중앙정부 및 지방자치단체가 보유한 각종 공공 데이터를 안전하게 연계하여

누구나 활용케 하는 K - 통계 시스템 (안)

통계청장 류 근 관

통계청

CONTENTS

통계청

요 약

요약

- 4차 산업혁명 시대에 데이터 이용 활성화를 위해 통계작성, 과학적 연구 등의 경우 정보주체의 동의 없이 가명정보의 활용이 가능토록하는 데이터 3법이 개정되면서 데이터 활성화에 대한 기대가 큼

- 그러나 개인정보 유출 및 데이터 오남용과 데이터 재식별화에 대한 우려가 커 현재의 가명처리 및 비식별화는 최선의 대안이 아님. 또한 중앙 및 지방정부에 분리 보관되어 있는 양질의 데이터의 활용 가능성을 배가할 필요

요약

- 이에 최근 연구가 활발하게 진행되고 있는 동형암호 등 최신의 암호기술을 활용하면 데이터 집중화에 따른 개인정보 유출이나 빅 브라더 출현 등 우려를 불식하면서도 규모의 경제와 범위의 경제로 대표되는 데이터의 잠재적 가치를 극대화할 수 있음

- 요컨대, 통합통계등록부를 이용한 각종 공공자료 결합의 법적, 제도적 틀을 구축하고, 이어 최신의 암호기술을 이용하여 최고의 보안 수준 유지한 채 각종 공공 자료를 결합 활용하는 한국판 공공 빅데이터 체계, 가칭 K-통계 체계 구축하고자 함

요약

- 이는 공공정책의 효율성과 형평성 제고는 물론 민간부분의 4차산업 활성화에도 크게 기여할 것임. 나아가 공공 빅데이터로 신규 창업 기업, 벤처기업을 적극 지원함으로써 기존 빅데이터 보유 플랫폼 기업과 신규 진입하는 기업간 자료 격차를 축소하여 경제 전반적으로 경쟁과 혁신의 생태계 구축에도 도움

- 또한 K-통계 체계는 암호 상용화를 위한 학문 발전, 산업계의 동형암호 가속기 등 개발에도 적극적인 인센티브를 제공하여 ICT 강국인 한국이 빅데이터 및 암호 상용화 분야에서도 세계에서 선도적 위치를 차지할 하나의 큰 계기

1. 데이터 활용을 위한 법개정

- 4차 산업혁명 시대에 전통적 데이터 외에도 인터넷, SNS, IoT 등 다양한 경로로 데이터가 수집됨

- 2019년 12월, 데이터 이용 활성화를 위한 데이터 3법(개인정보보호법, 정보통신망법, 신용정보법) 개정 및 제도 개선

  과학적 연구, 통계작성(상업적 목적), 공익적 기록보존 등의 목적으로 정보주체의 동의 없이 가명정보 활용 허용

  (개인정보보호법 : 개인정보 보호를 위한 법체계를 일원화하고 개인의 권익 보호를 강화하기 위한 법)

- 법 개정으로 대량의 데이터 저장 및 활용 확산될 것임

2. 한계

- 개인정보를 익명·가명화해도 다양한 원인에 의한 재식별화로 인해 프라이버시 침해, 법적 책임 문제 대두
  - 가명 내지 비식별화 처리로 인해 현 시점에서는 재식별 위험 없어도 추후 다른 정보와 결합하면 미래 시점에서 재식별될 위험은 상존
  - 이는 비모수 통계 분석에서의 curse of dimensionality와 같은 문제 (고차원에서 여러 조건을 모두 만족하는 cell 크기는 급감)
- 데이터 관리, 이동 과정에서 정보 유출 위험 상존

- 정보 활용 내지 결합 시 정보를 많이 남기면 개인 정보 침해 내지 재식별 위험 증대

- 정보 활용 내지 결합 시 coarse grouping 사용하면 정보 가치 훼손

현재의 가명처리 내지 비식별화 처리는 최선의 대안이 아님.
4차 산업혁명시대 데이터 확보, 공유, 정보보호를 위한
새로운 정보 보호/활용 체계 필요

