

K-ICT 빅데이터센터  
개인정보 가명·익명조치기술 전문교육

NIA 한국정보화진흥원

# 개인정보 가명·익명 처리 기술

1일차 | 기본4

김 승 환

인하대학교 데이터 사이언스학과

[swkim4610@inha.ac.kr](mailto:swkim4610@inha.ac.kr)

이 노트는 라곰 소프트 오형섭 이사 강의자료를  
편집하였음을 밝혀 둡니다.

**PART 1. 개인정보 가명처리 기법**

**PART 2. 개인정보 가명처리 기술**

**PART 2. 프라이버시 보호 모델**

## PART 2

# 개인정보 가명처리 기술



## 2. 개인정보 가명처리 기술

### 1. 삭제기술

#### 마스킹

- 정보의 일부 혹은 전부를 다른 글자로 대체하여 식별하지 못하도록 하는 기법

대체

특정 문자로 변경

스크램블링

기존 문자열을 **순서만 바꿔서 변경**

암호화

암호화된 형태로 변경, 일반적으로 **형태보존 암호화 기법** 사용

데이터 블러링

숫자 값에 대해 **노이즈를 추가한 값으로 변경**

삭제

특정 값을 **NULL값**으로 변경(데이터를 완전히 삭제)

## 2. 개인정보 가명처리 기술

### 1. 삭제기술

#### 마스킹

- 정보의 일부 혹은 전부를 다른 글자로 대체하여 식별하지 못하도록 하는 기법

대체

홍길동 → 홍\*\*

스크램블링

홍길동 → 동홍길만 바뀌서 변경

암호화

4121-0314-1345-6745 → 1234-4567-1235-0978 사용

데이터 블러링

10 → 20, 5 → 10 (\* 원본 데이터에 2를 곱해줌)

삭제

홍길동 → ' ' 값으로 변경(데이터를 완전히 삭제)

## 2. 개인정보 가명처리 기술

### 1. 삭제기술

컬럼 삭제	<ul style="list-style-type: none"><li>▪ <u>직접 식별자나 식별 가능성이 높은 간접 식별자</u>, 중복된 정보 등을 삭제하는 것으로 일반적으로 대상 컬럼을 삭제</li></ul>
부분 삭제	<ul style="list-style-type: none"><li>▪ <u>컬럼의 일부를 삭제</u>하여 데이터의 식별성을 낮추는 기법</li><li>▪ 예) 주소 일부, 날짜 일부 삭제</li></ul>
레코드 삭제	<ul style="list-style-type: none"><li>▪ 식별성이 높은 레코드에 대해 해당 레코드를 삭제하는 방법</li><li>▪ <u>이상값에 해당하는 데이터</u>는 일반적으로 분석 결과에 나쁜 영향을 줄 수 있으며 식별성도 매우 높아짐</li></ul>
식별자 전부 삭제	<ul style="list-style-type: none"><li>▪ <u>식별성이 있는 요소를 전부 삭제</u>하는 방법</li><li>▪ 실제 사용되는 사례는 HIPAA Privacy Rule의 Safe Harbor 방식이 대표적인 사례임</li></ul>

## 2. 개인정보 가명처리 기술

### 1. 삭제기술

#### 컬럼 삭제

#### 부분 삭제

#### 레코드 삭제

#### 식별자 전부 삭제

일련번호	이름	성별	생년	주소	나이	수신 총 잔액	신용대출 한도	신용 등급
10010785	조미선	F	1985	대전 동구 용운동	33	817,250	66,300,000	3
10011953	홍길병	M	1957	경북 안동시 용상동	61	4,559,120	327,700,000	2
10012231	김영심	F	1968	경남 진주시 옥봉동	50	13,601,564	41,300,000	3
10012598	이미정	F	1948	서울 강서구 가양3동	70	979,118	64,600,000	7
10013649	김경태	M	1978	서울 은평구 역촌1동	40	5,501,809	2,300,000	5
10014221	유영근	M	1975	경기 고양시 고양동	43	609,622	13,900,000	7
10015665	박을규	M	1995	경기 수원시 고색동	23	3,885,329	37,700,000	2
10016386	문정은	F	1951	경기 고양시 일산4동	67	23,992,801	3,500,000	3
10016675	오한근	M	2010	경기 고양시 성사동	7	185,878,354	0	1
10017321	전태홍	M	1971	서울 금천구 독산1동	47	274,489	17,600,000	5
10017383	이현주	F	1943	경기 평택시 합정동	75	7,185,105	3,200,000	6
10018757	백지연	M	1939	서울 은평구 증산동	79	1,606,685	436,800,000	3
10018880	민영기	M	1973	강원 춘천시 근화동	45	868,878	34,900,000	4
10019912	김수복	F	1946	전남 광양시 진상면	72	5,260,714	3,500,000	3
10022529	엄경아	F	1957	서울 성북구 보문동	61	24,375,307	16,000,000	2

상적으로 대상 컬럼을 삭제

도 매우 높아짐

2. 개인정보 가명처리 기술

1. 삭제기술

컬럼 삭제

부분 삭제

레코드 삭제

식별자 전부 삭제

성별	주소	나이	수신 총 잔액	신용대출 한도	신용 등급
F	대전 동구 용운동	33	817,250	66,300,000	3
M	경북 안동시 용상동	61	4,559,120	327,700,000	2
F	경남 진주시 옥봉동	50	13,601,564	41,300,000	3
F	서울 강서구 가양3동	70	979,118	64,600,000	7
M	서울 은평구 역촌1동	40	5,501,809	2,300,000	5
M	경기 고양시 고양동	43	609,622	13,900,000	7

식별성이 있는 요소를 전부 삭제하는 방법

실제 사용되는 사례는 HIPAA Privacy Rule의 Safe

성별	주소	나이	수신 총 잔액	신용대출한도	신용등급
F	대전	33	817,250	66,300,000	3
M	경북	61	4,559,120	327,700,000	2
F	경남	50	13,601,564	41,300,000	3
F	서울	70	979,118	64,600,000	7
M	서울	40	5,501,809	2,300,000	5
M	경기	43	609,622	13,900,000	7

