



# 아이디어 제안서

## 보고서 및 논문 윤리 서약

1. 나는 보고서 및 논문의 내용을 조작하지 않겠습니다.
  2. 나는 다른 사람의 보고서 및 논문의 내용을 내 것처럼 무단으로 복사하지 않겠습니다.
  3. 나는 다른 사람의 보고서 및 논문의 내용을 참고하거나 인용할 시 참고 및 인용 형식을 갖추고 출처를 반드시 밝히겠습니다.
  4. 나는 보고서 및 논문을 대신하여 작성하도록 청탁하지도 청탁받지도 않겠습니다.
- 나는 보고서 및 논문 작성 시 위법 행위를 하지 않고, 명지인으로서 또한 공학인으로서 나의 양심과 명예를 지킬 것을 약속합니다.



보고서명 : 금융 빅데이터 비식별화를 활용한 통신사 미납 해결

학 과 : 컴퓨터공학과

응모분야 : 빅데이터 아이디어

담당기관 : 금융보안원

학 번 : 60122461

이 름 : 이 승 현 (서명)



## Solution about Calling Plan

제안 아이디어 : 금융 빅데이터 비식별화를 활용한 통신사 미납 해결

보고기관: 금융보안원

팀 (개인): 이 승 현(명지대학교, 경기도 용인시 처인구 명지로 116, 031-330-6780)

작성일자: 2017.07.29.

문서 버전: V1.0

## 요약문

이 문서는 금융 빅데이터 활용한 통신사 미납 문제를 해결하는 금융 빅데이터 아이디어 공모전 제안서 이다. Python 라이브러리에서 제공하는 암호 방식을 이용한 데이터 결합과 ARX 데이터 비식별화 툴을 활용한 데이터 가공을 통해 바람직한 금융 데이터 활용 및 구현된 데이터를 활용한 경제성 향상에 이바지함을 그 핵심가치로 둔다.

## Abstract

*This document is a proposal-report for the 'Solution about calling plan' project that generates the proper way inside the building. And it's based on python libraries and ARX.*

*So, this project aims at improving utilization about financial data. Overall, this project will contribute to construct desirable financial utilizing-culture and materialize the calling plan economically.*

---

## <목 차>

1. 제안 개요 .....	1
1.1. 목적 .....	1
1.2. 주안점 .....	1
1.3. 팀 구성 .....	1
1.4. 목표 .....	1
1.5. 보고서의 구성 .....	2
2. 배경 및 필요성 .....	3
3.1. 배경 .....	3
3.2. 시장성 및 필요성 .....	4
3. 제안 아이디어 .....	6
3.1. 관련 이론 .....	6
3.2. 구현 .....	11
4. 기대효과 .....	19
5.1. 기대효과 .....	19
5.2. 마무리 .....	20
5. 참고자료 .....	21
5.1. 출처 .....	21
5.2. 기타자료 .....	22

# 1. 제안 개요

## 1.1 목적

본 제안서는 파이썬 라이브러리 및 데이터 비식별화 툴 ARX를 활용한 금융 빅데이터 가공 및 통신사 미납 활용 방안이다. 금융 업체는 통신사 관계자들이 요구하는 빅데이터를 제공할 수 있게 가공한다. 이를 필요로 하는 관계자들은 금융회사에 돈을 지급하고 데이터를 얻는다. 통신사는 이를 바탕으로 새로운 요금제를 설립하여 고객들의 미납 문제를 해결한다. 이는 최근 빅데이터에 대한 수요가 증가하는 시점에서 통신사 관계자들에게 활용할 수 있는 데이터를 제공하여 시간 절약은 물론 체계적인 요금제 설계가 가능하게 할 것이다. 또한, 금융 업체는 데이터를 제공하여 경제적인 이윤을 얻고 수요자에 대한 피드백을 통해 고객과의 소통이 더욱 원활해질 수 있다. 새롭게 정립된 요금제를 이용하는 고객은 새로 정립된 요금제를 통해 혜택을 누리고 요금 미납 문제를 해소한다. 이를 통해 더욱 바람직한 금융 문화 조성 및 소비자 가치를 창출하는 것이 본 제안서의 근본적인 목적이다.

## 1.2 주안점

금융 빅데이터를 활용하여 금융회사들의 이익에 이바지한다. 또한, 이를 활용하는 통신사 관계자들에게 데이터를 제공하여 새로운 요금 기준을 세워 고객들의 미납 문제를 해결한다. 고객은 이를 통해 혜택을 얻는다.

## 1.3 팀 구성

직책	성 명	학 번	이메일	연락처	경험 및 능력
학생	이승현	60122461	abbc020948@gmail.com	010-7228-2686	C, JAVA, Web, Python, ARX

< 팀 구성도 및 개발경험 >

## 1.4 목표

아이디어 제안 이후 데이터 결합 시 암호 알고리즘 선택 및 데이터 비식별화에 대한 솔루션 제작을 위해 노력을 해왔으나 비용적인 문제와 시간적 제약이 문제시되었다. 이에 따라 솔루션 제공이 아닌 두 기술을 병렬적으로 제공하여 사용자가 선택적으로 기능을 보안을 유지하여 사용할 수 있도록 가이드 라인 제시에 초점을 맞췄다. 이를 위해 암호화 라이브러리는 스크립트 언어인 파이썬을 이용하고 대표적인 데이터 비식별화 툴인 ARX를 활용한다.

## 1.5 보고서의 구성

### 가) 제안 개요

: 제안서의 목적, 주안점, 지원 주체와 사용자 및 기타 관련자 목록, 팀 구성, 목표 등 전반적인 내용을 설명한다.

### 나) 배경 및 필요성

: 현실적 제한조건을 바탕으로 배경 및 필요성을 설명한다.

### 다) 제안 아이디어

: 본 단계(제안 아이디어)에서 문제해결에 적용된 이론을 설명하고 업무를 수행한다.

### 라) 기대효과

: 업무 수행 후 발생하는 현실적인 기대효과에 관해 기술한다.

### 마) 참고자료

: 참고자료들의 출처를 기술한다

## 2. 배경 및 필요성

### 2.1 배경

2015년 말 기준 SK텔레콤 KT LG유플러스 등 국내 이동통신 3사에서 20대들이 통신요금을 체납한 규모는 13만9185건, 511억6100만원에 달했다. 그동안 SK텔레콤은 미납자 가운데 1년 이상, 100만원 이상 통신요금을 연체한 가입자를 채무불이행자로 신용평가회사(나이스신용정보)에 등록해 왔다. 회사 관계자는 “통신요금이 3개월 연체되면 바로 이용 정지를 하는데 100만원 이상 체납했다는 것은 정상적인 사용이 아닌 것으로 판단할 수 있다”고 설명했다.

회사 측은 다른 사람 명의를 불법 도용해 쓰거나 노숙자 등 사회취약계층의 개인정보를 활용해 이른바 ‘대포폰’(불법 개통 휴대폰)으로 사용하는 사례가 적지 않다고 밝혔다. 신용평가사에 과도한 연체자의 정보를 제공한 것은 사회적인 경각심을 심어주기 위한 불가피한 조치였다는 것이다. 김 의원실에 따르면 SK텔레콤이 2012~2015년 신용평가사에 등록한 채무불이행자는 6만7356명이었다.

하지만 이날 청년 신용불량자를 양산한다는 지적이 나오자 SK텔레콤은 KT와 LG유플러스처럼 연체정보의 신용평가사 제공을 중단하기로 했다. 통신3사 모두 앞으로는 한국정보통신진흥협회를 통해 연체 정보를 관리한다.

SK텔레콤이 일부 부정적 여론 탓에 연체 정보 등록을 중단하기로 하자 금융당국과 신용평가업계는 물론 금융업계도 “신용사회 흐름에 역행하는 것”이라는 반응을 내놨다.

SK텔레콤이 연체 정보 등록을 중단함에 따라 신용 관리가 더욱 어려워질 것이란 우려도 나온다. 앞으로는 금융회사들이 통신사 가입자의 연체 정보를 조회할 수 있는 길이 없어지기 때문이다. 금융감독원이 추진 중인 비(非)금융정보의 신용평가 반영 방안에도 찬물을 끼얹을 것이란 지적도 있다.



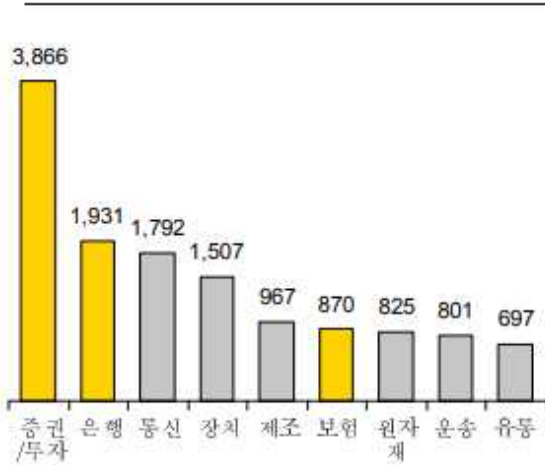
출처 : 한국 경제 매거진

## 2.2 시장성 및 필요성

통신사들은 미납자로 발생하는 금전적인 손실을 감수해야 하는 상황에서 금융 데이터 활용은 위와 같은 문제를 해결하는 방안이 될 수 있다. 나이나 성별과 같은 정보에 따른 신용등급이나 소비력 등을 참고하여 요금제를 개정하는 것이 그 예가 될 것이다. 실제 금융업은 타 산업 대비 데이터 보유량이 많고 증가 속도가 빠른 것으로 분석되었다.

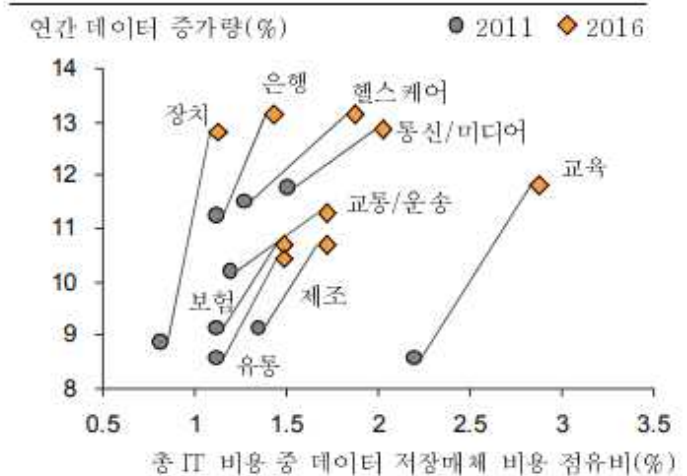
- 증권투자, 은행, 보험 순으로 현재 데이터 보유량이 많은 것으로 집계
- 은행은 향후 데이터 증가량이 타 산업을 상회할 것으로 전망

[그림 12] 산업별 기업의 평균 보유  
데이터 보유량 (미국 사례, 단위: 테라바이트)



