Transform Blockchain into Distributed Parallel Computing Architecture for Precision Medicine

Zonyin Shae
Department of Computer Science and Information Engineering
Asia University
Taichung, Taiwan
zshae1@gmail.com

Jeffrey J.P. Tsai
Department of Bioinformatics and Medical Engineering
Asia University
Taichung, Taiwan
jjptsai@gmail.com

Abstract—*This paper provides a vision and proposes mechanisms to transform the blockchain duplicated computing into distributed parallel computing architecture by transforming smart contract which features data as the first class of citizen to support moving computing to data strategy. This new distributed parallel computing architecture can be employed to build a large size of data set from distributed hosted medical data sets which consists of personal electronic medical record (EMR) and various medical data. This large medical data set will enable researchers to jump start the deep learning research for medical domain. Distributed data management, distributed data sharing, and distributed computing analytics and learning are the core mechanisms in the new architecture. The new researches and developments required to employ Google federated learning and transfer learning algorithms in this new architecture are also discussed. The approach and mechanism enabled by the new architecture is illustrated to build a real world evidence of clinical trial toward personal and precision medicine. Research issues and technical challenges are provided.*

Keywords—*blockchain; distributed and parallel computing; data ownership; data integration; health information exchange; big data analytics; deep learning; clinical trial; precision medicine;*

I. INTRODUCTION

In order to achieve peer to peer trust transaction among untrusted infrastructure and users across Internet, blockchain [1-3] is originally designed as a duplicated computing architecture that every blockchain node running the identical code with the same ledger. As a result, it suffers the scalability issue [4] that the performance (transaction latency and throughput) cannot scale up propositionally along with the number of nodes increasing. On the contrary, the performance of a single node is better than multiple nodes due to the faster consensus that blockchain broadcasts all the transactions of intent ledger modifications to all participants. It is through this broadcasting protocol that consensus of the distributed ledger modifications is agreed upon.

Most concerned issue of the duplicated computing is the waste of electricity. According to Digiconomist [5] the estimated power used for miner to verify bitcoin blockchain transactions is 30.14TWh a year, which exceeds electricity a year than Ireland as well as other 19 European countries. The duplicated transaction validation is repeatedly executing the hash computation of the validation puzzle by all miners in order to achieve consensus and maintain the data synchronization of blockchain distributed ledger. It would be very useful to channel these computing power for something which can provide some other benefits to society.

"Proof of Stack" [6] is one proposal to improve the wasting electricity in mining. It replaces the hash computation mining mechanism with the virtual mining mechanism in which the winning probability of block mining of each miner node is proportional to the amount of the virtual currency balance of the miner node. This approach resolves the wasting energy issue, but it is still a duplicated computing mechanism. Lightning network [7] approach reduces the number of transactions needed to be recorded in the distributed ledger for improving the performance. It creates a channel between two accounts. The created channel will be used by multiples of following up transactions between these two accounts. These intermediate transactions will not be broadcasted and recorded in the distributed ledger, but only the final results. From the distributed ledger point of view, it only sees one final transaction occurred. As such, lightning network reduces the loading of the number of transactions to improve the system overall performance. This approach does reduce some wasting energy by reducing the number of the transaction validations required, but it is still a duplicated computing mechanism. Sharding [8] approach does not let all the nodes to validate every single transaction, instead, it dynamically distributes the validation tasks for a given single transaction to a group of nodes. Such that the validations of different transactions can be done in parallel to improve the system performance. However, the parallel transaction validation, if not carefully managed, will be at risk of the double spending problem that the blockchain is designed to solve. Sharding approach provides a degree of parallelism in the transaction validation, but it only addresses the duplicated computing issue of transaction validation in mining space, but not provides a distributed and parallel computing architecture for arbitrary computation. Both ethereum enterprise alliance [9] and hyperledger Fabric [10] define a new category of smart contract, named private contract [11], in which the smart

This research is supported by the Ministry of Science and Technology under the grants MOST 106-2632-E-468-002 and MOST 106-2632-E-468-003.

contract data is kept privately in the local node, and not stored in public distributed ledger. This resolves the data privacy issue, and also improve the smart contract performance since less nodes are involved in the private contract. Private contract approach does not provide a distributed and parallel computing architecture for arbitrary computation.

There are some existed works in transforming the blockchain duplicated computing into parallel computing by transforming the consensus protocol of distributed ledger. For examples: FoldingCoin [12] from Stanford University for atomic-level simulations of protein folding, and GridCoin [13] from UC Berkeley. FoldingCoin created a "Proof of Fold" and GridCoin created a "Proof of Research" concept to verify contributed computational power of each participated blockchain node. SETI@home [14] from UC Berkeley accepts GridCoin. Volunteers of SETI@home are given small portions of observed radio signals to scan for potential patterns to identify signs of extraterrestrial life. Participants contribute their computing cycles to scientific research on the blockchain network instead of "Proof of Work" tasks on a traditional blockchain. However, both FoldingCoin and GridCoin need to replace the current blockchain consensus protocol and so are not compatible with the current existed commercial blockchain networks. Hoping that as ASIC miners begin pushing GPU users off the Bitcoin network, they could pick up some spare computing power from cryptocurrency miners and put it into a good use, for example, battle Alzheimer's disease. These two new distributed consensus protocols employ different consensus computing tasks in each node which are executed in parallel and not duplicated. However, only a rare set of computing tasks can be employed as consensus computation [15]. Such that the use cases for parallel computing is very limited.

In addition to the duplicated computing in maintaining the consensus of distributed ledger, there are also duplicated computing in the smart contract [16]. This paper will address the distributed parallel computing transforming aspect by transforming the smart contract. Smart contract is a new technology made possible by new generation of blockchain [17]. A traditional lawful contract written in paper states the terms of a relationship which can be enforced by law officials. A smart contract enforces a relationship with trusted code running inside the blockchain. Smart contract is created by users and deployed inside the blockchain and executed to enforce the rules exactly as the code is written. Smart contracts can generate events and are triggered by specific events that can be traceable and auditable. Smart contracts suffer from the even more severe duplicated computing issue since smart contracts need to be deployed into all the blockchain nodes, and the identical smart contract codes are executed at the same time in all the nodes. Since smart contract is a user created program code which can be any Turing complete computing intensive codes, e.g. big data analytics or artificial intelligent code, the waste of duplicated computation power is much more than the distributed consensus protocol. On the other hand, smart contract is critical to enable applying blockchain technology across various industrial sectors [18], for example, precision medicine. As such, various innovative mechanisms, especially those that are compliant with the existed commercial blockchain protocol, to effectively transform the smart contract duplicated computing into distributed and parallel computing are very much worthwhile to be investigated.

