개인정보보호론

[실시간 수업]

바이오인식정보

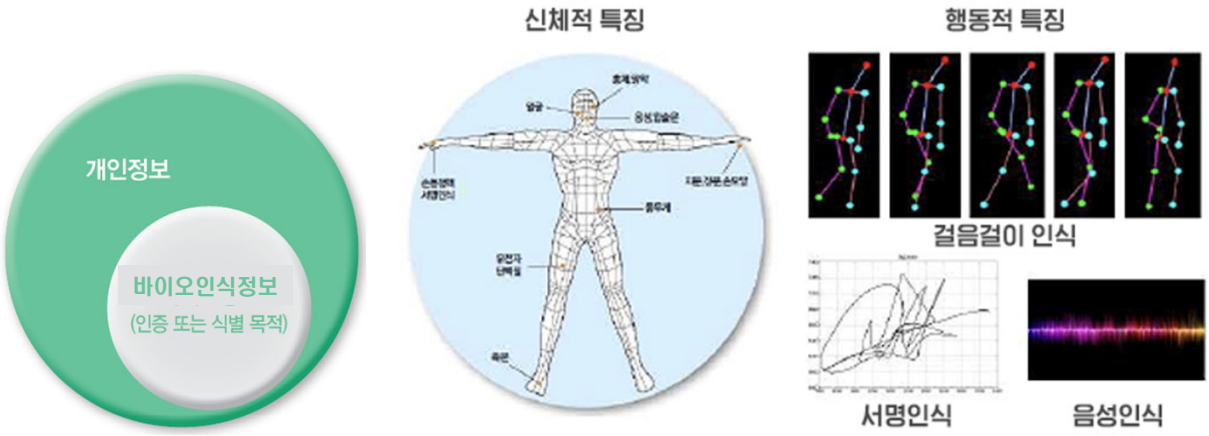
바이보정보

- 생체정보뿐만 아니라 유전정보, 건강 관련 정보를 포괄하는 개념

바이오인식정보 = 개인정보

- 지문, 홍채, 음성, 필적 등 개인의 신체적.행동적 특성에 관한 정보로서 개인을 인증 또는 식별하기 위하여 기술적으로 처리되는 개인 정보

신체적·행동적 특성에 관한 정보



개인 인증 및 식별

- 인증(검증, verification, authentication) – 제시된 개인정보와 기존 정보, 둘 사이의 동일성의 비교를 통해 (같다/다르다)를 판단 ⇒ 정말 그 사람인가?

Verification(Authentication) 인증
- ID given => yes / no
- "Decision boundary" is the issue

- 식별(인식, recognition, identification) – 등록된 정보와 제시된 개인정보의 비교를 통해 등록된 정보가 있는지 확인 ⇒ 누구인가?

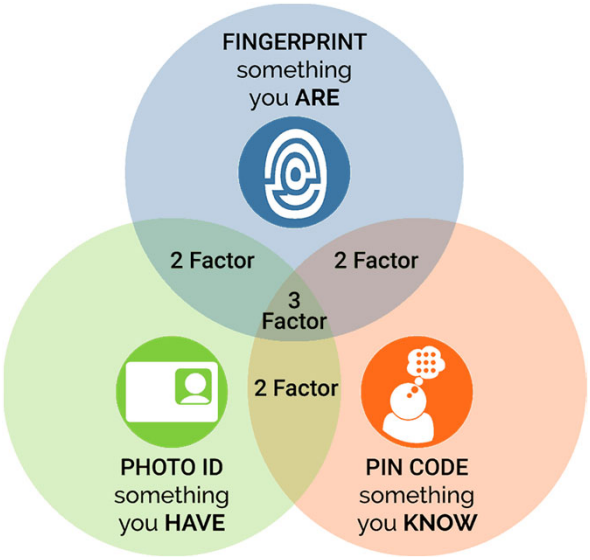
Recognition(Identification) 확인
- Who is the most likely in the DB?
- "Search" is the issue

- **인증은 기기에 입력된 바이오정보와 대조해 특정 개인을 확인하는 것**
  - 지문·홍채·안면인식 등을 이용한 스마트폰 잠금 해제

- **식별은 데이터베이스에 저장된 다수 바이오정보와 대조해 여러 사람 중 특정인을 확인하는 것**
  - 페이스북에 사진을 올리면, 안면인식을 통해 특정 개인을 태그 하는 서비스

인증(Authentication)

- **지식 기반 : 사용자가 알고 있는 것 (something you know)**
  - Ex) PW, PIN
- **소유 기반 : 사용자가 소유하고 있는 것 (something you have)**
  - Ex) 스마트카드, 토큰
- **존재 기반 : 사용자만의 고유한 특징 (something you are)**
  - ex) 홍채, 지문



지식 기반 인증

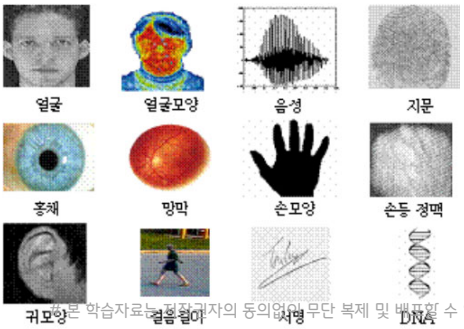
- **사용자가 알고 있는 것**
  - 패스워드, 개인 식별 번호(PIN), 자물쇠 번호 등
- **단점**
  - 사용자가 개인 정보를 잊어버리거나, 다른 사람이 인증에 사용 정보를 입수하여 시스템에 불법적으로 접근 가능
- **장점**
  - 설치비용 적음

소유 기반 인증

- **사용자가 소유하고 있는 인증방법**
  - 열쇠(카드 키), 티켓, 패스포트, 토큰, 스마트카드, 액세스 카드, 배지 등
- **사용자가 도구들을 잃어버려서 분실하거나 도난 될 경우 불법적인 시스템 접근에 악용될 수 있음**

