



## 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

**Thesis for the Degree of Master of Engineering**

**A User-Defined Privacy System with  
Collaborative Filtering for a Precision  
Medicine**



by

Seolah Jang

Department of Artificial Intelligence Convergence

The Graduate School

Pukyong National University

February, 2022

A User-Defined Privacy System with  
Collaborative Filtering for a Precision Medicine  
(정밀의료를 위한 협업필터링기반의 사용자  
정의 프라이버시 시스템)

Advisor: Prof. Kyung-Hyune Rhee

by

Seolah Jang

A thesis submitted in partial fulfillment of the requirements  
for the degree of

Master of Engineering

in Department of Artificial Intelligence,  
The Graduate School,  
Pukyong National University

February 2022

A User-Defined Privacy System with Collaborative Filtering for  
a Precision Medicine

A dissertation

by

Seolah Jang


Approved by:



(Chairman) *Chang Soo Kim*



(Member) *Bong-Kee Sin*



(Member) *Kyung-Hyune Rhee*

February 25, 2022

## Contents

<b>List of Figures</b> .....	<b>ii</b>
<b>List of Tables</b> .....	<b>iii</b>
<b>Abstract</b> .....	<b>iv</b>
<b>I. Introduction</b> .....	<b>1</b>
1.1. Motivation .....	3
1.2. Overview and Contributions .....	4
<b>II. Preliminary</b> .....	<b>6</b>
2.1. Precision Medicine .....	6
2.2. Decentralized E-Health System .....	8
2.2.1. Blockchain in Healthcare .....	11
2.2.2. Smart Contract .....	13
2.3. Collaborative Filtering .....	14
2.4. Nudge theory .....	17
<b>III. Proposed System</b> .....	<b>20</b>
3.1. System Overview .....	20
3.2. System Architecture .....	22
3.3. Proposed System Transaction .....	26
3.3.1. Participate Agreement .....	27
3.3.2. Transaction Between Hospital and Patient .....	28
3.3.3. Collaborative Filtering and Nudge theory .....	29
3.3.4. Transaction Between Hospital and Researcher .....	32
3.3.5. Incentive Transaction .....	33
<b>IV. Implementation and Evaluation</b> .....	<b>35</b>
4.1. Environment Setup .....	35
4.2. Collaborative Filtering for Classifying Privacy Level .....	37
4.3. Deploying Smart Contract .....	42
4.4. Inter Planetary File System .....	43
4.5. On-Chain Data .....	46
<b>V. Conclusion</b> .....	<b>49</b>
<b>References</b> .....	<b>50</b>

## List of Figures

Figure 1. Precision medicine as a new therapy system .....	8
Figure 2. Blockchain structure .....	11
Figure 3. Smart Contract structure .....	13
Figure 4. Amazon.uk using default strategy on delivery options .....	19
Figure 5. Medical data classification with CF and Nudge theory .....	23
Figure 6. Data sharing on blockchain network .....	23
Figure 7. Hospital – patient flow modeling for transaction update .....	27
Figure 8. Hospital – researcher flow modeling for transaction update .....	32
Figure 9. Matrix factorization .....	38
Figure 10. fit function in SVD.py .....	39
Figure 11. gradient function in SVD.py .....	40
Figure 12. SVD matrix factorization results .....	41
Figure 13 Deploying the contract .....	42
Figure 14 Inter Planetary File System .....	43
Figure 15 Example of the EMR.txt data .....	44
Figure 16 Stored in the decentralized web .....	44
Figure 17 Uploading flow for a new content .....	45
Figure 18 Transaction of hospital .....	46
Figure 19 Completed transaction of hospital .....	46
Figure 20. Hospital – requester gas consumption .....	47
Figure 21 Hospital cost (ETH) .....	48

## List of Tables

Table 1. Benchmarks of the conventional and decentralized healthcare environments .....	9
Table 2. The attributes of blockchain in several healthcare use cases .....	10
Table 3. Prominent blockchain consensus after the appearance of PoW in Bitcoin .....	12
Table 4. Tool of Nudge strategy .....	18
Table 5. Common healthcare data classification .....	25



# 정밀의료를 위한 협업필터링기반의 사용자 정의 프라이버시 시스템

## 장설아

부경대학교 대학원 인공지능융합학과

## 요 약

궁극적인 질병의 근원(root of disease)과 정확한 치료법을 개발하기 위해 최신 기술과 축적된 의료기록을 활용하는 정밀 의료의 가장 기본적인 개념은 최근 가속화되고 있는 유전자 분석기술의 발전과 IT기술의 진보와 함께 현실화되고 있다. 이와 함께 2020년 COVID-19로 팬데믹이 시작되면서 바이러스로 인해 누군가는 합병증으로 사망에 이르렀다면 또 다른 누군가는 무증상으로 회복되기도 하여 개인 유전체에 대한 연구는 전 세계적으로 주목받기 시작했다. 이에 ‘정밀의료(Precision Medicine)는 사람마다 동일한 병에 걸렸더라도 개인의 유전적 특성에 따라 양상이 다르고, 그에 따라서 치료법에 대한 부작용이 달라지는 것에 주목하여 연구되었다. 하지만 정밀의료 실현을 위해 대규모의 코호트를 구축하면서 개인 정보 동의의 문제와 데이터 접근과 사용 권한 부여에 대한 논의가 지속적으로 제기되어 왔다. 많은 연구자들이 의료 데이터의 투명성과 신뢰성을 위해 정밀의료의 인프라 기술로 블록체인 기술을 제안하였으나 의료데이터 거래 환경에서 병원과 병원간의 데이터 거래만 이루어져 데이터 주체자인 환자의 동의와 주권은 고려되지 않았고 규제들로 인해 데이터의 연계와 공유가 활성화되지 못했다. 본 논문에서는 데이터의 주체자에게 데이터 소유권을 보장함과 동시에 협업필터링과 넷지이론을 활용하여 데이터 주체자가 자신의 데이터가 활용되는 범위와 기대되는 이익 등을 인지하기 쉽게 하고, 이를 통해 데이터의 프라이버시 레벨에 따라 공개 항목을 개인이 설정할 수 있는 모델을 제안한다. 본 모델은 기존의 정밀의료 생태계에서 제외되었던 정보주체자의 주권과 동의과정을 보다 고도화 시키고 장기적인 정밀의료 생태계 구축에 자발적인 데이터 수집 참여를 유도한다. 따라서 본 논문에서는 병원, 환자(정보 주체자), 연구자, 그리고 제3기관 사이에서 데이터를 생성하고 정보주체자의 프라이버시 레벨에 따른 공개범위 설정 후 블록체인기반의 정밀의료 생태계에 기록 및 공유되는 방안을 제안한다.



# I. Introduction

With the start of the COVID-19 Pandemic era in 2020, the virus has become prevalent worldwide. Many victims occurred at the rapid transmission rate. However, not all infected people showed the same symptoms during the same incubation period. If someone died after medical complications, someone recovered asymptomatic. Likewise, in today's various diseases, individual differences exist in their symptoms or treatments. In other words, even if each person has the same disease, the pattern varies depending on the individual's genetic characteristics, and accordingly, the side effects of the treatment vary. Focusing on these problems, the 'Precision Medicine' technology was studied. Precision medical care is a technology that accurately diagnoses and predicts an individual's health by collecting and analyzing medical data with different values together [1]. The goal of medical technology is to provide more precise medical care by analyzing the relationship between diagnostic information on individual patients scattered widely in each medical institution and genetic information, lifestyle, environmental information, and daily activity information produced by other patients. When this technology becomes a reality, the medical paradigm shifts from many to individuals, from treatment to prevention, focusing only on treatment based on clinical trials, to dealing with the entire medical and healthcare process from prevention to diagnosis and treatment. It is expected to reduce risks such as drug side effects and medical costs, increase

treatment effects, and improve inequality in medical accessibility. However, precision medical technology is based on collecting and analyzing individual sensitive information data, and some steps should be taken to set up a system that can systematically manage and safely distribute data while also encouraging individual voluntary participation. In other words, when collecting and processing personal data, data must be collected and analyzed with the clear consent of the data subject. Already, foreign countries are establishing an environment in which data subjects take the lead in managing and agreeing to data by giving ownership of data to data subjects. In the case of the United States, The right to request copies of information subjects' health records was stipulated in the Medical Information Technology Act for Economic and Clinical Health of Electronic Health Records Utilization, beginning with the smart disclosure system in 2009. Since 2010, the "Blue Button" service has made it simple to download personal medical information in a single file format. It explains how to download data from various sources, including medication, allergies, medical information, and insurance claims. As such, by establishing an environment where data subjects can manage their data if they want, the consent form should empower data subjects, as well as general information such as data collection scope and research purpose, to protect their privacy and induce voluntary system participation.

Precision medicine, called future medical services, is accelerating research due to the development of medical big data, mobile healthcare, and various emerging technologies, such as artificial intelligence(AI), the

Internet of medical things(IoMT). Amid convergence research with various high-tech technologies, blockchain technology is attracting attention as a technology that promotes the realization of precision medical care, and research on healthcare data sharing is actively underway.

## **1.1. Motivation**

Patients' medical records have become critical components of today's healthcare environment. Many terms refer to patient medical record and data that can be accessed online or offline, such as electronic medical record (EMR), personal health record (PHR), continuity of care record (CCR), open electronic health records (openEHR), and so forth[2]. Security techniques in accessing the medical patient's data also support the significant increase in the data amount from several resources over time. In some cases, patient data is stored and scattered in different storage/services, making it more challenging to get complete access to the data. Therefore, the blockchain-based solution is used as a platform for aggregating data from diverse sources and allowing authorized parties to access it simultaneously within a single system. Thus, multiple aspects have begun to be further investigated, such as security, communication, and system effectiveness.

