발 간 등 록 번 호
11-1240245-000057-14

OPEN 공공누리 출처표시 상업용금지 변경금지 공공저작물 자유이용허락

2020년 연구보고서

데이터 거래 활성화를 위한 구독통계 플랫폼 구축 사전 연구

2021. 4.

http://kostat.go.kr/sri

ISSN 2288-1166(Print)
ISSN 2733-4120(Online)

통계청
통계개발원

연구보고서 2020-05

데이터 거래 활성화를 위한 구독통계 플랫폼 구축 사전 연구

변준석 · 박민정 · 천정희

Statistics Korea
Statistics Research Institute

발간사

통계개발원(Statistics Research Institute, SRI)은 2006년에 개원한 이래 국내에 유일한 「국가통계 전문연구기관」으로서 국가통계・조사 방법론의 발전과 데이터기반 정책 연구・개발 분야에서 선도적인 역할을 하고 있습니다. 지난 한 해 동안 SRI는 "광출어혁(光出於革)"의 정신, 즉 혁신에(革) 기반한 실용적인 「팀연구」를 통해서 국가통계・데이터과학의 미래를 밝히고자(光) 노력하였습니다.

SRI의 「2020년 연구보고서」는 이러한 혁신연구의 결과물입니다. 특히 코로나19의 경제・사회 영향을 측정하는 지표 설계 연구를 비롯해서 「데이터기반 인구・사회・경제정책」을 뒷받침하는 심도 깊은 연구가 진행되었습니다. 예를 들어서 저혼인・저출산・초고령 사회에 대응하기 위한 데이터 심층 분석, 균형발전 정책수립을 지원하는 지역사회지표 체계구축, 그리고 SDG를 활용한 남북한 통합통계 방안 등입니다. 또한 2020년 2월에 출범한 「SDG 데이터연구센터」는 지속가능발전목표(SDG)의 이행현황 점검과 SDG 대국민 데이터 서비스를 준비하는 실용적인 연구를 진행하였습니다.

국가통계・데이터과학・조사방법론 분야에서는 「데이터 정보보호센터」를 중심으로 개인정보보호 방법의 통계지리정보서비스(SGIS) 적용 등 실용적인 연구가 진행되었습니다. 또한 「조사표연구센터」를 활용하여 빅데이터 활용 홈페이지 개선을 위한 사용성 평가를 비롯해서 「조사표 인지실험」을 적용한 국가통계의 품질개선이 이루어졌습니다. 특히 지난해에는 데이터과학(Data Science)을 국가통계에 활용하고자하는 기초 연구를 추진함으로써 「데이터 혁신 방법론」의 새로운 전기를 준비하였습니다.

금년에 개원 15주년을 맞는 SRI는 본 연구보고서가 증거기반정책 입안자의 데이터 활용에 실제적인 도움이 될 뿐만 아니라 국가통계 생산자의 혁신적인 조사방법론 개발에 유용한 자료로 활용될 수 있기를 기대합니다. SRI가 「국가통계 싱크탱크」로서 국내외적으로 선도적인 역할을 할 수 있도록 독자 여러분의 지속적인 격려를 부탁드립니다. 국가통계의 도약과 혁신적 개발을 위하여 제언이 있다면 언제든지 말씀하여 주십시오. 겸허히 귀 기울이겠습니다.

이번 「2020년 연구보고서」를 위하여 실용적이고 품질 높은 연구 결과를 도출하기 위해 광출어혁(光出於革)의 정신으로 최선을 다한 연구진과 대내외적으로 협력·공동연구에 참여한 민·관의 연구자들에게 따스한 감사를 전합니다. 끝으로 본 연구에서 제시된 내용 및 결과는 저자 개인의 의견이며, 통계청 또는 통계개발원의 공식적인 견해가 아님을 밝혀둡니다.

2021년 4월

통계청 통계개발원장

전 영 일

목 차

제1장 서론 ····· 1

제2장 동형암호 이해와 활용 ····· 3

1. 동형암호의 이해 ····· 3
2. 동형암호의 활용 ····· 15

제3장 데이터 정보보호 ····· 26

1. 데이터 정보보호의 필요성 ····· 27
2. 노출위험과 정보손실 ····· 33
3. 재현자료 활용 ····· 40
4. 차등정보보호 적용 ····· 44
5. 플랫폼 맞춤형 알고리즘 개발 ····· 48
6. 분석결과물 반출 가이드라인 운영 ····· 51

제4장 구독 통계 플랫폼 실행방안 ····· 53

1. 데이터 거래 현황 ····· 53
2. 데이터 거래 시 애로사항 및 정책진단 ····· 54
3. 데이터 구독 플랫폼 ····· 58
4. 기대효과 ····· 64

제5장 결론 및 시사점 ····· 71

참고문헌 ····· 73

Abstract ····· 75

요 약

데이터 활용 여부가 기업경쟁력의 척도가 되는 시대가 되었다. 그러나 데이터 활용이 높아질수록 개인정보보호의 중요성은 더더욱 부각되고 있다. 데이터 활용과 개인정보보호라는 두 마리 토끼를 잡기 위한 혁신적인 플랫폼이 필요한 시점이다. 본 보고서에서는 이를 실현하기 위한 데이터 구독 플랫폼을 제안한다. 구독통계 플랫폼은 데이터 수집·유통 과정에서의 동형암호화, 제공 이후의 데이터 정보보호를 핵심 솔루션으로 하여 안전하고 효율적인 데이터 거래 생태계 형성을 지향한다.

제2장에서는 차세대 암호, 제 4세대 암호로 알려진 동형암호에 대하여 논의한다. 동형암호는 기존의 암호체계와 달리 암호화된 데이터 자체를 가지고 연산이 가능한 특성이 있다. 2016년 Cheon, Kim, Kim과 Song(2016)이 암호화된 상태에서 덧셈과 곱셈뿐만 아니라 반올림이라는 세 번째 연산까지 가능한 혜안(HEaaN)이라는 동형암호를 발표하면서, 이러한 형태의 암호가 실현 가능하게 되었다. 동형암호는 이미 군 보안통신, 미사일 해킹 방지 등의 전통적인 암호화 활용 영역뿐 아니라 신용정보 계산, DNA분석 등 일반 산업계에도 응용된 바 있다. 심지어 암호화된 상태에서 기계학습을 진행하는 동형기계학습도 가능하다.

제3장에서는 데이터 정보보호 필요성 및 방법론에 대하여 논의한다. 우선 독자들의 이해를 돕기 위하여 예제를 통하여 그 필요성에 대하여 이야기하고, 노출위험 및 정보손실 측정 개념의 중요성을 설명한다. 데이터 정보보호 방법론(기술)으로 재현자료, 차등정보보호, 대량의 빈도표 제공을 위한 비밀보호 알고리즘 BSCA 개발을 다룬다. 나아가 데이터센터 운영에 있어 반출물에 대한 통계적 노출제어 가이드라인 운영이 필요함을 강조한다.

제4장에서는 이상의 두 방법을 혁신적으로 접목한 데이터 구독 플랫폼에 대하여 설명한다. 이 플랫폼에서 데이터 소유자는 동형암호 처리된 데이터를 플랫폼에 제공하며, 데이터 구독자는 제공된 데이터를 암호화된 상태로 자유롭게 분석할 수 있다. 단 최종 결과는 데이터 정보보호 처리가 된 상태로 반출되어 데이터 수집·유통·반출 과정에서 개인정보가 철저하게 보호된다.

손쉬운 이동과 복제로 인한 데이터 재제공, 공급자와 수요자 간 시각차에서 오는 가격 괴리 등으로 데이터 거래시장은 자연발생적으로 생성되기 어렵다. 반면 데이터 구독 플랫폼에서는 이미 데이터가 동형암호화되어 있으므로 재제공이 아무런 의미가 없고, 회원 가입제 구독 시스템이므로 수요/공급자 간 개별 가격을 협상할 필요가 없다. 즉 안전한 구독 플랫폼에서는 시장 기능 작동이 자연스럽게 개시될 수 있으므로, 보안 및 가격 괴리 문제로 꿈적하지 않던 데이터 거래 생태계 형성이 가능해진다.

데이터 암호화 그리고 다양한 유형의 반출물에 대한 데이터 정보보호 처리는 역설적이게도 데이터 이동 및 활용을 자유롭게 한다. 동형암호화된 데이터는 암호화된 상태로 유통되기 때문에 데이터 담당자 또는 관리자도 정보를 확인할 수 없다. 데이터 정보보호 검증을 거친 데이터 반출은 개인정보 노출위험을 제어한다. 따라서 이에 따른 자유로운 데이터 거래는 효율적인 데이터 거래시장을 만든다.

주요용어 : 동형암호, 데이터 정보보호, 차등정보보호, 재현자료, 데이터 구독, 구독 통계 플랫폼, 데이터 보험

제 1 장

서 론

□ 4차 산업혁명 시대에 있어서 기업·산업의 경쟁력은 데이터를 얼마나 잘 활용하는가 또는 데이터를 활용하여 사용자의 경험을 얼마나 만족하게 하는가에 달려 있다고 해도 과언이 아니다.

○ 아마존, 구글 등 이른바 빅테크 기업의 경쟁력은 빅데이터 수집 능력과 활용에 있다는 것은 잘 알려진 사실이다.

○ 정부 정책도 이제는 거시적 증거기반정책(Evidence Based Policy)을 넘어 데이터·통계기반 지역 간·세대 간 맞춤형 정책 추진에 힘쓰고 있다.

□ 코로나로 침체된 경기를 회복하기 위하여 우리 정부는 데이터 댐 등 한국판 뉴딜 종합계획(2020.7.14.)을 추진하고 있다.

○ 한국판 뉴딜은 디지털 뉴딜, 디지털 그린·융복합, 그린 뉴딜로 구분된다.

○ 데이터 관련 사업은 디지털 뉴딜의 데이터 댐 분야에 속해 있다.

○ 데이터 댐 분야에서는 분야별 데이터 플랫폼을 10개에서 30개로 확대, 공공데이터 142만 개 신속 확대 정책을 펴고 있다.

○ ‘22년까지 총사업비 3.2조(국비 3.1조), 일자리 16.1만 개 창출, 나아가 ’25년까지 총사업비 6.7조(국비 6.4조), 일자리 29.5만 개를 목표로 하고 있다.

□ 데이터의 중요도 및 활용도의 증가에 따라 최근 데이터 정보보호도 매우 중요한 이슈로 떠오르고 있다.

○ 개인정보를 활용한 스팸 수준의 광고, 보이스 피싱 등의 범죄 등이 횡횡하는 사회적 분위기 등에 따라 개인정보보호 의식은 점차 높아지고 있다.

○ 특히 개인정보가 담긴 데이터의 연계 또는 통합 관리 시 데이터 해킹, 개인정보 유출 등에 대한 우려의 목소리가 높다.

□ 이에 정부에서도 법률지원 등을 통하여 개인정보보호에 나서고 있다.

○ 개인정보보호법(2011.3.11.)을 제정하여 그해 9월부터 시행에 들어갔다.

○ 이후 개인정보 위원회에서는 개인정보지침 가이드라인 등을 꾸준히 발간하고 있다.

○ 2016.6월 개인정보 비식별화 조치 가이드라인(관계부처 합동)에 이어 2020.9월에는 가명정보 처리 가이드라인(개인정보보호위원회)을 배포하여 개인정보보호가 가능한 데이터 활용 방안을 적극 지원하고 있다.

□ 구독 통계 플랫폼은 데이터 활용과 개인정보보호 두 마리 토끼를 잡을 수 있는 혁신적인 데이터 거래 플랫폼이다. 뿐만 아니라 자연스러운 데이터 거래 시장 생태계 형성에도 기여할 수 있을 것으로 기대한다.

○ 플랫폼은 데이터 수집·유통 과정에서의 동형암호화와 더불어 다양한 유형의 반출물에 대한 차등정보보호 등의 개인정보보호를 위한 기술적인 조치로 개인정보를 철저하게 보호한다.

○ 암호화된 상태에서 데이터를 유통시킴으로써 데이터 거래를 자유롭게 할 수 있는 장점이 있다.

○ 수요/공급자 간 데이터 가격 괴리를 해소하는 회원제 구독 시스템, 데이터 제재공을 원천 차단하는 동형암호를 통하여 자연스러운 데이터 거래 시장 형성에도 기여할 수 있다.

□ 제2장에서는 동형암호의 이해와 활용을 다룬다. 제3장에서는 데이터 정보보호에 대하여 심도 있게 논의한다. 제4장에서는 데이터 구독 플랫폼을 설명하고 실행방안에 대하여 논한다. 제5장에서는 시사점을 정리하며 보고서를 마무리 한다.

제 2 장

동형암호의 이해와 활용

1. 동형암호의 이해

1.1 암호의 개요

□ 암호는 패스워드(password, 인증기술)나 키(key)와 같은 비밀정보를 알지 못하는 사람은 해당 정보를 알 수 없도록 만드는 기술을 말한다.

○ 일반적으로 암호 알고리즘은 키생성(key generation), 암호화(encryption) 그리고 복호화(decryption)의 세 단계로 이루어진다. 키생성 단계에서는 암호화키와 복호화키를 생성하고 암호화 과정에서는 평문에 암호화키를 적용하여 암호문으로 바꾸며 복호화 과정에서는 암호문에 복호화키를 적용하여 원래의 평문을 복구한다.

○ 암호화키와 복호화키가 같은 경우를 대칭키 암호(symmetric key encryption)라 하고, 동일하지 않은 경우를 비대칭키 암호(asymmetric key encryption) 혹은 공개키 암호(public key encryption)라고 부른다.

□ 공개키 암호의 출현 전에는 어떤 정보를 저장하거나 메시지를 공유하고자 하는 경우 당사자들 사이에 동일한 비밀키(secret key)를 먼저 설정하고 이후 그 비밀키를 이용하여 정보를 암호화하여 저장하거나 암호화한 메시지를 공유할 수 있었다.

○ 이러한 대칭키 암호 방식은 규모가 작은 조직이나 군대와 같이 긴밀한 조직관리가 가능한 조직에서는 사용할 수 있지만, 오늘날과 같은 대규모 네트워크 환경에서는 당사자들 사이에 비밀키를 안전하게 공유한다는 것 자체가 어려운 문제가 된다.

□ 1976년 Diffie 와 Hellman은 공개키 암호화라는 새로운 아이디어를 제공하였다.

○ 이것은 암호화에 사용할 키를 공개하여 누구나 암호화를 할 수 있게 하고, 복호화키는 비밀로 하여 이 복호화키를 소유한 사람만이 정보를 알 수 있도록 하는

알고리즘으로 당사자들 간에 사전에 비밀키를 공유하지 않고도 안전하게 통신을 할 수 있도록 하였다.

○ 이때, 암호화에 사용되는 키를 공개키(public key)라고 하고 복호화에 사용되는 키를 비밀키(secret key)라고 부른다.

1.2 기존 암호의 한계

가. 비밀키 유출

□ 현대암호는 엄격하고 엄밀한 안전성 개념을 바탕으로 하고 있어, 안전성 가정에 맞추어 암호알고리즘을 적절히 사용한다면 제시된 안전성은 수학적으로 보장된다.

○ 암호의 안전성 가정에는 기반문제가 안전해야 한다는 것 이외에 당연하게도 비밀키가 안전하게 보관되어야 한다는 것이 포함되어 있다. 이 가정은 중요문서에만 암호가 적용되던 80년대 이전에는 쉽게 달성될 수 있는 것이었지만 현재의 암호체계에서는 가장 보편적인 공격의 목표가 되고 있다.

○ 컴퓨터와 인터넷의 발달로 암호의 사용이 빈번해지고 이에 따라 비밀키를 사용해야 하는 일이 많아지면서 공격자가 비밀키를 획득할 가능성이 더욱 커지고 있는 것이다.

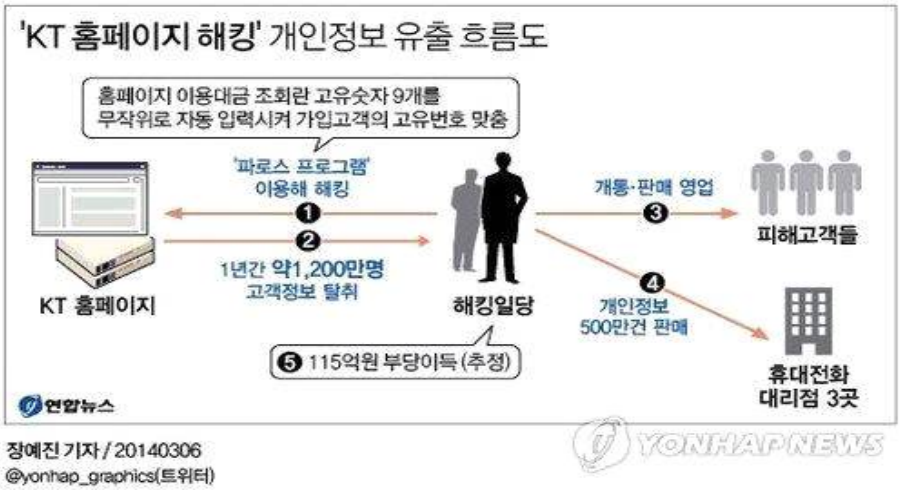
□ 실제로 Snowden이 발표한 리포트[1]에서는 미국 NSA가 악성코드 등을 통해 세계 각 나라의 암호에 쓰이는 비밀키를 획득하고 이를 이용하여 주요 정보들을 해독하여 수집하고 있다고 폭로하고 있다.

○ 기업이 보관하고 있는 데이터 역시 암호화하면 안전성이 높아질 것이라 기대하지만 현실에서는 그렇지 않은 경우가 많다.

○ 우리나라에서는 개인정보보호법에 따라 주민번호 암호화 의무조치를 시행하고 있고 (100만 명 이하의 데이터의 경우 2017년 7월 1일부터 적용되고 이 이상의 경우 2018년 1월 1일부터 적용), 민감한 개인정보 및 주민번호를 암호화하여 보호하고 있지만 개인정보 유출사고는 멈추지 않고 있다(<그림 2-1>).

---

1) Snowden(2013.6.10.), ‘Code Name ‘Verax’: Snowden, in Exchanges with Post Reporter, Made Clear He Knew Risks.’, The Washington Post.



<그림 2-1> KT 홈페이지 해킹 개인정보 유출 흐름도

□ 이는 저장된 데이터를 활용할 때마다 복호화를 실행하기 위해 비밀키를 꺼내야 하는데 이때 악성코드 등을 이용하는 해커에게 비밀키가 탈취될 수 있기 때문이다.

○ 실제로 현실에서는 비밀키가 안전하게 보관된다는 가정과 달리 사용의 편의를 위해 메모리에 상주하거나 백엔드 데이터베이스(backend database)보다 안전성이 떨어지는 웹서버에 보관되기도 하여 비밀키가 유출될 가능성이 높다.

○ 이에 따라 비밀키의 사용빈도를 줄이기 위해 비밀키 없이도 암호화한 상태로 여러 가지 연산을 수행할 수 있는 암호가 현대암호학의 주요 연구목표가 되었다.

○ 이 경우 사용자가 데이터를 볼 때에는 비밀키가 필요하나 그 이외에는 비밀키를 이용하지 않고 처리함으로써 비밀키를 안전한 곳에 보관할 수 있고 해커의 공격에 대비할 수 있는 것이다.

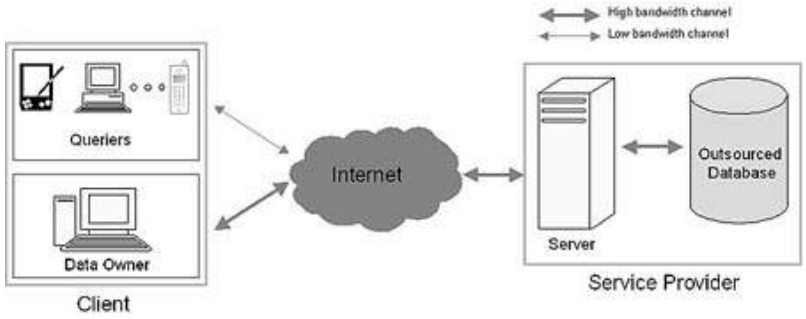
나. 관리자에 의한 자료유출

□ 비밀키 암호는 대량의 데이터를 효율적으로 암호화할 수 있는 장점이 있어서 비밀통신뿐만 아니라 외부 저장공간(outsourced database) 등에 자료를 암호화 저장하는 데도 사용되어 왔다.

○ 그러나, 최근 기업 데이터베이스에서의 고객 정보 유출 사례나 개인 홈페이지에 저장된 사진 등의 개인 정보 유출 사례가 보고되면서, 이러한 외부 저장 공간에 저장된 정보에 대한 보안문제가 이슈가 되고 있다.

○ 외부 저장공간에서의 보안 문제는 개인이 독립된 저장공간을 이용하여 스스로

정보를 관리하는 것과 달리, 정보의 소유자(data owner)와 저장공간을 관리하는 주체가 서로 다르기 때문에 발생한다.



<그림 2-2> 데이터베이스 시스템에서 서비스 제공

□ 일반적으로 데이터베이스 시스템의 서비스 제공자(Service Provider)는 자료(data)를 저장한 사람(owner)이든 시스템에 질의(query)를 하는 사람(client)이든 누군가로부터 서비스를 요청받게 되면 필연적으로 암호화된 데이터를 복호화하게 된다.

○ 이러한 복호화 과정에 개입하게 되는 관리자(administrator)는 데이터 자체나 질의 그리고 질의에 대한 결과를 미리 알 수밖에 없다. 접근제어(access control)나 키 관리(key management) 기법들은 외부 침입자를 막는 데는 유효한 방법이지만, 저장 공간의 소유자(administer)가 저장되어 있는 자료를 열람하는 것은 근본적으로 방지하지는 못한다.

□ 서비스 제공 과정이나 관리자들에 의한 자료 유출을 방지하기 위하여 암호화된 상태에서 검색이 가능한 암호화 방법(searchable encryption)이 제안되기도 하였다.

○ 이러한 검색가능 암호는 자료 자체를 암호화하는 암호문과 자료에 포함된 키워드 정보를 포함한 인덱스(index) 그리고 주어진 키워드에 해당하는 암호문을 검색하기 위한 트랩도어(trapdoor) 등이 요구되는데, 연산속도나 안전성과 무관하게 사전에 지정한 키워드, 인덱스, 트랩도어로 제한된 검색기능만 가능하다는 문제가 있다.

○ 따라서 데이터베이스에 암호화한 상태로 자료을 저장하고, 사용자가 원하는 검색뿐 아니라 다양한 통계적 처리를 할 수 있는 방법이 요구되고 있다.

다. 개인정보보호법 등 관련 법령 강화

□ 지난 2016년 세계경제포럼에서 클라우스 슈밥(Klaus Schwab)이 4차 산업혁명을 언급한 이후, 빅데이터는 그 핵심기술로 인정받고 있다.

○ 그런데, 최근 페이스북 이용자 개인정보 유출 사례 등 개인정보 유출 사고가 빈번하게 발생하면서 세계 각국에서는 개인정보보호법을 강화[2]하고 있어서 프라이버시를 보호함과 동시에 데이터를 분석 처리할 수 있는 기술이 절실히 요구되고 있다.

○ 따라서, 국내 기업 및 정부의 해외 진출이나 해외 정부, 기관과의 협업에 있어, HIPAA, GDPR을 준수하는 동시에 개인정보를 활용할 수 있는 기술적 대책이 절실히 요구되고 있다.

□ 그런데, 개인정보보호법을 준수하면 개인 프라이버시 침해를 방지할 수는 있지만, 사용가능한 데이터 종류를 강력히 제한하고 있어 데이터의 활용도를 현저히 떨어뜨린다.

○ 그 해결책으로 차별화된 프라이버시(Differential Privacy) 등 데이터 비식별화 기술이 제시되고 있으나, 데이터의 비식별화 과정에서 데이터의 손실이 발생하고, 비식별화 정도에 대한 자동화된 평가가 어려우며, 다른 정보와 결합 시 재식별의 위험성이 존재한다.

1.3 4세대 암호기술의 개념 및 특성

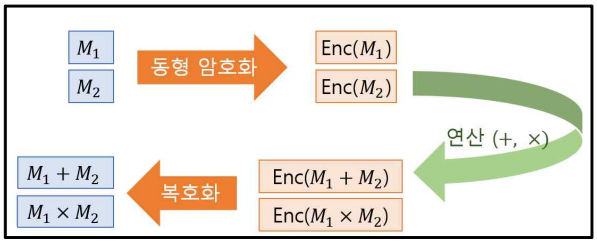
□ 동형암호[3]는 이러한 요구에 부응하는 것으로 암호화된 데이터에 대해 복호화하지 않고 마음대로 연산을 수행할 수 있는 암호화 방법이다.

○ 암호문들을 이용한 연산의 결과는 새로운 암호문이 되며, 이를 복호화하여 얻은 평문은 암호화하기 전 원래 데이터의 연산 결과와 같다.

