Scalable Signal Data Processing for Measuring Functional Connectivity in Epilepsy Neurological Disorder

Arthur Gershon, PhD[1], Samden D. Lhatoo, MD[2], Curtis Tatsuoka, PhD[2], Kaushik Ghosh, PhD[2], Kenneth Loparo, PhD[4], Satya S. Sahoo, PhD[1,2,4]

[1] Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, OH, USA.
[2] Department of Neurology, School of Medicine, Case Western Reserve University, Cleveland, OH, USA.
[3] Department of Mathematical Sciences, College of Sciences, University of Nevada Las Vegas, Las Vegas, NV, USA.
[4] Department of Electrical Engineering and Computer Science, Case School of Engineering, Case Western Reserve University, Cleveland, OH, USA.

Abstract

The accurate characterization of how different brain structures interact in terms of both structural and functional networks is an area of active research in neuroscience. A better understanding of these interactions can potentially lead to targeted treatments and improved therapies for many neurological disorders, such as epilepsy, which alone affects over 65 million people worldwide. The study of functional connectivity networks in epilepsy, which is characterized by abnormalities in brain electrical activity, will help to provide new insights into the onset and progression of this complex neurological disorder. In this chapter, we discuss statistical signal processing techniques and their use in determining functional connectivity among brain regions exhibiting epileptic activity. We also discuss computational challenges associated with deriving functional connectivity measures from neurological Big Data, and we introduce our highly scalable signal processing pipeline for quantifying functional connectivity with the goal of addressing these challenges and potentially advancing understanding of the underlying mechanisms of epilepsy. This pipeline makes use of a novel signal data format that facilitates storing and retrieving data in a distributed computing environment. We conclude the chapter by describing our current activities and proposed plans for improving our computational

pipeline, such as the inclusion of biomedical ontologies for semantic annotation in order to facilitate the integration and retrieval of signal data.

1 Introduction

1.1: Functional Networks in the Brain

The human brain is one of the most important organs in the body; it is responsible for many critical functions, including cognition, memory, language, and execution [1]. The brain has been the subject of a large body of research; nevertheless, we still have a rather limited understanding of various aspects of the brain and its constituents, as well as the interactions between brain structures, especially in the context of neurological disorders. Previous research efforts, which were focused solely on understanding the complexities of the brain's structure and functions, have identified numerous substructures in various regions of the brain that are responsible for a wide variety of different cognitive and physiological functions [2, 3].

In addition to understanding the functions of each substructure as an individual unit, it is important to determine how distinct regions in the brain interact and work together. Physical connections between distinct brain regions form *structural networks*, whereas associations that are formed over time and developed during processes such as speech, memorization, or other physiological events are referred to as *functional networks* [4]. *Functional connectivity measures* are computed by evaluating statistical correlations of physiological signals recorded from different brain regions [2]. Analyses of these latter connections have proven to be quite approachable using signal processing methods.

There are several important applications of functional connectivity analysis; for example, it can be used toward the goal of better understanding the mechanisms that underpin various

capabilities of the human brain. The measurement of functional connectivity in the brain can also be used to study neurological disorders, such as epilepsy. Epilepsy is one of the most common serious neurological disorders, affecting more than 65 million individuals worldwide in various forms [5].

Epileptic disorders are characterized by seizures resulting from the generation and propagation of abnormal electrical activity in the brain [6, 7]. These characteristic phenomena of epilepsy are commonly recorded and observed through the use of electroencephalograms (EEGs), and functional connectivity analysis can be applied to EEG data to aid in the understanding of the development and progression of epilepsy. There has been extensive research focused on the use of signal processing and data mining methods on intracranial recordings to determine case-specific collections of intracranial sites involved in seizure onset and that constitute the *epileptogenic zone* [2, 8]. Similarly, other research has focused on computing functional connectivity measures between different brain regions that constitute an epileptogenic network [8].

Several different categories of correlation measures have been used to evaluate functional connectivity [9, 10]. One example is the *linear* correlation measures that include cross-correlation, Pearson's correlation coefficient, and coherence. These measures quantify the *linear dependence* the time series, assuming that information propagates directly from one site to another without interference from ambient noise. By contrast, functional connectivity measures such as the Average Amount of Mutual Information (AAMI) index and other dynamical system analysis approaches do take into account interference from other signals, and thus do not make any direct assumption on how signal information propagates as a function of time. Measures in this category are often referred to as *non-linear* measures of functional connectivity [2, 10]. Non-

linear functional connectivity measures have been shown to provide more accurate results vis-à-vis linear approaches [11]. It is important to note, however, that existing non-linear correlation techniques are based on the temporal characteristics of the signal (e.g. the signal amplitude) instead of frequency domain features, such as those derived from the Fourier transform [8], whereas coherence, a linear correlation measure, uses signal frequency values for analysis [8, 10].

In addition to the development of new theoretical models to correlate signal data, the massive volume, wide variety, and rapid rate of signal data generation requires the development of highly scalable computational tools and platforms [1, 12, 13]. The computational challenges associated with this "Big Data" in neuroscience requires the development of algorithms and data structures that can leverage high performance distributed computing resources (e.g., cloud platforms) to store, analyze, and visualize large-scale datasets [1]. In particular, neuroscience Big Data requires the development of new multi-modal data representation formats that can effectively address the limitations of existing file formats and leverage distributed computing infrastructure for scalability and efficient analysis.

It is important to address the limitations of present data formats, such as the European Data Format (EDF) that is widely used for storing physiological signal data [14, 15]. Files using EDF consist of two major components: (1) a header containing patient data and recording metadata, including the names of each channel, recording times, and units of measurement, represented as ASCII strings; and (2) the data record with a list of all signal values, stored in a binary format, corresponding to each time sample of the data. EDF files store signal data as a collection of data recordings organized by the time of recording and although this significantly reduces the overall size of the file, it makes it can be difficult to extract and analyze recordings

from individual channels, e.g. in the computation of functional connectivity measures.