자료: IDC, McKinsey & Company

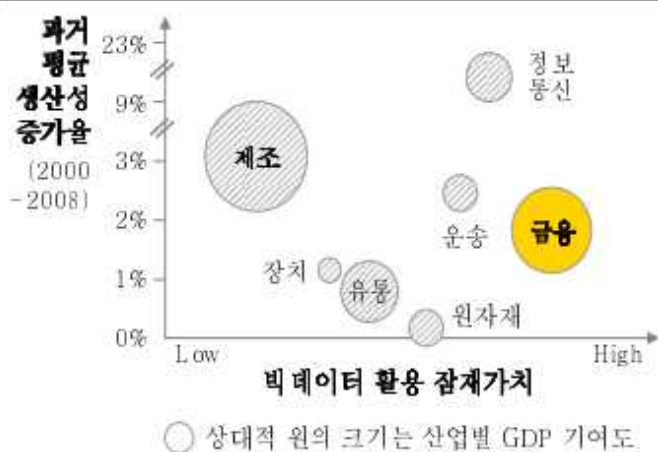
[그림 13] 산업별 데이터 증가 추세  
(전 세계 주요 기업 대상 설문조사 결과)



자료: Gartner

- 빅데이터의 활용을 통한 기업의 생산성 증가 정도를 산업별로 살펴보면, 정보통신, 금융, 운송 산업의 생산성 증가 효과가 클 것으로 기대
- 금융업은 2000년 대 들어 생산성 개선 수준이 타 산업 대비 낮은 편이었으나, 향후 빅데이터 활용으로 높은 수준의 생산성 증가가 기대

[그림 14] 산업별 빅데이터 활용 잠재가치(미국 사례)



자료: McKinsey & Company

출처 : KB 경영연구소

- **(신규 수익원)** 빅데이터 분석 결과의 판매, 타사와 전략적 제휴 통한 데이터 공유
  - JP모건은 신용카드 사업을 통해 축적된 고객의 거래 내역 등 내부 데이터와 미국의 경제지표 통계 등 외부 데이터를 통합해 소비 동향 분석 보고서를 작성했으며 이를 은행 고객에게 판매
  - 씨티그룹은 글로벌 데이터베이스를 통해 통합된 고객 거래 정보를 스페인의 의류회사와 공유하기로 했으며, 의류회사는 공유된 데이터를 활용해 생산시설 및 판매매장 위치 계획 수립 등에 활용할 계획

하지만, 데이터 활용에 대한 부정적인 시선도 있다. 건강사회를위한약사회, 민주노총, 민주사회를위한 변호사모임, 참여연대 등 12개 시민단체는 2017년 11월 9일 오후 1시 서울 서초동 서울중앙지검 앞에서 기자회견을 열고 한국인터넷진흥원, 한국정보화진흥원, 금융보안원, 한국신용정보원 등 4개 비식별 전문기관과 SK텔레콤 등 20개 기업을 개인정보보호법 위반 등의 혐의로 고발한다고 밝혔다.

조지훈 민주사회를위한변호사모임 디지털정보위원회 위원장은 “이른바 인증기관이라는 공공기관이 SK 텔레콤, 삼성생명 등 여러 민간 기업으로부터 받은 비식별화된 개인정보를 정보 결합물로 다시 제공하는 과정이 얼마나 심각한 위법 행위인지 따져보고자 한다”며 고발 취지를 밝혔다. 조 위원장은 “이 행위가 토대로 한 ‘개인정보 비식별 가이드라인’이 어떤 법적 근거와 타당성도 없다는 것을 알리기 위한 고발이기도 하다”고 설명했다.

시민단체들은 기자회견문을 통해 “지난 9월12일 국정감사를 통해 박근혜 정부가 설립한 비식별 전문기관이 통신, 금융 대기업들의 고객 정보를 고객 모르게 결합시켜주었다는 사실이 밝혀졌다. 지난 1년간 비식별 전문기관과 기업들이 3억4000만건에 이르는 개인정보를 무단으로 결합한 행위를 고발한다”고 밝혔다.

또 “비식별조치 가이드라인은 개인정보를 좁게 해석한다. 정부가 권장하는 방식대로 비식별조치만 취하면 더 이상 개인정보가 아닌 것으로 추정해주겠다는 것이 요지다. 그래서 기업이 개인정보를 정보주체인 이용자, 소비자, 환자 동의 없이 영리적 목적으로 얼마든지 자유롭게 제3자에게 제공하고 결합하고 심지어 판매할 수 있도록 하였다”고 주장했다.



출처 : 경향신문

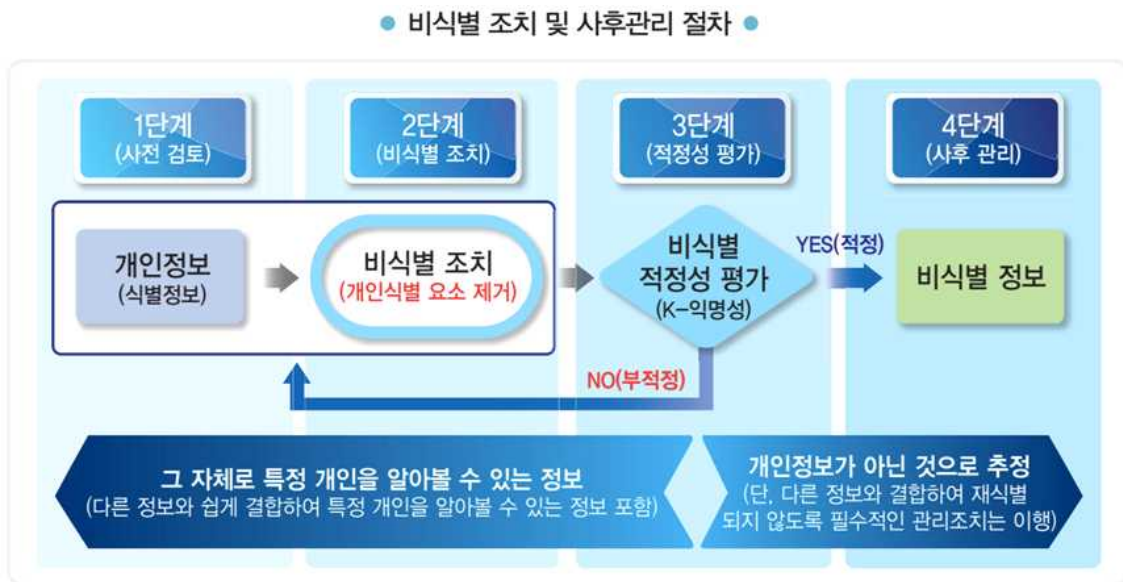


### 3. 제안 아이디어

#### 3.1. 관련 이론

비식별 전문기관은 2016년 6월 박근혜 정부 당시 방송통신위원회와 미래창조과학부(현 과학기술정보통신부) 등 6개 관계부처가 합동으로 발표한 ‘개인정보 비식별 가이드라인’에 따라 개인정보를 가명화·익명화·범주화해 개인정보가 아닌 것으로 ‘비식별화’하는 공공기관을 말한다.

이처럼 비식별화를 거쳐 기업이 마케팅 등에 활용할 수 있도록 하는 조치를 ‘정보집합물 결합서비스’라고 부른다. 이렇게 비식별화된 개인정보는 기업 등에 의해 활용하거나 유통될 수 있다.



출처 : 2016 비식별화 가이드라인

단계별 조치사항

- ① (사전 검토) 개인정보에 해당하는지 여부를 검토 후, 개인정보가 아닌 것이 명백한 경우 법적 규제 없이 자유롭게 활용
- ② (비식별 조치) 정보집합물(데이터 셋)에서 개인을 식별할 수 있는 요소를 전부 또는 일부 삭제하거나 대체하는 등의 방법을 활용, 개인을 알아볼 수 없도록 하는 조치
- ③ (적정성 평가) 다른 정보와 쉽게 결합하여 개인을 식별할 수 있는지를 「비식별 조치 적정성 평가단」을 통해 평가
- ④ (사후관리) 비식별 정보 안전조치, 재식별 가능성 모니터링 등 비식별 정보 활용 과정에서 재식별 방지를 위해 필요한 조치 수행

〈 예시 〉 식별자

- 고유식별정보(주민등록번호, 여권번호, 외국인등록번호, 운전면허번호)
- 성명(한자·영문 성명, 필명 등 포함)
- 상세 주소(구 단위 미만까지 포함된 주소)
- 날짜정보 : 생일(양/음력), 기념일(결혼, 돌, 환갑 등), 자격증 취득일 등
- 전화번호(휴대전화번호, 집전화, 회사전화, 팩스번호)
- 의료기록번호, 건강보험번호, 복지 수급자 번호
- 통장계좌번호, 신용카드번호
- 각종 자격증 및 면허 번호
- 자동차 번호, 각종 기기의 등록번호 & 일련번호
- 사진(정지사진, 동영상, CCTV 영상 등)
- 신체 식별정보(지문, 음성, 홍채 등)
- 이메일 주소, IP 주소, Mac 주소, 홈페이지 URL 등
- 식별코드(아이디, 사원번호, 고개번호 등)
- 기타 유일 식별번호 : 군번, 개인사업자의 사업자 등록번호 등

※ 美 HIPAA 프라이버시 규칙을 참고하여 작성

• 〈 예시 〉 속성자 •

개인 특성	<ul style="list-style-type: none"> <li>• 성별, 연령(나이), 국적, 고향, 시·군·구명, 우편번호</li> <li>• 병역여부, 결혼여부, 종교, 취미, 동호회·클럽 등</li> <li>• 흡연여부, 음주여부, 채식여부, 관심사항 등</li> </ul>
신체 특성	<ul style="list-style-type: none"> <li>• 혈액형, 신장, 몸무게, 허리둘레, 혈압, 눈동자 색깔 등</li> <li>• 신체검사 결과, 장애유형, 장애등급 등</li> <li>• 병명, 상병(傷病)코드, 투약코드, 진료내역 등</li> </ul>
신용 특성	<ul style="list-style-type: none"> <li>• 세금 납부액, 신용등급, 기부금 등</li> <li>• 건강보험료 납부액, 소득분위, 의료 급여자 등</li> </ul>
경력 특성	<ul style="list-style-type: none"> <li>• 학교명, 학과명, 학년, 성적, 학력 등</li> <li>• 경력, 직업, 직종, 직장명, 부서명, 직급, 전직장명 등</li> </ul>
전자적 특성	<ul style="list-style-type: none"> <li>• 쿠키정보, 접속일시, 방문일시, 서비스 이용 기록, 접속로그 등</li> <li>• 인터넷 접속기록, 휴대전화 사용기록, GPS 데이터 등</li> </ul>
가족 특성	<ul style="list-style-type: none"> <li>• 배우자·자녀·부모·형제 등 가족 정보, 법정대리인 정보 등</li> </ul>

출처 : 2016 비식별화 가이드라인

비식별 조치 단계 : 비식별 조치기법 적용 식별자(Identifier) 조치 기준 정보집합물에 포함된 식별자는 원칙적으로 삭제 조치

‘식별자’란 개인 또는 개인과 관련한 사물에 고유하게 부여된 값 또는 이름 다만, 데이터 이용 목적상 반드시 필요한 식별자는 비식별 조치 후 활용 속성자(Attribute value) 조치 기준 정보집합물에 포함된 속성자도 데이터 이용 목적과 관련이 없는 경우에는 원칙적으로 삭제

‘속성자’란 개인과 관련된 정보로서 다른 정보와 쉽게 결합하는 경우 특정 개인을 알아볼 수도 있는 정보

데이터 이용 목적과 관련이 있는 속성자 중 식별요소가 있는 경우에는 가명처리, 총계처리 등의 기법을 활용하여 비식별 조치 희귀병명, 희귀경력 등의 속성자는 구체적인 상황에 따라 개인 식별 가능성이 매우 높으므로 엄격한 비식별 조치 필요