---

2) 대표적으로 미국의 HIPAA(의료정보에 관한 개인정보보호법, 2003년, 의료정보를 사용하는 기관이 민간인의 건강정보를 공개, 사용, 접근하는데 허가가 필요하도록 함.), 유럽연합의 GDPR(일반정보보호규정, 2018년, 주소, 주민등록번호, 개인의 IP주소나 쿠키데이터 등 광범위한 개인정보를 보호) 등이 있다.

3) 동형암호(homomorphic encryption)에서 동형을 뜻하는 homomorphic은 수학에 흔히 다루고 있는 homomorphism(동형성)에서 온 것으로 동일한 유형의 두 대수 구조 사이의 연산을 보존하는 사상(map)을 말한다. 즉 동형암호는 평문공간의 원소를 암호문 공간의 원소로 대응시키는 사상으로 덧셈과 곱셈 등 특정한 연산을 보존하는 암호시스템이다.



<그림 2-3> 동형암호의 개념

○ <그림 2-3>과 같이 평문 M_1, M_2 를 동형암호화한 암호문 $Enc(M_1)$, $Enc(M_2)$가 주어져 있을 때, 이 두 암호문을 복호화하지 않고도 두 평문을 더한 (M_1+M_2)에 대한 암호문 $Enc(M_1+M_2)$이나 두 평문을 곱한 $(M_1\times M_2)$의 암호문 $Enc(M_1\times M_2)$을 얻을 수 있다.

○ 또한, 이 암호문들은 애초의 평문을 더하거나 곱한 값의 암호문과 동일한 결과를 포함하고 있어서, 이를 복호화하면 평문 자체를 더하거나 곱한 결과와 동일한 값을 얻을 수 있는 암호화 방법이다.

○ <그림 2-4>와 같이 컴퓨터의 모든 불함수(Boolean Funcions) 연산을 배타적 논리합(XOR)과 논리곱(AND)으로 표현할 수 있는 것과 마찬가지로 암호화된 데이터를 복호화하지 않고 덧셈과 곱셈이 가능한 동형암호의 기능을 확장하면, 통계연산이나 기계학습 등 다양한 연산의 수행도 가능하게 된다.

A	B	–	Y_0	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}
0	0		0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
0	1		0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	0		0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	1		0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1

$Y_0 = A \cdot \bar{A}$
$Y_1 = (\bar{A} \cdot \bar{B})$
$Y_2 = (\bar{A} \cdot B)$
$Y_3 = (\bar{A} \cdot B) + (\bar{A} \cdot \bar{B})$
$Y_4 = (A \cdot \bar{B})$
$Y_5 = (A \cdot \bar{B}) + (\bar{A} \cdot \bar{B})$
$Y_6 = (A \cdot \bar{B}) + (\bar{A} \cdot B)$
$Y_7 = (A \cdot \bar{B}) + (\bar{A} \cdot B) + (\bar{A} \cdot \bar{B})$

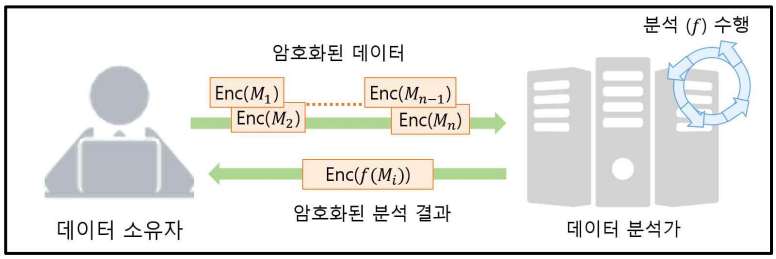
$Y_8 = (A \cdot B)$
$Y_9 = (A \cdot B) + (\bar{A} \cdot \bar{B})$
$Y_{10} = (A \cdot B) + (\bar{A} \cdot B)$
$Y_{11} = (A \cdot B) + (\bar{A} \cdot B) + (\bar{A} \cdot \bar{B})$
$Y_{12} = (A \cdot B) + (A \cdot \bar{B})$
$Y_{13} = (A \cdot B) + (A \cdot \bar{B}) + (\bar{A} \cdot \bar{B})$
$Y_{14} = (A \cdot B) + (A \cdot \bar{B}) + (\bar{A} \cdot B)$
$Y_{15} = (A \cdot B) + (A \cdot \bar{B}) + (\bar{A} \cdot B) + (\bar{A} \cdot \bar{B})$

<그림 2-4> XOR, AND 그리고 NOT 게이트로 표현한 Boolean Functions

○ 예를 들면 <그림 2-5>와 같이 데이터 소유자가 데이터 분석을 의뢰하되 데이터를 노출시키고 싶지 않을 때, 데이터 소유자는 데이터를 동형암호화하여 데이터

분석가에게 제공한다. 데이터 분석가는 동형암호 연산을 통하여 데이터 분석 결과에 대한 암호문을 산출하고 데이터 소유자에게 전달한다.

○ 이때 비밀키를 데이터 소유자만 가지고 있다면, 데이터 분석가는 산출된 데이터 분석 결과를 알 수 없고, 비밀키를 가지고 있는 데이터 소유자만 복호화를 할 수 있기 때문에 자신의 데이터를 노출시키지 않고 분석 결과를 얻는 것이 가능하게 된다.



<그림 2-5> 암호화된 상태에서의 데이터 분석

1.4 암호의 세대 구분

□ 1차 산업혁명은 증기기관이 주도하였으며, 2차 산업혁명은 전기와 컨베이어벨트가 주도하였다. 3차 산업혁명은 컴퓨터와 자동화 시스템 등으로 상징되며, 4차 산업혁명은 빅데이터(Big Data), 클라우드(Cloud), 사물인터넷(Internet of Things) 등으로 대표되는 상황이다.

○ 암호도 이와 같은 구분이 가능하다. 1세대 암호는 앞서 언급한 비밀번호(password)나 군대에서 사용하는 암구호와 같은 단순한 인증 수단이 될 것이다. 2세대 암호는 비밀키 암호(symmetric key encryption), 3세대는 공개키 암호(asymmetric key encryption)로 구분할 수 있을 것이다. 그러한 맥락에서 동형 암호는 4세대 암호라고 불린다.

1.5 동형암호의 역사[4)]

□ 동형암호는 1978년 RSA 암호의 저자 중 두 명인 Rivest, Addleman과 MIT의 Dertouzos가 공동으로 저술한 논문 RAD78[5)]에서 최초로 소개 되었다. 저자들은

4) 이화여자대학교 수리과학연구소 뉴스레터 창간호(2020.11월)에 게재된 '동형암호의 과거와 현재(천정희)'를 요약하였다.

논문에서 동형암호의 개념을 소개하고 몇 가지 후보 스킴(scheme)과 한 가지의 응용을 제시하였다.

○ 응용이 흥미로운데 데이터베이스를 운영하기 어려운 소규모 대부업자가 외부 업체에 데이터의 저장을 위탁하고 이 데이터를 가공하여 얻고 싶은 계산 결과를 프라이버시를 보존하면서 수행하고 싶을 때 동형암호를 사용한다는 시나리오였다. 요즘 널리 사용되는 외주계산(outsourced computation) 혹은 클라우드(cloud)의 개념으로도 볼 수 있다.

□ 동형암호는 암호화한 상태로 수행할 수 있는 연산의 종류에 따라 부분동형암호(partial homomorphic encryption)와 유한동형암호(somewhat homomorphic encryption, SHE)로 나눌수 있다. 사실 가장 널리 사용되는 RSA와 ElGamal도 곱셈연산을 암호화한 상태에서 수행할 수 있으므로 부분동형암호의 일종인 곱셈동형암호라고 할 수 있다.

○ 1982년 발표된 최초의 증명 가능한 암호인 Goldwassser-Micali82 스킴도 동형암호인데 배타적 논리합(XOR)연산을 보존한다. 최초의 정수 덧셈을 보존하는 암호는 1998년 발표된 Okamoto-Uchiyama 스킴과 이를 개선한 Paillier00 스킴을 들 수 있다.

○ 반면 RAD78에서 제시된 스킴은 암호화한 상태에서 덧셈과 곱셈을 모두 수행할 수 있는 암호체계로서 이를 이용하면 임의의 연산을 암호화한 상태에서 수행할 수 있다.[6] 이는 유한체상 임의의 연산은 덧셈과 곱셈의 반복으로 이루어진 다항식으로 표시할 수 있기 때문이다.

○ 다만 RAD78에 제시된 스킴들은 모두 공격이 되었고, 오히려 Boneh-Lipton96처럼 동형암호가 불가능하다는 것을 증명하려는 방향의 연구들이 다수 출판되었다. 이중 최종판으로는 유한체 간의 동형연산을 보존하는 복호화함수는 다항식시간(Polynomial Time)[7]에 공격이 된다는 결과가 있다.

□ 2009년 Gentry는 암호화한 상태에서 임의의 연산을 무한번 반복할 수 있는 동형

---

5) Rivest, Adleman과 Dertouzos(1978), On Data Banks and Privacy Homomorphisms, Foundations of secure computation 4(11), pp. 169-180.

6) 이를 튜링완전성이라 하는데, 이처럼 튜링머신이 수행하는 모든 연산을 생성할 수 있는 연산집합에는 {NAND}, {NOR}, {AND, NOT}, {OR, NOT} 등이 있다. 1980년대 초반에 NAND나 NOR를 보존하는 동형암호를 직접 설계하려는 시도가 있었지만 성공하지 못하였다.

7) 컴퓨터가 문제를 계산하는데 걸리는 시간$O(n)$이 문제의 크기 n의 다항식 함수보다 크지 않은 것을 말한다.

암호를 발표하였다. Gentry의 암호는 대수적인 연산에 제한되어 있던 기존의 접근 방식에서 벗어나 대수적 연산이 불완전하여 다수의 연산 후에는 더 이상 대수적 연산이 불가능한 암호체계를 제안하였다.

○ 이를 유한동형암호(somewhat homomorphic encryption, SHE)라 부르는데, 이 유한동형암호를 재부팅(bootstrapping)[8]이라는 과정을 통해 추가적으로 대수적 연산이 가능하게 되는 과정도 제시하였고 이를 완전동형암호(fully homomorphic encryption, FHE)라 부른다.

○ Gentry의 논문 이후 동형암호에 대한 기대와 관심이 커지면서 학계에서의 연구가 폭발적으로 이루어지게 되었다. 동형암호는 이후 MIT Technical Review에서 선정하는 2011년 10대 혁신기술[9]로 선정되었다.

□ 2011년 Brakerski-Gentry-Vaikuntanathan에 의해 발표된 BGV11 스킴에서는 곱셈시 생기는 잡음의 크기를 줄이는 기법이 제안되어 재부팅없이 가능한 곱셈 연산의 횟수를 암호문 길이의 선형(linear) 수준에서 지수(exponential) 수준으로 향상시켰다. 소위 modulus/key-switching이라 부르는 과정인데 이러한 과정을 도입한 스킴들을 2세대 동형암호라 부른다.

○ 2013년 Gentry-Sahai-Water 등이 곱셈과정의 잡음을 한층 줄이고 동형 곱셈연산을 할 때 비용이 많이 드는 재선형화(relinearization) 과정을 없애는 기술을 소개하였다. 이를 활용하는 동형암호를 3세대 동형암호라고 부르는데, 특히 평문을 1 비트로 제한하는 대신 매우 빠른 재부팅을 가능하게 한 Ducas-Micciancio14의 FHEW 스킴과 Chillotti-Gama-Georgieva-Izabachene16의 TFHE 스킴이 이에 속한다.

□ 2016년 Cheon-Kim-Kim-Song은 덧셈과 곱셈 외에 반올림이라는 세 번째 연산이 암호화 상태에서 가능한 혜안(HEaaN)이라는 동형암호를 발표하였다.

○ 기존의 동형암호도 반올림 연산을 수행할 수는 있지만 재부팅에 준하는 매우 복잡한 연산을 수행하여야 하나, 혜안 스킴에서는 이를 덧셈 수준으로 빠르게 처리하게 되어 산술연산이 대폭 개선되었다.

○ 실제로 비트당 30분이 걸리던 재부팅 연산이 2019년에는 0.5ms가 걸리게 되어

---

8) 암호문에 암호화된 비밀키를 이용하여 노이즈가 감소된 새로운 암호문 생성 후 다시 연산을 수행하는 설계 원리로 암호화된 키와 이중 암호화된 암호문을 입력한 후 복호화 알고리즘을 실행하여 새로운 암호문을 생성(Recrypt)하였다.

9) Homomorphic Encryption, Social Indexing, Smart Transformers, Gestural Interfaces, Cancer Genomics, Solid-State Batteries, Cloud Streaming, Crash-Proof Code, Separaing Chromosomes, Synthetic Cells.

300만 배의 개선을 이루었고 이는 매년 8배의 고속화에 해당하는데 이중 최근 3년간의 개선은 혜안 스킴에 힘입은 바가 크다.

○ 이에 따라 기계학습 등 근사연산을 활용하는 응용분야의 상용화가 가능해지게 되었고, 이를 4세대 동형암호로 부른다.

□ 2020년에는 미국 고등국방연구계획국(DARPA)에서 3,000만 달러 수준의 DPRIVE라는 과제를 공모 중인데 이 과제는 평문 연산 대비 동형암호의 연산이 10배인 하드웨어 가속기를 개발하는 것이다. 하드웨어 가속기를 활용하면 다양한 데이터 분석뿐만 아니라 자율주행 자동차 등 실시간 연산이 필요한 분야까지 더 많은 응용들이 현실화될 전망이다.

1.6 완전동형암호의 이해

가. 근사계산과 동형암호

□ 과학계산, 금융/의료 계산과 같이 실제 세상에서 요구되는 계산들은 주로 근사값의 계산이다. 이는 떠돌이소수점수(floating point number)로 표현되고 연산할 때 효율적이다.

○ 떠돌이소수점수는 부호부(sign)와 정규화된 가수부(normalized significand) 그리고 지수부(exponent)로 이루어져 있는데 예를 들어 -0.1101×2^4는 각각 -, 1101, 4가 그것이다. 부호부는 양수와 음수를 표현하는 한자리 수이고, 정규화된 가수부는 유효숫자(significand)를 표현하는데 1로 시작하도록 하며, 지수는 기저(base) 2의 몇 승인지를 표현한다.

○ 유효숫자가 10비트인 두 개의 수를 곱하면 유효숫자가 9비트인 수가 나오는데 이는 10비트를 두 개 곱해서 20비트를 얻은 후에 하위 11번째 이하 비트는 버리고 상위 9비트만을 택하는 과정이다. 이러한 과정을 통해 숫자가 계속 커지는 것을 방지하고 의미 있는 유효숫자만을 유지하며 계산을 지속할 수 있다.

□ Helib(IBM)과 SEAL2.4v(Microsoft)[10]의 경우 평문공간을 Z_k^n로 잡으면 평문은 $-k/2$ 이상 $k/2$ 미만의 정수들로 이루어진 길이 n인 벡터가 된다. 만일 $k=2$이면 이때는 모든 연산이 비트 단위로 이루어져서 편리한 반면 산술(arithmetic) 연산의 복잡도가 매우 높아지고, k가 커지면 큰 수 연산을 한 번에 처리할 수 있

10) HElib는 IBM에서 SEAL은 Microsoft에서 각각 개발 중인 동형암호 관련 오픈소스 프로젝트이다.

어 유리한 반면 비트를 다루는데 어려움이 있어 활용이 제한적이다. 이때 $k=2$인 경우를 비트연산 동형암호(bitwise homomorphic encryption)라 하고, $k>2$인 경우를 숫자연산 동형암호(digitwise homomorphic encryption)라 한다.

□ 숫자연산 동형암호는 $k=2$ 비트 덧셈과 곱셈을 한 번의 동형연산으로 처리할 수 있어 유리한 반면, 버림 연산과 같이 평문의 일부 비트를 추출하는 것은 매우 어렵다.

○ 이는 일반컴퓨터 연산과 달리 비트추출 인스트럭션(instruction)이 주어져 있지 않아서 원하는 비트추출 연산을 Z_k 위의 덧셈과 곱셈을 합성하여 만들어야 하고 이는 임의의 k차 함수를 대수적으로 계산하는 복잡도를 가지기 때문이다. 비트추출을 하지 않고 연산을 지속하면 순차적인 d번 곱셈 후에는 평문의 크기가 2^d까지 증가할 수 있어 재부팅 없이는 작은 d에 대해서만 곱셈이 가능하다.

□ 비트연산 동형암호에서는 각각의 평문비트가 따로따로 암호화되어 있기 때문에 상위 9번째까지의 비트에 해당하는 암호문만 계산하고 나머지는 생략함으로써 간단히 해결할 수 있다. 다만 이 경우 n 비트 수 두 개의 덧셈을 수행하였을 때 필요한 동형곱셈이 순차적으로 n 번이 되는 문제가 있다.

○ 동형암호가 수행할 수 있는 동형연산은 보통 동형용량(homomorphic capacity)이라 부르는 순차적인 곱셈연산 회수로 제한되며, 보통 최대 40회 정도를 잡는다. 이 경우 40비트 이상 숫자의 덧셈을 비트연산 동형암호로 연산하려면 재부팅을 해야 한다.

○ 실제 동형암호의 암호문은 보통 동형용량의 제곱에 비례하여 증가하므로 이 이상이면 재부팅을 수행하여 연산을 지속하는 것이 더 효율적이다.

□ 요약하면 숫자연산 동형암호는 큰 수의 덧셈과 곱셈을 수행하는 데 유리하나 곱셈의 순차적 반복 횟수가 커지면 재부팅을 해야 하는 단점이 있고, 비트연산 동형암호는 덧셈과 간단한 산술연산도 주어진 수의 커짐에 따라 효율이 급격히 떨어져서, 둘 다 곱셈 깊이가 40 이상인 연산을 수행하려면 재부팅이 필요하다.

○ 이에 따라 동형암호에서도 떠돌이소수점수 연산에서와 같이 효율성을 달성하기 위해서는 숫자단위 연산을 수행하지만 효율적으로 버림을 수행할 수 있어야 한다.

□ 서울대학교 천정희 등(2017)[11]은 이렇게 떠돌이소수점수 연산을 지원하는 동형암

11) Cheon, A. Kim, M. Kim과 Song(2017), <Homomorphic encryption for arithmetic of approximate

호를 근사동형암호(Approximate Number Homomorphic Encryption)라고 하였고, 근사동형암호는 이들이 개발한 HEAAN(homomorphic encryption for arithmetics of approximate numbers)이 유일이다.

○ HEAAN은 기존의 동형암호가 가지고 있는 평문의 덧셈, 곱셈 연산 이외에 버림 연산이 기본 연산으로 주어져 있다. 즉, 숫자연산 동형암호와 같이 큰 크기 정수를 원소로 하면서도 비트연산 동형암호처럼 동형 용량만큼의 순차적 동형연산을 수행할 수 있다.

○ 이를 이용하면 x의 역원계산이나 x의 지수승 등을 재부팅 없이 수행할 수 있으며, 나아가 삼각함수 $\sin x$나 초월함수 e^x, $\log x$ 등의 근삿값의 암호문도 x의 암호문으로부터 복호화키 없이 계산할 수 있다.

○ 이는 기계학습에서 널리 사용되는 활성화 함수인 시그모이드 $sigmoid(x) = 1/(1+e^{-x})$를 암호화한 상태에서 계산할 수 있어 동형기계학습을 효율적으로 수행하는 핵심기술이 된다.

나. 함수암호기술

□ 함수암호(functional encryption)는 특별한 기능을 가지는 암호화 기술로써 복호화키가 주어졌을 때 암호문의 일부 정보만 복호화하는 암호화 기술이다.

○ 공개키 암호방식의 함수암호는 공개키 암호와 동일하게 키 생성 과정에서 공개키와 비밀키가 생성되고 공개키는 암호화 과정에 사용되고 비밀키가 있으면 암호문을 복호화할 수 있다.[12)]

○ 추가되는 기능은 제한된 기능을 가지는 비밀키를 만들 수 있다는 데에 있다. 원래 비밀키와 함수 f가 주어지면 함수 f에 제한되는 비밀키 sk_f가 만들어지고 이를 이용하여 평문 m의 암호문을 복호화하면 m 대신 $f(m)$을 얻을 수 있다.

○ 예를 들어, $f(x)$가 $x = m^*$이면 1, 아니면 0을 출력하는 지표함수(indicator function)인 경우를 살펴보자. 암호문을 sk_f로 복호화하면 평문이 m^*이면 1, 아니면 0을 출력하여 sk_f로는 평문이 m*인지만을 확인할 수 있다. 이를 조금 일반화하면 서버에 생체정보를 주지 않고 생체인증을 하여 개인정보를 보호하는 생체인증에 이용할 수 있다.

---

numbers. <<Takagi T., Peyrin T.(eds) Advances in Cryptology – ASIACRYPT 2017>>. ASIACRYPT 2017. Springer, Cham. 409-437쪽.

12) 대칭키 방식의 함수암호의 경우 비밀키가 암호화키가 된다

○ 이때 $f(x)$는 입력되는 x가 고정된 m^*와 “가까울” 경우에만 1을 출력하는 함수로 정의된다. 여기서 가깝다는 것은 응용마다 달라질 수 있는데 홍채인증(iris authentication)에 쓸 경우에는 m^*와 x를 비트열로 적었을 때 해밍거리(hamming distance)가 짧다는 것이고 지문인증(fingerprint authentication)에 쓸 경우에는 가까운 특징점(minutiae)들 간의 유클리드 거리(euclidean distance)가 가깝다는 것이다.

○ 서버가 이런 f에 대해 sk_f를 가지고 있으면 암호화된 생체정보를 받았을 때 sk_f를 적용하여 등록된 생체정보와 가까운지 여부만을 판정할 수 있다.[13] 이에 비해 함수암호가 아닌 일반 공개키 암호를 이용하는 경우에는 미리 생체정보와 비밀키를 서버에 제공해야 되고 서버는 이 비밀키로 암호문을 복호화한 후 등록된 생체정보와 비교하는 과정을 거치므로 개인의 생체정보를 보호할 수 없다.

□ 이론적으로 함수암호는 임의의 함수에 대해 존재하지만 대부분은 매우 비효율적이다. 그러나 함수 f가 벡터 a에 대해 $f(x) = < a, x >$로 주어지는 내적함수인 경우에는 매우 효율적이다.

○ 따라서 선형회귀(linear regression)모형을 쓰는 기계학습에서는 함수암호를 실용적으로 이용할 수 있다. 또한 로지스틱 회귀(logistic regression)모형의 경우에도 우선 $< a, x >$를 복호화한 후에 sigmoid 함수를 취할 수 있으므로 어렵지 않다.

○ 실제 함수암호의 응용은 어떤 함수 f를 고르느냐에 따라 매우 다양하다. 이러한 함수 암호는 공개키를 ID로 쓸수 있는 신원기반암호(ID-based encryption)를 비롯하여 속성기반암호(attribute-based encryption), 술어암호(predicate encryption) 등을 만들 수 있다.

2. 동형암호의 활용

□ 2020년 1월 9일 데이터 3법[14]이 국회 본회의를 통과하면서 4차 산업혁명 시대의 핵심 자원인 데이터의 이용 활성화를 통한 신산업 육성이 국가적 과제로 대두되고 있다.

---

13) 좀 더 정확히는 이를 달성하기 위해서 sk_f에서 f에 대한 정보를 얻어낼 수 없어야 하는데 이런 함수암호는 model-hiding 성질을 만족한다고 말한다.

14) 개인정보보호법, 정보통신망 이용촉진 및 정보보호 등에 관한 법률(약칭 : 정보통신망법), 신용정보의 이용 및 보호에 관한 법률(약칭 : 신용정보법)

○ 개정된 데이터 3법은 다소 모호했던 개인정보의 판단 기준을 명확화하여, 개인정보 처리를 동반하는 사업 추진 시 혼란을 줄이고 익명정보의 이용을 활성화할 것으로 기대된다.

○ 암호화된 데이터의 연산이 가능한 동형암호는 빅데이터 활용과 개인정보 유출의 딜레마를 동시에 해결할 수 있는 특성으로 인해 데이터의 보호와 활용이 필요한 모든 분야로 적용이 확대할 수 있을 것으로 기대된다.

○ 현재 가장 가까운 동형암호의 응용분야는 마이데이터 사업[15]과 같은 프라이버시 보존 데이터분석으로 암호화를 통해 데이터에 대한 유출이 원천적으로 차단되고, 원문의 연산결과와 암호화된 연산결과가 동일하므로 데이터에 대한 다양한 분석이 가능하기 때문이다.

□ 금융, 의료, 교육 등 다양한 분야의 데이터를 보호하면서 활용하는 데 쓰일 수 있는데, 특히 2019년 12월 동형암호가 금융위원회의 혁신금융서비스로 지정되어 금융분야 데이터 결합을 위한 방법으로 사용될 수 있는 제도가 임시로 마련되었다.

○ 이에 따라, 신용평가회사인 코리아크레딧뷰로우(KCB)는 230만 명의 KCB 금융정보와 국민연금의 연금납부 데이터를 암호화한 상태로 결합하여 통계분석을 수행하였고 이에 따른 결과를 발표[16]하기도 하였다.

