1. General Information

1.0 Project Title: Bridging High Performance Computing and Machine Learning for Big Data Centric Neuroscience

1.1 Priority Areas and Sub-Area: Data Science Research for CPS. Sub area: Machine Learning at local level, Deep Learning, Distributed/parallel algorithms for learning.

1.2 Duration (in months): 36

1.3 Total Cost (in Rs Lakhs): 53.27

1.4 Foreign Exchange Component if any:

1.5 Principal Investigator: Dr. Dip Sankar Banerjee, Indian Institute of Information Technology, Guwahati.
Co-PI: Dr. Dipanjan Roy, Center of Behavioural and Cognitive Sciences University of Allahabad.

1.6 Designation: PI: Assistant Professor,
Co-PI: Assistant Professor.

1.7 Department and Institute Name:
PI: Department of Computer Science and Engineering, Indian Institute of Information Technology, Guwahati.
Co-PI: Center of Behavioural and Cognitive Sciences, University of Allahabad.

1.8 Address:
PI: Indian Institute of Information Technology, Guwahati, Ambari, GNB Road, Guwahati 781001. Assam.
Co-PI: University of Allahabad Senate Hall Campus 211002 UP.

1.9 Date of Birth:
PI: 24th August, 1986
Co-PI: 9th October, 1977

1.10 Gender:
PI and Co-PI: Male
Aadhar of PI: 252631957114
Aadhar of Co-PI: 548146887410

1.11 Telephone, Mobile, Fax and email:
PI: Mob: +91 96743 57187 email: dipsankarb@iiitg.ac.in
Co-PI: +91 8297518507 email: dipanjan@cbcs.ac.in

1.13 Project Summary:

Neuroscience, and in general, biomedical research is a grand challenge of the 21st century. The rapid development of highly capable neurotechnologies is creating massive volumes of data in a variety of modalities such as neuroimaging, large scale simulations, visualization, patient specific clinical platforms. This Neuroscience Big Data is rapidly becoming a significant challenge for researchers in terms of storage, data processing, and integrative analysis. One prominent direction towards current neuroscience research is towards the analysis, modelling and understanding of brain activities to come up with potential clinical biomarkers. Broadly data obtained from popularly used functional magnetic resonance imaging (fMRI) is used for performing functional segregation which correlates specific areas of the brain to specific behavior. On the other hand, structural scans of neuroimaging modalities provide valuable information about brains anatomical or *structural connectivity*. A recent trend amongst the researchers in the neuroscience community is towards performing structure-function correlations in the whole brain to understand the structural and functional connectivity differences in the health and disease. This *functional* approach is basically a spatio-temporal correlations between different brain regions and is an approach primarily based on *network science*. In parallel to the brain network approach, machine learning approach has evolved to become key to data science research. High dimensional fMRI data hinders the use of many classical statistical methods leading to increasing number of researchers relying on regularization methods that are commonly addressed through machine learning. Further, inference, at the level of single subjects are gaining paramount importance with several recent advancements such as the development of neurological markers. Predictive modelling of brain networks is thus particularly suitable through machine learning and thus finding widespread use cognitive, clinical, affective, and social neuroscience.

The traditional approach of building Neuroscience applications solves the Neuroscience problems in an ad-hoc manner which is hard to parallelize and scale to large scale environments. Recently, Neuroscience researchers have started to explore the benefits of parallelism offered by Big Data technologies such as Hadoop. Although these promise to achieve high productivity, they may not deliver the best performance and scalability on modern clusters that are enabled with multi-/many-core architectures, and Non-Volatile RAM (NVRAM) or Non-Volatile Memory Express (NVMe)-SSDs. These architectural trends are leading the design of Big Data middleware for Neuroscience applications and the applications themselves to a new era where high performance and scalability are equally important as high productivity.

On the other hand, as a part of the March 2015 executive order announced by Prime Minister Shri Narendra Modi for creating the National Supercomputing Mission [9], there has been an increased national focus on the use of High Performance Computing (HPC) to drive economic competitiveness and scientific discovery. The deployment of the various multi-petaflop HPC systems that are a part of this renewed national focus is being fueled by the recent advances in compute, networking, and storage technologies. On the compute front, multi-/many-core architectures (Intel Xeons, Intel Many Integrated Core (MIC), Xeon Phis, and NVIDIA Graphics Processing Units (GPGPUs) are the primary drivers.

The rapid development of highly capable neurotechnologies are creating **massive volumes** of data in a **variety of modalities** with a **high velocity** of generation rate. This Neuroscience "Big Data" is rapidly becoming a significant challenge for researchers in terms of storage, data processing, and integrative analysis. In several of the open access data sources, competent laboratories have made available nearly 30 Tera bytes of multimodal electrophysiological data in the past three years with the rate of data collection increasing each year. Traditional approaches in

such cases are hence becoming obsolete and newer technology stacks such as Apache Hadoop [2], and its derivatives such as Apache YARN [3], Pig [5], and Tez [7] are gaining importance. While these technologies are mostly focused towards distributed computation and load balance, it is also necessary to explore techniques for efficient representation of data for easy modification, retrieval, and storage. Since graphs often provide a very intuitive and flexible structure for representation and analysis of data, newer graph mining platforms such as Pregel [44], Apache Giraph [1], PEGASUS [41], and HADI [40] are gaining importance.

1.14 Objectives:

In this inter-disciplinary endeavor, we plan in a two dimensional plane where on one dimension we plan on investigating the application of machine learning and deep learning techniques for different challenges in modelling,and analyzing brain networks for individual classification. On the second dimension we plan on investigating acceleration of the modelling mechanisms through detailed systems level exploration, parallel computing, and exploitation of next generation of high performance computing processors and storage for better and faster results. This leads to the following fundamental challenges:

1. How can multimodal data obtained via EEG or fMRI be analyzed so as to arrive at simple graphs that represents statistical dependencies between the different regions of the brain?

2. How can this data be pre-processed to make it amenable to graph embeddings techniques where statistical machine learning can be applied?

3. How can machine leaning techniques be used for meaningful discriminative classifications that can be critically interpreted?

4. Can a scalable framework be designed to take advantage of advances in cluster technologies (multi-core, networking, and storage) in the most-optimized manner for Neuroscience applications?

5. How can the existing graph representation platforms such as Pregel be used for efficient storage and retrieval of neuroscientific information?

6. Is it possible to identify the computation, communication, and I/O characteristics and requirements of driving Neuroscience applications/benchmarks on modern high end computing systems?

7. Can we perform co-design of the neuroscience applications with the programming runtimes available for high performance processors such as GPUs and MICs for ensuring best performance and maximum system efficiency?

2.0 Problems intended to be addressed by proposed project :

Problem Statement: Modern computational neuroscience have some fundamental bottlenecks in terms of efficient representation and visualization of multimodal brain data. There lacks mechanisms through which this data can be efficiently modelled which can be subjected to predictive classifications and functional analysis. Due to heavy computational overhead, it is important that the predictive mechanisms be accelerated through the use of modern commodity processors that can provide high accuracy rates in a feasible training time. The training should be also able to take advantage of large training sets that are available for higher accuracy and better interpretations.

