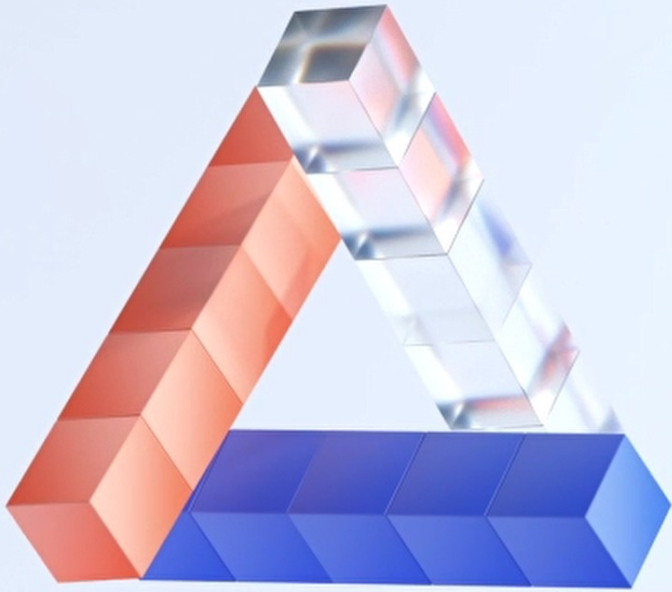




가명/익명처리 기술과 적정성 검토

SK텔레콤 정보보호팀 이창엽





목차

1. 가명/익명처리의 개념
2. 가명/익명처리 기술
3. 가명/익명 처리 적정성 검토



본 발표는 ‘기술’에 초점을 맞추었으며, 본 발표에서의 설명한 가명/익명처리 기술을 적용하였다고 하여 가명/익명정보라고 할 수는 없음.

가명처리 기술 외에도 목적 적합성, 처리 환경 등 실제로 가명/익명 처리 후 가명/익명 정보를 활용하기 위해서는 각 회사의 법무/정보보호 검토 및 데이터위원회(DRB)의 검토가 필요함.



가명/익명 정보의 개념



- 2020년 데이터 3법
- 개인정보를 동의 없이 당초 수집목적외 가명정보 형태로 활용가능
- 기존에는 신규 서비스 개발 목적으로 고객에게 선택동의를 받아야 활용 가능

제28조의2(가명정보의 처리 등) ① 개인정보처리자는 통계작성, 과학적 연구, 공익적 기록보존 등을 위하여 정보주체의 동의 없이 가명정보를 처리할 수 있다.

※ (주의) 「가명정보 처리에 관한 특례」에 따라 정보주체의 동의 없이 처리가 가능한 가명정보는 통계 작성, 과학적 연구, 공익적 기록보존에 한정되므로 처리 목적이 설정되지 않은 상황에서 보유하고 있는 개인정보를 가명처리하여 보관하는 것은 「가명정보 처리에 관한 특례」에 근거한 처리로 볼 수 없음

가명/익명 정보의 개념 - 개인정보 보호법



- 개인정보 보호법 제2조제1호

개인식별정보
(식별자)

개인식별
가능정보
/준식별자
/간접식별자
/속성자

가명정보

- 가. 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보
- 나. 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보. 이 경우 쉽게 결합할 수 있는지 여부는 다른 정보의 입수 가능성 등 개인을 알아보는 데 사용되는 시간, 비용, 기술 등을 합리적으로 고려하여야 한다.
- 다. 가목 또는 나목을 제1호의2에 따라 가명처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보(이하 "가명정보"라 한다)

- 개인정보보호법 제2조제1호의2

가명처리

- "가명처리"란 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보가 없이는 특정 개인을 알아볼 수 없도록 처리하는 것을 말한다.

- 익명정보

- 개인정보 보호법 제58조의2

- 이 법은 시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는 정보에는 적용하지 아니한다.



- 가명처리(Pseudonymization)
 - “식별”이 되지 않게
 - 특정한 개인을 알아볼 수 있게 되는 것
 - 정확한 정보주체를 알아보는 것에 한정, 구별과는 다름
 - 데이터의 이용 환경/목적 등도 고려
 - 처리를 덜 하여 유용성을 살리는 대신 다른 변수를 컨트롤
- 익명처리(Anonymization)
 - “Single out(선정)”이 되지 않게
 - 데이터 주체를 고유 식별하기 위해 데이터 셋의 특성 집합(동질집합)을 관찰하여 개인에 속한 레코드를 격리(isolation)해 낼 가능성
 - 특정 데이터에서 다른 데이터와 구별되는 유일한 행을 찾아내는 경우 선정되었다고 함.
 - Linkability 연결가능성 X
 - Inference 추론가능성 X
- 참고) 비식별 처리(De-identification) - 일반적으로는 가명처리와 익명처리를 모두 포함하는 개념
 - 익명처리에 가깝게 표현되는 경우가 많음. (익명이 상위개념인 경우도 있음)



가명처리

- 1) 구별 0
- 2) 연결* 0



개인정보 0

*신용정보법에서는 연계/연동으로 표기)

- 1) 식별가능성 X
- 2) 복원가능성 X

익명처리



개인정보 X

- 1) Single out X
- 2) 연결가능성 X
- 3) 추론가능성 X



개인정보(원본정보), 가명정보, 익명정보 예시

개인(원본)정보

이름	생년월일	핸드폰	직장전화	자택전화	주소	직업	가족	예금 평균잔액	대출액
홍길남	67년1월3일	010-9999-3333	055-747-0088	055-2345-6789	경남 창원시 마산구 합포동 123	국회의원	배우자 아들1 딸1	4,567,900원	45,896,789원

(잘못된) 가명정보 예시

※ 이름, 생년월일 등을 삭제하고 핸드폰 번호를 암호화하였으나, '창원시 마산구에
주소를 둔 국회의원'이라는 극소수 사람에 관한 정보로 개인 식별 가능

이름	생년월일	핸드폰	직장전화	자택전화	주소	직업	가족	예금 평균잔액	대출액
- • 삭제	- • 삭제	ajpejkc93 • 암호화	- • 삭제	- • 삭제	경남 창원시 마산구 • 범주화	국회의원	배우자 아들1 딸1	4,567,900원	45,896,789원

(제대로 된) 가명정보 예시

※ 사실상 해당 정보만으로는 개인식별 불가능. 다만, 핸드폰 암호 시 이용한 키
(추가정보)를 알면 복원가능하므로 가명정보에 해당

이름	생년월일	핸드폰	직장전화	자택전화	주소	직업	가족	예금 평균잔액	대출액
- • 삭제	- • 삭제	ajpejkc93 • 암호화	- • 삭제	- • 삭제	경남 창원시 마산구 • 범주화	- • 삭제	배우자 아들1 딸1	4,567,900원	45,896,789원

익명정보 예시

※ 그 자체로도 다른 정보와 결합해서도 개인을 식별할 수 없음

이름	생년월일	핸드폰	직장전화	자택전화	주소	직업	가족	예금 평균잔액	대출액
- • 삭제	- • 삭제	- • 삭제	- • 삭제	- • 삭제	경남 창원시 마산구 • 범주화	- • 삭제	가족 3명	400만원 ~500만원	4,000만원 ~5,000만원



가명/익명처리 기술

1. 가명/익명처리 기술/방법 소개

1. 국제 표준과 해외의 기술들
2. 가명정보 처리 가이드라인의 부록에 있는 기술들
3. 기타



- 국제표준과 유럽 기관의 리포트에서 수많은 기술들이 소개되어 있음

- ISO/IEC 20889:2018 기술

- Statistical Tools(통계도구)
 - Sampling(샘플링)
 - Aggregation(총계처리)
- Pseudonymization Techniques(가명화 기술)
- Anatomization(해부화)
- Cryptographic Tools(암호화 도구)
 - Deterministic Encryption(결정성 암호화)
 - Order-Preserving encryption(순서보존 암호화)
 - Format-Preserving encryption(형태보존 암호화)
 - Homomorphic encryption(동형 암호화)
 - Homomorphic secret sharing(동형비밀분산)
- Generalization Techniques(일반화 기술)
 - Rounding/Controlled Rounding(라운드/제어라운드)
 - Top and Bottom Coding(상하단 코딩)
 - Combining a set of attributes into a single attribute(특성집합을 단일특성으로 결합)
 - Local Generalization(로컬 일반화)
- Randomization Techniques(무작위화 기술)
 - Noise Addition(잡음추가)
 - Permutation(순열)
 - Micro-aggregation(부분총계)
- Suppression Techniques(삭제 기술)
 - Masking(마스킹)
 - Local Suppression(로컬 삭제)
 - Record Suppression(레코드 삭제)
- Synthetic Data(재현데이터)

- ENISA Report에 나온 기술

- Hashing without key
- Hashing with key or salt
- Encryption as a pseudonymisation technique
- Other Cryptography-based techniques
 - Homomorphic encryption
 - Polymorphic encryption
- Tokenisation
- Other approaches
 - Masking
 - Scrambling
 - Blurring, Barcodes, QR Codes
- Counter
- Random Number Generator (RNG)
- Cryptographic hash function
- Message Authentication Code(MAC)
- Encryption

.....

기타 등등...



•삭제기술

- 삭제
- 부분삭제
- 행 항목 삭제
- 로컬 삭제

•삭제/대체

•특 ※ 아래 분류는 이해를 돕기 위해 ISO/IEC 20889, 그리고 EU ENISA에서 발간한 보고서 등
국내·외 자료들을 참고하여 작성했으며 표준이 아님

•C클라우드(가명처리) 기술

- 일반 라운딩
- 랜덤 라운딩
- 제어 라운딩
- 상하단 코딩
- 로컬 일반화
- 범위 방법
- 문자데이터 범주화

•암호화

- 양방향 암호화
- 일방향 암호화 - 암호학적 해시함수
- 순서보존 암호화
- 행태보존 암호화
- 동형 암호화

- 조인화
- (의사)난수생성기
- 가명/익명처리를 위한 다양한 기술(기타기술)
 - 표본 추출
 - 해부화
 - 재현데이터
 - 동형비밀분산
 - 차분 프라이버시



•삭제기술

- 삭제
- 부분삭제
- 행 항목 삭제
- 로컬 삭제

•삭제/대체

- 마스킹

•통계도구

- 총계처리
- 부분총계

•일반화(범주화) 기술

- 일반 라운딩
- 랜덤 라운딩
- 제어 라운딩
- 상하단 코딩
- 로컬 일반화
- 범위 방법
- 문자데이터 범주화

•암호화

- 양방향 암호화
- 일방향 암호화 - 암호학적 해시함수
- 순서보존 암호화
- 행태보존 암호화
- 동형 암호화
- 다형성 암호화

•무작위화 기술

- 잡음 추가
- 순열(치환)
- 토큰화
- (의사)난수생성기

•가명/익명처리를 위한 다양한 기술(기타기술)

- 표본 추출
- 해부화
- 재현데이터
- 동형비밀분산
- 차분 프라이버시

가명/익명 처리 기술 소개 - 개보위 가명정보 처리 가이드라인



• 삭제기술

- 삭제/부분삭제 : 원본정보에서 개인정보 컬럼을 삭제 or 부분삭제
 - 목적 - 식별자나 혹은 목적에 필요하지 않는 항목은 삭제/부분삭제

성명	성별	나이	핸드폰번호	주소	통신료	단말기금액	누적포인트
김철수	남	41세	010-6666-8888	서울특별시 중구 무교동	98,700	1,198,700	356,800
이영희	여	61세	010-9999-2222	부산광역시 북구 화명동	69,400	505,400	203,000
박민호	남	30세	010-2222-7777	광주광역시 서구 금호동	104,400	1,604,400	198,000
이윤정	여	57세	010-3333-4444	전라남도 나주시 빛가람동	954,800	3,954,800	20,532,000
최동욱	남	28세	010-5555-6666	세종특별자치시 어진동	83,600	883,600	400,900