• 〈 예시 〉 비식별 조치 기법 적용 •

원본데이터	주민등록번호	성별	입원날짜	연령	병명
	770914-1234567	남	2015/06/23	39	독감
	850930-1234567	남	2015/10/01	31	독감
	710119-2345678	여	2014/01/21	45	고혈압
	770619-2345678	여	2014/09/23	39	고혈압
	830425-1234567	남	2015/04/16	33	간염
	860804-2345678	여	2014/11/11	30	간염



바식발  
데이터

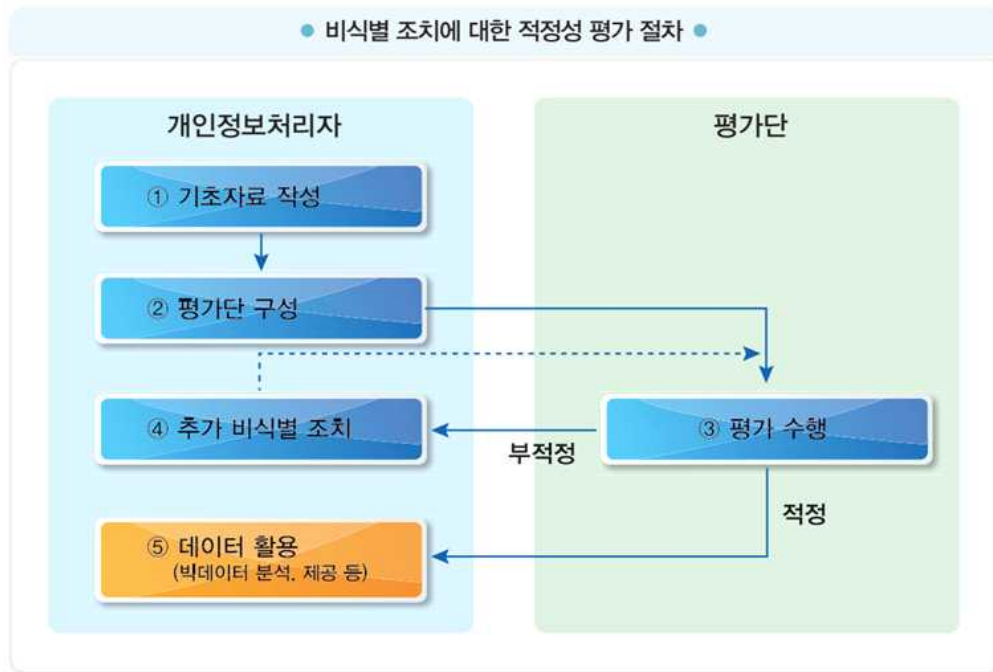
① 데이터 삭제(주민등록번호)

주민등록번호	성별	입원날짜	연령	병명
	남	2015/06/23	39	독감
	남	2015/10/01	31	독감
	여	2014/01/21	45	고혈압
	여	2014/09/23	39	고혈압
	남	2015/04/16	33	간염
	여	2014/11/11	30	간염

② 데이터 마스킹(주민등록번호, 입원날짜, 총계처리(평균 연령))

주민등록번호	성별	입원날짜	연령	병명
7*****-1*****	남	2015/**/**	35	독감
8*****-1*****	남	2015/**/**	35	독감
7*****-2*****	여	2014/**/**	35	고혈압
7*****-2*****	여	2014/**/**	35	고혈압
8*****-1*****	남	2015/**/**	35	간염
8*****-2*****	여	2015/**/**	35	간염

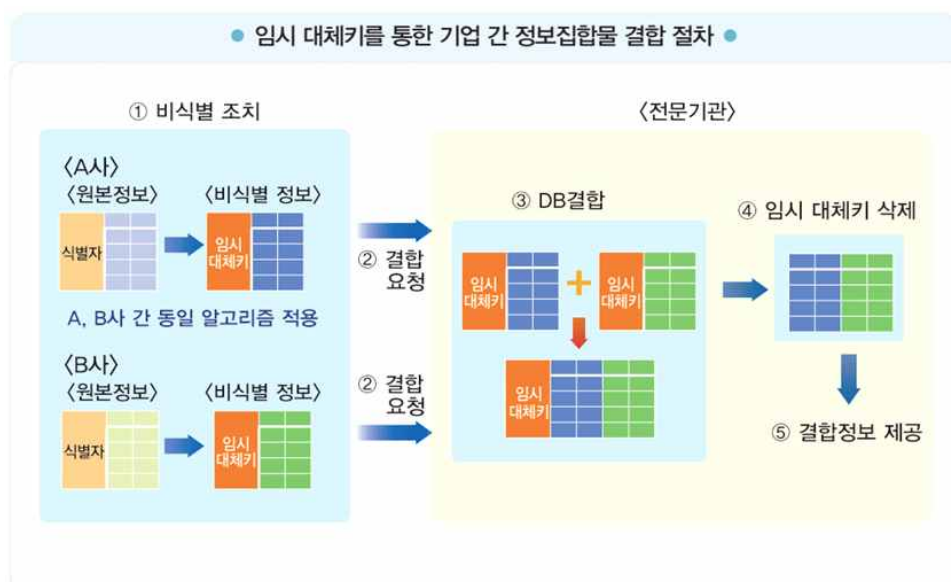
출처 : 2016 비식별화 가이드라인



출처 : 2016 비식별화 가이드라인

#### 적정성 평가 절차

- ① (기초자료 작성) 개인정보처리자는 적정성 평가에 필요한 데이터 명세, 비식별 조치 현황, 이용기관의 관리 수준 등 기초자료 작성
- ② (평가단 구성) 개인정보 보호책임자가 3명 이상으로 평가단을 구성(외부전문가는 과반수 이상)
- ③ (평가 수행) 평가단은 개인정보처리자가 작성한 기초자료와 k-익명성 모델을 활용하여 비식별 조치 수준의 적정성을 평가
- ④ (추가 비식별 조치) 개인정보처리자는 평가결과가 '부적정'인 경우 평가단의 의견을 반영하여 추가적인 비식별 조치 수행
- ⑤ (데이터 활용) 비식별 조치가 적절하다고 평가받은 경우에는 빅데이터 분석 등에 이용 또는 제공이 허용



출처 : 2016 비식별화 가이드라인

## 결합 절차

- ① 같은 알고리즘을 적용하여 식별자를 임시 대체키로 전환하고, 결합대상 정보집합물도 비식별 조치 및 적정성 평가 수행
- ② 비식별 조치된 정보를 전문기관에 제공 및 결합 요청
- ③ 임시 대체키를 활용, 전문기관에서 결합 수행 20 개인정보 비식별 조치 가이드라인
- ④ 임시 대체키 삭제
- ⑤ 결합 DB를 필요한 기업에게 제공(전문기관은 제공 후 파기 조치)

## 프라이버시 보호 모델

### k-익명성(k-anonymity)

- ① k-anonymity는 신분 노출 공격 방어를 위한 개념, 연결공격(linkage attack) 등 취약점을 방어
- ② k-anonymity는 같은 QI 값으로 이루어져 있는 레코드들을 묶는다
- ③ 다른 k-1 사람들과 구분될 수 없도록 익명성 보장
- ④ VGH(Value Generalization Hierarchy) 기법 사용

k = 2

Race	DOB	Gender	ZIP	Problem
black	1965/*/*	male	0214*	diabetic
black	1965/*/*	male	0214*	chest pain
black	1965/*/*	female	0213*	painful eye
black	1965/*/*	female	0213*	wheezing
black	1964/*/*	female	0213*	obesity
black	1964/*/*	female	0213*	chest pain
white	1964/*/*	male	0213*	short of breath
black	1965/*/*	female	0213*	hypertension
white	1964/*/*	male	0213*	obesity
white	1964/*/*	male	0213*	fever
white	1967/*/*	male	0213*	vomiting
white	1967/*/*	male	0213*	back pain

$\ell$ -다양성( $\ell$ -diversity)

- ① k-익명성의 취약점을 보완한 프라이버시 보호 모델
- ② Homogeneity attack(그룹핑 한 것이 모두 값은 값)
- ③ Background Knowledge Attack ex. 여자는 전립선염에 걸리지 않는다.

## $\ell$ -diversity Model

### • 기본형 $\ell$ -diversity

Race	DOB	Gender	ZIP	Problem
black	1965	male	0214*	diabetic
black	1965	male	0214*	chest pain
black	1965	female	0213*	painful eye
black	1965	female	0213*	wheezing
black	1964	female	0213*	obesity
black	1964	female	0213*	hypertension

$\ell$ -diversity with  $\ell=2$

t-근접성(t-closeness)

- ① 값의 의미를 고려하여  $\ell$ -다양성의 취약점을 보완하기 위해 모델
- ② 질 집합에서 특정 정보의 분포와 전체 데이터 집합에서 정보의 분포가 t이하의 차이를 보여야 함
- ③ 분포 간의 유사성은 EMD(Earth Moving Distance) 방법을 사용

• <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 질병으로 안전
3	4767*	≤ 40	50	만성 위염	
8	4767*	≤ 40	90	폐렴	
4	4790*	≥ 40	60	급성 위염	급여의 분포와 다양한 질병으로 안전
5	4790*	≥ 40	110	감기	
6	4790*	≥ 40	80	기관지염	
2	4760*	3*	40	급성 위염	급여의 분포와 다양한 질병으로 안전
7	4760*	3*	70	기관지염	
9	4760*	3*	100	만성 위염	

출처 : 2016 비식별화 가이드라인



### 3.2. 구현

위의 이론을 바탕으로 금융 데이터를 통신사 데이터와 결합하여 비식별화 조치하는 과정을 간단하게 진행했다. 우선 샘플 데이터를 얻기 위해 데이터 비식별화를 위한 오픈소스 소프트웨어를 운영하는 ARX에 데이터를 요청하였다.

이승현

Password request

받는 사람: [redacted]

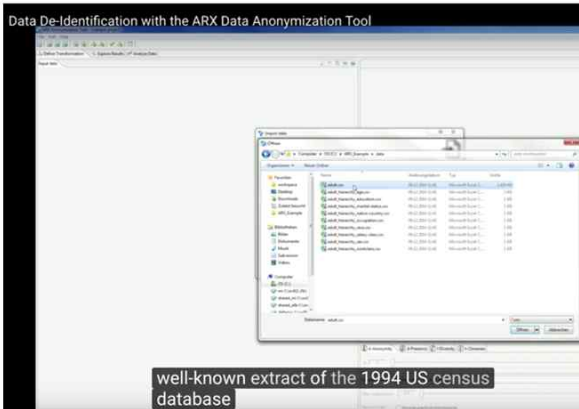
Hello Im student and I need your ARX sample dataset to study data de-identification.

Can I get the sample?

Have a nice day!

If you don't want to give all sample files, can I get the sample file, 'adult.csv'?

Data De-Identification with the ARX Data Anonymization Tool



Fabian Prasser

Re: Password request

받는 사람: 이승현

Hi,

thank you for your interest in ARX!