As indicated earlier, in this proposal we aim to address some of the fundamental challenges in computational neuroscience through novel application of modern network science, machine learning and acceleration through the application of high performance computing. With the massive advancements made in the data gathering techniques and open sourced availability of neuroscientific data, it is imperative that research effort is focused towards the following broad problems:

1. *Graph based representations:* In order to visualize the multimodal data captured through popular mechanisms such as fMRI, functional modelling is necessary. It is required that this data is represented in a way that intuitively provides an abstraction for the statistical dependencies and reduce inter-class variabilities.

2. *Predictive modelling of graph representations:* In order to perform classical classification of the data, the graph structures are required to be critically interpreted. Towards this, application of modern machine learning techniques are already explored. However, higher accuracy rates are needed to be achieved which may lead to a performance cost.

3. *Fast classification and visualization:* Due to high compute involved in the computational implementation of machine learning methods especially deep learning, often such methods are not feasible due to high amounts of time spent in training. It is a fundamental challenge that needs to be addressed through the exploitation of modern HPC architectures.

4. *Handling large data:* Success of the machine/deep learning methods lie in the use of large training datasets which is widely available. However, storage and retrieval of such data is a fundamental challenge. It is necessary that the data be stored in an efficient manner and staged to the compute pipeline in a way which minimizes the gap between computation and I/O.

3.0 Applicability/ usage and future potential of the outputs/Technologies of proposed project :

3.1 Who has identified the problem and its relevance to the objectives of ICPS?

The domain of computational neuroscience has gained importance in the recent years due to its high impact in diagnosis of several diseases and illnesses. With the maturity of high precision data gathering equipment such as fMRI, MEG and EEG along with the open sourcing efforts adopted by several leading research communities across the world, computational methods in neuroscience has emerged as an exciting domain for research. With the relevance of different high performance computational techniques, machine learning, and mechanisms for Big Data challenges, computational neuroscience has truly emerged as the champion for inter-disciplinary research efforts. The problems identified in the project has been jointly identified by the PI and the co-PI each of whom

carry expertise and research experience in the field of computer science and neuroscience. Through several meetings we have been for see the natural intersection for the challenges faced by the computational neuroscience community that can be approached by the experts in the pure computer science domain from an applications point of view that can lead to high impact results. The collaborative effort will lead to results and software that can benefit communities in both domains equally.

3.2 How will the project outputs dovetail into the overall Research strategy of ICPS?

The problems proposed to be solved as a part of this proposal naturally falls in Theme-2 Data Science Research strategy of ICPS. As already indicated earlier, several of the sub-problems to be tackled will require the application of learning, distributed and parallel computing, dimensionality reduction, data visualization, and interpretation of data. Namely, the proposal has components where we will need to explore the application of machine learning/deep learning in performing predictive modelling of the brain networks. The brain network will be efficiently formed as a data visualization mechanism and will be possible only after performing successful de-noising, and dimensionality reduction. The final component involves implementation of the said mechanisms on modern high end computing systems involving development of parallel algorithms and distributed computing.

3.3 What are the likely impacts foreseen in the ICPS subject area or related areas?

A three dimensional impact can be expected from the output of this proposal in the ICPS subject area. In the first place, we shall be developing innovative applications of machine learning techniques on data that is gathered from openly sourced databases. This is even more relevant in India due to the high availability of subjects and a mature clinical neuroscience community who are constantly generating high volumes of data which awaits a scalable framework for analysis. In the second place, output from the project will showcase the advantage of modern high performance computing systems on computational neuroscience where scant work has taken place internationally. And finally, the proposal is expected to accrue sufficient material for offering inter-disciplinary computational neuroscience as a course for undergraduate and postgraduate students which will equip them with knowledge in neuroscience, machine learning, and big data analytics.

3.4 If the aim of the project is to develop an Operational System/Deployment then when the project will become self-sustaining. Who are its potential users (Govt/ PSUs/Private Industry/ Academics/start-ups etc) and suggest the enablement path.

With the output generated from this project, several communities will benefit. Clinical neuroscientists, neurologists, docotors, companies dealing with products related to big data research, and academia stand to benefit from the results and frameworks generated. The enabling will be done through open source release of the computational framework with proper documentation. Additional workshops and short-term courses may be offered for direct training and adoption.

4.0 International Review Status

The topic of computational neuroscience has attracted some of the most exciting research endeavors in the past decade. With the emergence of network based modelling, machine learning and advanced high end computing systems at low cost, the field has thrown open multiple research challenges. The concept of "networks" has now encompassed almost all research fields from pure natural sciences to social science and neuroscience. In [24], the authors first presented the concept of complex networks. The authors introduced the initial definitions of networks science and demonstrated how the study of different properties of the networks from the perspective of topology, connectivity and such could lead to the discovery of several critical information about the underlying physical/meta-physical phenomenon that the network is trying to capture. The term "network" has been used to neuroscience to identify different regions of the brain that are active during a generic mental state. The connectivity of two different cortical regions of the brain during generic cognitive tasks was demonstrated in [34]. The authors showed how the connectivity patterns between two different regions of the human brain can be similar during functional modelling. In [25], the authors demonstrated how structural and functional modeling can be achieved through graph representations using data that is obtained from common imaging sources such as fMRI, MEG, and EEG. A whole survey showing the success of networks science and graph theory in computational neuroscience was shown in [31] where the authors studied the benefits that can be obtained in functional modeling through the application of graph analysis.

Semantic web technologies has found applications in several scientific fields including neuroscience [50, 52, 42], where it is being employed to integrate neuroscience knowledge and to make such integrated knowledge more easily accessible to researchers. Emerging technologies for Map Reduce-based Big Data systems, such as Hadoop [2] and Spark [6], take advantage of scalable and fault-tolerant distributed file systems such as HDFS, are widely being leveraged in research in various fields of science have become available, including translational research in neuroscience. The research work by Tanimura et. al. [51] propose a scalable RDF data processing based on the general data processing platform of Pig [5] and Hadoop with several extensions, that portray the performance improvement achievable for bulk load and store operations, including the schema conversion cost from conventional RDF file formats, as compared to existing single-node RDF databases.

SPARQL [17] is a standard query language for RDF datasets. Most of the SPARQL query handling engines use a transaction-based query language, and focus at optimizing single queries in execution. In order to overcome these limitations, Liu et. al designed HadoopSPARQL [43], a Hadoop-based Engine for multiple SPARQL query answering engine. Similarly, Apache Jena [4] is a free and open source Java framework for building semantic web and Linked Data applications. It helps web developers to manage the various semantic components of the semantic web and linked-data application to conform with the standards of the W3C. An increasing amount of data is being represented using Resource Description Framework (RDF) [14] on the Semantic Web, but traditional semantic web frameworks lacked the ability to handle large amounts of data in a scalable manner. The research work [32] describes a framework built using Hadoop to store and retrieve large number of RDF2 triples. The framework consists of a Jena-based Semantic Web Framework for data pre processing and Map Reduce framework for query processing using Pellet OWL Reasoner6. Along similar lines, Apache Spark with GraphX [10] is being considered an ideal platform for distributed processing of RDF graphs [16]. SparkRDF [30] implements SPARQL query based on Spark, a novel in-memory distributed computing framework to reduce the high I/O and communication costs for distributed platforms.