This paper proposes a new blockchain distributed parallel computing architecture for big data analysis, especially for the precision medicine which data sets are owned by different hospitals, patients, and health service providers and resided in separate locations. Data is the king for the big data analytics, deep learning and artificial intelligence (AI) applications. Data quality decides completely the analytics and learning results. It is reported [19] that 80% of time will be spent in the data preparation in the big data analytics and learning tasks. Our system architecturally positions the data as the first class of citizen of the system as it should be. The design strategy moves the computing engine to the native data sets which is different from the Hadoop mechanism which only moves the computing engine to the internally generated intermediate data during the code executing process. This new distributed parallel computing architecture can be employed to build a large size of distributed hosted medical data set which integrated personal EMR and various medical data. This large data set enables researchers to learn a set of core features and models for the medical domain. Transfer learning algorithm [20] can then be employed to extend these learned core features and models to jump start the deep learning research for medical domain. This approach is similar to employ the ImageNet [21] data set to jump start the transfer learning of convolutional neural network (CNN) [22] for image domain. The researches and developments required to employ Google federated learning [23] and transfer learning algorithms for this new architecture are also discussed. This paper will also illustrate mechanisms leveraging the new architecture to build a real world evidence clinical trial to fulfill USA FDA vision [24] of next generation of clinical trial toward personal and precision medicine.

Our main technical contributions in this paper are: (a) Provide a vision and propose mechanisms to transform original blockchain duplicated computing into distributed parallel computing architecture by transforming the smart contract, (b) Devise a new distributed parallel computing paradigm which positions data as the first class of citizen with moving computing to data strategy, (c) Illustrate mechanisms to build a large medical data sets from various distributed data sources owned and hosted by various hospitals, patients, and service providers to jump start the deep learning research for medical domain, (d) Investigate mechanisms for standardized data sharing to support real word evidence clinical trial toward personal and precision medicine, and (e) Provide a list of research issues and technical challenges to build such a blockchain distribute parallel computing architecture.

This paper is organized as follow. A brief description of the real world evidence clinical trial with precision medicine is provided at section 2, followed by the description of a new blockchain distributed parallel computing paradigm in section 3, in which the core mechanisms: distributed data management, distributed data sharing, and distributed computing analytics and learning using clinical trial as example are illustrated. The mechanism to transform blockchain smart contract into distributed parallel computing architecture for precision medicine is discussed in section 4, followed by a list of research issues and technical challenges in section 5. A summary is provided in section 6.

II. PRECISION MEDICINE

Barack Obama, president of USA, spoke about the Precision Medicine Initiative in his 2015 State of the Union Address [25, 47], “Tonight, I'm launching a new Precision Medicine Initiative to bring us closer to curing diseases like cancer and diabetes — and to give all of us access to the personalized information we need to keep ourselves and our families healthier.”. Its goal is to take individual differences in genome, environments, and lifestyles into account to better understand the complex mechanisms underlying a patient’s health and disease, and to better predict which personalized treatments will be most effective.

It was also reported in a recent Nature article [26] that people are taking medications that will not help them. The article reported that the top ten highest grossing drugs in the United States only help between 4% and 25% of the people who take them. For some drugs, such as statins routinely used to lower cholesterol, as few as 2% may benefit. There are even drugs that are harmful to certain ethnic groups because of the bias towards white western participants in classical clinical trials. The problem is due to the many issues related to current classical clinical trial process for new drug and needs to be enhanced to next generation of real world evidence clinical trial toward personal and precision treatment as the USA FDA proposed [27]. The real world evidence clinical trial will access the real personal EMR data directly from various hospitals and service providers as the trial go on, and keep on monitoring the efficacy and possible side effects after the drug is approved and used in public. To achieve this next generation of clinical trial process, it requires the collaboration among blockchain, AI, and medical domain expertise.

The precision medicine research is full of challenges. It needs to integrate various data sets and analytics tools, for examples, NGS-related data (bioinformatics) with analytic tools for people’ genome, EMR-related data (clinical informatics) with analytic tools for patient’ health, personal activity record with analytic tools for environments and lifestyles, and big data analytics and deep learning tools for real-time actionable report for treatment. As such, there are technical challenges across domains of computer science, biomedical science, and informatics science. In this paper, we emphasize from the computer science aspect. The technical challenges for computer scientists for precision medicine include (a) lack of common data format as well as real time data sharing standard, (b) lack of large and integrated complexity of heterogeneous data sets of various ownerships (EMR, wearable device health data, environment data, genome data, lifestyle data, and clinical trial data, experiment and research data) for big data analytics and deep learning research, (c) lack of distributed and parallel computing analytical tools for heterogeneous and distributed data, (d) lack of domain knowledge of bioinformatics and medical informatics (mechanism to enable cross domain and international collaboration is therefore extremely needed).

In this paper, we propose and discuss approaches using blockchain technology to address these above technical challenges. Blockchain smart contract technology integrated with AI is used to build a large scale heterogeneous medical data set by distributed management of various data sets of various ownerships and makes it easily available for big data analytics and AI research. Blockchain technology provides mechanisms to achieve real time data sharing standard. Blockchain can also enable a distributed parallel computing environment for heterogeneous and distributed data. Moreover, blockchain can provide a secure, transparent, trust, and easily auditable environment to promote cross domain and international collaboration.