There has been a significant amount of work in the development of neuroscience data storage formats to address the limitations of EDF and other existing file formats. In the next section, we describe various approaches used to manage neuroscience data and the computation of functional connectivity measures derived from the signal data.

1.2: Related Work

Although we focus on computing functional connectivity measures derived from signal data in this chapter, there are a variety of other approaches used to determine functional networks in the brain. For example, instead of using EEG electrical signal data, a substantial amount of neurological research uses blood oxygen level dependency (BOLD) signals, measured using functional magnetic resonance imaging (fMRI), to derive functional connectivity measures both in general neuroscience [2] and specifically in epilepsy research [16, 17].

There has also been extensive work in the development of file formats to address the challenges in storing and analyzing neurological data [18]. For example, the Neuroscience Electrophysiology Object (NEO) format is an object-oriented file format based on the Python programming language. The NEO format is proposed as a natural method for storing neurological data due to its object-oriented nature, which makes it suitable for use across computing platforms [19]. Similarly, the Hierarchical Data Format (HDF5) has been developed as a general scientific data storage format with implementations in a variety of programming environments. The HDF5 project has developed various optimization techniques for data storage and access [20], which has made it popular as a file format for storing neurological data.

1.3: Outline

The remainder of this chapter is structured as follows. In Section 2, we describe the

computation of functional connectivity measures in epilepsy using techniques of signal analysis, with a particular focus on various statistical models used to derive correlations among signal data recorded from different channels. In addition, we introduce a novel computational pipeline that uses a new data format, called the *Cloudwave Signal Format (CSF)*, to process and analyze signal data using a non-linear correlation technique. In Section 3, we give a broad overview of the results we have obtained from the use of this pipeline. Then, in Section 4, we describe the broader application of our techniques and tools used to compute functional connectivity in epilepsy patients, and discuss proposed enhancements. We conclude in Section 5 with a summary of our work.

2: Methods

In this section, we give an overview of the CSF format and its role in enabling the distributed storage of signal data, we describe techniques of signal data analysis used to derive functional connectivity measures, and finally we introduce the multi-step computational pipeline we have developed that implements the CSF format and the above data analytic techniques for measuring functional connectivity.

2.1: Recording of Brain Electrical Activity

The study of electrical activity in the brain has been of fundamental importance in neurology since Galvani's experiments on electrical activity in frogs and the subsequent development of an electronic theory of the nervous system [21]. We recall that the brain is comprised of special cells known as *neurons* that regulate various processes according to location, and a cognitive function is understood to result from the transmission of information between two neurons [22]. There are a number of ways in which signals can be relayed among

neurons, including the use of chemicals such as neurotransmitters for processes that require a lossless transmission of information [23].

Electrical signals comprise another category of methods for neuronal communication. They are used to rapidly convey information among different regions of the brain in the execution of reactive and motor skills, and in the synchronization of cognitive functions that form the basis of processes such as learning and perception [23]. The science of *electroencephalography* (EEG) encompasses a variety of methods of gathering data on intracranial electrical activity, including the use of scalp electrodes and magnets. *Niedermeyer's Electroencephalography* [6] provides a comprehensive overview of the subject.

EEG signal data are recorded using electrodes placed on the scalp according to the 10-10 system of placing electrodes at 10% intervals. The 10-10 placement scheme is a standard developed by the American Electroencephalographic Society and can be viewed as an amelioration of the International Standard 10-20 system that instead uses 20% spacing [24]. We note that there is ongoing research related to the optimal placement of electrodes to record brain electrical activity [25]. In contrast to scalp electrodes, depth electrodes are implanted in the brain (penetrating gray matter), and signals are recorded by one or more electrical contacts on each electrode. The specific number of contacts on the electrode depends on the position of the electrode and the depth of its implant [26, 27].

Depth electrodes are often implanted using a stereotactic approach, and the corresponding method of recording signal data is called *stereotactic electroencephalography*, or *SEEG* [28]. Although SEEG is an invasive recording technique, the quality of data is robust with brain electrical activity recorded at a high resolution. The analysis of SEEG data is therefore used as a gold standard in the diagnosis and treatment of epilepsy [29].

2.2: Managing Signal Data: The Cloudwave Signal Format

The storage and the management of signal data are significant challenges in brain connectivity research since effective analysis of data requires the use of both data and essential contextual metadata, for example instrument parameters, sampling rate, and study protocol. EDF is one of the most widely used signal storage formats in neuroscience applications [15]; however, it is not well suited for developing efficient data integration and analysis techniques. In addition, the EDF format does not support the FAIR principles that facilitate data sharing and reusability [30]. The FAIR principles allow efficient discovery of and access to scientific data using the following properties associated with datasets:

- **Findable:** Data should be easy to locate and easy to identify through the use of persistent identifiers and appropriate contextual metadata.
- **Accessible:** Data should be easy to access using existing network protocols and associated metadata information.
- **Interoperable:** Data should be annotated using a standard ontology term that allows easy sharing and analysis of data aggregated from different sources.
- **Reusable:** Data and any associated metadata should be made available with a clearly defined license to allow secondary use of datasets.

EDF files have limited to no support for these FAIR principles; for example, it is difficult to locate specific segments of signal data in an EDF file due to the format's minimal use of metadata information and lack of semantic annotations using standardized terminology. Similarly, the storage of signal data in EDF files as collection of temporally ordered recordings make it difficult to analyze channel-specific signals over a period of time. In particular, the retrieval of channel-specific signal data for a specific time period in an EDF file requires

multiple “look ups” for each timestamp, significantly increasing the number of computations required for time-series analysis.