4.1 National Review Status

In India several prominent research groups are actively investigating several of the emerging topics in complex networks. In this context we can broadly divide the contributions made into two significant sub groups where one type deals with mostly algorithmic and mathematical formulations for capturing and retrieving meaningful information from complex networks while the other has dealt with approaches aimed at more efficient systems implementations.

Understanding the structure of networks especially those in the region of social networks, transportation, academic, and scientific networks has been broadly studied across different research groups across the country. Especially the problem of understanding the community structure existing in static and dynamic networks has been studied for citation networks [28, 27] and social networks [23]. In each of these works the authors have broadly used popular graph parameters [26] such as degree distributions, clustering co-efficient, and neighbors. The presence of constant communities where a community structure is not affected by vertex re-ordering has also been studied in this context [29]. Studies pertaining to centrality measures have been carried out in [33] where the authors have proposed faster algorithms for updating betweeness measures in static and dynamic networks. Understanding human navigation is also a part of the research interests where the human navigation patterns could be understood from the analysis of transportation and human navigation patterns. These works are broadly discussed in [38, 39].

High performance computing and research in understanding emerging muticore and manycore architectures has been pursued actively in different research communities across India. Work along the implementation of better cache hierarchies and coherence protocols for multicores has been shown in [49, 45, 46, 35]. Better scheduling and work partitioning in the context of homogeneous and heterogeneous manycore computing has always been an open area for research. Works such as the ones proposed in [22, 37, 48] has proposed new mechanisms for better scheduling and work partitioning on heterogeneous systems. Thread and work scheduling for GPUs as shown in [20] provide new runtime mechanisms for better thread scheduling and higher system utilization. Control flow in massively parallel architectures such as GPUs and Intel MICs often form a major performance bottleneck which has to resolved through intelligent scheduling. Works such as the one proposed in [19, 47] proposes solutions which minimizes divergence and ensures higher system efficiency. In both of these works, the authors propose enhancements in runtime systems via the compiler and subsequent assembly level code which reelects good benefits at the user level application. Apart from these works which broadly lie in the context of this proposal, there has been an wide array of work that has been pursued in parallel computing and high performance computing by several Indian academic and industrial research groups.

Understanding graph primitives which as discussed earlier forms an integral part of complex network analysis has also been pursued in India. One of the earliest work for large graph traversals (BFS) and computing the shortest paths on massively parallel architectures such as GPUs was proposed in [53, 36]. Each of these early works mostly focused on pure GPU (homogeneous) algorithms which required re-visitation once heterogeneous computation gathered higher interest. In the recent years, heterogeneity in high performance computing has emerged as a leading paradigm. Towards this different communities have actively looked a re-design, and evaluation of algorithms suited for heterogeneous platforms where multiple muticore and manycore processors may reside. Different programming models based on MPI such as MPI+OpenMP, and MPI+X in general provides state of the art programmability on such platforms. Towards this, in [21] authors proposed mechanisms for efficient connected components, BFS and shortest path computations on heterogeneous CPU+GPU systems. In all of these works, authors have only investigated the problems on

single standalone heterogeneous node and not on distributed memories. Although most of these algorithms were targeted at shared memory single node HPC systems, it is of high interest to re-investigate these problems with larger data, dynamic conditions, and observing impact of heterogeneity.

4.2 Gap areas identified between International and National

From the existing literature work we can see that there has been significant diversified efforts towards network based modelling for neuroscientific data internationally. Nationally significant research has been carried out for the implementation and understanding of complex networks as well as high performance computing. From the said research, we can identify the gap areas that are existing as follows:

1. There can be wide variabilities that can be introduced in the data that is captured through popular mechanisms such as fMRI, MEG, and EEG. These variations can be due to variations on modalities, velocities of capture, and subject based. How can existing network based modeling techniques be improved to take into consideration these variabilities?

2. Deep learning techniques have found wide spread success in recent times owing to two major factors. One is easy availability of large volumes of training data and the second factor being the advent of low cost high performance processors which can be programmed easily. In the context of neuroscience however, the application of deep learning is not explored. How can Deep Learning techniques be applied to substitute existing machine learning based predictive modelling techniques so as to gain better representations and accuracy?

3. In order to have an efficient implementation for the application of both network science and deep learning, it is essential to have a common framework that will provide a holistic solution having both upper level algorithms and under lying systems designs. Such an effort is currently unknown to us both nationally and internationally. Hence, how can such a framework such as the SGPS framework proposed in the earlier sections be designed?

4. The application of modern high performance processors such as Graphics Processing Units (GPU), and Intel Many Integrated Cores (MIC) are largely unexplored both nationally and internationally. Is it possible to exploit the raw compute power offered by such high performance processors through the development of parallel algorithms and efficient runtimes that can provide good benefits to next generation neuroscientific applications?

5. Because we will be working with a problem that will be requiring large volumes of data it is important that we investigate modern storage and networking technologies. What kind of benefits (in terms of performance and scalability) can be achieved by the proposed designs for representative Neuroscience applications and benchmarks on modern and next-generation high end computing systems with RDMA-enabled interconnects, NVRAM/NVMe-SSD and parallel file systems?

4.3 How this project will address the identified gap at the end-of-project

A synergistic and comprehensive research plan, with cross-layer focus and exploiting complimentary expertise of two different investigators, is proposed to address the above challenges. In order to address these challenges we broadly propose the SGPS framework which will model and accelerate Neuroscience applications. The broad activities to be performed by the proposed framework are:

1. Develop new techniques for more efficient representation of neuroscientific data through the use of modern network science.

2. Propose the exploitive Deep Learning mechanisms for arriving at higher accuracy and enhance predictive models for brain graphs.

3. Study the implementation, scalability and performance of the proposed SGPS framework on modern high performance computing platforms by taking advantage of performance and features of the latest multi-core, networking, and storage technologies (NVRAM/NVMe-SSD/parallel file system).

4. Study and explore the benefits offered by modern storage technologies targeted towards applications oriented around large volumes of data.

5. Demonstrate the benefits of the enhanced Big Data middleware as well as the SDPS related designs by using a set of driving Neuroscience applications and benchmarks.