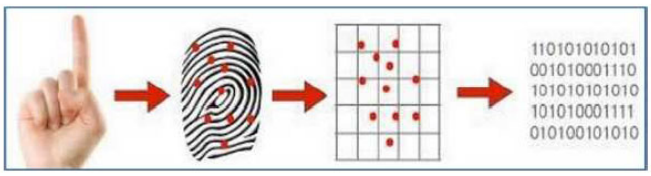
존재 기반 인증

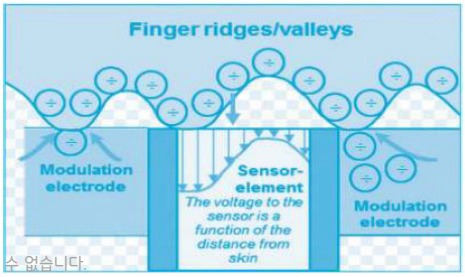
- **사용자만의 고유한 특징을 이용한 인증방법**
  - 바이오인식 기술
- **유니크한 정보로 한번 노출되면 회복 불가**



지문 인식(fingerprint recognition)

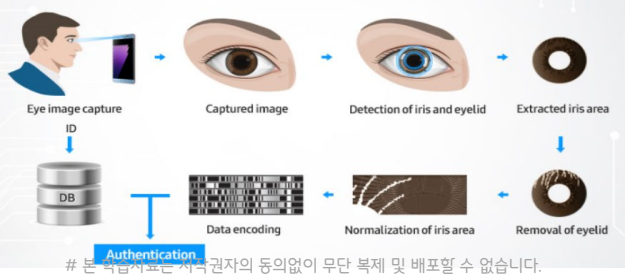
- **지문인식은 보통 지문 융기의 분기점, 끝점 등으로 구성되는 특징점의 위치와 속성을 추출, 저장, 비교하는 알고리즘을 채용**
  - 지문인식은 인식방법에 따라 정전용량 방식, 광학 방식, 초음파 방식 등으로 구분





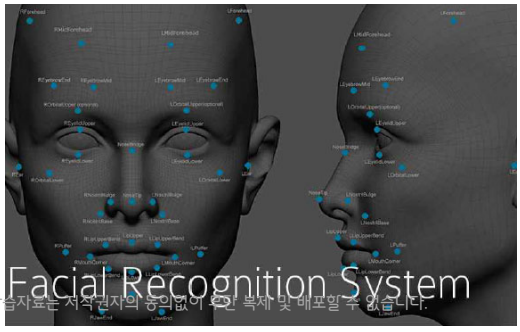
홍채 인식(Iris Recognition)

- **홍채인식은 안구 중앙의 검은 동공과 흰자위 사이에 있는 도넛 모양의 홍채를 이용한 인증 기술**
  - 홍채인식 장치의 적외선 카메라가 홍채를 이미지화한 후, 홍채인식 알고리즘으로 사용자 고유의 홍채 코드를 생성, 등록 후 비교하는 방식



얼굴 인식

- **얼굴 인식은 각 개인 얼굴의 특징을 이용**
  - 카메라를 통해 입력된 화상으로부터 각 개인마다 독특한 부위를 측정 단위로 추출하는 것으로, 독특한 부위가 어떠한 곳인지 결정하는데 이 기술의 정확도가 달려있음



필체인식 (Signature)

- **필체인식 혹은 서명인식은 개인서명의 고유한 특징을 이용하여 인증하는 기술**
  - 이미 작성된 서명을 인식하는 정적인 방법과 서명하는 과정을 동적으로 파악하는 방법으로 구분
    - ✓ 동적서명 인식은 새로운 서명 샘플과 원본 데이터 서명의 모양을 단순히 비교하는 방법이 아니라 원본 데이터와 샘플링된 데이터가 쓰여지는 방법을 비교하는 것으로 서명시간, 속도, 압력, 종이로부터 펜이 떨어진 횟수 등을 이용

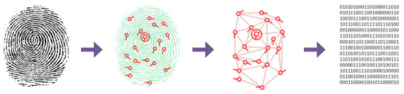
바이오인식정보 특성

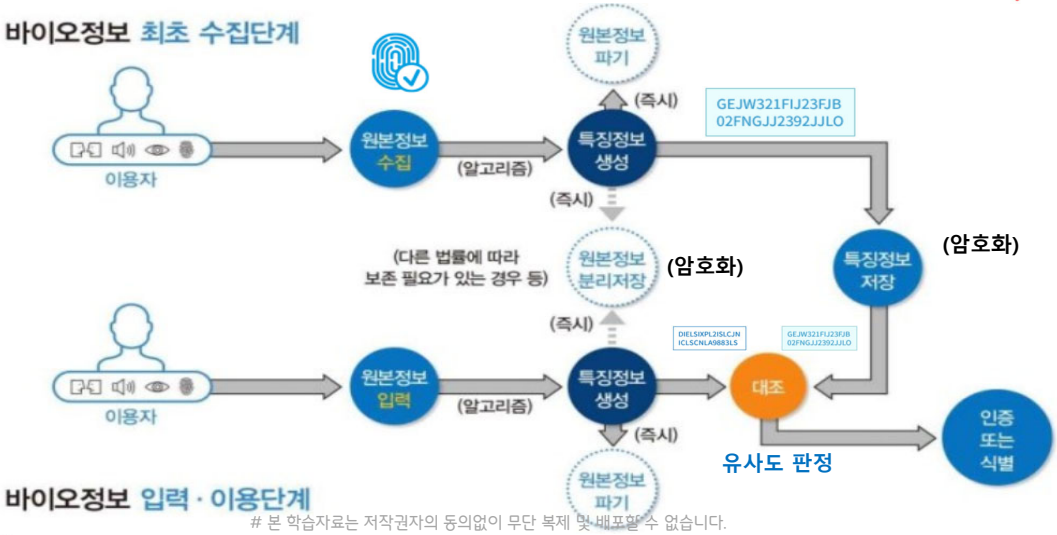
- **보편성(Universaility) : 모든 사람에게 있는 특징이어야 한다**
  - 그렇지 않다면 바이오인식을 처리하지 못하고 오류가 발생할 것이다.
- **고유성 (Uniqueness) : 서로 다른 개인을 식별할 수 있기 위해서는 사람마다 가지고 있는 정보가 달라야 한다**
  - 즉, 그 개인에게 고유한 정보여야 한다
- **영구성(permanency) : 평생 변하지 않는 특성을 가지고 있어야 한다**
  - 그렇지 않고 시간이 지나면서 혹은 어떤 영향으로 변화한다면 해당 개인을 더 이상 식별하거나 인증할 수 없게 될 것이다

바이오인식정보 시스템 4단계

- 1. 획득 : 바이오 특성을 디지털 형태로 변환
- 2. 특징추출 : 사람마다 고유하면서 변별력이 높은 특징점 추출
- 3. 비교 : 등록된 특징과 입력된 특징을 신속 정확하게 비교
- 4. 유사도 판정 : 비교된 두 특징들이 동일인의 특성인가를 판단