을 삭제하는 것으로 일반적으로 대상 컬럼을 삭제

을 줄 수 있으며 식별성도 매우 높아짐



2. 개인정보 가명처리 기술

1. 삭제기술

컬럼 삭제

부분 삭제

레코드 삭제

식별자 전부 삭제

성별	주소	나이	수신 총 잔액	신용대출한도	신용등급
F	대전	33	817,250	66,300,000	3
M	경북	61	4,559,120	327,700,000	2
F	경남	50	13,601,564	41,300,000	3
F	서울	70	979,118	64,600,000	7
M	서울	40	5,501,809	2,300,000	10
M	경기	43	609,622	13,900,000	7
M	경기	23	3,885,329	37,700,000	7
F	경기	67	23,992,801	3,500,000	8
M	경기	7	8,185,878,354	0	9
M	서울	47	274,489	17,600,000	8
F	경기	75	7,185,105	3,200,000	6
M	서울	79	1,606,685	436,800,000	3
M	강원	45	868,878	34,900,000	4
F	전남	72	5,260,714	3,500,000	8
F	서울	61	24,375,307	16,000,000	4

로 일반적으로 대상 컬럼을 삭제

식별성도 매우 높아짐

임

## 2. 개인정보 가명처리 기술

### 1. 삭제기술

컬럼 삭제

직접 식별자나 식별 가능성이 높은 간접 식별자, 중복된 정보 등을 삭제하는 것으로 일반적으로 대상 컬럼을 삭제

부분 삭제

레코드 삭제

식별자 전부 삭제

HIPAA Privacy Rule의 Safe Harbor

①이름	⑦사회보장번호	⑬각종 장비 식별번호
②주소 정보*	⑧의료기록번호	⑭인터넷 주소(URL 정보)
③날짜 정보*	⑨건강보험번호	⑮IP주소
④전화번호	⑩계좌번호	⑯생체정보(지문, 음성 등)
⑤팩스번호	⑪자격취득번호	⑰전체 얼굴사진 및 유사 이미지
⑥이메일주소	⑫자동차번호 (차량식별번호, 등록번호 등)	⑱기타 특이한 식별번호 또는 코드

■ 이상값에 해당하는 데이터는 일반적으로 분석 결과에 나

- 식별성이 있는 요소를 전부 삭제하는 방법
- 실제 사용되는 사례는 HIPAA Privacy Rule의 Safe Harbor

000병원의 HIPAA 18 PHI를 참고로 21개 개인건강정보 정의

No	개인식별정보
1	이름
2	읍/면/동 이하 상세 주소
3	전화번호 일체(Fax번호 포함)
4	이메일주소
5	주민등록번호
6	외국인등록번호
7	여권번호
8	건강보험증번호
9	은행계좌번호
10	신용카드번호
11	자격증번호/면허번호/학번
12	차량번호
13	환자등록번호
14	회원ID(홈페이지,ARC 등)
15	사번
16	IP 주소
17	URLs
18	바이오정보 : 지문, 홍채, 정맥, 음성, 필적, 개인식별이 가능한 유전 정보 등
19	얼굴의 전판 사진 또는 이에 상응하는 이미지
20	기타 개인식별이 가능한 정보(예 : 병리번호)
21	생년월일(생년월까지 허용)

## 2. 개인정보 가명처리 기술

### 2. 총계처리

#### 총계처리

- 특정 컬럼을 **통계적으로 처리하는 기법**으로 데이터 전체 또는 부분을 집계로 처리
- 집계 방법은 일반적으로 **‘평균값, 최대값, 최소값, 최빈값, 중앙값’ 중 하나로 처리**함

#### 평균값

일반적으로 통계에 많이 사용되는 기법으로 전체의 평균값으로 대체  
이상값에 의한 전체 데이터 왜곡이 발생할 수 있음

#### 중앙값

모든 데이터 값을 일렬로 세워 정확하게 중간에 있는 값으로 대체

#### 최대값

대상 데이터 중 가장 큰 값으로 대체  
일반적인 정규 분포에서 하단의 이상값에 대해 경계값으로 처리하는 경우 최댓값으로 변환하여 처리

#### 최소값

대상 데이터 중 가장 적은 값으로 대체  
일반적인 정규 분포에서 상단의 이상값에 대해 경계값으로 처리하는 경우 최솟값으로 변환하여 처리

#### 최빈값

대상 데이터 중 **가장 많은 빈도로 나타난 값**으로 대체

## 2. 개인정보 가명처리 기술

### 2. 총계처리

#### 총계처리

- 특정 컬럼을 **통계적으로 처리하는 기법**으로 데이터 전체 또는 부분을 집계로 처리
- 집계 방법은 일반적으로 **‘평균값, 최대값, 최소값, 최빈값, 중앙값’ 중 하나로 처리**함

평균값

= AVERAGE

일단 평균값에 많이 사용되는 기법으로 전체의 평균값으로 대체  
이상값에 의한 전체 데이터 왜곡이 발생할 수 있음

중앙값

= MEDIAN

모 = MEDIAN 을 일렬로 세워 정확하게 중간에 있는 값으로 대체

최대값

= MAX

데이터 중 가장 큰 값으로 대체  
일반적인 정규 분포에서 하단의 이상값에 대해 경계값으로 처리하는 경우 최댓값으로 변환하여 처리

최소값

= MIN

데이터 중 가장 작은 값으로 대체  
일반적인 정규 분포에서 상단의 이상값에 대해 경계값으로 처리하는 경우 최솟값으로 변환하여 처리

최빈값

= MODE.SNGL

데이터 중 가장 많은 빈도로 나타난 값으로 대체

## 2. 개인정보 가명처리 기술

### 3. 일반화 기술

- 하위의 공통된 특성을 찾아 **상위 개념으로 묶는 기법**, 특정 정보를 해당 **그룹의 대표 값이나 구간 값으로 변환**하는 기법
- 특정 정보의 명확한 값을 숨길 수 있기 때문에 **감추기**라고도 함

일반 라운딩	<ul style="list-style-type: none"> <li>세세한 정보 보다는 전체 통계 정보가 필요한 경우 많이 사용되며, 범주화의 실 기법으로도 사용할 수 있음</li> <li><b>반올림, 올림, 내림</b></li> </ul>
랜덤 라운딩	<ul style="list-style-type: none"> <li>라운딩의 <b>자리수와 기준이 되는 수를 자유롭게 지정</b>할 수 있는 라운딩 기법</li> </ul>
제어 라운딩	<ul style="list-style-type: none"> <li><b>원본의 행, 열의 합과 라운딩 적용 후 행, 열의 합이 동일</b>하게 만드는 라운딩 기법</li> <li>일반적인 라운딩에서는 특정한 수를 기준으로 라운딩을 적용하지만 제어라운딩의 경우 계산에 의해 적절한 수에서 라운딩을 적용</li> </ul>
상하단 코딩	<ul style="list-style-type: none"> <li>정규분포의 특성을 가진 데이터에서 양쪽 끝에 치우친 정보는 적은 수의 분포를 가지게 되어 개인의 식별성을 가질 수 있음</li> <li>따라서 <b>적은 수의 분포를 가진 양끝단의 정보를 범주화</b>하여 개인의 식별성을 낮추는 기법</li> </ul>
로컬 일반화	<ul style="list-style-type: none"> <li><b>이상치에 해당하는 레코드에 대해서만 일반화</b>를 적용하는 기법</li> </ul>