To address these limitations of the EDF format, we have developed the Cloudwave Signal Format (CSF) that allows for the efficient storage, retrieval, and processing of signal data [31, 32]; Figure 1 illustrates the overall structure of a CSF file. The CSF format has been developed using the JavaScript Object Notation (JSON) framework that associates “values”, such as text data, numerical data, or other JSON objects, with textual strings known as “keys” [33]. EDF files can be transformed into CSF files without any loss of information or any other difficulty in the reusability of the signal data. On the contrary, the CSF format enables significant improvements over the EDF format in terms of signal data accessibility and interoperability. For example, we recall how the signal processing of data in EDF files requires several steps; each involving some computation of byte offset values in order to access the data. By contrast, CSF files can be easily processed with a single invocation of an appropriate value retrieval function in a programming language (e.g., a “getter” function in Java) with the associated key string as the function’s input.

In addition to greater accessibility, CSF files also support the interoperability of signal data generated from different sources through the use of ontology terms for data annotation. This feature allows for the reconciliation of data heterogeneity and improves the integration of data to allow researchers to query and analyze large repositories of signal data. In addition, the use of ontology terms for the annotation of signal data in CSF also enables the greater reusability of data and supports the creation of efficient indices for data segments. Although the use of descriptive “keys” and “values” in the CSF format leads to an increase in the storage size of the resulting files, we believe that the increasing availability of cheap data storage infrastructures

will address this challenge and allow CSF files to be used in practical data management systems.

```
{
    "Header": {
        "firstFragment": A,
        "lastFragment": B,
        "epochDuration": 30.0,
        "fragmentNoA": {
            "fragmentNo": A,
            "startDate": "MM.DD.YY",
            "startTime": "HH.MM.SS"
        },
        "fragmentNo(A+1)": {
            "fragmentNo": A+1,
            "startDate": "MM.DD.YY",
            "startTime": "HH.MM.SS" // 30 seconds later
        },
        ...,
        "fragmentNoB": {
            "fragmentNo": B,
            "startDate": "MM.DD.YY",
            "startTime": "HH.MM.SS" // 30*(B-A) seconds later
        }
    },
    "studyMetadata": {
        "edfFileName": "eegRecord.edf",
        "dataFormatVersion": 0,
        "localPatientID": "patientIDString",
        "recordingStartDate": "11.22.33",
        "dateFormat": "MM.DD.YY",
        "recordingStartTime": "12.34.56",
        "recordingStartTimeFormat": "HH.MM.SS",
        "numberHeaderBytes": "56064",
        "numberDataRecords": "14400",
        "dataRecordDuration": "0.1",
        "dataRecordDurationUnit": "seconds",
        "numberSignals": "N+1"
    },
    "clinicalAnnotationList": {
        "timestamp_1": "Annotation_1",
        "timestamp_2": "Annotation_2",
        ...
    },
...
...
"channelMetadata": {
    "channelName_0": {
    "channelNumber": "0",
    ... // other signal metadata listed in the EDF header
},
"channelName_1": {
    "channelNumber": "1",
    ... // other signal metadata listed in the EDF header
    },
    ...
    "channelName_N": {
        "channelNumber": "N",
        ... // other signal metadata listed in the EDF header
    },
    "channelList": "[channelName_0, channelName_1,
                    ..., channelName_N]"
},

"dataRecords": {
    "fragmentNumberA": {
        "channelName_0": "[ ... ]", // values are decimal arrays
        "channelName_1": "[ ... ]",
        ... ,
        "channelName_N": "[ ... ]"
    },
    "fragmentNumber(A+1)": {
        "channelName_0": "[ ... ]",
        "channelName_1": "[ ... ]",
        ... ,
        "channelName_N": "[ ... ]"
    },
    ... ,
    "fragmentNumberB": {
        "channelName_0": "[ ... ]",
        "channelName_1": "[ ... ]",
        ... ,
        "channelName_N": "[ ... ]"
    }
}
```

Figure 1: An example of the structure of the CSF format. Each value is associated to a plaintext key that can be used to easily retrieve data.

2.3: Processing Signal Data to Compute Functional Connectivity Measures

The CSF format supports the efficient computation of functional connectivity measures between different channels using various statistical techniques. The Pearson linear regression coefficient is a common statistical technique used to measure correlation between two datasets [34]. However, measures of *linear* correlation assume that electrical signals in the brain propagate as a *linear* function of time, which is not corroborated by clinical data [8]. To address

this limitation, *non-linear* regression techniques have been explored in the brain connectivity research community, such as the measurement of the Average Amount of Mutual Information (AAMI) shared by the signals. These regression techniques have led to the creation of non-linear functional connectivity measures that address certain limitations present in linear regression techniques [11].

A correlation metric developed by Pijn et al. [11] called the *non-linear correlation coefficient* has been found to be useful for computing functional connectivity in epilepsy patients [8]. This non-linear correlation measure views discretely-recorded signals as continuous functions of time, and uses the well-known mathematical fact that any continuous function can be approximated by a piece-wise linear function, where the error in the approximation is controlled by both the number and the locations of the endpoints of each linear piece [35]. The correlation coefficient of Pijn et al. uses linear regression on each (linear) piece of the approximation, and the average of the corresponding linear correlation coefficients is computed as an approximation to a "true" correlation.