III. BLOCKCHAIN DISTRIBUTED COMPUTING PARADIGM

Recently Hadoop computing [28], Grid computing [29], and Cloud computing [30] are among the most popular parallel computing paradigms. Hadoop computing paradigm is a centralized architecture. Each computing node requires high performance CPU and memory. The data needs to be resided at the centralized location of Hadoop Distributed File System (HDFS) hosted by a group of servers. During the Hadoop application code execution process, there are some intermediate data will be generated in each computation circle and will be temporarily stored in some of the HDFS nodes for the next computing circle to use. Hadoop will smartly select the right local computing engine to compute the right intermediate local data since Hadoop knows exactly which HDFS nodes store which particular intermediate data. In such a way, Hadoop employs the strategy of moving the computing to data strategy. However, the strategy is limited only in its internally computing process during the code executing time, and only deal with the intermediate data generated internally from the code executing process. Grid computing paradigm is a distributed and parallel architecture. Basically every node can ask to join the grid computing network and contribute its unused computing power for the aggregated computing. Each node performs a different task/application. Coordinating tasks on grids across distributed computing resources is a complex task. Grid computing is suited to multiple tasks which can be executed independently without internal communication. Cloud computing paradigm is a centralized computing architecture, in which the computing resources can be virtualized into virtual machines for parallel but individual computing model. The cloud computing resources featuring with the elasticity property. There is hybrid computing paradigm [31] in combing the cloud elasticity property into the grid computing using super computer grid.

All these above approaches architecturally treat the computing engines and data sets separately and independently. These approaches all assume that they own all the data sets and the data is available and placed at the right location for the computing engine to freely access. In fact, architecturally their designs consider only the computing engine parts. However, in the big data computing world of big data analysis, AI, and deep learning, data is the king. Data sets can be owned by different entities and organizations. Analytics results will depend on the data used. Same data will possibly have different result using different brand of analytics tools. It is therefore logical that computing engine and data sets should be considered as an integral entity in the design of a distributed and parallel architecture right at the beginning of the system architecture design.

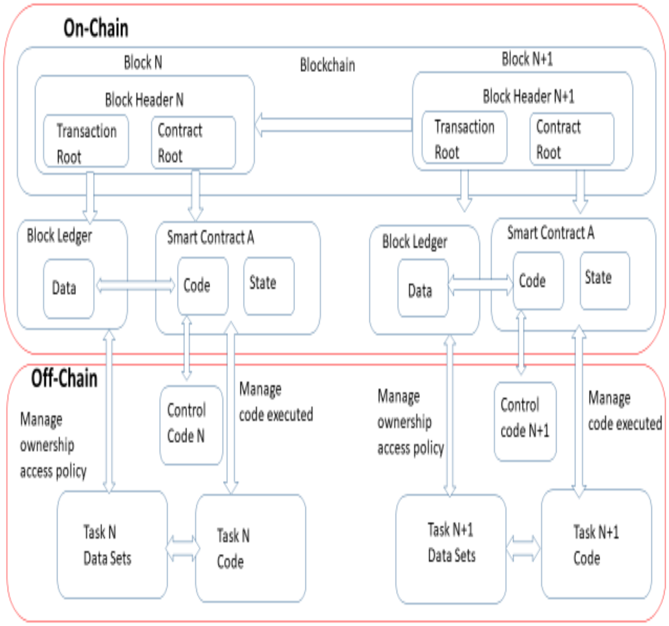

Figure 1. System architecture for blockchain distributed parallel computing paradigm

The mechanism we propose for investigation in this paper is to design smart contracts as light weight as possible and only function as the access policy control point which in turn manage the off-chain real arbitrary computation codes and data sets. The system architecture is illustrated in Figure 1. The smart contract code is deployed in blockchain and run the same code in all the nodes (smart contract A in Figure 1).

The on-chain smart contract code communicates with the different off-chain control codes (control code N and control code N+1 in Figure 1). For a given on-chain smart contract code which runs on all blockchain nodes are identical. Such that our system is compliant with the existed commercial blockchain protocol. However, the off-chain control code which communicate with on-chain smart contract of each node is different. Each individual control code will feed different data to the smart contract. As a result, each smart contract on each node will effectively behave differently. Effectively, the behavior of smart contract of individual node will be independently managed by its correspondent control node. The control node will redirect the smart contract to utilize the various off-chain analytics tools for various off-chain data sets at the right time. The on-chain smart contract will be used to enforce the ownership right and fine grain access policy of off-chain data and analytics code. Each off-chain data and analytics code will need to register and record its ownership right and access policy in the blockchain.

Leveraging this innovative seamless collaboration of on-chain and off-chain technology, smart contract in each node therefore accesses the different off-chain data sets (task N data sets and task N+1 data sets in Figure 1) and coordinate the different off-chain arbitrary computation codes (task N code and task N+1 code in Figure 1). As such, our mechanism will be compliant with the existed commercial blockchain network protocol and effectively transform blockchian into distributed parallel computing architecture. This architecture positions the data as the first class of citizen and is designed to seamlessly integrate code and data. The data sets can be resided virtually everywhere. We assume that the data sets will be owned and hosted by separated and independent organizations, for examples, hospitals, patients, and service providers. Data sets will be protected securely inside each secure infrastructure of hosted sites. We assume that all the data hosted sites will build its own computing infrastructure to run analytics tools. The system will automatically detect which computing tools are required and then deploy and run the analytics tools for the right data sets at the hosted site. In this case, innovative distributed and parallel computing algorithms are needed to avoid that expensive computing facility is required for each individual data hosted sites. For examples, the researches and developments of new innovative decomposition mechanisms are required to decompose a complicated analytics into distributed and parallel tasks which can be run in the blockchain distributed parallel smart contract environment. The new mechanisms allowing the blockchain smart contract to manage and enforce its integrity of the off-chain data and code are also needed to be developed.

In the case that we need for data exchange among sites for the analytics, real world evidence clinical trial, and medical practice purpose, or in the case where the required computing is too expensive to be deployed in all individual data hosted sites, we can then build a centralize computing entity which run by trustful 3rd party, for example, government agency like FDA. Our system can then provide a standard secure and real time health information exchange mechanism that the required data sets can then be dynamically exchanged to each other. In this case, Figure 1 describes the system architecture of each data hosted site. Each data hosted site represents a medical blockchain node in Figure 2 which will communicate with each other and form a global medical blockchain network illustrated in Figure 2. Government agency (e.g., FDA) can be as a special blockchain node in the medical blockchain which can facilitate as a trusted or law-required middleman for health data exchange and can be equipped with powerful computing engine and large data storage infrastructure. The system architecture to support the global medical blockchain is illustrated in Figure 2.

