개인정보보호론

[12주차. 비식별화]

비식별화

비식별화(De-identification)

- 비식별화란 데이터 셋에서 개인을 식별할 수 있는 요소를 전부 또는 일부삭제하거나 대체하는 등의 방법을 활용, 개인을 알아볼 수 없도록 하는 과정

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

- Anonymous Data : 정보의 수집단계에서 근원적으로 개인을 식별할 수 없는 형태로 수집한 정보
- de-identification Data : 개인을 식별할 수 있는 상태에서 수집한 정보를 비식별화 과정을 통하여 개인을 식별할 수 없게 처리한 정보

➢ 익명정보와 비식별화정보의 차이는 개인정보수집자가 정보주체를 인지할 수 있는 식별자를 포함하여 수집한 정보인가 이다. 여기서 익명 정보와 비식별화 정보는 표면적으로는 개인을 식별할 수 없으나, 비식별화는 재식별화의 위험성이 내포되어 있고, 개인을 작은 단위로 그룹화하여 개인을 관리하는 것이 가능하다

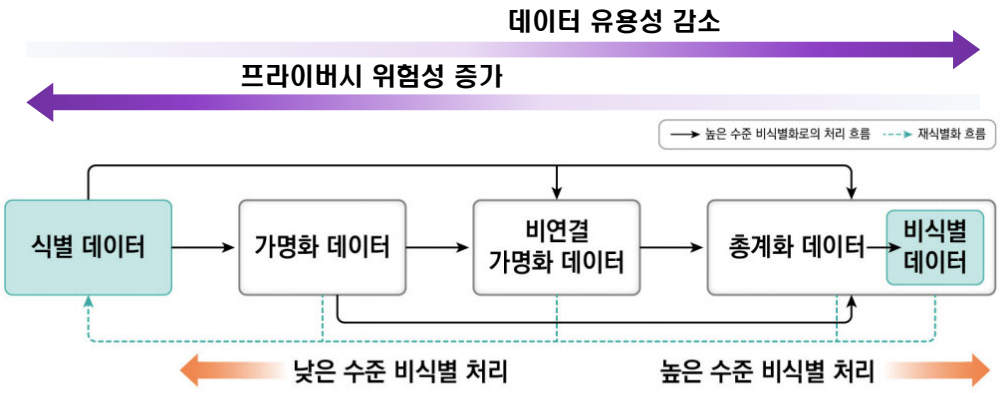
비식별화(De-identification) 판별 특성

- 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 구간값으로 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 개인식별에 중요한 값을 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
    - ✓ 연예인. 정치인 등의 가족 정보(관계정보), 판례 및 보도 등에 따라 공개되어 있는 사건과 관련되어 있음을 알 수 있는 정보
  - (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
    - ✓ 예를 들면 (a)의 원본 레코드 A4에 대응되는 (b)의 비식별 레코드 X4의 (A4,X4) 쌍의 "수입"속성값 유사도를 다음와 같이 계산

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

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 ㅣResearchㅣ 익명화 데이터의 익명 결합 방법 표현하므로 (A4,X4)쌍의 연령 속성값 유사도는 다음과 같이 계산

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

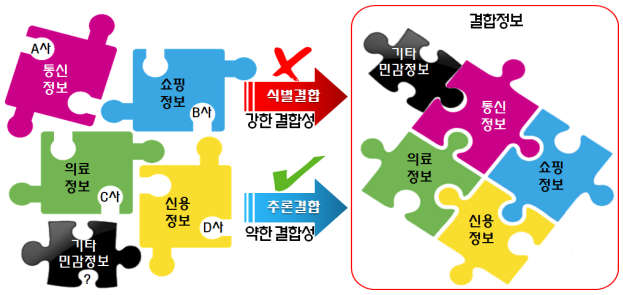
원본ID	성별	수입	나이
A1	여	1400	23
A2	남	1700	32
A3	여	2900	43
A4	남	2100	25
A5	여	3000	40
A6	여	1900	28