삭제

성별	나이	통신료	단말기금액	누적포인트
남	41세	98,700	1,198,700	356,800
여	61세	69,400	505,400	203,000
남	30세	104,400	1,604,400	198,000
여	57세	954,800	3,954,800	20,532,000
남	28세	83,600	883,600	400,900

성명	성별	나이	핸드폰번호	주소	통신료	단말기금액	누적포인트
김철수	남	41세	010-6666-8888	서울특별시 중구 무교동	98,700	1,198,700	356,800
이영희	여	61세	010-9999-2222	부산광역시 북구 화명동	69,400	505,400	203,000
박민호	남	30세	010-2222-7777	광주광역시 서구 금호동	104,400	1,604,400	198,000
이윤정	여	57세	010-3333-4444	전라남도 나주시 빛가람동	954,800	3,954,800	20,532,000
최동욱	남	28세	010-5555-6666	세종특별자치시 어진동	83,600	883,600	400,900

삭제

성명	성별	나이	핸드폰번호	주소	통신료	단말기금액	누적포인트
김	남	41세	8888	서울특별시 중구	98,700	1,198,700	356,800
이	여	61세	2222	부산광역시 북구	69,400	505,400	203,000
박	남	30세	7777	광주광역시 서구	104,400	1,604,400	198,000
이	여	57세	4444	전라남도 나주시	954,800	3,954,800	20,532,000
최	남	28세	6666	세종특별자치시	83,600	883,600	400,900

※ (주의) 「가명정보 처리에 관한 특례」에 따라 정보주체의 동의 없이 처리가 가능한 가명정보는 통계작성, 과학적 연구, 공익적 기록보존에 한정되므로 처리 목적이 설정되지 않은 상황에서 보유하고 있는 개인정보를 가명처리하여 보관하는 것은 「가명정보 처리에 관한 특례」에 근거한 처리로 볼 수 없음

가명/익명 처리 기술 소개 - 개보위 가명정보 처리 가이드라인



• 삭제기술

- 행 항목 삭제 : 다른 정보와 뚜렷하게 구별되는 행 항목을 삭제 -> 특이치 제거
- 로컬 삭제 : 특이정보를 해당 행 항목에서 삭제 -> 특이치 제거

성명	성별	나이	핸드폰번호	주소	통신료	단말기금액	누적포인트
김철수	남	41세	010-6666-8888	서울특별시 중구 무교동	98,700	1,198,700	356,800
이영희	여	61세	010-9999-2222	부산광역시 북구 화명동	69,400	505,400	203,000
박민호	남	30세	010-2222-7777	광주광역시 서구 금호동	104,400	1,604,400	198,000
이윤정	여	57세	010-3333-4444	전라남도 나주시 빛가람동	954,800	3,954,800	20,532,000
최동욱	남	28세	010-5555-6666	세종특별자치시 어진동	83,600	883,600	400,900

삭제

성명	성별	나이	핸드폰번호	주소	통신료	단말기금액	누적포인트
김철수	남	41세	010-6666-8888	서울특별시 중구 무교동	98,700	1,198,700	356,800
이영희	여	61세	010-9999-2222	부산광역시 북구 화명동	69,400	505,400	203,000
박민호	남	30세	010-2222-7777	광주광역시 서구 금호동	104,400	1,604,400	198,000
최동욱	남	28세	010-5555-6666	세종특별자치시 어진동	83,600	883,600	400,900

성명	성별	나이	핸드폰번호	주소	통신료	단말기금액	누적포인트
김철수	남	41세	010-6666-8888	서울특별시 중구 무교동	98,700	1,198,700	356,800
이영희	여	61세	010-9999-2222	부산광역시 북구 화명동	69,400	505,400	203,000
박민호	남	30세	010-2222-7777	광주광역시 서구 금호동	104,400	1,604,400	198,000
이윤정	여	57세	010-3333-4444	전라남도 나주시 빛가람동	954,800	3,954,800	20,532,000
최동욱	남	28세	010-5555-6666	세종특별자치시 어진동	83,600	883,600	400,900

삭제

성명	성별	나이	핸드폰번호	주소	통신료	단말기금액	누적포인트
김철수	남	41세	010-6666-8888	서울특별시 중구 무교동	98,700	1,198,700	356,800
이영희	여	61세	010-9999-2222	부산광역시 북구 화명동	69,400	505,400	203,000
박민호	남	30세	010-2222-7777	광주광역시 서구 금호동	104,400	1,604,400	198,000
이윤정	여	57세	010-3333-4444	전라남도 나주시 빛가람동	954,800	3,954,800	
최동욱	남	28세	010-5555-6666	세종특별자치시 어진동	83,600	883,600	400,900



- 삭제/대체
 - 마스킹 : 특정 항목의 일부 또는 전부를 공백 또는 문자(*, _ 등)로 대체

성명	성별	나이	핸드폰번호		성명	성별	나이	핸드폰번호
김철수	남	41세	010-6666-8888		김**	남	4*세	***-****-****
이영희	여	61세	010-9999-2222		이**	여	6*세	***-****-****
박민호	남	30세	010-2222-7777		박**	남	3*세	***-****-****
이윤정	여	57세	010-3333-4444		이**	여	5*세	***-****-****
최동욱	남	28세	010-5555-6666		최**	남	2*세	***-****-****



- 통계도구

- 총계처리 : 평균값, 최댓값, 최솟값, 최빈값, 중간값 등 통계값으로 대체
 - 데이터 전체가 유사한 특징을 가진 개인으로 구성되어 있을 경우 그 데이터의 대푯값이 특정 개인의 정보를 그대로 노출시킬 수도 있으므로 주의 필요

통신료 98,700 69,400 104,400 954,800 83,600	→ 평균값 →	통신료 262,180 262,180 262,180 262,180 262,180	→ 최댓값 →	통신료 98,700 69,400 104,400 954,800 83,600	→ 최솟값 →	통신료 98,700 69,400 104,400 954,800 83,600	→ 최솟값 →	통신료 69,400 69,400 69,400 69,400 69,400
통신료 98,700 69,400 104,400 954,800 104,400	→ 최빈값 →	통신료 104,400 104,400 104,400 104,400 104,400	→ 정렬 →	통신료 98,700 69,400 104,400 54,800 83,600	→ 중간값 →	통신료 54,800 69,400 83,600 98,700 104,400	→ 중간값 →	통신료 83,600 83,600 83,600 83,600 83,600



- 통계도구
 - 부분총계 : 정보집합물 내 하나 또는 그 이상의 행 항목에 해당하는 특정 열 항목을 총계처리. 즉, 다른 정보에 비하여 오차 범위가 큰 항목을 평균값 등으로 대체
 - 동질 집합* 내의 특정 항목을 총계처리 하거나 특정 조건에 너무 특이한 값이 있어 개인의 식별 가능성이 높지만 분석에 꼭 필요한 값인 경우 처리

지역	나이	소득금액
서울	30대	5,987,900
서울	30대	28,169,700
서울	30대	3,009,600
나주	30대	4,607,300
나주	30대	3,560,800
나주	30대	2,940,100
세종	30대	6,088,400
세종	30대	2,789,200
세종	30대	5,048,300



지역	나이	소득금액
서울	30대	12,389,067
서울	30대	12,389,067
서울	30대	12,389,067
나주	30대	4,607,300
나주	30대	3,560,800
나주	30대	2,940,100
세종	30대	6,088,400
세종	30대	2,789,200
세종	30대	5,048,300

가명/익명 처리 기술 소개 - 개보위 가명정보 처리 가이드라인



- 일반화(범주화)기술
 - 라운딩 : 올림, 내림, 반올림 등의 기준을 적용하여 집계처리 하는 방법
 - 랜덤라운딩 : 수치 데이터를 임의의 수인 자리수, 실제 수 기준으로 올림, 내림 하는 기법

나이	올림	내림	반올림
33세	40세	30세	30세
61세	70세	60세	60세
47세	50세	40세	50세
66세	70세	60세	70세
40세	40세	40세	40세

※ 적절하지 않은 라운딩의 경우 라운딩 후에도 남은 값의 유일성이 남게 될 수 있으며, 적용하는 단위에 대한 판단이 중요

금액	백 단위 라운딩	금액	백만 단위 라운딩
983,116,785	983,117,000	983,116,785	980,000,000
984,715,591	984,716,000	984,715,591	980,000,000
984,932,383	984,932,000	984,932,383	980,000,000
985,660,262	985,660,000	985,660,262	990,000,000
986,047,778	986,048,000	986,047,778	990,000,000

적절하지 않은 라운딩

적절한 라운딩

금액		금액
869,250	만 단위 라운딩	900,000
4,559,120	십만 단위 라운딩	4,000,000
13,601,564	십만 단위 라운딩	14,000,000
979,118	만 단위 라운딩	900,000
122,848,878	백만 단위 라운딩	120,000,000



- 일반화(범주화)기술
 - 제어라운딩 : 라운딩 적용 시 값의 변경에 따라 행이나 열의 합이 원본의 행이나 열의 합과 일치하지 않는 단점을 해결하기 위해 원본과 결과가 동일하도록 라운딩을 적용하는 기법

(설명) 나이에 대한 평균 분석 시 원본의 경우 평균이 51세가 되나 일반 라운딩을 적용한 경우 평균이 50세가 되어 결과가 다르게 되고, 이에 일부 값을 다르게 라운딩(제어)하여 평균 나이가 원본과 일치되도록 함

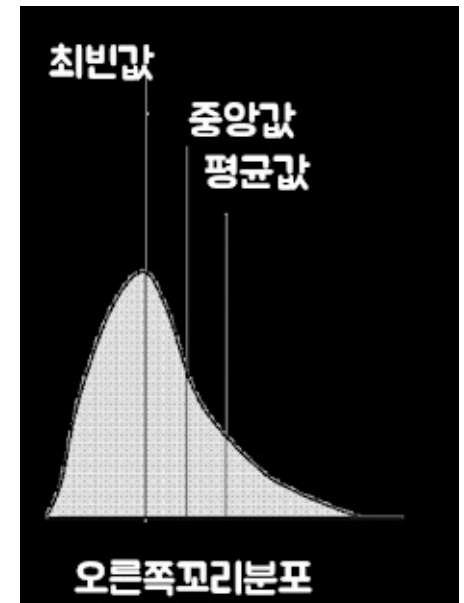
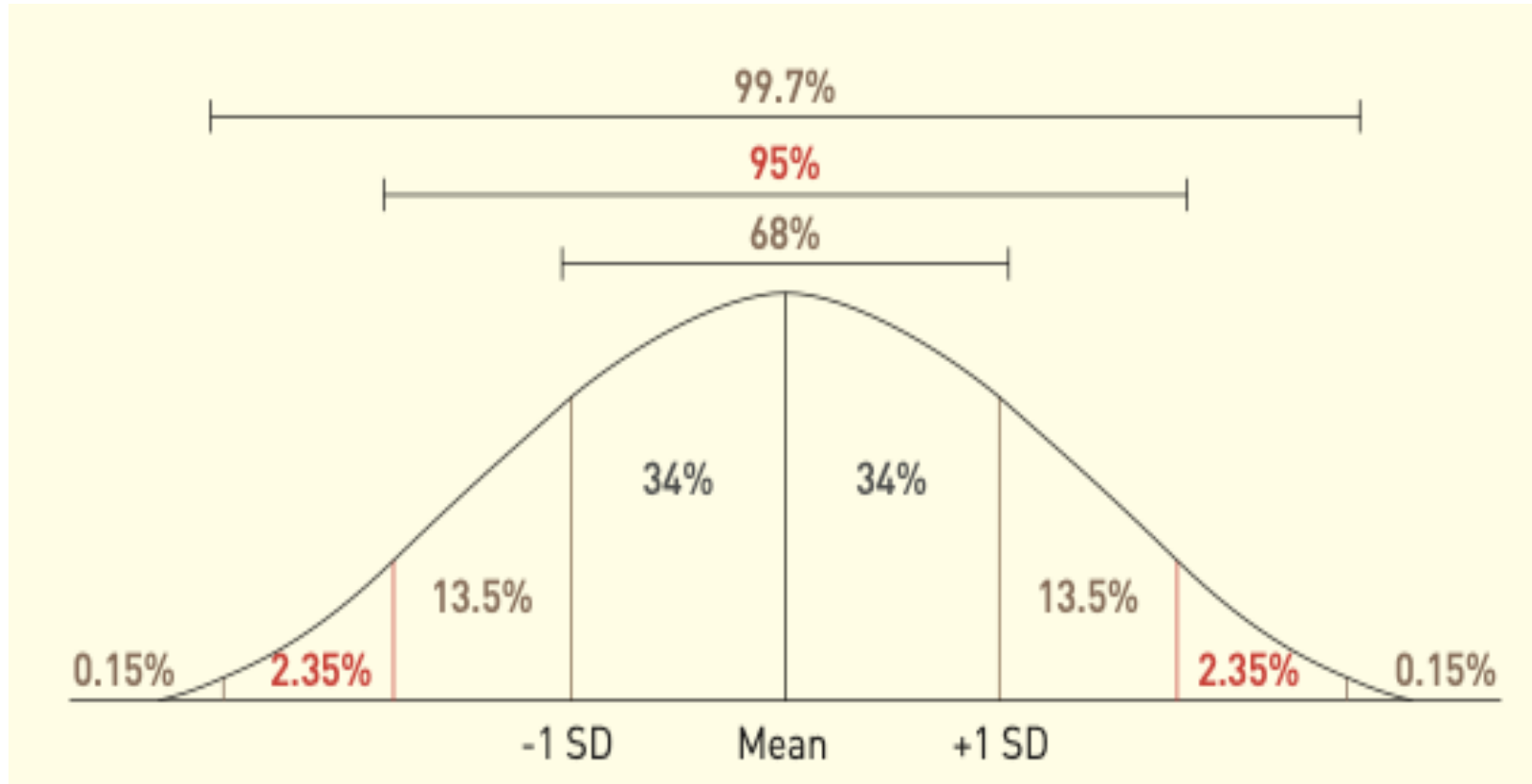
원본(나이)
33세
61세
50세
72세
43세
44세
23세
67세
68세
49세
평균: 51세
합계: 510