5.0 Work Plan:

Our detailed work plan is as follows:

Table 1:: Work Plan

Sl No.	Task	Months
1.	Literature Survey and Applications Overview	1-3
2.	Graph Based Modelling for Inter-class variabilities and modalities	3-9
3.	Exploration of Machine Learning and Deep Learning for Predictive Modelling	9-15
4.	Re-visitation of graphical representations for large datasets	15-18
5.	Algorithms development and implementation on HPC architectures	18-30
6.	Burst Buffer, and NVRAM/NVMe based staging	24-30
7.	Designing efficient runtime for distributed execution, staging and RDMA	24-33
8.	Iterative and Integrated Evaluation	33-36

5.1 Approaches/Detailed Methodologies:

Motivation: As indicated earlier, recent advances made in the domain of machine learning, network science and high performance systems engineering is paving the way towards the development of more realistic applications and designs enabling better realization of brain functions. These features and the associated performance and scalability benefits offer great potential for significantly accelerating modern and next-generation Neuroscience algorithms/applications. However, current generation Neuroscience algorithms/applications and the associated execution frameworks have not been able to fully take advantage of the capabilities of these technologies in a significant manner even though most of these technologies have been available in the market for the last several years as open sourced libraries and commodity components at low cost.

Recent advances made in the domain of artificial intelligence and machine learning can be broadly focused in the area of Deep Learning. Although Deep Learning as a concept was proposed in the early 1990s, it has found renewed interests in the recent years due to the emergence of large amounts of data and cheap hardware. The success of Deep Learning can be attributed towards the availability of large amounts of data which allows the underlying architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) among the other techniques, provide accuracies that can match human identification and classification performances. Each of the underlying implementations involve heavy compute requirements which was one of the primary impediments towards its success in the early years. Recent revolutions brought about by parallel architectures such as multi-core CPUs many-core Graphics Processing Units (GPU) and high speed interconnects such as InfiniBand has led to wide spread development, research and adoption. In the context of Neuroscience however, there has been a limited application of the emerging Deep Learning mechanisms where by functional classification of brain activities could be modeled. With the combination of the modeling of large Neuroscientific data as graphs and applying deep learning principles on such models leads to remaining questions in the context of

pre-processing of data, performance and analysis.

With the emergence of high-performance and scalable data processing models (e.g. Tez, Spark), it is becoming critical to provide novel designs for Neuroscience algorithms/applications to achieve high-performance. As the Neuroscience systems aim to process large volumes of multi-modal data, such as electrophysiological signals and Magnetic Resonance Imaging (MRI) data, it is also critical to re-design the underlying data processing and data storage mechanisms to take advantage of modern Non-Volatile RAM (NVRAM), Non-Volatile Memory Express (NVMe)-SSDs, Remote Direct Memory Access (RDMA)-enabled high-performance interconnects, and memory within multi-core servers in an intelligent manner to reduce data access and data processing time. If the massive volumes of data are distributed across the nodes and stored in various storage mediums, such as DRAM, NVRAM, SSD, etc., it is essential for data storage frameworks to be aware of the hierarchy of the underlying architectures. This situation will require a holistic understanding about the computation, communication, and I/O characteristics of Neuroscience algorithms/applications and the interaction across different layers of applications, middleware, and hardware. Based on these observations, we identify that intelligent data processing approaches in Neuroscience execution frameworks need to be proposed to fully utilize the provided resources on high end computing systems.

A synergistic and comprehensive research plan, with cross-layer focus and exploiting complimentary expertise of three different investigators, is proposed to address the above challenges. The proposed research project, as indicated in Figure 1, takes on these fundamental challenges the objectives being the items enlisted in Table 1.

Problem Description and Research Challenges:

The development of scalable intelligent neuroscience processing systems is important for the communities to study the large electrophysical and multi modal datasets. Towards solving the broad questions that we are posing as a part of this proposal, it is necessary to carefully look at each of the questions in a cohesive manner. Through the following points we describe the challenges and the proposed research methodologies that we wish to undertake towards finding effective and efficient solutions.

In Figure 2, we show the three dimensions encompassing the modern approach to neuroscience. Through the use of machine learning, statistics and network science, we plan on investigating neuroscientific data with high

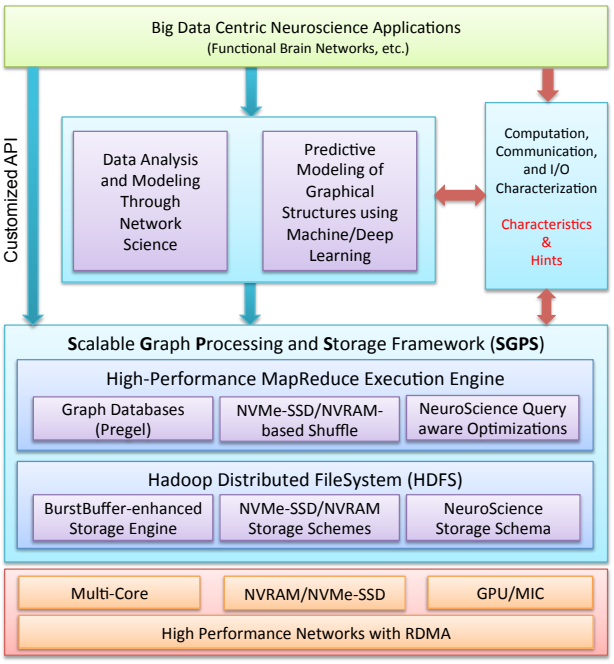

Figure 1:: The proposed data modeling and analysis framework for Big Data centric Nonscientific Applications.

1. Modelling and Analysis of Neuroscientific Data through Graphs:

As indicated earlier, the human brain can be visualized as a connected system of nodes or units that are interlinked through connections that represent the different communication pathways between the different regions of the brain. This abstract representation of the brain as a graph has allowed to visualize functional brain networks and describe their nontrivial topological properties in a compact and objective way. By

measuring the magnitude of temporal dependence between regional activities using functional modalities such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), or magnetoencephalography (MEG) functional connectivity patterns describe how N different brain areas interact with each other. With the advancements made in each of these data capturing mechanisms, there has been a huge burst in the volume of data that is produced and at the same time it has become a complex challenge to analyze these graphs due to the high compute requirement. In this regard it remains to be seen how traditional graph theoretic approaches such as traversals, computing of components, connectivity, and clustering stays relevant. There is also a need to apply modern theories proposed by in complex network science to model the data so as to gain meaningful insights. In this regard, it is imperative to also understand that given the large compute requirement in each of these mechanisms, we need to explore the domain of parallel algorithms in order to re-design the existing mechanisms for graph analysis and re-design them. This re-design catered towards the analysis of neuroscientific data has to be then evaluated against benchmark results keeping the accuracy values intact.

2. Application of Modern Machine Learning/Deep Learning on graphical data:

Data obtained from popular sources such as fMRI is often complex from the point of low signal-to-noise ratio, spatial correlations, long range temporal dependencies and high dimensionality. The importance of capturing spatio temporal dependencies make it highly desirable to arrive at a level of abstraction from where inference can be made easily. Graphs provide that abstraction due to its semantics and flexible representations through nodes and edges. The intersection of statistical machine learning techniques and graph representations has been of interest for several years in fields such as computer vision, pattern recognition, and data mining (as evidenced by regular workshops such as GbR (Graph-Based Representations in Pattern Recognition), SSPR (Structural and Syntactic Pattern Recognition), or MLG (Mining and Learning with Graphs)), but has only relatively recently started to be exploited in the context of brain networks, and formalizing neuroscientific questions as graph classification problems is a very recent trend. Given the current appeal of graphs for brain data representation and the simultaneous enthusiasm for machine learning approaches in the neuroimaging community it is of high interest to study different learning approaches in the neuroscience domain.