3. 새로운 시도

- 각종 공공데이터는 가장 손쉽게 확보할 수 있는 공공재로서의 빅데이터 원천

- 양질의 데이터가 각 중앙부처 및 지방정부 독자 서버에 분리된 채 쌓여 있음
  - 하지만, 데이터 소스별 보관 방식, 표준화 등의 차이 데이터 보유기관 간 부처 이기주의 및 상호간 신뢰 부족

- 각 공공기관 보유 자료의 연계 및 결합 활용 확대 필요성
  - 자료 활용에 있어 자료가 모일수록 가치가 증가하는 규모의 경제는 물론 각종 자료를 결합할수록 가치가 배가되는 범위의 경제도 존재

- 중앙 부처, 지방 정부 보유 자료를 각각 암호화한 뒤 통합등록부 이용, 개인별 및 사업체별로 결합, 활용하는 K-통계 체제 구축 필요
- 추후 민간 보유 빅데이터와도 연결
- 각종 공공 자료를 보다 효율적으로 활용하려면 제반 자료 출처에 걸쳐 개인이나 사업체별로 결합하여 사용하는 게 필요
- 이를 위해 통합등록부를 축으로 하는 통계법 개정 필요
  (5월 국회 제출 일정으로 정부 입법 추진 중)

※ 통계등록부는 인구, 사업체 등에 대해 조사자료와 행정자료를 연계하여 작성한 통계단위별 모집단 자료.

1. 개요

- K-통계는 암호화 기법을 이용하여 중앙 부처, 지자체 보유 각종 공공 데이터를 안전하게 보관, 연결, 결합 활용케 하는 하나의 바람직한 국가 통계 체계
- 민감한 정보를 암호화한 뒤 암호화한 상태서 결합 및 각종 연산 수행
- 각 중앙 부처 내지 지방자치단체의 자료를 하나의 기관 내부로 전부 이관하거나 하나의 댐에 모두 통합할 필요 없이 통합등록부상의 암호화된 고유키로 연계
- 연결 활용된 분석 결과는 사회적 합의 기구를 거친 이후 복호화 함으로써 자료의 오남용을 방지하고 개인 정보를 보호할 수 있음

2-1. 평문(plain text) 자료 이용시의 단점

- 데이터 해킹 시 심각한 개인정보 침해 우려
- 데이터 주권이 개개 국민에게 존재하지 않음
- 자료 사용에 대한 포괄적 동의로 인해 제공되는 데이터의 범위, 종류, 용도 등을 특정하기 어려움

2-2. 암호문(cipher text) 자료 이용시의 장점

- 정보주체(data subject)인 국민의 데이터 권한 강화 및 보안성 개선

- 이는 장기적으로 개개 국민 내지 사업체 자료 이용에 대한 국민의 신뢰를 회복하여 공공기관의 자료 수집, 축적 및 공공 목적 활용에 기여할 것임

3. 암호문 체계의 양대 도구

3-1. 차분정보보호

Differential Privacy

3-2. 동형암호

Homomorphic Encryption

3-1. 차분 정보보호 Differential Privacy

- 원 자료나 결과값에 적절한 잡음(noise)을 추가. 원하는 분석 결과는 추가된 잡음의 영향을 거의 받지 않게 하면서 개개인의 정보만큼은 보호하는 방법

- 예컨대, 표본평균을 구하는 경우, n 명의 평균과 n-1 명의 평균을 상호 비교하면 제외된 한 개인의 사적 정보를 알게 됨.
  - 이때 표본평균의 결과 값에 약간의 잡음(e의 크기)을 추가해 발표하면 결과값에는 큰 영향이 없지만 두 결과값을 비교하여 한 개인의 값을 복원하려는 경우에는 아주 큰 잡음(n*e 규모의 크기)에 직면하게 되어 개인 정보는 보호됨