You can download the files here:

<https://drive.google.com/open?id=0B1QMEQleBZ9zSHBoUHeEZ2hnd00>

If you have any further questions, please don't hesitate to contact us.

Best regards

Fabian

P.S.: To support our efforts, please cite one of our papers if you happen to use ARX for scientific work. Recommended reference: Fabian Prasser, Florian Kohlmayer, Klaus A. Kuhn, A Benchmark of Putting Statistical Disclosure Control Into Practice: The ARX Data Anonymization Tool. In: Gioulalas-Divanis, Aris, Loukides, Grigoris (Eds.), Medical Data Privacy Handbook, Springer, November 2015. ISBN: 978-3-319-29832-2.

If you used our benchmark datasets for your research, please cite the following paper: Fabian Prasser, Florian Kohlmayer, Klaus A. Kuhn, A Benchmark of Globally-Optimal Anonymization Methods for Biomedical Data. Proceedings of the 27th IEEE International Symposium on Computer-Based Medical Systems, May 2014, New York City, USA.

Am 9. Januar 2018 20:24:38 GMT+05:30, schrieb "이승현"

Hello Im student and I need your ARX sample dataset to study data de-identification.

Can I get the sample?

Have a nice day!

Dr. rer. nat. Fabian Prasser  
Computer Scientist

Chair of Medical Informatics  
Institute of Medical Informatics, Statistics and Epidemiology  
University Hospital rechts der Isar  
Technical University of Munich

Grilparzerstr. 18 (3rd floor)  
81675 Munich  
Germany

## KDD Cup 1998 Data

### Abstract

This is the data set used for The Second International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-98 The Fourth International Conference on Knowledge Discovery and Data Mining. The competition task is a regression problem where the goal is to estimate the return from a direct mailing in order to maximize donation profits.

### Usage Notes

The KDD-CUP-98 data set and the accompanying documentation are now available for general use with the following restrictions:

1. The users of the data must notify Ismail Parsa ([iparsa@epsilon.com](mailto:iparsa@epsilon.com)) and Ken Howes ([khowes@epsilon.com](mailto:khowes@epsilon.com)) in the event they produce results, visuals or tables, etc. from the data and send a note that includes a summary of the final result.
2. The authors of published and/or unpublished articles that use the KDD-Cup-98 data set must also notify the individuals listed above and send a copy of their published and/or unpublished work.
3. If you intend to use this data set for training or educational purposes, you must not reveal the name of the sponsor PVA (Paralyzed Veterans of America) to the trainees or students. You are allowed to say "a national veterans organization"...

For more information regarding the KDD-Cup (including the list of the participants and the results), please visit the KDD-Cup-98 web page at: <http://www.epsilon.com/new>. While there, scroll down to Data Mining Presentations where you will find the KDD-Cup-98 web page.

Ismail Parsa  
Epsilon  
50 Cambridge Street  
Burlington MA 01803 USA  
TEL: (781) 685-6734  
FAX: (781) 685-0806

#### Information files:

- [readme](#). This list, listing the files in the FTP server and their contents.
- [instruct.txt](#). General instructions for the competition.
- [cup98doc.txt](#). This file, an overview and pointer to more detailed information about the competition.
- [cup98dic.txt](#). Data dictionary to accompany the analysis data set.
- [cup98que.txt](#). KDD-CUP questionnaire. PARTICIPANTS ARE REQUIRED TO FILL-OUT THE QUESTIONNAIRE and turn in with the results.
- [valtarget.readme](#). Describes the valtarget.txt file.

#### Data files:

- [cup98lrn.zip](#) PKZIP compressed raw LEARNING data set. (36.5M; 117.2M uncompressed)
- [cup98val.zip](#) PKZIP compressed raw VALIDATION data set. (36.8M; 117.9M uncompressed)
- [cup98lrn.txt.z](#) UNIX COMPRESSED raw LEARNING data set. (36.6M; 117.2M uncompressed)
- [cup98val.txt.z](#) UNIX COMPRESSED raw VALIDATION data set. (36.9M; 117.9M uncompressed)

요청하여 얻은 데이터 중 미국 컴퓨터 협회 데이터 마이닝 학회에서 제공하는 데이터를 활용했고, 데이터 Column 명은 다음과 같다.

1. ZIP : Zipcode (a nominal/symbolic field)
2. AGE Overlay Age 0 = missing
3. STATE : State abbreviation (a nominal/symbolic field)
4. INCOME : HOUSEHOLD INCOME
5. GENDER : Gender
  1. M = Male
  2. F = Female
  3. U = Unknown
  4. J = Joint Account, unknown gender
6. RAMNTALL : Dollar amount of lifetime gifts to date
7. NGIFTALL : Number of lifetime gifts to date
8. MINRAMNT : Dollar amount of smallest gift to date

01 개인신용등급별 인원분포

신용등급	2017년 3월	2017년 6월	2017년 9월	2017년 12월	2018년 3월	2018년 6월
1등급	10,526,412	10,755,042	10,972,327	11,189,431	11,471,196	11,693,031
2등급	7,834,362	7,870,832	7,900,990	7,914,238	7,940,455	7,996,002
3등급	3,380,806	3,395,069	3,404,880	3,405,575	3,376,218	3,396,729
4등급	6,301,541	6,287,949	6,319,070	6,326,688	6,643,768	6,666,483
5등급	7,374,528	7,253,270	7,125,426	7,027,257	6,718,378	6,646,940
6등급	5,070,370	5,092,554	5,108,206	5,152,183	5,174,011	5,185,271
7등급	1,395,154	1,362,137	1,335,672	1,279,462	1,251,389	1,235,576
8등급	1,264,307	1,253,905	1,247,619	1,241,936	1,236,060	1,202,134
9등급	1,306,554	1,281,986	1,266,178	1,243,579	1,227,773	1,199,074
10등급	370,512	366,905	358,952	364,729	376,330	377,146
전체	44,824,546	44,919,649	45,039,320	45,145,078	45,415,578	45,598,386

\* 2011년 10월 이후 신규 서비스 개시

출처 : NICE 평가정보

데이터 신용등급에 대한 정보가 없으므로 'INCOME' 칼럼 정보를 바탕으로 임의로 신용등급을 추가하고 ZIP 칼럼을 임시 대체키로 활용하였다. (위 연구 과제에서 SHA-256과 MD-5 알고리즘을 사용하였지만, 실제에서는 어떤 알고리즘을 사용했는지 알기 어렵다)

```

1 import csv
2 import hashlib
3
4 cin = open('cup.csv', 'r', encoding='utf-8')
5 rdr = csv.reader(cin, delimiter=',')
6
7 # print("##### 원본 데이터 출력 #####")
8 #
9 # for line in rdr:
10 #     print(line)
11
12 rows = [row for idx, row in enumerate(rdr) if idx != 0]
13
14 num_rows = len(rows)
15
16 print("전체 칼럼 수 : ", num_rows)
17
18
19 sensitive_records = set()
20 quasai_list = []
21 quasai_column = []
22 for col in rows:
23     quasai_column = [col[0], col[1], col[2], col[3], col[4], col[5], col[6], col[7]]
24
25     if quasai_column not in quasai_list:
26         quasai_list.append(quasai_column)
27
28
29
30
31
32 sha_256 = hashlib.sha256()
33 md_5 = hashlib.md5()
34
35 for line in quasai_list:
36     sha_256.update(line[0].encode('utf-8'))
37     m = sha_256.hexdigest()
38     md_5.update(m.encode('utf-8'))
39     line[0] = sha_256.hexdigest()
40
41
42
43 # print("\n##### 수정된 데이터 출력 #####")
44 #

```

Run ARX

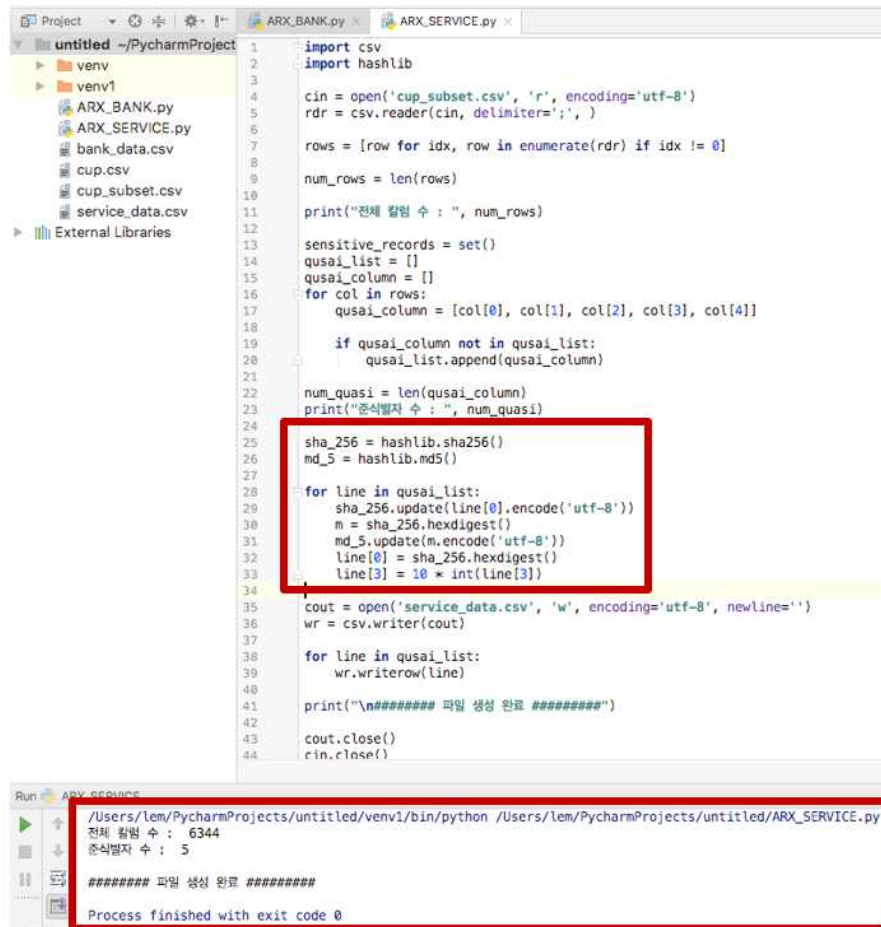
/Users/lem/PycharmProjects/untitled/venv/bin/python /Users/lem/PycharmProjects/untitled/ARX\_BANK.py  
전체 칼럼 수 : 63441