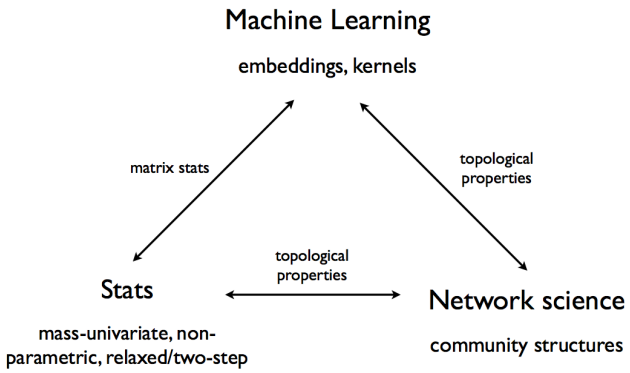

Figure 2:: Different dimensions for analytical neuroscience.

While conventional machine learning mechanisms such as neural networks, Support Vector Machines, Regression, and component analysis have found wide spread success, it is of immediate necessity to investigate recent successes in the AI community namely Deep Learning. In Deep Learning conventional machine learning structures have been added with another level of abstraction. Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and several other similar techniques can be applied to data obtained from fMRI, EEG and others to obtain brain space visualizations followed by the application of statistical confidence intervals to arrive at meaningful inferences. Figure 3 provides one such method where predictive modelling of brain graphs can be performed.

3. Computation, Communication, and I/O Characterization:

In order to efficiently design the proposed framework, the upper layer query APIs and engines to accelerate Neuroscience Big Data applications, their innate performance characteristics in terms of computation, communication and I/O behavior on modern clusters, need to be well understood. As cluster architectures are evolving rapidly, we first need to understand the interaction between Neuroscience applications and the different levels of I/O hierarchy and data movement costs. Then we need to classify their computation/communication and I/O ratios. To do so, we propose to use standard tracing and profiling tools such as Yourkit [18], iostat [11], jstat [12], btrace [8], and SL4J [15], to track application communication and data access patterns. In addition to the driving applications, we propose to design a set of microbenchmark suites for capturing the communication and I/O characteristics of Neuroscience applications, including popular datasets such as that of W3C Linked Data Sets [13] for RDF data processing and storage. Resource Description Framework (RDF) is a metadata data model which was originally proposed for data resources implemented in the web. In recent years RDF has also found applications in knowledge management applications and is widely used for managing neuroscientific data.

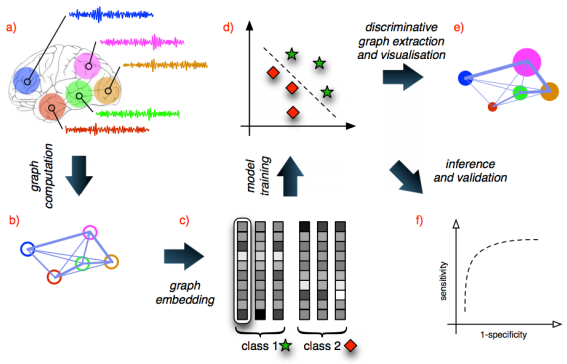

Figure 3:: Overall mechanism for using machine learning for predictive modelling of brain graphs.

Understanding and identifying these performance governing patterns will give more insights to the bottlenecks specific to these applications as they execute on modern clusters. For instance, by tracing and profiling the data store and access pattern of Neuroscience RDF data, a Neuroscience RDF storage schema will be designed to achieve fast I/O performance. Further, upon understanding predictable characteristics, execution time estimation for different computation and communication patterns will form the basis for designing the RDMA-based communication and NVRAM/NVMe-SSD-based shuffle techniques.

4. Optimized Neuroscience Data Analysis Libraries:

Many of the new generation neurotechnologies allow recording of high resolution data to study patterns of brain interconnections and functional networks at unprecedented levels of granularity, for example electrophysiological signal data and magnetic resonance imaging (MRI). Brain connectivity measures are derived from both functional and structural network data. Brain functional connectivity represents correlated brain activity recorded from different brain regions using multiple data modalities including EEG. The physical interconnections between brain regions represent structural connectivity and are derived from diffusion MRI data using fiber tractography algorithms. Brain connectivity data are intuitively represented as graph structures consisting of **nodes** representing single neuron or a brain region and **edges** representing structural or functional connectivity. Although graph network analysis is being increasingly used to study functional connectivity, the use of neurological big data requires the development of scalable graph analysis library that can leverage modern cluster technologies. A necessary innovation for Big Data processing based brain network analysis is the development of fundamental operations for distributed graph traversal, isomorphism, and partitioning. The proposed research plans for innovative use of the Resource Description Framework (RDF) with formal semantics, graph partitioning, and RDMA-based high speed interconnect techniques, including: (i) modeling connectivity data over a new semantic RDF graph processing tool, (ii) developing distributed graph analysis algorithms that easily scale over modern clusters, and (iii) defining new graph-based measures for characterizing

spatio-temporal properties of brain networks based on both structural and functional connectivity. These RDF data analysis libraries will allow the neuroscience research community, especially researchers focused on brain connectivity research to leverage the Neuroscience Big data to advance our understanding of brain networks. We propose to integrate these optimized Neuroscience data analysis libraries into Pig execution engine so that it can automatically and transparently work with NeuroPigPen.

5. Burst Buffer-enhanced Storage Engine for HDFS:

Although parallel file systems are optimized for concurrent access by large scale applications, write overheads can still dominate the run times of data-intensive applications. Besides, there are applications that require significant amount of reads from the underlying file system. In our case we wish to explore the Hadoop Distributed File System (HDFS) for distributed processing. For example, performances of neuroscience applications that analyze the interactions of the neurons in human brain are highly influenced by the data locality in the Hadoop cluster (when the data related to the interacting neurons are found locally, the processing can be done much faster). As a result, applications that have equal amounts of reads and writes suffer from poor write performance when run on top of Hadoop Distributed File System (HDFS); whereas, because of inadequate locality, read performs sub-optimally while these applications run entirely on top of a parallel file system. Such applications need a new design that efficiently integrates these two file systems and can offer the combined benefits from both of these architectures. Moreover, the limited local storage space on the compute nodes makes the deployment of HDFS challenging on HPC systems. The integration of HDFS with parallel file systems through the burst buffer not only eliminates the bottlenecks of shared file system during data write, but also offers locality through local copy of data in HDFS. Data-intensive neuroscience applications can leverage the benefits of this design via placement of the input schema data as well as the output of different queries. If the output of one query is needed by subsequent queries, then the output can be stored to the local storage devices to enhance data locality; otherwise this output data can be directly written to file system through the burst buffer. The burst buffer can take advantage of the emerging NVM also. Moreover, RDMA-based data transfer in both HDFS and key-value store ensures faster placement and access of the application data.