바이오인식정보 시스템





원본정보 분리 저장

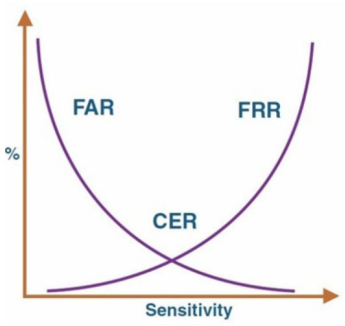
A모델

B모델

유사도 판정 : 바이오인식정보 정확성 평가 지표

- FRR(False Rejection Rate) **오거부율**
  - 오거부율은 바이오인식 시스템의 에러로 개인을 식별하지 못하는 등 인식을 거부할 확률
- FAR(False Acceptance Rate) **오인식률**
  - 오 인식률은 타인의 바이오인식 정보를 특정인의 것으로 잘못

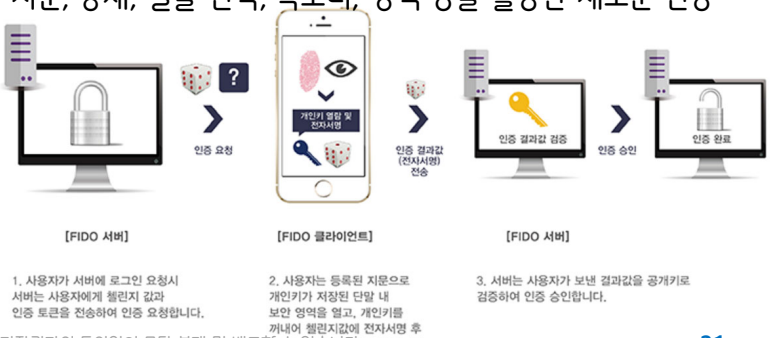
➢ **CER(Crossover Error Rate) / EER(Equal Error Rate)**
  - ➢ **FRR과 FAR이 일치하는 지점으로 수치가 낮을수록 정확**
  - ✓ **사용자 편의성이 높아지는 경우 FRR은 낮아지고 FAR이 높아진다**
  - ✓ **보안성을 강화할 경우에는 FAR은 낮아지고 FRR이 높아진다**



FIDO (Fast IDentity Online)

- **FIDO 인증 : 온라인 환경에서 사용자의 신원을 편리하고 안전하게 인증하기 위한 기술 표준으로, 주로 사용자 개인의 고유한 바이오정보를 이용하는 인증 기술**
  - 아이디와 비밀번호 조합 대신 지문, 홍채, 얼굴 인식, 목소리, 정맥 등을 활용한 새로운 인증

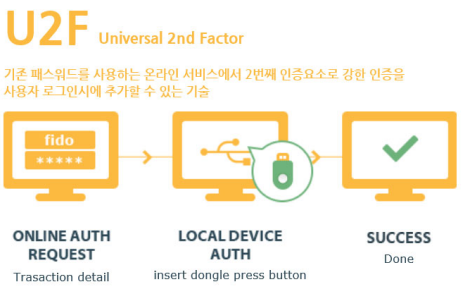
구분	서버저장 방식	FIDO 방식
방식	• 개인 생체정보를 **서버에 저장** • 바이오 인식 단말에서 추출한 정보와 비교	• 개인 생체정보를 **단말기에 저장** • 전자서명 방식으로 단말기에서 인증 과정 진행
사례	• 신한은행 디지털 키오스크 • 기업은행 홍채인증 ATM	• 우리은행, 신한은행, KEB하나은행 등 주요 은행 모바일뱅킹



FIDO 방식 : UAF & U2F

- UAF(Universal Authentication Framework)
  - 사용자의 단말기에서 제공하는 인증방법을 온라인 서비스와 연동하여 인증하는 기술로 패스워드 없이(passwordless) 바이오정보만으로 인증을 완료하는 것
- **UAF 방식은 스마트폰과 같은 모바일 환경에 적합**
  - 모바일 기기에는 지문 인식 모듈, 홍채 인식 카메라, 마이크 등이 탑재돼 바이오 정보를 인식하기 위한 기반이 마련되어 있기 때문

- U2F(Universal 2nd Factor)
  - 기존 패스워드를 사용하는 지식기반 인증에서 USB, NFC 보안키, 바이오인증 등의 두 번째 인증요소를 추가 하는 것
- 기존 PC 기반 온라인 서비스에 적합
  - PC 기반 온라인의 경우 ID/패스워드 기반의 개인 인증 시스템이 주로 사용되어 바이오인증 방식으로의 갑작스런 전환은 사용자의 편의성을 저해할 우려가 있으며, 바이오인증 시스템 구축을 위한 전환비용 등의 문제 역시 존재

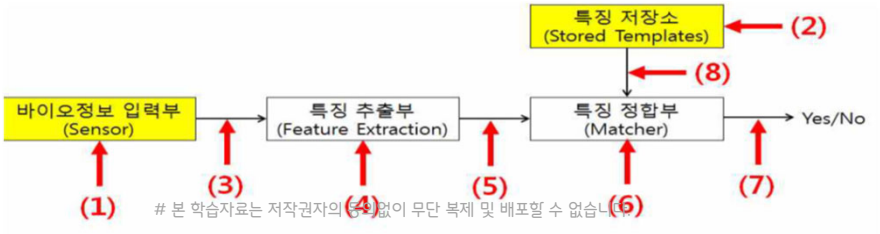


바이오인식정보 해킹

- Chaos Computer Club(CCC) 해커가 독일 국방장관 폰데어라이엔 지문 획득
- 독일의 해커단체 CCC는 구글검색을 통해 러시아 대통령 푸틴의 고해상도 사진을 출력하여 홍채 복제(Print attack)*
- 독일의 시큐리티리서치랩스는 목재용 접착제에 사용자 지문을 복제하여 지문인식 잠금장치 해제

바이오인식정보 인증시스템의 보안상 취약점

- **바이오인식정보 인증시스템은 크게 4개의 모듈로 구분됨**
  - 바이오정보 입력부 : 센서를 통해 바이오정보를 취득
  - 특징 추출부 : 취득된 바이오정보로 부터 특징정보를 추출
  - 특징 저장소 : 특징정보 및 개인정보 등을 저장하는 저장소
  - 특징 정합부 : 저장된 특징정보와 새로 입력된 특징정보를 비교하여 인증여부를 결정