일반 라운딩	제어 라운딩
30세	30세
60세	60세
50세	50세
70세	70세
40세	40세
40세	50세
20세	20세
70세	70세
70세	70세
50세	50세
평균: 50세	평균: 51세
합계: 500	합계: 510



- 일반화(범주화)기술
 - 상하단코딩 : 정규분포의 특성을 가진 데이터에서 양쪽 끝에 치우친 정보는 적은 수의 분포를 가지게 되어 식별성을 가질 수 있으며, 이를 해결하기 위해 적은 수의 분포를 가진 양 끝단의 정보를 범주화,삭제 등의 기법을 적용하여 식별성을 낮추는 기법 -> 특이치 제거





- 일반화(범주화)기술
 - 로컬 일반화 : 전체 정보집합물 중 특정 열 항목(들)에서 특이한 값을 가지거나 분포상의 특이성으로 인해 식별성이 높아지는 경우 해당 부분만 일반화를 적용하여 식별성을 낮추는 기법
 - 범위 방법 : 수치 데이터를 임의의 수 기준의 범위로 설정하는 기법으로, 해당 값의 범위 또는 구간으로 표현

(설명) 서울 지역의 30대 중 분포 상 다른 금액에 비해 특이한 값을 동질집합 내 범주화
 ※ 특이한 로컬(28,169,700)에만 3,009,600 ~ 28,169,700으로 범주화 할 수 있음

지역	나이	소득금액
서울	30대	5,987,900
서울	30대	28,169,700
서울	30대	3,009,600
나주	30대	4,607,300
나주	30대	3,560,800
나주	30대	2,940,100
세종	30대	6,088,400
세종	30대	2,789,200
세종	30대	5,048,300



지역	나이	소득금액
서울	30대	3,009,600~28,169,700
서울	30대	3,009,600~28,169,700
서울	30대	3,009,600~28,169,700
나주	30대	4,607,300
나주	30대	3,560,800
나주	30대	2,940,100
세종	30대	6,088,400
세종	30대	2,789,200
세종	30대	5,048,300

(예시) 소득 3,300만원을 소득 3,000만원 ~ 4,000만원으로 대체 표기



- 일반화(범주화)기술
 - 문자데이터 범주화 : 문자로 저장된 정보에 대해 상위의 개념으로 범주화 하는 기법

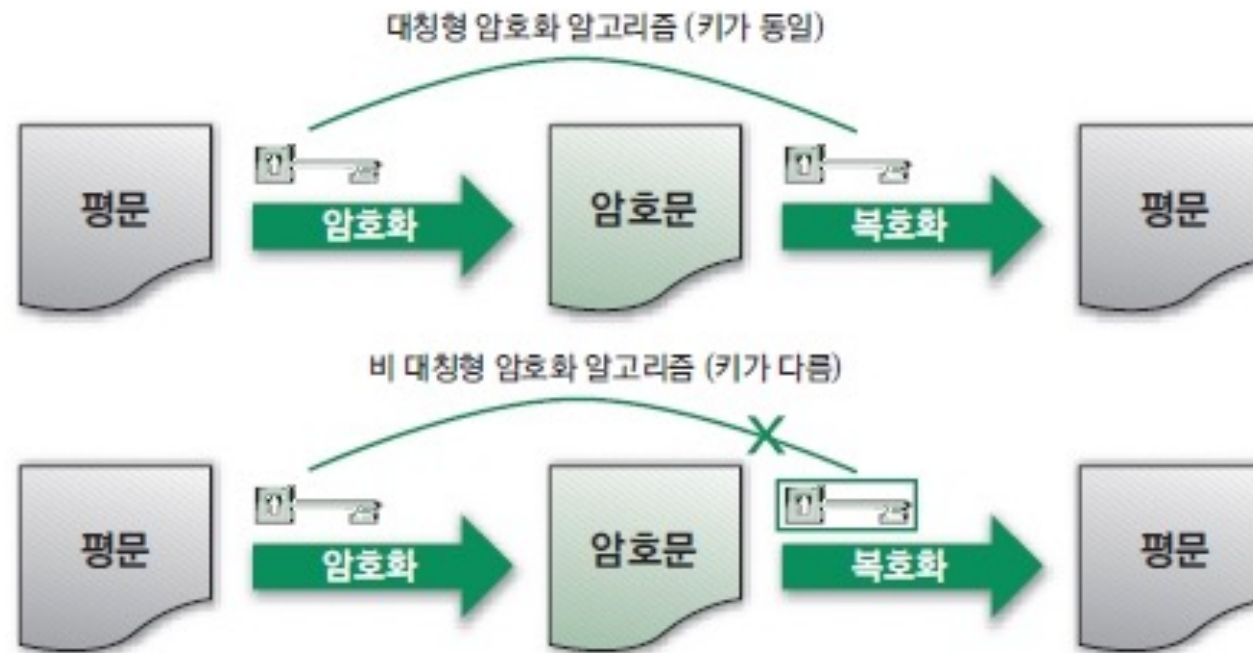
품목	품목
분유	육아용품
기저귀	육아용품
젖병	육아용품
샤워타올	육아용품
욕실화	육아용품

<주소 시군구 이하 삭제>
서울특별시 중구 을지로 65

<우편번호 앞2자리>
서울특별시 (01 ~ 09)
경기도 (10 ~ 20)
인천광역시 (21 ~ 23)
강원도 (24 ~ 26)
충청북도 (27 ~ 29)
세종특별자치시 (30)
충청남도 (31 ~ 33)
대전광역시 (34 ~ 35)
경상북도 (36 ~ 40)
대구광역시 (41 ~ 43)
울산광역시 (44 ~ 45)
부산광역시 (46 ~ 49)
경상남도 (50 ~ 53)
전라북도 (54 ~ 56)
전라남도 (57 ~ 60)
광주광역시 (61 ~ 62)
제주특별자치도 (63)

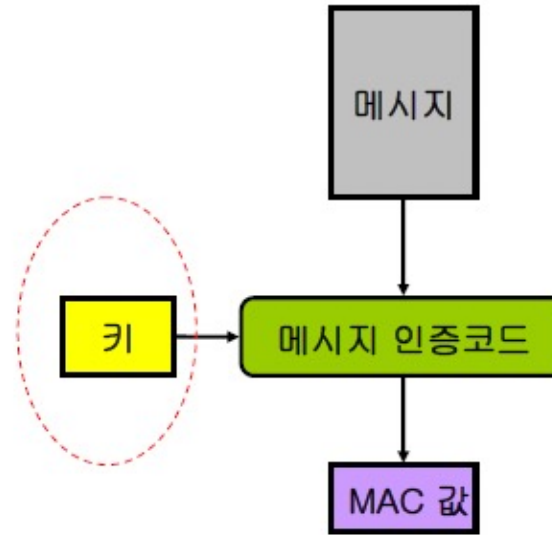
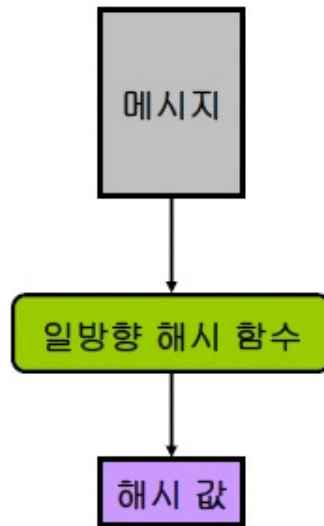


- 암호화
 - 양방향 암호화 : 특정 정보에 대해 암호화와 암호화된 정보에 대한 복호화가 가능한 암호화 기법
 - 대칭키 암호화
 - 공개키 암호화
 - 키 삭제를 통한 결정적 암호화





- 암호화
 - 일방향 암호화 - 암호학적 해시함수
 - 원문에 대한 암호화의 적용만 가능하고 암호문에 대한 복호화 적용이 불가능한 암호화 기법
 - Hashing without key
 - Hashing with key or salt





- 암호화
 - 순서보존 암호화 : 원본정보의 순서와 암호값의 순서가 동일하게 유지되는 암호화 방식
 - 암호화된 상태에서도 원본정보의 순서가 유지되어 값들 간의 크기에 비교 분석이 필요한 경우 안전한 분석이 가능

값	Order-preserving encryption 적용 값
1	0x0001102789d5f50b2beffd9f3dca4ea7
2	0x000422a6142ece2a405cd7530afed84
3	0x009fb3c8fc4e75330df0155dfcdb48sd
4	0x010ce5aaa81ae738d2c530f15e55az7
5	0x04jo294lkksjlrops2365t9lk2jkml4m22
6	0x04zwoekrjfl3422ljfsadnml8s903ljkasjj
7	0x05e75330df0155dfcdb230afed84c5a
8	0x06002b6e8187c6733f40e573b76f99a
9	0x07234e75330df0155dfcdb260d39eac