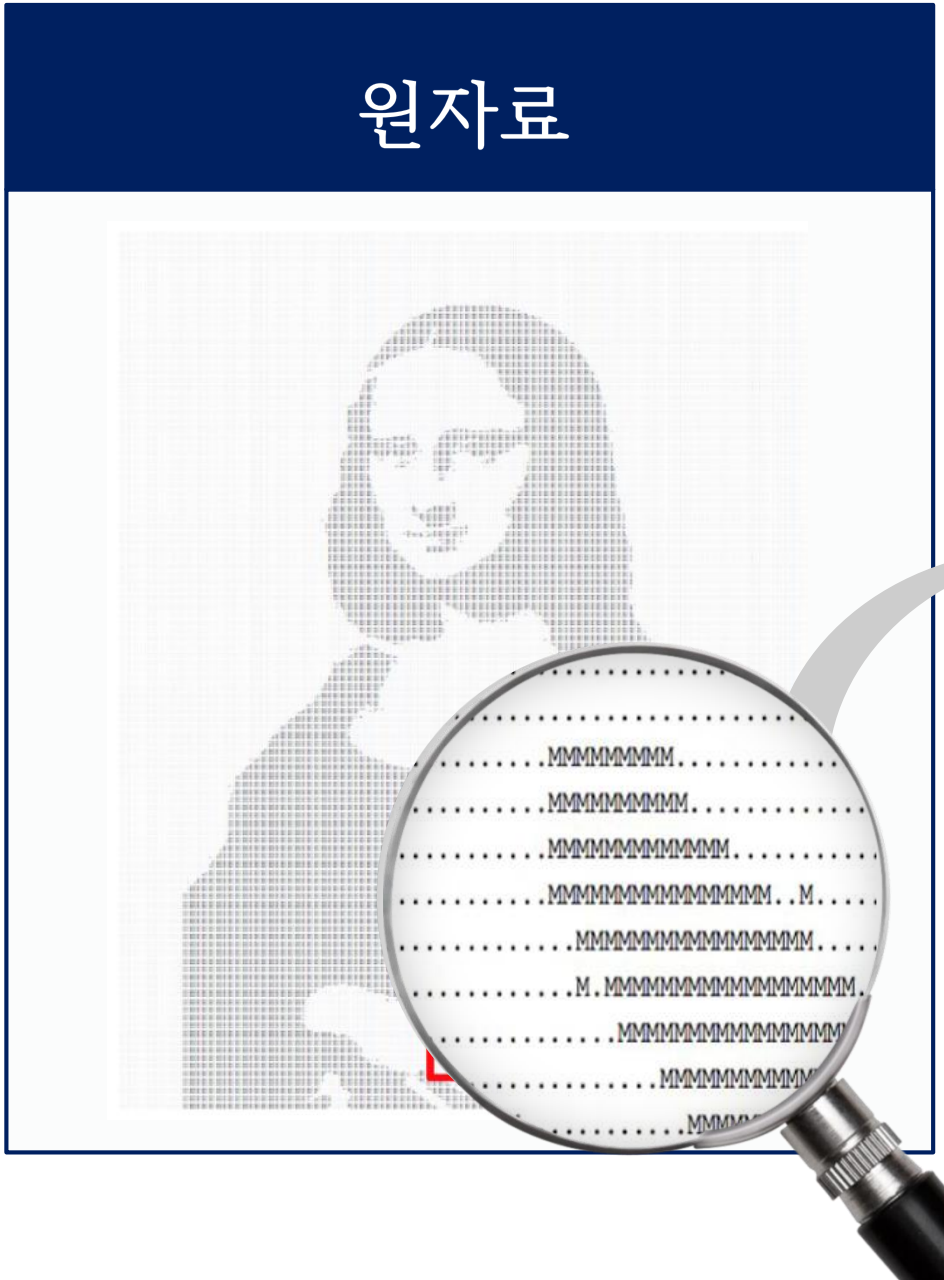
차분 정보보호의 적용사례

Global DP	2020년 US Census에 도입
Local DP	Google Chrome이 사용자 검색어를 수집할 때 사용