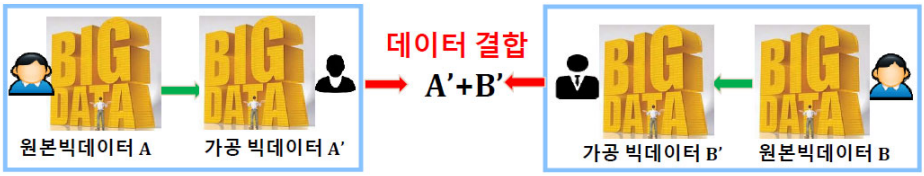
결과ID	성별	수입	나이
X1	*	1532	20대
X2	*	1697	30대
X3	*	2835	40대
X4	*	2133	20대
X5	*	3013	40대
X6	*	1858	20대

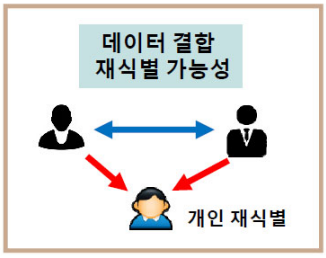
데이터 결합

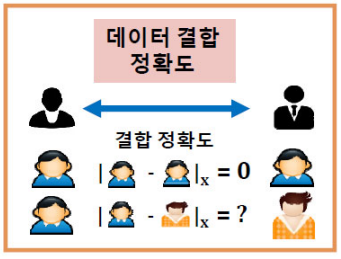
비식별 정보의 결합 ⇒ 빅데이터

- 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 교통상황을 예측할 수 있음









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 데이터셋에서 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

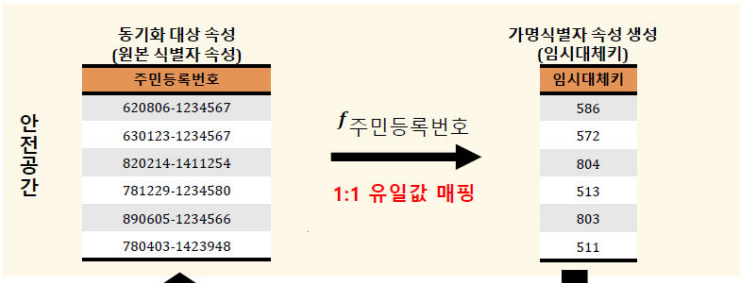
개인식별자1	개인식별자2	이름
790812-	790812-	홍
860923-	860923-	이
700723-	700723-	허
861121-	861121-	김