만약 $x < y$ 이면, $Enc(x) < Enc(y)$

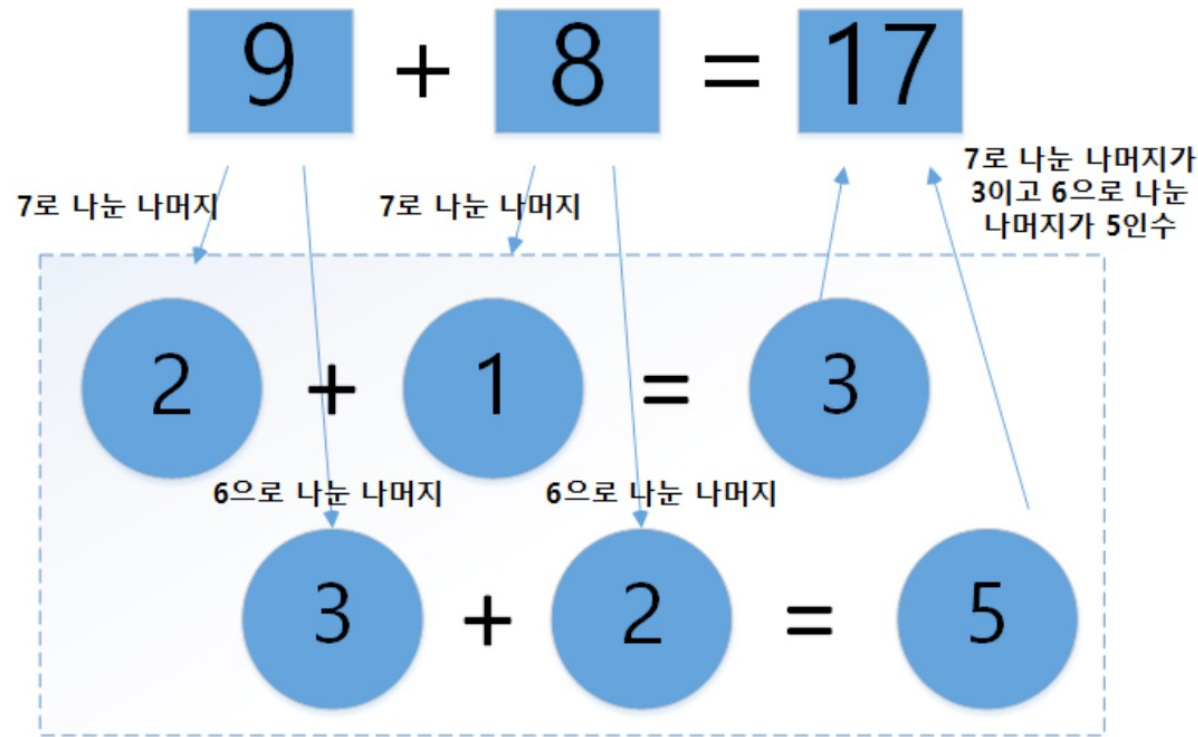


- 암호화
 - 형태보존 암호화 : 원본정보의 형태와 암호화된 암호값의 형태가 동일하게 유지되는 암호화 방식
 - 원본정보와 동일한 크기와 구성 형태를 가지기. 때문에 일반적인 암호화가 가지고 있는 저장공간의 스키마 변경 이슈가 없어 저장 공간의 비용 증가를 해결할 수 있음
 - 암호화로 인해 발생하는 시스템의 수정이 거의 발생하지 않아 토큰화, 신용카드 번호의 암호화 등에서 기존 시스템의 변경 없이 암호화를 적용할 때 사용

	<사례> 블록 암호화 적용 ==> sFSIEJlsdkiwjeovv.serijf== FPE 적용(동일 카드번호 형식) ==> 1738 4237 2345 9814
---	---



- 암호화
 - 동형 암호화 : 암호화된 상태에서의 연산이 가능한 암호화 방식
 - 원래의 값을 암호화한 상태로 연산처리를 하여 다양한 분석에 이용 가능
 - 암호화된 상태의 연산한 값을 복호화 하면 원래의 값을 연산한 것과 동일한 결과를 얻을 수 있는 4세대 암호화 기법

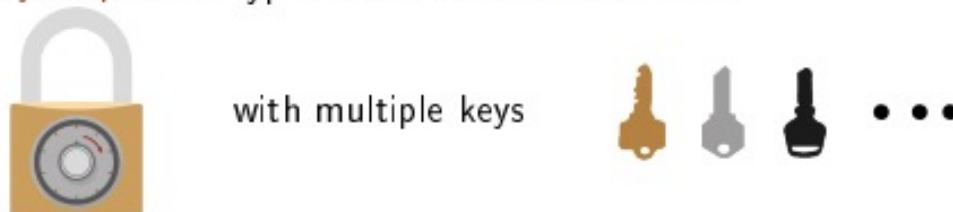




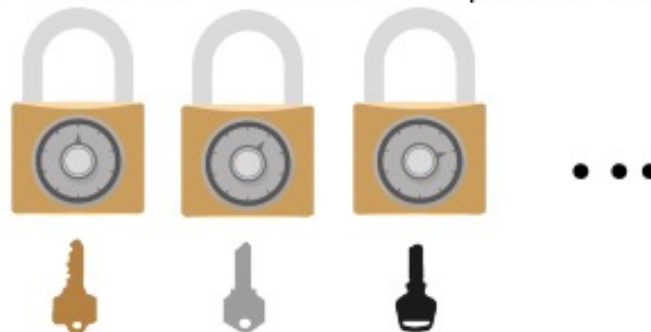
- 암호화
 - 다형성 암호화 : 가명정보의 부정합 결합을 차단하기 위해 각 도메인 별로 서로 다른 가명처리 방법을 사용하여 정보를 제공하는 방법
 - 정보 제공시 서로 다른 방식의 암호화된 가명처리를 적용함에 따라 도메인 별로 다른 가명정보를 가지게 됨

Polymorphic locks

- ▶ Traditionally, only the owner of the **private key**  can decrypt
- ▶ In **polymorphic** encryption we use **malleable** locks:



- ▶ By turning the wheel, the lock can be morphed to a specific key:





- 무작위화기술
 - 잡음추가 : 개인정보에 임의의 숫자 등 잡음을 추가(더하기, 곱하기) 하는 방법
 - 지정된 평균과 분산의 범위 내에서 잡음이 추가되므로 원 자료의 유용성을 해치지 않으나, 잡음값은 데이터 값과는 무관하기 때문에 유효한 데이터로 활용하기 곤란하여, 중요한 종적 정보는 동일한 잡음을 사용해야 함. (입원일자 예시)

생년월일	잡음추가	잡음추가생년월일
2011-12-05	+3	2011-12-08
2016-08-09	-2	2016-08-07
2009-02-11	-5	2009-02-06
1998-05-27	-6	1998-05-21
1991-06-18	+9	1991-06-27

가명/익명 처리 기술 소개 – 개보위 가명정보 처리 가이드라인



- 무작위화기술
 - 순열(치환) : 기존 값은 유지하면서 개인이 식별되지 않도록 데이터를 재배열하는 방법
 - 개인정보를 다른 행 항목의 정보와 무작위로 순서를 변경하여 전체 정보에 대한 변경 없이 특정 정보가 해당 개인과 연결되지 않도록 하는 방법

(설명) 원본과 비교하여 평균 분석 시 전체 재배열은 결과가 다르며
동질집합 내 재배열 결과는 동일

지역	나이	소득금액(원본)	소득금액(전체 재배열)	소득금액(동질집합 내 재배열)
서울	30대	5,987,900	2,789,200	3,009,600
서울	30대	8,169,700	4,607,300	5,987,900
서울	30대	3,009,600	5,987,900	8,169,700
나주	30대	4,607,300	2,940,100	2,940,100
나주	30대	3,560,800	8,169,700	4,607,300
나주	30대	2,940,100	5,048,300	3,560,800
세종	30대	6,088,400	3,009,600	2,789,200
세종	30대	2,789,200	3,560,800	5,048,300
세종	30대	5,048,300	6,088,400	6,088,400

원본 분석결과	지역	서울	나주	세종
	평균소득	5,722,400	3,702,733	4,641,967
전체 재배열 분석결과	지역	서울	나주	세종
	평균소득	4,461,467	5,048,300	4,219,600
동질집합 내 재배열 분석결과	지역	서울	나주	세종
	평균소득	5,722,400	3,702,733	4,641,967



- 무작위화기술
 - 토큰화 : 개인을 식별할 수 있는 정보를 토큰으로 변환 후 대체함으로써 개인정보를 직접 사용하여 발생하는 개인에 대한 식별 위험을 제거하여 개인정보를 보호하는 기술
 - 토큰 생성시 의사난수생성이나 일방향 암호화, 순서보존 암호화를 주로 사용





- 무작위화기술
 - 의사난수생성기 : 주어진 입력 값에 대해 예측이 불가능 하고 패턴이 없는 값을 생성하는 매커니즘으로 임의의 숫자를 개인정보에 할당
 - 난수는 원칙적으로 규칙적인 배열순서가 없는 임의의 수를 의미하며 컴퓨터는 원천적으로 입력에 의한 처리 결과를 반환하는 것으로 처리의 방법과 입력이 동일하면 항상 동일한 출력이 발생하기 때문에 완전한 난수의 생성은 불가능

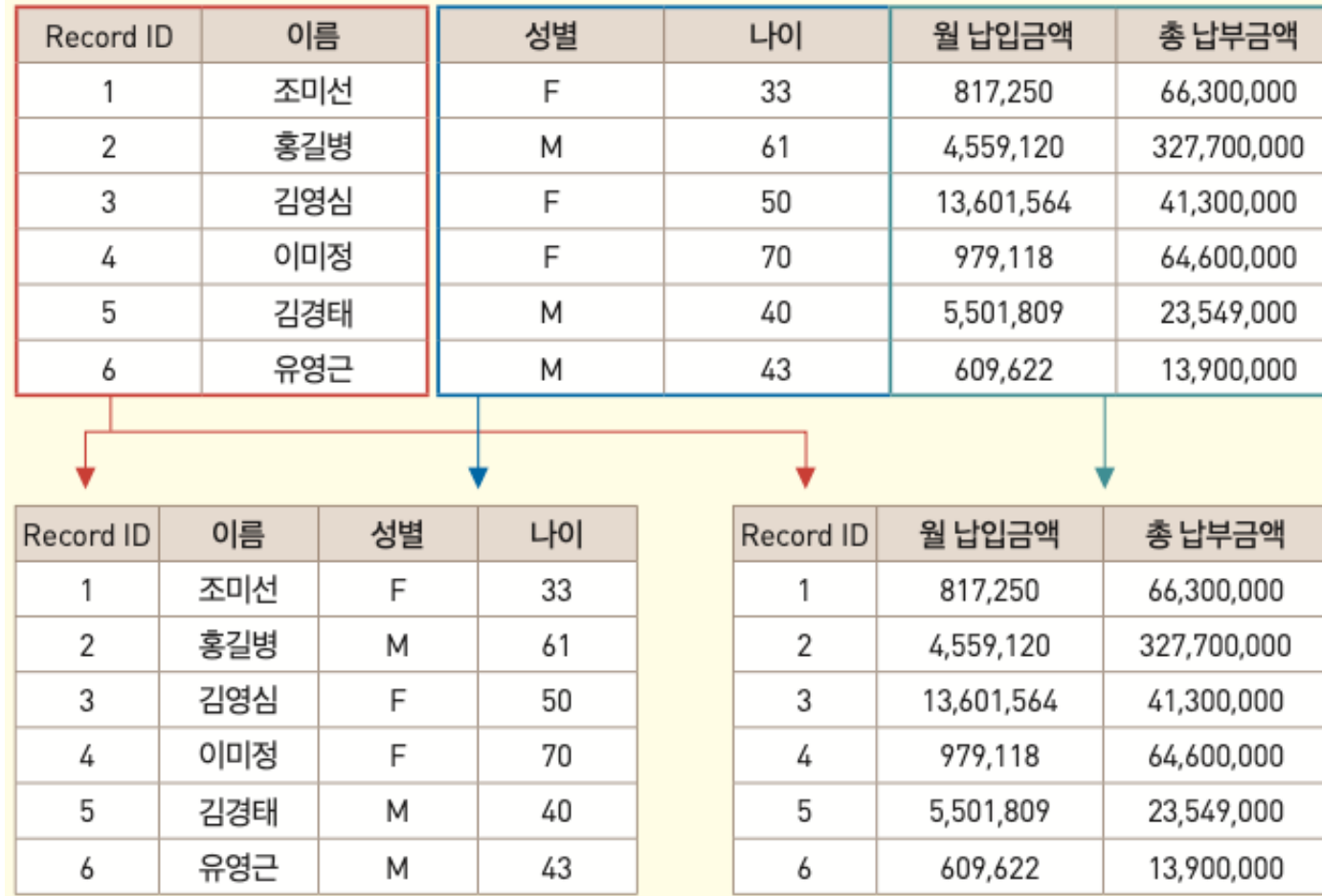


- 가명/익명처리를 위한 다양한 기술(기타기술)
 - 표본추출 : 일명 샘플링으로 데이터의 주제별로 전체 모집단이 아닌 표본에 무작위 레코드 추출 등의 기법을 통해 모집단의 일부를 분석하여 전체에 대한 분석을 대신하는 기법
 - 확률적 표본 추출
 - 무작위 표본추출(복원, 비복원) – 동일 확률($1/n$) 추출
 - 계통적 표본추출 – 매 X번째 구성요소를 추출
 - 층화 표본추출
 - 집락 표본추출