(1) 위조지문, 고해상도 사진 등 위조된 바이오 정보를 센서에 입력하여 인증을 우회

(2) 저장소에 침투하여 기 저장된 특징정보를 조작, 삭제, 유출

(3) 불법 취득한 바이오 정보를 재생(replay)하여 인증

(4) 위조된 특징정보를 임의로 생성

(5) 정상적인 특징정보를 임의의 위조된 특징정보로 대체

(6) 특징 정합부에서 인증 결과값을 임의로 변경

(7) 최종 인증결과를 조작

(8) 저장소에서 정합부로 전송되는 특징정보를 절취 또는 타인의 정보로 대체

바이오정보 보호 가이드라인

- **대상 정보**
  - 정보통신망법상 바이오정보(지문, 홍채, 음성, 필적 등 개인을 식별할 수 있는 신체적 또는 행동적 특징에 관한 정보)
- **바이오정보보호 원칙**
  - 1) 비례성 원칙
  - 2) 수집·이용 제한의 원칙
  - 3) 목적제한의 원칙
  - 4) 통제권 보장의 원칙
  - 5) 투명성 원칙
  - 6) 바이오정보 보호 중심설계 및 운영원칙

바이오정보 보호 6원칙

- **비례성 원칙**
  - 바이오정보를 활용함에 따라 수반되는 위험이 사업 상 바이오정보의 필요성 및 예상되는 편익에 비해 과도하지 않은지 등을 검토 후, 수집·이용 여부를 판단하여야 한다

원칙	세부 원칙	원칙 설명
① 비례성 원칙	- 위험성 검토	- 바이오 정보 사용 시 위험과 편익 검토
	- 위험성 최소화	- 위험 최소화 바이오 정보 사용

- **수집.이용 제한의 원칙**
  - 바이오정보의 수집·이용 목적, 항목, 보유기간을 이용자에게 명확히 알리고 동의 받아야 한다.
  - 인증·식별 목적에 필요한 최소한의 바이오정보를 수집·이용해야 한다.

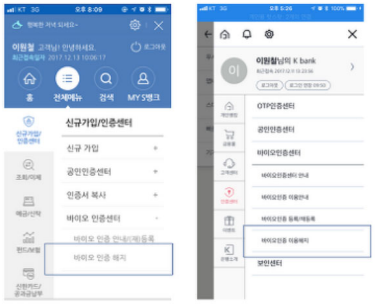
원칙	세부 원칙	원칙 설명
② 수집/이용 제한 원칙	- 수집/이용 정보 명시 및 동의	- 바이오정보 동의필요 - 목적, 항목, 보유기간
	- 특징 생성 후 파기 원칙	- 원본정보 즉시 파기 - 민감정보 추출방지

- **목적 제한의 원칙**
  - 바이오정보는 이용자에게 동의 받은 인증 또는 식별 이외의 목적으로 무단으로 활용되어서는 아니 된다.

원칙	세부 원칙	원칙 설명
③ 목적 제한 원칙	- 동의받은 내용 외 활용금지	- 이용자의 동의 외 무단 활용 금지

- **통제권 보장의 원칙**
  - 이용자가 바이오정보를 수정하거나 삭제할 수 있도록 다양한 통제 수단을 제공해야 한다.
  - 이용자가 바이오정보의 제공을 원하지 않거나 신체적 장애 등으로 제공할 수 없는 경우를 대비하여 가능한 대안을 마련하는 것이 바람직하다.

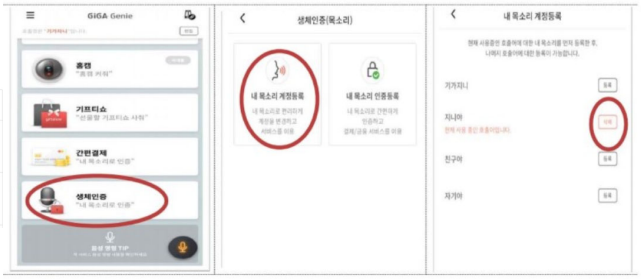
원칙	세부 원칙	원칙 설명
④ 통제권 보장 원칙	- 통제수단 제공	- 수정/삭제 가능수단 - 기기 통제권 행사
	- 대안 마련	- 미동의, 제공불가 시 다른 정보 활용



신한은행: 신한S뱅크 K뱅크: K뱅크

- **투명성 원칙**
  - 바이오정보 보호에 관한 사항을 이용자에게 적극적으로 안내해야 한다.
  - 바이오정보 서비스와 관련된 이용자의 문의 및 침해 민원 등을 처리하기 위한 피해구제 기능을 마련·운영해야 한다.

원칙	세부 원칙	원칙 설명
⑤ 투명성 원칙	- 관련내용 적극 안내	- 바이오정보 종류 - 보호조치, 행사방법
	- 이용자 문의 민원 기능	- 통제권행사 피해신고 - 처리부서, 연락처



▲ 아이폰 바이오정보 관리 방법

- **바이오정보 보호 중심설계 및 운영 원칙**
  - 바이오정보를 활용한 서비스의 개발·설계 단계부터 이용자의 바이오정보 보호를 고려하도록 권고한다.
  - 대량의 바이오정보를 서버로 전송하여 처리하는 경우, 사전에 이용자의 프라이버시에 미칠 영향 및 개인정보 위험 요인 등을 조사·분석·평가하는 절차를 마련하는 것이 바람직하다.