Existing e-health systems utilize a default data privacy policy for each patient, which is determined by the hospital's policies or the interests of healthcare stakeholders. Thus, there are various types of default privacy

policies in e-health environments, such as presented in [3], [4], and [5]. More precisely, the paper in presented the scheme in protecting consumers' privacy and personal data from e-health's most common service and online reservation services. Some of the current e-health systems do not provide full flexibility for patients to update their privacy, let alone offer new privacy updates for old patients who have been registered in the system in advance. A conventional e-health system provides privacy updates for patients. However, it is still performed manually, and information is stored in a logbook (online and offline) which can be changed by irresponsible parties, leading to disputes in the future. Therefore, e-health needs an innovative arrangement that can provide recommendations for periodic privacy updates by patients themselves without any intervention. The updated version of privacy information is then recorded on the blockchain.

## **1.2. Overview and Contribution**

In this thesis, we propose a collaborative system called UDPM (User-Defined Privacy Management) with the nudge theory concept in decentralized e-health environments to address the issues mentioned earlier. UDPM provides the best possible data privacy recommendations for patients based on accumulated data processed using the nudge theory concept. The patient fully determines the privacy management based on the recommendations given by the system. In addition, the defined data privacy management is implemented on-chain. Every update can be

inspected by authorized entities in a confidential manner. We leverage the Ethereum platform with smart contract feature as an open-source and decentralized software platform.

In summary, this thesis provides the following contributions:

- i. We construct a secure architecture that enables patient-defined data privacy management using the nudge theory concept in decentralized e-health environments.
- ii. We proffer the state of the art of our decentralized e-health system in facilitating patient-defined data privacy control.
- iii. We formulate the UDPM model and evaluate the performance based on the simulation results.
- iv. We note several requisite concerns and remarks based on our modeling of the UDPM system.

## II. Preliminary

In this chapter, we present the core system components of our proposed approach. We systematize core system components into three groups: blockchain-based smart contract (based on the Ethereum platform), collaborative filtering, the concept of nudge theory in general, and decentralized e-health, which are related studies in this thesis. All core system components are united to achieve the desired objectives that we further utilize in chapter 4.

### 2.1. Precision Medicine

Precision medicine care, called future medical care, refers to predicting the risk of disease based on genetic information and performing customized treatment for individuals, such as personally performing different types of medication [6]. If drugs for previous diseases are uniformly prescribed, precision medicine prescribes patient personalized drugs according to the type of disease, clinical results, and the patient's individual genetic characteristics. Precision medicine technology is expected to expand its scope to diagnosis and treatment of patients' diseases, as well as prediction and prevention of diseases, and healthcare. It is to redesign lifestyle and eating habits according to omics information, lifestyle habits, and individual medical history, such as the results of genetic analysis of individuals, which have become active due

to the growth of private medical companies today [7]. In other words, the scope of precision medicine refers to the entire medical service, including analysis and prediction of personal genetic information, disease diagnosis, prevention, treatment, and life management. It may also include predicting drug side effects and resistance as well as the probability of getting sick. Precision medicine research begins with building a cohort of scale[8]. The cohort used in epidemiological research refers to a group of people of a scale who share certain characteristics. It is used in a study to determine the source cause of the disease by discovering the difference between the group exposed to factors and the group not exposed through a cohort built in hospitals for a long time. However, the development of established medical Big Data technology is raising concerns about personal information protection along with the realization of precision medicine care[9]. In principle, the collection of personal information is based on the acquisition of patient consent. However, hospitals, related institutions, or researchers using personal information may collect information through consent exemption or collect information without consent on the premise that de-identified information is not personal information. Considering the characteristics of information collected in precision medicine, it is practically difficult to conduct precision medicine research through consent exemption methods, and protection of personal information de-identified in big data analysis is also necessary.



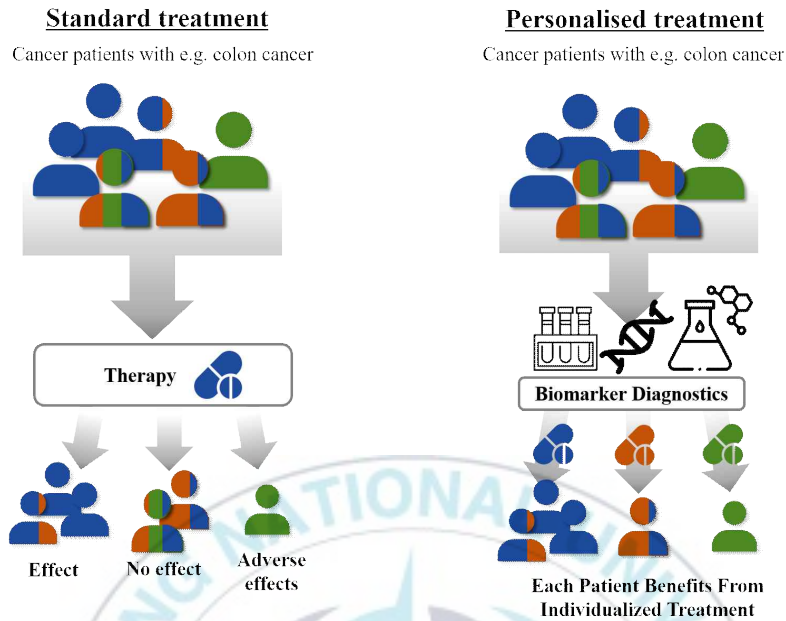


Figure 2. Precision medicine as a new therapy system[10]

## 2.2. Decentralized E-Health System

E-health refers to the services and activities associated with the systematic delivery of medical treatment to individuals or groups of people via the internet. There are many terms for e-health that refer to identical functions such as personal health information (PHI), personal health record (PHR), electronic medical record (EMR), and so on. However, all of these terms have something in common: authorized parties can access their health record via the internet provided by healthcare providers. Due to many providers involved in an e-health environment, it causes data to be fragmented and difficult to be accessed by the patient, as stated by a research paper in [11].



Table 1. Benchmarks of the conventional and decentralized healthcare environments

Benchmark	Architecture	Security
Hierarchical distributed EH R(HDEHR)	DE, P2P	N/D
m-Health	DE	N/D
Ubiquitous PHR (uPHR)	DE	N/D
Conceptual Framework(CF)	CS,DO	CIA, HIPAA
HealthVault	CS	CP-ABE
DEPR	DC	N/D
My HealthVet	DE	Security Policies
SNOW	DC	Privacy Policies

N/D: not defined; DE: distributed electronic; P2P: peer-to-peer; CS: client-server; DO: distributed object; CIA: confidentiality, integrity, and Availability; HIPAA: health insurance portability and accountability; CP-ABE: ciphertext-policy attribute based encryption; DC: distributed components.

A decentralized healthcare system that involves many providers into a single system has begun to be formed by relying on blockchain technology with smart contract features. The objectives of decentralized e-health are spawned with a powerful idea to organize and improve healthcare services worldwide. Table 1 presents the conventional and decentralized healthcare environments in single and multiple collaborated servers. Academia and industry have developed a variety of methodologies and procedures to provide healthcare providers, doctors, and patients with services that are not limited by time or location.

Table 2. The attributes of blockchain in several healthcare use cases [12]

Use cases Attributes	EHR & PHR	Insurance & Market	IoT & Health Monitoring	Supply Chain	Research & Trial
Timestamped immutable data history	✓	✓	✓	✓	✓
Autonomous contracts	✓	✓	✓		✓
Decentralized verification	✓	✓		✓	
Interoperability	✓		✓	✓	✓
Transparency	✓	✓		✓	✓
Gamification		✓	✓		
Decentralized value transfer		✓	✓		

Table 2 illustrates the use of blockchain in multiple applications with different goals. The goals are divided into decentralized value transfer, gamification, transparency, interoperability, non-centralized verification, autonomous contracts, and immutable transactions. Meanwhile, healthcare use cases are divided into three categories, namely EHR & PHR, insurance & market, and internet-of-things & monitoring. Several healthcare providers achieve the main objectives of implementing blockchain technology using different platforms and techniques [13]. The benefits of decentralized e-health such as transparency, decentralized value transfer, interoperability, and

immutable data history records are still the main goals to be attained.

### 2.2.1. Blockchain in Health

Blockchain, a fundamental technology of Bitcoin with unique features has revolutionized the paradigm of transacting on the internet[14]. Transactions no longer rely solely on an intermediary or third parties in organizing the goals the parties want to accomplish. The transactions' verification process is carried out in a decentralized manner (not centralized to one node) [15]. Every node in the same blockchain network propagates the ledgers. As a result, every node has the most up-to-date version of the ledger state.

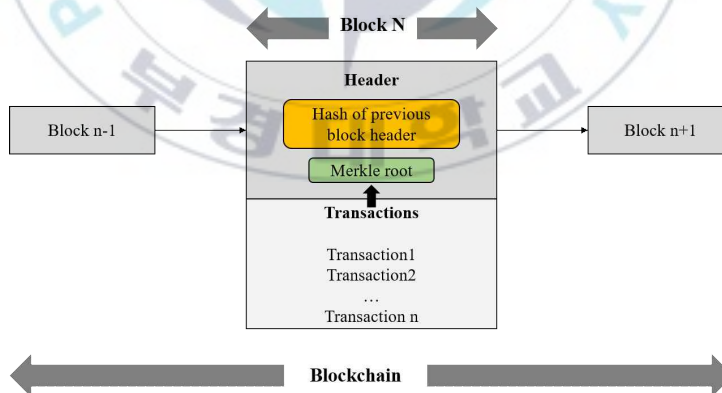


Figure 2. Blockchain structure

Table 3. Prominent blockchain consensus after the appearance of  
Proof-of-Work in Bitcoin

Benchmark	Public Access	Private Access
Distributed Validation	Proof-of-work (PoW), Proof-of-stake (PoS), PoW based derivatives, Federated Byzantines agreement	Proof-of-work (PoW), Proof-of-stake (PoW), PoW based derivatives, Federated Byzantines agreement
Concentrated validation	Delegated Proof-of-stake (DPoS)	Redundant Byzantine fault tolerance, Ripple consensus bilateral node- to-node (N2N), RAFT and derivatives, Delegated Proof-of-stake (DPoS)

Blockchain transactions are time-stamped and recorded in chronological order that the parties who are granted authority can inspect the transaction. Data that has been successfully verified and stored is tamper-proof; hence, there is no possibility of altering data, data destruction, or data deletion by malicious parties. Miners or validators validate transactions that contain data through a particular consensus mechanism (see Table 1) that eliminates the risk of twofold entry, counterfeit data, or fraud. Explicitly speaking, blockchain technology provides decentralization, immutability, security, and transparency [16].