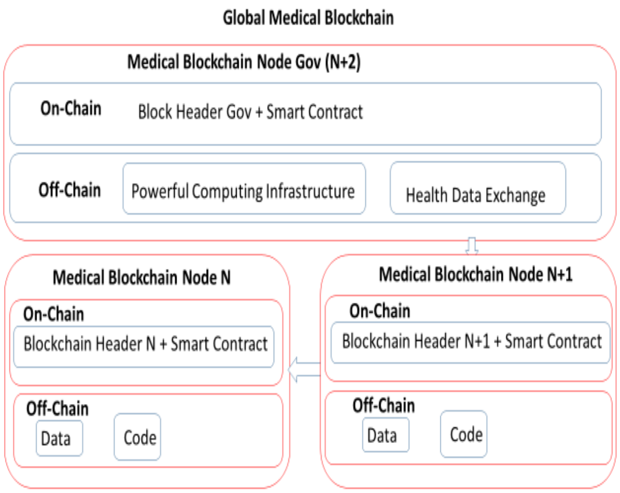
Global Medical Blockchain
Medical Blockchain Node Gov (N+2)
On-Chain
Block Header Gov + Smart Contract
Off-Chain
Powerful Computing Infrastructure
Health Data Exchange
Medical Blockchain Node N
On-Chain
Blockchain Header N + Smart Contract
Off-Chain
Data
Code
Medical Blockchain Node N+1
On-Chain
Blockchain Header N+1 + Smart Contract
Off-Chain
Data
Code

Figure 2. System architecture for Global Medical Blockchain

Our architecture makes use of blockchain smart contract trust mechanism to protect the data ownership as well as privacy, and provide fine grain access policy of the data sets. Data sets can be distributed resided as it was and no need to make copy or move the data sets. Moreover, smart contract can select the most suitable analytics tools and learning models for the particular data sets at the right time to achieve the optimal results.

In order to achieve this goal, the system needs at least to be able to perform these innovative core mechanisms: (a) distributed data management, (b) distributed data sharing, and (c) distributed computing analytics and learning. Each individual core mechanism is full of its own complicated challenges and many innovations are required. Since these mechanisms and systems should be different with different application domains. In order to have in depth investigation, we will investigate this transforming mechanism starting from the precision medicine domain.

A. Distributed Data Management

There is a positively proportional relationship in the amount of training data required and the size of parameters of the model. The number of parameters in the model should be large enough to capture relations in the training data. If the complexity of the problem is high, the number of parameters and the amount of training data required is also very large. It is therefore generally required to have a large amount of training data for a successful deep learning of a realistic application. For examples, it requires 1.2 M images for VGGNet [32] to train 1000 image categories of 140M parameters; requires 1.1M spot videos for DeepVideo [33] to train video categories of 100M parameters; and requires 6M sentence pairs and 340M words for GNMT (Google translation) [34] to train English text to French text translation. Moreover, most recent innovative deep learning CNN (convolutional neural Network) networks, such as, AlexNet [35], VGG [36], Network in Network [37], GoogleLeNet [38], and ResNet [39], they are all performing training and learned models using a very large common core dataset (e.g. ImageNet [40], which contains 1.2 million images with 1000 categories). Due to the availability of the large ImageNet core common data sets in the image domain, researchers can build a set of core learned features to advance the deep learning research in the image domain by leveraging transfer learning mechanisms. In practice, very few people train an entire CNN from scratch with random initialization for image domain, because it is relatively rare to have a training dataset of sufficient size.

TCGA [53] is currently the most important and popular database for the cancer and precision medicine research around the world. TCGA is managed by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) of Unite States. "TCGA collected and characterized high quality tumor and matched normal samples from over 11000 patients. The data set contains (a) clinical information about participants, (b) metadata about the samples, (c) histopathology slide images from sample portions, and (d) molecular information derived from the samples.". It is very expensive and huge efforts to be able to build TCGA database for cancer research. Currently, the TCGA database is therefore served as a corner stone for the cancer and precision medicine research. Huge numbers of academic research papers had been published based on it. Although TCGA comprises more than two petabytes of genomic data, but the size of the sample points (11000 patients) is very small, by TCGA itself is far from sufficient to support reasonable deep learning researches for medical domain under normal circumstances.

We need to have mechanisms to establish such large size of an initial core big data training sets to jump start the deep learning for medical domain. The availability of large core training data sets creates a high barrier in almost every domain, the huddle is even much higher in the medical domain due to the huge size of the distributed data sets, ownership, privacy, and administrative and government regulation/policy imposed to the medical data. Most biomedical, medical, personal healthcare related wearable IoT devices databases have been developed and owned in silos. Patients went to various healthcare providers and hospitals through the course of their lives and leave their EMR scatted around in various medical databases. Biomedical and public health researchers require the ability to analyze information from many sources in order to identify public health risks, develop new treatments and cures, and enable precision medicine.

A recent study by Greg Irving and John Holden [48] showed that using the blockchain is a low-cost independent verification method for verifying the report data integrity of scientific research. They proposed to create a hash for the raw data set and create a transaction in the public bitcoin blockchain distributed ledge to store the hash value of raw data in the created blockchain transaction. As such, the data modification can be easily detected by any peer. This approach can be used to link the off-chain data sets to blockchain. Alternatively, the authors of this paper also previously described approaches [18] to integrate various data sets by creating a virtualized SQL data based on the schema request from user's query.

We propose in this paper that Blockchain smart contract technology and monitor node will be applied to securely and privately build correlated personal healthcare records from various locations. Blockchain technology can be a solution to build such a large size core initial training data sets for medical domain by enabling the individual and distributed EMR data sets hosted by various hospitals and service providers to be securely available as a group for deep learning and advance the transfer learning in medical domain. The system architecture for heterogeneous data integration is illustrated in Figure 3. A monitor node is used to monitor all the related smart contract events which would like to access the managed heterogeneous data sets. The monitor mode is a mechanism for our system to securely bridge the smart contract and the external world by remote procedure calls which will return a standard format to smart contract access.