가명/익명 처리 기술 소개 - 개보위 가명정보 처리 가이드라인



- 가명/익명처리를 위한 다양한 기술(기타기술)
 - 해부화 : 기존 하나의 데이터셋(테이블)을 식별성이 있는 정보집합물과 식별성이 없는 정보집합물로 구성된 2개의 데이터셋으로 분리하는 기술





- 가명/익명처리를 위한 다양한 기술(기타기술)
 - 재현데이터 : 원본과 최대한 유사한 통계적 성질을 보이는 가상의 데이터를 생성하기 위해 개인정보의 특성을 분석하여 새로운 데이터를 생성하는 기법
 - 머신러닝, 통계학적 기법 등을 통해서 생성
 - 완전 재현 데이터
 - 부분 재현 데이터
 - 선택된 민감한 속성의 값만 재현값으로 대체



- 가명/익명처리를 위한 다양한 기술(기타기술)
 - 동형비밀분산 : 식별정보 또는 기타 식별가능정보를 메시지 공유 알고리즘에 의해 생성된 두개 이상의 share(기밀사항을 재구성하는데 사용할 수 있는 하위집합)로 대체
 - 재식별은 가명/익명처리된 데이터의 쉼어를 소유한 모두가 동의하는 경우만 가능
- 차분 프라이버시 -> 프라이버시 보호모델



HIPAA (The Health Insurance Portability and Accountability Act)

- Privacy Rules De-identification Methods
 - Expert Determination
 - 적절한 지식과 경험을 갖춘 전문가가 정보주체의 식별 위험성이 매우 낮다고 결정할 경우 비식별화 된 것으로 인정하는 방식
 - Safe Harbor : 규칙에서 명시한 18가지의 식별자를 제거

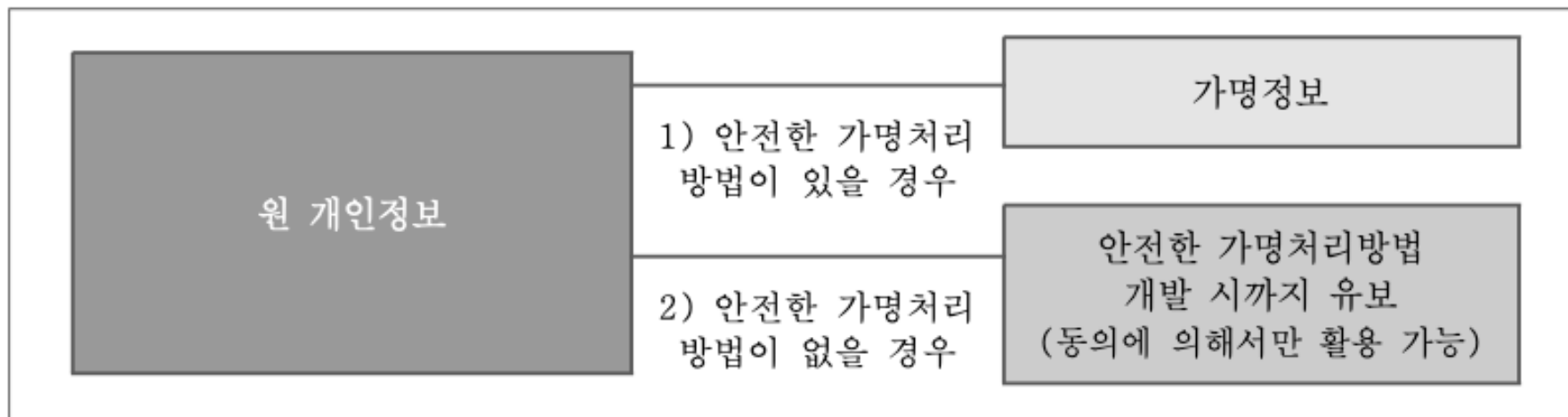
세이프하버 방식에서 정하는 18가지 식별자 및 준식별자 항목

- | | |
|---|---------------------------|
| (A) 이름 | (I) 건강보험 등록번호 |
| (B) 주(state)보다 작은 지리적 단위(시, 군, 구역),
우편번호 및 이와 상응하는 지역코드 | (J) 계좌번호 |
| (C) 개인과 직접적으로 연관되는 모든 날짜로 생일,
입원일, 퇴원일, 사망일, 89세 이상 모든 연령과
그 연령을 나타내는 모든 일자 | (K) 수료증 혹은 자격증 번호 |
| (D) 전화번호 | (L) 자동차 번호판을 포함한 차량번호 |
| (E) 팩스번호 | (M) 기기 아이디 및 시리얼 넘버 |
| (F) 이메일 주소 | (N) 웹페이지 주소(URLs) |
| (G) 사회보장번호 | (O) 인터넷 IP주소 |
| (H) 의무기록번호 / 환자번호 | (P) 신체특성 지표자, 지문 및 음성포함 |
| | (Q) 얼굴 정면 사진 및 이에 상응하는 사진 |
| | (R) 기타 모든 고유한 숫자, 문자, 코드 |



보건의료 데이터 활용 가이드라인

- 처리 상세 가이드
 - 속성값별 "일반적" 기준
 - 측정수치 정보 : 별도의 조치 불필요
 - 알고리즘이 생산한 건강정보 : 별도의 조치 불필요
 - 체내/외를 촬영한 영상정보
 - 눈, 코, 입, 문신, 기타 외양적 특징을 모두 삭제 혹은 마스킹
 - 메타데이터 상 식별자 삭제, 영상에 식별자 모두 삭제
 - 음성정보 : 가명처리 가능여부 유보 (본인 동의 기반으로만 사용가능)
 - 진료과정 음성녹음, 판독과정음성녹음 등





가명/익명처리 적정성 검토

1. (검토를 받는 입장에서) 적정성 검토
2. 프라이버시 보호 모델

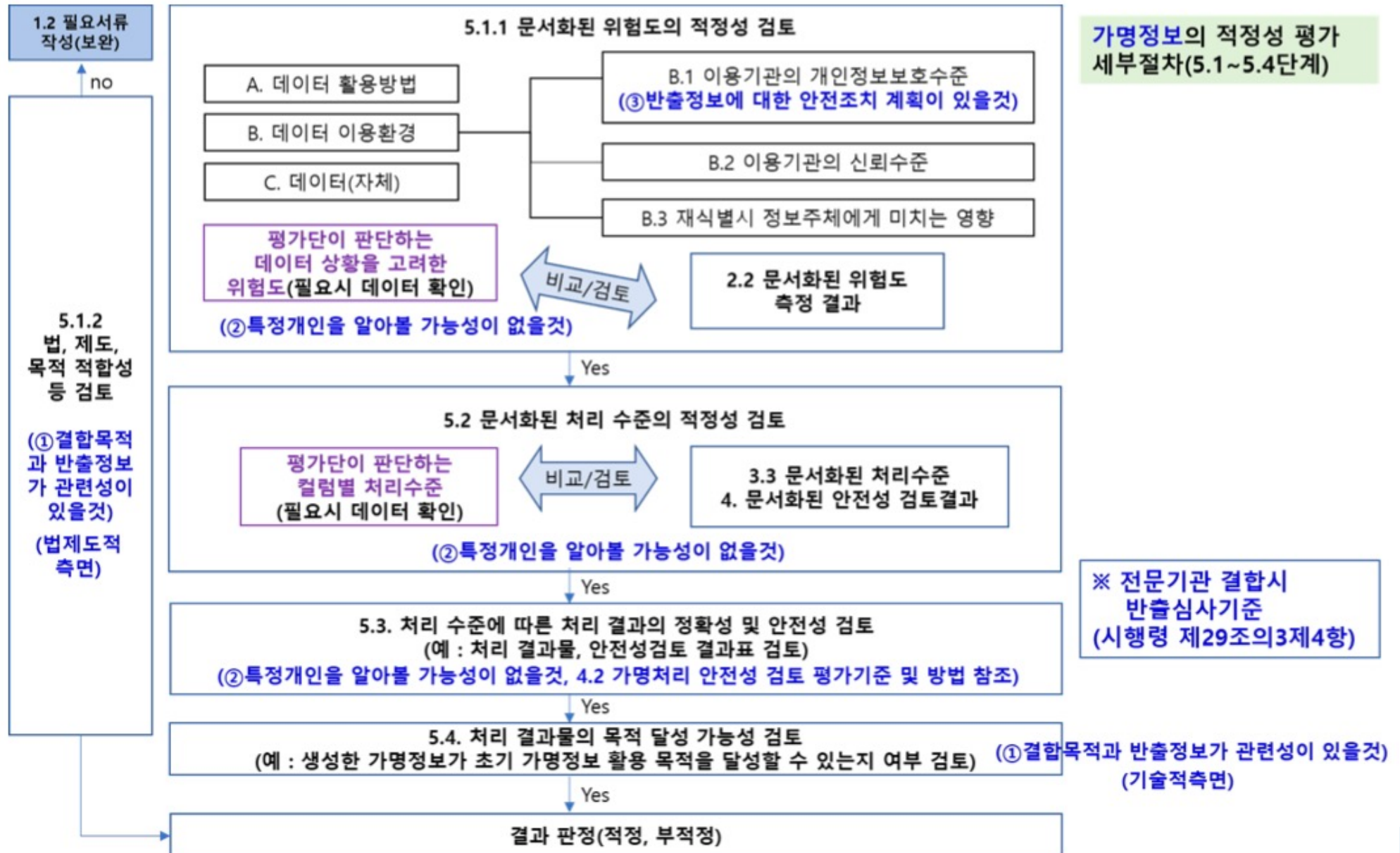


- 개보위 가명정보 처리 가이드라인 적정성 검토
 - 사전 준비 서류 및 위험성 검토 -> 문서화된 위험도
 - 가명처리 방법 및 수준 적정성 검토 -> 문서화된 처리 수준
 - 가명처리 적정성 검토 -> 문서화된 처리 결과의 정확성
 - 처리 목적 달성 가능성 검토 -> 유용성