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

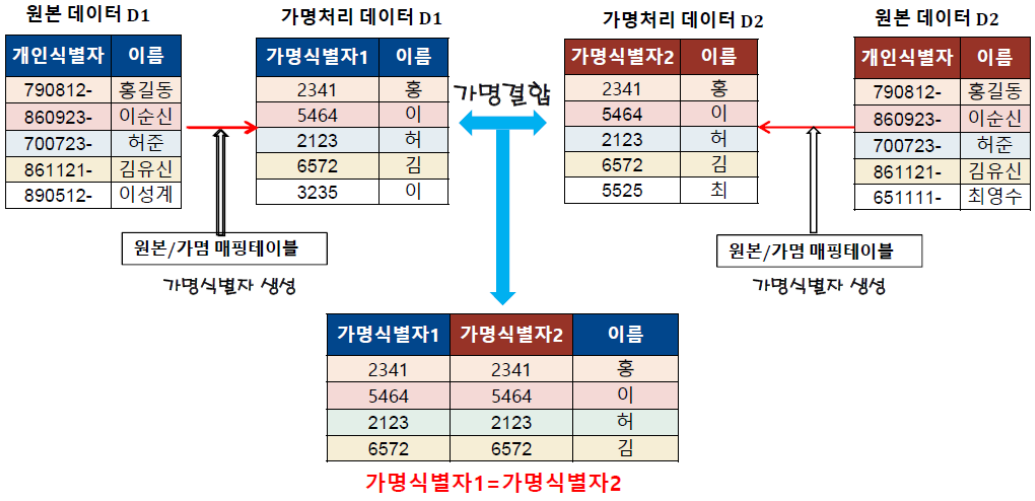
<질병 현황 개인정보 원본 데이터셋>

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>

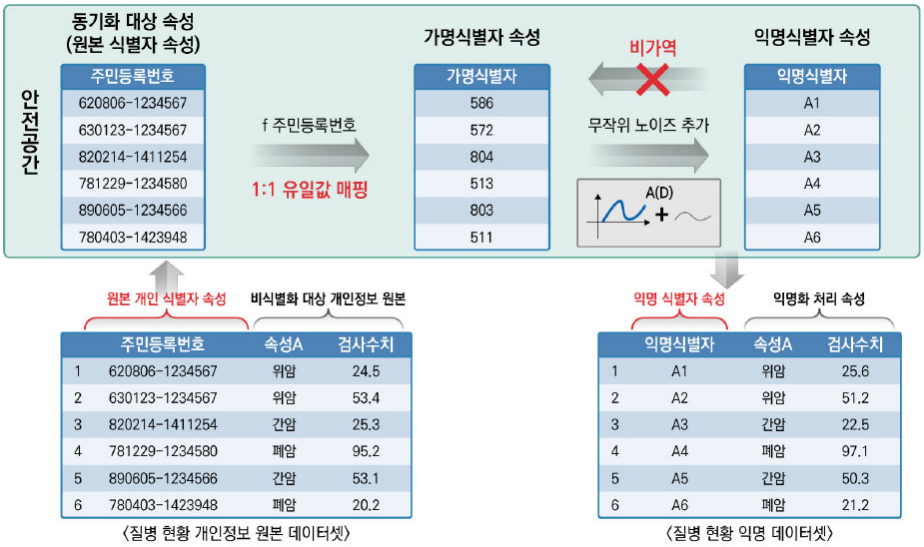
- **가명 결합 : 동일 개인의 가명식별자간 결합**



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**

▪ 익명식별자 생성방법



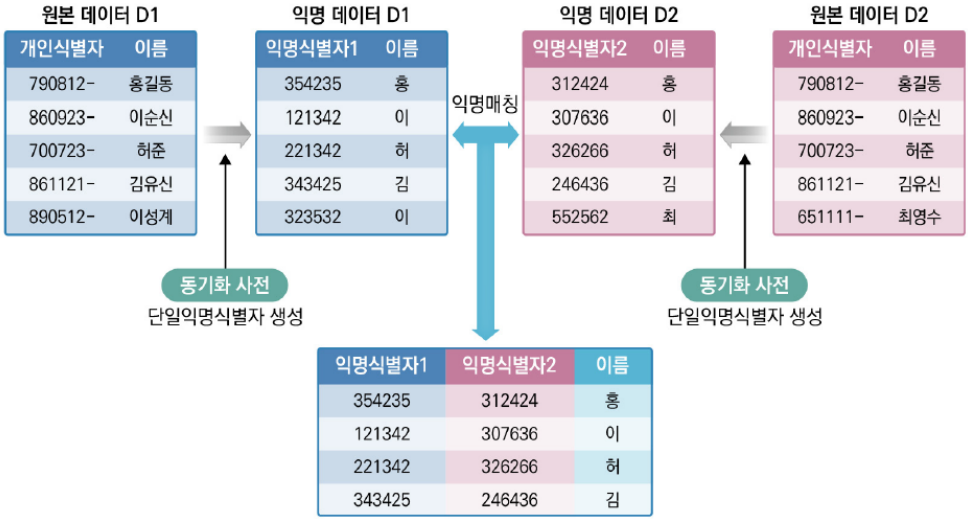
	주민등록번호	속성A	검사수치
1	620806-1234567	위암	24.5
2	630123-1234567	위암	53.4
3	820214-1411254	간암	25.3
4	781229-1234580	폐암	95.2
5	890605-1234566	간암	53.1
6	780403-1423948	폐암	20.2

〈질병 현황 개인정보 원본 데이터셋〉

	익명식별자	속성A	검사수치
1	A1	위암	25.6
2	A2	위암	51.2
3	A3	간암	22.5
4	A4	폐암	97.1
5	A5	간암	50.3
6	A6	폐암	21.2

〈질병 현황 익명 데이터셋〉

▪ 익명결합



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

익명식별자1	이름
354235	홍
121342	이
221342	허
343425	김
323532	이

익명식별자2	이름
312424	홍
307636	이
326266	허
246436	김
552562	최

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

익명식별자1	익명식별자2	이름
354235	312424	홍
121342	307636	이
221342	326266	허
343425	246436	김