2.1 신용정보 계산

□ 신용도란 금융시장에서 채무를 얼마나 잘 갚을 수 있는지를 평가하여 나타낸 지표이다. 이러한 신용도는 국가기관인 금융위원회로부터 허가를 받은 신용조회회사들이 산정한다.

○ 우리나라 신용조회회사는 KCB(Korea Credit Bureau)와 NICE 신용평가정보가 있다. 이러한 신용조회회사는 총대출금액 등의 부채수준, 최장연체일수와 같은 연체정보, 금융거래기간 등을 신용평가항목으로 입력하여 복합적으로 신용평가 모형을 유도한다. 데이터 융합은 데이터 집합을 합치는 기법 또는 처리과정을 의미한다.

○ 데이터 융합을 통해 여러 데이터 속성을 통합하여 개별 데이터 속성이 제공하는

15) 개인이 직접 금융기관과 통신사, 병원 등에 있는 자신의 개인정보를 제 3의 업체에 전달해 새로운 서비스를 받도록 하는 사업이다.

16) KCB, 동형암호 기술로 국민연금 가입자 신용데이터 분석(연합뉴스, 2020.6.16.)

정보보다 정확하고 유용한 정보를 생성할 수 있다. 최근에는 금융기관에서도 이종기관의 데이터를 융합하여 데이터의 속성을 풍부하게 하여 보다 정확한 개인 신용도를 평가하려는 시도가 있다.

○ 한 예로 '비금융정보 신용평가 반영' 제도를 이용해서 통신요금, 공공요금, 건강보험료, 아파트 관리비 납부 데이터들을 신용평가 모델에 적용하여 신용도를 산정하려는 시도가 있었다. 이 외에도 방대한 비정형화된 데이터를 이용하여 신용도를 평가하려는 시도는 계속되고 있다.

○ 그러나 방대한 데이터인 만큼 그에 비례하여 개인의 민감한 정보를 포함하고 있는 경우가 많다. 개인정보가 수집되어 처리되는 환경에서 개인정보 유출이 발생하게 되면 개인의 사생활이 노출되고 또 급속하게 퍼져나가 통제 불능의 상황으로 전개될 수가 있다.

□ 개인정보를 안전하게 보호하면서 데이터를 활용하는 방안으로 지난 2016년 6월 개인정보 비식별화 조치 가이드라인이 제시되었다. 개인 식별정보는 개인을 직접적으로 알 수 있거나 유추하여 알 수 있는 모든 정보를 말한다.

○ 구체적으로 식별자(identifier)는 이름, 주소, 이메일 주소, 신용카드번호, 주민등록번호 등 직접적으로 개인을 식별할 수 있는 데이터이고, 식별자와 달리 자체로는 개인을 식별할 수 없지만 다른 데이터와 결합하여 간접적으로 추론하는 데 사용되는 속성인 우편번호, 성별, 나이 등을 준식별자(quasi-identifier)라고 한다.

○ 비식별화 조치 가이드라인의 핵심은 빅데이터의 활용성 증대와 개인정보 보호 강화, 개인의 사생활 침해를 방지하는 안전장치 마련과 비식별화된 정보를 산업적으로 활발히 활용하는 데 있다. 이를 위해 우리나라에서는 개인정보보호법에 따라 주민번호 암호화 의무조치 시행과 개인정보 비식별화 기법을 활용하고 있다.

□ 비식별화 기법으로는 식별자를 삭제하는 일반적인 기법인 가명처리, 총계처리, 데이터 삭제, 데이터 범주화, 데이터 마스킹 등의 방법과 재식별 가능성을 검토하는 기법인 프라이버시 보호 모델로 k-익명성(k-anonymity), l-다양성(l-diversity), t-근접성(t-closeness)이 있다.

○ 여기서 k-익명성은 공개된 정보들의 조합으로 개인을 추론할 수 있는 연결공격(linkage attack)에 대한 방어로 공개된 준식별자 속성이 같은 k개의 데이터를 확보하여 개인정보를 보호하는 모델이다. 그러나 k-익명성은 이종 데이터 병합으로 인해 민감 정보가 노출되는 사례가 발생하기도 한다.

○ 이를 보완하기 위한 방법으로 l-다양성이 제시되는데, l-다양성은 k-익명성을 만족하는 데이터에서 적어도 l개의 서로 다른 민감 정보를 가지고 있어야 한다는 모델이다. 준식별자의 속성이 k개가 같은 데이터 집합에서 적어도 l개의 서로 다른 민감 정보를 포함해야 한다는 것이다.

○ 또한 동일한 준식별자의 속성으로 익명화된 데이터 집합에서 민감한 정보의 분포와 전체 데이터 집합에서 민감한 정보의 분포를 t 이하로 제한하는 t-근접성은 두드러진 특이성을 제거하여 k-익명성과 l-다양성의 문제점을 보완하는 모델이다.

□ 그러나 이러한 보완에도 불구하고 식별 가능한 데이터를 수정하여 특정 개인을 유추하거나 식별할 수 없도록 하는 프라이버시 보호 모델로는 데이터의 유용성이 감소하는 문제점을 해결할 수 없다.

○ 이러한 문제점을 개선할 수 있는 가능한 방안으로는 프라이버시 보호 모형으로 데이터를 융합하고 신용도를 계산하는 프로토콜과 식별자를 메시지인증 코드로 가명화하고 각 속성은 동형암호화한 동형기계학습으로 신용도를 계산하는 것이다.

2.2 DNA 분석

□ 2003년 ‘휴먼게놈 프로젝트’를 통해 인간 유전자 지도가 완성되면서 유전정보를 이용한 의학발달의 초석이 마련되었으며, 이를 토대로 최근까지 유전자에 관한 연구가 활발히 진행되고 있다.

○ TCGA[17], ENCODE[18] 프로젝트 등 초대형 유전자 기반 프로젝트들이 성공적으로 완료됨에 따라 인간 유전자에 대한 지식이 축적되었고, 이에 2015년 2월 미국 오바마 대통령은 연두교서에서 정밀의료 이니셔티브(Precision Medicine Initiative, 총 2억 1500만 달러, 2016년 미국 R&D 예산 우선정책 선정)를 통해 개인의 유전체 및 의료정보, 생활 환경 및 습관 등 다양한 정보를 통합적으로 분석해 적절한 시기에 의료 서비스를 제공하려는 새로운 보건의료 패러다임을 제시하였다.

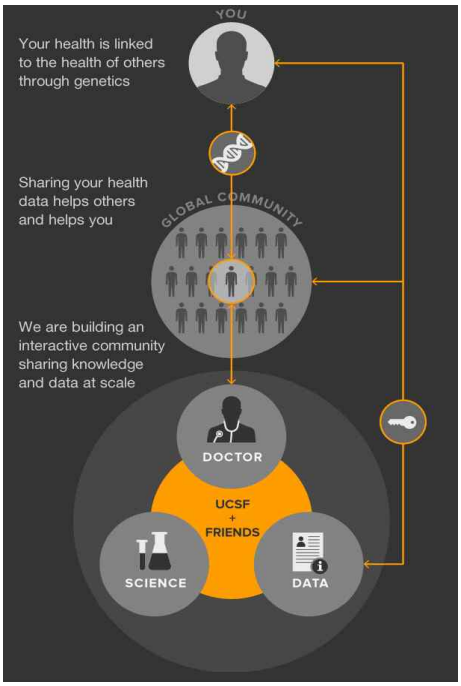
○ 인간 유전정보에 대한 이해도의 증가는 유전정보의 공통적 특성을 넘어 개개인의 차이를 파악하려는 시도로 이어지고 있다. 특히 의료분야에서 기존의 환자의

17) The Cancer Genome Altals (TCGA), http://cancergenome.nih.gov/

18) Encylopedia of DNA Elements (ENCODE), https://www.encodeproject.org/

증상 및 상태 기반의 치료에서 오는 환자와 유전자 간 약물 비적합성 부작용 문제를 해결해 줄 것으로 기대된다.

○ 하지만 이를 위하여 필수적인 개인의 유전정보 수집, 분석, 공유에 있어 한 개인의 식별 및 개인의 상태 정보를 모두 포함하는 유전정보 특성상 이를 보호하기 위한 새로운 암호화 기술의 연구가 요구된다. 정밀의료를 위해 수집된 개인의 유전정보는 병원, 연구기관 등과 공유되고 분석된다(<그림 2-6>).[19]



<그림 2-6> 유전정보 암호화를 통한 개인정보 보호 기반의 정밀의료 개념도

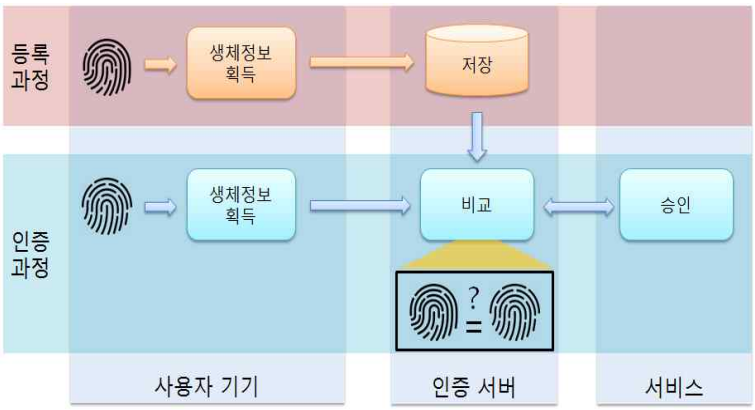
○ 이 과정에서 개인의 유전정보의 보호를 위한 암호화가 필요하고, 암호를 풀 수 있는 키(key)를 가진 개인만이 자신의 유전정보에 접근할 수 있다. 따라서 이러한 유전정보를 분석하려는 여러 기관들에게 개인키를 매번 제공하는 것은 정보 유출의 위험도가 증가하므로 암호화한 생태에서 여러 가지 비교 분석 연산을 가능하게 하는 동형암호화 기술이 요구된다. 이와 같은 정밀의료에 있어 동형암호화 기술은 키 유출을 최소화하고 개인의 유전정보를 보호할 수 있는 방법을 제공한다.

19) 미국 UCSF 대학의 정밀의료 서비스, http://meforyou.org

2.3 생체인증

□ 생체인증은 사용자의 신원을 검증하기 위하여 홍채, 얼굴, 지문 인식 등 개인의 고유한 생체 정보를 이용하여 본인을 확인하는 기술이다. 현재 많이 사용되는 비밀번호나 안전 토큰(Secure token)에 의한 인증방식은 망각/분실, 비밀번호 노출, 도난의 위험이 있다.

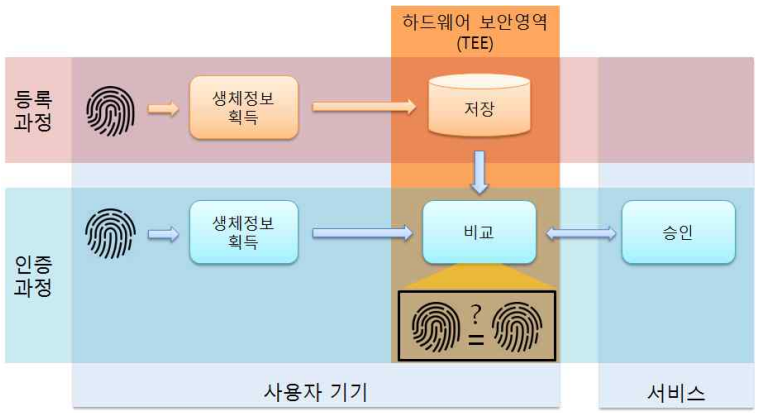
○ 생체인증 방식은 크게 서버 저장 방식과 사용기기 저장 방식(예. Fast IDentity Online, 이하 FIDO)으로 양분된다. 서버 저장방식을 보여주는 <그림 2-7>은 사용자의 생체정보를 인증서버(Authentication Server)에 저장한 뒤, 사용자가 자신의 생체정보를 전송하면 인증서버에 저장된 값과 일치하는지 여부를 확인하여 인증을 수행한다.[20]



<그림 2-7> 서버 저장 방식

○ 사용자 기기 저장방식을 나타내는 <그림 2-8>에서는 안전한 하드웨어 보안영역(Trusted Execution Environment, 이하 TEE)에 생체정보를 보관한 뒤 사용자가 자신의 생체정보를 입력하면 사용자 기기가 안전한 영역에 등록된 값과 비교한 결과를 서버에 전송하여 인증을 수행한다.

20) 생체정보에서 개인마다 다른 고유의 특징을 추출해 낸 정보를 '템플릿'이라고 부르는데 서버 저장 방식에서는 이러한 템플릿을 서버에 저장한 뒤 인증에 사용한다.

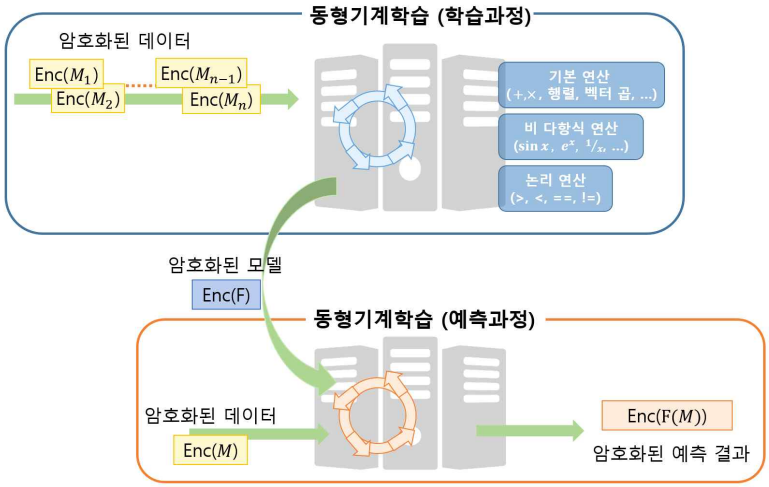


<그림 2-8> 사용자 기기 저장 방식

○ 현재 애플이나 삼성, 알리바바에서 생체인증을 이용한 제품과 서비스를 제공하고 있고, 국내 시중은행에서도 홍채인증과 지문을 이용하여 현금을 인출하는 서비스를 선보이고 있다. 동형암호는 침체된 생체인증 시장의 활성화에도 크게 기여할 것이다.

2.4 동형기계학습

□ 기계학습은 사람이 무언가를 학습하듯 컴퓨터에게 많은 데이터를 주고 스스로 학습하게 하는 기술이다.



<그림 2-9> 동형기계학습

○ 컴퓨터는 특정한 학습 과정을 통해 주어진 데이터를 가장 잘 표현하는 모델을 만들어낸다. 이를 통해 새로운 데이터에 대한 결과값을 예측한다. 일반적으로 기계학습은 주어진 데이터를 통해 학습하여 모델을 만들어내는 '학습단계(Training phase)'와 만들어낸 모델을 통해 새로운 데이터를 분석하는 '예측단계(Prediction phase)'로 구성된다.

○ 그런데, 기계학습을 진행하기 위해서는 많은 양의 데이터가 필요하고 민감한 개인정보를 많이 포함할수록 더 정확한 기계학습 결과를 얻을 수 있다. 그러나, 데이터 분석과정에서 민감한 개인정보가 필연적으로 노출되는 문제가 발생한다.

□ 최근 많은 환자 데이터를 소유하고 있는 대형병원이나 다수의 고객 신용정보를 가지고 있는 금융 업계들은 기계학습 기술의 활용에 큰 관심을 보이고 있다. 그러나 의료/금융 데이터들은 민감한 개인정보를 포함하고 있어 개인정보보호법 등의 규제를 받고 있고, 개인의 불신과 비동의 등으로 인해 개인정보 활용이 곤란한 실정이다.

○ 동형암호와 동형기계학습 기술은 이러한 개인정보보호 문제를 해결하고, 다양한 산업분야에서 민감한 데이터를 활용한 기계학습 기반을 마련해 줄 것이다.

○ 동형기계학습은 동형암호화된 데이터 혹은 모델에 대해 기계학습을 수행하는 것을 의미한다. <그림 2-9>와 같이 전체 학습 과정에서 데이터는 처음부터 끝까지 암호화되기 때문에 학습 과정에서 데이터의 유출을 막을 수 있다. 이때 동형암호는 암호화된 데이터 간의 산술 연산을 지원하며, 기계학습의 학습 및 예측 단계는 수학적 연산들로 구성되어 있기 때문에 동형기계학습은 충분히 가능하다.

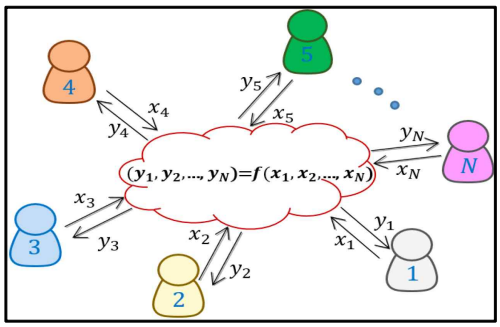
2.5 다자간 계산(Multiparty Computation)

□ 다자간 계산은 데이터 보호와 계산을 동시에 가능하게 하기 위해 연구되고 있는 암호 프로토콜이다. 다자간 계산은 참여자 간의 지속적인 정보교환을 바탕으로 한다.

○ 그런데, 다자간 계산은 참여자들이 모두 공동의 노력으로 얻고자 하는 결과에 맞게 프로토콜을 만들어야 하고 거기에 맞는 값을 서로 주고받아야 한다. 따라서, 목적으로 하는 것이 무엇이든 참여 인원이 많아질수록 계산량과 통신량이 최소 인원수의 제곱 혹은 그 이상으로 증가하기 때문에 다수 인원의 데이터의

분석에는 한계가 있다.

○ 또한, 효율성과 보안성(참여자가 공모할 경우 프라이버시 침해 가능) 간의 반비례 관계로 인해, '다수의 참여자가 공모하지 않는다'거나 '참여자 모두가 인정하는 믿을 수 있는 제 3자(Trusted Third Party)가 있다'는 다소 강한 가정이 필요하다.



<그림 2-10> 다자간 계산

○ 이에 비해, 동형암호는 다자간 계산의 참여자들이 공통으로 사용할 공개키만 공유하게 되면 각자 자신의 데이터를 암호화하여 다른 참여자나 제 3자에게 제공하고, 이후에는 기존의 다자간 연산과 달리 별도의 프로토콜을 설계할 필요도 없이 필요한 연산을 할 수 있다. 동형암호는 다양한 다자간 연산을 수행할 수 있는 틀이 될 수 있다.

2.6 군 보안통신 및 미사일 해킹 방지

□ 국방분야는 암호기술의 '텃밭'으로 불린다. 군용 통신에서 암호기술이 시작됐고, 전쟁을 치르면서 발전했기 때문이다. 제 2차 세계대전까지만 해도 암호기술은 군사적 목적으로 먼저 개발된 뒤 다른 분야로 확대됐다.

○ 동형암호는 군사적인 목적으로 개발된 것은 아니지만 뛰어난 보안 성능이 확인되면서 군 기술로의 적용이 활발히 검토되고 있다. 미국 국방부 고등연구계획국(DARPA)은 지난 2011년부터 지속적으로 동형암호 기술 연구개발을 지원해오고 있다.

○ 쿠르트 롤로프(Kurt Rohloff) 미국 뉴저지공대 컴퓨터과학과 교수팀은 여러 사람이 인터넷전화(VoIP)로 통신회의를 할 때 메시지를 동형암호로 암호화시켜 보호하는 방법을 개발해 2016년 12월 국제전기전자공학회에서 발행하는 보안분야 학술지(IEEE Transaction on Information Forensics and Security)에 발표했다.

□ 기존 기술로는 1대 1로 통신하는 경우에만 보안을 유지할 수 있다는 한계가 있다.

○ 최근 군에서도 여러 사람이 동시에 통화하면서 회의하는 경우가 많은데, 기존 보안기술로는 암호화된 상태에서 이들의 메시지를 합친 뒤 각자에게 전송하기가 어렵다.

○ 그렇게 하려면 일단 서버에서 암호를 푼 뒤 내용을 합치고, 이를 다시 암호화시킨 뒤 전송해야 한다. 하지만 서버의 경우 해킹위험이 도사리고 있어 중요한 군사기밀이 유출될 가능성이 있다.

□ 연구팀은 통신내용을 동형암호로 보호하면 서버에서 암호를 풀지 않은 상태에서 메시지를 합칠 수 있을 것으로 예상했다. 이를 위해 메시지를 동형 암호화시킨 뒤 복호화시키는 알고리즘, 서버에서 동형암호화된 메시지를 합치는 알고리즘 등을 개발했다.

○ 그런 뒤 이를 구현하는 아이폰용 소프트웨어를 만들어 실험했다. 2G와 3G, 4G LTE 등 다양한 방식의 통신시스템을 이용해 성능 시험을 실시한 결과 기존 기술과 비교해서, 송신 속도는 비슷하지만 수신 속도는 최대 5분의 1 수준까지 빨라졌다고 한다.

□ 헐리우드 블록버스터 영화에는 미사일이 발사된 뒤 목적지에 도달하기 전 다급히 미사일에 공격 취소 명령을 전송해 일촉즉발의 위기에서 벗어나는 장면이 종종 등장한다.

○ 실제로 최근 미사일 기술은 통신을 이용해 제어한다. DARPA는 이런 미사일 통신이 해킹당하지 않도록 동형암호로 보호하는 연구를 지원하고 있다. 이를 '코드난독화 기술'이라고 한다.

○ DARPA가 동형암호를 이용한 코드난독화 기술을 연구하는 이유는 미국이 다른 나라에 미사일을 판매하고 있기 때문이다. 판매되는 미사일 안에는 폭발 명령을 조종하는 알고리즘이 들어가 있고, 미사일을 구입한 나라나 미사일을 획득한 적군은 이를 해킹할 수 있다.

○ 결국 미사일을 팔았다가 국가안보에 심각한 위기를 초래할 수 있다. DARPA는 이를 해결하는데 동형암호가 효과적일 것으로 기대하고 연구를 지원하고 있다. 동형암호로 미사일 알고리즘을 암호화시키고, 공격 명령 역시 암호화된 상태로 전송하면 해킹 위험을 최소화시킬 수 있다.

□ 그동안 동형암호의 속도가 개선된 것은 암호학적 연구와 수학적 알고리즘의 개선에 힘입은 바가 크다. 이에 따라 암호문 상태의 연산이 평문 상태의 연산 대비 평균 1000배 빨라졌으나, 산업화가 활성화되기 위해서는 추가적인 개선이 필요하다.

○ 최근 소프트웨어적인 최적화에 더불어 동형암호 하드웨어 가속기가 연구되고 있다. DARPA는 최근 3000만 달러 수준의 DPRIVE 과제를 공모 중인데 이 과제의 목표는 평문연산 대비 10배 수준의 하드웨어 가속기 개발이다.

○ 하드웨어 가속기를 활용하면 다양한 데이터 분석뿐 아니라 자율주행 자동차 등 실시간 실행이 필요한 분야까지 동형암호의 더 많은 응용들이 현실화될 것으로 기대된다.

제 3 장

데이터 정보보호

□ 데이터 정보보호(data privacy)는 국내에서 아직은 생소한 분야이나, 4차 산업혁명, 디지털 뉴딜 정책 제안 등으로 그 중요성이 날로 부각되고 있다. 데이터 활용이 필수적인 시대이지만 개인정보보호 역시 중요하게 추구해야 할 가치이기 때문이다. 이 장에서는 데이터 정보보호를 소개한다. 이는 자료를 제공할 때 개인정보보호를 위하여 통계적으로 자료를 처리하는 기술 및 개념을 설명하는 것이다.

○ 1절에서는 데이터 유형에 따른 데이터 정보보호의 필요성을 알아보고, 2절에서는 노출위험과 정보손실을 설명한다. 다음으로 데이터 정보보호를 위한 구체적인 방법으로써 재현자료, 차등정보보호, 플랫폼 맞춤형 빈도표 비밀보호 알고리즘 개발, 분석결과물 반출 가이드라인 운영을 3절에서 7절까지 하나씩 차례로 다룬다. 구독통계 플랫폼 구축을 제안하는 본 보고서에서 데이터 정보보호 관련 개념 및 기술을 동형암호 활용과 더불어 주요 기술적 도구로서 제안하는 바이다.

○ 참고로 매스킹[21] 기법을 포함하여 노출위험과 정보손실을 측정하는 방법을 통틀어 통계적 노출제어(statistical disclosure control) 혹은 통계적 비밀보호(statistical data confidentiality)라고 부른다. 또한 최근에는 차등정보보호(differential privacy)[22]의 대두와 더불어 데이터 정보보호(data privacy)라는 용어를 해외에서 활발하게 사용하는 것으로 파악된다. 이 장에서는 노출제어, 비밀보호, 정보보호[23]를 유사한 개념으로서 구분 없이 사용한다.