##### 파일 생성 완료 #####

Process finished with exit code 0

bank_data.csv	AGE	GENDER	GRADE	STATE	RAMNTALL	NGIFTALL	MINRAMNT
ZIP							
26cf5165946e8846d48d3cc9e797e82f4a859c18b9a5899b824d641fba2afcd	46	Male		4 CA	47	3	10
cd4a884e38762e8b9335b091651bd79101b84d1024fe7609bad364bf29c7793	70	Female		9 CA	109	16	2
6b12521990013a82106c15e68d827e68566c5ad6eb5f556c2d8bea719dbf2f6e	78	Female		7 FL	254	37	3
3b69ed027315945a075812ad18ecd8dd8cafc8d6720ff61e6153f37cbb2283e	38	Female		6 IN	107	14	3
eb11b4d9f6413e54bbca39f3362d0062210a81149500e9144e716de9f42aa932	75	Male		9 IN	25	2	10
a14f4edf141a185e3ec3174abc080544ba2ea6751904e7515708b60ad65592e	72	Female		6 MN	48	9	4
0b11b59f9248f98dce8fa21176aa0b923ed908b22e029057f22d844500bbf82c	70	Male		6 UT	74	7	5
ca6c77c2bbb8821c6ee8e2c6e0bf65ad3d78a9dd443478752a0a7587449f47b5	44	Male		9 MI	37	3	10
480dc3b799ba91027a937a2a88e99601a8f26fa3d9d0a2e3f9cb46801fb6ec	46	Female		3 IL	100	4	15
d0fb20a734eac223c0c1b459262109b25b7786ee3da70f6cf04dbdeaa814ee7	62	Female		6 FL	81	10	5
c3b1063b4a248bd136cf549cf07f02ea97ef31326c388bacefd601676a963bc	62	Female		7 MN	61	15	2
633c0e9262104b27224664d5ac0a1ba8d58fb326ebe143a382940027e32110f9	82	Male		8 MI	25	1	25
139417733f74e629f747f690d8c1eb0d8c659ec4eca89b753afd9c6133d00da3	46	Female		3 MN	36	4	5
f873890310ab72e1157295350d3d46d963950cec26c9055a688e175411392a66	54	Male		3 CA	107	6	10
dc8f8522f584f983b6190b9eb8304127e8bb400e557352ee9e15c794d5d51d	42	Female		3 IL	50	1	50
fe91dabdd118f6dc2e6206578f3c4768e0c603f2e0d738a3ee277e6081aa32d	84	Male		3 MO	71	8	3
0fd0c68175f5dcca360acd00d9069471a46543c12c8627caff4ab83c0c3227e	49	Female		3 TX	102	12	5
c062aa4c1f4343ee86f671681783f3137b9fcc7a90c648dce5d6b0110e392c74	38	Female		6 IL	57	7	5
2b6a02509d06ec10053b204dc9b7b4bf1e088c765297d08436d3f1f57919	72	Male		9 MO	40	2	20
9488942c73c1c0795d032d6d10fa293f8f8c587b0bf975f98d9ccbf633e6c83c	84	Female		6 NC	40	2	15
bcc1e88956932e113f8f8ca02d091ba7903509968da83475bb6fec6227f6b89c	69	Female		6 FL	220	13	5
b9bfe7f599e9e608c1247519f7b8e210fafec603368e6356745e365df293d9b2	69	Male		6 TX	25	1	25
24e89bf1fb9c86f63d31986793b03a0e1da4ec5109b9739f74bfc0ef78192c7f	88	Female		5 CA	291	18	5
8540b28b7c9f60db18c54c64c71d990b23294071c9f6ac58854df8451367769f	75	Female		4 MN	96	8	10
1c43ea75db56a2120e1243f3c4373def9844e8a62369cb7cfb4cd12ae8ccb84	84	Female		7 MN	94	10	3
ade105c8b06ae1f763c6db8bd4c750f30c7d74d83a30305392869a6512476fbfb	30	Male		3 WA	30	2	10
3cca75e43e754be5cc6b98e4459c7c8798048b766e275cf02e168b97b25e8e9e	44	Male		3 FL	68	4	10
887d57456748c98277e52ebb6ab00804f010cb8feb440a577db4d97968881230	51	Female		5 WI	338	29	5





```

1 import csv
2 import hashlib
3
4 cin = open('cup_subset.csv', 'r', encoding='utf-8')
5 rdr = csv.reader(cin, delimiter=';', )
6
7 rows = [row for idx, row in enumerate(rdr) if idx != 0]
8
9 num_rows = len(rows)
10
11 print("전체 칼럼 수 : ", num_rows)
12
13 sensitive_records = set()
14 quasi_list = []
15 quasi_column = []
16 for col in rows:
17     quasi_column = [col[0], col[1], col[2], col[3], col[4]]
18
19     if quasi_column not in quasi_list:
20         quasi_list.append(quasi_column)
21
22 num_quasi = len(quasi_list)
23 print("준식별자 수 : ", num_quasi)
24
25 sha_256 = hashlib.sha256()
26 md_5 = hashlib.md5()
27
28 for line in quasi_list:
29     sha_256.update(line[0].encode('utf-8'))
30     m = sha_256.hexdigest()
31     md_5.update(m.encode('utf-8'))
32     line[0] = sha_256.hexdigest()
33     line[3] = 10 * int(line[3])
34
35 cout = open('service_data.csv', 'w', encoding='utf-8', newline='')
36 wr = csv.writer(cout)
37
38 for line in quasi_list:
39     wr.writerow(line)
40
41 print("\n##### 파일 생성 완료 #####")
42
43 cout.close()
44 cin.close()

```

Run ARX\_SERVICE

```

/Users/lem/PycharmProjects/untitled/venv1/bin/python /Users/lem/PycharmProjects/untitled/ARX_SERVICE.py
전체 칼럼 수 : 6344
준식별자 수 : 5

##### 파일 생성 완료 #####
Process finished with exit code 0

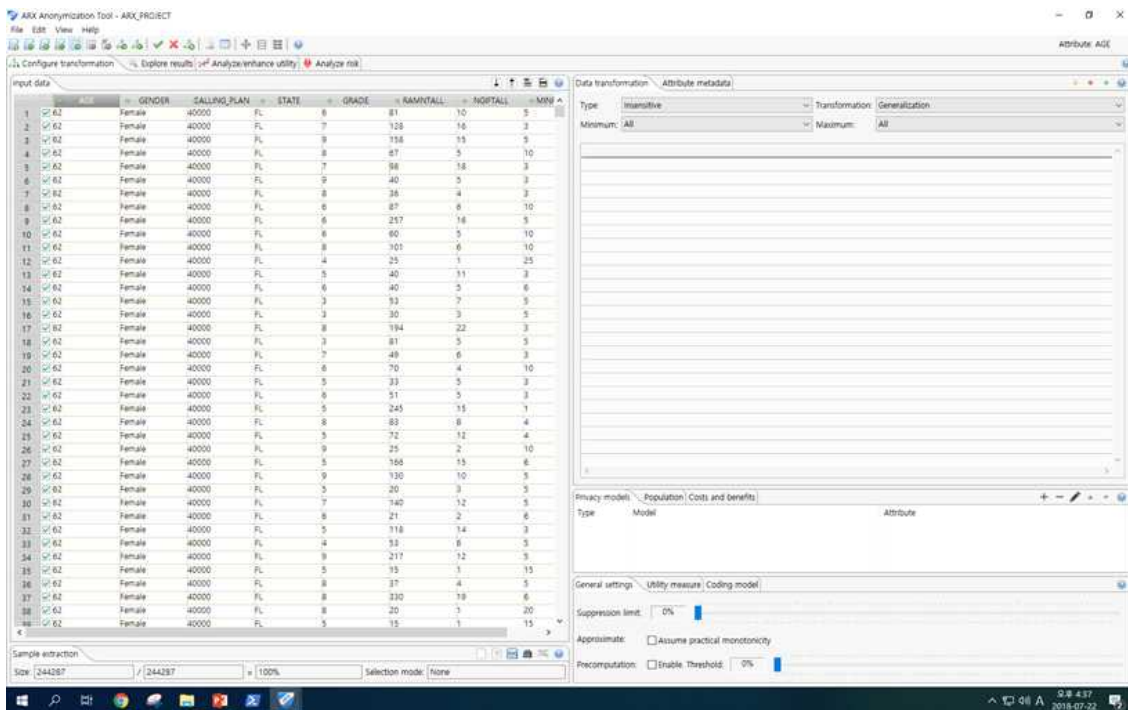
```

service_data.csv				
ZIP	AGE	GENDER	CALLING PLAN	STATE
627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2		62 Female	40000	FL
789c715f50245801fe3604782f93036bbf9a99ae1e57e87db8599fd82fc0f406		88 Male	30000	IN
567af847e60dd8828cd327554040abdb5e58d02d7b008ef659382298a73d1240		62 Female	30000	AL
6f9d6ed2b153cb2772ced4be69d628d14bbffcf47219cd9a9456a16bec72d52d		80 Female	50000	GA
d4b5970291c84fe17160e3e76228994b83da7babc7d6c13f3628ed6e68b6cfb0		84 Male	20000	IN
59a94fecf0333bcc36c228cbf84832146d4a9c34023b791247bd2248d0b14625		72 C	40000	SC
502aafafa1c0d8decea5c92ab07c37ac8d29e5125b748eff6f907a36535c41a4		63 Male	30000	KY
b4faa7ced28ca9f32e113ff9cbfb0aecf6b8af4d3eca08689f12a221def6599		73 Female	10000	CA
991e46508ba672dca194caf2975abfb7788af9fd54b3247831ab44bdaff043		75 Male	40000	TX
f838ac04defd87271a107c3f6241de9ce916e45402d96c5c616b7daf08508acc		73 Female	20000	FL
345eab4ce7e1987411c715cb4cb4bda465881de142e9a9ad8cbb910fa9928c9f		88 Female	10000	FL
3ee44953da1d1b4c4507efa041ed713a15387d135f25b82039adf9516c58f6		48 Female	10000	CA
1580f6c68098fb52595e4732048b9b3cd8e57126540d66488043a2257c83d140		64 Male	70000	WA
458b49298f80e090fd00e567fe85119544151c73b110b249faee86aca821a439		41 Male	40000	WA
9a3f779210d49c63182e0937d0e503ee2bab3585d9b9173596f2f33dc7142882		36 Female	60000	CA
f68b94e6b53209add701cac05a2654ff507fe04a988542d83ac3213c4058072a		53 Female	30000	CA
b97a9a34ff48e2a3111a3e284742759b378df747b72db076062bdfb39d004c268		82 Female	20000	CA
da90d686eacfa2cee9465269ca8de9136a10a6b34cf1175b9be5b13143759903		70 Female	60000	CA
5ce30d5fb5bc2acac4eab56edd09918f98e85d494cb113d2c327c99de3a7a57		58 Female	10000	WI
ea3502de63d8f9f5ecfe710b077d666f4c5dbe39706b3f314529bee89b55a93e		65 Male	30000	TX
1129cd47236e3421e5396d5b71eaa672dd8e4da6b5cf89cb52305285e042ba7e		49 Male	50000	NC
fc68acac37c7048cd70c18c870a0b3bfd2e24d2dc96f98b8d690b02a5eee340b		42 Male	70000	NC
0b128a5ecb425d49a06a3971271ff74f88936caa430035acd4f3d32e5b01e2be		48 Male	40000	IL
c05bc4a9b0965ed58a914ab54f66e3b753dc0c75407fd2eb37907e473bc1d769		56 Female	50000	CO
ba3fbb1b2648910271bd90af10145bd592b896ebfb885a26a535eed5f749cbb9		49 Male	50000	IN
cefd47f8381deac03d8da167d1bf28880e48f17d86fbd1e1cd7ba7d52515567		62 Male	70000	AZ
c0100eb404e121accab0f54fc23eec7e380c350442557cb7c9cc4cfb84e6c80		71 Male	20000	MI
cb4af20aa7c00d91e7442803a42c0e6130b9ede7c005571aa2632be3a63a189		31 Female	70000	MI

임시 대체키를 이용하여 두 데이터를 결합하였다.