### 2.2.2. Smart Contract

Blockchain with the smart contract feature began to be widely recognized by the public through Ethereum platforms in 2015, categorized into Blockchain 2.0 (starting from Bitcoin for Blockchain 1.0). The development has continued ever since, as evidenced in 2017 by updating the Ethereum features into Distributed Applications ICOs (categorized into Blockchain 3.0) [17]. There are various types of smart contract platforms, such as Ethereum and Hyperledger Burrow (Solidity, Serpent, Mutant, and Vipe), Hyperledger Fabric (Golang, Java, JavaScript), Quorum (Solidity), and Open Transactions (ChainScript).

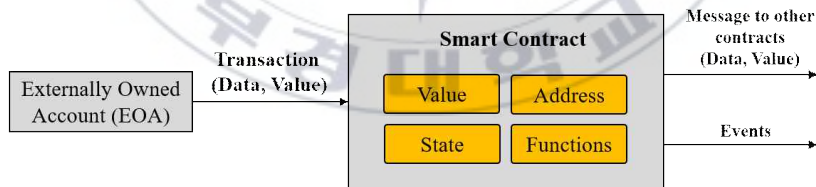


Figure 3. Smart Contract structure [18]

Smart contracts are verified in real-time, based on conditions stated in the contract. The contracts holder can execute the contract's functions without any interference, eliminating the long-winded attestation process.

Regarding the business perspective, merging tasks and automating the contract's function can assist streamline business services and boost profitability. The smart contracts-based application provides clear communication among the entities. Eventually, smart contracts by design are paperless, lower cost, faster run, and secure, making use of encryption at the blockchain level. With these merits, smart contracts are extensively adopted in various businesses and applications, Therefore, blockchain-based smart contracts are also suitable to be adopted in the healthcare environment, such as e-health, personal health information (PHI) and electronic medical records (EMR).

### **2.3. Collaborative filtering**

The recommender system is a system that analyzes preference for items through user rating information or user transaction pattern, and proposes new items that users are likely to prefer [19]. Therefore, a well-established recommendation system provides user-friendly services and enables the establishment of a powerful system through the influx of new users.

The recommender system can be divided into content-based filtering(CB) and collaborative filtering(CF)[20]. The content-based filtering method is a method of recommending another item with content similar to that item when the user prefers a specific item. The

collaborative filtering is one of the representative technologies of the recommendation system, and the recommendation target is determined based on the similarity between users and items. The CF analyzes information such as usage patterns and consumption records of recommended system users and predicts users' preferred information. The core assumption of collaborative filtering is that users with similar tastes will have similar preferences for certain items. In [21], analyzes the tendency to disclose their financial information according to the age, occupation, and final education of the system users, leading to a wider range of data to be recorded in the blockchain environment.

Basically, the collaborative filtering is largely divided into memory-based approaches and latent factor approaches.

#### (1) Neighborhood based method (Memory based)

It's the most traditional approach. The similarity between users/items can be stored in memory. When recommendations are needed for a specific user, items consumed by k users similar to that can be recommended, or k items identical to that item can be estimated based on rating[22].

#### (2) Latent Factor Collaborative Filtering method (Matrix factorisation)

Matrix factorisation is a method of decomposing or lowering vectors representing user or product information into algorithms like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). In other words, matrix factorisation aims to extract the latent for the user



and the latent for the product by decomposing it into two matrices when there is a matrix that rows the user and columns the evaluation of the product. Latent is a vector that represents each user's characteristics and is created using the method and number that the machine understands. It can be used to recommend similar users or products using distance between latent vectors [23].

One of the collaborative filtering technologies, Singular Value Deposition (SVD), employs a dimensional reduction technique recommendation system. It is one of the "matrix decomposition" methods, in which a matrix of any dimension can be decomposed as follows.

$$A = U\Sigma V^T \quad (1)$$

Each matrix has the following properties.

- $U = User \times latent factor$ , An unitary matrix with size  $m \times m$
- $V^T$  is  $item \times latent factor$ . conjugate transpose of  $V$ , an unitary matrix with size  $n \times n$
- $\Sigma$  is  $latent factor$ , the values of the elements on the rectangular diagonal matrix are not negative with all the values of the remaining elements being zero.

In summary,  $U$  and  $V$  are matrices representing singular vectors and  $\Sigma$  is matrices for singular values. By specifying the size of  $\Sigma$ , it is also possible to specify the size of the grant vector. Subsequently, to represent a matrix that infers empty matrix spaces, It uses the



decomposed matrices to create a matrix of the same size as the original matrix  $A$ . In other words, the completed matrix  $A'$  is created by learning unknown  $U$ ,  $V$  and  $\Sigma$  using the data rated in the matrix  $A$ .

As mentioned above, memory-based filtering uses similarity coefficients of similar users or items to fill empty spaces in the matrix. It is explanatory and easy to apply recommendations. Still, it requires a lot of operations to forge recommendation results and makes it difficult to respond to the scale and low sensitivity to data scarcity [24]. Therefore, in this thesis, a model-based recommendation system is applied to apply a scalable recommendation system considering accurate privacy policy recommendations and efficient computation to system users.

## 2.4. Nudge theory

Nudge theory is a concept within the field of "behavioural economics" [25]. In traditional economics, it has often been found that it is not reasonable to make individual's everyday choices in theory, and behavioural economics was born when economists actively accepted the findings of psychology to understand them in the framework of economics. This concept is not economics based on the premise of rational and rational economic human beings, but economics to study practical human behaviour in determining the causes and consequences of decision-making of choice behaviour.

Recently, several Western countries have made various attempts to

incorporate ideas and insights from behavioural economics, including nudge, into public policy. At a time when economic incentives, a practical approach to health care policy, do not achieve much, behavioural economics, which points to errors in "economic humans" introduced by mainstream economics and explains human behaviour more plausible based on a solid study of psychology, seems to be an attractive approach.

Table 4. Tool of Nudge strategy [26]

Regulation	Financial			Non-Financial			
Non	Encourage and provide options.						
	Incentive and information			Nudge			
Laws and regulations.	Financial Incentive	Non-Financial incentive	Provide information	Simplifying information and framing.	Changes in the physical environment	Changes in default policy.	The use of social norms.

Nudge strategies can be classified as financial and non-financial means of inducing change by providing financial means such as incentives. The classification is given in Table3. The financial instruments of the nudge strategy include information simplification, changes to the physical environment, default policy, and social norm. The default option to be used in this thesis is a strategy to induce changes in the behaviour of policy subjects by setting the options that policy designers think are desirable as default. Changing the behaviour of policy subjects with changes in default options can be observed in many nudge policies. This default option follows the theory that "human beings usually tend to choose the first option". In Figure 4, *Amazon.co.uk* also applied a default strategy to set the First Class option, which requires a certain fee rather

than free shipping in delivery options. In other words, *Amazon.co.uk* compelling the desired consumer behavior by applying default options.



Figure 4. *Amazon.uk* using default strategy on delivery options

As mentioned earlier, various policies incorporating nudge based on behavioral insights have different forms, but their basic characteristics can be understood through the concept of “liberal paternalism” presented by Sailor and Thaler & Sunstein in 2008. liberal paternalism can be said to be “paternalism in terms of interfering (inducing or intervening) individuals to make the right choice. However, it provides individuals with options of choice and does not force specific choices. In addition, in many cases, it can be seen as liberalism in that there are no restrictions on changing choices.” Likewise, policies incorporating nudge have the characteristics of this liberal paternalism.

## III. Proposed System

In this chapter, we describe the design of the proposed system, which enables patients-defined data privacy management in a decentralized e-health environment and the data flow between the main participants of the system. This section is divided into three sections. First, we describe assumption of the proposed model. Secondly, we put forward the state of the art that presents UDPM at a high level. Finally, the thesis elaborates data privacy classification management, and nudging with collaborative filtering model is also given afterwards.

### 3.1. System Overview

UDPM is a framework for achieving patient-defined data privacy in decentralized e-health environments using the nudge theory concept. Our approach focuses on managing the patient's data privacy, classification and utilizing the concept of nudge theory to obtain better services and recommendations for the patient inspired by research in [21]. However, the final decision remains entirely up to the patient to organize and determine his medical data accessible by certain groups. The nudging results' output data is then stored on-chain via an Ethereum smart contract, allowing authorized parties to access the patient's historical data privacy records. Meanwhile, UDPM leverages end-to-end encryption techniques for communication, where only the communicating e-health

entities can read the messages.

Figure 6 illustrates the state of the art of the UDPM framework in enabling patient-defined data privacy. The UDPM system commences by storing encrypted data by patients or other authorized entities that use cloud-based storage services, which is IPFS. By design, the data are collected by authorized parties whose access is governed by the patient or the personal doctor representing the data owner. To deploy the UDPM framework, the authorized doctors or stakeholders roughly classify the patient's data privacy by default, such as lifelong information, medical activities, insurance, and so forth.