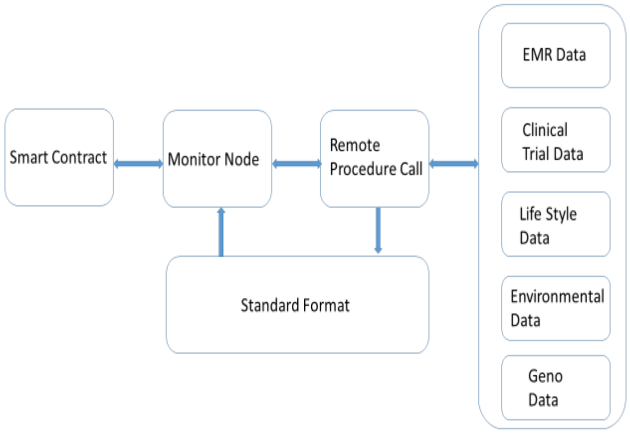

Figure 3. System for heterogeneous data integration

B. *Distributed Data Sharing*

Since our data sets is virtually composed from distributed data sets as previously described in the distributed data management section. We need to have a mechanism to ensure the data integrity and share the data for global modeling. To make this point clear, we discuss in more detail using clinical trial as example. For the data integrity issue, since 2007, US drug regulators have asked all clinical trials that test recruited subjects must be registered in the publicly accessible database ClinicalTrials.gov [41]. However, despite mandates for open access to the protocols and data captured in clinical trials, the issue of clinical trial data integrity remains. According to COMPare [42], a recent project to monitor clinical trials, just nine in 67 trials it studied (13 percent) had reported results correctly. There are near 20000 registrations per year at ClinicalTrails.gov, COMPare project findings show only the tip of the iceberg of the data integrity problem. China government reported [43] about 80% of clinical trial data performed in China is falsified. Blockchain can provide transparency and integrity to all data that could dissuade those who attempt to selectively report only good outcomes. For the data sharing issue, IT systems that now support the US Health Information Exchange (HIE) medical data exchange standards [44] are both opaque and un-auditable. Even there are many complaints reported to USA government about the data sharing problems encountered, and the USA government understood and acknowledged there are violations of the HIE law, but USA government cannot decide which involved parties to blame due to the complexity of the process as well as the opaque and un-auditability in nature of the current IT system. USA government is therefore reported [45] that it is the current IT system to be blamed. In addition, HIE medical data exchange is conducted through secure e-mail. As a result, various medical data sources cannot be integrated, and cannot directly be used for AI analysis to support accurate medical treatment. Blockchain can be used to explore medical data sharing mechanisms that can be standardized, transparent, auditable, and directly interfaced with analytics tools and AI computing.

In order to achieve the FDA vision of future real world evidence clinical trial, personal clinical trial data needs to be directly accessed from various hospitals and service providers. Such that it can be easily traceable to audit the trial, recruit the unbiased trial participants, and continuously monitor in near real time for any personal side effects and drug efficacy. Blockchain can provide mechanisms for standardized medical data sharing. Currently, there is an initial discussion of standardized activity for clinical trial based on blockchain from IEEE standard association [46] supported by USA FDA, and we had submitted a proposal for standardization. This blockchain for clinical trial standardization activity is a good starting point to investigate further the distributed data sharing mechanism.

C. *Distributed Computing Analytics and Learning*

Standard machine learning approaches require centralized the training data on a location where the computing engine co-located. However, in most of the cases it is impossible to copy or move the medical related data sets to a centralized location. Medical data sets are created and owned by different service providers and individual patients and hosted by various hospitals and service providers at different locations due to privacy, data ownership, security, huge size of the medical data sets, and various access policy requirements. The data sets become more decentralized for the precision medicine research since the analytics needs to combine all the genome, life style, and environment data sets. These data sets could also be generated from various wearable devices and hosted virtually everywhere.

Transfer learning mechanism formed the basis of the CNN deep learning models for image domain. Previously we had discussed to leverage blockchain distributed parallel computing architecture to build a core distributed data sets to serve as a basis for the transfer learning for medical domain. However, the ImageNet data set for the transfer learning in the image domain is centralized located. So are the current transfer learning algorithms which are centralized. There is no distributed transfer learning algorithm available. In our case, the data sets are distributed located, so there is a need to investigate distributed transfer learning algorithms that can be executed in distributed and parallel fashion. Especially, it should be a good fit that the distributed transfer learning mechanisms become an integrated part of the distributed learning mechanism in the new blockchain distributed parallel computing environment.

Google researchers introduced a distributed learning approach, named federated learning, that enables mobile phones to collaboratively learn a shared prediction model while keeping all the training data on local devices. The models learned from individual devices can then be composed into an improved completed learned model in a centralized location. All the training data remains on devices locally. Google federated learning allows for smarter models, lower latency, and less power consumption, all while ensuring privacy.

Google federated learning has many algorithmic and technical challenges. Google federated learning approach makes use of the millions of heterogeneous phones as computation and data storage devices. In addition, these devices have significantly higher-latency, lower-throughput network connections and are only intermittently available for training. Google federated learning research are currently focus on solving the high latency and low throughput network connection challenges.

We will apply the core idea of Google federated learning and transfer learning to transform blockchain smart contract into distributed parallel computing architecture. However, the technical challenges encountered in the Google federated learning and transfer learning are different from our situation. Our system assumes each individual computing server is a very powerful computing engine which is needed since most of the analytics and machine learning algorithms in the medical domain demand complicated computation. In our setting, the computing engine and data are distributed located. Therefore, our main challenge is to innovate a new distributed transfer learning algorithm, as well as to innovate a modified distributed federated learning with an optimization of query decomposition algorithm and implemented into the smart contract environment to meet the user query demand.

IV. TRANSFORM BLOCKCHAIN INTO DISTRIBUTED PARALLEL COMPUTING ARCHITECTURE FOR PRECISION MEDICINE

This is an international collaboration project among Asia University, China Medical University, Taiwan, University of Missouri, University of California, Irvine, and National Institute of Health (NIH), USA. The goal is to design and build an international AI medical blockchain data platform for AI research in healthcare and precision medicine. We begin our investigation with clinical trial, brain stroke and cancer diseases. The research focus is to transform blockchain smart contract into distributed parallel computing architecture to achieve the patient centric medicine by exploring standardized, trusted and secure data exchange methodologies with deep learning analysis for personal healthcare and precision medicine. The overall smart contract management architecture is illustrated in Figure 4.