---

21) 각 나라의 통계기관을 중심으로 자료를 제공할 때 자료 내에서 개인이 식별되지 않도록 자료를 처리하는 방법은 매스킹(masking)이라는 이름으로 연구되어 왔다. 반면 통계청을 제외한 우리나라 각 분야에서는 이러한 매스킹 기법을 비식별화라는 용어로 지칭하고, 데이터 일부를 '*' 등의 표식으로 가리는 매스킹 분야의 한 가지 기법(local suppression)을 매스킹이라 부르는 경향이 있다. 익명화(anonymization) 또한 비식별화와 유사한 의미로 쓰이고 있다.

22) 국내 컴퓨터 과학(compuater science) 분야 등에서 차등정보보호를 차분 정보보호, 차분 프라이버시 등으로 번역하기도 한다.

23) 컴퓨터 과학(compuater science) 분야에서 정보보호라는 번역이 보안(security)을 포괄하는 개념으로 사용되고 있으나, 이 장에서는 security와 구분하여 (semantic) privacy에 대한 번역으로서 정보보호라는 용어를 사용한다. 또한 정보보호(privacy)를 privacy/confidentiality protection의 줄임말로도 이해한다.

1. 데이터 정보보호의 필요성

□ 데이터를 공유하고 활용할 때 개인정보가 노출될 수 있다. 이 절에서는 데이터 유형에 따라 개인정보 노출이 어떻게 발생할 수 있는지 이해하고, 데이터 유형에 따른 정보보호(privacy/confidentiality protection) 전략을 논의한다.

1.1 데이터의 유형에 따른 개인정보 노출 가능성

□ 동일한 항목(변수)에 대한 값들이 개체별로 쌓여있는 자료를 마이크로데이터라고 한다. 컴퓨터 과학(computer science) 분야에서는 이를 테이블 데이터(table data)라고 부르기도 한다. 마이크로데이터는 행렬 형식을 취하며 각 행은 변수를, 각 열은 개체별 자료를 나타낸다. 각 열은 레코드(record)라고도 부른다. 모든 레코드에서 자료의 값이 숫자, 간단한 문자 등 일정한 규칙을 가지는 경우에는 마이크로데이터를 정형 자료라고 부르며, 그렇지 않으면 비정형 자료라고 부른다. 이 장에서는 비정형 자료는 논외로 하며, 정형 마이크로데이터에서 개인정보 노출 및 정보보호 전략을 논의한다.

□ 이 장에서는 마이크로데이터를 분석한 결과물을 매크로데이터라고 지칭하기로 한다. 매크로데이터의 한 예로 통계청에서 KOSIS 등을 통해 공표하는 다양한 집계자료를 생각할 수 있으며, 그 종류는 다양하다. 이 외에도 회귀분석 결과로 산출되는 계수와 같은 분석 결과물도 매크로데이터로 통칭한다. 매크로데이터의 경우 개인정보가 노출되지 않는다는 선입견을 가질 수 있으나, 생각보다 복잡한 상황과 개인정보 노출이 발생할 수 있음을 이해하자. 아래 예는 4개 변수에 대하여 총 N개(예. 5천만)의 개체 자료가 모인 마이크로데이터 및 이에 대한 총 K개(예. $2\times18\times5\times8=1{,}440$)의 빈도로 구성된 매크로데이터를 보여준다.

id	G	A	M	E	area
1	1	4	3	7	1068020001
...	2	17	5	1	2057040003
...	1	15	1	3	3070020402
...					
...					
N	2	18	5	8	5060020479

마이크로데이터 예제

case	G	A	M	E	frequency
1	1	1	1	1	$f_{1,1,1,1}=f_1$
2	2	1	1	2	$f_{2,1,1,2}=f_2$
...	...				...
K	2	18	5	8	$f_{2,18,5,8}=f_K$
total					N

매크로데이터(빈도표) 예제

가. 마이크로데이터에서 개인정보 노출 가능성

□ 마이크로데이터에서는 특정 레코드가 유일한 특징을 가지는 경우 해당 개체가 식별될 수 있고 나아가 민감한 값이 노출될 수 있다. 통계청 가계금융복지조사의 형식을 차용하여 작성된 다음의 가상 사례에서, k번째 개인은 전국에서 “유일한” 특징을 가진다고 가정하자. 이 개인을 잘 알고 지내는 이웃은 키변수인 지역, 연령, 성별, 직업의 값을 보고 자료에서 이 개인을 식별하며, 이 개인의 민감변수인 자산, 부채, 소득을 파악할 수 있다.

id	지역	연령	성별	직업	자산	부채	소득
1	서울	83	남	무직	28,874	1,784	570
2	부산	37	여	의사	1,158,547	27,478	8,745
…	…	…	…	…	…	…	…
k	**제주**	**57**	**여**	**군인**	**74,475**	**2,547**	**4,785**
…	…	…	…	…	…	…	…

○ 마이크로데이터에 유일한 개체가 포함되어 있을 때 노출위험이 높다. 만약 동일한 특징을 가지는 개체가 2개라면 둘 중 한 명이 다른 한 명의 정보를 알 수 있다. 따라서 동질한 특징을 가지는 개체가 얼마나 있는지를 기준으로 노출위험을 판단할 수 있으며, 이는 뒤에서 설명할 k-익명성의 개념과 연결된다. 참고로 표본자료는 한 개체가 여러 명을 대표하므로 유일성의 의미가 전수자료와 다르고, 상대적으로 안전하다고 판단한다. 샘플링은 하나의 매스킹 세부 기법이기도 하다.

□ 마이크로데이터를 공표할 때는 전통적으로 매스킹(비식별화) 처리를 실시한다. 위 예에서는 각 세 연령을 5세 혹은 10세 단위로 설정하고(global recoding, 전반적 재범주화), 유일한 k번째 개체의 직업을 N/A 처리(local suppression, 국소 감추기)할 수 있다. 이러한 매스킹 처리는 개인 식별 위험성을 감소시키나, 연령을 이용한 회귀분석을 어렵게 하거나 여성 군인에 대한 평균 소득 추정값을 산출하는데 영향을 미칠 수 있다. 즉 매스킹 처리를 통해 노출위험을 감소시키면, 정보손실이 발생하여 자료의 유용성이 떨어진다.

○ 매스킹을 넘어서는 마이크로데이터의 정보보호 방법으로는 뒤에 설명하는 재현자료 활용을 우선적으로 고려할 수 있다.

○ 마이크로데이터 매스킹 처리에 관한 보다 자세한 내용은 ‘개인정보 비식별화 조

치 가이드라인(관계부처 합동, 2016)’ 및 개정안(개인정보보호위원회, 2020), ‘마이크로데이터 매스킹 기법 실무 활용 안내서(박민정, 2015)’ 등을 참고할 수 있다.

나. 매크로데이터(분석 결과물)에서 개인정보 노출 가능성

□ 빅데이터 시대의 자료 대용량화, 자료 분석 플랫폼의 발전 등으로 인하여 마이크로데이터를 대신하여 이용자 맞춤형 매크로데이터 제공이 증가하는 추세이다. 예를 들어 통계청은 SGIS라는 통계지리정보서비스 플랫폼을 운영하고 있으며, 이용자는 전국민의 마이크로데이터를 다운로드 받아 분석하는 대신에 변수 범위를 선택하여 특정 지역의 분포(빈도표)를 파악할 수 있다. 다음 그림의 예는 어린이집을 개소하려는 이용자를 가정하여 소지역(집계구, 인구수 500명 내외) 단위로 해당 특징을 가지는 인구수를 표출한 예이다. 정보보호를 위하여 매스킹(5 미만은 N/A로 표기) 처리하여 매크로데이터를 제공한다. 이용자가 복지부의 보육복지포털 어린이집 현황과 함께 이용하면 어린이집 개소 관련 의사결정에 도움이 되는 정보를 얻을 수 있다. 이처럼 SGIS를 통하여 일반 국민이 인구주택총조사 자료 등을 손쉽게 이용할 수 있도록 돕는 것이 플랫폼 운영의 주요 목적이다.

SGIS 플랫폼에서 변수 선택 지도위에 빈도 표기

<그림 3-1> SGIS 플랫폼의 이용자 맞춤형 매크로데이터 제공 예

○ 만약 위의 인구주택총조사에서 성별, 각 세 연령, 교육수준, 혼인상태를 선택할 수 있다고 하자. 집계구 기준은 아파트 1개 동이 포함되는 경우도 많다는 것을 감안하여 다음의 가상 사례를 살펴보자. 이용자가 다음과 같이 조건을 선택하면서 어느 아파트 A동에서 해당 인구수 정보를 얻는다고 하자.

조건 선택	빈도수
남 + 50~65세 + 박사 + 기혼	5
남 + 50~65세 + 박사 + 기혼, 이혼	6
남 + 51~65세 + 박사 + 기혼, 이혼	5
남 + 50세 + 박사 + 이혼	N/A

○ N/A 처리로 개인정보 노출을 충분히 제어할 수 있을 것 같지만, 위의 가상 사례 결과를 정리하여 다음과 같은 빈도표를 만들 수 있다. 극단적인 예로 아침에 주차장에서 말다툼한 이웃집 남자 교수에 대한 혼인 상태를 파악할 수도 있고, 이러한 정보가 분쟁 과정에서 악용될 수도 있다. 따라서 빈도를 제공할 때 작은 값이 특정되는 것은 국내외 모든 통계기관에서 허용하지 않는다. 작은 값의 기준은 사회마다 다를 수 있다. 통계청은 대부분의 매크로데이터에서 사업체의 경우 3을, 개인 및 가구에 대하여 5를 기준값으로 사용하고 있다. 이 기준이 커지면 노출위험이 낮아지지만 제공받는 자료의 유용성도 감소한다. 즉 하나의 마이크로데이터에서 대량의 매크로데이터를 제공할 때, 노출위험 기준은 유지하면서 정보손실을 발생시키지 않는 것은 불가능하다.

	남, 51~65세, 박사	**남, 50세, 박사**	남, 50~65세, 박사 합(margin)
기혼	5	0	5
이혼	0	N/A → **1**	N/A → 1
합(margin)	5	N/A → **1**	6

○ SGIS에서는 연령을 각 세로 제공하지 않고 5세 단위로 제공하며, 현재 교육수준 변수와 혼인상태 변수를 제공하지 않는다. 즉 매스킹 기법 중에서 전반적 재범주화(global recoding) 및 감추기(suppression) 처리를 실시하고 있다. 노출위험을 고려하면 항목 제공범위 확대를 쉽사리 결정할 수 없음을 통계 이용자는 이해할 필요가 있다. 자료 제공기관 역시도 이용자가 요구한다고 해서 제공범위를 확대하면, 예기치 못한 개인정보 침해 문제를 발생시킬 수 있으니 주의가 필요하다.

○ 위의 간단한 예를 통하여 매크로데이터를 제공할 때 정보보호(semantic privacy) 문제는 동형암호 등 최신 보안(security) 기술을 사용한다고 해도 해결하기 어렵다는 것을 알 수 있다. 암호화된 두 개의 자료를 연산한 결과를 복호화하면 매

크로데이터의 참값을 얻을 수 있으며, 대량으로 제공되는 참값 매크로데이터에서 정보보호 실패 현상은 동일하게 발생하기 때문이다.

○ 매스킹을 넘어서는 매크로데이터의 정보보호 방법으로는 뒤에 설명하는 차등정보보호 적용, (차등정보보호를 만족하는) 재현자료 활용, 맞춤형 알고리즘 개발, 반출물의 통계적 노출제어 가이드라인 운영 등을 고려할 수 있다.

○ 다음 추가 예제를 통해 매크로데이터에서 정보보호 문제를 좀 더 이해하자.

□ 한편 통계기관은 데이터센터를 운영하여 자료를 제공할 수도 있다. 이용자는 자신의 자료를 센터로 반입하여 센터가 제공하는 자료와 연계하여 분석하고, 분석결과물을 반출할 수 있다. 반출은 전문 직원의 심의를 거쳐야 한다. 이러한 매크로데이터의 제공은 센터 방문이라는 이용자의 비용과, 전문 직원의 심의라는 자료 제공기관의 비용 발생을 감수한다면 바람직한 자료제공 방식으로 여겨질 수 있다. 그러나 대용량 마이크로데이터를 분석하고 얻어진 대용량 매크로데이터를 반출하는 사례가 증가하여 자료 처리 한계를 넘어서는 상황을 가정하고, 특정 개체에 대한 민감정보 노출 가능성을 검토할 필요가 있다. 다음의 사례를 살펴보자.

	업종	급여여부	지역	고객수	민감변수_평균	분산	중앙값	...	
반출요청	**0 (1+2)**	**1_급여**	**A**	**42**	1.93	0.07	1	...	
	0 (1+2)	2_비급여	A	103	2.84	1.3	3	...	집계표 길이
	1	2_비급여	A	7	2	0.7	2	...	
	2	**1_급여**	**A**	**41**	1.9	0.28	1.5	...	**13만**
	2	**2_비급여**	**A**	**3**	2.91	0.82	1.5	...	**이상**
	...	...	...	...	...	...	...	...	

○ 파란색으로 표기된 두 분석결과 차이에 의하여 다음과 같이 어느 고객의 정보가 노출될 수 있다. 현실적으로 위와 같은 형식의 대용량 반출요청 자료를 검토하면서 노출위험을 제어하기는 어렵다.

	업종	급여여부	지역	고객수	민감변수_평균	분산	중앙값	...
사례 1	**1**	**1_급여**	**A**	**1**	**3.16**	...	...	...

○ 붉은색으로 표기된 행은 고객 3명에 대한 매크로데이터를 나타낸다. 반면 반출

을 요청하는 분석 결과 항목은 평균, 분산, 중앙값 등이다. 즉 레코드의 개수에 비하여 반출을 요청하는 정보의 수가 많다. 따라서 간단한 계산을 통하여 각 개체에 대한 “마이크로데이터를 복원”할 수 있다. 이는 개인정보보호를 위하여 마이크로데이터를 제공하지 않고 센터 내에서만 이용 가능하게 하려는 통계기관의 노력이 물거품이 될 수 있음을 보여준다.

사례 2	업종	급여여부	지역	고객수	**민감변수_평균**	**분산**	**중앙값**	...
	2	2_비급여	A	**3**	2.91	0.82	1.5	...

○ 데이터센터를 이용한 반출물에 대하여 전문 직원이 심의할 경우, 위와 같은 사례들을 방지할 수 있어야 하며 적절한 비용으로 그러한 작업이 가능해야 한다.

1.2 데이터의 유형에 따른 정보보호 전략

□ Reiter(2004)는 데이터를 제공할 때 필요한 3대 요건을 다음과 같이 정리하였다.

<표 3-1> 데이터 제공의 3대 요건(Reiteer, 2004)

- 응답자의 식별 정보나 민감한 정보를 알려는 외부 시도에 대한 **안전성**
- 광범위한 통계적 분석을 뒷받침할 수 있는 **정보의 충분성**
- 보편적인 통계적 방법론들을 사용하는 **이용자의 접근/분석 편의성**

○ 그러나 자료제공 방식에 따른 개별정보 노출제어 처리의 한계는 다음과 같이 정리된다. 노출위험을 축소하기 위한 매스킹(비식별화) 처리가 정보손실을 유발하여 노출위험과 정보손실 사이에 상충관계가 있다. 따라서 위의 세 요건을 모두 충족하여 통계자료를 제공하는 방법은 현실적으로 존재하기 어렵다.

<표 3-2> 자료제공 방식에 따른 노출제어 처리의 한계

		안전성	유용성	편의성
• 노출위험 축소 기법 적용	☞	○	**X**	○
• 정보손실 축소를 위한 원자료 제공	☞	**X**	○	○
• 안전성 확보 및 충분한 정보	☞	○	○	**X**

○ 이에 다음과 같이 이용자의 유형에 따른 자료 배포 전략을 모색할 필요가 있다.

<표 3-3> 이용자에 따른 자료 배포 전략/기법

<table>
<tr><td>이용자</td><td colspan="2">심층 이용자</td><td colspan="4">불특정 다수</td></tr>
<tr><td>종류</td><td colspan="2">물리적 노출제어</td><td colspan="4">통계적 노출제어(SDC)</td></tr>
<tr><td rowspan="2">배포
형식</td><td>microdata</td><td>macrodata
(analysis output)</td><td>microdata</td><td colspan="3">macrodata
(analysis output)</td></tr>
<tr><td>인가파일</td><td>가공 통계표
원격 접속(RAS)
데이터센터(RDC)</td><td>공공이용파일</td><td colspan="3">온라인 분석 시스템
(OAS)</td></tr>
<tr><td>전략/
기법</td><td colspan="2">접근제어 및 결과통제
➢ 결과물에 SDC 처리 필요</td><td>매스킹
기법</td><td>재현자료
(SD)</td><td>플랫폼 맞춤형
(ex. SGIS)</td><td>차등정보보호
(DP)</td></tr>
<tr><td>현재</td><td colspan="2">(예) 40%</td><td colspan="4">(예) 60%</td></tr>
<tr><td>↓</td><td colspan="6"></td></tr>
<tr><td>미래</td><td>(예) 20%</td><td colspan="5">(예) 80%</td></tr>
</table>

○ 이제 다음 절에서 노출위험과 정보손실을 기준으로 하여 비밀보호 방식을 평가하는 개념을 살펴보고, 이어 현재 사용가능한 비밀보호 기법들을 차례로 알아보자.

2. 노출위험과 정보손실

□ 이 절에서는 비밀보호 기법 적용 결과를 평가(assessment)하기 위하여 반드시 필요한 노출위험 측도와 정보손실 측도에 대하여 알아보고, 위험-유용성 지도(risk utility map)를 통해 각 기법을 평가하는 방식을 설명한다.

2.1 매스킹(비식별화) 처리와 KLT 모형

□ 전통적으로 개인이 식별된다는 것은 모집단 내에서의 개체 유일성에 근거한다. 따라서 데이터 익명화, 비식별화란 먼저 식별자(identifier)를 제거하여 유일한 정보를 삭제하고, 준식별자(quasi-identifier)를 변형 및 조작하여 유일한 개체를 남기지 않는 것을 목표로 한다. 이를 통해 민감정보를 공유하거나 공개하는 것을 방

지하고자 한다.

○ 식별자는 특정 개인과 1:1 대응이 되는 변수를 말하며, 주민등록번호 등을 생각할 수 있다. 자료를 공개할 때 반드시 삭제한다.

○ 준식별자는 그 자체로서는 식별자가 아니지만 다른 정보와 결합하면 개인을 특정할 수 있는 변수들을 말한다. 앞 절에서 키변수라고 언급한 지역, 성별, 연령, 직업 변수 등을 생각할 수 있다.

○ 민감정보는 개인이 드러내기 원하지 않는 정보를 말하며 그 기준은 사회마다 다를 수 있다. 예를 들어 소득 수준, 부채 총액, 질병번호 등의 변수를 생각할 수 있다. 다만 궁극적으로는 민감정보를 다시 준식별자로 사용하여 자료를 공격하는 경우도 고려할 필요가 있다. 즉, 준식별자와 민감정보의 구분은 자의적일 수 있다.

○ 다음은 여러 논문에서 언급되는 예제이다. 왼쪽은 원자료를, 오른쪽은 매스킹 처리하여 2개 변수(ZIP, Age) 기준으로 익명성을 확보한 자료를 나타낸다.

No	ZIP	Age	Desease
1	47677	29	AIDS
2	47672	22	Ebola
3	47678	27	Cancer1
4	47905	53	AIDS
5	47909	52	Ebola
6	47906	57	Cancer2

원본자료

No	ZIP	Age	Desease
1	4767*	≤40	AIDS
2	4767*	≤40	Ebola
3	4767*	≤40	Cancer1
4	4790*	≥50	AIDS
5	4790*	≥50	Ebola
6	4790*	≥50	Cancer2

비식별화/익명화/매스킹 처리된 자료

□ (k-익명성) 노출위험을 평가하는 가장 널리 알려진 기준으로 k-익명성(anonymity)이 있다. 앞의 예에서 우편번호와 연령대를 매스킹(global recoding) 처리하여 두 변수를 기준으로 익명성 3을 충족하였다.

□ (ℓ-다양성) k-익명성은 동질성 공격에 취약한 단점이 있다. 아래 예 왼쪽에서 1~3번 개체는 익명성 3을 충족하지만 민감변수인 질병 항목에서 모두 동일한 속성을 가진다. 따라서 1~3번 개체 모두 민감한 정보가 드러나게 된다. 이를 방지하기 위한 것이 ℓ-다양성(diversity)이다. 이는 준식별자가 동일한 k개의 개체가 적어도 ℓ개의 다양한 속성을 가지고 있어서 그룹 내에서 동질성이 존재하지 않음을 의미한다. ℓ-다양성을 만족시키려면 보통은 더 많은 개체를 매스킹 처리한다. 따라서

ℓ-다양성을 충족하기 위해서는 정보손실이 추가로 발생한다.

No	ZIP	Age	Desease
1	4767*	≤40	AIDS
2	4767*	≤40	AIDS
3	4767*	≤40	AIDS
4	4790*	≥50	Ebola
5	4790*	≥50	Ebola
6	4790*	≥50	Ebola

익명성 충족, 다양성 미충족

No	ZIP	Age	Desease
1	4767*	≤40	AIDS
2	4767*	≤40	Ebola
3	4767*	≤40	Cancer1
4	4790*	≥50	AIDS
5	4790*	≥50	Ebola
6	4790*	≥50	Cancer2

익명성과 다양성 충족

□ (t-근접성) 만약 속성이 좁은 범위에 다르게 분포해 있다면 ℓ-다양성만으로는 여전히 프라이버시 침해가 발생할 수 있다. 이를 보완하기 위한 것이 t-근접성(closeness)이다. 이는 준식별자가 동일한 k개 개체의 속성 분포와 전체 자료에서 해당 속성 분포와의 차이를 t 이하로 만드는 것이다. t-근접성까지 만족시키려면 자료에서 정보손실이 추가로 발생한다.

□ 이상을 KLT 모형이라 하며 기준을 추가하여 충족시킬수록 노출위험은 낮아지나 정보손실이 증가하여 자료의 유용성이 떨어진다. KLT 각 모형을 어느 기준으로 충족하는지를 측정하여 매스킹 처리한 자료를 평가할 수 있다.

○ k-익명성을 예로 들면 익명성이 1인 개체가 몇 개, 2인 개체가 몇 개, 3 이상인 개체가 몇 개와 같은 방식으로 파일 전체에 대한 노출위험을 측정할 수 있다. 또한 각 개체에 대해서는 익명성이 얼마인지 표기할 수 있다. R패키지 sdcMicro나 ARX 등의 프로그램을 이용하여 이러한 계산을 손쉽게 할 수 있다.

2.2 통계적 모형을 이용한 유일성 기준 노출위험 측정

□ 유일성을 기준으로 노출위험을 측정하는 통계적 모형을 소개한다. 참고로 연속형 변수의 경우에는 이러한 통계적 모형이나 앞의 k-익명성 등을 사용할 수 없다. 유일성 기준 방식의 노출위험 측정은 범주형 변수를 대상으로 한다.

○ 조사자료의 경우 일반적으로 가중값이 존재한다. 전수자료의 경우에는 가중값을 1로 처리하면 된다. 가중값을 감안하여 조사된 표본자료에서 유일한 개체가 모집단에서 유일할 확률을 다음과 같이 나타낼 수 있다. 소문자는 표본자료 빈도

를, 대문자는 모집단의 빈도를 나타낸다.

$$r_{1k} = P(F_k = 1 | f_k = 1)$$

id	G	A	M	E
1	1	4	3	7
2	2	17	5	1
3	1	15	1	3
...	...			
...	...			
n	2	18	5	8
total				

표본 마이크로데이터

case	G	A	M	E	frequency
1	1	1	1	1	f_1
...	...				...
k	...				f_k
...	...				...
K	2	18	5	8	f_K
total					n

표본 자료에 대한 빈도표

○ 위 노출위험의 계산은 음이항 모형, 로그 선형 모형 등 다양한 모형을 이용할 수 있으며, R패키지 sdcMicro는 네덜란드 통계청이 제안한 음이항 모형을 이용하여 개체별 노출위험 및 파일 전체에 대한 노출위험의 평균과 합계를 제공한다.

□ 참고로 연속형 변수의 노출위험은 매스킹 처리 전후 값의 변화를 반영하도록 정의할 수 있다. 사전에 참값을 기준으로 일정한 구간을 정의하고, 매스킹 처리 이후의 값이 해당 구간에 없으면 노출위험을 0으로, 구간에 있으면 노출위험을 1로 정의한다. 이제 전체 레코드에서 1의 비율로 노출위험을 측정한다. 이는 엄밀한 수준의 노출위험 측도라고 하기는 어렵지만, 직관적으로 이해하고 계산하기에 용이하여 활용되고 있다. 또한 R패키지 sdcMicro에서 사용하고 있는 방법이기도 하다.

○ 보다 자세한 내용은 '가계금융복지조사 매스킹 방안 평가(박민정, 2014)' 등을 참고할 수 있다.