The initial classification of patient data makes UDPM more manageable and convenient to use in a collaborative filtering-based algorithm so that the algorithm's output can be more precise. Collaborative filtering protocol can be used in the UDPM framework as a filtering method used by the recommender system to make automatic predictions about patient privacy management and preferences from any inputs of data. Several inputs of data are derived from the different number of patients with their respective data that has been processed beforehand. The underlying premise of the collaborative filtering in UDPM is that if patient X has the same interest as patient Y on a use case, patient X is more likely agreeing with Y's selection on a different use case than that of a randomly chosen patient. Finally, the nudge theory's concept makes the collaborative filtering output more likely that patients will perform a particular choice or behave distinctly by adjusting the conditions so that automatic cognitive rules are triggered to favour

the craved outcome.

### 3.2. System Architecture

As shown in Figure 5 and Figure 6, the model proposed in this thesis consists of 1) patients as data subjects, 2) hospitals (doctor) as medical data generators, 3) IPFS nodes as data distributed storage, and 4) researchers and other hospitals requesting data.

- **Patient:** In proposed system, the patient provides the hospital with data generated by the hospital or data generated by wearable devices. In this case, the patient may control the disclosure (privacy level) of data according to the privacy policy defined by the patient.
- **Hospital:** The hospital receives the patient's medical data generated during the treatment process and the patient's life log data and stores them in IPFS storage. In this case, even the same medical data may be classified and stored according to the privacy policy defined by the patient. In addition, the stored data can be shared with other institutions other than the hospital for research purposes.
- **IPFS node:** All hospitals participating in the system configure a private IPFS network to store and share data.
- **Researcher (data requestor):** Researchers can request and share medical data of patients collected and stored in hospitals for research purposes. At this time, the entire process of data sharing, including access authority authentication, is performed on the IPFS network.

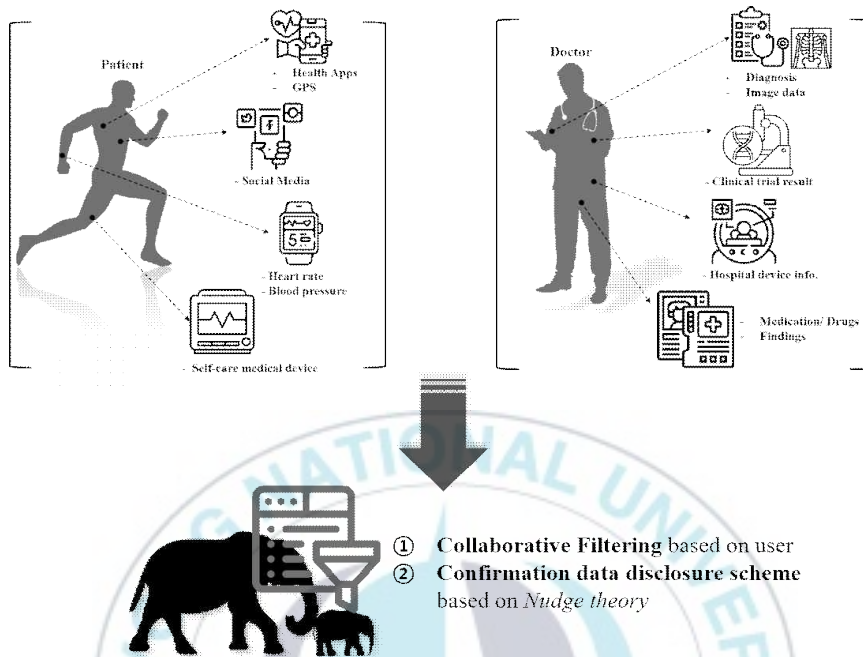


Figure 5. Medical data classification with CF and Nudge theory

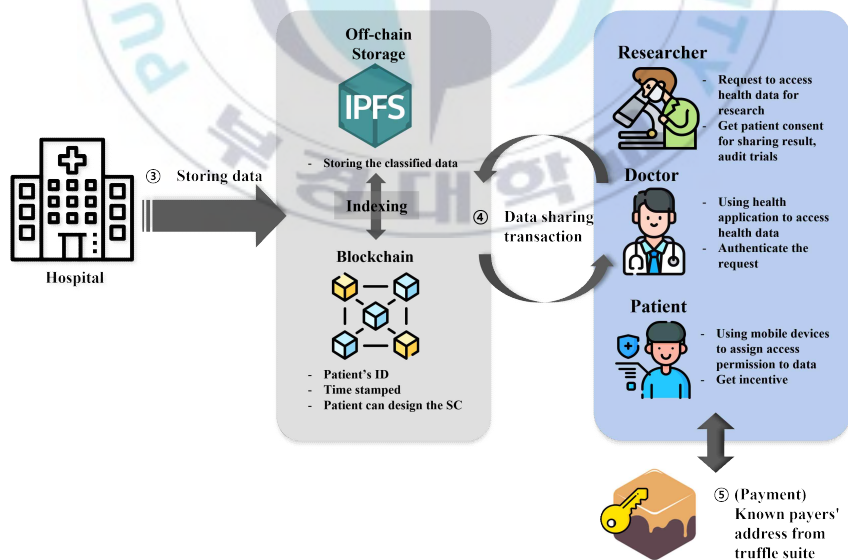


Figure 6. Data sharing on blockchain network



According to GDPR and HIPPA, patients' medical sensitive data should not be used identifiable in the sharing process. Therefore, in this thesis, it is assumed that there is a separate process of de-identifying shared medical data before sharing. In addition, reliable medical institutions are assumed to be IPFS nodes in this thesis because nodes in general IPFS networks are not suitable for storing medical data, which is sensitive information of individuals. This can play a role in preventing patients from directly participating in the system as IPFS nodes and spreading unreliable data into the system.

In the proposal system, a policy recommendation system based on collaborative filtering was used to help patients define personal privacy policies. Collaborative filtering learns the selection result of a user group with a tendency similar to the user's tendency, predicts appropriate selection, and recommends it to the user. At this time, the collaborative filtering model requires user propensity data (user's age, educational background, income level, etc.) to use the recommendation system as input. In this thesis, users are required to data and use these tendencies through surveys before participating in the system for the first time. In addition, data used for learning a collaborative filtering model are assumed to be user selection results defining privacy policies without using a recommendation system.

Medical data can be classified and organized into several types according to the identifiability of information contained in the data. The classified data is selectively stored in the blockchain according to the



privacy policy defined by the user. In this thesis, data classification is defined as shown in Table 5.

Table 5. Common healthcare data classification

<b>MSPI</b> (Most Stringent Protection Information)	<b>PHI</b> (Personal Health Identifier )	<b>HSI</b> (Hospital Sensitive Information)	<b>II</b> (Important information)	<b>PI</b> (Public information)
<u>Omics data</u>	<u>Identifying information</u>	<u>Image data information</u>	<u>Medical treatment information</u>	<u>Social Media health information</u>
<u>Clinical trial information</u>	<u>Attribute value information</u>	<u>Hospital device information</u>	<u>Patient medical information</u>	<u>Lifelog data</u>
		<u>Image data related with Clinical trial information</u>		
		<u>Hospital employee information</u>		

In [27] classified the medical data based on the similarity, usage, and purpose like ‘User Credentials’, ‘Participant characteristics’, ‘Medical Data’, ‘Project/Internal Data’ and ‘Meta Data’. In addition, sensitivity levels presented in this paper were classified into three categories ( Low, Moderate, High) based on the possibility that the data subject could be identified. For example, although the information of the data subject is not included directly, the sensitivity was set to ‘Moderate’ if the data subject could be inferred. As such, there is no discipline that accurately defines the sensitivity of current medical data, and sensitivity can be defined by classification of identifiable. However, there are limitations in separating and managing vast amounts of medical data at the three levels presented in the paper above. In this thesis, the sensitivity of medical data, including the possibility of identifying patients and equipment information of hospitals, was classified into five categories.

### 3.3. Proposed System Transaction

This section describes the detailed protocol of the UDPM system proposed in this thesis. The protocol is consist of 1) Participate Agreement, 2) Between Hospital and Patient transaction, 3) Collaborative filtering and Nudge theory 4) Between Hospital and Researcher Transaction and 5) Incentive Transaction. All protocols are divided into Figure 7 and Figure 8 based on the primary transaction entity.

Notion	
Agreement	$A_g$
Smart Contract	$SC_x$
Hospital	$H_{px}$
Patient	$U_x$
Patient's data	$U_{xdata}$
Researcher	$RE_x$
IPFS storage	$IPFS_{stor}$
Private ETH	$ETH_{priv}$
Privacy Level	$P_{lv}$
Collaborative Filtering	$Col_{fx}$

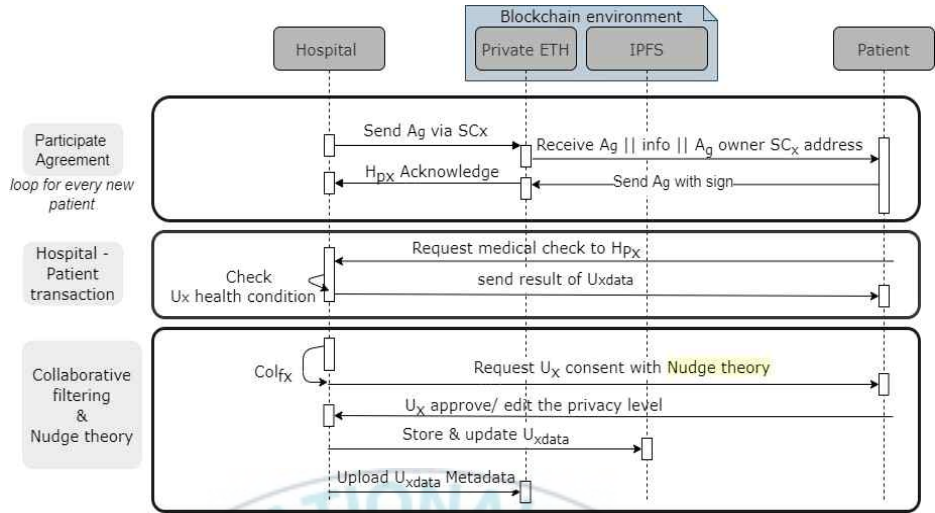


Figure 7. Hospital – patient flow modeling for transaction update

In figure 7 illustrates the overview of the system transactions between the patient and the hospital, consist of participate agreement protocol, Diagnosis result protocol between hospital and patient and collaborative filtering and nudge protocol.