- 다만, 차분정보보호 방식으로는 확장된 계산 수행 시 안전한 결과값 도출 어려움

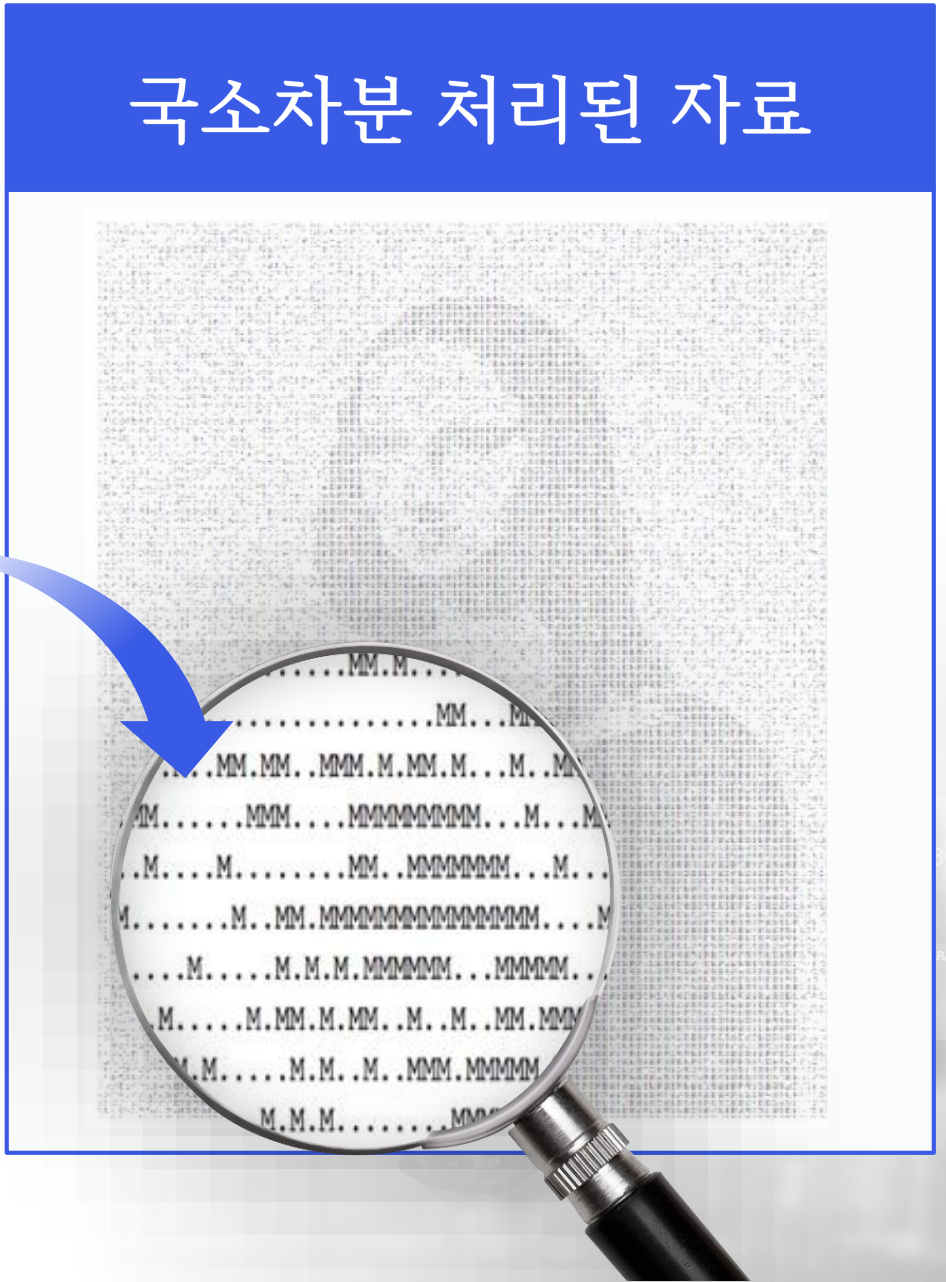
- 하나의 연산에서 개인 정보 보호하는 차분정보보호가 다른 연산에서도 개인 정보 보호하게 되는 것은 아님