3. 일반화 기술

- 하위의 공통된 특성을 찾아 **상위 개념으로 묶는 기법**, 특정 정보를 해당 **그룹의 대표 값이나 구간 값으로 변환**하는 기법
- 특정 정보의 명확한 값을 숨길 수 있기 때문에 **감추기**라고도 함

일반 라운딩	나이		반올림	올림	내림
	33		30	40	30
랜덤 라운딩	61		60	70	60
	50		50	50	50
	70		70	70	70
제어 라운딩	40	→	40	40	40
	43		40	50	40
상하단 코딩	23		20	30	20
	67		70	70	60
	66		70	70	60
로컬 일반화	47		50	50	40

하의 실 기법으로도 사용할 수 있음

기법

라운딩 기법

라운딩의 경우 계산에 의해 적절한 수에서 라운

분포를 가지게 되어 개인의 식별성을 가질 수 있

성을 낮추는 기법

3. 일반화 기술

- 하위의 공통된 특성을 찾아 **상위 개념으로 묶는 기법**, 특정 정보를 해당 **그룹의 대표 값이나 구간 값으로 변환**하는 기법
- 특정 정보의 명확한 값을 숨길 수 있기 때문에 **감추기**라고도 함

일반 라운딩	수신 총 잔액		자릿수 기반 랜덤 라운딩
	869,250		869,000
랜덤 라운딩	4,559,120	→	4,560,000
	13,601,564		13,600,000
	979,118		979,000
	5,501,809		5,500,000
제어 라운딩	609,622		610,000
	3,885,329		3,890,000
	23,992,801		23,990,000
	185,878,354		186,000,000
상하단 코딩	274,489		274,000
	7,185,105		7,190,000
	1,606,685		1,610,000
	868,878		869,000
로컬 일반화	5,260,714		5,260,000
	761,039		761,000
	13,595,307		13,600,000
	6,722,935		6,720,000

주화의 실 기법으로도 사용할 수 있음

기법

라운딩 기법  
제어라운딩의 경우 계산에 의해 적절한 수에서 라운

의 분포를 가지게 되어 개인의 식별성을 가질 수 있  
별성을 낮추는 기법

3. 일반화 기술

- 하위의 공통된 특성을 찾아 **상위 개념으로 묶는 기법**, 특정 정보를 해당 **그룹의 대표 값이나 구간 값으로 변환**하는 기법
- 특정 정보의 명확한 값을 숨길 수 있기 때문에 **감추기**라고도 함

일반 라운딩	나이		반올림	제어라운딩
	33		30	30
	61		60	60
랜덤 라운딩	50		50	50
	72		70	70
	43		40	40
제어 라운딩	44	→	40	50
	23		20	20
	67		70	70
상하단 코딩	68		70	70
	49		50	50
로컬 일반화	합계 510		합계 500	합계 510



## 2. 개인정보 가명처리 기술

### 3. 일반화 기술

- 하위의 공통된 특성을 찾아 **상위 개념으로 묶는 기법**, 특정 정보를 해당 **그룹의 대표 값이나 구간 값으로 변환**하는 기법
- 특정 정보의 명확한 값을 숨길 수 있기 때문에 **감추기**라고도 함