### 3.3.1. Participate Agreement

The first Algorithm 1 is for registering a new peer, it can only be executed by the system manager (Hospital). It is embedded by ETH\_Pubkey, which is used for authentication every time the user invokes the function within the contract.

---

**Algorithm 1: Registering a New Peer**

---

**Procedure CONTRACT OWNER PERFORMS:**

**Input :** ETH\_Pubkey 128 Chars / 64 bytes

**Output:** GUI\_notifications

**If** *execute.contract* **is not the contract owner** **then**  
    *throw*; **end**

**else**

**If** *addNewUser* **exist** **then**

**return** false;

**send** notifications to peer; **end**

**else**

        [addNewPeer.ETH\_Pubkey]

        New\_Peer\_+1 = true;

**return** true;

**End**

---

### 3.3.2. Transaction Between Hospital and Patient

In this transaction, a total of three transactions will occur. (1) It is divided into transactions in which patients receive medical treatment from hospitals, (2) transactions in which hospitals provide medical services to patients with equipment from various hospitals, and (3) transactions in which patients provide data.

The patient visits the hospital and receives medical services from the doctor. The doctor provides the patient with the treatment results with medical image data or video data, at which time the patient's data is recorded in the UDPM system, explaining the benefits the patient can receive. If the patient approves the participation in the system, the hospital recommends the scope of data disclosure of other users with

similar tendencies to the patient through collaborative filtering and recommends some information to the patient in public settings according to the system's policy.

### **3.3.3. Collaborative Filtering and Nudge theory**

Healthcare data over the internet is unique and challenging to measure. Combining e-health data from multiple sources and the interpretation of that information is essential to optimize patient care. A decent e-health data improves communication between healthcare entities with an in-depth understanding of particular health conditions, insurance, long-term planning, and so forth. Therefore, determining the correct data to be filtered before being processed using nudge theory needs to be considered, such as applying a collaborative filtering protocol.

---

**Algorithm 2: User Defined Privacy Model**

---

```
Procedure  $\forall$  PEERS PERFORMS:  
Define list of entities:  
    List  $\rightarrow$  Hospital(Hpx), Blockchain(BCE), contd ...  
        Patient(Ux), Researcher(REx);  
for participate_agreement do  
    Hpx send Ag via SCx;       //Ag: agreement  
    BCE receive Ag||info||Ag owner SCx addrs.;  
    Ux send Ag.signed;  
    end for  
for Hospital_Patient_Activities do  
    Activities  $\rightarrow$  Private ETH (SCx)  
    Ux request check to Hpx ;  
    Hpx send result.Uxdata;  
    Deliver Colfx to Ux  
    //Colfx: Collaborative Filtering procedure)  
    (Defined_Privacy_Level)  
    end for  
End
```

---

Collaborative filtering protocol in the UDPM system can be defined into two steps. Firstly, the UDPM searching for patients who participate in the same rating patterns with the active patients (the patient whom the prediction is for). Secondly, the UDPM applying ratings from those like-minded patients observed in the earlier step to compute a forecast for the active patient. The model-based types of collaborative learning rely on patient rating data to execute patient information similarity. Representative models of this method are model-based CF with patient-based top-N recommendations. For instance, in patient-based procedures, the value of rating patient  $u$  gives to item  $i$  is determined as an aggregation of some similar patient's rating of the item is  $r_{u,i} = \text{aggr}_{u'} \in U r_{u',i}$ ; where  $U$  depicts the set of top N patients that are most similar to patient  $u$  who rated item  $i$  [28]. Remarkable models of the aggregation function include:

$$r_{u,i} = \frac{1}{N} \sum_{u' \in U} r_{u',i} \quad (1)$$

$$r_{u,i} = k \frac{1}{N} \sum_{u' \in U} \text{simil}(u, u') r_{u',i} \quad (2)$$

Where  $k$  is a normalizing factor defined as  $k = 1 / \sum_{u' \in U} \text{simil}(u, u')$ . The model-based algorithm computes the relationship between two patients or items while producing a forecast for the patient by considering the weighted average of all the ratings. In this approach, models are developed using different data mining, machine learning algorithms to predict patients' rating of the unrated medical data item. There are many model-based CF algorithms. Bayesian networks, clustering models, latent semantic models such as singular value decomposition, probabilistic latent semantic analysis, multiple multiplicative factor, latent dirichlet allocation, and Markov decision process based models. One benefit of applying this method for the UDPM system is that possessing a high dimensional matrix consisting of abundant missing values. Futuremore, we can develop CF for extended the UDPM system. Likeness computation between items or patients is essential in the UDPM system. Various models, such as Pearson correlation and vector cosine-based similarity, can be adapted to achieve a better result.



### 3.3.4. Transaction Between Hospital and Researcher

In this transaction, the researcher participates in this system and asks the hospital to access the data. In figure 8 illustrates the overview of the system transactions between the hospital and the researcher,

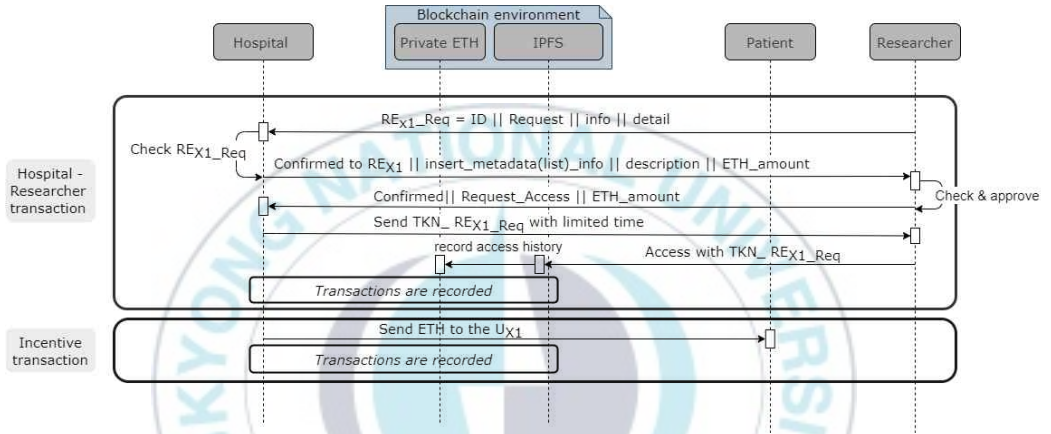


Figure 8. Hospital – researcher flow modeling for transaction update

The last algorithm as shown in Algorithm 3 is for the hospital and the researcher data request which have been uploaded onto the IPFS and the Blockchain.

---

#### Algorithm 3: BCE Transactions (Hpx and REx)

---

**Procedure Hpx and REx on BCE PERFORMS:**

**Define:**

TKN\_REx1\_req.: Token with limited time

**for** Hpx – REx transactions **do**

**Activities** → **Private ETH (SCx)**

REx\_request = ID||Req.||info||detail;

Hpx\_confirm REx\_request;

**If** REx\_request = Correct; **then**

---



---

```
    Send TKN_REx1 to REx1;  
    Record access history;  
else  
    throw;  
Send reward for data provider;  
Records transactions on-off chain;  
end for  
End
```

---

Researchers ask the hospital for information on their researcher ID and necessary data to access the data uploaded by the hospital to IPFS. The hospital issues tokens to researchers that allow them to access data for a certain period of time after undergoing a licensed researcher recognition process.

### 3.3.5. Incentive Transaction

Blockchain with smart contract features can be deployed as a decentralized incentive mechanism since it provides fair transactions that are recorded in the distributed and immutable database system. On the other hand, the incentive might be the tools to motivate patients to participate actively and share their data to improve the data quality of researchers. In this sense, the participated patients will receive rewards from the system based on their positive contributions. Thus, this incentive mechanism will give advantages to either patients or researchers in the UDPM system.

In order to realize the decentralized incentive transactions, UDPM leverage the Ethereum smart contract that records every transaction securely. Here, Ether (ETH) as the Ethereum coin is used for

participating patients in data sharing transactions. This incentive scheme will execute after all data transactions are completed. Patients will receive rewards for their data utilization and incentives from hospitals for continuously providing their lifelog data. The hospital will allocate incentives (ETH) for the patient's data received from the researcher.



## IV. Implementation and Evaluation

In this chapter, we examine the implementation methods of the model discussed in the previous chapter and evaluate the model implementation results. Here, patients define the sensitivity level of data for the scope of data disclosure, which is classified through user-based collaborative filtering and nudge policy. Subsequently, a blockchain network is established, the operation process and data sharing transactions are checked through a smart contract, as well as the assessment and IPFS upload results. In order to implement the proposed model, Ethereum platform was adopted as blockchain prototype for realizing a decentralized data sharing transaction system.

### 4.1. Environment Setup

In UDPM implementation, hospitals first upload data from patients classified through collaborative filtering to IPFS by empowering off-chain transactions. Here, we use IPFS as a distributed P2P file system that can deliver high-capacity files quickly and efficiently as well as promise a stable ecosystem. Then, only metadata of transactions will be recorded on the blockchain with an on-chain data management approach. On the other side, the researcher, which needs the particular data example to be further analyzed, requests data and calculates the gas fee for the hospital and patient as data providers. Finally, the patients who provide

data-sharing transactions will obtain a certain amount of reward from the UDPM system.

Our proposed model implementation consists of several steps, including privacy level classification using collaborative filtering, developing the distributed data management system by combining IPFS and blockchain technology, and deploying a fair incentive mechanism based on Ethereum smart contract. It is worth noting that the UDPM system is composed of different framework implementations. The detailed technical implementation and simulation setup is described as follows.