The non-linear correlation coefficient generates accurate results with respect to correlation measures for signal data as demonstrated by Pijn et al. [11] and Wendling et al. [8]. In addition, the non-linear correlation coefficient is applicable in scenarios where the signals are linearly correlated as a function of time; in such cases, the value of the measure identically matches the value of Pearson's linear correlation coefficient. Moreover, the proposed non-linear correlation coefficient is *asymmetric*; that is, the value $h^2(X, Y)$ of the non-linear correlation coefficient comparing signal X to signal Y may differ from the value $h^2(Y, X)$ used to correlate signal Y to signal X. (The notation h^2 for, and indeed the name of, the non-linear correlation coefficient is meant to be analogous to the notation r^2 for, and the name of, Pearson's linear

correlation coefficient [8].) This asymmetric property of the correlation coefficient measure introduces a notion of *directionality* and allows us to evaluate if the activity at location *X* influences the activity at location *Y,* if:

$h^2(X, Y) \geq h^2(Y, X)$), (Eq. 1)

if conversely activity at *Y* influences activity at *X,* if:

$h^2(X, Y) \leq h^2(Y, X)$), (Eq. 2)

or if the activities at each site have some mutual influence on one another [11], if:

$h^2(X, Y) \approx h^2(Y, X)$), (Eq. 3)

We note, however, that the statistical measures described above address the issue of correlating signal amplitude values only, whereas clinicians often use signal frequency analysis to identify epilepsy related events [36]. Although linear functional connectivity measures of signal frequencies such as coherence are used in the research community, there are no non-linear functional connectivity measures available to correlate frequencies of signal data [8].

2.4: A Computational Pipeline for Analyzing Signal Data

In the previous section, we described various components of the signal analysis pipeline we have developed in order to analyze neurological signal data. We now describe our implementation and use of this signal processing pipeline to evaluate functional connectivity.

In the first phase, de-identified signal data recorded from an epilepsy patient and stored as a collection of EDF files are processed and transformed into CSF files. This process involves subdividing the entire duration of the signal data recording into smaller segments (typically 30 seconds in duration). For each segment, data is extracted from the EDF files by parsing the file and computing the byte location of each data element as described by the EDF specifications, and the extracted data is stored in an intermediate data structure. A predetermined number of

segments (provided as user input) are aggregated in a CSF file, which is generated using the Java JSON Application Programming Interface (API) [37].

The process of rewriting EDF files as CSF files involves a transformation of the layout of signal data from a collection of signals recorded during a given time period into a collection of time series data corresponding to each recording channel. In effect, this is a "transposition" of the time value-recording site matrix to a channel-oriented layout by measuring the byte offset to locate and extract the information. We present a schematic diagram of the transformation process in Figure 2.

The computational pipeline takes as input a list of user-defined parameters including start and end time stamps for a seizure event (or ictal period) under investigation, and a list of recording channels. In the next step, the tool iterates over pairs of signal channels listed in the input parameters, and extracts the data for each pair of signals over the given ictal period. The channel-oriented layout of the signal data in the CSF files facilitates the retrieval of relevant signal data during this step. In the next phase, the pipeline computes Pijn's non-linear correlation coefficient for each pair of signal channel using the extracted data. We note that this step is performed for all pairs of signal recording channels, which ensures that we perform all relevant computations in both directions of signal propagation (as discussed earlier in Section 2.3). The output of the computational pipeline is a two-dimensional matrix $\{h^2(X, Y)\}$ of non-linear correlation coefficient values for a given ictal period. In the next step, the matrix values are analyzed to qualify the correlation between different channels of signal data. We evaluate the relative strengths of each correlation during the ictal period by computing the average value, denoted by μ, and the standard deviation, denoted by σ, of all of the values in the matrix. These values are used to compute $N_\sigma(X, Y) = (h^2(X, Y) - \mu)/\sigma$ for each pair of signals, a value that

measures the number of standard deviations between a specific (i.e., local) correlation coefficient $h^2(X, Y)$ and the global average μ (a similar method is used by Wendling et al. [8]).