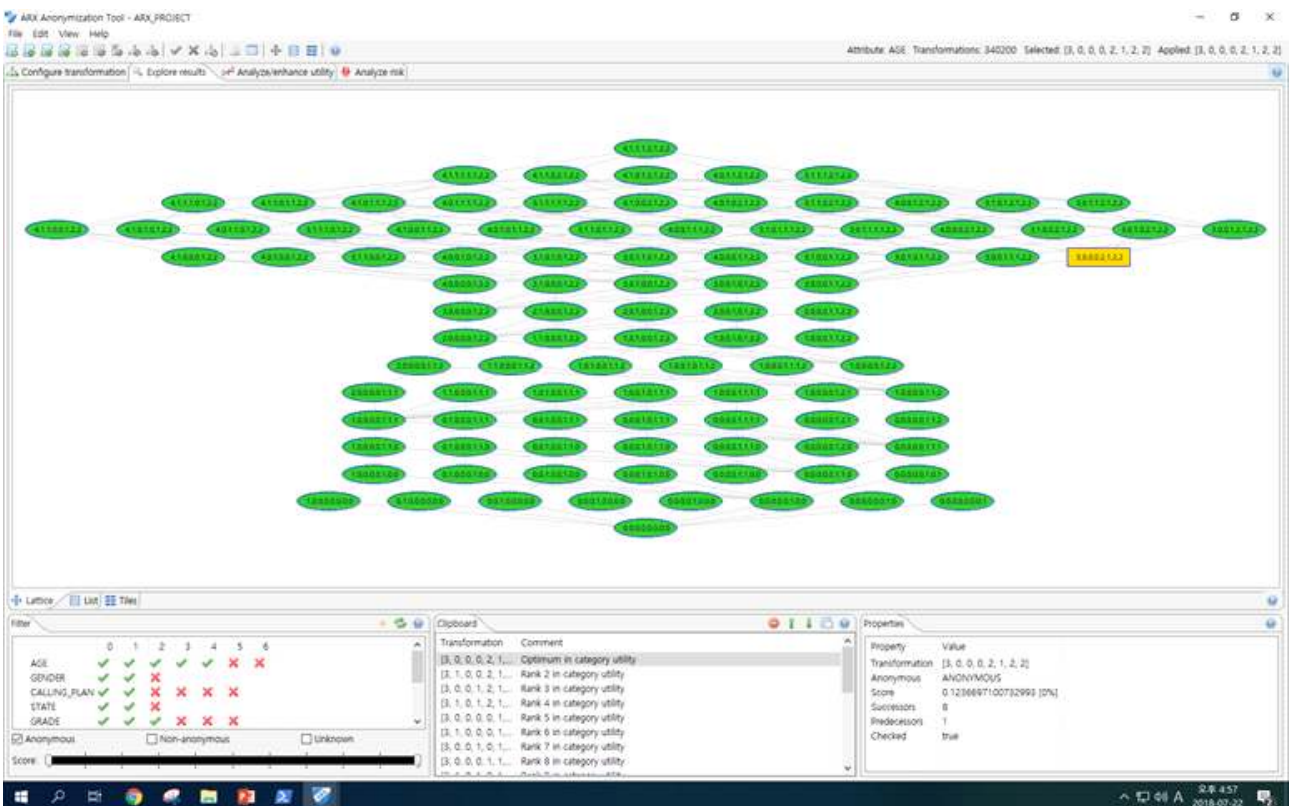
그 후, 임시 대체키를 제거하고 데이터 비식별화를 위해 ARX tool에 import 하였다.

	A	B	C	D	E	F	G	H	I
1	ZIP	AGE	GENDER	CALLING_PLAN	STATE	GRADE	RAMNTALL	NGIFTALL	MINRAMNT
2	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		6	81	10	5
3	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		7	128	16	3
4	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		9	158	15	5
5	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		8	67	5	10
6	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		7	98	18	3
7	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		9	40	5	3
8	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		8	36	4	3
9	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		6	87	6	10
10	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		6	257	16	5
11	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		6	60	5	10
12	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		8	101	6	10
13	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		4	25	1	25
14	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		5	40	11	3
15	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		6	40	5	6
16	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		3	53	7	5
17	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		3	30	3	5
18	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		8	194	22	3
19	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		3	81	5	5
20	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		7	49	6	3
21	627172078097cf37b30d60a0c086865518af4c9349e3b33cd35b4bde0cc423e2	62	Female	40000 FL		6	70	4	10

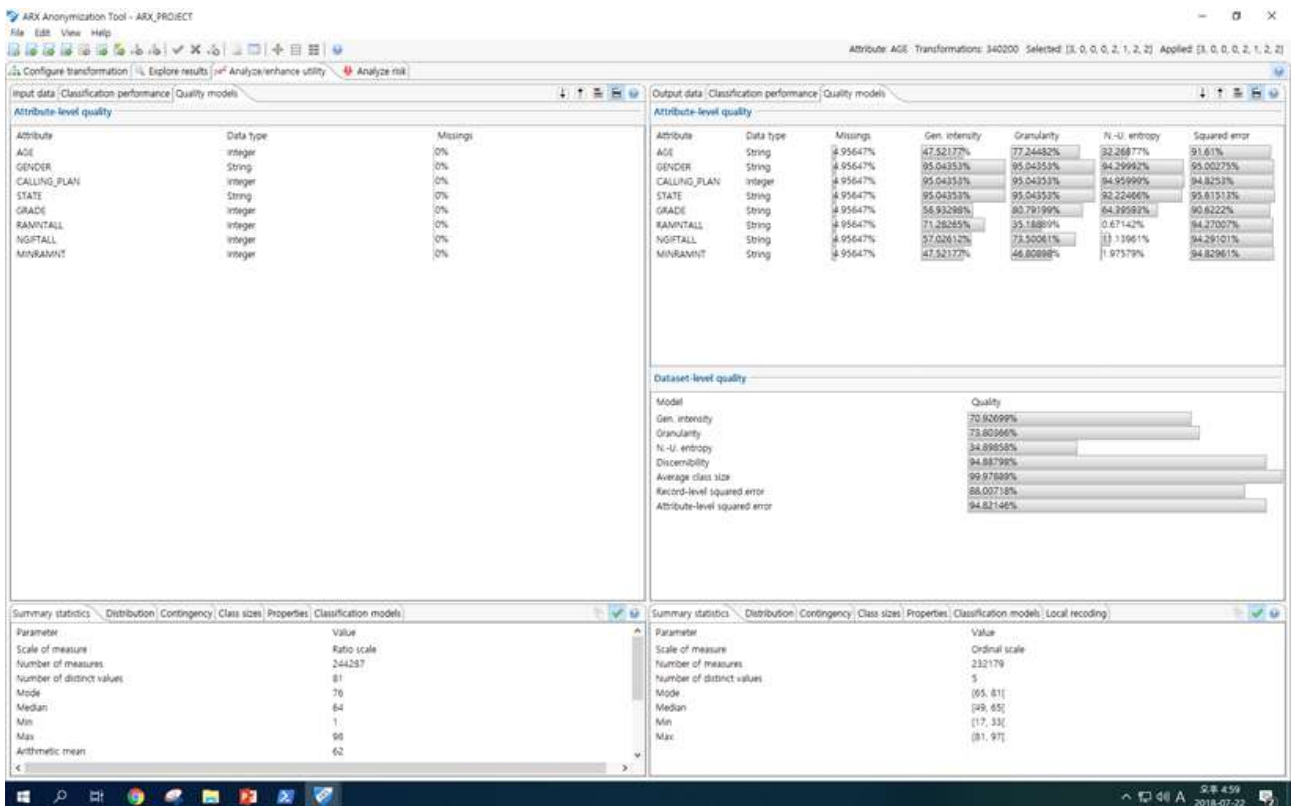
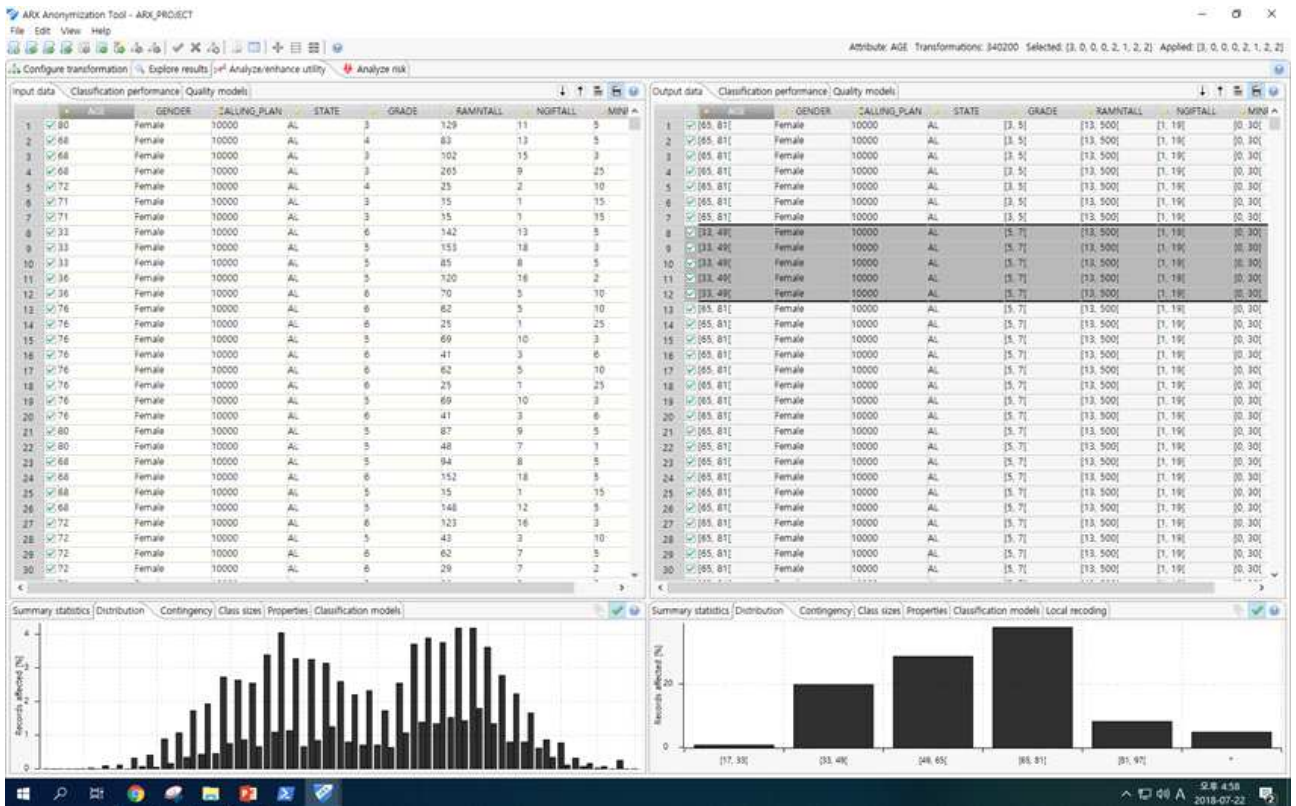