원칙	세부 원칙	원칙 설명
⑥ 바이오정보 보호중심 설계 및 운영원칙	- 설계단계부터 정보보호 고려	- default값 보호 설정 - 특징정보 암호화
	- 프라이버시고려 - 위험요인 조사	- PIA 개인정보영향평가 - 위험분석, 개선도출

비식별화

비식별화(De-identification)

- 비식별화란 데이터 셋에서 개인을 식별할 수 있는 요소를 전부 또는 일부삭제하거나 대체하는 등의 방

익명정보 vs 비식별화정보

익명 정보		비식별화 정보
	≠	
Anonymous Data	vs	de-identification Data
특정개인을 식별할 수 없는 형태로 정보를 수집한 자료		개인을 식별할 수 있는 상태로 수집한 정보를 비식별화 과정을 통하여 개인을 식별할 수 없게 처리한 자료

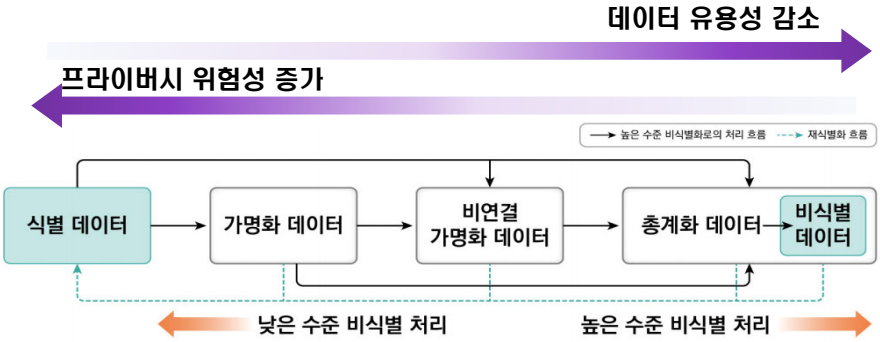
비식별화(De-identification) 판별 특성

- 비식별화란, 본질적으로 개인정보를 구성하는 세가지 요인인 개별성, 연결가능성, 추론가능성 중 일부 혹은 전부를 제거하는 과정

➢ 개별성(Single out) : 특정 정보가 특정 개인과 일대일 대응 정도
➢ 연결가능성(Linkability) : 특정 정보와 특정 개인을 연결할 수 있는 정도
➢ 추론가능성(Inference) : 특정 정보로부터 특정 개인을 추론할 수 있는 정도

비식별 처리과정의 데이터 형태와 비식별 수준

- 식별 데이터 형태는 비식별 처리가 수행됨에 따라 점차적으로 가장 높은 수준인 비식별 데이터 형태로 변환



출처 : ITU-T draft Recommendation X.fdip(2017.3)

- **식별 데이터**
  - 식별 데이터 형태에서는 데이터에 포함된 정보가 개인의 것이라는 것을 관찰 가능하기 때문에 데이터가 특정 개인과 명확하게 연관될 수 있다
- **가명화 데이터**
  - 가명화된 데이터 형태에서는 모든 식별자가 다른 값으로 대체되기 때문에 대체 처리를 수행한 당사자가 아닌 사람은 특정인과 연결될 수 있는 원래의 데이터를 알 수 없다
- **비연결 가명화 데이터**
  - 비연결 가명화 데이터 형태에서는 모든 식별자를 지우거나 혹은 가명화를 위한 대체 방법도 유지하지 않기 때문에 비식별 처리를 수행한 당사자도 비식별 처리 이전의 원래 데이터로 복구가 불가능하다

- **총계화 데이터**
  - 총계화 데이터 형태에서는 특정 개인을 식별할 수 있는 값들을 포함하지 않도록 서로 다른 사람에 대한 정보를 구성한다. 총계화 방법을 통해 형성된 데이터는 특정 값을 통해서 식별할 수 있는 사람들의 수(예, k-익명화의 k값 등)를 설정하고 그 수 미만으로 데이터를 형성하여 특정 사람을 식별할 수 없도록 한다
- **비식별 데이터**
  - 비식별 데이터는 특정인에 해당하는 데이터 값을 변경하여 직·간접적으로 다른 데이터와 결합이 불가능한 형태로써 데이터 자체 혹은 다른 데이터와 결합을 통해서도 재식별이 어려운 형태이다

비식별 처리를 위한 사전 단계 : 식별자 구분

- **데이터셋에서 정보를 표현하는 최소 단위를 속성이라 하는데, 특정 개인을 식별하게 하는 것을 개인 식별자 속성이라 함 → 식별자, 준식별자**

- **식별자 (Identifiers)**
  - 개인을 식별할 수 있는 속성들 (1:1 대응이 가능한 모든 정보)
    - ✓ 주민번호, 전화번호, 이메일, 이름, 구글 ID, 계좌번호, 유전자 정보 등
    - ✓ 암호화된 값도 식별자로 분류됨.
  - 비식별 조치시 가능한 무조건 “삭제”

- **준식별자 (QI : Quasi-Identifiers)**
  - 자체로는 식별자가 아니지만, 다른 데이터와 결합을 통해 특정 개인을 간접적으로 추론하는데 사용될 수 있는 속성들
    - ✓ 거주 도시명, 몸무게, 혈액형 등
  - 비식별 처리를 통한 변형/조작의 대상이 됨.
- **민감정보 (SA : Sensitive Attributes)**
  - 개인의 사생활을 드러낼 수 있는 속성
    - ✓ 병명, 예금 잔고, 카드 결제 액, 종교, 소속 정당 등
  - 데이터 분석시 주로 측정되는 대상 속성으로, 저장 시 비식별 처리로 데이터 처리

➢ **“지리산에 오직 한 명의 해녀가 산다”는 그 해녀가 누구인지를 유일하게 특정할 수 있으므로 개인 식별자 속성을 지닌다**

비식별 처리기법

처리기법	조치전	비식별조치후
가명처리	홍길동, 35세, 서울 거주, 한국대 재학	임꺽정, 30대, 서울 거주, 국제대 재학
총계처리	임꺽정 180cm, 홍길동 170cm, 이콩쥐 160cm, 김팥쥐 150cm	물리학과 학생키 합 660cm, 평균키 165cm
데이터삭제	주민등록번호 901206-1234567	90년대생, 남자
데이터범주화	홍길동, 35세	홍씨, 30~40세
데이터 마스킹	홍길동, 35세, 서울 거주, 한국대 재학	홍OO, 35세, 서울 거주, OO대학 재학

처리기법	내용
가명처리 (Pseudonymisation)	▪ 개인 식별이 가능한 데이터에 대하여 직접적으로 식별할 수 없는 다른 값으로 대체
총계처리(Aggregation)	▪ 개인정보에 대하여 통계값(전체 혹은 부분)을 적용하여 특정 개인을 판단할 수 없도록 함
데이터 값 삭제 (Data Reduction)	▪ 개인정보 식별이 가능한 특정 데이터 값 삭제
데이터 범주화 (Data Suppression)	▪ 단일 식별 정보를 해당 그룹의 대표값으로 변환(범주화)하거나 구간 값으로 변환(범위화) 하여 고유 정보 추적 및 식별 방지
데이터 마스킹 (Data Masking)	▪ 개인 식별 정보에 대하여 전체 또는 부분적으로 대체값 (공백, '*', 노이즈 등)으로 변환