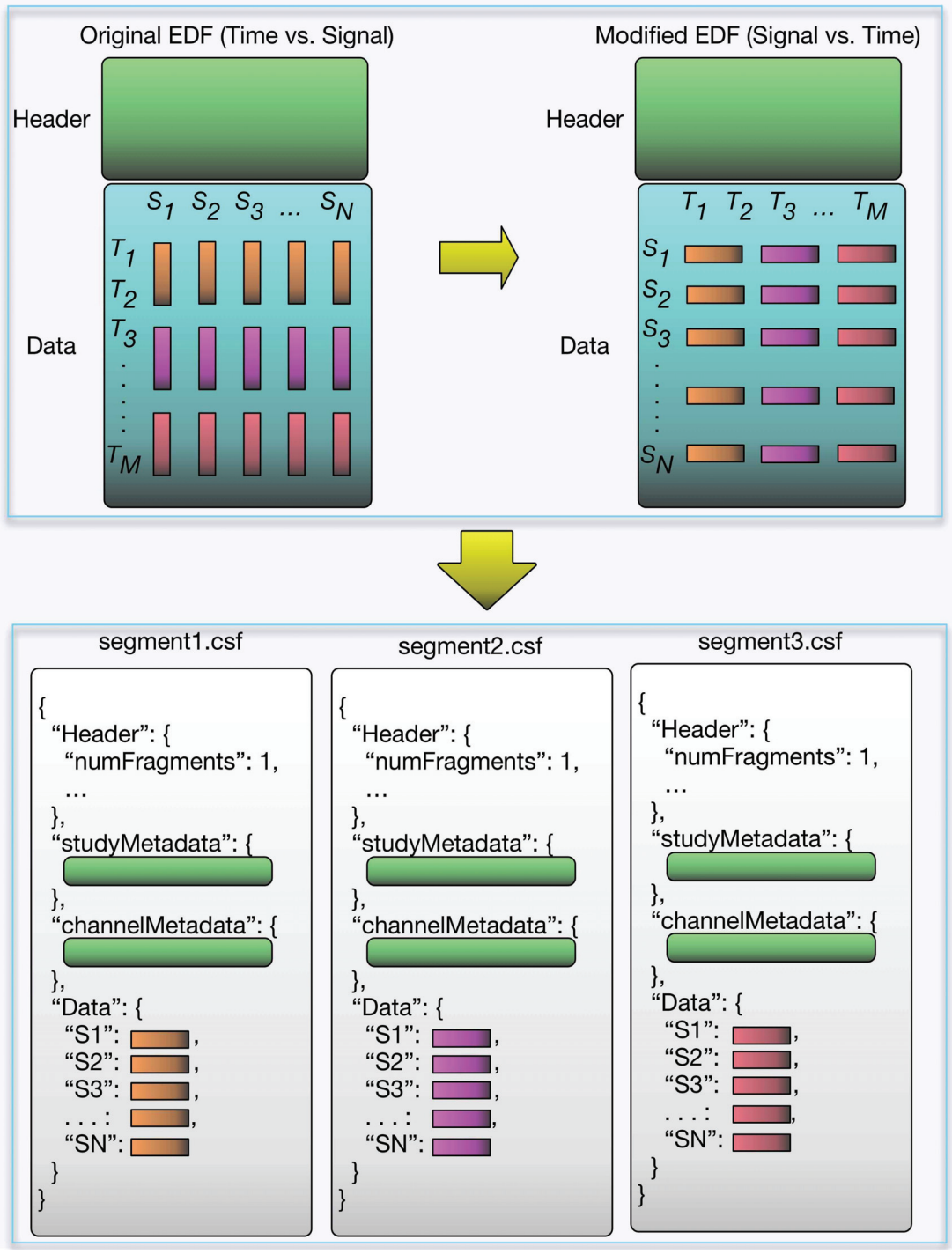

Figure 2: A schematic diagram of the signal processing pipeline. The first arrow represents the transposition of time/site data as written in EDF to site/time data stored in an intermediate object. The second arrow denotes the fragmentation of the data into multiple CSF files for use in distributed

computing environment.

The final output of the computational pipeline is a visualization of the data as a network graph with a set of vertices corresponding to the set of signal recording sites in the brain and edges corresponding to the matrix of N_σ values. It is common practice in statistics to characterize those values in a given set that differ from the average of all the values in that set by more than two standard deviations as *statistically significant*, as this behavior is observed for at most 5% of the values in a set with a Gaussian distribution [38]. While EEG signal data does not typically satisfy a Gaussian distribution, a similar proportion of signal values is observed to lie more than two standard deviations from the mean [8]. In accordance with these observations, we have therefore added directed edges in our output graph from vertex X to vertex Y for each pair of signals (X, Y) with $N_\sigma(X, Y) \geq 2$.

3: Results

In this section, we describe the preliminary results from analyzing de-identified signal data of an epilepsy patient using the computational pipeline described in Section 2. Figure 3 provides a schematic representation of the computational pipeline used to generate the matrix of h^2 values, the matrix of N_σ values, and the corresponding directed network graph. A preliminary review of the results shows that correlation of signals is not transitive. For example, given three electrode contacts A, B, and C, if activity at A is correlated with activity at B and activity at B is correlated with activity at C, then it is not necessarily true that A is correlated with activity at

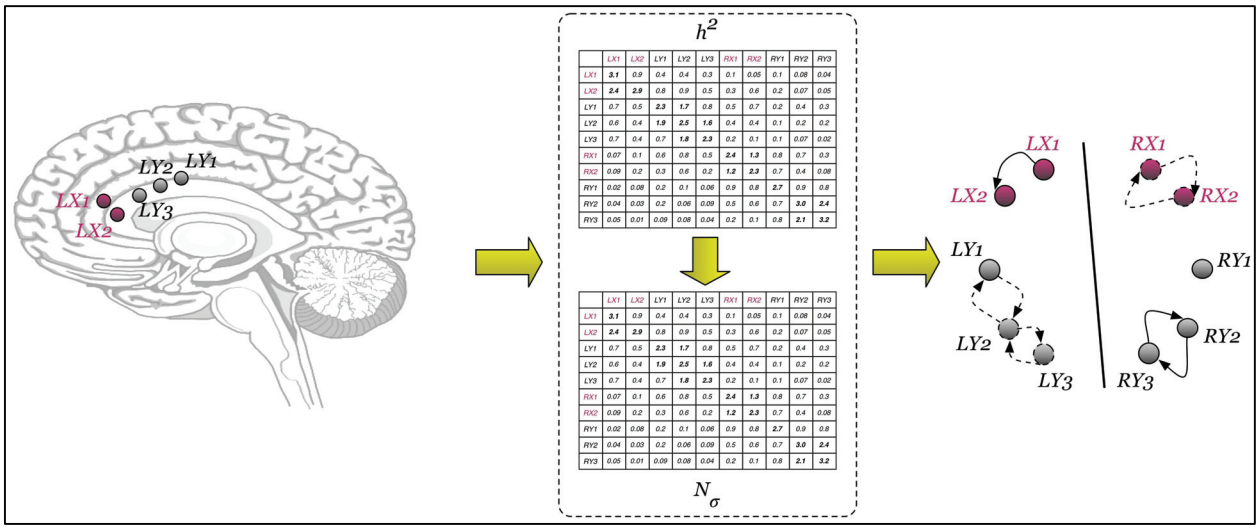

Figure 3: A conceptual overview of the data acquisition, processing, and analysis phases of the pipeline are illustrated. The diagram on the left displays an example placement of depth electrodes within the brain (brain image created by N. Byrd [39]). The middle shows the resulting matrix of h^2 correlation values and its conversion to a matrix of N_σ values. Finally, the right-most image gives an example of the network graph output created from the N_σ matrix. The notation X_i and Y_i (i=1, 2, 3) represent electrode contacts with LX_i and RX_i corresponding to placement of the electrode in the left and right hemisphere respectively. The directed edges connecting the electrode contact nodes represent correlation measures computed by the computational pipeline with solid lines and dashed lines used to differentiate between different correlation measure values.