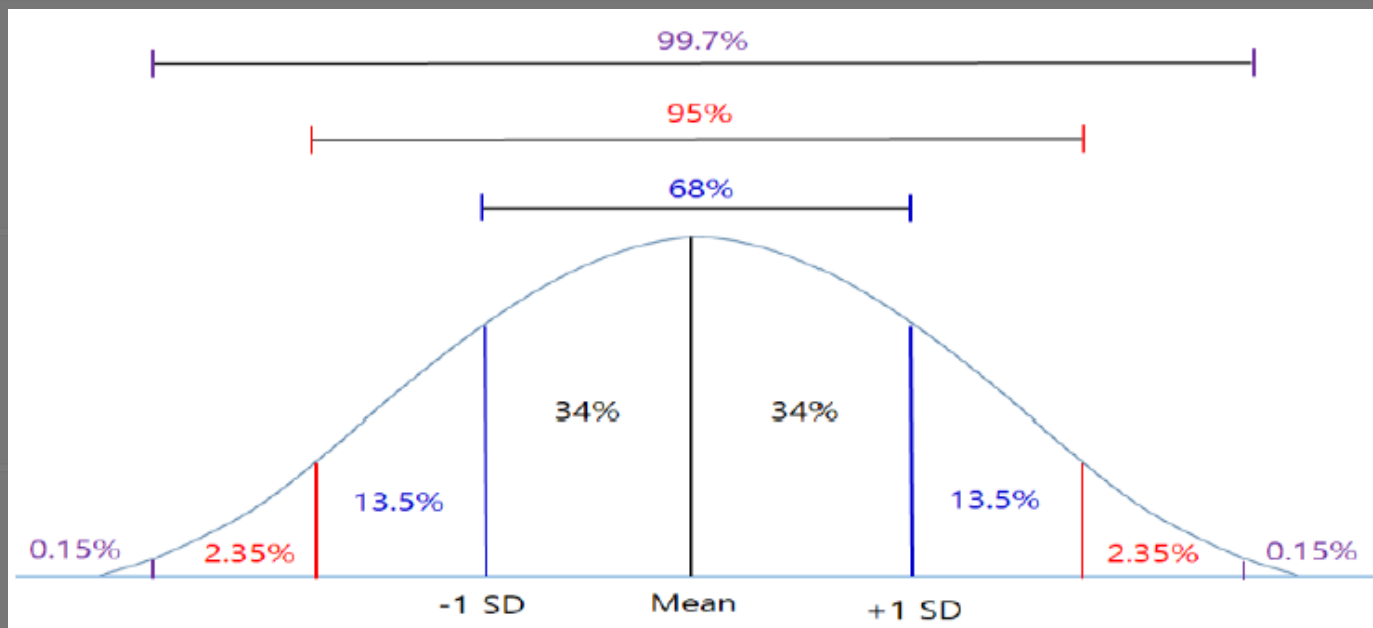
일반 라운딩

랜덤 라운딩

제어 라운딩

상하단 코딩

로컬 일반화

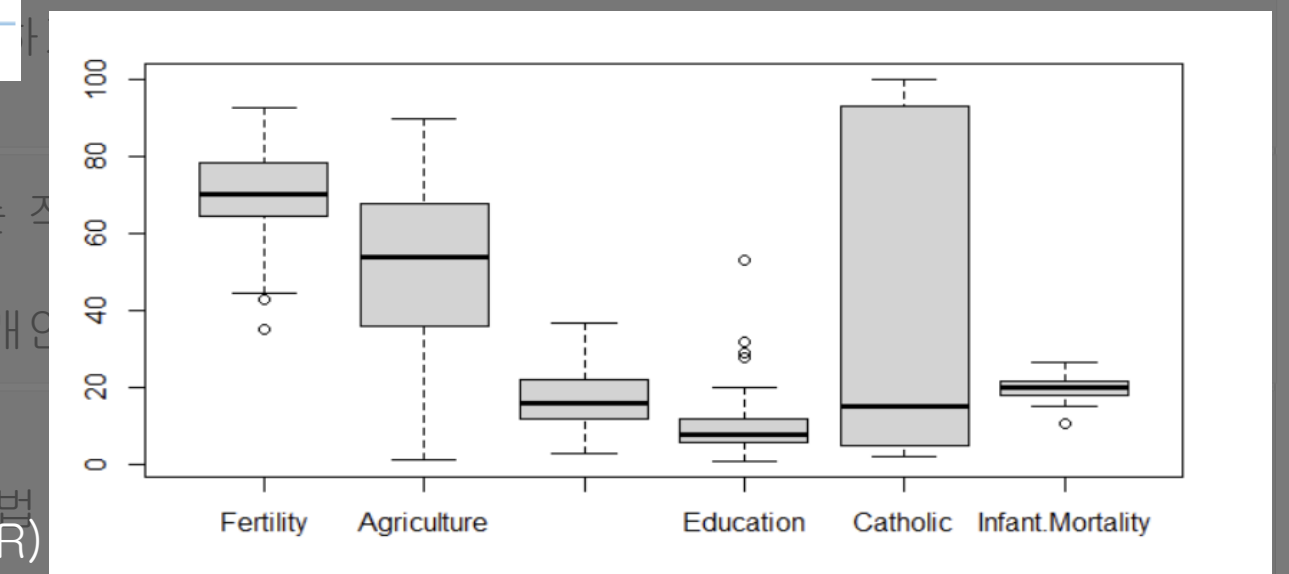


정규분포

이며, 범주화의 실 기법으로도 사용할 수 있음

라운딩 기법

만드는 라운딩 기법



## 2. 개인정보 가명처리 기술

### 4. 암호화

#### 결정성 암호화

- 동일한 알고리즘과 동일한 키로 동일한 값을 암호화 한 경우 암호화된 값이 항상 **일정한 값으로 생성되는 암호화** 기법

\* 비밀키 암호화(대칭키 방식), 공개키 암호화(비대칭키 방식)은 모두 **양방향 암호화** 방식임

#### 비밀키 암호화

암호화 할 때와 **복호화 할 때 같은 키(비밀키)를 사용**하는 암호화 방식  
AES, SEED, ARIA 등

#### 공개키 암호화

암호화 할 때와 **복호화 할 때 서로 다른 키(공개키, 개인키)를 사용**하는 암호화 방식  
RSA, ECC 등

#### 일방향 암호화 (암호학적 해쉬함수)

암호화문에서 본문으로의 **복원(복호화)이 불가능한 방식**으로 암호문의 크기가 매우 작아짐(축약)  
**SHA-256/512** 등

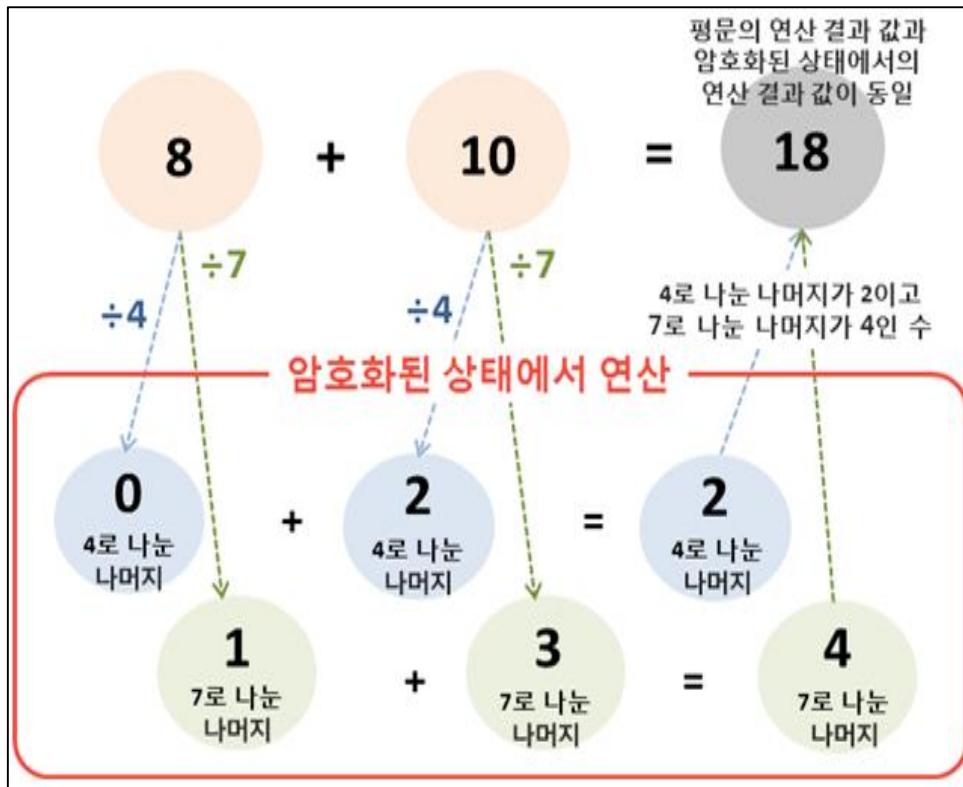
## 2. 개인정보 가명처리 기술

### 4. 암호화

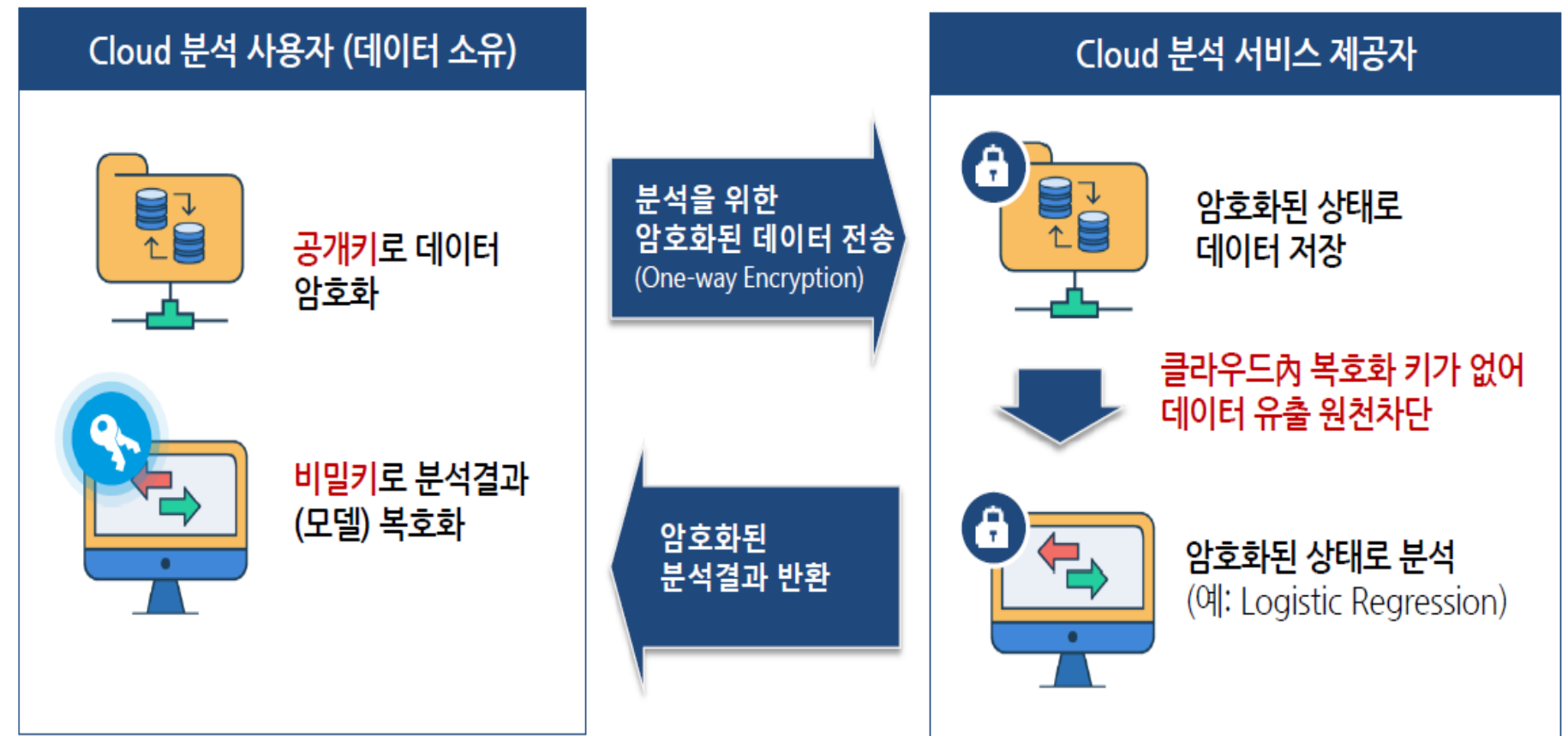
#### 동형 암호화

- Homomorphic Encryption
- 1970년대 이론 연구 시작, 2009년 IBM Gentry가 기술적 가능성 증명
- 평문에 대한 연산 결과와 암호문에 대한 연산 결과가 같은 값을 가져, 암호화된 개인정보를 풀어보지 않고도 통계분석이 가능한 기술

※ (출처) 과학기술정보통신부, “안전한 데이터 활용을 위한 동형암호 기술 실증”



※ (출처) 삼성SDS, “데이터분석을 위한 동형암호 기술”