원자료



+잡음

국소차분 처리된 자료



3-2. 동형암호 (One of 10 Emerging Technologies, MIT Technical Review, 2011)

- 암호화 후 연산한 값=연산 후 암호화한 값
- 즉, 암호화와 연산간 교환법칙 성립하는 암호 체계
- 동형암호의 이러한 특성은 연산을 위해 암호를 풀 필요가 없고 최종 연산 후 필요한 경우에만 복호화 하면 되므로 정보보호를 획기적으로 개선
- 통합등록부의 통계아이디 기반으로 각 공공기관에 산재된 각종 마이크로 데이터를 암호화한 채 결합하여 활용

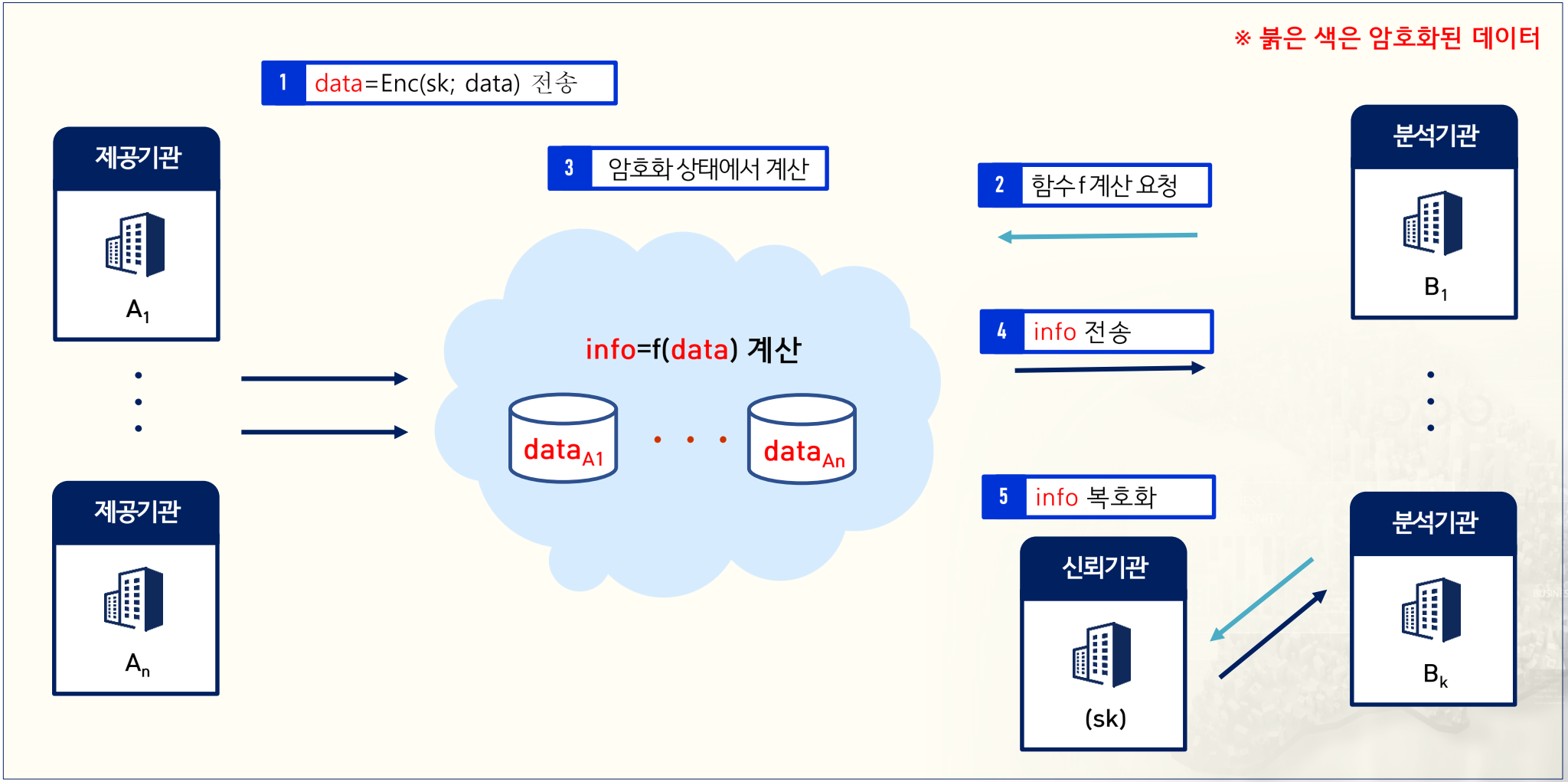
3-2. 동형암호 (One of 10 Emerging Technologies, MIT Technical Review, 2011)