2.3 침입자 의사결정론을 이용한 노출위험 측정

□ 앞에서 설명한 유일성에 대한 통계적 모형은 키변수 조합에 따라 다르게 계산되는 노출위험 측도이며, 외부인이 가지고 있는 자료의 특징을 반영하지 못한다. 또한 통계기관이 (연속형) 민감변수에 시행한 매스킹 처리를 평가하지 못한다. 이에

의사결정이론(decision theory)을 이용하여 외부인의 표적 t가 통계기관이 공표하는 자료에 있는지 없는지를 판단하는 확률로 노출위험을 정의하기도 한다. 이는 다음의 식으로 표현할 수 있다. Z는 매스킹 처리된 자료를 나타내며 다음 식은 자료의 특정 j번째 개체가 표적 t이라고 판단할 확률을 나타낸다.

$$\Pr(J=j|\vec{t},Z) \quad j=1,...,n$$

○ 이러한 방식은 후에 설명하는 재현자료의 노출위험 측정에서 적용 가능한 방법으로써 모색되고 있기도 하다. 보다 자세한 내용은 '가계금융복지조사 매스킹 방안 평가(박민정, 2014)' 등을 참고할 수 있다.

2.4 정보손실 측정

□ 정보손실의 측정은 노출위험에 비하여 용이하다. 기존의 자료와 달라진 정도를 측정하여 정보손실로 정의한다. 다만 그 기준으로 자료의 값, 자료의 구조를 표현하는 공분산 벡터, 주성분 벡터 등을 사용할 수 있다. 또한 분포의 변화를 측정하거나, 매스킹 전후 신뢰구간 중복 등을 이용하기도 한다.

	평균제곱오차 Mean square error	평균절대오차 Mean abs. error	평균변동 Mean variation
자료행렬 $X-X'$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$
공분산 벡터 $V-V'$	$\frac{\sum_{j=11\le i\le j}^{p}\sum(v_{ij}-v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=11\le i\le j}^{p}\sum\lvert v_{ij}-v'_{ij}\rvert}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=11\le i\le j}^{p}\sum\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p+1)}{2}}$
...	...	...	...

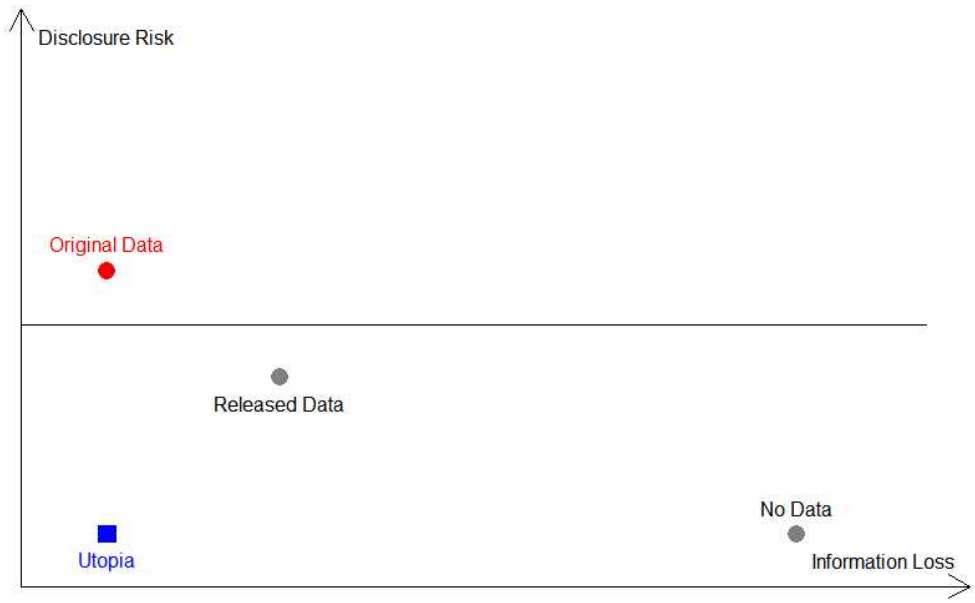
○ 보다 자세한 내용은 '가계금융복지조사 매스킹 방안 평가(박민정, 2014)' 등을 참고할 수 있다.

2.5 노출위험과 정보손실 평가

□ 전통적으로 각국 통계기관에서는 매스킹 처리한 자료의 노출위험과 정보손실을

측정하여 적절한 비식별화 방안을 선정하고자 하였다. 그러나 노출위험과 정보손실을 측정하는 실무적인 방식이 모두 매스킹 전후 차이를 이용하므로, 각 측도가 서로 상충관계(trade off)에 있다. 즉 노출위험을 낮추기 위하여 자료를 많이 바꾸면 자료의 구조에 변형이 생기며 정보손실이 커진다. 매스킹 처리의 목표는 노출위험과 정보손실을 동시에 낮추는 것이나, 매스킹으로 하나의 측도 값을 낮추면 다른 측도의 값이 증가한다.

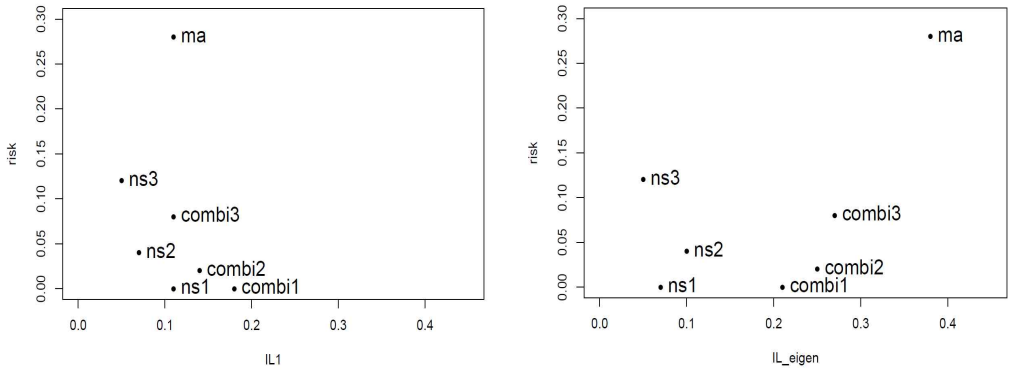
○ 이를 다음과 같이 그림으로 나타낼 수 있다(Hundepool et al., 2012). 가로축에 정보손실을, 세로축에 노출위험을 표기하였다. 원자료(Original Data)를 정보손실 없이 노출위험만 감소시켜 새로운 자료(Utopia)를 만들고 싶으나 노출위험을 낮추면 정보손실이 발생하여 공개하는 자료(Released Data)가 생성된다. 아무 자료도 제공하지 않으면(No Data) 개인정보 위험이 전혀 없으나, 정보손실이 최대로 커지면서 자료의 유용성을 확보할 수 없다.



<그림 3-2> 위험-유용성 지도(노출위험과 정보손실)

○ 실제로 가계금융복지조사 마이크로데이터에 다양한 매스킹 처리를 한 후 이를 평가하였다. 다음 그림은 그 결과를 보여준다. 왼쪽은 정보손실 측도로 자료 사이의 거리를, 오른쪽은 고유값을 이용하였다. 만약 ns2와 combi3 방식을 비교한다면 양쪽 그림 모두 ns2 방식이 효율적이라고 평가할 수 있다. 그러나 동일한 노출위험 측도에 대하여 ns1와 ns2를 비교한다면 왼쪽 그림의 방식으로는 상충관계를, 오른쪽 그림의 방식으로는 ns1이 우수함을 나타낸다. 이렇듯 매스킹 처

리로는 노출위험과 정보손실 사이의 상충관계가 존재할 뿐만 아니라 측도가 바뀔 때마다 달라지는 결과에 대한 의사결정 논란 등의 문제가 발생할 수 있다. 노출위험을 측정하는 기준이나 정보손실을 측정하는 기준이 수학적으로 다양할 수 있음을 주지하기 바란다.



<그림 3-3> 노출위험-정보손실 지도

○ 매스킹(비식별화) 처리는 근본적인 노출위험과 정보손실 사이의 상충관계를 해결하지 못하며, 각 측도에 대한 기준 또한 모호하다. 따라서 패러다임을 전환한 새로운 개념과 방법론이 발전하였다. 최신 방안으로서 재현자료(synthetic data)와 차등정보보호(differential Privacy)를 다음 절에 소개한다. 이어서 그 이후에는 통계청에서 개발한 대용량 빈도표 비밀보호를 위한 알고리즘 및 데이터센터 반출물에 대한 통계적 노출제어 가이드라인을 소개하여 참고할 수 있도록 한다.

○ 이 절이 설명하는 자료 평가 방식 전체적인 흐름을 좀 더 이해하기 원하는 독자는 다음 연구보고서의 서론 등을 참고할 수 있다.

- 통계적 노출제어 최신동향 검토 – 차등정보보호를 중심으로-(박민정, 2016)
- 차등정보보호에 관한 연구(박민정 외, 2018)
- 차등정보보호 적용 실험 연구(박민정 외, 2019)
- 통계데이터센터 DB를 활용한 재현자료 생성 방법 연구(박민정 외, 2020)

---

3. 재현자료 활용

□ 이 절에서는 재현자료를 소개한다. 재현자료는 다중대체 방식을 이용하도록 제안되었으며(Little, 1993; Rubin, 1993), 최근 딥러닝 기법을 활용한 방법도 제시되었다(Park et al., 2018). 재현자료의 노출위험은 통계적 방식으로 측정할 수도 있고, 차등정보보호를 만족하는 재현자료를 생성하려는 노력도 있었다(Machanavajjhala, 2008). 현재 학계에서 관련 연구가 활발히 진행 중이다.

○ 재현자료는 원자료를 대체하여 제공하기도 하며, 원자료 분석을 위하여 데이터센터를 방문하는 이용자에게 사전분석용으로 제공하는 용도로도 사용될 수 있다.

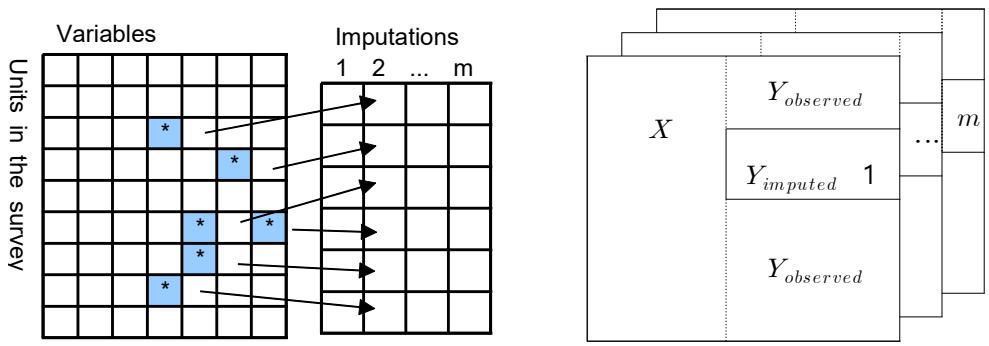
3.1 다중대체와 재현자료

□ 매스킹 처리의 한계를 극복하기 위하여 다중대체(multiple imputation)를 이용하여 재현자료(synthetic data)를 생성하는 방식이 제안되었다. 먼저 다중대체를 소개한다.

○ 단일대체는 자료에서 결측값이 있을 때 이를 적절한 하나의 값으로 대체하여 이용자가 완전한 자료를 이용할 수 있도록 한다. 그러나 다음과 같이 통계량의 분산을 계산할 때 분모에서 실제보다 큰 결측값이 포함된 개수를 사용하게 되며, 이는 분산 추정에 영향을 미친다. 이러한 단점을 극복하기 위하여 대체하는 값의 불확실성을 고려한 다중대체가 제안되었다.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

○ 다중대체는 베이지안 사후예측분포로부터 결측값이 가질 수 있는 다수의 자료를 생성한다. 따라서 결측값을 대체하는 파일이 여러 개 생성된다.



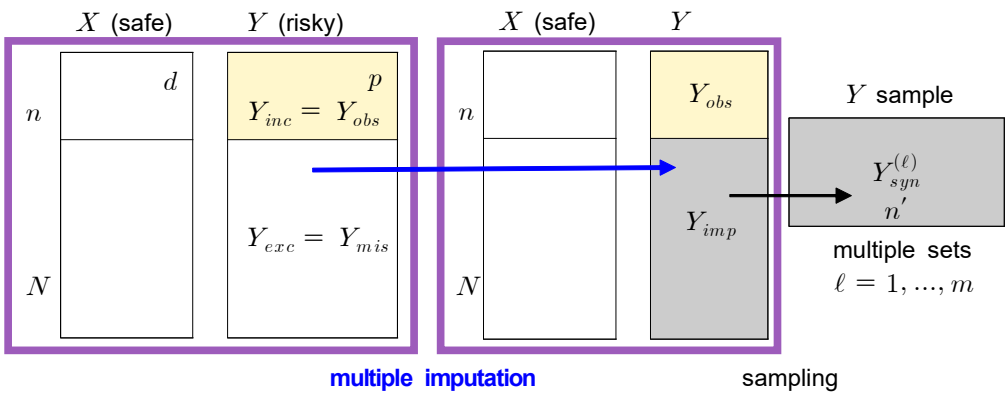
<그림 3-4> 다중대체를 이용한 재현자료 생성 예

○ 이용자는 다음과 같은 결합규칙을 통해 최종 통계량 및 분산을 계산한다. Q 및 U는 각 세트에서 구한 통계량과 분산을, B는 각 세트 사이의 변동을 나타낸다.

$$\overline{Q}_m = \frac{1}{m}\sum_{\ell=1}^{m} Q^{(\ell)}, \qquad T_m = (1+m^{-1})B_m + \overline{U}_m$$

□ 다중대체에서 재현자료를 생성하는 아이디어는 다음과 같다. 먼저 모집단에서 조사되지 않은 개체($Y_{exc} = Y_{mis}$)를 결측값으로 간주하여 다중대체 한다. 다음으로 다중대체로 생성한 자료에서 표본을 추출하여 조사한 자료를 대신하여 제공한다.

○ 이러한 아이디어를 도식화하면 다음과 같다.



<그림 3-5> 다중대체 시 재현자료 생성 아이디어

○ 위의 그림에서 X에 해당하는 변수들은 표본 설계용 변수 등 이용자가 알고 있는 변수를 말한다. Y는 민감변수들이다. 먼저 관측되지 않은 자료들을 결측값으로 간주하고, 관측된 자료에서 얻은 베이지안 사후예측분포로부터 자료를 생성하여 관측되지 않은 모집단을 채우고, 모집단에서 표본추출하여 만든 복수의 자료 세트를 이용자에게 제공한다. 따라서 이용자에게는 원본 자료가 전혀 제공되지 않는다. 이용자는 각 자료 세트로 원하는 분석을 수행하고 결합규칙을 이용하여 최종 통계량을 계산한다.

○ 이와 같이 모든 변수 및 모집단에 대하여 재현자료를 생성하는 방식을 완전 재현자료(fully synthetic data)라고 한다. 완전 재현자료에는 원자료를 사용하지 않으며, 원자료는 재현자료를 생성하는 모형을 구축하는 용도로만 사용된다. 그러나 정규분포가 아니라면 모수적 모형으로 완전 재현자료를 생성하면 정보손실이 크게 발생한다. 이에 비모수적 모형의 활용이나 노출위험이 낮은 일부 변수 혹

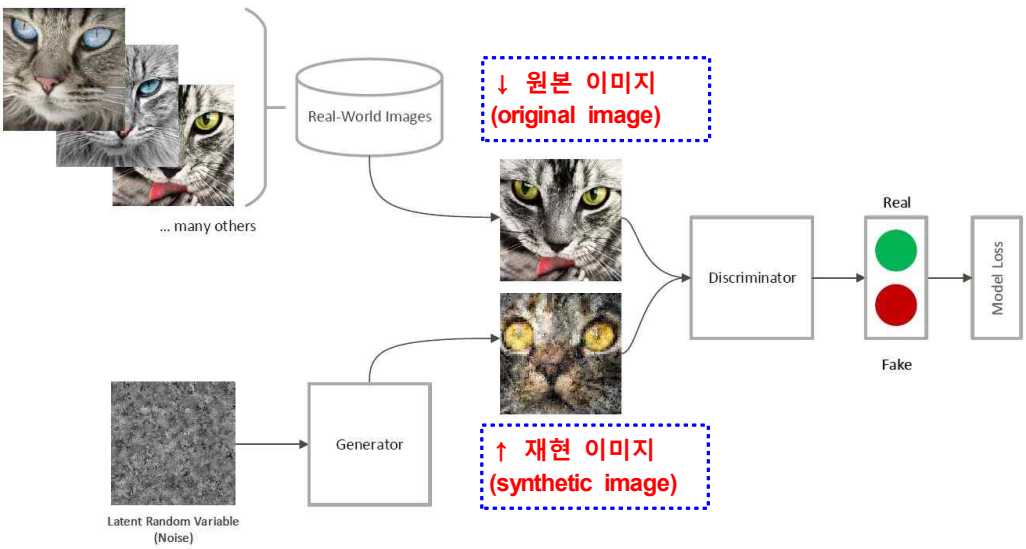
은 레코드만 생성하는 부분 재현자료(partially synthetic data) 방식이 제안되었다. 각 방식에 따라 결합규칙에 약간의 차이가 있다.

3.2 딥러닝을 이용한 재현자료 생성

□ 최근 적대적 생성망(Generative Adversarial Networks, GAN)을 이용한 재현자료 생성 기법이 제안된 바 있다(Park et al., 2018).

○ GAN의 주요 아이디어는 두 개의 머신을 경쟁시킴으로써 더 나은 분류 기능을 가지도록 하는 것이다. 온라인에서 얻은 아래 그림에서 Discriminator(D)라는 머신은 원본과 위조 이미지를 판별하며, Generator(G)라는 머신은 원본과 유사한 이미지를 생성하여 머신 D를 속이도록 학습한다.

○ GAN을 이용하여 재현자료를 생성할 때는 머신 G가 생성하는 자료를 재현자료로서 이용하는 방식을 취한다. 이미지가 아닌 마이크로데이터의 경우에는 각 레코드를 행렬 형식으로 만들어서 이미지처럼 GAN에 입력한다. 따라서 이용자는 G가 생성한 재현자료를 이용하며, 원본자료를 받지 않는다.



<그림 3-6> GAN를 이용한 재현 이미지 생성

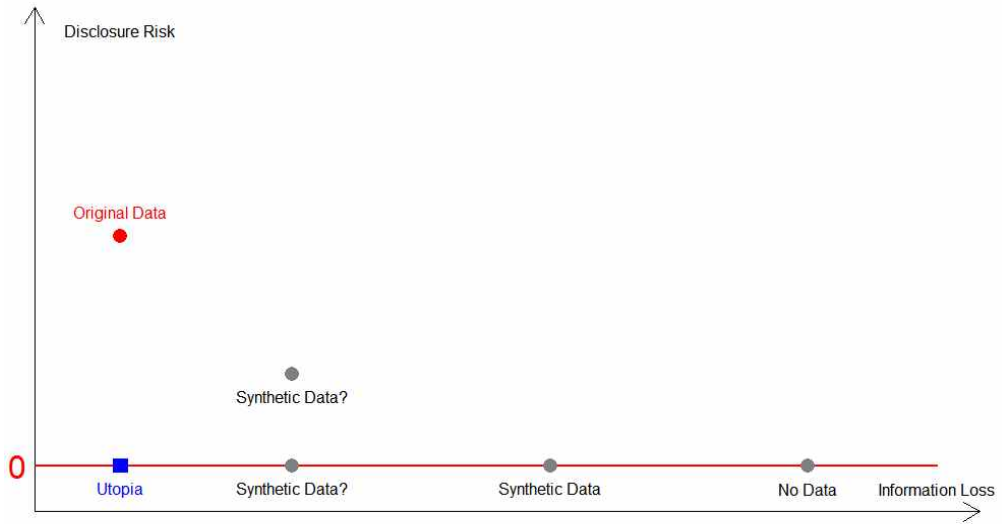
○ 나아가 조건부 적대적 생성망(Conditional GAN)을 이용하여 재현자료를 생성하는 방식이 제안되어 자료의 특성을 더욱 잘 반영할 수 있게 되었다(Xu et al., 2019). 베이지안 모형 및 딥러닝을 이용하는 양측 접근 방식 모두에서 자료에 대하여

조건부 분포 추정의 문제를 주요하게 다루는 공통점이 발견된다. 관련 주제는 현재 학계에서 활발한 연구가 진행 중이다.

○ 베이지안 재현자료 생성 및 GAN을 이용한 재현자료 생성에 관한 보다 자세한 내용은 '재현자료 작성 방법론 검토(박민정과 김정연, 2017)' 및 '통계데이터센터 DB를 이용한 재현자료 생성 방법 연구(박민정 외, 2020)' 등을 참고할 수 있다.

3.3 재현자료의 노출위험과 정보손실 측정

□ 완전 재현자료의 경우 노출위험이 0이라고 판단한다. 따라서 재현자료를 생성하는 모수적 모형을 통계적으로 잘 구축하는 것이 관건이었다. 그러나 정보손실이 적게 발생하도록 실제 자료의 모형을 추정하는 것은 쉬운 일이 아니다. 이에 베이지안 붓스트랩을 이용하는 비모수적 재현자료 생성 방안이나 부분 재현자료 생성 방법이 연구되었다. 이러한 유형의, 원자료가 포함되는 재현자료는 노출위험이 0이라고 단정하기가 어렵다는 지적이 있다. 따라서 정보손실을 줄이다보니 노출위험이 증가하는 매스킹과 유사한 상충관계 문제가 발생한다. 이러한 재현자료에 대하여 노출위험을 측정하는 것은 쉽지 않은 문제이다.



<그림 3-7> 이상적인 완전 재현자료의 위험-유용성 지도

○ 이에 차등정보보호를 만족하는 재현자료 생성 방안이 제안된 바 있다. 이러한 방법으로는 노출위험은 정량적으로(formal privacy) 제어되나 정보손실이 과하게

발생하는 측면이 있다. 관련 논의는 ‘차등정보보호에 관한 연구(박민정 외, 2018)’를 참조할 수 있다.

○ 현실적으로 엄밀성은 부족하더라도 재현자료의 노출위험을 간단하게 계산하는 노력이 있으며, 관련 논의는 ‘통계데이터센터 DB를 이용한 재현자료 생성 방법 연구(박민정 외, 2020)’를 참고할 수 있다.

□ 마지막으로 재현자료의 정보손실 측정은 매스킹 처리된 자료의 경우와 유사하다. 이 주제 역시 관련 연구가 활발하게 진행되고 있다.

4. 차등정보보호 적용

□ 관련 기술의 발전으로 대용량 데이터 파일을 공유하는 대신에 데이터 분석을 가능하게 하는 플랫폼을 활용하는 방안이 대두되고 있다. 기존 관점에서는 마이크로데이터의 제공을 어떻게 할 것인가를 주요 이슈로 다루는데 비하여, 차등정보보호는 분석 결과를 제공하는 플랫폼을 가정한다.

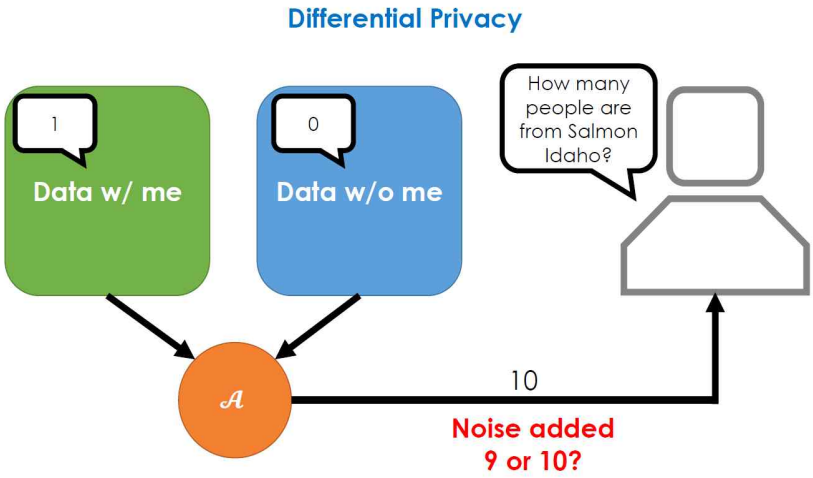
○ 차등정보보호는 정보보호를 생각하는 기존 방식에 의문을 제기하고, 새로운 관점을 제안하고 있다. 따라서 차등정보보호의 개념과 정의를 잘 이해할 필요가 있다.

○ 마이크로소프트 연구진에 의하여 제안된 차등정보보호는 미국 센서스국뿐만 아니라, 구글, 애플 등의 기업에서도 그 활용 방안이 활발하게 논의되고 있다. 다만 대용량 데이터의 분석 결과가 아닌 경우에는, 노출위험을 강하게 제어할 때 정보손실이 크게 발생할 수 있음을 주의할 필요가 있다.

4.1 개념 및 수학적 정의

□ 차등정보보호의 주요 아이디어는 “한 개체”만 다른 두 개의 데이터베이스(DB)에 쿼리(질의)를 요청할 때 얻어지는 결과물의 차이를 제한하여 해당 개체에 대한 정보가 드러나지 않게 하는 것이다. 이는 노출위험에 대한 새롭고 강력한 정의이면서도 수학적 엄밀성을 확보하여 주목받고 있다.