※ 추가적인 동형암호에 관한 정보는 서울대학교 천정희 교수 홈페이지 참고([www.math.snu.ac.kr/~jhcheon](http://www.math.snu.ac.kr/~jhcheon))

## 2. 개인정보 가명처리 기술

### 4. 암호화

#### 동형 비밀 분산

- Homomorphic Secret Sharing
- 식별자 또는 기타 특성정보를 메시지 공유 알고리즘에 의해 생성된 두 개 이상의 쉼어(share)로 대체하는 기법
- 수학 연산을 이용하여 식별자 또는 기타 속성 값들을 여러 개의 쉼어(share)로 분할하여 쉼어 소유자(share-holder)들에게 배포, 정보를 여러 명의 쉼어 소유자들이 공유
- 계산에 관한 성능 오버헤드가 상대적으로 낮지만, 쉼어 소유자와 쉼어를 교환할 때 발생하는 추가적인 오버헤드가 발생하며 이용
  - 기법에 따라 **상당한 성능 비용이 발생**할 수 있음
- 공유 데이터의 통제된 재식별화는 비식별화된 데이터의 쉼어를 소유한 쉼어 소유자가 정해진 수 만큼 재식별화에 모두 동의할 경우만 가능
- 관련 기법과 연산에 대한 설명은 ISO/IEC 19592와 ISO/IEC 29101에 표준화되어 있음

## 2. 개인정보 가명처리 기술

순열

잡음 추가

부분 총계

가입일자	노이즈	노이즈가입일자
2001-11-05	3	2001-11-08
2007-09-27	-1	2007-09-26
2002-06-11	-5	2002-06-06
2002-10-27	-6	2002-10-21
2006-01-18	3	2006-01-21
2007-06-17	4	2007-06-21
2005-10-10	-4	2005-10-06
2002-08-13	4	2002-08-17
2008-08-08	-4	2008-08-04
2006-04-18	-7	2006-04-11
2004-05-06	-3	2004-05-03
2007-10-10	0	2007-10-10
2005-03-25	5	2005-03-30