6. Integrated and Iterative Evaluation:

Design of the proposed scalable graph processing and storage (SGPS) framework, as well as the optimization of target upper layer query engine i.e., Pig is not independent from each other. The development process will follow a spiral model. For instance, along with designing the advanced features to leverage NVRAM/NVMe-SSD, parallel file systems (e.g. Lustre), and RDMA technologies, we need to redesign the storage engine and schema for HDFS as indicated earlier. In the meantime, in order to achieve the optimal data processing performance, enhancing existing Map Reduce engines, such as Apache Tez and Spark, need to be done by fully taking advantage of NVRAM/NVMe-SSD and RDMA technologies on modern clusters. For achieving both high-performance and high-productivity, we need to extend upper-layer query interface (e.g. Pig Latin) for Neuroscience applications as well as to enhance the query engines by highly optimized Neuroscience RDF analysis libraries. To evaluate the effectiveness of the scalable designs outlined in this proposal, we need to enhance and evaluate with current/next generation Neuroscience applications based on the underlying capabilities provided by the SGPS framework . Similarly, we propose to design each component in an iterative manner, evaluating its impact on and interoperability with other components in the framework.

There exist several benchmarks and applications that attempt to capture the behavior of various Neuroscience computing environments. Some general benchmarks like Logistic Regression, Latent

Dirichlet Allocation (LDA), TPC, PageRank, and TeraSort, which are widely used to characterize the performance of computation and communication for query engines as well as Map Reduce (Tez) and Spark engines. We also propose to design a set of new microbenchmark suites for capturing the communication and I/O characteristics of Neuroscience applications, which can be used to evaluate our designs.

In this proposal, we have identified a set of real-world Big Data applications from the fields of Neuroscience, that involves large scale electro-physical signal data management. They represent different data processing and storage paradigms, and highlight the impact of specific optimizations proposed in the framework. We plan to use these benchmarks and applications for performing a holistic evaluation of the proposed research framework, and demonstrate the effectiveness of the SGPS libraries and framework. We plan to evaluate with RDF data sets for Brain and Neuroscience data, that are part of the emerging web of Linked Data, and ontology-based scalable computing applications based on Apache Pig.

7. Timeline for proposed tasks

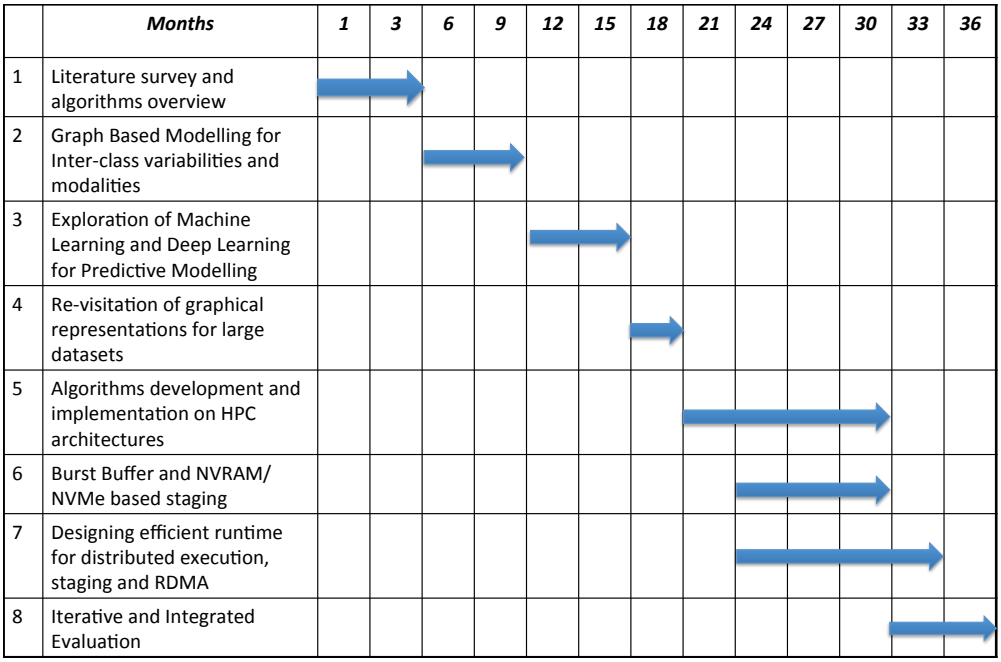

Figure 4:: Timeline for the tasks proposed.

5.3 Relevance of the project to the work already going on in the organization:

At the Indian Institute of Information Technology, Guwahati, we have already set up a high performance computing laboratory which is equipped with one workstation having multi-core and many-core processors. We are in the process of exploring complex network algorithms from the perspective of heterogeneous implementation and distributed storage. There has been two papers that have been published in this direction and we are in the process of exploring newer research directions in machine learning, deep learning, and parallel implementations of such problems.

5.4 Implementation arrangements proposed for the project:

For smooth execution of this project, it is necessary to have a small cluster with high speed interconnects at the PI's location. There is already a in-house machine which will be also connected to the cluster's final configuration. It is necessary that we equip ourselves with state-of-the-art processors so as to perform experiments and provide solutions that can be acceptable to the larger communities in general and also to stay competitive with other groups pursuing research along similar directions. The investigators also plan on gaining access from the large scale clusters deployed and maintained by C-DAC such as Param Yuva II for performing scalability studies and performance engineering.

5.5 Suggested plan of action for utilization of expected outputs from the project :

The expected output from the project can have utilization on multiple fronts as follows: 1. The computational framework that is released can be used by biomedical research teams for better analysis of neuro-imaging data and data gathered from other sources such as fMRI and EEG.
2. The software APIs proposed can be utilized for customized analytics of large scale neuroscientific data by the various research groups.
3. The high performance software can be used for deployment in large scale systems installed under the National Supercomputing Mission for scalable usage by different practitioners.
4. Knowledge gained from the project can be used for developing state-of-the-art courses at undergraduate and postgraduate levels for training new manpower in machine learning and big data analysis.

5.6 Suggestions for replicability and/or scaling-up of the research outcomes from the project:

As already indicated earlier we envisage to build our SGPS framework in a manner which can scale to multiple thousands of compute cores across multiple nodes that can support larger volumes of data and have support for larger number of users who expect quick training over large volumes of datasets. Also since the whole software package will be released as an open sourced project, other groups can easily modify and customize the package as per different data sources and computational requirements. It will also provide the mechanism for easy reproduction of the claimed results.

5.7 Expected/ Foreseen risks of implementation of the project, if any :

No risk in general is expected during the execution of the project.