C (in either direction). We propose to investigate the underlying cause for this result in collaboration with clinical researchers as part of our ongoing research.

Furthermore, we noted in our analysis that correlated sites of activity are more likely to be located on the same electrode, whereas correlated activity rarely occurs between contacts on different electrodes, especially across the two brain hemispheres. We found that our conclusions concur with those obtained in a clinical setting using evoked potentials [27]. We emphasize, however, that our results are computed using data from a single patient, and additional analysis is required to understand the underlying causes for these characteristics in the signal data.

4: Discussion

In this section, we discuss some of the limitations and proposed improvements to the signal data correlation coefficient measure proposed by Pijn et al. In addition, we discuss the use of parallel and distributed computing techniques for the goal of improving the performance of our computational pipeline. Finally, we describe the importance of using common terminological systems to facilitate interoperability of signal data with patient data stored in Electronic Health Record (EHR) systems.

4.1: Developing an Accurate Measure of Signal Correlation in Neurological Disorders

As we described in Section 3, the non-linear correlation coefficient developed by Pijn et al. is effective in corroborating some of the clinical findings related to neurological disorders such as epilepsy. In spite of this perceived effectiveness, we believe that there are several areas of improvement that will enable signal analysis to provide better insights into brain functional connectivity in both patients with neurological disorders and persons who do not have neurological disorders.

In Section 2, we noted that the correlation coefficient of Pijn et al. is based on a discrete approximation of the signal; it is therefore plausible that this measure represents a discrete approximation of a more accurate connectivity measure. Thus, an area of potential improvement lies in determining how to compute this “true” correlation measure using different techniques, such as through the use of some type of limiting process. Such a development could resolve some of the issues described in Section 2, such as an accurate determination of those pairs of signals for which correlation is statistically significant.

The results of our evaluation agree with previous findings by Wendling et al. [8] that the non-linear measure developed by Pijn et al. [11] effectively measures functional connectivity.

This suggests that intracranial signals propagate in a fashion that is *non-linear* with regard to time, which in turn implies the existence of some kind of signal interference that influences the transmission of electrical signals during epileptic events. However, the non-linear correlation coefficient cannot accurately determine the nature of this interference. Further investigation into this matter may require the incorporation of other techniques, such as dynamical system analysis.

Finally, we note that our current use of the non-linear correlation coefficient does allow us to determine the direction of influence among pairs of signals due to the inherent asymmetric properties of the measure. However, this correlation measure only computes *instantaneous* correlation; that is, we do not know how long it takes for a signal to reach some other site in the brain. The incorporation of additional features in the correlation coefficient measure for signal data, such as a method to compute any *time lag*, will significantly help advance our understanding of functional connectivity in epilepsy.

4.2: Improving the Performance of the Computational Pipeline

The statistical measures used to compute correlations within signal datasets require the pairwise processing of data recorded from different locations. With rapid technological advances in recording brain activities, the number of data points that are available to be processed for the computation of functional connectivity measures has increased dramatically in the past few years. For example, current SEEG recording techniques can record data at a rate of 10kHz from 256 electrode contacts. In addition, the volume of signal data is expected to keep increasing. Although the processing of such vast amounts of data is useful for advancing functional connectivity research, it presents significant computational challenges. The current implementation of our computational pipeline, for example, requires several hours to process data for an ictal event lasting only 30 seconds. As signal recording technology continues to

improve and the volume of data correspondingly increases, there is a clear need to develop efficient computational approaches to analyze signal data on a large scale.

The use of high performance parallel and distributed computing approaches, including the use of a cloud computing infrastructure, will allow us to improve the performance of the computational pipeline used to derive functional connectivity measures. In particular, the use of Apache Hadoop [40] or Apache Spark [41] will allow multiple ictal periods to be analyzed simultaneously. Apropos, the inherent ability of CSF to fragment and store signal data across multiple sites is ideally suited for use with a cloud computing infrastructure. We successfully developed a proof-of-concept implementation of our computational pipeline using Apache Pig that processed 750 gigabytes (GB) of EDF file data into CSF files using a 31-node Hadoop cluster [42]. Following this pilot implementation, we plan to develop an Apache Spark-based implementation of our computational pipeline to significantly improve the performance time for large-scale signal data processing.

4.3 The Use of Ontologies for Standardizing Terminology in Signal Data Analysis

Terminological heterogeneity in data generated from multiple sources arises due to the use of disparate terms to describe similar physiological events (e.g., signal complexes in EEG recordings), and it represents a key challenge in integrating large scale neuroscience data [1]. To address this critical challenge, we use terms modeled in existing biomedical ontologies to annotate signal data in CSF files as part of the computational pipeline. Ontologies are knowledge models that represent terms in a domain of discourse using formal knowledge representation languages, such as the description logic-based Web Ontology Language (OWL2) [43]. Biomedical ontologies have been widely adopted and used to reconcile data heterogeneity and support data integration and querying. For example, Gene Ontology (GO) is widely used to

annotate genomic data to facilitate the use of common terminology across different data sources and also enable users to easily query the integrated data [44].