■ **Platform: Ethereum (EVM 1 Bytecode)**

- Smart Contract (Solidity)
- Pragma solidity  $\geq 0.4.21$   $< 0.6.0$ ;
- Truffle Suite (rapid development DApps)

■ **Cryptocurrency wallet & Ethereum gateway**

■ **IPFS Desktop : 0.17.0 (0.17.0)**

- UI v.2.13.0
- Running on a public gateway

■ **The decentralized medical environment is running on:**

- macOS Big Sur (ver\_11.6.1)
- Memory : 1600Mhz DDR3
- Processor Speed: 1.3 GHz
- L2 Cache (per Core): 256 KB, L3 Cache: 3 MB

■ **Basic prerequisites are installed (nodejs, npm, xcode, vscode, etc.)**

## 4.2. Collaborative Filtering for Classifying Privacy Level

The UDPM model utilizes the concept of CF for classifying the privacy level of patients' data based on Table 5. As described in Section 3.3.3, CF will find the active patients in the same rating patterns with the model-based CF method representation. On the other hand, the likeness computation between items or patients is essential in the UDPM system. Therefore, it is essential to encrypt the non-disclosure of patient information in the system. Thus, those several steps can be described as follows:

- Step 1. Input the computed disclosure schemes include the default schemes.
- Step 2. Output of step 1 is encrypted patient's information.
- Step 3. Check the computed disclosure schemes, if it equals non-disclosure, then go to step 2, otherwise go to step 4.
- Step 4. Search the nearest smart ledger of blockchain, if computed disclosure schemes are changed then go to step 3, otherwise go to
- Step 5. Using the blockchain proprietary key to encrypt the patient's information.
- Step 6. Update patient's information to the chain for consensus confirmation.

According to the data classification provided by the hospital, patients can

set sensitivity levels ranging from 1 to 5. Personal information such as the patient's name, password, and mobile phone number may be the most sensitive information at this time. Furthermore, medical information can include diagnosis information, prescription records, and clinical records, and individuals can control the level of sensitivity by classifying their data.

When two different types of entities are multiplied, matrix factorisation is used to generate latent features. It is used in CF to determine the link between item and user entities.

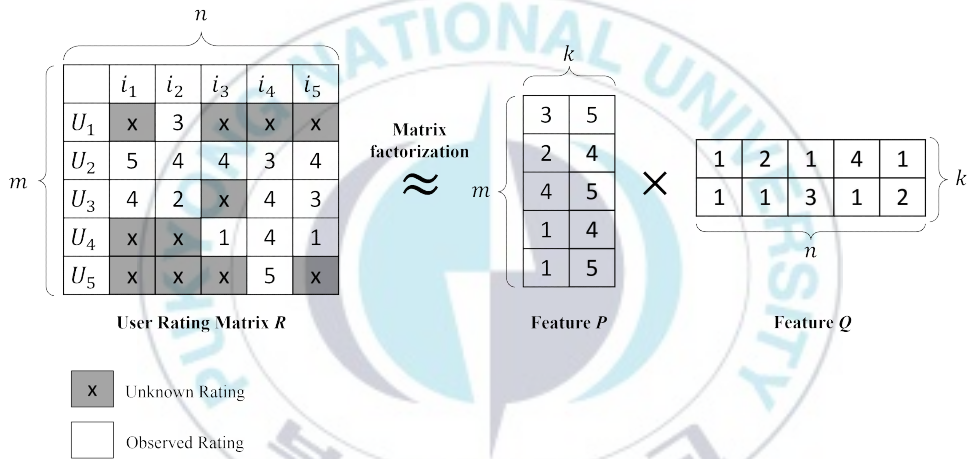


Figure 9. Matrix factorisation

Matrix factorisation is a method in which the actual users' rating matrix is divided into two main feature (user and item) matrices, with the new two matrices being multiplied to get the actual matrix. The expected matrix gets results that are comparable to the true values, and the 0 ratings in the actual matrix are replaced with predictions based on similar user preferences (see Fig. 9).

We implemented collaborative filtering in our proposed system by python.

```

25     def fit(self):
26         self._P = np.random.normal(size=(self._num_users, self._k))
27         self._Q = np.random.normal(size=(self._num_items, self._k))
28
29         self._b_P = np.zeros(self._num_users)
30         self._b_Q = np.zeros(self._num_items)
31         self._b = np.mean(self._R[np.where(self._R != 0)])
32
33         self._training_process = []
34         for epoch in range(self._epochs):
35             xi, yi = self._R.nonzero()
36             for i, j in zip(xi, yi):
37                 self.gradient_descent(i, j, self._R[i, j])
38             cost = self.cost()
39             self._training_process.append((epoch, cost))
40
41             if self._verbose == True and ((epoch + 1) % 50 == 0):
42                 print("Iteration: %d ; cost = %.4f" % (epoch + 1, cost))
43
44

```

Figure 10. fit function in SVD.py

Firstly, the input user-item rating matrix  $R$  is defined, and it can be decomposed into two latent vectors: user  $P$  and item  $Q$ . As a result, we have the following equation:

$$R \approx P * Q^T \quad (3)$$

Multiplying the decomposed matrix proceeds with the original shape  $R$ , and the question patterns are refilled with the implied values. In this context, the elements of  $P$  and  $Q$  can be considered as variables that must be discovered. Consequently, the implied value can be guided to as the predicted rating, and the formula is as follows. We aim to train  $P_{ik}$  and  $Q_{jk}$  to spit out the predicted rating.

$$R_{ij} = P_i^T Q_j = \sum_{k=1}^K P_{ik} Q_{jk} \quad (4)$$

Then, we define the above formula to transalte it into code based on [29]. We use the initial global bias `self._b` as the average value of



ratings in input  $R$ . In addition, normalization is performed to remove the negative numbers entering the final rating.

```

54 def gradient(self, error, i, j):
55
56     dp = (error * self._Q[j, :]) - (self._reg_param * self._P[i, :])
57     dq = (error * self._P[i, :]) - (self._reg_param * self._Q[j, :])
58     return dp, dq
59
60
61 def gradient_descent(self, i, j, rating):
62
63     prediction = self.get_prediction(i, j)
64     error = rating - prediction
65
66     self._b_P[i] += self._learning_rate * (error - self._reg_param * self._b_P[i])
67     self._b_Q[j] += self._learning_rate * (error - self._reg_param * self._b_Q[j])
68
69     dp, dq = self.gradient(error, i, j)
70     self._P[i, :] += self._learning_rate * dp
71     self._Q[j, :] += self._learning_rate * dq
72
73
74 def get_prediction(self, i, j):
75
76     return self._b + self._b_P[i] + self._b_Q[j] + self._P[i, :].dot(self._Q[j, :].T)
77
78
79 def get_complete_matrix(self):
80
81     return self._b + self._b_P[:, np.newaxis] + self._b_Q[np.newaxis:, ] + self._P.dot(self._Q.T)
82
83
84

```

Figure 11. gradient function in SVD.py

If the SVD is decomposed as it is, a speed problem may occur while the entire user-item matrix and the missing data may not be well processed. Therefore, we handle the problem by get the loss based on the square error of rating and execute `gradient_descent`. The `gradient_descent` is minimize the sum of square error.

$$\frac{1}{n} \sum_{i=1}^n (y_i - t_i)^2 \quad (5)$$

In other words, SVD is optimized in the form of minimizing Mean

Squared Error (MSE), equation (5). The division by  $N$  in MSE is monotonic and does not affect the direction of gradient, so it is excluded from the optimization. In this implementation, gradient descent without batch adjustment was used, and variables were initialized to completely random variables that were not corrected using Numpy. The same equation as the equation (5) was used for the normalization parameter, and global matrix bias was used for latent correction. The result can be referred to Figure 12.

```

Iteration: 50 ; cost = 0.4490
Iteration: 100 ; cost = 0.1205
Iteration: 150 ; cost = 0.0485
Iteration: 200 ; cost = 0.0259
Iteration: 250 ; cost = 0.0181
Iteration: 300 ; cost = 0.0162
Iteration: 350 ; cost = 0.0160
Iteration: 400 ; cost = 0.0163
Iteration: 450 ; cost = 0.0165
Iteration: 500 ; cost = 0.0167
Iteration: 550 ; cost = 0.0168
Iteration: 600 ; cost = 0.0169
Iteration: 650 ; cost = 0.0169
Iteration: 700 ; cost = 0.0169
Iteration: 750 ; cost = 0.0169
Iteration: 800 ; cost = 0.0170
Iteration: 850 ; cost = 0.0170
Iteration: 900 ; cost = 0.0170
Iteration: 950 ; cost = 0.0170
Iteration: 1000 ; cost = 0.0170
[[1. 2. 3. 1. 3.]
 [2. 0. 3. 1. 1.]
 [1. 2. 5. 5. 6.]
 [1. 0. 5. 4. 4.]
 [2. 1. 5. 4. 4.]
 [5. 1. 5. 4. 2.]
 [2. 3. 3. 1. 3.]]
✧

```

Figure 12. SVD matrix factorization results

### 4.3. Deploying Smart Contract

```
[9:06:52 AM] Starting server with initial configuration:
{ "gasLimit": 6721975, "gasPrice": 20000000000, "hardfork": "muirGlacier", "hostname": "127.0.0.1", "port": 7545, "network_id": 5777, "default_balance_ether": 100, "total_accounts": 10, "unlocked_accounts": [], "locked": false, "vmErrorsOnRPCResponse": true, "verbose": false, "db_path": "/Users/admin/Library/Application Support/Ganache/workspaces/Quickstart/chaindata" }
```

```
Starting migrations...
> Network name: 'development'
> Network id: 5777
> Block gas limit: 6721975

1_initial_migration.js
=====
Deploying 'Migrations'
> transaction hash: 0xa8d36a6879e40c3a13a56e0cb8c475fd61fa7ceb015b41f526a1c
> Blocks: 0
> contract address: 0x9B28326C413AB9aa3CEa34D90FCDf040aF3685F7
> account: 0xCE5bb154c297C239663BD15B3883F6c59a01323A
> balance: 99.99518728
> gas used: 244636
> gas price: 20 gwei
> value sent: 0 ETH
> total cost: 0.00489272 ETH

> Saving artifacts
> Total cost: 0.00489272 ETH
```

```
gasPrice: 20000000000
},
currentProvider: HttpProvider {
  host: 'http://127.0.0.1:7545',
  httpAgent: [Agent],
  timeout: 0,
  headers: undefined,
  connected: true,
  send: [Function (anonymous)],
  _alreadyWrapped: true
},
network_id: '5777'
},
methods: {
  'name()': [Function (anonymous)] {
    call: [Function (anonymous)],
    sendTransaction: [Function (anonymous)],
    estimateGas: [Function (anonymous)],
    request: [Function (anonymous)]
  },
}
```

Figure 13. Deploying the contract

After the privacy level of data is classified, data sharing transactions are conducted based on blockchain technology. Figure 13 shows the smart contract deployment. The first one is ‘transaction hash,’ showing that the transaction is executed, where the block ‘0’ indicates that it is the first block and refers to the genesis block. In the Ethereum environment, the cost of migrating smart contracts is calculated in units of ‘gas.’ Hence, in Figure 13, the ‘gas used’ is 244636 gas, which can be converted into Ethereum’s payment unit of ‘0.00489272ETH’.