6.0 Budget Estimates

Table 2:: Budget Estimates

Sl. No.	Budget Head	Budget in Rs. Lakhs			
		1st Year	**2nd Year**	**3rd Year**	**Total**
1.	**Project Staff:**				
	1. JRF/SRF - 1	3,67,200	3,67,200	4,10,400	11,44,800
	2. JRF/SRF - 2	3,67,200	3,67,200	4,10,400	11,44,800
	Sub Total	**7,34,400**	**7,34,400**	**8,20,800**	**22,89,600**
2.	**Equipment:**				
	1. Graphics Processing Units (GPU)	10,00,000			10,00,000
	2. Many Integrated Core (MIC)	9,00,000			9,00,000
	3. Solid State Drives	60,000			60,000
	4. Host Control Adapters	1,20,000			1,20,000
	5. Workstation	2,00,000			2,00,000
	Sub-Total	**22,80,000**			**22,80,000**
3.	**Domestic Travel**	50,000	50,000	50,000	1,50,000
4.	**Consumables**	25,000	10,000	10,000	45,000
5.	**Contingencies**	10,000	10,000	10,000	30,000
6.	**Institutional Overheads**	3,44,378	89,378	98,978	5,32,733
7.	**Grand Total**	**34,43,778**	**8,93,778**	**9,89,778**	**53,27,333**

6.1 Justifications:

6.2 Justification for project staff

Since the project broadly has two components. Two research students are envisaged. The first student shall be responsible for carrying out the advanced research, coding, and development of the graph based modelling and machine learning aspects for nonscientific data. The other student will be responsible for pursuing research, coding and evaluation of the high-performance runtime to support the algorithmic implementations and data centric optimizations. Both the students shall be recruited as JRF for first two years and SRF for the last year.

6.3 Justification for Equipment

1. Two workstations to form the chassis of our high-performance nodes are required that will be able to host and house the GPU, MIC and InfiniBand with the least power overhead.

2. Since a majority of work will be focused on high-performance architectures we need a local server hosting the latest GPUs. We understand that several of the supercomputers currently installed in India has GPUs installed, does not host the configurations required for our research. GPUs connected via InfiniBand will SSD support is the configuration we are envisaging to work on. Latest technologies for GPUs such as RDMA, GPUDirect RDMA and GDR Async is available supported on latest models.

3. Another important architecture requiring evaluation is the Intel Many Integrated Core (MIC). Since they are sparingly available in the national installations we plan on installing it on our local compute node for implementation and experimentation.

4. PCIe based SSDs are required for the pursuing the component of I/O staging. Due to this we plan on installing a PCIe FUSION IO 1.2 TB SSD which will provide the best configuration for staging the large datasets.

5. The InfiniBand backbone will be essential for the setup of the local cluster composed of two nodes. The Extended Data Rate (EDR) architecture from Mellanox currently is capable of providing the best latency and bandwidth along with the PCIe-based Host Control Adapters. ConnectX-3 onwards Mellanox provides support for GPUDirect RDMA.

6.4 Justification for Consumables:

Since experimental work will start from Year-1, we will need to purchase several cables, storage devices, printer cartridges, and small fittings for setting up the hardware. This will require help from local electricians and carpenters. Second year onwards the setup cost will no longer be required.

6.5 Justification for Contingencies:

Contingencies are necessary for extraneous expenses such as maintenance of hardware, purchase of books and access to online study materials.

6.6 Justification for Domestic Travel

Since two PIs are located on geographically different locations, it will be necessary for the project staff to visit each other on a regular basis. Initial travel also required to attend workshops, seminars and summer schools all held nationally on algorithms and parallel computing. Second year onwards will require travel for attending conferences and workshops for attending and presenting papers.

7. Is Institute registered with Public Financial Management System (PFMS, formerly called CPSMS) i.e., at cpsms.nic.in ? Yes

If Yes then provide unique code: AGENCYADM

References

[1] Apache Giraph. http://giraph.apache.org/.

[2] Apache Hadoop. http://wiki.apache.org/hadoop/.

[3] Apache Hadoop Yarn. http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html.

[4] Apache Jena. http://jena.apache.org.

[5] Apache Pig. https://pig.apache.org.

[6] Apache Spark. https://spark.apache.org.

[7] Apache Tez. https://tez.apache.org/.

[8] BTrace. https://kenai.com/projects/btrace.

[9] Govt to launch Rs 4,500-cr National Supercomputing Mission. http://www.cdac.in/index.aspx?id=pk_itn_spot948.

[10] GraphX. http://spark.apache.org/graphx/.

[11] iostat. http://linux.die.net/man/1/iostat.

[12] jstat. https://docs.oracle.com/javase/8/docs/technotes/tools/unix/jstat.html.

[13] Linked Data - W3C. https://www.w3.org/standards/semanticweb/data.

[14] Resource Description Framework (RDF). https://www.w3.org/RDF/.

[15] Simple Logging Facade for Java (SLF4J). http://www.slf4j.org.

[16] Spark and SPARQL; RDF Graphs and GraphX. http://www.snee.com/bobdc.blog/2015/03/spark-and-sparql-rdf-graphs-an.html.

[17] SPARQL Query Language for RDF. https://www.w3.org/TR/rdf-sparql-query/.

[18] YourKit Java Profiler. https://www.yourkit.com/features/.

[19] J. Anantpur and R. Govindarajan. Taming control divergence in gpus through control flow linearization. In *Compiler Construction - 23rd International Conference, CC 2014, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2014, Grenoble, France, April 5-13, 2014. Proceedings*, pages 133–153, 2014.

[20] J. Anantpur and R. Govindarajan. PRO: progress aware GPU warp scheduling algorithm. In *2015 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2015, Hyderabad, India, May 25-29, 2015*, pages 979–988, 2015.

[21] D. S. Banerjee, A. Kumar, M. Chaitanya, S. Sharma, and K. Kothapalli. Work efficient parallel algorithms for large graph exploration on emerging heterogeneous architectures. *J. Parallel Distrib. Comput.*, 76:81–93, 2015.

[22] T. Beri, S. Bansal, and S. Kumar. A Scheduling and Runtime Framework for a Cluster of Heterogeneous Machines with Multiple Accelerators. In *2015 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2015, Hyderabad, India, May 25-29, 2015*, pages 146–155, 2015.

[23] P. Bhattacharya, M. B. Zafar, N. Ganguly, S. Ghosh, and K. P. Gummadi. Inferring user interests in the twitter social network. In *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014*, pages 357–360, 2014.

[24] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(45):175 – 308, 2006.

[25] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[26] T. Chakraborty, A. Dalmia, A. Mukherjee, and N. Ganguly. Metrics for community analysis: A survey. *CoRR*, abs/1604.03512, 2016.

[27] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee. On the categorization of scientific citation profiles in computer science. *Commun. ACM*, 58(9):82–90, 2015.

[28] T. Chakraborty, S. Sikdar, N. Ganguly, and A. Mukherjee. Citation interactions among computer science fields: a quantitative route to the rise and fall of scientific research. *Social Netw. Analys. Mining*, 4(1):187, 2014.

[29] T. Chakraborty, S. Srinivasan, N. Ganguly, S. Bhowmick, and A. Mukherjee. Constant communities in complex networks. *Scientific Reports*, 3, 2013.