The National Center for Biomedical Ontologies (NCBO) lists more than 500 open source biomedical ontologies that can be used for a semantic annotation of biomedical data and an automated reconciliation of heterogeneous terms used to describe similar data values [45]. At present, we are building on our experience in the development and application of a domain ontology for epilepsy called the Epilepsy and Seizure Ontology (EpSO) [46] to integrate additional neuroscience-specific ontologies in the computational pipeline for a semantic annotation of signal data. These semantic annotations are expected to significantly improve the integration and retrieval of signal data aggregated from multiple sources, including data generated in multi-center research studies. In addition, the use of terminological systems such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for semantic annotation is also expected to facilitate the interoperability of signal data with related clinical data stored in EHR systems. This will enable neuroscience researchers to perform clinical research studies.

5: Conclusion

The determination of dynamic properties of functional networks in the brain, in both healthy individuals and persons suffering from neurological disorders, is an important and challenging research problem. Visualizing the brain as an interactive and interconnected network of structures, we can create maps of functionally connected brain regions by observing the generation and propagation of electrical activity. In this chapter, we have outlined the use of statistical correlation techniques to compute functional connectivity measures from SEEG signal data. We have also described a computational pipeline that incorporates the new CSF signal data

representation format, along with other data processing and signal analysis functionalities. We expect that our pipeline will help to analyze signal data on a large scale, and thereby potentially advance our understanding of complex neurological disorders such as epilepsy. Our computational pipeline makes effective use of the novel CSF file format for signal data representation and storage. The CSF format has been designed to be effectively support time-series signal analysis and parallel processing techniques. We believe that integrating new functionalities and improving correlation measures for signal data will allow us to effectively leverage the growing volume of signal data for further research in neurological disorders.

Acknowledgement: This work is supported in part by the NIH-NIBIB Big Data to Knowledge (BD2K) 1U01EB020955 grant and NSF grant# 1636850.

References

[1] C. Bargmann, Newsome, W., Anderson, D., et al., "BRAIN 2025: a scientific vision. ," in "Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group Report to the Advisory Committee to the Director NIH," US National Institutes of Health. 2014.

[2] K. J. Friston, "Functional and effective connectivity: a review," *Brain Connect,* vol. 1, no. 1, pp. 13-36, 2011.

[3] S. Seung, *Connectome: How the Brain's Wiring Makes Us Who We Are*. Boston: Houghton Mifflin Harcourt, 2012.

[4] O. Sporns, "Structure and function of complex brain networks," *Dialogues in Clinical Neuroscience,* vol. 15, no. 3, pp. 247-262, 2013.

[5] P. O. Shafer. *About Epilepsy: The Basics*. Available: http://www.epilepsy.com/start-

here/about-epilepsy-basics (Accessed on January 15, 2015)

[6] D. L. Schomer and F. H. Lopes da Silva, "Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields." Philadelphia: Lippincott Williams & Wilkins, 2011.

[7] P. O. Shafer. *What Happens During A Seizure?* . Available: http://www.epilepsy.com/start-here/about-epilepsy-basics/what-happens-during-seizure (Accessed on January 15, 2015)

[8] F. Wendling, F. Bartolomei, and L. Senhadji, "Spatial analysis of intracerebral electroencephalographic signals in the time and frequency domain: identification of epileptogenic networks in partial epilepsy," *Philos Trans A Math Phys Eng Sci,* vol. 367, no. 1887, pp. 297-316, Jan 28 2009.

[9] O. David, D. Cosmelli, and K. J. Friston, "Evaluation of different measures of functional connectivity using a neural mass model," *Neuroimage,* vol. 21, no. 2, pp. 659-73, Feb 2004.

[10] K. Ansari-Asl, L. Senhadji, J.-J. Bellanger, and F. Wendling, "Quantitative evaluation of linear and nonlinear methods characterizing interdependencies between brain signals," *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics,* vol. 74, no. 3 Pt 1, pp. 31916-31916, 09/26 2006.

[11] J. P. Pijn and F. Lopes da Silva, "Propagation of Electrical Activity: Nonlinear Associations and Time Delays between EEG Signals," in *Basic Mechanisms of the EEG*, S. Zschocke and E.-J. Speckmann, Eds. Boston: Birkhauser, 1993.

[12] S. S. Sahoo, Jayapandian, C., Garg, G., Kaffashi, F., Chung, S., Bozorgi, A., Chen, C., Loparo, K., Lhatoo, S.D., Zhang, GQ, "Heartbeats in the Cloud: Distributed Analysis of

Electrophysiological “Big Data” using Cloud Computing for Epilepsy Clinical Research," *Journal of American Medical Informatics Association (Special Issue on Big Data),* vol. 21, no. 2, pp. 263-71, 2014.

[13] Editorial-Introduction, "Challenges and Opportunities," *Science,* vol. 331, no. 6018, pp. 692-692, 2011.

[14] B. Kemp, A. Värri, A. C. Rosa, K. D. Nielsen, and J. Gade, "A simple format for exchange of digitized polygraphic recordings," *Electroencephalography and Clinical Neurophysiology,* vol. 82, no. 5, pp. 391-393, 5// 1992.

[15] B. Kemp and J. Olivan, "European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data," *Clin Neurophysiol,* vol. 114, no. 9, pp. 1755-61, Sep 2003.

[16] C. Kesavadas and B. Thomas, "Clinical applications of functional MRI in epilepsy," *Indian J Radiol Imaging,* vol. 18, no. 3, pp. 210-7, Aug 2008.

[17] L. Maccotta *et al.*, "Impaired and facilitated functional networks in temporal lobe epilepsy," *Neuroimage Clin,* vol. 2, pp. 862-72, 2013.

[18] A. Schlögl, "An overview on data formats for biomedical signals," in *World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany: Vol. 25/4 Image Processing, Biosignal Processing, Modelling and Simulation, Biomechanics*, O. Dössel and W. C. Schlegel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 1557-1560.