## 4.4. Inter Planetary File System

Inter Planetary File System (IPFS) is a distributed P2P file system that attempts to connect all computers. In other words, it is a faster, safer, and open network by realizing it with P2P communication of nodes without a centralized server. Unlike the HTTP web, which has fatal consequences when large servers are disconnected, IPFS maintains a stable ecosystem even if some nodes are disconnected. IPFS can deliver high-capacity files quickly and efficiently, and because duplicate files can be known, efficient storage management is possible. Each uploaded file consists of several blocks, and each block has its own name expressed in the hash.

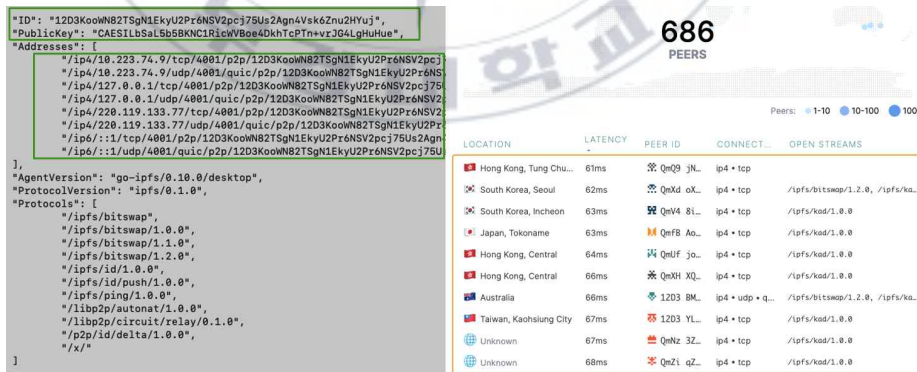


Figure 14. Inter Planetary File System

In Figure 14, the Inter Planetary File System (IPFS) has an ID, a

public key, and several created addresses. When we accessed IPFS, there were 686 nodes worldwide, and a hash of the data is shared and stored for several of these nodes.

```
<EPCOSBody>
<EventList>
<ObjectEvent>
  <!--0000 0000-->
  <eventTime> 2021-02-06T17:10:20.196-09:00</eventTime>
  <!--Timezone-->
  <eventTimezoneOffset>-09:00</eventTimezoneOffset>
  <epcList>
    <!--000 ID: PKNU-->
  <epc>urn:epc:id:pknu:2020131.1234567890</epc> </epcList>
  <action>OBSERVE</action>
  <!--0000000 ID-->
  <id>urn:epc:id:sgln:2020131.67321</id>
  </bizLocation>
  <!--0000-->
  <AutoIDLabs:ehr>
  <!--0000 ID: lisia-->
  <AutoIDLabs:documentID>urn:epc:id:lisia:2020131.28731.200</AutoIDLabs>
<documentID>
  <AutoIDLabs:diagnosis>
  <!--0000-->
  <AutoIDLabs:accuracy>presumptive clinical
<diagnosis</AutoIDLabs:accuracy>
  <!--0000-->
  <AutoIDLabs:mainInjuryAndDisease>a sprain of cervical
spine</AutoIDLabs:mainInjuryAndDisease>
  <!--0000-->
  <AutoIDLabs:subordinateInjuryAndDisease>diabetes</AutoIDLabs:subordinateInjuryAndDisease>
  <!--00 00 00-->
  <AutoIDLabs:mainCode>S134</AutoIDLabs:mainCode>
  <!--0000 00-->
```

Figure 15. Example of the EMR.txt data



dag-pb UnifxFS View on IPFS Gateway

CID QmRyDxxxyfWpuGfU1TFN7EQhJYxYHxaPbZ1B1TmcJFbp

SIZE 1 KB

LINKS 0

DATA

Object {type: "file", data: Uint8Array, blockSize: Array(0)}



**CID INFO**

QmRyDxxxyfWpuGfU1TFN7EQhJYxYHxaPbZ1B1TmcJF...  
base58btc - cidv0 - dag-pb - sha2-256-256-E...  
BASE - VERSION - CODEC - MULTIHASH

**MULTIHASH**

0x1220E01FE0078067196605BE3C5F7976A3E6  
B51ADE33CB33764A7DD40F68C5BAF2B3  
HASH DIGEST

0x12 = sha2-256  
0x20 = 256 bits

Figure 16. Stored in the decentralized web

Figure 16 shows that when the hospital uploaded the patient's EMR data (Figure 15) on IPFS as a .txt file type. The hospital that created the content registers new content in the proposal system, as shown in Figure 13.

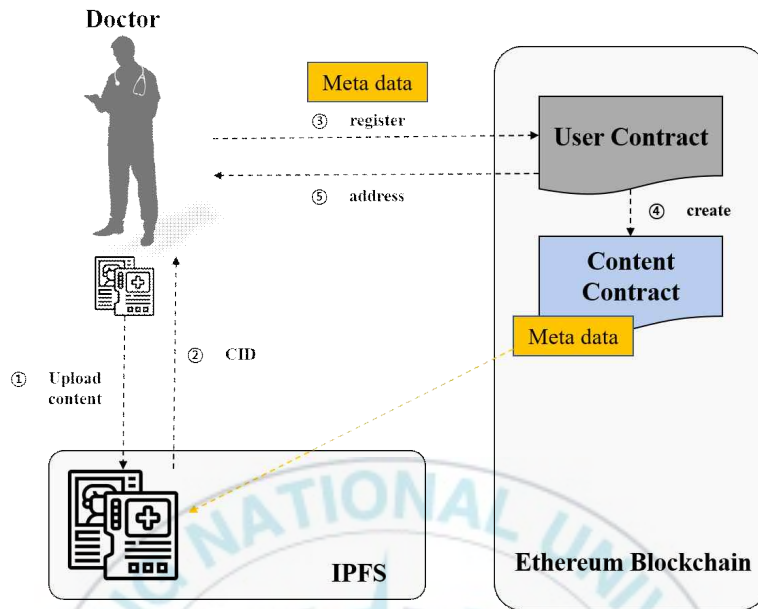


Figure 17. Uploading flow for a new content

At this time, rather than directly storing and registering content on the blockchain, it uses a method of storing content in an IPFS distributed storage and registering a 'Content Identifier (CID)' that can be accessed to the blockchain. This is to reduce the cost and network load incurred when directly registering large-sized content on the blockchain. In Figure 15, CID is a unique value that allows access to content through IPFS, and hash is a hash value of content.



## 4.5. On-Chain Data

```
{
  tx: '0x610f0763532f129250e8e9d33b05f072833f472d8037a41cfdbf3c',
  receipt: {
    transactionHash: '0x610f0763532f129250e8e9d33b05f072833f472',
    transactionIndex: 0,
    blockHash: '0x79ef8a80ce896b2518fbcdec2a5bde271819237bfe638',
    blockNumber: 5,
    from: '0xce6bb154c297c239663bd15b3883f6c59a01323a',
    to: '0xfb122c95f538e4157bfe4e08b8978c5bfee7ee1d',
    gasUsed: 111471,
    cumulativeGasUsed: 111471,
    contractAddress: null,
    logs: [ [Object] ],
    status: true,
  },
  {
    logIndex: 0,
    transactionIndex: 0,
    transactionHash: '0x610f0763532f129250e8e9d33b05f072833f4',
    blockHash: '0x79ef8a80ce896b2518fbcdec2a5bde271819237bfe6',
    blockNumber: 5,
    address: '0xfb122c95f538e4157bfe4e08b8978c5bfee7ee1d',
    type: 'mined',
    id: 'log_b0c3b816',
    event: 'PostCreated',
    args: [Result]
  }
}
```

Figure 18. Transaction of hospital

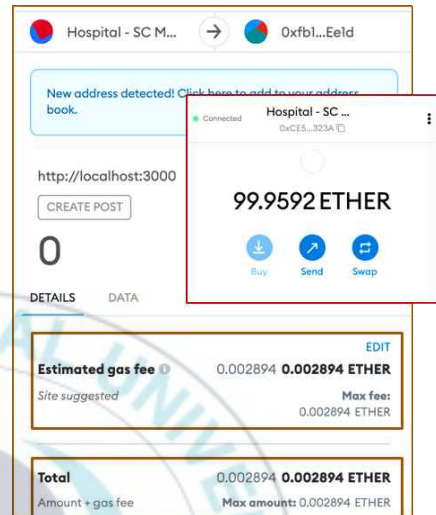


Figure 19. Completed transaction of hospital

Figure 18 is the illustration for the hospital that records the patient's data on the on-chain. Based on our simulation, the gas used to record on-chain is '111471' gas, which can be verified as '0.0002894 ETH', as shown in Figure 18. Thus, after the transaction was completed, the system confirms that '99.9592 ETH' left after the cost was paid from the MetaMask wallet of the hospital. It is worth to be noted that there have been 100 ETH before.