[30] X. Chen, H. Chen, N. Zhang, and S. Zhang. Sparkrdf: Elastic discreted rdf graph processing engine with distributed memory. In M. Horridge, M. Rospocher, and J. van Ossenbruggen, editors, *International Semantic Web Conference (Posters & Demos)*, volume 1272 of *CEUR Workshop Proceedings*, pages 261–264. CEUR-WS.org, 2014.

[31] F. De Vico Fallani, J. Richiardi, M. Chavez, and S. Achard. Graph analysis of functional brain networks: practical issues in translational neuroscience. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1653), 2014.

[32] M. Farhan Husain, P. Doshi, L. Khan, and B. Thuraisingham. Storage and Retrieval of Large RDF Graph Using Hadoop and MapReduce. In M. Jaatun, G. Zhao, and C. Rong, editors, *Cloud Computing*, volume 5931 of *Lecture Notes in Computer Science*, pages 680–686. Springer Berlin Heidelberg, 2009.

[33] K. Goel, R. R. Singh, S. Iyengar, and S. Gupta. A faster algorithm to update betweenness centrality after node alteration. *Internet Mathematics*, 11(4-5):403–420, 2015.

[34] M. D. Greicius, B. Krasnow, A. Reiss, and A. L. Menon. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100:253–258, 2002.

[35] N. D. Gulur, M. Mehendale, R. Manikantan, and R. Govindarajan. Bi-modal DRAM cache: Improving hit rate, hit latency and bandwidth. In *47th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO 2014, Cambridge, United Kingdom, December 13-17, 2014*, pages 38–50, 2014.

[36] P. Harish and P. J. Narayanan. Accelerating Large Graph Algorithms on the GPU Using CUDA. In *High Performance Computing - HiPC 2007, 14th International Conference, Goa, India, December 18-21, 2007, Proceedings*, pages 197–208, 2007.

[37] S. R. K. B. Indarapu, M. K. Maramreddy, and K. Kothapalli. Architecture- and workload-aware heterogeneous algorithms for sparse matrix vector multiplication. In *Proceedings of the 7th ACM India Computing Conference, COMPUTE 2014, Nagpur, India, October 9-11, 2014*, pages 3:1–3:9, 2014.

[38] S. Iyengar, C. E. V. Madhavan, K. A. Zweig, and A. Natarajan. Understanding human navigation using network analysis. *topiCS*, 4(1):121–134, 2012.

[39] S. Iyengar, N. Zweig, A. Natarajan, and V. Madhavan. A network analysis approach to understand human-wayfinding problem. In *Proceedings of the 33th Annual Meeting of the Cognitive Science Society, CogSci 2011, Boston, Massachusetts, USA, July 20-23, 2011*, 2011.

[40] U. Kang, C. E. Tsourakakis, A. P. Appel, C. Faloutsos, and J. Leskovec. HADI: Mining Radii of Large Graphs. *ACM Trans. Knowl. Discov. Data*, 5(2), Feb. 2011.

[41] U. Kang, C. E. Tsourakakis, and C. Faloutsos. PEGASUS: A Peta-Scale Graph Mining System Implementation and Observations. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, ICDM '09, pages 229–238, Washington, DC, USA, 2009. IEEE Computer Society.

[42] H. Y. K. Lam, L. Marenco, T. Clark, Y. Gao, J. Kinoshita, G. Shepherd, P. Miller, E. Wu, G. Wong, N. Liu, C. Crasto, T. Morse, S. Stephens, and K. hoi Cheung. Semantic Web Meets e-Neuroscience: An RDF Use Case. In *In Proceedings of the ASWC International Workshop on Semantic e-Science*, pages 158–170. University Press, 2006.

[43] C. Liu, J. Qu, G. Qi, H. Wang, and Y. Yu. HadoopSPARQL: A Hadoop-Based Engine for Multiple SPARQL Query Answering. In E. Simperl, B. Norton, D. Mladenic, E. Della Valle, I. Fundulaki, A. Passant, and R. Troncy, editors, *The Semantic Web: ESWC 2012 Satellite Events*, volume 7540 of *Lecture Notes in Computer Science*, pages 474–479. Springer Berlin Heidelberg, 2015.

[44] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski. Pregel: A system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 135–146, New York, NY, USA, 2010. ACM.

[45] R. Manikantan, K. Rajan, and R. Govindarajan. Nucache: An efficient multicore cache organization based on next-use distance. In *17th International Conference on High-Performance Computer Architecture (HPCA-17 2011), February 12-16 2011, San Antonio, Texas, USA*, pages 243–253, 2011.

[46] S. Pai, R. Govindarajan, and M. J. Thazhuthaveetil. Fast and efficient automatic memory management for GPUs using compiler-assisted runtime coherence scheme. In *International Conference on Parallel Architectures and Compilation Techniques, PACT '12, Minneapolis, MN, USA - September 19 - 23, 2012*, pages 33–42, 2012.

[47] S. Pai, R. Govindarajan, and M. J. Thazhuthaveetil. Preemptive thread block scheduling with online structural runtime prediction for concurrent GPGPU kernels. In *International Conference on Parallel Architectures and Compilation, PACT '14, Edmonton, AB, Canada, August 24-27, 2014*, pages 483–484, 2014.

[48] R. Prabhakar, R. Govindarajan, and M. J. Thazhuthaveetil. Cuda-for-clusters: A system for efficient execution of CUDA kernels on multi-core clusters. In *Euro-Par 2012 Parallel Processing - 18th International Conference, Euro-Par 2012, Rhodes Island, Greece, August 27-31, 2012. Proceedings*, pages 415–426, 2012.

[49] S. Rai and M. Chaudhuri. Exploiting dynamic reuse probability to manage shared last-level caches in CPU-GPU heterogeneous processors. In *Proceedings of the 2016 International Conference on Supercomputing, ICS 2016, Istanbul, Turkey, June 1-3, 2016*, pages 3:1–3:14, 2016.

[50] M. Samwald, H. Chen, A. Ruttenberg, E. Lim, L. Marenco, P. Miller, G. Shepherd, and K.-H. Cheung. Semantic SenseLab: Implementing the vision of the Semantic Web in neuroscience. *Artificial Intelligence in Medicine*, 48(1):21–28.

[51] Y. Tanimura, A. Matono, S. Lynden, and I. Kojima. Extensions to the pig data processing platform for scalable rdf data processing using hadoop. In *Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on*, pages 251–256, March 2010.

[52] R. Verbeeck, T. Schultz, L. Alquier, and S. Stephens. Relational to RDF mapping using D2R for translational research in neuroscience. In *Bio-Ontologies Meeting, ISMB*, 2010.

[53] V. Vineet, P. Harish, S. Patidar, and P. J. Narayanan. Fast minimum spanning tree for large graphs on the GPU. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS Conference on High Performance Graphics 2009, New Orleans, Louisiana, USA, August 1-3, 2009*, pages 167–171, 2009.