- **가명처리**
  - 개인정보 중 주요 식별요소를 다른 값으로 대체하여 개인식별을 곤란하게 함
    - ✓ 홍길동, 35세, 서울 거주, 한국대 재학 → 임꺽정, 30대 서울 거주, 국제대 재학
  - (휴리스틱 가명화, heuristic pseudonymization) 데이터를 정해진 규칙으로 가명처리하여 실제 누구 데이터인지 알 수 없게 하는 기술
  - (암호화, encryption) 암호화 알고리즘을 기반으로 개인정보를 암호화하여 숨기는 기술
  - (교환 방법, swapping) 민감한 데이터를 사전에 정해진 외부 데이터로 치환하는 기술

- **총계처리 또는 평균값 대체**
  - 데이터의 총합 값을 보임으로서 개별 데이터의 값을 보이지 않도록 함
    - ✓ 임꺽정 180cm, 홍길동 170cm / 이콩쥐 160cm, 김팥쥐 150cm → 물리학과 학생 키 합 : 660cm, 평균키 165cm
  - (총계처리 기본방식, aggregation) 데이터의 총합이나 평균으로 개인의 실제 정보를 숨기는 기술
  - (부분총계, micro aggregation) 다른 속성 값에 비해 오차 범위가 크거나 특징적인 경우 해당 속성 값에 대해서만 통계 값을 적용하여 개인을 식별하지 못하게 하는 기술
  - (라운딩, rounding) 올림, 내림, 반올림 등의 방법을 사용하여 개인의 실제 정보를 숨기는 기술
  - (재배열, rearrangement) 그룹 내 데이터를 임의로 섞어 특정 데이터와 개인 간 연결성을 끊는 기술

- **데이터 값(가치) 삭제**
  - 데이터 공유. 개방 목적에 따라 데이터 셋에 구성된 값 중에 필요없는 값 또는 개인식별 값 삭제
    - ✓ 홍길동, 35세, 서울 거주, 한국대 졸업 → 35세, 서울 거주
    - ✓ 주민등록번호 901206-1234567 → 90년대 생, 남자
    - ✓ 개인과 관련된 날짜 정보(자격 취득일자, 합격일 등)는 연 단위로 처리
  - (속성값 삭제) 속성값을 완전하게 삭제하는 기술
  - (속성값 부분삭제, reducing partial variables) 속성의 일부 값을 삭제하여 대표성을 가진 값으로 보이게 하는 기술
  - (레코드 삭제) 대표성을 가진 값을 삭제하는 기술
  - (식별요소 전부삭제) 식별가능한 요소를 완전하게 삭제하는 기술

- **데이터 범주화**
  - 데이터의 값을 범주의 값으로 변환하여 명확한 값을 감춤
    - ✓ 홍길동, 35세 → 홍씨, 30-40세
  - (감추기) 데이터의 평균 또는 범주값으로 변환해 일반화하는 기술
  - (랜덤 라운딩, random rounding) 임의의 값을 기준으로 해당 값을 올리거나 내려 민감성이 높은 정보를 대표값으로 처리하는 기술
  - (범위 방법, data range) 개인 수치데이터를 범위나 구간으로 표현
  - (제어 라운딩, controlled rounding) 행과 열의 합이 일치되도록 고려하여 값을 라운딩(rounding)하는 기술

- **데이터 마스킹**
  - 공개된 정보 등과 결합하여 개인을 식별하는데 기여할 확률이 높은 주요 개인식별자가 보이지 않도록 처리하여 개인을 식별하지 못하도록 함
    - ✓ 홍길동, 35세, 서울 거주, 한국대 재학 → 홍**, 35세, 서울 거주, **대학 재학
  - (임의 잡음 추가 방법, adding random noise) 임의의 노이즈(random noise) 값을 넣어 식별정보 노출을 방지하는 기술
  - (공백과 대체, blank and impute) 속성 값 일부를 공백처리하고 특수문자 등으로 채우는 기술 등

비식별 처리 정보의 활용성 판단 지표 ⇒ 원본 유사도

- **원본 유사도는 비식별 데이터셋의 활용성을 나타내는 지표**
- **원본 데이터셋과 이를 비식별 처리한 비식별 데이터셋이 얼마나 유사한지를 나타내는 지표**

- **레코드 잔존도과 레코드 유사도로 측정 → 잔존도과 유사도가 높으면 활용성이 높은 것으로 판단**

- **① 레코드 잔존도**
  - 원본 데이터셋의 총 레코드 수 대비 비식별 데이터셋의 총 레코드 수를 나타낸 지표
  - 비식별 처리 과정에서 원본 데이터셋에서 삭제되지 않고 비식별 데이터셋에 남은 레코드들의 비율
  - 예를 들면 그림 (a)의 원본 데이터셋에 대한 그림 (c)의 비식별 데이터셋의 잔존율은 5/8 = 0.65

구성	성별	이름	연령
1	남	이지언	39
2	여	김영희	35
3	여	이지언	35
4	여	임순희	35
5	여	김영희	35
6	여	이지언	39
7	여	김영희	35
8	여	이지언	39

(a) 원본 데이터셋 →

구성	성별	이름	연령
1	남	이**	39
2	여	김**	35
3	여	이**	35
4	여	임**	35
5	여	김**	35
6	여	이**	39
7	여	김**	35
8	여	이**	39

(b) 이름 비식별화 →

구성	성별	이름	연령
2	여	김**	35
5	여	김**	35
6	여	이**	39
7	여	김**	35
8	여	이**	39

(c) 유일한 속성값 조합 레코드 삭제 후

➢ **"지리산에 오직 한 명의 해녀가 산다"는 그 해녀가 누구인지를 유일하게 특정할 수 있으므로 개인 식별자 속성을 지닌다**

- **② 레코드 유사도**
  - 원본 레코드와 비식별 레코드 쌍 간의 통계적 유사성을 0과 1 사이의 값으로 표현한 지표
  - 속성의 유형(수치형, 명목형)에 따라 두 레코드의 속성값 유사도를 먼저 계산

$$(\text{레코드 유사도}) = \frac{\sum(\text{속성 유사도})}{\text{속성 수}} = \frac{\text{성별 속성 유사도} + \text{수입 속성 유사도} + \text{나이 속성 유사도}}{3}$$