[19] S. Garcia *et al.*, "Neo: an object model for handling electrophysiology data in multiple formats," *Front Neuroinform,* vol. 8, p. 10, 2014.

[20] M. T. Dougherty *et al.*, "Unifying Biological Image Formats with HDF5,"

Communications of the ACM, vol. 52, no. 10, pp. 42-47, 2009.

[21] E. Niedermeyer and D. L. Schomer, "Historical Aspects of EEG," in *Niedermeyer's Electrencephalography: Basic Principles, Clinical Applications, and Related Fields*, D. L. Schomer and F. H. Lopes da Silva, Eds. Philadelphia: Lippincott Williams & Wilkins, 2011.

[22] I. Singh, *Textbook of Human Neuroanatomy.*, Seventh ed. New Dehli: Jaypee Brothers Medical Publishers, 2006.

[23] S. G. Hormuzdi, M. A. Filippov, G. Mitropoulou, H. Monyer, and R. Bruzzone, "Electrical synapses: a dynamic signaling system that shapes the activity of neuronal networks," *Biochimica et Biophysica Acta (BBA) - Biomembranes,* vol. 1662, no. 1–2, pp. 113-137, 3/23/ 2004.

[24] J. Malmivuo and R. Plonsey, "Electroencephalography," in *Bioelectricomagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields*, J. Malmivuo and R. Plonsey, Eds. New York: Oxford University Press, 1995.

[25] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems," *Neuroimage,* vol. 34, no. 4, pp. 1600-11, Feb 15 2007.

[26] J. Gonzalez-Martinez *et al.*, "Stereotactic placement of depth electrodes in medically intractable epilepsy," *J Neurosurg,* vol. 120, no. 3, pp. 639-44, Mar 2014.

[27] N. Lacuey *et al.*, "Homotopic reciprocal functional connectivity between anterior human insulae," *Brain Struct Funct,* vol. 221, no. 5, pp. 2695-701, Jun 2016.

[28] H. Lüders, Engel, J. Jr., Munari, C., "General Principles.," in *Surgical Treatment of the Epilepsies*, J. J. Engel, Ed. New York, NY: Raven Press, 1993, pp. 137-153.

[29] D. Cosandier-Rimélé, J.-M. Badier, P. Chauvel, and F. Wendling, "Modeling and interpretation of scalp-EEG and depth-EEG signals during interictal activity," *Conference Proceedings,* vol. 1, pp. 4277-4280, 2007.

[30] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data,* Comment vol. 3, p. 160018, 03/15/online 2016.

[31] C. P. Jayapandian, Chen, C.H., Bozorgi, A., Lhatoo, S.D., Zhang, G.Q., Sahoo, S.S., "Cloudwave: Distributed Processing of "Big Data" from Electrophysiological Recordings for Epilepsy Clinical Research Using Hadoop.," in *American Medical Informatics Association (AMIA) Annual Symposium*, Washington DC, 2013, pp. 691-700: AMIA.

[32] C. Jayapandian, Wei, A., Ramesh, P., Zonjy, B., Lhatoo, S.D., Loparo, K., Zhang, GQ, Sahoo, S.S., "A Scalable Neuroinformatics Data Flow for Electrophysiological Signals using MapReduce," *Frontiers in Neuroinformatics,* vol. 9, no. 4, 2015.

[33] ECMA International, "The JSON Interchange Format," no. ECMA-404*,* Available: http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf (Accessed on January 15, 2015)

[34] S. W. Scheff, *Fundamental Statistical Principles for the Neurobiologist: A Survival Guide*. London: Academic Press, 2016.

[35] E. Süli and D. F. Mayers, *An Introduction to Numerical Analysis*. Cambridge, UK: Cambridge University Press, 2003.

[36] M. Zijlmans, P. Jiruska, R. Zelmann, F. S. Leijten, J. G. Jefferys, and J. Gotman, "High-frequency oscillations as a new biomarker in epilepsy," *Ann Neurol,* vol. 71, no. 2, pp. 169-78, Feb 2012.

[37] *Java API for JSON Processing (JSON-P)*. Available: http://json-processing-spec.java.net/

(Accessed on January 15, 2015)

[38] J. Rosenblatt, *Basic Statistical Methods and Models for the Sciences*. Boca Raton, FL: Chapman & Hall, 2002.

[39] N. Byrd, "A 2D vector drawing of the brain sliced down the center viewed from the side. Created using Adobe Illustrator CS3.," ed: Wikimedia Commons, 2014.

[40] The Apache Software Foundation. *Welcome to Apache Hadoop!* Available: http://hadoop.apache.org/ (Accessed on January 15, 2015)

[41] The Apache Software Foundation. *Apache Spark™ -- Lightning-Fast Cluster Computing*. Available: http://spark.apache.org/ (Accessed on January 15, 2015)

[42] S. S. Sahoo *et al.*, "NeuroPigPen: A Scalable Toolkit for Processing Electrophysiological Signal Data in Neuroscience Applications Using Apache Pig," *Front Neuroinform,* vol. 10, p. 18, 2016.

[43] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, "OWL 2 web ontology language primer," *W3C recommendation,* vol. 27, no. 1, p. 123, 2009.

[44] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nature genetics,* vol. 25, no. 1, pp. 25-29, 2000.

[45] M. A. Musen *et al.*, "The national center for biomedical ontology," *Journal of the American Medical Informatics Association,* vol. 19, no. 2, pp. 190-195, 2012.

[46] S. S. Sahoo, Lhatoo, S.D., Gupta, D.K., Cui, L., Zhao, M., Jayapandian, C., Bozorgi, A., Zhang, GQ., "Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care.," *Journal of American Medical Informatics Association,* vol. 21, no. 1, pp. 82-9, 2014.