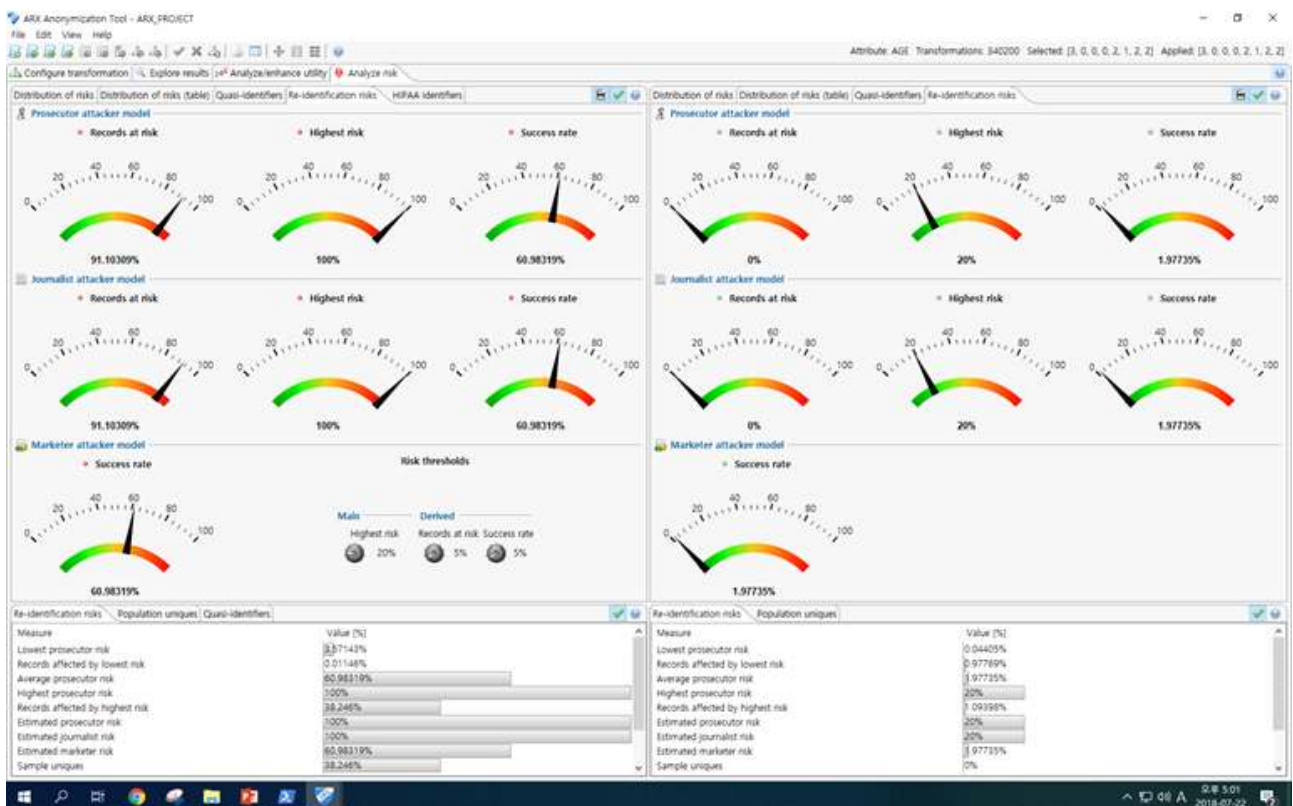
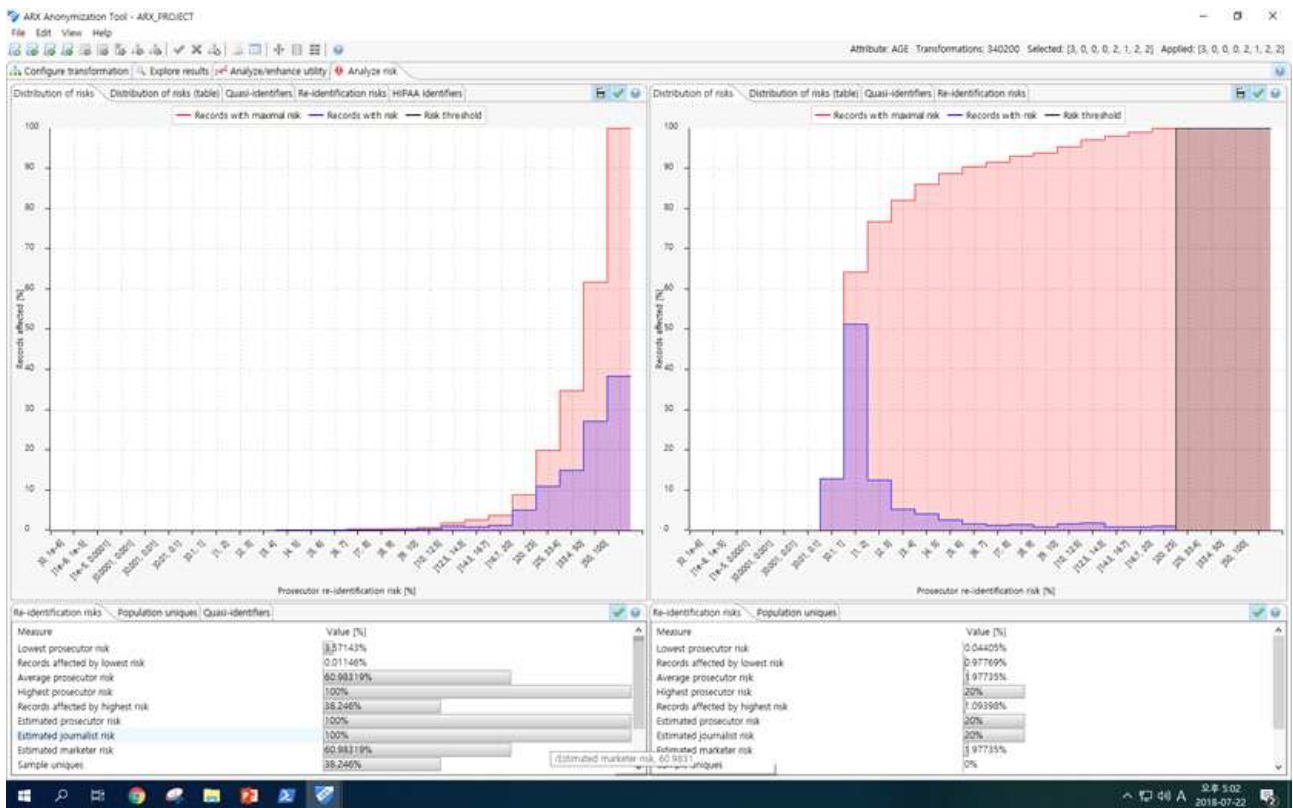
위 툴을 사용하여 K-익명성, I-다양성, T-근접성 이외에도  $\delta$ -presence, Differential Privacy 등 다양한 비식별화 조치를 할 수 있다.

위 프로젝트는 VGH 기법을 이용한 K-익명성 조치를 하였고 그 결과에 대한 최적을 찾고 데이터 손실 및 위험도 등에 대해 분석하였다.









## 4. 기대효과

### 4.1. 기대효과

데이터 기반 금융혁신을 촉진하고 관련 산업의 발전에 이바지한다. 위 예에서는 통신 관련 산업로 한정하였지만, 병원 등의 의료, 자율주행차 등의 제조 산업 분야와 데이터 결합을 통해 융합 신산업 성장을 촉진한다.

소비자 중심 금융 데이터 활용을 통해 소비자에게 혜택이 돌아간다. 금융 데이터로 소비자를 다각적으로 분석하여 더 좋은 서비스가 가능한 환경을 마련한다.

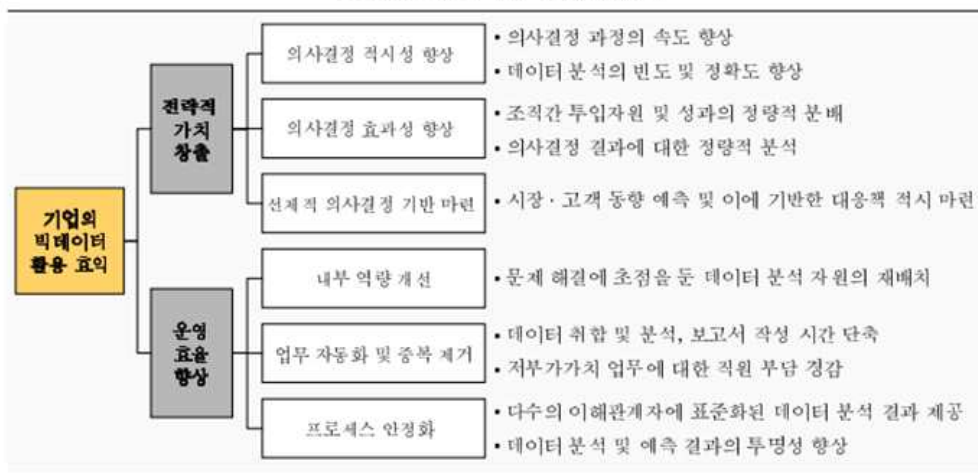
데이터 생태계를 형성하여 중소기업, 핀테크 스타트업에 기회를 제공하는 금융시스템의 공정성을 강화한다. 풍부한 데이터 인프라를 통해 정보독점이 완화되고 정보격차가 해소된다.

다양한 개인정보에 대한 자기 결정권이 도입되고 정보보호가 실질적으로 강화된다. 프라이버시 위험에 대응하여 정보 주체를 보다 실질적으로 보호할 수 있다.

#### ■ 기업의 빅데이터 활용 효과

- 기업은 빅데이터의 활용을 통해 의사결정의 적시성과 효과성을 높이고 나아가 선제적인 의사결정의 기반을 마련할 수 있음. 또한 내부 역량 향상, 업무 자동화 및 중복 제거, 프로세스의 안정화 등을 통한 생산성의 증가를 기대할 수 있음

[그림 7] 기업의 빅데이터 활용 효과



출처 : KB 경영연구소

#### <주요과제 추진일정>

추진 과제		추진 일정
금융분야 빅데이터 활성화	① 빅데이터 분석·이용의 법적인거 명확화	'18.上 신청법 개정 추진
	금융정보 DB 분석·이용	'18.下 서비스 실시 및 신청법·보험업법 시행령 개정 추진
	② 빅데이터 인프라 구축·운영	'19.上 시범서비스 실시
	③ CB사·카드사의 시장선도 역할 강화	CB사 '18.上 신청법 개정 추진 카드사 '18.下 카드업계 간담회 후 부수업무 신고 등 추진

출처 : 금융위원회

## 4.2. 마무리

대부분 사람은 개인정보 노출에 대해 둔감하다. 지문을 노출하는 지하철 광고 등에서 예를 쉽게 발견할 수 있다. 하지만 비식별화된 데이터를 상업적으로 활용하고 제3의 단체에 공유하는 행위에 대해 불신한다. 2017년 11월, KISA 등 개인정보 비식별화 전문기관 및 20개 기업 고발당한 사례가 대표적이다. 이런 사람들의 인식은 개선이 필요하다. 사람들에게 개인정보의 보호 필요성을 알리고 이를 비식별화한 데이터를 활용하여 사람들에게 혜택을 제공하여 데이터 비식별화에 대한 가능성을 보여줘야 한다.

다양한 비식별화 모델 및 랜덤 임시키를 사용하여 본 식별자와의 연관성을 완전히 차단하는 기법데이터를 소개하고 활용을 위해 다양한 비식별화 모델 적용 및 유럽의 GDPR과 일본의 개정된 개인정보 보호법을 국내에 비식별화 정보에 적용하여야 한다.