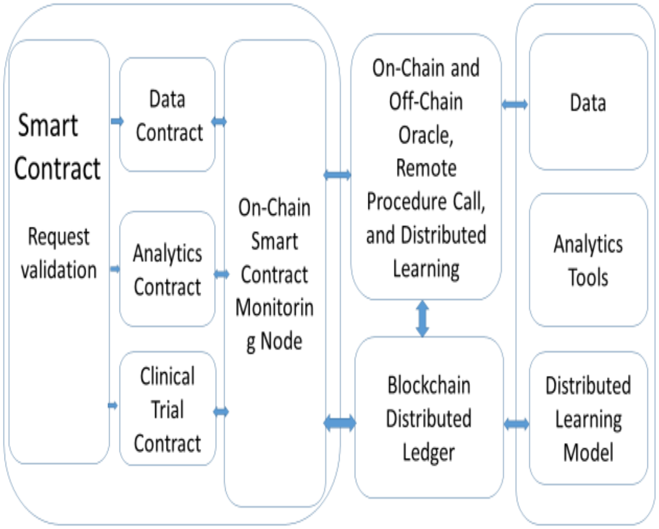
Smart Contract
Request validation
Data Contract
Analytics Contract
Clinical Trial Contract
On-Chain Smart Contract Monitorin g Node
On-Chain and Off-Chain Oracle, Remote Procedure Call, and Distributed Learning
Blockchain Distributed Ledger
Data
Analytics Tools
Distributed Learning Model

Figure 4. Blockchain smart contract management architecture

The smart contract will first validate each input request before executing. There are 3 categories of smart contract request: data contract request, analytics contract request, and clinical trial contract request. Data contract is used for requesting the data sets, analytics contract is used for requesting the computing of analytics tools or learning models, and clinical trial contract is used for managing clinical trial participants' recruitment and continuously trail monitoring. For security reason, on-chain smart contract is strictly limited or without direct external communication capability with outside world, and so we need to design a special data oracle mechanism by remote procedure call to bridge on-chain smart contract monitoring and off-chain data sets, analytics tools, and learning model as shown in Figure 4.

The huge size of the medical data set renders the operations of copying or moving data around for the analytics computing very expensive and impossible most of the time due to the privacy and ownership issues and medical data policy requirements. Medical data sets are therefore needed to be hosted locally inside the IT infrastructure premises of each individual hospitals or medical service providers. The transformed blockchain smart contract mechanism will apply under such environment and move the computing engine to the data for analytics and various distributed learning models.

Architecturally, the transform consists of three steps: (a) decompose the data query and analytics request into local systems as illustrated in Figure 5, (b) execute the transformed blockchain smart contract in each local system as illustrated in Figure 6, and (c) compose the local models and results into completed model and result as illustrated in the data service modules in Figures 5 and 6.

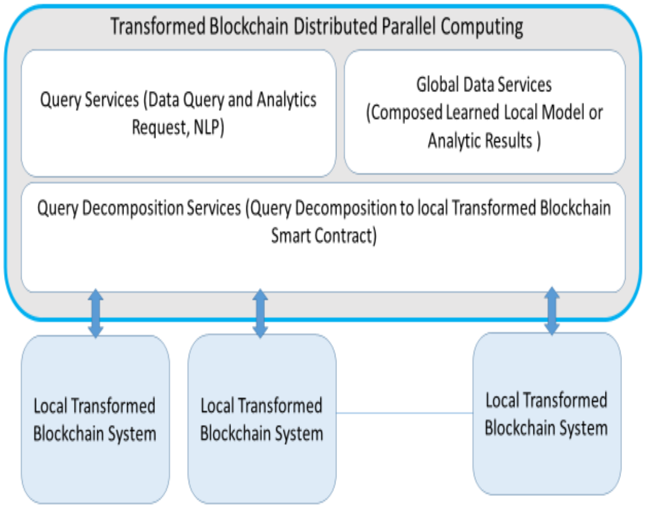
Transformed Blockchain Distributed Parallel Computing
Query Services (Data Query and Analytics Request, NLP)
Global Data Services (Composed Learned Local Model or Analytic Results)
Query Decomposition Services (Query Decomposition to local Transformed Blockchain Smart Contract)
Local Transformed Blockchain System
Local Transformed Blockchain System
Local Transformed Blockchain System

Figure 5. Decompose query into local system smart contract

The overall system consists of two layers, the top layer (Figure 5) is used for users to submit the request, and the bottom layer (Figure 6) is used for the individual participated medical entity or organization to execute the request. The system provides mechanisms for users to submit requests for retrieving the data or executing analytics tools. If the users' submitted requests are retrieving data, the system will return the encrypted data which only the requesting user can decrypt. Users (e.g., researchers) do not need to know where is the data physically resided, in fact, users can not know where is the data resided most of the cases. The system will automatically return optimal data retrieved and compiled from various distributed data sets. The returned data format will be based on users' requested schema. If the users' submitted requests are executing analytics tools for particular data. The system will execute the requested analytics tools locally with the locally data resided in each individual premise (see Figure 6). The individual results will then be composed and models updated before returning to users (see Figure 5).

We need to transform the data sets hosted by individual hospital and service provider as illustrated in the "local transformed blockchain system" component in Figure 5. The transform mechanism has been described in the section 3. The function of the query service component in Figure 5 is to understand the user request by nature language processing (NLP) and decompose the requests into various local transformed blockchain system to access data and execute the request. Users can also submit the requests in the form of query vector which consists of various parameters expressing the users' query interest. After individual data or learned models have been returned from individual local systems, the models will be composed and optimally updated by global data services component before returning to users.

The request decomposition and composition operation is technically very challenging, and full of research topics. For examples, how to convert and map NLP to the query vector, what is the optimal parameters in the query vector for a given query topic, how to convert the query vector into smart contract, how to compose and updated the global model from various distributed local models (federated learning), how to extend the distributed learned models into the other model in the medical domain (distributed transfer learning).