성이 높지만 분석에 꼭 필요한

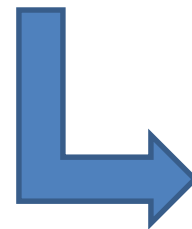
## 2. 개인정보 가명처리 기술

### 6. 해부화 기법

#### 해부화

- 하나의 테이블을 두 개 이상의 테이블로 분할하여 개인의 식별성을 낮추는 기법
- 일반적으로 해부화를 적용할 때 **식별성이 있는 컬럼과 분석 대상 컬럼을 분할**

임시 일련번호	이름	성별	나이	수신 총 잔액	신용대출한도	신용등급
1	조미선	F	33	817,250	66,300,000	3
2	홍길병	M	61	4,559,120	327,700,000	2
3	김영심	F	50	13,601,564	41,300,000	3
4	이미정	F	70	979,118	64,600,000	7
5	김경태	M	40	5,501,809	2,300,000	10
6	유영근	M	43	609,622	13,900,000	7



임시 일련번호	이름	성별	나이
1	조미선	F	33
2	홍길병	M	61
3	김영심	F	50
4	이미정	F	70
5	김경태	M	40
6	유영근	M	43

임시 일련번호	수신 총 잔액	신용대출한도	신용 등급
1	817,250	66,300,000	3
2	4,559,120	327,700,000	2
3	13,601,564	41,300,000	3
4	979,118	64,600,000	7
5	5,501,809	2,300,000	10
6	609,622	13,900,000	7

## 2. 개인정보 가명처리 기술

### 7. 재현 데이터

#### 재현 데이터

- 원자료와 다르지만 원자료와 동일한 분포를 따르도록 통계적으로 생성한 자료

#### 완전 재현자료

1993년 다중대체 기법을 기반으로 Rubbin이 제시한 비밀보호(Data Confidentiality) 방안

- ① 표본틀(모집단)에서 조사되지 않은 모든 값들을 결측값으로 취급하여 다중대체하고,
- ② 대체되어 채워진 재현 모집단에서 단순랜덤추출로 표본을 추출하여 제공은 방법