- 이때 암호키 관리 체계는 사회적 합의 하에 수립

- 개별 공공기관의 자료가 해당 기관의 서버를 떠나지 않은 상태서도 기관간 자료의 결합 활용이 가능

동형암호 기반 데이터 공유하는 데이터댐

– 댐 속 물과는 다르게 각 기관 자료는 분리 보관됨



4. 차분정보보호+동형암호

- 다만 암호화하면 자료 사이즈 증대, 연산 속도 하락
- 해서, 차분정보보호 이용한 재현 데이터(synthetic data)로 자료 구조 및 속성 파악, 자료 전처리, 모형 구축 등 충분히 연습
- 이후 동형암호 상태서 전체 자료에 대해 모형 추정

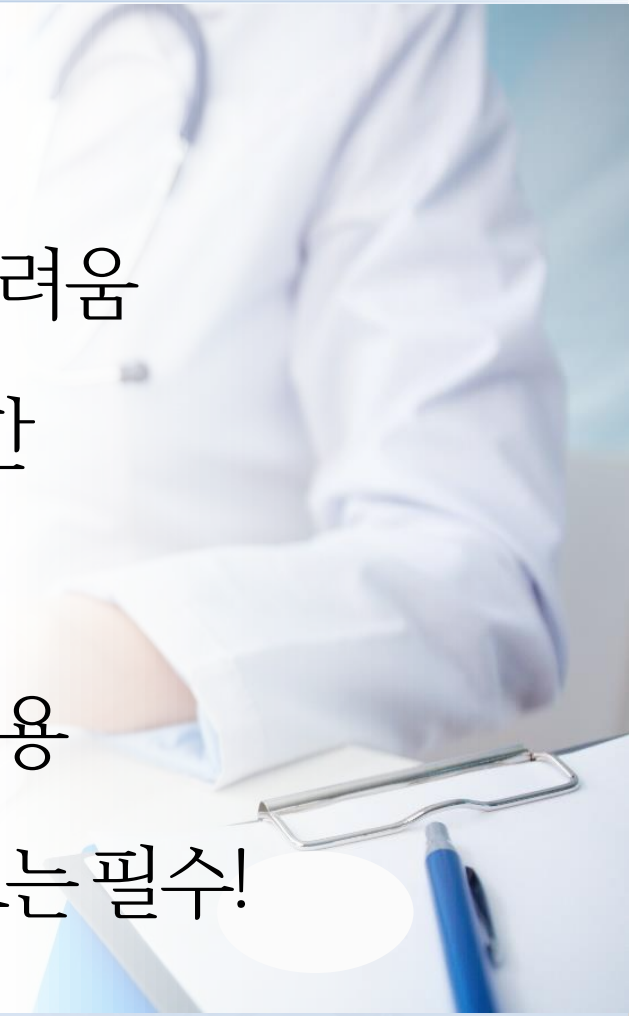
Ⅲ 활용 사례 및 기대효과

1. 의료 정보 결합

2. 국민 건강과 사생활 모두 보호

3. 근거 기반 정책 집행

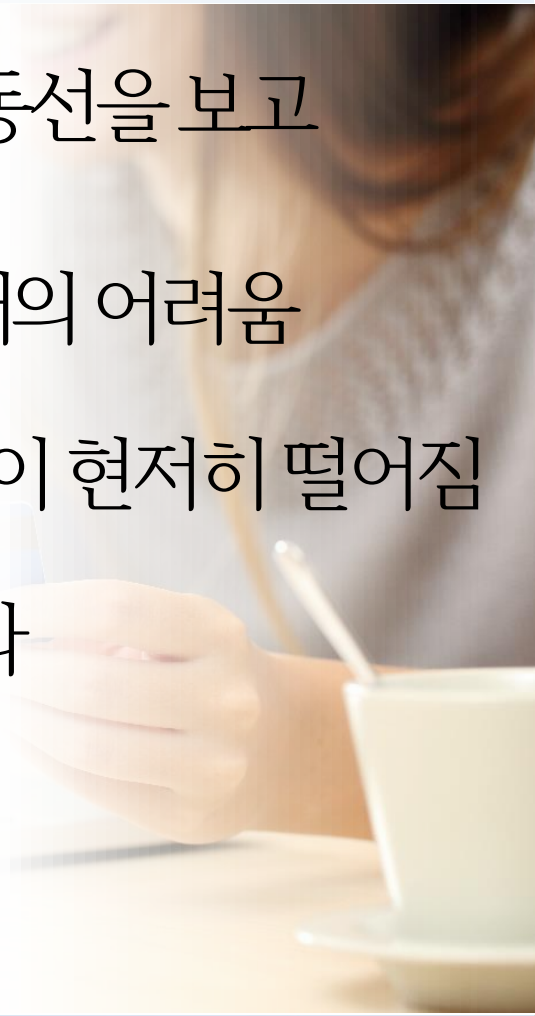
1. 의료정보활용



- 국내 의료데이터의 양과 질은 세계 최고 수준
- 환자의 프라이버시 보호 문제로 데이터 공유 어려움
- 의료법: 타 기관간 진료데이터 공유, 엄격히 제한
- 의료데이터는 모을수록 가치 배가
- 암호화한 상태서 전국민 의료데이터 공유 및 활용
- Family lineage와 연계해 유전병 연구. 정보보호는 필수!

2. 국민 건강과 사생활 모두 보호