○ 다음 그림은 차등정보보호(differential privacy)의 개념을 함축적으로 보여준다. 전체 인구수가 얼마인지 묻는 쿼리에 대하여 한 개체가 포함되어 있을 때와 그렇지 않을 때 해당 개체의 정보보호를 달성하기 위한 방법을 논의한다.



<그림 3-8> 차등정보보호의 예시

□ 차등정보보호 개념 이해를 위하여 다음의 예를 살펴보자(Nissim et al., 2018).

Alice와 Bob은 연구 활동이 활발한 교수입니다. 그들은 한 대학에 대하여 각 학생이 받는 재정 지원과 관련된 정보 등, 학생에 관한 개인 정보가 들어있는 **데이터베이스에 접근(access)**할 수 있습니다. 데이터베이스에는 개인 정보가 포함되어 있기 때문에 허가 없이는 접근이 제한됩니다. 접근 권한을 얻기 위해 Alice와 Bob은 **기밀 유지 교육**을 받고, 데이터베이스에서 얻은 개인 정보의 사용 및 공개를 금지하는 **데이터 사용 계약서에 서명**했습니다. 이로써 개인정보 처리를 위한 이용자 규제를 따르고 있습니다.

3월에 Alice는 이 데이터베이스의 정보를 기반으로 한 연구 결과를 발표했고, 그 중에 "사립대학은 **3,005명**의 신입생이 있으며, 이중 **202명은 연간 35만 달러 이상**의 수입을 올리는 가정입니다."라는 내용이 포함되어 있습니다. Alice는 그녀가 **3,005명이 넘는 집단에 대한 총 통계를 발표**했으므로 어떤 개인의 개인 정보도 공개되지 않았다고 주장합니다.

다음 달 Bob은 다음 통계를 포함하는 별도의 연구 성과를 발표합니다. "사립대학 신입생 클래스 **3,004명** 중, **201가구가 연간 35만 달러를 초과**하는 가계 소득을 가지고 있습니다." Alice와 Bob은 둘 다 비슷한 정보를 발표했음을 모르고 있습니다.

영리한 학생 Eve는 이 두 기사를 읽고 관찰합니다. Eve는 3월과 4월 사이에 한 신입생이 사립대학을 탈퇴했으며, 학부모는 연간 35만 달러 이상의 수입을 얻는다고 결론을 내립니다. Eve는 주위에 묻고 John이라는 학생이 3월 말쯤에 중퇴했다는 것을 알 수 있습니다.

이제 Eve는 학급 친구들에게 **John의 부모님이 아마도 1년에 35만 달러 이상을 벌어들일 것**이라고 알려줍니다.

○ 이 예에서 이용자는 데이터센터를 방문하고 집계된 결과만을 공개하였으나, 외부 정보 및 DB의 차이로 인하여 한 개인의 프라이버시가 침해되었음을 알 수 있다.

○ 이렇게 한 개체만 차이가 나는 DB를 가정하는 것은 일반적으로 생각할 수 있는 가장 강력한 수준의 위험에 해당한다. 따라서 차등정보보호는 가장 강력한 수준의 정보보호를 지향한다고 할 수 있다.

□ 차등정보보호를 '개체 차이를 기준으로 하는 수학적 정보보호 정의' 혹은 '한 개체의 차이에 따른 프라이버시 변동량(증분) 제어'라고 이해할 수 있다. DB에서 개체 하나의 차이를 반영하므로 차분 프라이버시/정보보호라는 용어도 일리가 있으나, 본 연구에서는 국내 초창기 연구 논문(이용희, 2013)의 번역을 따르도록 한다.

> (Wikipedia) Roughly, an algorithm is differentially private if an observer seeing its output cannot tell if a particular individual's information was used in the computation.

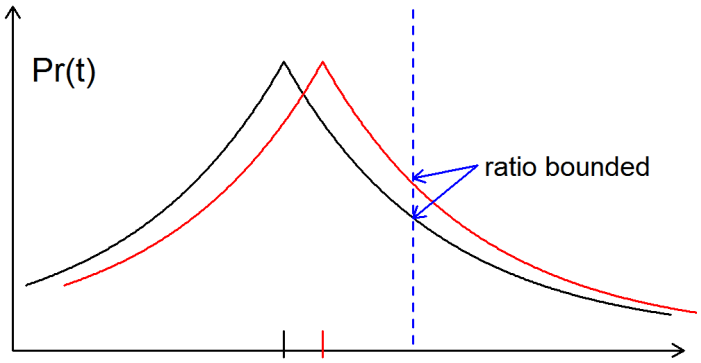
□ 차등정보보호의 수학적 정의는 다음과 같다. 레코드 하나만 다른 두 개의 데이터베이스 D 및 D'에 대하여 이용자가 수령하는 결과물이 가지는 분포의 비율이 일정한 기준 이하가 되도록 하는 것이다. 아래에서 K는 랜덤화 함수로 분석결과에 잡음을 추가하는 기능을 수행하며, 쿼리의 민감도와 정보보호 수준 파라미터에 따라 달라진다.

$$\frac{Pr[\,K(D) = t\,]}{Pr[\,K(D') = t\,]} \leq exp(\epsilon)$$

○ 아래의 그림을 통하여 수학적 개념을 이해할 수 있다. 주어진 쿼리 f에 대하여 데이터베이스 D를 분석한 결과를 가로축의 검은색 선으로 나타내었다. D'에 대한 분석 결과는 가로축의 붉은색 선으로 나타내었다. 만약 검은 선 및 붉은 선이 나타내는 값을 그대로 제공한다면 그 차이로 인하여 해당 개체의 정보가 노출된다.

○ 이때 아래 그림의 뾰족한 곡선과 같은 모양을 가지고 평균이 0인 분포를 따르는 잡음을 하나 발생시켜 분석 결과값에 더하여 제공한다. 쿼리가 요청될 때마다 제공되는 분석 결과를 모으면 위 그림과 같은 분포를 따르게 된다. D를 분석한 결과이든 D'를 분석한 결과이든, 이용자가 특정한 값을 얻을 확률이 유사한 것을 알 수 있다. 이처럼 차등정보보호는 결과값이 가지는 분포 비율의 차이를 제

한하여 한 개체의 차이에 따른 노출위험을 제어하고자 한다.



<그림 3-9> 차등정보보호의 수학적 개념 이해

○ 보다 자세한 내용은 ‘정보공개 환경에서 개인정보 보호와 노출 위험의 측정에 대한 통계적 방법(이용희, 2013)’, ‘통계적 노출제어 최신동향 검토(박민정, 2016)’, ‘Statistical disclosure control for public microdata: present and future(Park and Kim, 2016)’ 및 ‘차등정보보호에 관한 연구(박민정 외, 2018)’ 등을 참고할 수 있다.

□ 한편 실제 적용 사례 및 모의실험 결과는 ‘차등정보보호에 관한 연구(박민정 외, 2018)’ 및 ‘차등정보보호 적용 실험 연구(박민정과 권성훈, 2019)를 참고할 수 있다.

4.2 국소 차등정보보호 소개

□ 차등정보보호는 원자료가 수집되어 있는 DB를 가정한다. 그러나 자료 제공자가 원자료를 수집할 때부터 정보보호를 원할 수 있다. 이를 위한 방법을 국소 차등정보보호(local differential privacy)라고 하며, 주로 구글, 애플 등의 기업에서 그 적용 방안이 활발하게 연구되고 있다. 학계에서도 관련 연구가 활발하다.

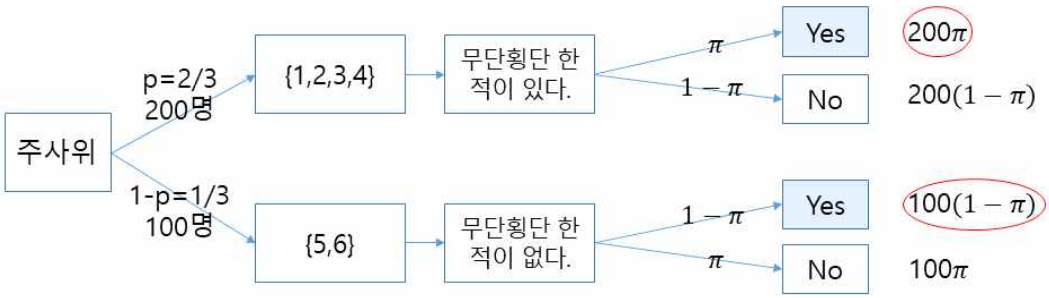
○ 앞에서 차등정보보호를 구현할 때 쿼리마다 잡음을 추가하는 방법이 다름을 언급하였다. 국소 차등정보보호도 유사한 특징을 가진다. 즉, 특정 쿼리를 분석할 용도로 국소 차등정보보호를 적용하여 수집한 자료로 다른 쿼리를 분석할 수 없다. 이는 목적 외 이용금지를 요구하는 유럽의 GDPR(일반개인정보보호규정)에 부합하는 특징이다. 한편 국소 차등정보보호를 적용하면 원자료의 참값은 부정확하게 수집되며, 그 노출위험 수준을 정량적으로 제어할 수 있다.

□ 국소 차등정보보호는 Warner(1965)가 제안하는 원리를 따른다. 이를 다음과 같은 예를 통하여 이해하자. 횡단보도 설치를 결정해야 하는 자료 수집기관이 시민들이 특정 위치에서 무단횡단을 한 적이 있는지 조사한다고 하자. 자료 제공자는 자신의 무단횡단 유무를 알리기 원하지 않는다고 가정한다. 무단횡단을 하는 비율을 추정하기 위하여 자료를 수집할 때 다음과 같은 질문을 준비한다.

A) 무단횡단을 한 적이 있다.	B) 무단횡단을 한 적이 없다.

○ 자료를 제공하는 사람은 주사위를 던져 1, 2, 3, 4가 나오면 A 질문에, 5, 6이 나오면 B 질문에 대하여 답을 한다. 자료를 수집하는 사람은 주사위 결과를 모르며, “예”와 “아니오”만 수집한다. 무단횡단을 한 사람의 실제 비율이 π이고 A 질문에 답하는 확률이 p라고 하자. 이 예제에서는 $p = 2/3$가 된다. 만약 n명의 사람을 조사하여 m명이 “예”라고 대답했다면 다음과 같이 π를 추정할 수 있다.

$$\hat{\pi} = \frac{p-1}{2p-1} + \frac{m}{(2p-1)n}$$



<그림 3-10> 정보제공자 보호 자료수집(국소 차등정보보호) 예제

5. 플랫폼 맞춤형 알고리즘 개발

□ 이 절에서는 통계청 SGIS의 대용량 빈도표 제공을 위한 플랫폼 맞춤형 비밀보호 알고리즘 BSCA 개발을 소개한다. 이를 통하여 플랫폼 맞춤형 비밀보호 기술의 적용이나 관련 시스템 개발이 필요함을 인식하고자 한다.

5.1 대용량 빈도표 제공과 노출제어 문제

□ 1절의 매크로데이터 개인정보 노출 가상 사례에서 다양한 조합에 대한 빈도를 제

공할 때 작은 값에 대한 N/A 처리만으로는 개인정보보호가 어려움을 살펴보았다.

○ 이러한 N/A 처리는 국소 감추기(local suppression, primary suppression)라고 불리며, 국소 감추기를 적용할 때는 총계 제공을 고려하여 2개 이상의 셀을 감추어야(secondary suppression) 함이 잘 알려져 있다(Hundepool et al., 2012).

○ SGIS(통계지리정보서비스) 플랫폼은 인구주택총조사와 같은 대용량 자료를 다운로드하여 분석하는 대신에, 이용자가 필요한 변수 범주를 선택하면 해당 빈도를 지도상에 표기하는 서비스를 제공한다. 따라서 이용자는 SGIS가 제공하는 항목 및 지역에 대한 모든 유형의 빈도표를 작성해 볼 수 있으므로, 국소 감추기를 2차까지 바르게 적용하게 되면 거의 대부분의 빈도가 N/A 처리되어 나타날 수 있다.

○ 빈도표의 비밀보호 기법은 각국 통계기관을 중심으로 연구되어 왔으며, 대부분 기법이 매우 복잡하고 정보손실이 크다. 차등정보보호를 만족하는 재현자료 생성(Machanavajjhala et al., 2008)은 설명하기에는 가장 단순하지만, 이론적으로 어렵고 정보손실이 크다는 단점이 있다. 한편 실무적으로 적용 가능하고 효율적인 방법으로 작은 셀 조정(small cell adjustments, SCA)이 있다(Hundepool et al., 2012).

5.2 대용량 빈도표 제공을 위한 BSCA 알고리즘 개발

가. 정보손실-노출위험 우선순위 패러다임 전환

□ 많은 연구에서 노출위험 제어를 우선적으로 고려하고, 정해진 노출위험 수준을 만족하는 범위에서 정보손실 감소 방안을 연구한다(Machanavajjhala et al., 2008).

○ 그러나 빈도표의 경우, 이용자가 큰 정보손실을 수용하기는 어려우므로, 정보손실 제어를 우선시한 알고리즘 개발을 도모하였다. 이에 정보손실이 유계한 작은 셀 조정(Bounded Small Cell Adjustments, BSCA) 알고리즘을 설계하였다.

나. BSCA 알고리즘 및 적용 결과

□ BSCA 알고리즘은 SGIS가 제공하는 빈도표의 위계구조를 고려한다. 또한 시스템이 제공하는 정보와 이용자가 유추하는 정보를 고려하여 작성되었다.

○ (기저 빈도표) 최소 지역 단위마다 제공하는 모든 변수의 경우를 고려한 빈도표로, 작은 셀을 랜덤 라운딩(SCA) 처리하여 제공한다.

○ (상위 빈도표) 상위 지역에서 집계하거나 특정 변수에 대해 집계한 빈도표로, 기저

빈도표의 참값 및 SCA 처리된 값을 BSCA 알고리즘으로 비교/집계해 제공한다.

○ BSCA 알고리즘으로 상위 빈도표의 셀 값을 결정하는 과정은 다음과 같다. 먼저 모든 자연수에 대하여 작은 셀 기준 B로 나누어진 구간을 만들고 아래 세 단계를 시행한다. 단계 ①에서는 정보손실이 B/2 이하로 발생한다. 단계 ②에서는 정보손실이 B만큼 증가할 수 있지만 그 비율은 1% 미만으로 적다.

① (정보손실 제어) 셀의 참값이 속하는 구간의 중앙값을 할당
② (노출위험 제어) 시스템 제공 정보 C 및 이용자 추론 정보 D를 비교해 C⊂D이도록 조정
③ (일관성 확보) B보다 작은 셀의 값은 B로 변경

○ 다음은 1개 변수에 대해 격자 단위로 작성한 기저 빈도표와 상위 빈도표의 예이다.

기저 빈도표

참값

X11	X12	...
0	0	...
3	4	...
4	5	...
2	1	...
0	0	...
8	0	...
0	0	...

SCA 적용 결과

X1	X2	...
0	0	...
3	4	...
4	5	...
3	3	...
0	0	...
8	0	...
0	0	...

상위 빈도표

참값

X11	X12	...
0	1	...
20	9	...
20	33	...
6	7	...
0	0	...
20	13	...
1	3	...

BSCA 적용 결과

X1	X2	...
0	0	...
22	9	...
21	31	...
5	6	...
0	0	...
19	14	...
0	3	...

○ SCA만으로 위계구조가 있는 빈도표를 제공하면, 집계 방식에 따라 정보손실이나 노출위험이 제어되지 않는다. BSCA는 노출위험이 높은 셀을 선별하고 노출위험을 제어하기 위하여 정보손실을 추가로 발생시킨다(박민정 외, 2018; Park, 2018). 따라서 노출위험과 정보손실을 정량적으로 제어한다. 한편 BSCA 최종 알고리즘 작성 및 SGIS 등에 적용하는 작업을 수행하여 관련 내용을 통계지리정보서비스(SGIS) 제공 자료의 비밀보호 적용 방안(홍영희와 박민정, 2020)에 수록하였다.

6. 분석결과물 반출 가이드라인 운영

□ 1절의 사례에서 데이터센터를 방문하는 이용자가 반출을 요청하는 분석결과물에 노출제어가 필요함을 살펴보았다. 통계청은 전수자료를 제공하는 (빅)데이터 통계센터를 2017년부터 운영하기 시작하여 관련 가이드라인(안)을 작성한 바 있다(박민정 외, 2017). 이는 미국 국립보건통계센터(NCHS) 등의 사례를 참조하여 작성되었다.

6.1 데이터센터 통계적 노출제어 원칙 및 반출유형별 반출 기준

□ 가이드라인 초기 버전에서는 통계적 노출제어를 위한 기본 원칙으로 자료의 표준화를 선정하였다. 여기에는 동일 변수 동일 범주의 사용, 반출물의 최소 단위 통일 등이 포함된다. 예를 들어 만 연령 기준 5세 단위 연령 변수 사용이나, 매출액의 백만 원 단위 반출 등을 생각할 수 있다. 다만 노출제어 원칙 운영에는 실제 반출현황을 반영하여 원칙 추가, 수정 및 보완 작업을 꾸준하게 지속해야 한다.

□ 가이드라인은 다음과 같이 반출목적을 2종류로, 반출유형을 4종류로 구분하였다.

○ 각 유형에 대하여 노출위험 및 정보손실을 정의하고 수정/보완 작업을 지속한다.

<표 3-4> 통계데이터센터 반출 가이드라인

반출목적		
① 일반이용	② 표본설계	③ 기타
▸ 반출물을 활용해 보고서, 논문 등을 작성하는 경우 ▸ 반출물을 신규통계 공표에 사용하는 경우 등	▸ 반출물을 표본설계에 이용하는 경우 ▸ 모집단 보완 및 구축 등 ▸ 반출물 직접 공표 금지	▸ 앞의 유형에 속하지 않는 경우 ▸ 심의위원회가 별도 관리

반출유형				
① 빈도표	② 총계표	③ 기술통계	④ 모형결과	⑤ 기타

□ 1절의 사례2를 참고하여 반출하는 통계의 개수와 분석에 사용된 레코드 개수에 대한 관리가 필요하며, 초기 가이드라인에 관련 내용을 새로 반영할 필요가 있다.

6.2 반출물에 대한 체계적 관리 시스템 운영

□ 가이드라인에서 자료이용 신청/반출 서식을 제안하였고, 이는 통계데이터센터 시스템에 반영되어 있다. 데이터센터마다 보유 DB 및 이용 현황이 다양하므로 각 센터의 반출 사례를 분석하여 가이드라인을 맞춤형으로 업데이트할 필요가 있다.

제 4 장

데이터 구독 플랫폼 실행방안

1. 데이터 거래 현황

□ 거래가 이루어지고 있으나 실제 거래 현황은 미미한 수준이다.

○ 공공부문에서는 공공데이터포털(행안부, 2011), 데이터스토어(데이터산업진흥원, 2012), 금융데이터거래소(금융보안원, 2020) 등 10여 개의 데이터 플랫폼이 있다.

○ 민간부문에서는 한국데이터거래소(민간, 2019), 네이버, SKT, LG CNS 등 10여 개를 들 수 있다.

○ 한국데이터산업진흥원 '데이터스토어'의 `19.10 기준 최근 5년간 누적 거래량 11억 원에 지나지 않는다.

○ 특히 데이터스토어 내 10만 2,097건 중 미술·사진 등 데이터가 98.6%로 대부분이며, 기업이 활용 가능한 파일 또는 오픈API 데이터는 1.4%(1,444)에 불과하다.

□ 해외에서는 세계 최대 데이터 브로커 시장을 가진 미국과 최근 급속히 성장하고 있는 중국의 사례를 살펴본다.

○ 미국의 데이터 브로커 시장 규모는 2018년 기준 약 1,832달러 한화로 약 220조 원에 달하는 것으로 알려져 있다. 특히 맞춤형 광고정보 등을 제공하는 대표적 데이터 브로커 기업은 엑시엄으로 연매출은 9억 달러(약 1조 원, 2018)에 달한다. 이 기업은 약 7억 명의 거래 데이터를 가지고 있는 것으로 알려져 있다.

○ 중국은 비교적 최근에 데이터 거래시장 활성화 정책을 펴면서 귀양빅데이터거래소를 오픈하였다. 귀약빅데이터거래소의 지난 4년간('14 ~'18.5)의 거래 규모는 약 3억 위안(약 505억 원) 정도인 것으로 알려져 있다.

2. 데이터 거래 애로사항 및 정책진단

□ 본 절에서는 먼저 데이터 거래에 있어서의 현실적 애로사항을 살펴본다. 그리고 이를 극복하기 위해서 현재의 정책을 진단하고 정책방향에 대하여 제언한다.

2.1 데이터 거래에 있어서의 애로사항

□ 과기부에서는 2015년 이후 매년 데이터 거래 현황의 애로사항 등을 진단하기 위하여 데이터 구매 경험자를 대상으로 데이터산업현황조사를 실시해오고 있다.

○ 주요 질의 사항은 크게 1. 데이터 구입 및 입수 시 애로사항 2. 데이터 가공 및 개인정보 관련 애로사항으로 나누어 볼 수 있다.

□ 주요 애로사항은 데이터의 질 및 가격 등 데이터 입수 등과 관련된 사항이다. 반면 데이터 가공 및 개인정보와 관련한 애로사항은 많지 않다.

○ ‘쓸 만한 양질의 데이터 부족’이 2018, 2019년 기준 각각 48.4%, 50.9%로 가장 높게 나타났다.

○ 다음으로는 ‘불합리한 데이터 가격 산정’이 각 연도별 애로사항의 40.7%, 35.8%로 나타났다. 아울러 ‘데이터 유통 채널 부족’이 역시 각각 40.7%, 30.2%로 세 번째로 높은 애로사항임을 알 수 있다.

<표 4-1> 데이터 거래 시 애로사항

	2015	2016	2017	2018	2019
쓸 만한 양질의 데이터 부족	-	-	-	48.4	50.9
불합리한 데이터 가격 산정	29.4	33.9	35.1	40.7	35.8
데이터 유통 채널 부족	11.8	11.9	33.0	40.7	30.2
데이터 소재파악 및 검색의 어려움	-	-	-	34.1	22.6
데이터 품질 문제	26.5	27.1	12.4	25.3	24.5
데이터 가공 문제	-	-	-	8.8	3.8
데이터 거래 절차 및 방법 미숙	14.7	13.6	2.1	8.8	1.9
개인정보 처리 기술력 및 예산 부족	2.9	3.4	5.2	1.1	5.7
개인정보 포함 데이터 유통/활용 시 법률지원	8.8	6.8	6.2	1.1	7.5
기타	2.9	1.7	-	8.8	3.8

출처 : 데이터산업현황조사(과학기술정통부)

○ ‘개인정보 처리 기술력 및 예산 부족’은 1.1%, 5,7%, ‘개인정보 포함 데이터 유통/활용 시 법률지원’은 1.1%, 7.5%로 가장 낮게 나타났다.

2.2 데이터 거래 활성화 정책의 한계

가. 데이터 특성

□ 데이터는 손쉬운 복사와 이동이 가능하다는 특성을 가진다.

○ 이는 데이터 고유의 특성이라기보다는 디지털화의 특성이다. 거의 모든 데이터가 디지털화되고, 클라우드 서비스와 초고속 인터넷 망의 발달로 데이터 복사와 이동에 매우 용이해졌다.

○ 반면 데이터 생산자는 데이터가 쉽게 뿌려지고 유통될 수 있기 때문에 오히려 데이터 제공을 꺼리는 현상도 발생하고 있다.

□ 개인정보 또는 개인들의 성향이 포함된 데이터가 많아 유출 시 위험이 매우 크다.

○ 의료, 금융데이터 등 민감한 개인정보를 포함한 데이터를 들 수 있다.

○ 또 다른 경우로는 서로 별도의 데이터 베이스로 관리될 때는 문제가 없으나 데이터의 결합으로 인하여 개인이 식별되어 개인정보가 유출되는 경우를 들 수 있다.

□ 데이터 거래시장이 형성되고 활성화되기 위해서는 데이터의 손쉬운 이동, 개인정보 포함 등의 특성이 반영한 거래 시장 또는 거래 생태계 개념이 필요하다.

○ 첫째, 데이터 손쉬운 이동으로 인한 무단, 불법 복제 사용을 막을 수 있어야 한다. 일반적인 소프트웨어와 달리 데이터의 경우에는 기업 단위에서의 무단 사용을 통제하는 것도 쉬운 일이 아니다.

○ 둘째, 적극적 개념의 개인정보 보호 대책이 필요하다. 데이터 노출 시점이 아니라 컴퓨팅 파워 등이 획기적으로 발전한 미래시점에 있어서도, 그리고 다른 데이터와의 결합 가능성이 있는 경우에는 결합 후에도 개인정보 유출로부터 안전한 데이터 보호 대책이 필요하다.

나. 데이터 특성을 반영한 거래 생태계 형성 및 활성화 전략

□ 데이터 특성을 반영한 데이터 거래 전략이 필요하다.