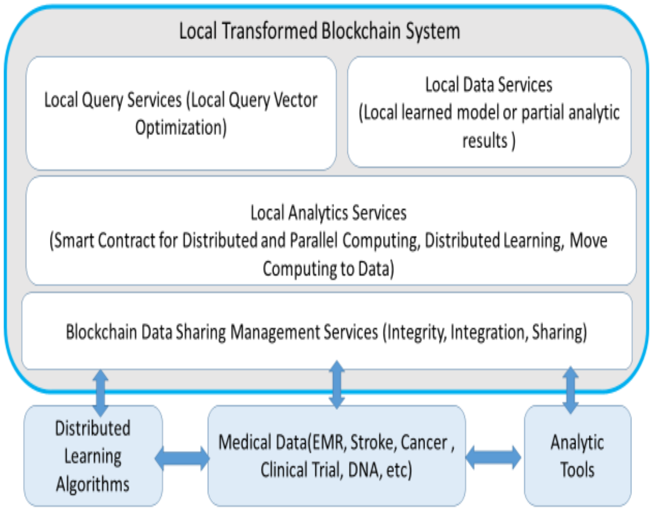
Local Transformed Blockchain System
Local Query Services (Local Query Vector Optimization)
Local Data Services (Local learned model or partial analytic results)
Local Analytics Services (Smart Contract for Distributed and Parallel Computing, Distributed Learning, Move Computing to Data)
Blockchain Data Sharing Management Services (Integrity, Integration, Sharing)
Distributed Learning Algorithms
Medical Data(EMR, Stroke, Cancer , Clinical Trial, DNA, etc)
Analytic Tools

Figure 6. Local transformed blockchain system

The "local transformed blockchain system" component in Figure 5 is zoomed into next level of details as illustrated in Figure 6 where it represents a single protected premise of individual hospital or service provider that hosted the data sets and analytics tools. The input to this system is the decomposed query vectors that sent from global query service. Each local system needs to transform and map the query vector into smart contracts for data access and analytics tools execution.

The main functions of various modules shown in both Figure 5 and 6 are briefly summarized as follows. **Blockchain Data Sharing Management Services s**afeguard data ownership, security and privacy to establish relevant personal health records from different locations and relevant biological data in different databases. Blockchain technology is used to integrate and enable secure and auditable data sharing among various localized databases hosted by various health service providers. This module will support standardization of HIE and enforcement of data integrity for clinical trial. **Data Services** provide good quality of data. The good analytics results of AI algorithms are from the quality of the data, not the amount of data. The good data quality needs to be based on a set of standard data collection systems and analysis tools. Each analytics tool and targeted disease research will need its own optimized data properties for better results. Utilize AI to optimize the common data format for integrating various EMR and medical data sets. **Analytics Services** move computing to data (medical big data size is not suitable to move data to computing strategy). Blockchain smart contract will manage the right computing tool to right data set at the right time. The analytics decision tree is based on the resulting data and condition of the results of previous computing step. The pipeline of these tools need dynamically established. Use the industry standard tools and frameworks like Tensorflow [49], Torch [50], Caffe [51], and Keras [52] to train data into models. Blockchain new distributed and parallel computing paradigm can bridge off-chain distributed parallel computing engine as well as data sets and on-chain control and pipeline management. Research and development on distributed federated learning and distributed transfer learning mechanisms can then be used to train the models. **Query Services** accept data query and analytics request. It supports NLP for natural language query. The main technical challenge is to invent innovative algorithms to convert the query request into optimized query vector, and map the query vector into smart contracts.

V. Research Issues and Technical Challenges

There are full of research issues and difficult challenges in fulfilling the vision of transforming blockchain into distributed parallel computing architecture for precision medicine described in this paper. For starter, this is a new area where leads to forward-looking research and development. Some of the research issues and technical challenges are listed below.

- Explore distributed federated learning and distributed transfer learning mechanisms within the blockchain distributed parallel computing paradigm framework to particularly apply for the precision medicine.
- Explore the use of blockchain smart contracts for medical data sharing mechanisms that can be standardized, transparent, scrutinized, and docked directly to AI analytics.
- Explore innovative algorithms to convert the query request into optimized query vector, and map the query vector into smart contracts to meet the user query demand.
- Explore blockchain distributed data management mechanisms to integrate data sets owned and hosted in all legacy systems, hospitals and service providers.
- Explore to achieve real time HIE and real world evidence next generation of clinical trial toward personal and precision medicine.
- Explore mechanisms to integrate various legacy EMR formats.
- Explore the optimized data query vector for a given research target and query request.
- Explore alliance AI technologies within the new blockchain distributed parallel architecture to dramatically boost the combined impact.

VI. Summary

A vision and mechanisms to transform original blockchain duplicated computing into distributed parallel computing architecture are proposed and briefly investigated. This results in a new distributed parallel computing paradigm which features data as the first class of citizen with moving computing to data strategy. Distributed data management, distributed data sharing, and distributed computing analytics and learning are the core mechanisms in the new architecture. Various benefits, approaches and use cases of this architecture are described. For examples, to build a large distributed medical data sets from various distributed data sources owned and hosted by various individual patients, hospitals and service providers to jump start the deep learning researches for medical domain; to discuss medical data sharing mechanisms that can be standardized, transparent, audited, and docked directly to AI analytics; and to fulfill the USA FDA vision of establishing a next generation of clinical trial with real world evidence toward personal and precision medicine. Research issues and technical challenges are also provided.

Acknowledgment

The authors thank all international researchers participating in this project for their collaborations. Particularly, Chi-Ren Shyu of University of Missouri, and Chung Hsu of China Medical University for their insight discussions.

References

[1] Satoshi Nakamoto , ”Bitcoin: A Peer-to-Peer Electronic Cash System”, https://bitcoin.org/bitcoin.pdf

[2] Vitalik Buterin, “Ethereum Whitepaper” https://cdn.relayto.com/media/files/QFTNsr6YSYCE9zyamwis_EthereumWhitePaper.pdf

[3] “Hyperledger white paper”, https://docs.google.com/document/d/1Z4M_qwILLRehPbVRUsJ3OF8Iir-gqS-ZYe7W-LE9gnE/edit#heading=h.m6iml6hqrnm2

[4] Croman, Kyle; Eyal, Ittay, “On Scaling Decentralized Blockchains”, International Conference on Financial Cryptography and Data Security, FC 2016

[5] Digiconomist, “Bitcoin Energy Consumption Index”, https://digiconomist.net/bitcoin-energy-consumption

[6] bitcoinwiki, “Proof of Stake”, https://en.bitcoin.it/wiki/Proof_of_Stake

[7] Joseph Poon, Thaddeus Dryja, “The Bitcoin Lightning Network: Scalable Off-Chain Instant Payments”, https://lightning.network/lightning-network-paper.pdf

[8] Mustafa Al-Bassam, etc., “Chainspace: A Sharded Smart Contracts Platform”, https://arxiv.org/pdf/1708.03778.pdf