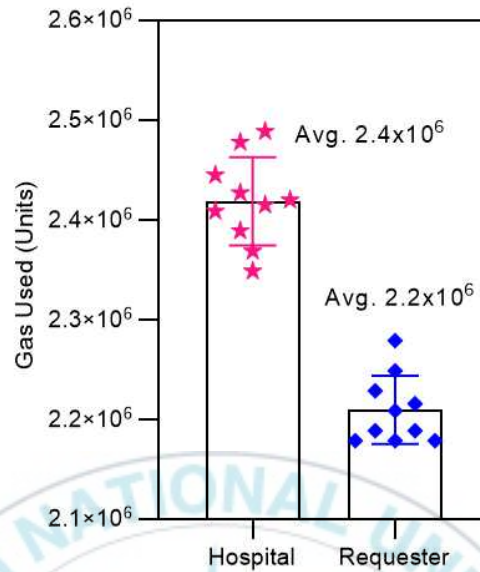


Figure 20. Hospital - Requester Gas Consumption

Figure 20 Shows the gas consumption of major entities.

Gas consumption between hospital and requester (e.g., researcher) was compared and analyzed with the stored data in IPFS according to the data sensitivity level defined by the patient. As a result, it can also be confirmed that gas consumption of hospitals is higher than the requester due to the intensity of mainly uploading data.

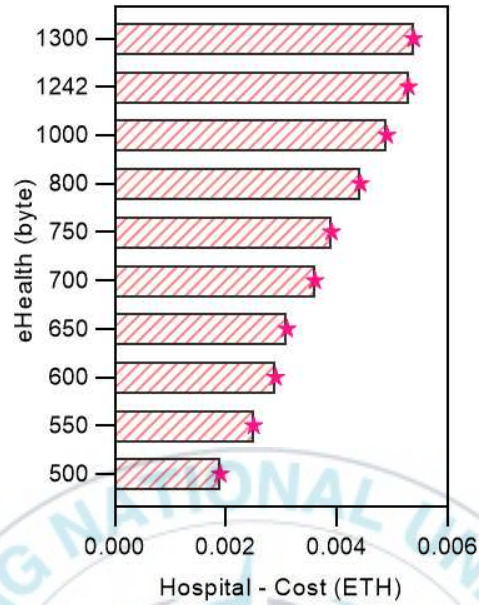


Figure 21. Hospital cost (ETH)

Figure 21 shows the ether cost required based on the data size when the hospital uploads data. We checked the results by dividing the data size by 500 bytes to 1300 bytes when considering the hospital's data from text files to image and video data. As a result, it can be seen that a high cost is required as the data size increases, and the '0.002 ETH' fee was incurred at least 500 bytes.

## V. Conclusion

We have presented the UDPM as a patient-defined data privacy management using nudge theory in decentralized e-health environments. Our objective focuses on designing state of the art UDPM empowered by nudge theory and blockchain technology in providing immutable patient's privacy management in an e-health system. The UDPM system we have designed is the initial stage of the overall design we want to achieve. As part of the embodiment of immutable data privacy management, Ethereum smart contracts promise to be implemented in the UDPM system due to the low transaction costs yet still having exemplary services. Therefore the system is expected to be a system to protect patients' privacy while medical data is collected in the COVID-19 era. However, the blockchain scalability trilemma must be considered more profoundly to balance security, decentralization, and scalability. Thus, we need a detailed estimate of the number of entity members of the UDPM system for future work. More precisely, our future work will be focusing on improving the performance of collaborative filtering with patient big data with a variant of collaborative filtering protocol and designing a more efficient Ethereum smart contract by minimizing the arbitrary inputs in the smart contract.

## Reference

- [1] Collins, F. S., & Varmus, H, “A new initiative on precision medicine”. *New England journal of medicine*, vol. 372 no.9, pp.793-795, 2015.
- [2] A. Roehrs, C. A. Da Costa, and R. da Rosa Righi, “Omniphr: A distributed architecture model to integrate personal health records,” *Journal of biomedical informatics*, vol. 71, pp. 70–81, 2017.
- [3] S. Becher, A. Gerl, B. Meier, and F. Bölz, “Big picture on privacy enhancing technologies in e-health: a holistic personal privacy work-flow,” *Information*, vol. 11, no. 7, pp. 356, 2020.
- [4] Y. Hong, T. B. Patrick, and R. Gillis, “Protection of patient’s privacy and data security in e-health services,” *2008 international conference on biomedical engineering and informatics IEEE*, vol. 1, pp. 643–647, 2008.
- [5] C. Butpheng, K.-H. Yeh, and H. Xiong, “Security and privacy in iot-cloud-based e-health systems—a comprehensive review,” *Symmetry*, vol. 12, no. 7, pp. 1191, 2020.
- [6] Ministry of Science and ICT., Korea Institute of Science & Technology Evaluation and Planning (KISTEP). (2020). "The Future of Precision Medical Technology.(정밀의료 기술의 미래)", 2020 Technology impact assessment book.
- [7] Froelicher, D., Misbach, M., Troncoso-Pastoriza, J. R., Raisaro, J. L., & Hubaux, J. P, "MedCo2: privacy-preserving cohort exploration and analysis." *Digital Personalized Health and Medicine*. IOS Press, pp. 317-321, 20

20.

- [8] Khoury, Muin J., and James P. Evans. "A public health perspective on a national precision medicine cohort: balancing long-term knowledge generation with early health benefit." *Jama* 313.21 pp. 2117-2118, 2015.
- [9] Azencott, C. A. "Machine learning and genomics: precision medicine versus patient privacy." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376 no. 2128, pp. 20170350, 2018.
- [10] B. Houtan, A. S. Hafid, and D. Makrakis, "A survey on blockchain-based self-sovereign patient identity in healthcare," *IEEE Access*, vol. 8, pp. 90478–90494, 2020.
- [11] H. Shafagh, L. Burkhalter, S. Ratnasamy, and A. Hithnawi, "Droplet: Decentralized authorization and access control for encrypted datastreams," in *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pp. 2469–2486, 2020.
- [12] B. Houtan, A. S. Hafid, and D. Makrakis, "A survey on blockchain-based self-sovereign patient identity in healthcare," *IEEE Access*, vol. 8, pp. 90478–90494, 2020.
- [13] A. Farouk, A. Alahmadi, S. Ghose, and A. Mashatan, "Blockchain platform for industrial healthcare: Vision and future opportunities," *Computer Communications*, vol. 154, pp. 223–235, 2020.
- [14] Nakamoto, Satoshi. "Re: Bitcoin P2P e-cash paper." *The Cryptography Mailing List*, 2008.

- [15] S. Rahmadika, M. Firdaus, S. Jang, and K.-H. Rhee, "Blockchain-enabled 5g edge networks and beyond: An intelligent cross-silo federated learning approach," *Security and Communication Networks*, vol. 2021, 2021.
- [16] S. Rahmadika and K.-H. Rhee, "Unlinkable collaborative learning transactions: Privacy-awareness in decentralized approaches," *IEEE Access*, 2021.
- [17] D. D. F. Maesa and P. Mori, "Blockchain 3.0 applications survey," *Journal of Parallel and Distributed Computing*, vol. 138, pp. 99–114, 2020.
- [18] Cong, L. W., & He, Z. "Blockchain disruption and smart contracts." *The Review of Financial Studies*, vol. 32, no. 5, pp. 1754-1797, 2019.
- [19] Resnick, P., & Varian, H. R. "Recommender systems." *Communications of the ACM*, vol. 40 no. 3, pp. 56-58, 1997.
- [20] Hema, P., & Pillai, N. S. "Efficient mining and recommendation of sparse data through collaborative filtering technique in medical transcriptions." in *International Conference on Information Communication and Embedded Systems (ICICES2014)*. IEEE, pp. 1-5, 2014.
- [21] Wang, H., Ma, S., Dai, H. N., Imran, M., & Wang, T. "Blockchain-based data privacy management with nudge theory in open banking." *Future Generation Computer Systems*, vol. 110, pp. 812-823, 2020.
- [22] Bell, R. M., & Koren, Y. "Improved neighborhood-based collaborative filtering." in *KDD cup and workshop at the 13th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 7-14, 2007.
- [23] Mnih, A., & Salakhutdinov, R. R. "Probabilistic matrix factorization." in

- Advances in neural information processing systems*, pp. 1257-1264, 2008.
- [24] Bokde, D., Girase, S., & Mukhopadhyay, D. "Matrix factorization model in collaborative filtering algorithms: A survey." *Procedia Computer Science*, vol. 49, pp. 136-146, 2015.
- [25] Graf, R. (2019). Nudging before the nudge? Behavioural traffic safety regulation and the rise of behavioural economics. In *Handbook of behavioural change and public policy*. Edward Elgar Publishing.
- [26] N.-H. Kwon, Korea Institute of Taxation and Finance. (2018) "Public Policy using Nudge: Current Status and Implications (넛지(Nudge)를 활용한 공공정책: 현황과 시사점)" Financial Performance Evaluation Center of the Korea Institute of Taxation and Finance.
- [27] Katarahweire, M., Bainomugisha, E., & Mughal, K. A. (2020). Data classification for secure mobile health data collection systems. *Development Engineering*, 5, 100054.
- [28] Wikipedia, "Collaborative filtering, 2021, [online]. available:[https://en.wikipedia.org/wiki/collaborative filtering](https://en.wikipedia.org/wiki/collaborative_filtering)," 2021.
- [29] Y.Lab, "[Recommender System]-Dimension reduction technique using Autoencoder, 2019, [online]. <https://yamalab.tistory.com/116?category=747907> " 2021.