○ 데이터에 내재되어 있는 개인정보를 보호하고, 데이터의 무단 복제 등을 통한 오남용을 통제할 수 있어야 한다. 그런데 디지털화된 데이터의 속성상 손쉬운 복제와 이동 특성으로 인하여 이를 제어하는 데에는 한계가 있다.

○ 그렇기 때문에 현재 데이터 관리는 대부분 데이터 외부반출을 통제한 상태에서 특정 물리적 공간에서 데이터를 분석하고 결과를 반출하는 방식으로 이루어진다.

○ 그런데 이러한 방식은 5G 시대, 플랫폼 사업 시대 등 현 경제 산업 트렌드와는 잘 어울리지 않는 방식이다.

○ 결국 데이터 정보보호를 하면서도 온라인상에서도 분석이 가능한 혁신적인 방법을 강구할 필요가 있다.

□ 데이터 거래 생태계라는 관점이 필요하다.

○ 데이터는 위의 특성으로 인하여 일반 재화와 차별화된다. 그렇기 때문에 데이터 거래 고유의 생태계를 구축하고, 이에 걸맞는 시스템을 구축할 필요가 있다.

○ 즉 단순히 물리적 데이터 거래소 숫자만 늘린다던가 아니면 데이터의 종류와 양을 늘린다고 자연스럽게 데이터 거래가 활성화되리라고 기대하기 어려운 것이다.

○ 데이터 거래 생태계에서는 다음과 같은 조건을 만족하여야 한다.

- 첫째, 데이터 제공자에게는 위험 부담없이 안심하고 데이터를 제공할 수 있는 환경을 만들어 주어야 할 것이다.
- 둘째, 데이터 이용자는 각자의 기업에 적합한 데이터와 그에 대한 활용 가능한 방법과 얻게 될 성과에 대한 정보를 쉽게 파악할 수 있어야 할 것이다.
- 셋째, 데이터 가공, 분석 업체는 다양한 양질의 데이터를 쉽게 입수하고, 분석할 수 있는 환경을 제공해야 할 것이다.

□ 플랫폼 경제 구조를 이해하고 적용할 수 있어야 한다.

○ 재화의 속성상 자연스럽게 유통시장이 형성되지 못하는 경우가 있다. 이 경우 재화의 거래 속성을 잘 이용할 수 있는 거래 시장(플랫폼)을 구축할 필요가 있다.

○ 예를 들어 스마트폰 어플리케이션 개발의 경우 초기 투자비용이 많이 들어 이용자가 많지 않다면 높은 가격을 책정할 수밖에 없다. 그런데 가격이 높아지면 소

비자(이용자)들의 외면을 받기 십상이다. 플랫폼을 구축하여 초기개발비용이나 마케팅 비용을 줄여준다면 낮은 가격을 책정할 수 있고, 낮은 가격을 책정하면 이용자가 많아지는 선순환 과정을 거치게 된다. 플랫폼은 이러한 선순환 과정[24]을 가능하게 해준다.

○ 데이터 거래 또한 마찬가지이다. 데이터 수요자가 많다고 하면 추가적인 데이터 복제에 거의 추가적인 비용이 들지 않는 만큼 데이터 가격을 낮출 수 있다. 즉 플랫폼은 수요자와 공급자 간의 가격 괴리를 줄여주는 역할을 할 수 있다.

○ 또한 데이터 거래 통합 플랫폼을 만듦으로 해서 다양한 데이터를 한 군데 플랫폼에서 검색하고 거래할 수 있도록 해야 한다.

□ 구독 경제의 개념을 적용하여 데이터 거래를 활성화할 필요가 있다.

○ 구독경제란 일정기간 동안 정기적으로 회비를 납부하면서 서비스를 제공받는 것을 의미한다. 많은 데이터는 시간이 흐르면서 새로운 데이터가 누적적으로 만들어진다. 1회성 거래보다는 구독을 통한 정기적인 데이터 업데이트가 데이터 이용자들에게도 더 합리적인 서비스가 될 수 있다.

○ 또한 구독 경제는 개별 데이터에 대한 가격 책정 부담으로부터 자유로울 수 있다. 데이터 가치는 보는 시각에 따라, 시점에 따라, 각자의 상황에 따라 천차만별이다. 이를 합리적인 하나의 잣대로 데이터 가치를 평가하고 가격을 매기는 것은 불가능에 가깝다. 그런데 구독 통계는 개별 데이터에 가격을 매기는 것이 아니라 소정의 회비를 납부하는 방식으로 운영하기 때문에 가격책정의 부담이 없다.

□ 데이터 정보보호 기반하에 시장 거래 생태계가 구성되어야 한다.

○ 데이터 분석 결과만 반출 시에는 차등정보보호 등의 개인정보보호 처리과정을 거쳐야 한다.

○ 마이크로 데이터를 제공할 경우에는 향후 발생할 수 있는 이종 간 데이터의 결합이 불가능하도록 조치하여야 할 것이다.

○ 암호화된 상태의 마이크로데이터 제공 시에는 미래 컴퓨팅 기술이 획기적으로 발전이 이루어진 이후에도 안전한 방식으로 암호화될 필요가 있다.

---

24) 이러한 효과를 네트워크 효과라고 부른다. 네트워크 효과란 공급시장과 수요시장이 상호 영향을 주는 것을 의미한다. 다시 말해 양 시장 간 선순환 과정으로 수요시장의 수요자가 많아지면 공급자도 많아지고, 공급자가 많아지면 수요자가 많아지는 것을 의미한다.

3. 데이터 구독 플랫폼

3.1 데이터 거래 시장 생성의 어려움

□ 구독 플랫폼이 필요한 첫 번째 이유는 데이터 공급자와 수요자 간의 큰 가격 괴리를 들 수 있다.

○ 기업단위 데이터 수요자는 데이터를 활용하여 얼마나 큰 추가적인 부가가치를 창출할 수 있을지 불확실하다. 이러한 상황 속에서 데이터에 높은 가격을 지불하려는 의사결정을 하기는 쉽지 않을 것이다.

○ 데이터 공급자 입장에서 실제 데이터 생산 비용은 높지 않을 수 있다. 기업 운영에 따른 부산물로 얻어질 수도 있기 때문이다.

○ 그러나 데이터를 외부에서 이용할 수 있도록 제공하는 것은 단순히 데이터 생산 비용만을 고려할 수 없다. 데이터 제공 시 향후 개인정보 유출 등의 위험에 노출 될 수 있기 때문이다. 다시 말해 데이터 제공은 일종의 우발채무 가능성을 높일 수 있어 데이터 생산 기업은 데이터 제공에 소극적이거나 높은 가격을 책정하여 데이터 활용 가능성을 낮출 유인이 있다.

□ 데이터 공유 또는 재제공에 따른 시장교란행위가 있을 수 있다.

○ 하나의 데이터 셋을 공동으로 구매하는 경우를 들 수 있다. 데이터를 공동으로 구매하여 공유하는 경우 사실상 데이터 구입가격을 절반으로 낮출 수 있다.

○ 구입한 데이터를 무상으로 재제공하는 경우이다. 이 경우에도 시장에서 데이터의 가격이 온전히 유지될 수 없다.

□ 데이터 정제·연계·분석의 어려움이다.

○ 수요자 입장에서 데이터를 분석하거나 기업의 마케팅에 실제로 이용하는 일은 쉬운 일이 아니다.

○ 데이터를 분석하고 해석하기 위해서는 통계학적 지식, 해당 도메인에 대한 지식, 그리고 코딩 역량이 필요하기 때문이다. 이러한 데이터를 회사의 업무에 적용하거나 분석 결과를 활용하는 것은 논외로 하고라도 말이다.

○ 즉 기업 활동 과정에서 많은 데이터가 자연스럽게 축적되거나 전문적 역량을 가진 인재들이 많은 대기업을 제외하고는 사실상 데이터를 직접적으로 활용하는데

는 어려움이 많을 수밖에 없다.

3.2 데이터 구독 플랫폼이란?

가. 개념

□ 데이터 구독 플랫폼은 동형암호를 기반으로 하여 데이터 거래에 구독경제 시스템을 도입한 것을 의미하며 다음과 같이 정의한다.

○ 데이터 구독 플랫폼은 데이터 사용자들이(기업·개인) '구독 데이터 플랫폼'에 회원으로 가입하고, 정기적으로 합리적 가격의 구독료를 지불하면서 암호화된 다양한 데이터를 연계·분석·활용 할 수 있는 거래 형태를 의미한다.

□ 구독경제란(Subscription Economy)는 제품이나 서비스를 구매·소유하는 것보다 적은 금액을 지불하고 일정 기간 동안 사용할 수 있는 방식을 의미한다.

○ 소비자는 필요할 때 필요한 만큼만 소비함으로써 ① 소비자 맞춤형 서비스를 받을 수 있고 ② 합리적 가격으로 이용이 가능하다.

○ 대표적인 예로는 SaaS(Software as a Service), BaaS(Battery as a Service) 등을 들 수 있다. 세일스포스는 고객관리 전문회사로 클라우드에서 소프트웨어를 고객이 필요할 때 필요한 만큼 사용하도록 서비스하는 SaaS를 구현하고 있다. 중국의 전기차 회사 니오는 베터리 충전시간을 단축하기 위하여 배터리를 구매하지 않고, 월정액을 내면 필요할 때마다 배터리 교환소에서 교체하는 BaaS를 제공하고 있다.

□ 동형암호[25]란 암호화된 데이터를 복호화 없이도 연산할 수 있는 암호 기술이다.

○ 암호화된 상태에서 연산한 결과 값을 복호화하면 평문 상태의 데이터를 연산한 결과와 동일한 값을 얻을 수 있다는 의미를 내포하고 있다.

○ 동형암호는 암호화된 상태에서 연산을 진행하더라도 그 결과 값은 암호문 형태로 남아 있어 결과 값을 보려면 복호화(암호문을 평문으로 해독하는 과정) 과정을 거쳐야 한다. 그런데 복호화 하려면 권한이 있는 자가 보유하고 있는 복호화 키가 필요하다.

○ 즉 데이터를 한번 입수하면 그 데이터에 대한 모든 권한을 가지는 것이 아니라

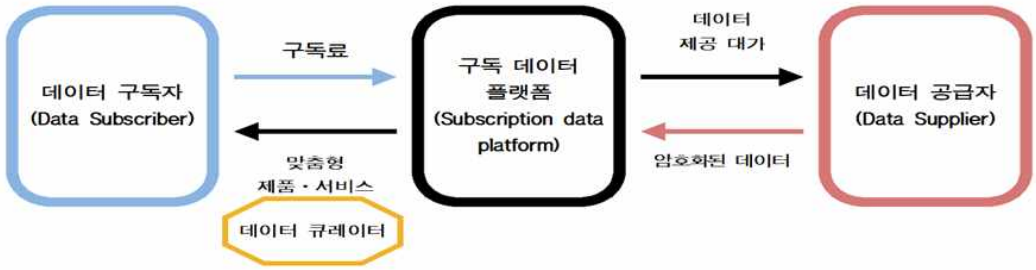
25) 동형암호에 대한 상세한 설명은 2장을 참고하라.

사실상 데이터를 연산할 수 있는 권한만을 가지는 것이다. 데이터 분석 이후에 다시 데이터 제공자에게 키를 요청해야 하기 때문이다.

○ 복호화 과정에서 다시 한번 거치는 과정은 정기적으로 데이터 제공자와 관계를 가지는 프로토콜인 구독 컨셉과 연결된다.

나. 기본 구조도

□ 아래 <그림 4-1>은 구독통계 플랫폼의 기본구조도를 나타낸다. 플랫폼 운영에 있어서의 핵심은 단순 데이터 제공이 아니라 다양한 데이터에 쉽게 접근하고, 편리하게 상용할 수 있도록 서비스 제공한다는 데에 있다.



<그림 4-1> 구독통계 플랫폼 기본구조도

□ 데이터 구독자(이용자)는 소정의 정기(년 또는 월) 구독료를 내고 데이터를 구독한다. 데이터 구독 등급은 데이터의 종류, 데이터의 양, 데이터 민감도, 데이터 시차특성에 따라 여러 가지 단계로 구분하여 운영할 수 있을 것이다. 여기에서는 가장 기본적인 3단계로 나누어 본다.

○ 첫 번째 단계는 무료회원 단계이다. 이 단계에서는 현재 통계청의 마이크로 데이터 제공 수준을 회원 가입 후 무료로 이용하는 것이다.

○ 두 번째 단계는 민간 데이터까지 접근 가능한 등급이다. 이 등급에서는 국민연금 등 양질의 공공 데이터뿐 아니라 신용카드 데이터 등 민간 데이터까지 접근이 가능하도록 한다. 다만 최신 데이터 이용에는 제한을 두어 최근 1년~2년 이내의 데이터에는 접근하지 못한다.

○ 마지막 세 번째 단계에서는 시차 없는 최신 데이터까지 모든 데이터에 접근이 가능한 단계이다. 최근 1년 이내에 갱신된 데이터까지 활용하여 사업현장에서 바로 적용할 수 있는 데이터의 접근 권한까지 가진다.

□ 데이터 공급자는 보유하고 있는 양질의 데이터를 동형 암호화하여 플랫폼에 제공한다.

○ 동형암호 형태로 데이터를 제공하도록 함으로써 데이터 제공에 따른 데이터 유출 등의 위험이 원천 차단[26]된다. 많은 공공기관 또는 기업이 데이터 외부 제공을 꺼리는 이유가 데이터 유출 등의 위험이라는 것을 고려하면 더 많은 데이터를 상대적으로 쉽게 구독 플랫폼으로 집적할 수 있는 환경을 만들 수 있다.

○ 데이터 제공 댓가는 데이터에 합당한 가격을 지불하는 방법과 현물로 지불하는 방법으로 나눌 수 있다.

- 합당한 가격을 지불하는 방식이 거래에 있어서는 합리적 선택으로 보인다. 그러나 여기에는 합리적 데이터 가격 책정 이슈가 다시 불거진다는 단점이 있다.

- 두 번째는 현물로 지불하는 방법이다. 현물로 지불하는 방법은 데이터를 제공하게 되면 구독 플랫폼에 있는 다른 데이터를 무상으로 이용할 수 있도록 하는 것이다. 이는 데이터의 품질에 따라 구독 데이터의 등급도 차별화되도록 한다. 물론 소정의 비용을 부담하면 구독 데이터의 등급은 상향 가능할 것이다. 또한 데이터 큐레이터 서비스를 이용할 수 있는 쿠폰을 지급하는 것이다. 실제 운영에 있어서는 다양한 방식의 조합이 가능할 것이다.

□ 플랫폼 운영자는 전체 데이터 거래 생태계가 효율적으로 운영될 수 있도록 관리한다.

○ 데이터 제공자들과 협력하여 데이터 정제가 완료된 데이터를 암호화하여 이후 사용 단계에서 문제가 최소화되도록 관리한다.

○ 복호화키 관리 역시 매우 중요하다. 데이터의 민감도에 따라 플랫폼 관리자가 복호화키 제공 권한을 가질 것인지, 데이터 원 제공자가 통제할 것인지, 아니면 상황에 따라 협의하여 진행할 것인지에 대한 관리가 필요하다.

○ 데이터 보험제도 운영이다. 플랫폼은 데이터 보험에 가입하던지 아니면 데이터 보험을 자체적으로 운영하는 방법도 생각해 볼 수 있다. 특정 기업 데이터 활용 중 개인정보 유출 등의 문제가 발생하거나 개인정보가 식별되는 상황이 발생할 수 있다. 이 경우 데이터 제공자의 잘못이 없음에도 불구하고 데이터 제

---

26) 앞서 2장에서 살펴본 바와 같이 동형암호는 암호화된 상태에서 연산, 처리가 가능하기 때문에 플랫폼을 관리하는 내부 직원조차도 복호화시키지 않고 관리할 수 있다. 현실적으로 모든 것은 암호화된 상태에서 처리할 수는 없겠지만 현재 암호화 기술보다는 복호화키가 필요한 상황이 현저히 줄어들 것임은 자명하다.

공 기업 이미지 등에는 큰 타격을 입을 수 있다. 이러한 상황에 유연하게 대응할 수 있도록 데이터 보험 제도를 운영한다고 하면 데이터 이용자나 데이터 제공자 모두 안심하고 데이터를 사용할 수 있을 것이다.

□ 구독 통계 플랫폼를 통한 데이터 거래 생태계의 완성 단계는 소위 '데이터 큐레이터'라고 불리우는 데이터 분석 기업이 자생적으로 활동할 수 있는 여건 조성과 연관이 있다.

○ 데이터 큐레이터는 단순히 데이터 목록, 위치 등을 확인하고 알려주는 부차적인 업무를 수행하는 기업이 아니다. 데이터 제공, 처리, 분석, 적용의 모든 단계에서 적극적으로 개입하여 제공자, 이용자 모두에게 서비스하는 데이터 처리 종합 솔류션 기업이라고 할 수 있다.

○ 데이터 제공 단계에서는 제공 기업이 플랫폼 프로토콜에 맞도록 효율적으로 데이터를 정제할 수 있도록 지원하고 서비스한다. 플랫폼에 잘 정제된 데이터가 업로드되어야 이후 연계, 결합 등 데이터 처리 과정이 매끄럽게 진행될 수 있다.

○ 데이터 처리 단계에서는 이용자들이 원하는 형태로 데이터를 연계해주고 메타데이터, 기초통계 산출 등을 지원한다.

○ 데이터 분석 단계에서는 이용자들의 요구사항을 듣고, 그에 합당한 분석 서비스를 진행한다.

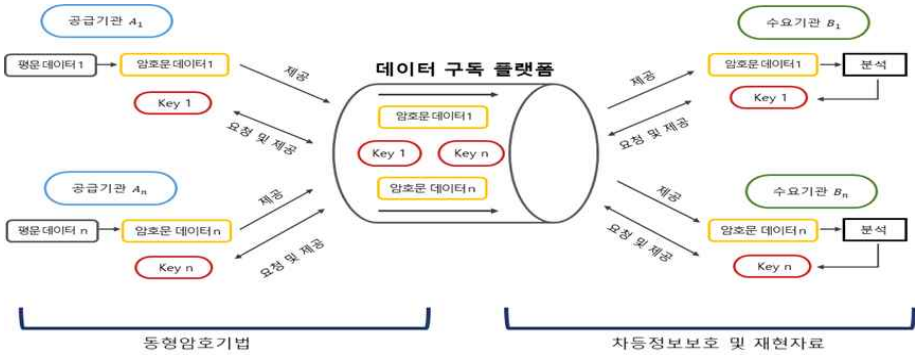
○ 데이터 적용 단계에서는 이용 기업의 데이터 활용 가능 상태 등을 진단하고, 실무에서 적용하기 위한 방안을 제시한다.

다. 데이터 중심 구조도

□ <그림 4-2>는 데이터의 흐름을 구체적으로 보여주는 데이터 중심 구조도이다.

○ 플랫폼을 중심으로 좌측(공급자측)은 동형암호 기반으로, 우측(수요자측)은 차등정보보호, 재현자료 솔루션이 중심이 된다.

○ 모든 데이터는 동형암호화된 상태로 온라인상에서 유통·활용되는 것을 지향한다. 효율적 데이터 분석을 위하여 민감정보가 아닌 일부 데이터는 평문 상태로 제공이 가능하다. 암호화 여부의 판단은 플랫폼 주관하에 데이터 제공자와 협의하여 결정한다.



<그림 4-2> 구독통계 플랫폼 데이터 중심 구조도

□ 공급기관은 제공할 데이터에 대하여 플랫폼에서 효율적으로 활용할 수 있도록 데이터를 가공한다.

○ 우선 평문상태에서 오기, 이상치 등 관련 1차 데이터 정제를 실시하고, 결측치 등 관련 정보를 수집한다.

○ 데이터 정제 후 동형암호화한다. 상황에 따라 민감정보는 동형암호화하고 그렇지 않은 데이터는 평문상태로 둔다.

○ 어떤 수준의 분석에서 복호화키를 허용할 것인가에 대한 내부적 가이드라인을 마련한다. 물론 플랫폼 운영기관에서 관련 정보를 제공하거나 협조한다.

○ 향후 데이터 연계 등이 가능하도록 식별자를 관리한다.

□ 수요기관은 데이터 활용 및 분석 목적을 세우고, 활용 데이터를 플랫폼에서 찾는다.

○ 데이터 탐색 후 연계가 필요한 경우에는 플랫폼에 연계를 요청한다.

○ 플랫폼에서 제공받은 재현자료를 이용하여 데이터 분석을 실시하거나 데이터 큐레이터 기업에 분석을 의뢰한다.

○ 재현자료를 이용하여 분석을 위한 코딩을 마치고 대략적인 결과를 예상한다. 또는 데이터를 실무에 적용할 수 있는 가능성 여부와 적용 시 효과를 판단한다.

○ 재현자료로 분석이 끝나면 플랫폼에 접속하여 암호화된 원자료를 분석한다. 또는 실무에 적용하기 위하여 동형암호화된 데이터를 요청한다.

○ 플랫폼에 분석된 자료의 결과 값을 요청하여 결과를 입수한다.

□ 플랫폼은 데이터 관리, 복호화키 관리, 재현자료 및 차등정보보호 적용 등의 기능을 수행한다.

○ 플랫폼은 다양한 종류의 데이터를 관리한다. 메타데이터를 체계적으로 관리하여 이용자들이 편리하게 데이터를 검색하도록 지원한다. 또한 연계가능한 데이터 목록 등을 미리 파악하고 공지하여 이용자 요구 시 즉시 대응할 수 있도록 한다.

○ 복호화키 관리 운영방안을 만들고 시행한다. 복호화키 제공 등급과 권한은 데이터 제공업체와 협의하여 최종 확정한다.

○ 제공된 자료에 대한 재현자료 특히 연계 후의 재현자료 등을 구현하여 이용자들이 데이터 요구 시 즉시 제공할 수 있도록 한다.

○ 이용자들의 분석결과 반출 요청 시 개인정보 유출 가능성을 검토하고 차등정보보호 솔루션을 적용하여 반출을 허가한다.

□ 데이터 큐레이터는 플랫폼 전·후방의 모든 과정에 관여할 수 있다.

○ 기업 단위에서 어떤 목적의 분석을 빠르고 효율적인 분석을 원할 경우 큐레이터는 현재 플랫폼에 있는 데이터로 실행 가능성을 검토하고 분석 방향을 제시한다.

○ 플랫폼과 협력하여 데이터 제공 기업에는 데이터 제공 시 잇점 등을 설명하고, 요청 시 데이터 제공을 위한 데이터 사전 정제 작업 등을 지원한다.

○ 데이터 큐레이터는 플랫폼 이용자들의 니즈(needs)를 반영할 수 있도록 지속적으로 활동한다.

4. 기대 효과

□ 본 절에서는 데이터 구독 플랫폼의 기대효과를 수요공급, 참여주체, 생태계 측면에서 살펴보고자 한다.

4.1 수요 공급 측면

가. 수요 측면의 애로사항

□ 첫째, 가장 높은 애로사항 응답률을 보인 ‘쓸 만한 양질의 데이터 부족’에 대하여 살펴보자.

○ 동형암호화된 상태로 데이터를 제공받기 때문에 ① 플랫폼 관리자조차도 데이터에 있는 정보에 접근이 불가능 하다는 점 ② 데이터 유출이 있더라도 암호화된

상태이기 때문에 개인정보 유출은 있을 수 없다는 점에서 데이터 제공자들의 제공 부담은 현격히 줄어든다. 즉 별도의 보안구역 또는 별도의 기관에서 관리하고 있는 복호화키가 동시에 유출되지 않는다면 개인정보유출로부터 안전하다.

○ 또한 데이터 큐레이터 서비스를 통하여 데이터 활용이 용이하도록 서비스한다. 데이터가 좋은 정보를 담고 있는 것도 중요하지만 데이터를 사용할 수 있도록 정제·연계 등의 서비스 확대를 통한 활용도 제고에도 무게를 실어야 한다.

□ 둘째, ‘불합리한 가격’에 대하여 고려해 보자.

○ 앞에서도 언급하였듯이 데이터 수요자와 공급자는 테이터를 바라보는 관점이 완전히 다르기 때문에 둘 간의 가격 괴리 현상은 필연적이다. 즉 공급자 입장에서의 합리적 가격 책정 방법은 수요자의 공감을 사기 어렵다.

○ 이 문제는 개별 데이터에 대한 가격책정이 아니라 구독이라는 형태로 데이터 공급 방법을 바꾸면 원천적으로 해결된다. 개별 데이터별 가격 협의가 아니라 소정의 구독료만 책정하고 지불하면 되기 때문이다.

□ 셋째, ‘유통채널 부족’, ‘데이터 소재파악 어려움’, ‘거래방법 미숙’이다.

○ 현재와 같이 금융 데이터, 공공기관 데이터 등 분절적으로 데이터 거래소를 운영하는 것이 아니라 구독통계 플랫폼이라는 단일 플랫폼에 다양한 종류의 데이터를 집적하여 서비스하기 때문에 위의 애로사항들은 쉬이 해결할 수 있다.

