proux D. , Pierre M. , Segond F. , Kergourlay I. , Darmoni S., Pereira S. , Gicquel Q. , and Metzger M. (2009), Natural Language Processing to detect Risk Patterns related to Hospital Acquired Infections. *http://www.aclweb.org/anthology/W09-4506*

Rouse M. (2012), Data Visualization. *Search Business Analytics,* [*https://searchbusinessanalytics.techtarget.com/definition/data-visualization*](https://searchbusinessanalytics.techtarget.com/definition/data-visualization)

Rouse M. (2014), clinical decision support system (CDSS). *Search HealthIT, https://searchhealthit.techtarget.com/definition/clinical-decision-support-system-CDSS*

Wocl L. (2017), Some Reasons Patients Are Misdiagnosed, [*Failure to Diagnose*](https://www.woclleydon.com/blog/failure-to-diagnose/)

[Working Party of the Royal College of Physicians](https://www.ncbi.nlm.nih.gov/pubmed/?term=Working%20Party%20of%20the%20Royal%20College%20of%20Physicians%5BCorporate%20Author%5D) (2005), Doctors in society. Medical professionalism in a changing world*,* *Pubmed, https://www.ncbi.nlm.nih.gov/pubmed/16408403*