– 수치형 속성값 유사도

✓ 수치형 속성의 경우 속성 도메인 크기 대비 원본 레코드 속성값과 비식별 레코드 속성값의 차이 비율로 정의

$$(A4,X4)\text{쌍의 수입 속성 유사도} = 1-\frac{|2100-2133|}{Range(\text{수입})} = 1-\frac{|2100-2133|}{\max(\text{수입})-\min(\text{수입})} = 1-\frac{|2100-2133}{3000-1400} = 1-\frac{33}{1600} = 0.9793$$

원본ID	성별	수입	나이
A1	여	1400	23
A2	남	1700	32
A3	여	2900	43
A4	남	2100	25
A5	여	3000	40
A6	여	1900	28

결과ID	성별	수입	나이
X1	*	1532	20대
X2	*	1697	30대
X3	*	2835	40대
X4	*	2133	20대
X5	*	3013	40대
X6	*	1858	20대

– 명목형 속성값 유사도

✓ 명목형 속성의 경우 원본 데이터셋에서 해당 속성의 유일한 속성값 개수 대비 비식별 데이터셋에서 해당 속성의 유일한 속성값 개수의 비율로 정의

✓ 예를 들면 “연령”에 대해 원본 데이터셋에서 연령의 유일한 속성값이 총 30개(20세-49세)라고 하면 비식별 데이터셋에서 연령의 유일한 속성값이 3개(20대, 30대, 40대)

✓ 20대는 총 10개의 서로 다른 나이값으로 | Research | 익명화 데이터의 익명 결합 방법 표현하므로 (A4,X4)쌍의 연령 속성값 유사도는 다음과 같이 계산

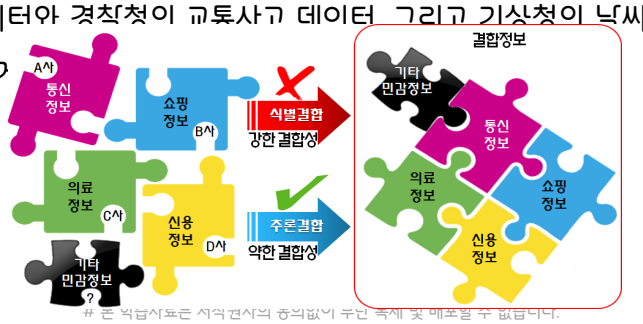
원본ID	성별	수입	나이
A1	여	1400	23
A2	남	1700	32
A3	여	2900	43
A4	남	2100	25
A5	여	3000	40
A6	여	1900	28

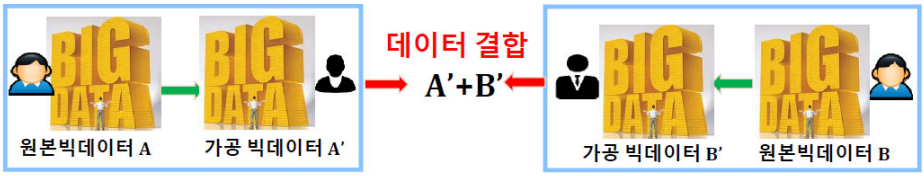
결과ID	성별	수입	나이
X1	*	1532	20대
X2	*	1697	30대
X3	*	2835	40대
X4	*	2133	20대
X5	*	3013	40대
X6	*	1858	20대

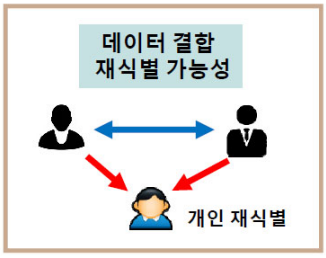
$$(A4,X4)\text{쌍의 나이 속성 유사도} = 1-\frac{\text{비식별화 결과의 원소 } count \text{ } - \text{ } 1}{\text{원본도메인의 } distinct \text{ } count} = 1-\frac{10-1}{30} = 0.7$$

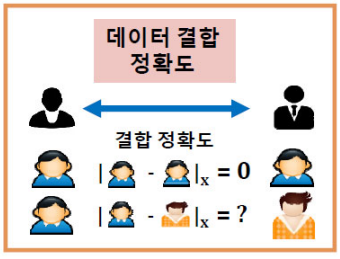
비식별 정보의 결합 ⇒ 빅데이터

- 빅데이터를 사용하는 가장 큰 장점은 서로 다른 영역의 빅데이터들을 결합하여 여러 영역의 거시적 현상을 세밀하게 분석할 수 있다는 점
  - 도로교통 상황 예측을 위해 빅데이터를 활용하고자 할 때, 한국도로공사의 교통소통데이터와 경찰청의 교통사고 데이터, 그리고 기상청의 날씨 데이터를 결합하면 보다 정확하게









데이터 결합 모델

- **원본 결합**
  - 개인식별자 기반 데이터 결합 (예: 주민등록번호)
- **가명 결합**
  - 가명식별자 기반 데이터 결합 (예: 임시대체키)
- **익명 결합**
  - 익명식별자 기반 데이터 결합

개인식별자 기반 원본결합

- **개인식별자란 주민번호나 전화번호와 같이 개인별로 1개의 유일한 값으로 정해지는 속성**

- **두 원본 데이터셋에 동일한 개인식별자가 있을 때 두 원본 데이터 셋에서 동일한 개인식별자를 갖는 두 레코드 쌍을 결합하는 작업**
  - 두 원본 데이터셋 A와 B에 모두 '주민번호' 속성이 있을 때 'A.주민번호=B.주민번호'를 만족하는 A와 B의 두 레코드 쌍을 각각 결합

원본 데이터 D1

개인식별자1	이름
790812-	홍길동
860923-	이순신
700723-	허준
861121-	김유신
890512-	이성계

원본 결합

원본 데이터 D2

개인식별자2	이름
790812-	홍길동
860923-	이순신
700723-	허준
861121-	김유신
651111-	최영수

개인식별자1	개인식별자2	이름
790812-	790812-	홍
860923-	860923-	이
700723-	700723-	허
861121-	861121-	김

개인식별자1=개인식별자2

결합 결과 데이터

가명식별자 기반 가명결합

- **가명식별자 생성**
  - 가명식별자란 개인정보를 가명처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 속성을 의미

원본 개인식별자 속성 | 개인정보 원본

	주민등록번호	진단명	검사수치
1	620806-1234567	위암	24.5
2	630123-1234567	위암	53.4
3	820214-1411254	간암	25.3
4	781229-1234580	폐암	95.2
5	890605-1234566	간암	53.1
6	780403-1423948	폐암	20.2