#### 부분 재현자료

1993년 Little이 제시

자료의 모든 정보가 민감하다고 보기는 어려운 경우도 많으므로,  
모든 변수가 아니라 노출제어 처리가 필요한 일부 변수만 다중대체하자는 방식

## PART3

# 익명화 프라이버시 보호 모델



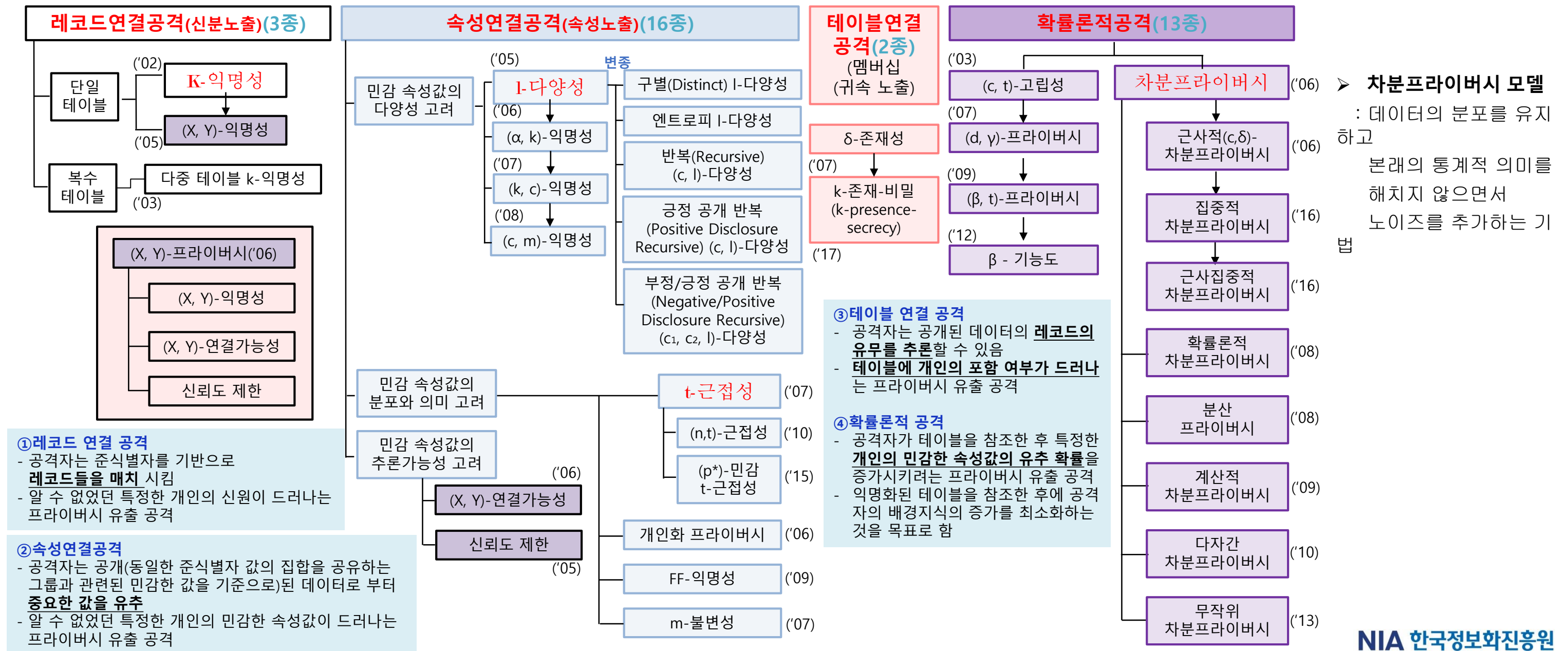


# 3. 프라이버시 보호 모델

## ■ 프라이버시 보호모델

- 가명/익명정보의 **재식별화 공격 위험을 최소화** 하기 위해 다양한 프라이버시 보호모델 개발 및 **재식별 위험성 정량적 측정** 시도

# ISO/IEC JTC1/SC27 WG5 20889, 2018.11-프라이버시 측정모델(34종)



### 3. 프라이버시 보호 모델

#### ■ 프라이버시 보호모델

기법	의미	적용례
k-익명성	<ul style="list-style-type: none"> <li>특정인임을 추론할 수 있는지 여부를 검토, 일정 확률수준 이상 비식별 되도록 함</li> </ul>	<ul style="list-style-type: none"> <li>동일한 값을 가진 레코드를 k개 이상으로 함</li> <li>이 경우 특정 개인을 식별할 최대확률은 <math>1/k</math> 임</li> </ul>
l-다양성	<ul style="list-style-type: none"> <li>k-익명성을 만족하여도 해당 정보가 한쪽으로 편중되어 있어 생기는 프라이버시 이슈를 해결하는 기법</li> </ul>	<ul style="list-style-type: none"> <li>각 레코드는 최소 l개 이상의 다양성을 갖도록 하여 추론 방지</li> </ul>
t-근접성	<ul style="list-style-type: none"> <li>특정집단의 분포가 전체집단과 차이를 보일 때, 분포 차이로 인한 추론 가능성을 낮추는 기법</li> </ul>	<ul style="list-style-type: none"> <li>전체 데이터 집합의 정보 분포와 특정 정보의 분포 차이를 t이하로 하여 추론 방지</li> </ul>

### 3. 프라이버시 보호 모델

#### ■ K-익명성(k-anonymity) : 프라이버시 보호를 위한 기본 모델

- (개념) 공개된 데이터에 대한 연결공격(Linkage Attack) 등 취약점을 방어하기 위해 제안된 프라이버시 보호모델

※ 연결공격

- 다른 데이터 셋의 특정 속성값을 결합하여 개인의 민감한 정보를 재식별할 수 있는 공격
- (예) 미국 매사추세츠 주지사 의료정보 식별(1997년), '선거인명부'와 '공개 의료데이터'가 결합하여 개인의 병명 노출 사례

- (정의) 주어진 데이터 집합에서 같은 값이 적어도 k개 이상 존재하도록 하여 쉽게 다른 정보로 결합할 수 없도록 함
  - 데이터 집합의 일부를 수정하여 모든 레코드가 자기 자신과 동일한(구별되지 않는) k 개 이상의 레코드를 가짐

■ K-익명성(k-anonymity) : 프라이버시 보호를 위한 기본 모델 (사례)

● <표 1> 공개 의료데이터 사례 ●				
구분	지역 코드	연령	성별	질병
1	13053	28	남	전립선염
2	13068	21	남	전립선염
3	13068	29	여	고혈압
4	13053	23	남	고혈압
5	14853	50	여	위암
6	14853	47	남	전립선염
7	14850	55	여	고혈압
8	14850	49	남	고혈압
9	13053	31	남	위암
10	13053	37	여	
11	13068	36	남	
12	13068	35	여	

● <표 2> 선거인명부 사례 ●				
구분	이름	지역코드	연령	성별
1	김민준	13053	28	남
2	박지훈	13068	21	남
3	이지민	13068	29	여
4	최현우	13053	23	남
5	정서연	14853	50	여
6	송현준	14850	47	남
7	남예은	14853	55	여
8	성민재	14850	49	남
9	윤건우	13053	31	남
10	손윤서	13053	37	여
11	민우진	13068	36	남
12	허수빈	13068	35	여

■ K-익명성(k-anonymity) : 프라이버시 보호를 위한 기본 모델 (사례)

● <표 1> 공개 의료데이터 사례 ●				
구분	지역 코드	연령	성별	질병
1	13053	28	남	전립선염
2	13068	21		
3	13068	29		
4	13053	23		
5	14853	50		
6	14853	47		
7	14850	55		
8	14850	49		
9	13053	31		
10	13053	37		
11	13068	36		
12	13068	35		

K-4 익명성 만족

● <표 3> k-익명성 모델에 의해 비식별된 의료데이터 사례 ●					
구분	지역 코드	연령	성별	질병	비고
1	130**	< 30	*	전립선염	다양한 질병이 혼재되어 안전
2	130**	< 30	*	전립선염	
3	130**	< 30	*	고혈압	
4	130**	< 30	*	고혈압	
5	1485*	> 40	*	위암	다양한 질병이 혼재되어 안전
6	1485*	> 40	*	전립선염	
7	1485*	> 40	*	고혈압	
8	1485*	> 40	*	고혈압	
9	130**	3*	*	위암	모두가 동일 질병(위암)으로 취약
10	130**	3*	*	위암	
11	130**	3*	*	위암	
12	130**	3*	*	위암	

### 3. 프라이버시 보호 모델

#### ■ L-다양성(l-diversity) : K-익명성의 취약점을 보완한 프라이버시 보호모델

- (개념) K-익명성에 대한 두 가지 공격, 즉 동질성 공격 및 배경지식에 의한 공격을 방어하기 위한 모델

- ※ 동질성 공격(Homogeneity Attack)

k-익명성에 의해 레코드들이 범주화 되었더라도 일부 정보들이 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

- (정의) 주어진 데이터 집합에서 함께 비식별되는 레코드들은 (동질 집합에서) 적어도 한 개 이상의 서로 다른 민감한 정보를 가져야 함
  - 비식별 조치 과정에서 충분히 다양한(1개 이상) 서로 다른 민감한 정보를 갖도록 동질 집합을 구성

### 3. 프라이버시 보호 모델

#### ■ L-다양성(l-diversity) : K-익명성의 취약점을 보완한 프라이버시 보호모델 (사례)

<div> <b>k = 4</b> <span>● &lt;표 3&gt; k-익명성 모델에 의해 비식별된 의료데이터 사례 ●</span> </div>					
구분	지역 코드	연령	성별	질병	비고
1	130**	< 30	*	전립선염	다양한 질병이 혼재되어 안전
2	130**	< 30	*	전립선염	
3	130**	< 30	*	고혈압	
4	130**	< 30	*	고혈압	
5	1485*	> 40	*	위암	다양한 질병이 혼재되어 안전
6	1485*	> 40	*	전립선염	
7	1485*	> 40	*	고혈압	
8	1485*	> 40	*	고혈압	
9	130**	3*	*	위암	모두가 동일 질병(위암)으로 취약
10	130**	3*	*	위암	
11	130**	3*	*	위암	
12	130**	3*	*	위암	