가명/익명처리 적정성 검토 - 가명 예시



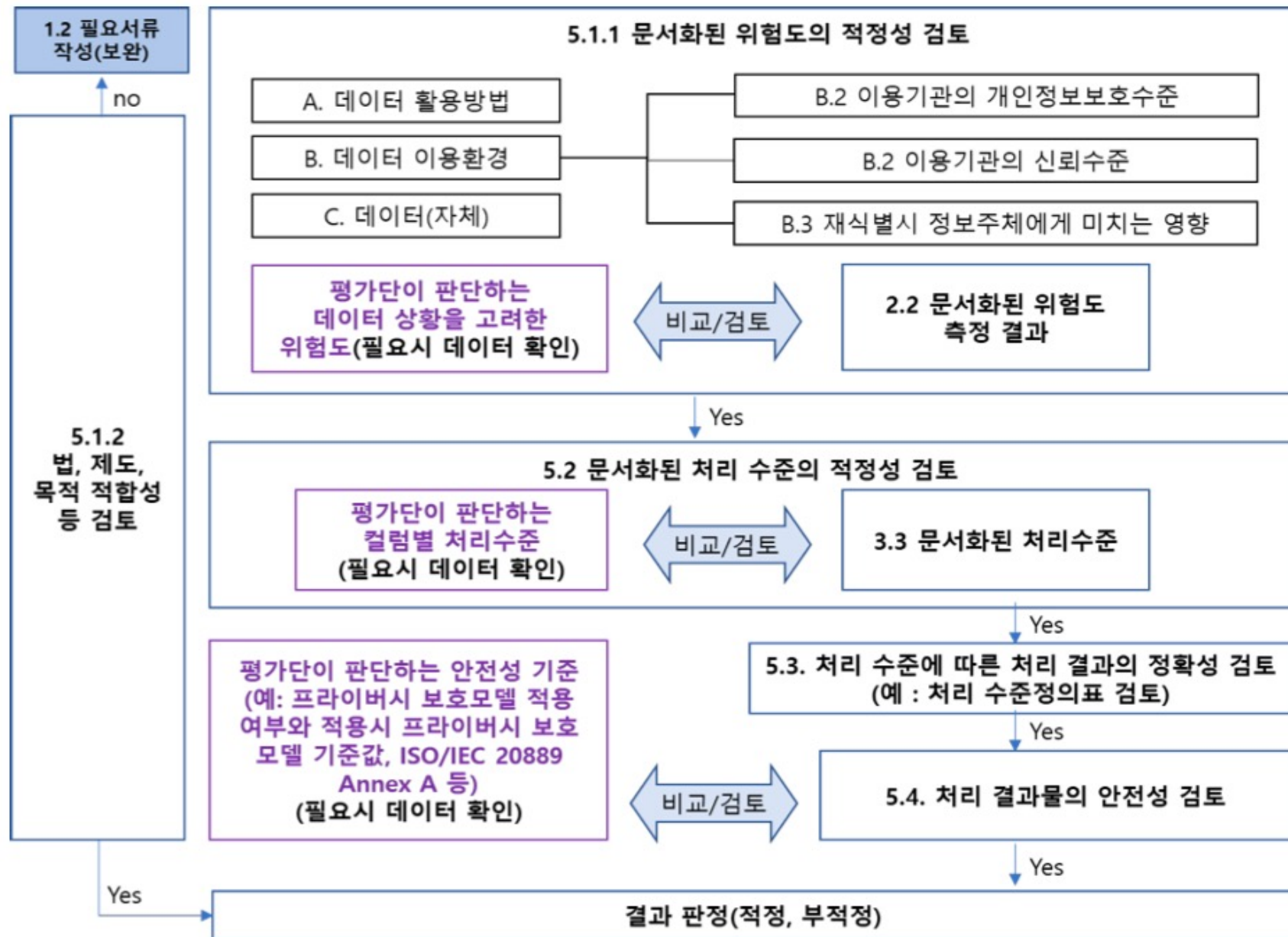
- 가명처리 적정성 검토 예시 (KISA 가명처리 전문가과정 자료)



가명/익명처리 적정성 검토 - 익명 예시



- 익명처리 적정성 검토 예시 (KISA 가명처리 전문가과정 자료)





- 적정성 검토 신청시 사전에 확인해야할 것들
 - 포함된 항목들은 모두 반드시 필요한 항목들인지(익명의 경우 필수는 아님)
 - 가장 좋은건 각 항목별로 사용 목적(필요사유) 명시
 - 식별정보는 모두 처리 되었는지(익명의 경우 모두 삭제)
 - 식별가능정보는 식별(선정)이 되지 않을 정도로 처리 되었는지(세분화가 많이 될수록 처리 수준이 높아야함)
 - 익명의 경우 k-익명성 모델에서 최소 k값 확인
 - 특이치는 모두 삭제 혹은 일반화 등을 통해 처리 하였는지
 - 처리 기술이 예외 없이 잘 적용되었는지
 - 위의 처리 수준이 처리 환경에 적합한지(가명만 해당. 익명의 경우 완전공개라고 가정)
 - 목적 달성을 할 수 있는지
 - 서류 : 데이터 명세서(각 항목별 분포) + 검토 신청서(구체적 목적, 처리환경 등) 등등





[처리환경별 위험도 분류 예시]

- ▶ ○○교육청이 A부서의 A1 개인정보를 가명처리하여 B부서에서 사용하며 보호수준은 높고, 재식별 가능성은 낮은 상태라고 가정할 경우

처리환경				기관의 개인정보 보호수준**	재식별수준 (다른정보 보유 등)**
활용장소	활용부서	처리유형	기관유형		
내부(1)	소유부서(1)	활용(1)	교육기관(1)	상(1)	상(3)
외부(2)	타부서(2)	제공(2)	민간기관(2)	중(2)	중(2)
		결합(3)		하(3)	하(1)

[표 4] 처리환경별 위험도 분류 측정 예시

- ※ 예시 표에서 점수는 처리환경의 다양성을 고려하여 가명정보처리자가 판단하여 점수 배정
- ▶ 각각의 분류 기준의 점수를 합산하여 처리환경 1+2+2+1=6점, 보호수준 1점, 재식별수준 1점, 총 8점으로 산정하고 점수가 높을수록 위험도가 높은 처리환경으로 판정할 수 있음
 - ▶ 처리환경별 위험도 점수별 가명처리 방안을 마련하여 해당 방안을 적용

가명/익명처리 적정성 검토 – 데이터 명세서 예시(금융보안원)



1. 데이터명세서

1-1) 개요

항목	내용
■ 의뢰기관명	A카드
■ 데이터 크기 (용량)	20.5 GB
■ 데이터 레코드 수	25,000,000
■ 데이터 컬럼 수	14
■ 데이터 생성 방법	18년~20년 A카드 이용자에 대한 카드상품 신청/이용/탈퇴 등의 내부정보 및 외부 조회 정보
■ 결합데이터 활용목적	B은행의 신규 신용카드 결합상품 개발시 참고하기 위한 통계작성
■ 결합데이터 이용기관	B은행

1-2) 정보 개요

No.	정보영역	항목건수	비중
1	KEY	1	7.1%
2	신청정보	6	42.9%
3	결제정보	2	14.3%
4	대출정보	2	14.3%
5	기타정보	3	21.4%
Total		14	100.0%

1 페이지

1-3) 정보항목별 상세

No.	정보원천	자료형태	정보영역	컬럼명	항목명(국문)	항목 설명	분포현황 작성	제출 데이터 예시	비고 (데이터 출처)
1	KEY	-	결합키	A_key	결합키	작성 불요		작성 불요	작성 불요
2	내부정보	수치형	신청정보	column2	연령			90	
3	내부정보	범주형	신청정보	column3	성별			남	
4	내부정보	범주형	신청정보	column4	직업코드	○○기관의 직업코드 활용	○	20201	
5	내부정보	수치형	신청정보	column5	신청일자			20180101	
6	내부정보	수치형	신청정보	column6	계좌개설일자			20180101	
7	내부정보	수치형	신청정보	column7	개설상품 건수		○	14	
8	내부정보	수치형	결제정보	column8	KK카드상품 결제금액	A카드의 KK카드상품	○	70000	
9	내부정보	수치형	결제정보	column9	MM카드상품 결제금액	A카드의 MM카드상품	○	70000	
10	외부정보	범주형	대출정보	column10	카드대출 연체액 등급	전체 카드대출상품 연체액을 15000단위로 1~7까지로 범주화	○	7	○○신용평가회사 데이터 조회
11	외부정보	수치형	대출정보	column11	카드대출 상환액	전체 카드대출상품	○	500000	○○신용평가회사 데이터 조회
12	외부정보	수치형	기타정보	column12	소득금액		○	500000	○○청 데이터 조회
13	외부정보	수치형	기타정보	column13	건강보험료납부액		○	500000	○○공단 데이터 조회
14	외부정보	수치형	기타정보	column14	지방세납부액		○	500000	○○청 데이터 조회

가명/익명처리 적정성 검토 - 데이터 명세서 예시(금융보안원)



2. 가명처리 요약

2. 가명처리 요약

2-1) 데이터 전체 적용

2-1) 데이터 전체 적용

가명처리 방법(데이터전체)			
No.	가명처리	대상 컬럼	설명
1	샘플링	전체 데이터	전체 데이터 80% 샘플링 적용
2			

데이터전문기관: 데이터 전반에 가명처리했을 경우 작성
(예) 샘플링 적용, 노이즈 추가 등

데이터전문기관:

가명처리했을 경우에만 처리방법을 작성
처리 방식이 여러 개 일 경우 셀 추가하여 작성

데이터전문기관:

가명처리를 적용하지 않은 경우, 미적용 사유를 작성

2-2) 데이터 컬럼별 적용

2-2) 데이터 컬럼별 적용

기본 정보						가명처리 방법										
No.	자료형태	정보영역	컬럼명	항목명(국문)	항목 설명	가명처리1	단위/기준	가명처리2	단위/기준	미적용 사유	(전) 원본 Max	(전) 원본 Max 빈도수	(후) 처리후 min	(후) 처리후 min 빈도수	(후) 처리후 Max	(후) 처리후 Max 빈도수
1	-	결합키	A_key	결합키	작성 불요	작성 불요	작성 불요	작성 불요	작성 불요	작성 불요	작성 불요	작성 불요	작성 불요	작성 불요	작성 불요	작성 불요
2	수치형	신청정보	column2	연령		상하단코딩	20/90				99	1	20	125	90	10
3	범주형	신청정보	column3	성별						성별 소변제단어 반드시 필요하여 미적용	-					
4	범주형	신청정보	column4	직업코드	OO기관의 직업코드 활용	범주화					20201	160			20201	160
5	수치형	신청정보	column5	신청일자						내부 원본정보만으로 분리되면 제식별 위험성 없음	-					
6	수치형	신청정보	column6	계좌개설일자						내부 원본정보만으로 분리되면 제식별 위험성 없음	-					
7	수치형	신청정보	column7	개설상품 건수						내부 원본정보만으로 분리되면 제식별 위험성 없음	-					
8	수치형	결제정보	column8	KK카드상품 결제금액	A카드의 KK카드상품	라운드(절삭)	10000				77,985	2			70,000	123
9	수치형	결제정보	column9	MM카드상품 결제금액	A카드의 MM카드상품	라운드(절삭)	10000				77,985	3			70,000	123
10	범주형	대출정보	column10	카드대출 연체액 등급	전체 카드대출상품 연체액율 1500만원 이하 1~7개치로 범주화	범주화	15000				222,000,000	1			7	1,231,231
11	수치형	대출정보	column11	카드대출 상환액	전체 카드대출상품	라운드(절삭)	10000	상단코딩	500000		222,000,000	2			500000	12,312,312
12	수치형	기타정보	column12	소득금액		라운드(절삭)	10000	상단코딩	500000		222,000,000	3			500000	12,341,234
13	수치형	기타정보	column13	건강보험료납부액		라운드(절삭)	10000	상단코딩	500000		222,000,000	1			500000	12,345,123
14	수치형	기타정보	column14	지방세납부액		라운드(절삭)	10000	상단코딩	500000		222,000,000	2			500000	12,121,212



3. 범주형 변수 분포현황

※ 최종 제출하는 데이터 중 민감 또는 주요 정보에 대하여 작성하며, 평가 중 추가 작성 요청이 있을 수 있음

No.	컬럼명	항목명(국문)	컬럼값	세부 설명	빈도수	구성비
4	column4	직업코드	20201	금융업	160	13.95%
			20202	건설업	170	14.82%
			20203	서비스업	153	13.34%
			20204	유통업	160	13.95%
			20205	정보통신업	186	16.22%
			20206	농업	164	14.30%
			20207	수산업	154	13.43%
10	column10	카드대출 연체액 등급	1	0~15000	160	13.95%
			2	15001~30000	170	14.82%
			3	30001~45000	153	13.34%
			4	45001~60000	160	13.95%
			5	600001~75000	186	16.22%
			6	75001~90000	164	14.30%
			7	90001~	154	13.43%