[9] Enterprise Ethereum Alliance, https://en.wikipedia.org/wiki/Ethereum

[10] Petr Vlasek, “Tutorial: Hyperledger Fabric v1.1 – Create a Development Business Network on zinux”, https://github.com/CATechnologies/blockchain-tutorials/wiki/Tutorial:-Hyperledger-Fabric-v1.1-%E2%80%93-Create-a-Development-Business-Network-on-zLinux

[11] David Cerezo Sánchez, “Raziel: Private and Verifiable Smart Contracts on Blockchains”, https://eprint.iacr.org/2017/878.pdf

[12] FoldingCoin, http://foldingcoin.net/

[13] GridCoin wiki, http://wiki.gridcoin.us/Proof-of-Research

[14] SETI@home wiki, https://setiathome.berkeley.edu/

[15] Arvind Narayanan, Joseph Bonneau, Edward Felten, Andrew Miller, Steven Goldfeder, “Bitcoin and Cryptocurrency Technologies”, Princeton University Press in 2016

[16] Vitalik Buterin, “Ethereum: Platform Review, Opportunities and Challenges for Private and Consortium Blockchains”, https://www.scribd.com/doc/314477721/Ethereum-Platform-Review-Opportunities-and-Challenges-for-Private-and-Consortium-Blockchains

[17] Fan Zhang, Ethan Cecchetti, Kyle Croman, “Town Crier: An Authenticated Data Feed for Smart Contracts”, ACM CCS'16, https://eprint.iacr.org/2016/168.pdf

[18] Zonyin shae, Jeffrey J.P., Tsai,” On the Design of a Blockchain Platform for Clinical Trial and Precision Medicine”, ICDCS 2017

[19] Gil Press, “Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says”, https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#70c4f4956f63

[20] Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson, “How transferable are features in deep neural networks? “, https://arxiv.org/pdf/1411.1792.pdf

[21] ImageNet, http://www.image-net.org

[22] Stanford CS class http://cs231n.stanford.edu/. “Convolutional Neural Networks for Visual Recognition”, http://cs231n.github.io/convolutional-networks/

[23] H. Brendan McMahan, etc, “Communication-Efficient Learning of Deep Networks from Decentralized Data”, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) (2017)

[24] Jill Wechsler, “FDA Moves to Broaden Acceptance of Real-World Evidence in Clinical Research”, Applied Clinical Trials, http://www.appliedclinicaltrialsonline.com/fda-moves-broaden-acceptance-real-world-evidence-clinical-research

[25] “THE PRECISION MEDICINE INITIATIVE”, HTTPS://OBAMAWHITEHOUSE.ARCHIVES.GOV/NODE/333101

[26] N. J. Schork, "Personalized medicine: time for one-person trials," Nature, vol. 520, pp. 609611, 2015.

[27] USA FDA, https://www.fda.gov/ScienceResearch/SpecialTopics/RealWorldEvidence/default.htm

[28] Hortonworks Hadoop computing tutorial, “Hadoop Tutorial: getting started with HDP”, http://hortonworks.com/hadoop-tutorial/hello-world-an-introduction-to-hadoop-hcatalog-hive-and-pig/"

[29] Michael Creel, William Goffe, “Multi-core, Clusters, and Grid Computing: a Tutorial”, Computational Economics, Vol.32, No 4, 353-382, January 2008.

[30] Peter Mell, Timothy Grance, “The NIST Definition of Cloud Computing”, Sep. 2011, http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

[31] Moustafa AbdelBaky, Manish Parashar, Hyunjoo Kim, Kirk E. Jordan, Vipin Sachdeva, James Sexton, Hani Jamjoom, and Zon-Yin Shae, Gergina Pencheva, Reza Tavakoli, and Mary F. Wheeler, “Enabling High Performance Computing as a Service”, IEEE Computer, Oct. 2012.

[32] Karen Simonyan ∗ & Andrew Zisserman, “VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION”, ICLR 2015

[33] Karen Simonyan Andrew Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos”, https://arxiv.org/abs/1406.2199,

[34] Yonghui Wu, et. al.,“Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”, https://arxiv.org/pdf/1609.08144

[35] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, Advances in Neural Information Processing Systems 25 (NIPS 2012)

[36] Visual Geometry Group, http://www.robots.ox.ac.uk/~vgg/research/very_deep/

[37] Min Lin1, Qiang Chen, Shuicheng Yan, “Network in Network”, https://arxiv.org/abs/1312.4400

[38] Christian Szegedy, et.al., “Going Deeper with Convolutions”, CVPR2015

[39] Kaiming He, ey. Al., “Deep Residual Learning for Image Recognition”, CVPR 2016

[40] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database”, CVPR2009

[41] The CArdiovasCulAr Diabetes & Ethanol (CASCADE) Trial. Tabular View ClinicalTrials.gov. 2016. https://clinicaltrials.gov/ct2/show/record/NCT00784433?term=NCT00784433&rank=1

[42] COMPare - Full results. 2016. http://compare-trials.org/results

[43] http://www.thepaper.cn/newsDetail_forward_1531175

[44] https://www.healthit.gov/HIE

[45] The office of National Coordinator for Health Information technology, “Report on health information blocking,” U.S. Department of HHS, Tech. Report, 2015.

[46] “Blockchain for Clinical Trial”, https://blockchain.ieee.org/

[47] “President Obama state of union address 2015”, https://www.youtube.com/watch?v=cse5cCGuHmE

[48] Greg Irving, John Holden, ˝How blockchain-timestamped protocols could improve the trustworthiness of medical science”, F1000research.8114.2 https://f1000research.com/articles/5-222/v2

[49] Martín Abadi , et, al,. “TensorFlow: A System for Large-Scale Machine Learning”, Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI ’16). November 2–4, 2016, Savannah, GA, USA

[50] Ronan Collobert, Laurens van der Maaten, Armand Joulin, “Torchnet: An Open-Source Platform for (Deep) Learning Research”, Proceedings of the 33 rd International Conference on Machine Learning, New York, NY, USA, 2016.

[51] Yangqing Jia, et. al., “Caffe: Convolutional Architecture for Fast Feature Embedding”, Proceedings of the 22^{nd} ACM international conference on Multimedia, 2014

[52] Francois Cholley, “Deep Learning with Python”, ISBN 13: 9781617294433

[53] TCGA, https://cancergenome.nih.gov/