→ L 값 : 2  
이지민 / 29세 / 여

→ L 값 : 3

→ L 값 : 1  
130\*\*의 지역코드 / 30대

전체 데이터셋의 L 다양성은 1

L-1

※ ‘이지민 / 29세 / 여’의 경우,

- 배경 지식(전립선염은 남자에 해당)을 통해 ‘고혈압’인 것과
- 동질성 공격으로 130\*\* 지역코드의 30대는 모두 ‘위암’ 환자인 것을 식별할 수 있음

■ L-다양성(l-diversity) : K-익명성의 취약점을 보완한 프라이버시 보호모델 (사례)

k = 4

● <표 3> k-익명성 모델에 의해 비식별된 의료데이터 사례 ●

구분	지역 코드	연령	성별	질병	비고
1	130**	< 30	*	전립선염	다양한 질병이 혼재되어 안전
2	130**	< 30	*	전립선염	
3	130**	< 30	*	고혈압	
4	130**	< 30	*	고혈압	
5	1485*	> 40	*	위암	다양한 질병이 혼재되어 안전
6	1485*	> 40	*	전립선염	
7	1485*	> 40	*	고혈압	
8	1485*	> 40	*	고혈압	
9	130**	3*	*	위암	모두가 동일 질병(위암)으로 취약
10	130**	3*	*	위암	
11	130**	3*	*	위암	
12	130**	3*	*	위암	

● <표 4> l-다양성 모델에 의해 비식별된 의료데이터의 예 ●

구분	지역 코드	연령	성별	질병	비고
1	1305*	≤ 40	*	전립선염	다양한 질병이 혼재되어 안전
4	1305*	≤ 40	*	고혈압	
9	1305*	≤ 40	*	위암	
10	1305*	≤ 40	*	위암	
5	1485*	> 40	*	위암	다양한 질병이 혼재되어 안전
6	1485*	> 40	*	전립선염	
7	1485*	> 40	*	고혈압	
8	1485*	> 40	*	고혈압	
2	1306*	≤ 40	*	전립선염	다양한 질병이 혼재되어 안전
3	1306*	≤ 40	*	고혈압	
11	1306*	≤ 40	*	위암	
12	1306*	≤ 40	*	위암	

※ ‘이지민 / 29세 / 여’의 경우,  
- 배경지식(전립선염은 남자에 해당)을 통해 ‘고혈압’인 것과  
- 공질성 공격으로 130\*\* 지역코드의 30대는 모두 ‘위암’ 환자인 가진 것  
을 식별할 수 있음

l = 3



### 3. 프라이버시 보호 모델

#### ■ t-근접성(t-closeness) : 값의 의미(분포도)를 고려하는 프라이버시 보호 모델

##### ■ (개념) I-다양성의 취약점(쏠림 공격, 유사성 공격)을 보완하기 위한 모델

###### ※ 쏠림 공격(skewness Attack)

- 정보가 특정한 값에 쏠려 있을 경우 I-다양성 모델이 프라이버시를 보호하지 못함
- (예) 임의의 '동질 집합'이 99개의 '위암 양성' 레코드와 1개의 '위암 음성' 레코드로 구성되어 있다 가정 시,  
공격자는 공격 대상이 99%의 확률로 '위암 양성'이라는 것을 알 수 있음

###### ※ 유사성 공격 (Similarity Attack)

- 비식별 조치된 레코드의 정보가 서로 비슷하다면 I-다양성 모델을 통해 비식별 된다 할지라도 프라이버시가 노출될 수 있음
- 동질 집합의 값(예: 병명)이 서로 다르지만 의미가 서로 유사함(예: 위궤양, 급성 위염, 만성 위염)

(정의) 전체 데이터에서 민감한 정보의 분포와 각 동질 집합에서 민감한 정보 분포의 차이가  $t$  값 이하임을 보장 (단,  $0 \leq t \leq 1$ )

- $t$ 가 '0' 가까울수록 전체 데이터에서의 민감한 정보와 동질 집합에서 민감한 정보의 분포의 차가 작아짐. 즉, 서로 비슷한 분포를 의미

■ t-근접성(t-closeness) : 값의 의미를 고려하는 프라이버시 보호 모델(사례)

● <표 5> l-다양성 모델에 의해 비식별되었지만 유사성 공격에 취약한 사례 ●					
구 분	속성자		민감한 정보		비고
	지역 코드	연령	급여(백만원)	질병	
1	476**	2*	30	위궤양	모두가 '위'와 관련한 유사 질병으로 취약
2	476**	2*	40	급성 위염	
3	476**	2*	50	만성 위염	
4	4790*	≥ 40	60	급성 위염	다양한 질병이 혼재되어 안전
5	4790*	≥ 40	110	감기	
6	4790*	≥ 40	80	기관지염	
7	476**	3*	70	기관지염	다양한 질병이 혼재되어 안전
8	476**	3*	90	폐렴	
9	476**	3*	100	만성 위염	

➤ K-3의 익명성 만족, L-3의 다양성도 만족 하나,,,  
데이터 내 포함된 지역코드 476\*\*, 20대는  
모두 “위”와 관련한 유사한 질병임을 식별할 수 있음  
(데이터 분포의 유사성/동질성으로 인한 취약점 발생 가능)

● <표 6> t-근접성 모델에 의해 비식별 조치된 데이터 사례 ●					
구 분	속성자		민감한 정보		비고
	지역 코드	연령	급여(백만원)	질병	
1	4767*	≤ 40	30	위궤양	급여의 분포와 다양한 질병 으로 안전
3	4767*	≤ 40	50	만성 위염	
8	4767*	≤ 40	90	폐렴	
4	4790*	≥ 40	60	급성 위염	급여의 분포와 다양한 질병 으로 안전
5	4790*	≥ 40	110	감기	
6	4790*	≥ 40	80	기관지염	
2	4760*	3*	40	급성 위염	급여의 분포와 다양한 질병 으로 안전
7	4760*	3*	70	기관지염	
9	4760*	3*	100	만성 위염	

### 3. 프라이버시 보호 모델

#### ■ Privacy beyond k-Anonymity and l-Diversity

[https://www.cs.purdue.edu/homes/ninghui/papers/t\\_closeness\\_icde07.pdf](https://www.cs.purdue.edu/homes/ninghui/papers/t_closeness_icde07.pdf)

**Original** Patients Table


	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

**Original** Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer


**3-Anonymous** Version

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

  
 K-3  
 익명성  
 보장  
 익명성은 L-1

**3-diverse** version

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

  
 K-3 익명성  
 L-3 다양성 보장

개인정보 가명·익명 처리 기술

End of Document

감사합니다.