4. 수치형 변수 분포현황

※ 최종 제출하는 데이터 중 민감 또는 주요 정보에 대하여 작성하며, 평가 중 추가 작성 요청이 있을 수 있음

No.	컬럼명	항목명(국문)	최소값	1분위수(25%)	2분위수(중앙값)	3분위수(75%)	최대값
7	column7	개설상품 건수	0	5	13	20	25
8	column8	KK카드상품 결제금액	0	10000	40000	70000	77985
9	column9	MM카드상품 결제금액	0	10000	40000	70000	77985
11	column11	카드대출 상환액	0	100000	250000	400000	500000
12	column12	소득금액	0	100000	250000	420000	500000
13	column13	건강보험료납부액	0	100000	250000	400000	500000
14	column14	지방세납부액	0	100000	250000	420000	500000

가명/익명처리 적정성 검토 - 데이터 명세서 예시(금융보안원)



2. 익명처리 요약

익명처리를 적용하지 않은 경우, 미적용 사유를 작성

2-1) 데이터 전체 적용

익명처리 방법			
No.	익명처리	대상 컬럼	설명
1	샘플링	전체 데이터	전체 데이터 80% 샘플링 적용
2	모델링(K-익명성)	연령, 성별, 직업코드, 신청일자, 계좌개설일자	k=3 적용, k=3 미만 레코드 삭제처리

데이터전문기관: 데이터 전반에 익명처리했을 경우 작성
(예) 샘플링 적용, 모델링(K-익명성, L-다양성, t-근접성), 노이즈 추가 등

데이터전문기관:

익명처리했을 경우 작성하며, 처리 방식이 여러 개 일 경우 셀 추가하여 작성

2-2) 데이터 컬럼별 적용

기본 정보						개인식별 가능정보	익명처리 방법			2 페이지							
No.	정보형태	정보영역	컬럼명	항목명(국문)	항목 설명		익명처리1	단위/기준	익명처리2	단위/기준	미적용 사유	(전) 원본 Max	(전) 원본 Max 빈도수	(후) 처리후 min	(후) 처리후 min 빈도수	(후) 처리후 Max	(후) 처리후 Max 빈도수
1	수치형	신청정보	column1	연령		0	상단코딩	20/90				99	1	20	125	99	1
2	범주형	신청정보	column2	성별		0					성별 소비패턴이 반드시 필요하여 미적용						
3	범주형	신청정보	column3	직업코드	OO기관의 직업코드 활용	0	범주화										
4	수치형	신청정보	column4	신청일자		0					내부 원본정보이므로 분리되면 재식별 위험성 낮음						
5	수치형	신청정보	column5	계좌개설일자		0					내부 원본정보이므로 분리되면 재식별 위험성 낮음						
6	수치형	신청정보	column6	개설상품 건수		-					내부 원본정보이므로 분리되면 재식별 위험성 낮음						
7	수치형	결제정보	column7	KK카드상품 결제금액	A카드의 KK카드상품	-	라운딩(절삭)	1000	상단코딩	70,000		77,985	2			70,000	123,123
8	수치형	결제정보	column8	MM카드상품 결제금액	A카드의 MM카드상품	-	라운딩(절삭)	1000	상단코딩	70,000		77,985	3			70,000	123,456
9	범주형	대출정보	column9	카드대출 연체액 등급	신용 카드대출상품 연체액율 1000만원 이하 1~7개월로 분류	-	범주화	15000				222,000,000	1			7	1,231,231
10	수치형	대출정보	column10	카드대출 상환액	전체 카드대출상품	-	라운딩(절삭)	1000	상단코딩	500,000		222,000,000	2			500000	12,312,312
11	수치형	기타정보	column11	소득금액		-	라운딩(절삭)	1000	상단코딩	1000/500000		222,000,000	3	1,000	12,345	500000	12,341,234
12	수치형	기타정보	column12	건강보험료납부액		-	라운딩(절삭)	1000	상단코딩	1000/500000		222,000,000	1	1,000	123,456	500000	12,345,123
13	수치형	기타정보	column13	지방세납부액		-	라운딩(절삭)	1000	상단코딩	1000/500000		222,000,000	2	1,000	123,123	500000	12,121,212



가명/익명처리 적정성 검토

1. (검토를 받는 입장에서) 적정성 검토
2. **프라이버시 보호 모델**



- 프라이버시 보호 수준을 통계적 기법을 활용하여 정량적으로 나타내는 방식
- 개인이 직접 식별되는 것뿐만 아니라 추론을 통해 식별되는 것도 방지
- 가명처리에는 거의 사용되지 않으며, 주로 익명처리 기법에 사용
- 각 재식별 공격별로 프라이버시 보호 모델이 존재하고 용도가 정해져 있으며, 하나의 프라이버시 보호 모델로 모든 재식별 위험을 제거할 수 없음
- 단점 재식별 위험성만 평가하고 데이터 유용성은 평가하지 않음



- ID : Identifier, 식별자, 개인을 식별할 수 있는 속성들. 암호화된 값도 식별자로 분류됨. 식별정보
- QI : Quasi-Identifier, 준식별자, 자체로는 식별자가 아니지만 다른 데이터와 결합을 통해 특정 개인을 간접적으로 추론하는데 사용될 수 있는 속성들(주로 인구통계학적으로 분류되는 속성들), 식별가능정보와 유사한 개념
- 동질 집합(EC; Equivalent Class) : QI의 조합의 집합
- SA : Sensitive Attribute, 민감속성. 민감한 정보, 식별자를 통해 신원이 공개되는 경우 개인에게 가볍게는 수치심으로부터 심각하게는 사회적 차별, 경제적 피해를 끼칠 수 있는 모든 정보. 개인이 스스로 공개되기를 원하지 않는 정보.
 - > 보호해야할 정보
 - 법률적 민감정보
 - 식별성 희귀정보
 - 낙인성 정보
 - 기타 민감한 정보

프라이버시 보호 모델 - k-Anonymity



- k-익명성 모델
 - 동일한 속성(동질 집합)을 가지는 레코드가 최소한 k개 이상 존재하도록 하여 프라이버시를 보호(k=3일 경우, 동일한 개인식별가능정보 중 식별가능성이 높은 정보를 가지는 사람이 3명 이상 존재하여 특정 개인 식별이 불가) -> single out 확인

〈k-익명성 모델 적용 전〉

[식별자 제거 데이터]			연결공격 (Linkage Attack)	[확보된 공개 데이터]		
연령	성별	카드 결제금액	1:1	이름	연령	성별
60	남	320,000	↔	김철수	60	남
62	여	600,000	↔	이민아	62	여
61	남	500,000	↔	안상태	61	남
27	남	1,500,000	↔	김상우	27	남
29	남	1,000,000	↔	한기범	29	남
27	여	1,750,000	↔	장아름	27	여
26	여	1,400,000	↔	양수지	26	여
60	여	150,000	↔	김다래	60	여
61	남	145,000	↔	윤영하	61	남
60	여	402,000	↔	김순자	60	여
28	남	1,330,000	↔	김민영	28	남
25	여	1,220,000	↔	유슬아	25	여

〈k개(3개) 이상의 레코드가 존재하는 동질집합을 구성하여 1:1 연결 방지〉

[k-익명성 적용 데이터]				[확보된 공개 데이터]		
연령	성별	카드 결제금액	k : 1	이름	연령	성별
60대	남	320,000	↔	김철수	60	남
60대	남	500,000	↔	이민아	62	여
60대	남	145,000	↔	안상태	61	남
60대	여	600,000	↔	김상우	27	남
60대	여	150,000	↔	한기범	29	남
60대	여	402,000	↔	장아름	27	여
20대	남	1,500,000	↔	양수지	26	여
20대	남	1,000,000	↔	김다래	60	여
20대	남	1,330,000	↔	윤영하	61	남
20대	여	1,750,000	↔	김순자	60	여
20대	여	1,400,000	↔	김민영	28	남
20대	여	1,220,000	↔	유슬아	25	여

- 레코드 연결 공격 -> k-익명성으로 방어
 - 공격자는 준식별자를 기반으로 레코드들을 매칭시킴
 - 알 수 없었던 특정한 개인의 신원이 드러나는 프라이버시 유출 공격



- k-익명성 모델
 - 취약점
 - 동질성 공격(Homogeneity attack) : k-익명성에 의해 레코드들이 범주화 되었다더라도 동질 집합 내에서 민감정보들이 모두 같은 값을 가질 수 있기 때문에 데이터 집합에서 동일한 정보를 이용하여 공격 대상의 정보를 알아내는 공격
 - 배경지식에 의한 공격(Background knowledge attack) : 주어진 데이터 이외의 공격자의 배경 지식을 통해 공격 대상의 민감한 정보를 알아내는 공격

[K-익명성 적용 데이터]

연령	성별	우편번호	신용등급
60대	남	180**	8
60대	남	180**	8
60대	남	180**	8
60대	남	180**	8
60대	여	180**	1
60대	여	180**	3
60대	여	180**	5
60대	여	180**	2

동질성 공격 (Homogeneity Attack)

○○○지역의 모든 60대 남자의
신용등급은 8등급
(추론 예) ○○○지역에 사는 남자,
박철수의 신용등급은 8등급
↓
동질집합 내에서 다양성이 부족하여
특정 개인의 정보 추론 가능

지역 코드	연령	성별	질병
130**	< 30	*	전립선염
130**	< 30	*	전립선염
130**	< 30	*	고혈압
130**	< 30	*	고혈압

선거인 명부에서 '김영희' 정보 획득(13068, 29세)
배경지식 : 여자는 전립선염에 걸릴 수 없음 -> 고혈압 환자



- I-다양성 모델
 - 동질집합의 민감속성정보가 최소한 1개의 다양한 속성을 가지도록 하여 k-익명성의 취약점(동질성 공격, 배경지식 공격)을 보완함

〈I-다양성 모델 적용 전〉

[K-익명성 적용 데이터]

연령	성별	우편번호	신용등급
60대	남	180**	8
60대	남	180**	8
60대	남	180**	8
60대	남	180**	8
60대	여	180**	1
60대	여	180**	3
60대	여	180**	5
60대	여	180**	2

동질성 공격 (Homogeneity Attack)

○○○지역의 모든 60대 남자의
신용등급은 8등급
(추론 예) ○○○지역에 사는 남자,
박철수의 신용등급은 8등급



동질집합 내에서 다양성이 부족하여
특정 개인의 정보 추론 가능

〈I-다양성 모델 적용 후〉

동질집합이 1개(3개)의 다양한 민감정보(신용등급)를 가지도록 조정

[I-다양성 적용 데이터]

연령	성별	우편번호	신용등급
60대	*	1803*	8
60대	*	1803*	8
60대	*	1803*	5
60대	*	1803*	2
60대	*	1804*	1
60대	*	1804*	3
60대	*	1804*	8
60대	*	1804*	8

60대 남자의 신용등급
추론 가능성 낮아짐
(|값= 3)



- I-다양성 모델
 - 취약점
 - 쓸림 공격(skewness attack): 정보가 특정한 값에 쏠려 있는 경우 I-다양성 모델이 프라이버시를 보호하지 못함
 - 유사성 공격(similarity attack) : 익명처리된 레코드의 정보가 서로 비슷하다면 I-다양성 모델을 통해 처리되었다 할지라도 프라이버시가 노출될 수 있음

[I-다양성 적용 데이터]

연령	성별	우편번호	소득
60대	남	180**	10,000
60대	남	180**	50,000
60대	남	180**	60,000
60대	남	180**	15,000
60대	여	180**	35,000,000
60대	여	180**	100,000,000
60대	여	180**	175,000,000
60대	여	180**	24,000,000

쓸림 공격 (Skewness Attack)