○ 물론 데이터 집적에 따른 정보 유출, 또는 빅브라더가 이슈가 될 수 있다. 그러나 동형암호의 특성상 데이터는 유출이 되어도 정보를 볼 수 없고, 유출된 정보를 연계시키는 것은 더더욱 불가능하다. 또한 플랫폼 관리자도 암호화된 데이터만 관리할 뿐 데이터의 정보를 확인할 수 없기 때문에 빅브라더 이슈로부터도 완전히 자유롭다.

□ 넷째, ‘데이터 품질, 가공문제’ 및 ‘개인정보 처리’의 어려움 등이다.

○ 데이터 수집, 분석, 해석은 다양한 분야의 전문적인 역량을 필요로 한다. 데이터 전문 분석가 기업인 데이터 큐레이터가 활동할 수 있는 데이터 거래 생태계를 형성하고 활성화시킴으로써 이 문제에 대하여 대응이 가능하다.

○ 데이터 큐레이터는 구독 플랫폼뿐 아니라 데이터 거래 활성화에 가장 핵심적인 기능을 하는 주체이다. 이들이 적극적으로 데이터를 분석하여 부가가치를 높일 수 있도록 특별한 지위를 부여하고, 이들에 대하여 적극적인 감독기능을 수행한

다면 안전하면서도 데이터 활용성을 높이는 두 마리 토끼를 잡을 수 있을 것으로 기대한다.

<표 4-2> 수요 측 애로사항 및 해결경로

애로사항	해결 경로
① 쓸 만한 양질의 데이터 부족	• 암호화된 상태의 데이터 제공으로 제공 부담 완화 • 데이터 큐레이터를 통한 맞춤형 서비스
② 불합리한 가격 ③ 유통채널 부족 ④ 데이터 소재파악 어려움 ⑦ 거래방법 미숙	• 합리적 구독료 책정으로 가격문제 및 거래문제 해소 • 데이터를 하나의 플랫폼에 집중하여 거래하게 함으로써 유통채널, 데이터 소재 파악 어려움 문제 해소
⑤⑥ 데이터 품질, 가공 문제	• 데이터 큐레이터스타트업 육성
⑧⑨ 개인정보 처리 등	• 동형암호 및 차등정보보호

나. 공급 측면의 애로사항

□ 첫째, ‘데이터 재제공을 통한 시장교란’ 행위이다.

○ 데이터 재제공은 정상적인 시장 거래를 불가능하게 만드는 요인이다. 시장 가격에 신뢰를 줄 수 없게 만들기 때문이다. 그런데 구독 플랫폼에서는 동형암호화된 상태로 서비스되기 때문에 데이터 재제공은 전혀 문제가 되지 않는다. 데이터 제공만으로는 그 어떠한 분석 결과도 얻지 못하기 때문이다. 결국 역설적이게도 암호화가 데이터를 자유롭게 만든다.

□ 둘째, ‘개인정보보호 유출 우려’이다.

○ 동형암호가 데이터 자체의 유출에 따른 위험을 원천적으로 차단하는 기능이라고 하면, 차등정보보호는 합법적인 데이터 분석으로 반출한 결과 값으로부터 유출될 수 있는 개인정보를 보호하기 위한 방법이다.

○ 특정 데이터를 반복적으로 분석하고 결과를 반출할 경우 그 결과 값들의 결합으로 개인정보를 식별 가능한 경우가 발생할 수가 있는데 이러한 상황을 미연에 방지하고자 하는 것이 차등정보보호 등 개인정보 보호 솔루션이다.

○ 따라서 동형암호와 데이터 정보보호를 플랫폼 전·후방에서 동시에 적용함으로써 개인정보 유출을 방지할 수 있다.

□ 셋째, ‘데이터 제공을 위한 데이터 표준화, 정제 등 추가적인 작업 소요’ 관련 사항이다.

○ 데이터를 공급하는 공공기관의 경우 현재 그 성과는 데이터제공 횟수나 개수 등을 중심으로 평가되는 경향이 있다. 이러한 상황에서는 데이터를 정제·연계해야 할 유인이 없을 뿐 아니라 활용하기 용이한 오픈 API 등을 제공할 유인이 전혀 없다. 이보다는 오히려 가지고 있는 데이터를 쪼개어 데이터 제공 횟수나 개수를 늘릴 유인만이 있을 뿐이다.

○ 그런데 동형암호화된 데이터의 경우는 최종 사용자가 결과를 확인하기 위하여 킷값을 요구하거나 최종 파일 복호화를 요청해야 한다. 즉 최종 활용 결과를 기준으로 실적을 파악할 수 있다는 장점이 있다.

○ 분석·활용 결과를 기준으로 실적을 평가하게 되면 데이터를 제공하는 기관들은 이용자들이 더 쉽게, 더 많이 사용할 수 있도록 제공할 유인이 생기게 된다. 이 경우 공공기관 등에서는 데이터 정제, 표준화, 연계 등에 더 많은 인력과 예산을 투입하여 활용도를 높이고자 할 것이다.

○ 여기에서도 데이터 큐레이터의 역할이 중요하다. 공공기관 등에서 활용성이 높은 데이터를 제공하려고 해도 그 방향성을 잡기가 어려울 것이다. 데이터 큐레이터들이 그 방향성을 제안하고, 서비스하는 영역에도 일정 부문 역할을 할수 있을 것이다.

<표 4-3> 공급 측 애로사항 및 해결경로

애로사항	해결 경로
ⓐ 데이터 재제공을 통한 시장교란	• 동형암호 기반 구독 방식으로 재제공 원천 금지됨
ⓑ 개인정보보호 문제 우려	• 동형암호 및 차등정보보호 • 데이터 보험(개인정보 유출대비)
ⓒ 데이터 제공을 위한 데이터 표준화, 정제 등 추가 작업 소요	• 데이터 활용량(제공량이 아닌)에 따른 공공기관 인센티브 제공 • 데이터 큐레이터스타트업 육성으로 별도 시장 개척

4.2 시장참여 주체별 기대효과

□ 데이터 생태계에서 자본의 흐름, 이윤의 배분 과정을 이해하는 것도 중요하다.

○ 첫 번째로 말하고 싶은 것은 데이터 거래의 현금 흐름 속에서 부가가치를 창출

하는 주체는 데이터 가공·분석 집단인 데이터 큐레이터여야 한다는 점을 명확히 하고 싶다. 데이터 큐레이터가 데이터 시장에서 기업의 영업활동을 통해서 이윤을 창출할 수 있을 때 데이터 구독 플랫폼 및 거래 시장은 점차 확대되어갈 수 있기 때문이다.

○ 그렇다면 나머지 참여 주체들은 어떻게 이익을 얻게 될 것인가? 구독자는 데이터 분석 및 활용을 통해서 자연스럽게 데이터 거래를 통한 이익을 추구할 수 있음을 알 수 있다. 데이터 공급자와 데이터 플랫폼 사업자는 데이터 거래 시장이 커지고, 데이터 이용 활성화에 따른 전반적인 시장 효율성 제고를 통해서 이익을 가져가는 구조를 갖추어야 한다. 데이터 큐레이터 기업이 부가가치를 창출할수록 데이터 거래 시장의 효율성은 높아지는 선순환구조를 만들어 내는 것이 중요하다.

□ 시장 참여 주체별로 애로사항과 구독통계 플랫폼이 어떻게 대응할 수 있는지에 대하여 정리하였다.

<표 4-4> 시장 참여 주체별 기대효과

참여자	주요 특징 및 장점
데이터 구독자	• 합리적 비용으로 다양한 데이터 연계 및 활용 ※ 시장 형성 초기에는 정책지원 등을 통한 합리적 비용 또는 무료 구독 데이터 제공 필요 • 하나의 플랫폼 내에서 데이터 검색 가능 • 서버접속이나 특정 공간 방문 없이 데이터 연계 및 사용 • 데이터 큐레이터를 통한 고급 서비스 이용 가능 • 암호화 상태로 기계학습이 가능해 활용도 높음 • 구독자 소유 데이터를 플랫폼 관리자에 제공 없이 연계 가능(외부 데이터 연계 시)
데이터 공급자	• 암호화된 형태의 데이터 제공으로 데이터 공급자 제공 부담 완화(완벽한 확인) • 플랫폼(또는 데이터 큐레이터와 연합)의 데이터 보험 제공으로 개인정보 유출 부담 경감 • 양질의 다양한 데이터 제공 가능성 향상 • 데이터 보안 등급에 따른 차등 권한 부여 가능 • 데이터 활용 현황(제공 현황이 아닌) 파악이 명료하므로 성과관리 용이
플랫폼 사업자	• 개별 데이터에 대한 가격 책정 부담 없음 • 개인정보(차등정보보호, 재현자료) 보호 기반 효율적, 안정적 서비스
데이터 큐레이터	• 데이터 정제, 연계, 검색, 추천 등 맞춤형 서비스 • 자연스러운 데이터 큐레이터스타트업 생태계 형성 → 일자리 창출
데이터 거래시장	• 데이터 거래시장 양립 가능(플랫폼 거래 및 자사 데이터 자체 거래 양립) • 데이터 재제공이 원천 차단됨에 따라 시장 교란 요소 제거됨

○ 앞서 수요자 및 공급자 애로사항 및 해결경로와 중복되는 부분은 생략하고 몇 가지 추가적인 사항들만 설명한다.

○ 동형암호화된 상태로 기계학습이 가능하다는 점이다. 즉 암호화된 상태에서 기계학습이 가능하다는 것은 거의 대부분의 작업이 가능하다는 것을 의미한다.

○ 또한 데이터를 이용자에게 직접 제공하는 제도인 만큼 이용자가 본인의 데이터를 플랫폼에 제공하는 절차 없이 데이터 연계가 가능하다. 매우 민감한 데이터를 가지고 있어 공공 플랫폼에서 데이터를 공유를 꺼리는 경우에도 동형암호 기반 구독 플랫폼 시스템하에서는 활용이 가능하다.

○ 데이터 보안 등급에 대한 차등권한 부여, 데이터 활용 현황에 따른 실적 집계 등은 앞서 언급한 바와 같다.

4.3 거래생태계 구축효과

□ 거래 생태계구축에 있어서는 네트워크 효과를 극대화시키는 것이 중요하다. 데이터 거래 생태계 구축 과정은 다음과 같다.

○ 초기 데이터 거래 및 이용을 활성화시키기 위하여 양질의 데이터를 거의 무료로 이용할 수 있는 환경을 구축한다.

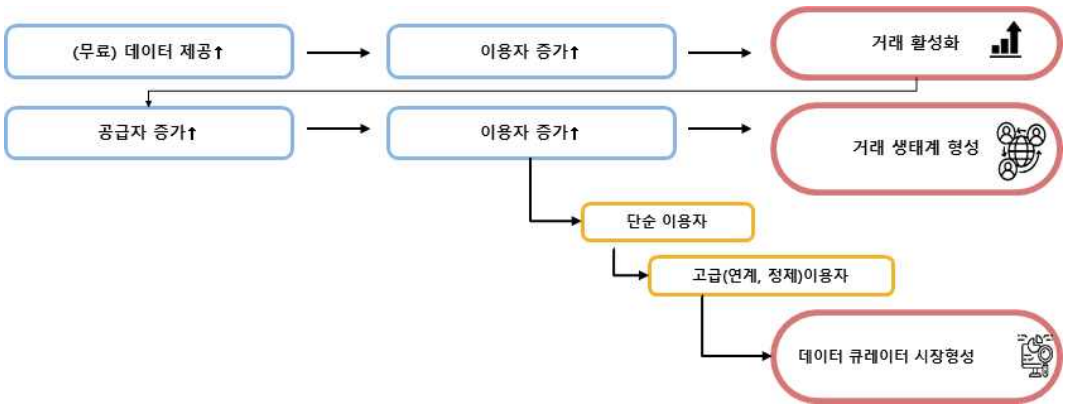
○ 양질의 데이터가 무료인 환경 속에서 다양한 데이터 분석 및 활용 결과들이 나오게 되면서 이용자는 자연스럽게 증가한다.

○ 이용자가 증가하게 되면 데이터 공급자 입장에서도 데이터 공급 시 수익 창출 등 더 많은 부가가치를 증가시킬 기회를 포착하게 되고 더 많은 데이터를 공급할 유인이 생긴다.

○ 이렇게 되면 다시 이용자가 증가하여 자연스럽게 데이터 거래 및 이용은 활성화된다.

○ 이 과정에서 단순히 데이터를 이용하거나 분석하는 차원을 넘어 직접 사업에 적용 또는 신속·정확한 분석 서비스에 대한 수요가 생기게 되면 데이터 큐레이터 시장이 형성된다.

○ 아래 <그림 4-4>에서는 위에 설명한 과정들을 잘 나타내주고 있다. 데이터 생태계의 형성의 마지막 단계는 데이터 큐레이터가 자생할 수 있는 환경이 만들어지는 것이라 할 수 있다.



<그림 4-4> 데이터 거래 생태계 구축과정

제 5 장

결론 및 시사점

□ 데이터 암호화 및 개인정보보호 조치는 역설적이게도 데이터 이동 및 활용을 자유롭게 한다.

○ 동형암호화된 데이터는 암호화된 상태로 연산이 가능하여 데이터 유출로 인한 개인정보 노출이나 빅브라더 이슈로부터 자유롭다.

○ 데이터 정보보호 검증을 거친 데이터의 반출은 개인정보 노출위험을 제어한다.

○ 데이터의 자유로운 활용은 궁극적으로 정보의 독점으로부터 야기되는 지대추구 행위 등의 비효율성을 제거하므로 효율적인 시장경제를 만들어 갈 수 있다.

□ 제2장에서는 동형암호에 대하여 논의하였다.

○ 기존 암호의 개요에 대하여 설명하고, 기존 암호의 한계를 설명한다. 암호화 상태로 연산 가능한 동형암호의 특성을 설명하고 지속적 연산이 가능한 완전동형암호에 대하여 설명한다.

○ 동형암호의 활용에서 암호기술의 텃밭으로 불리는 군 보안 시설뿐 아니라 신용정보 계산, DNA 분석 등 산업 현장에서의 활용 사례를 제시한다. 동형기계학습 과정과 다자간 계산의 가능성에 대하여도 설명한다.

□ 제3장에서는 데이터 정보보호를 다루었다.

○ 데이터를 통한 개인정보 노출에 대해 경각심을 일깨우기 위하여, 데이터 유형에 따른 정보노출 가능성을 예시를 들어 논하였다. 이러한 가능성에 대한 일반적인 인식을 제고할 필요가 있기 때문이다. 다음으로 노출위험과 정보손실을 설명하였다. 이는 정보보호 정도를 판단하기 위해 필수적인 개념이다.

○ 한편 정보보호를 위한 신기술로 활발하게 연구되고 있는 재현자료와 차등정보보호를 설명하였다. 더불어 대용량 빈도표 제공 시 비밀보호를 위하여 개발된 알고리즘도 소개하였다. 이러한 기술들에 관한 내용을 상세히 다루고 있는 국문

보고서도 함께 소개하였으므로 전문적인 내용을 더욱 숙지하기 원하는 독자는 언급한 각 국문 보고서에 수록된 참고문헌을 활용할 수 있다.

○ 마지막으로 데이터센터의 반출물에 대한 통계적 노출제어 가이드라인 활용을 설명하였다. 이러한 가이드라인의 작성 및 시스템화는 데이터센터 운영에 있어 필수적이며, 반출사례를 관리하고 가이드라인을 꾸준히 업데이트해 나가는 것은 노출 사고 방지를 위하여 매우 중요하다. 3장의 설명은 데이터 정보보호에 관한 전체적인 맥락을 파악하기 원하는 일반 독자를 위하여 작성되었으며, 전문성을 가진 독자들은 언급된 문헌들을 활용하기 바란다.

□ 제4장에서는 위의 두 방법을 접목한 데이터 구독 플랫폼에 대하여 설명하였다.

○ 본 플랫폼을 통하여 거래함으로써 ‘쓸 만한 양질의 데이터 부족’ 등 거래 시 발생할 수 있는 애로사항 9가지를 모두 해결할 수 있는 해결 경로를 제시하였다.

○ 데이터 구독 플랫폼 기본 구조도와 데이터 중심 구조도 설명을 통하여 플랫폼 내에서 수요자·공급자·데이터 큐레이터 등 각 주체와 데이터의 움직임에 대하여 설명하였다.

○ 마지막으로 데이터 거래 생태계 구축 과정을 보여주면서, 생태계가 자생적으로 움직이기 위해서는 수요시장과 공급시장 간 네트워크 효과를 극대화시키는 방향으로 정책방향을 설정해야 함을 설명하였다.

□ 데이터 구독 플랫폼은 데이터 거래 활성화뿐 아니라 시장 생태계를 형성하기 위한 핵심적인 기능을 담당할 수 있다.

○ 정책적 드라이브로 일시적으로 데이터 거래시장이 형성될 수는 있다. 그러나 하나의 경쟁력 있는 산업으로 발돋움하고 장기적으로 성장해나가기 위해서는 시장 경제 내에서 스스로 작동할 수 있는 자생적 거래 생태계 형성이 필수적이다.

참고문헌

관계부처 합동(2016). 개인정보 비식별화 조치 가이드라인 - 비식별 조치 기준 및 지원・관리 체계 안내-.

관계부처 합동(2020).『한국판 뉴딜』종합계획 - 선도국가로 도약하는 대한민국 대전환 -.

개인정보위원회(2020). 가명정보처리 가이드라인.

김명환(2016). **DNA 분석 및 생체인증 보호를 위한 동형암호 실용화 기술개발**, 서울대학교 산학협력단.

박민정(2014). 가계금융·복지조사 매스킹 방안 평가, **2013년 연구보고서**, 통계개발원.

박민정(2015). 마이크로데이터 매스킹 기법 실무 활용 안내서, **2014년 연구보고서**, 통계개발원.

박민정(2016). 통계적 노출제어 최신동향 검토 - 차등정보보호를 중심으로-, **2015년 연구보고서**, 통계개발원.

박민정, 김정연(2017). 재현자료 작성 방법론 검토, **2016년 연구보고서**, 통계개발원.

박민정, 박영옥, 황영자(2017). 통계데이터센터 반출 결과물의 통계적 노출제어 가이드라인(안), **2016년 연구보고서**, 통계청 내부발간.

박민정, 신우람, 박천영, 서영오(2018). 그리드별 통계서비스의 비밀보호 방안, **2017년 연구보고서**, 통계개발원.

박민정, 이용희, 권성훈(2018). 차등정보보호에 관한 연구, **2017년 연구보고서**, 통계개발원.

박민정, 권성훈, 정재호(2019). 차등정보보호 적용 실험 연구, **2018년 연구보고서**, 통계개발원.

박민정, 한정임, 박노성(2020). **통계데이터센터 DB를 활용한 재현자료 생성 방법 연구**, 통계개발원.

보도자료(2019). 한국데이터산업진흥원 '데이터스토어'의 누적거래량 11억원 수준, 송희경 의원실.

유성준, 박나리(2020). CART 기법을 이용한 개인신용정보 재현자료 생성, **통계연구, 25(1)**, 1-30.

이용희 (2013). 정보공개 환경에서 개인정보 보호와 노출 위험의 측정에 대한 통계적 방법, **한국데이터정보과학회지, 24(5)**, 1029-1041.

천정희(2012). 동형암호의 과거와 현재, **이화여자대학교 수리과학연구소 뉴스레터(창간호)**. 32-35

천정희, 어윤희, 김재윤(2018). **개인정보가 보호되는 동형암호기반 금융데이터분석**, 금융정보연구.

천정희(2020). **동형암호화된 데이터의 심층신경망 연산을 지원하는 완전 동형암호 기계학습 알고리즘 개발 및 라이브러리 구현**, 서울대학교 산학협력단.

한국데이터산업진흥원(2019). 2019 데이터 산업 백서, 한국데이터산업진흥원.

홍영희, 박민정(2020). **통계지리정보서비스(SGIS) 제공 자료의 비밀보호 적용 방안**, 통계개발원.

Cheon J., Kim A., Kim, M., & Song Y.(2017). *Homomorphic encryption for arithmetic of approximate numbers*. Takagi T., Peyrin T.(eds) Advances in Cryptology – ASIACRYPT 2017. ASIACRYPT 2017. Springer, Cham. 409-437.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & de Wolf, P. P.(2012). *Statistical Disclosure Control*, John Wiley & Sons Ltd.

Little, R. J. A.(1993). Statistical Analysis of Masked Data, *Journal of Official Statistics, 9(2)*. 407-426.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L.(2008). *Privacy: Theory Meets Practice on the Map, in Proceedings of the 24th International Conference on Data Engineering, Istanbul*, Turkey, 277—286.

Nissim, K., Steinke, T., Wood, A., Altman, M., Bembenek, A., Bun, M., & Vadhan, S.(2018). *Differential Privacy: A Primer for a Non-technical Audience, Privacy Tools for Sharing Research Data Project,* Harvard Univ., Boston, MA, USA, Tech. Rep. TR-2017-03.

Park, M. & Kim, H.(2016). Statistical disclosure control for public microdata: present and future, *Korean Journal of Applied Statistics, 39(6)*. 1041-1059.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y.(2018). Data Synthesis based on Generative Adversarial Networks, *VLDB, 11(10)*, 1071-1083.

Reiter, J. P.(2004). New approaches to data dissemination: A glimpse into the future, Chance, 17:3 (Summer 2004). 12-16.

Rubin, D. B.(1993). Discussion : Statistical Disclosure Limitation, *Journal of Official Statistics, 9,* 461-468.

Warner, S. L.(1965). Randomised response: a survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association, 60(309)*, 63–69.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K.(2019). *Modeling Tabular Data using Conditional GAN*, In NeurIPS.

https://blog.openmined.org/what-is-homomorphic-encryption/

Abstract

For data transaction activation Preliminary research on building subscription statistics platform

Junseok Byun, Minjeong Park, Jeonghee Cheon

The age when data utilization becomes a measure of corporate competitiveness. However, as the use of data increases, the importance of personal information protection is becoming more prominent. It is time for an innovative platform to catch the two rabbits, data utilization and privacy. In this report, we propose a data subscription platform to realize this. The subscription statistics platform aims to form a safe and efficient data transaction ecosystem by using homomorphic encryption in the process of data collection and distribution process, and data information protection after delivery as key solutions.

Chapter 2 discusses next-generation cryptography and homomorphic cryptography known as fourth-generation cryptography. Homomorphic encryption has the characteristic that it can be operated with the encrypted data itself, unlike existing encryption systems. In 2016, when Cheon-Kim-Kim-Song (2016) announced a homomorphic cipher called HEaaN, which enables not only addition and multiplication, but also the third operation of rounding in an encrypted state, this type of encryption became possible. Homomorphic cryptography has already been applied not only to traditional cryptography applications such as military security communications and missile hacking prevention, but also to general industries such as credit information calculation and DNA analysis. Even homogeneous machine learning, which performs machine learning in an encrypted state, is possible.

Chapter 3 discusses the necessity and methodology of data information protection. First, to help readers understand the necessity, through examples, the importance of the concept of measuring exposure risk and information loss is explained. Data information protection methodology (technology) deals with the development of BSCA, a confidentiality protection algorithm for providing reproduction data, differential information protection, and providing a large number of frequency tables. Furthermore, it emphasizes the need to operate statistical exposure control guidelines for exports in data center operation.

Chapter 4 describes a data subscription platform that innovatively combines the above two methods. In this platform, data owners provide homomorphic-encrypted data to the platform, and data subscribers can freely analyze the provided data in an encrypted state. However, the final result is exported with data information protection processing, and personal information is thoroughly protected during the data collection, distribution, and export process.

The data transaction market is difficult to spontaneously generate due to the re-provision of data due to easy movement and duplication, and the price gap resulting from the difference in perspective between the supplier and the consumer. On the other hand, in the data subscription platform, since the data is already homogeneous encryption, re-provision is meaningless, and since it is a membership subscription system, there is no need to negotiate individual prices between demanders and suppliers. In other words, since market functions can be started naturally on a secure subscription platform, it is possible to form a data transaction ecosystem that was not dreamed of due to security and price gap issues.

Data encryption and data protection processing for various types of exports, paradoxically, frees data movement and utilization. Homomorphic-encrypted data is distributed in an encrypted state, so neither the person in charge of the data nor the administrator can check the information. Data export that has been verified for data protection controls the risk of personal information exposure. Free data trading creates an efficient data trading market.

Key Terms: Homomorphic Encryption, Data Information Protection, Differential Information Protection, Reproduced Data, Data Subscription, Subscription Statistics Platform, Data Insurance

집필진

- 변준석 (통계청 통계개발원 경제사회통계연구실 사무관)
- 박민정 (통계청 사무관)
- 천정희 (서울대학교 자연과학대학 수리과학부)

연구보고서 2020-05

데이터 거래 활성화를 위한 구독통계 플랫폼 구축 사전 연구

인 쇄	2021년 4월 12일
발 행	2021년 4월 13일
발 행 인	통계개발원장 전영일
발 행 처	통계청 통계개발원 35220 대전광역시 서구 한밭대로 713 TEL.(042)366-7100 FAX.(042)366-7123
홈페이지	http://sri.kostat.go.kr
ISSN(Online)	2733-4120

통계청
통계개발원