- 현재의 Covid 19 정보 전달체계는 공개된 확진자의 동선을 보고 개개인이 확진자와의 접촉여부 판단. 문제는 확진자 동선 공개로 인한 개인정보 침해 및 정밀한 동선 공개의 어려움
- 문자 서비스로 확진자에 대한 정보 제공되나 유용성이 현저히 떨어짐
- 동형암호 이용하여 개인정보 보호하면서도 확진자와 개개 국민의 접촉 여부 확인 가능
- 2021. 1. 21. 통계청-경기도-서울대간 MOU 체결



3. 근거 기반 정책 집행

- 전국민 소득 파악하여 긴급 재난 지원금 지급 대상자 선정에 활용. 피해자 집중 지원
- 누락과 중복을 줄여서 fairness 증대에 기여
- 재난 지원금 신청 절차 없이 찾아가는 서비스
- 사업자 매출, 소득 파악하여 맞춤형 소상공인 지원에 활용

① 우리나라는 K-통계 시스템 구축에 최적의 여건

② 양질의 데이터 보유

③ 법제도적 발판 : 데이터 3법 통과, 혁신금융서비스, 규제샌드박스

④ 기술 인프라 : **ICBM** (IoT, Cloud, Bigdata, Mobile) 강국

⑤ 디지털 뉴딜이 포함된 한국판 뉴딜의 범 정부적 추진

⑥ 세계적 동형암호 기술 보유

⑦ 부처별로 산재한 각종 공공자료를 모을수록 데이터 가치 배가

⑧ 자료 축적/자료 결합과 자료 보안간 강한 보완 관계 존재

⑨ 암호 이용한 K-통계 체계 구축은 우리 경제 혁신의 큰 계기