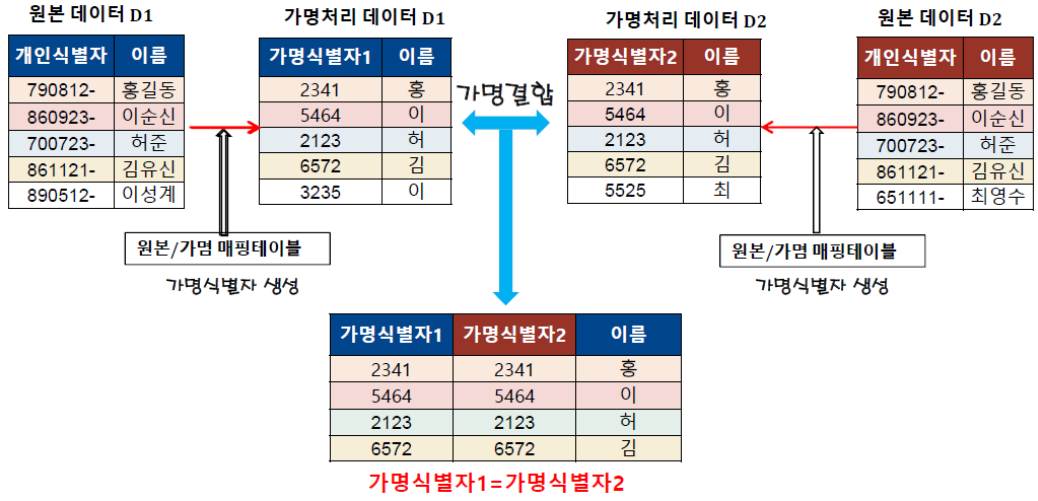
<질병 현황 개인정보 원본 데이터셋>

가명식별자 속성 | 개인정보 원본

	임시대체키	진단명	검사수치
1	586	위암	24.5
2	572	위암	53.4
3	804	간암	25.3
4	513	폐암	95.2
5	803	간암	53.1
6	511	폐암	20.2

<질병 현황 익명 데이터셋>

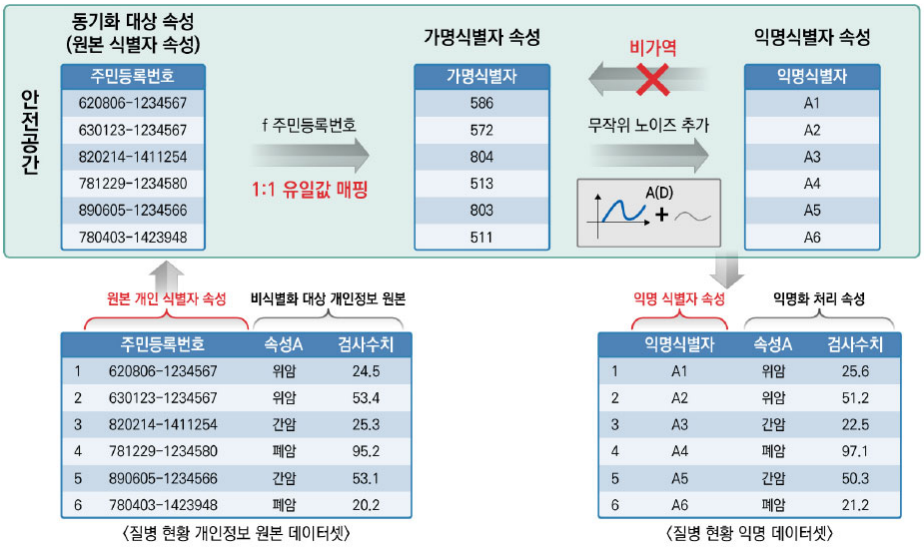
- 가



익명식별자 기반 익명결합

- 원본에 있는 개인별로 개인식별자에 1:1 대응되는 가명식별자를 생성
- 개인별 가명식별자에 무작위(Random)한 노이즈를 추가하여 해당 개인의 익명식별자를 생성
- 개인별로 익명식별자를 생성할 때 무작위 노이즈가 추가되므로 익명식별자값으로 자신의 원래 가명식별자 값을 복원하는 것은 불가능
  - 결과적으로 익명식별자를 기반으로 원본의 개인을 재식별하는 것도 불가능
  - 또한 한 개인의 익명식별자를 새로 생성할 때마다 추가되는 무작위 노이즈 값이 다르므로 한 개인에 대해 많은 수의 서로 다른 익명식별자들을 생성 가능

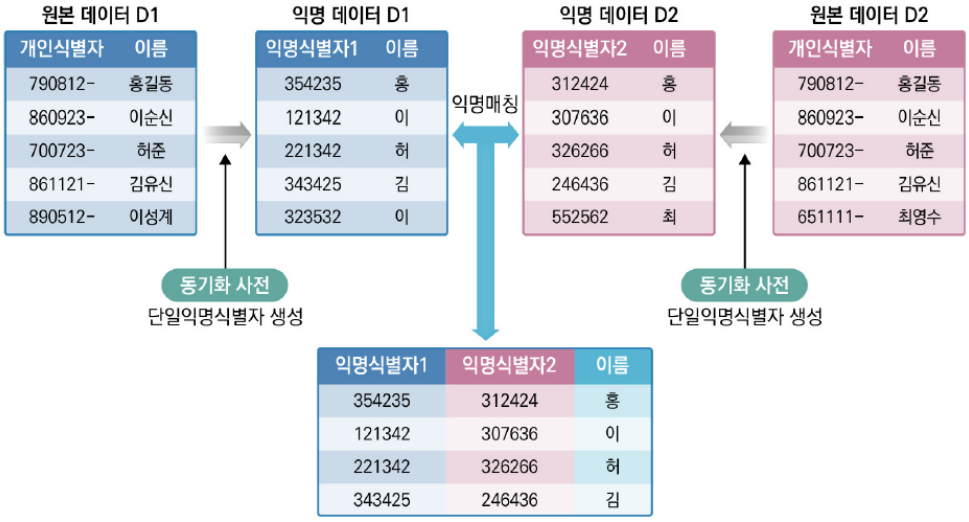
▪ 익명



〈질병 현황 개인정보 원본 데이터셋〉

〈질병 현황 익명 데이터셋〉

▪ 익명



감사합니다.