○○○지역의 60대 남자의
소득은 매우 낮다
(추론 예) ○○○지역에 사는 남자,
박철수의 소득은 매우 낮다



특정한 값에 쏠린 특성을 이용하여
개인의 정보 추론 가능

	준식별자		민감 정보	
	우편번호	연령	급여 (백만원)	질병
1	476**	2*	30	위궤양
2	476**	2*	40	급성 위염
3	476**	2*	50	만성 위염
4	4790*	≥ 40	60	급성 위염
5	4790*	≥ 40	110	감기
6	4790*	≥ 40	80	기관지염
7	476**	3*	70	기관지염
8	476**	3*	90	폐렴
9	476**	3*	100	만성 위염



• t-근접성 모델

- 특정 동질집합의 개인식별가능정보 분포와 전체 데이터의 개인식별가능정보 분포 차이를 t 이하가 되도록 조정(t 가 0에 가까울수록 분포가 유사하며, 이를 통해 특정집단의 개인식별가능정보 추론문제 보완)

Earth Mover's Distance (EMD)를 사용하여 원본 테이블과 동질 클래스의 민감한 속성 값의 분포가 얼마나 가까운지(closeness)를 계산하며, EMD가 $t(0 \leq t \leq 1)$ 이하일 때 t-근접성을 만족

[정의] 원본 테이블을 T , 익명화한 테이블의 어떤 동질 클래스를 EC 라 하자. 모든 동질 클래스에서 $EMD(T, EC) \leq t$ 일 때, t-근접성을 만족

[문제점] 모든 동질 클래스에서 EMD를 t 이하로 강제하기 때문에 익명화 과정에서 준식별자와 민감 속성간의 관계가 크게 손상될 가능성이 높음

〈 t-근접성 모델 적용 전 〉

[I-다양성 적용 데이터]

연령	성별	우편번호	소득
60대	남	180**	10,000
60대	남	180**	50,000
60대	남	180**	60,000
60대	남	180**	15,000
60대	여	180**	35,000,000
60대	여	180**	100,000,000
60대	여	180**	175,000,000
60대	여	180**	24,000,000

쓸림 공격
(Skewness Attack)

○○○지역의 60대 남자의
소득은 매우 낮다
(추론 예) ○○○지역에 사는 남자,
박철수의 소득은 매우 낮다



특정한 값에 쓰린 특성을 이용하여
개인의 정보 추론 가능

〈 t-근접성 모델 적용 후 〉

[t-근접성 적용 데이터]

연령	성별	우편번호	소득
60대	*	1803*	10,000
60대	*	1803*	50,000
60대	*	1803*	175,000,000
60대	*	1803*	24,000,000
60대	*	1804*	35,000,000
60대	*	1804*	100,000,000
60대	*	1804*	60,000
60대	*	1804*	15,000

기타속성자 분포의 특성을
이용한 추론 방지

60대 남자의 소득수준
추론 가능성 낮아짐



ARX

Data Anonymization Tool

ARX is a comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analyzing the usefulness of output data.

The software has been used in a variety of contexts, including commercial big data analytics platforms, research projects, clinical trial data sharing and for training purposes.

ARX is able to handle large datasets on commodity hardware and it features an intuitive cross-platform graphical user interface. You can find further information [here](#), or directly proceed to our [downloads](#) section.



- QI : Quasi-Identifier, 준식별자, 자체로는 식별자가 아니지만 다른 데이터와 결합을 통해 특정 개인을 간접적으로 추론하는데 사용될 수 있는 속성들(주로 인구통계학적으로 분류되는 속성들), 식별가능정보와 유사한 개념

- SA : Sensative Attribute, 민감속성. 민감한 정보, 식별자를 통해 신원이 공개되는 경우 개인에게 가법게는 수치심으로부터 심각하게는 사회적 차별, 경제적 피해를 끼칠 수 있는 모든 정보. 개인이 스스로 공개되기를 원하지 않는 정보.

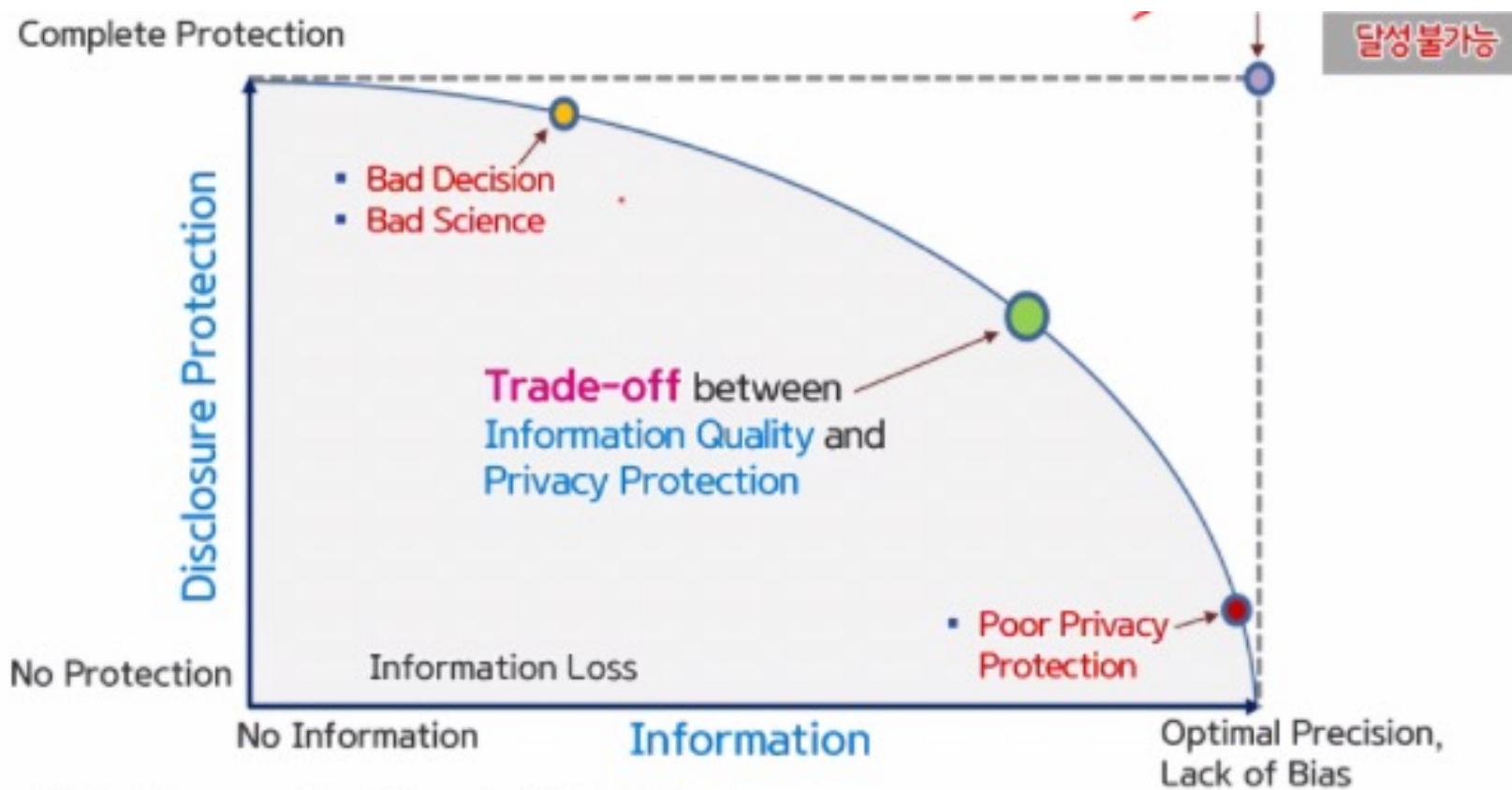
-> 보호해야할 정보

- 법률적 민감정보
- 식별성 희귀정보
- 낙인성 정보
- 기타 민감한 정보

가명/익명처리 적정성 검토 방법론 - 유용성



- 유용성과 프라이버시 보호는 트레이드 오프 관계



(참고 : twitter.com/dbarthjonse, Daniel Barth Jones)



가명/익명처리 기술과 적정성 검토

감사합니다.



가명/익명처리 적정성 검토 - 안전성



- ISO/IEC 20889:2018 Annex. A
하나의 처리 기술만을 사용하는 것이 아니라 여러 기술을 함께 사용해야 함.

기술의 종류	공격에 대해 방어가능 여부		
	Single out	연결 공격	추론 공격
통계도구			
총계처리	O	O	O
암호도구			
결정성 암호/ 순서 보존 암호	X	△	X
동형 암호/ 동형비밀분산기법	X	X	X
삭제			
마스킹	O	△	X
로컬 마스킹/ 레코드 제거/ 표본 추출	△	△	△
가명화	X	△	X

기술의 종류	공격에 대해 방어가능 여부		
	Single out	연결공격	추론 공격
일반화			
라운드/상하단 코딩	X	△	△
무작위화			
잡음추가/순열	△	△	△
마이크로어그리 게이션	X	△	△
차분프라이버시	O	O	△
k-익명성	O	△	X
표본추출의 경우, 표본에 단일 속성이 미포함 되었을 경우에 한함			

Technique name	Data truthfulness at record level	Applicable to types of values	Applicable to types of attributes	Reduces the risk of		
				Singling out	Linking	Inference
Statistical tools						
Sampling						
Aggregation	N/A	Continuous, discrete	All attributes	Yes	Yes	Yes
Cryptographic tools	Yes					
Deterministic encryption	Yes	All	All attributes	No	Partially	No
Order-preserving encryption	Yes	All	All attributes	No	Partially	No
Homomorphic encryption	Yes	All	All attributes	No	No	No
Homomorphic secret sharing	Yes	All	All attributes	No	No	No
Suppression	Yes					
Masking	Yes	Categorical	Local identifiers	Yes	Partially	No
Local suppression	Yes	Categorical	Identifying attributes	Partially	Partially	Partially
Record suppression	Yes	N/A	N/A	Partially	Partially	Partially
Sampling	Yes	N/A	N/A	Partially ^a	Partially	Partially
Pseudonymization	Yes	Categorical	Direct identifiers	No	Partially	No
Generalization	Yes	All, subject to meaning	Identifying attributes			
Rounding	Yes	Continuous	Identifying attributes	No	Partially	Partially
Top/bottom coding	Yes	Continuous, ordinal	Identifying attributes	No	Partially	Partially
Randomization	No		Identifying attributes			
Noise addition	No	Continuous	Identifying attributes	Partially	Partially	Partially

^a If the data principal record is not included in the sample.

^b Unless K-anonymity is implemented using microaggregation.

N/A Not applicable.

ISO/IEC 20889 Annex A.

Technique name	Data truthfulness at record level	Applicable to types of values	Applicable to types of attributes	Reduces the risk of		
				Singling out	Linking	Inference
Permutation	No	All	Identifying attributes	Partially	Partially	Partially
Micro aggregation	No	Continuous	Indirect identifiers, and all attributes	No	Partially	Partially
Differential privacy	No	All	Identifying attributes	Yes	Yes	Partially
K-anonymity	Yes ^b	All	Quasi identifiers	Yes	Partially	No

^a If the data principal record is not included in the sample.

^b Unless K-anonymity is implemented using microaggregation.

N/A Not applicable.



- ISO/IEC 20889:2018
 - Privacy enhancing data de-identification terminology and classification of techniques
- ENISA(유럽네트워크정보보호원) – The European Union Agency for Cybersecurity
 - Recommendations on shaping technology according to GDPR provisions - An overview on data pseudonymization, 2018
 - Pseudonymisation techniques and best practices, 2019
 - Data Pseudonymisation - Advanced Techniques and Use Cases, 2021
- 개보위, 금융위, 교육부, 보건복지부, 행안부 등에서 발행한 가이드라인
- 이 외에도....
 - NIST IR 8053 (2015)
 - De-identification of Personal Information
 - ISO 25237 (2017)
- 등등