현재 '개인정보 비식별 조치 가이드라인'은 산업계에서 비식별 조치의 까다로움과 그 복잡한 절차 및 비용, 그리고 비식별화 조치에 따라 손상된 데이터의 활용가치에 대해 우려가 있다. 반면, 민간 소비자 단체와 시민단체는 개인의 정보를 목적과 달리 활용하고 제3의 단체에 공유하는 행위에 대한 근본적인 우려 및 가이드라인에 기술된 비식별 조치기준의 효과에 대한 불신이 있다. 위 견해차를 해결하기 위해 산업계에는 비식별화의 필요성에 대한 교육을 시민에게는 비식별화 조치된 데이터에 안정성 및 혜택에 관한 교육 및 세미나를 제공해야 한다. 또한, 기술적으로 다양한 비식별화 모델 도입을 해야 한다.

### ■ 빅데이터 시대에 대응하기 위한 금융기관의 적극적인 준비 필요

- 금융업은 타 산업 대비 기업의 데이터 보유량이 많고 빅데이터의 활용 잠재가치가 높은 것으로 평가되므로 이에 대해 충분한 **전사적 공감대 형성** 필요
- 빅데이터의 활용에 대한 **장기적이고 종합적인 로드맵** 수립 필요
  - 빅데이터는 신규 투자를 수반할 수 있으며, 단기간에 어느 한 부서에서 추진되는 일회성 프로젝트가 아니므로 장기적 로드맵에 따라 추진되어야 함
  - 영업, 마케팅, 고객응대, 재무, 리스크관리 등 경영활동의 다양한 영역에서 빅데이터 활용을 통한 개선 기회 검토
- 빅데이터 관련 **역량의 단계적 확보** 필요
  - 빅데이터의 올바른 활용을 위해 인력, 기술, 시스템 등의 고도화가 필요하며, 활용 목적에 따라 우선적으로 필요한 역량의 단계적 확보 필요
  - 특히 구조화되지 않은 대규모 데이터 속에서 숨겨진 통찰을 찾아내야 하는 빅데이터 환경에서는 데이터 분석가의 역할이 매우 중요하므로 우수 인재 양성에 힘써야 함
- 빅데이터 **관련 규제 모니터링 및 대응** 필요
  - 최근 여신전문금융업법 개정안<sup>15</sup>에 따라 그동안 신용카드사가 고객 서비스 차원에서 비영리로 제공해온 빅데이터 관련 서비스를 수익사업으로도 추진 가능해지는 등 최근 빅데이터 관련 규제가 구체화될 움직임
  - 국내보다 빅데이터 도입이 앞선 유럽 등에서 빅데이터 관련 규제가 점차 세분화되고 있는만큼, 국내외 빅데이터 관련 규제 동향에 대한 지속적 모니터링과 대응이 필요

출처 : KB 경영연구소



## 5. 참고자료

### 5.1. 출처

-[http://jobnjoy.com/portal/joylife/campus\\_life\\_view.jsp?nidx=130827&depth1=2&depth2=1&depth3=3](http://jobnjoy.com/portal/joylife/campus_life_view.jsp?nidx=130827&depth1=2&depth2=1&depth3=3)  
일단 쓰고 '배짱 연체'...20대 밀린 통신비만 500억 한국 경제 매거진

-[http://www.hani.co.kr/arti/economy/economy\\_general/726956.html](http://www.hani.co.kr/arti/economy/economy_general/726956.html)  
'통신비 연체' 신용 추락' 20대 최다 SKT도 채무불이행 등록 중단 밝혀 한겨레

-[https://www.kbfg.com/kbresearch/index.do?alias=report&viewFunc=default\\_details&categoryId=1&boardId=&rBoardId=101&articleId=1002377](https://www.kbfg.com/kbresearch/index.do?alias=report&viewFunc=default_details&categoryId=1&boardId=&rBoardId=101&articleId=1002377)  
금융업의 빅데이터 활용 KB 금융지주 경영연구소

-[http://news.khan.co.kr/kh\\_news/khan\\_art\\_view.html?artid=201711091626001&code=940100](http://news.khan.co.kr/kh_news/khan_art_view.html?artid=201711091626001&code=940100)  
12개 시민단체, 고객정보 무단 결합해 데이터로 만든 공공기관·기업 고발 경향신문

-[https://www.eprivacy.or.kr:40018/UploadAction.do?cmd=download&file\\_seq=1&attach\\_idx=1468302445524\\_1HKHmX8M3f](https://www.eprivacy.or.kr:40018/UploadAction.do?cmd=download&file_seq=1&attach_idx=1468302445524_1HKHmX8M3f)  
2016 비식별화 가이드라인 정보보호인증마크제도

-<https://arx.deidentifier.org/>  
ARX - Data Anonymization Tool

-<http://www.kdd.org/kdd-cup>  
KDD CUP ARCHIVES 샘플 데이터

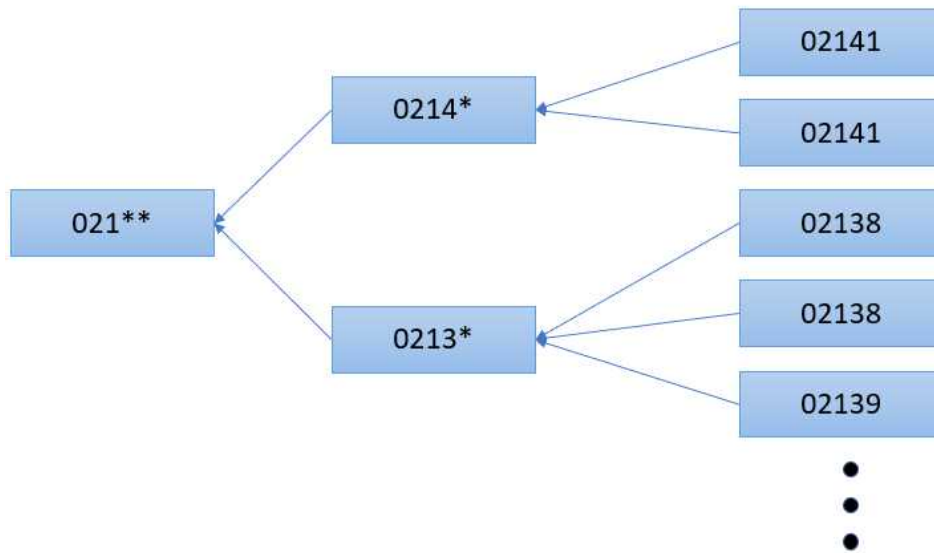
-[http://www.niceinfo.co.kr/creditrating/cb\\_score\\_3.nice](http://www.niceinfo.co.kr/creditrating/cb_score_3.nice)  
개인신용평가 관련 통계자료 NICE 평가정보

-[http://www.fsc.go.kr/info/ntc\\_news\\_list.jsp?menu=7210100&bbsid=BBS0030](http://www.fsc.go.kr/info/ntc_news_list.jsp?menu=7210100&bbsid=BBS0030)  
금융 빅데이터 활용 기대효과 및 향후 추진계획 금융위원회



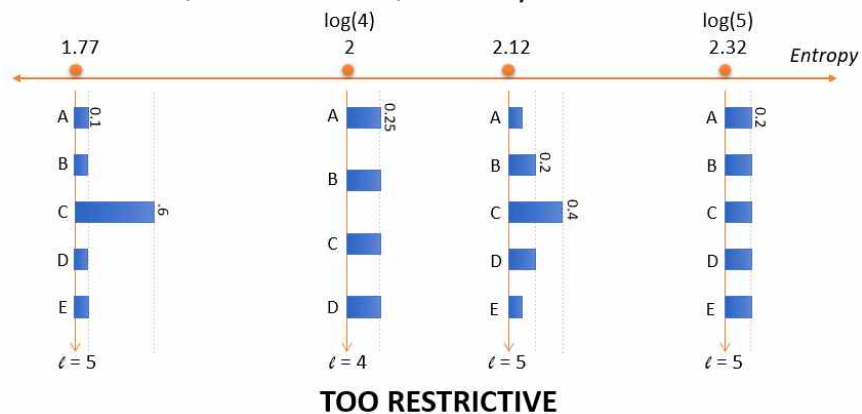
## 5.2. 기타자료

# VGH 기법



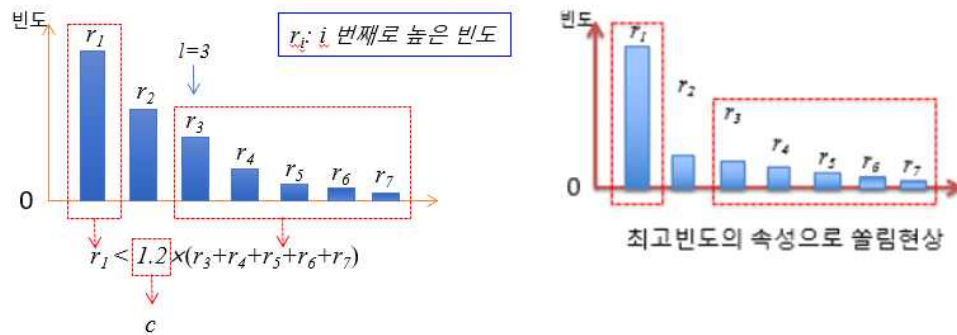
## Entropy l-diversity

- Entropy of table should be  $> \log(l)$
- 장점 : 속성 노출 공격률  $\leq 1/l$

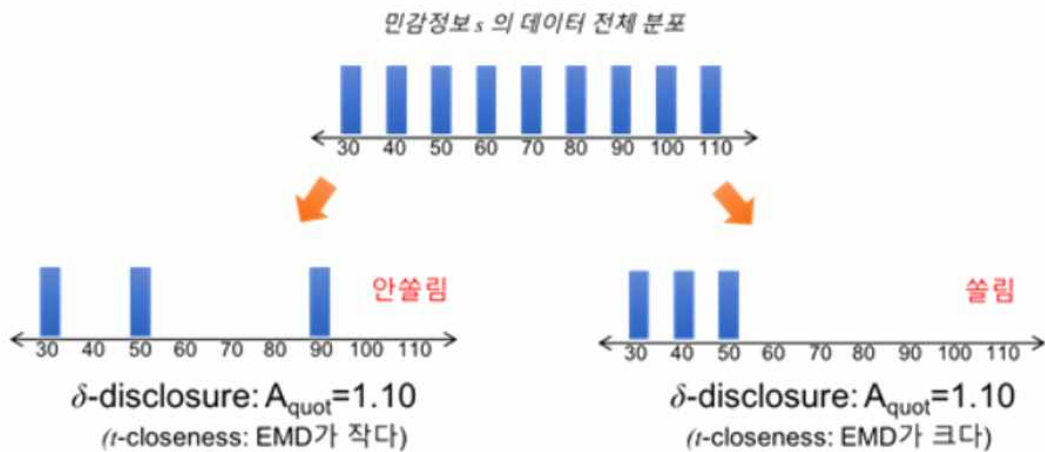


## Recursive ( $c, l$ )-Diversity

- Entorpy 달성이 어렵고 제한적임
- 가장 빈도가 높은 속성과 빈도가 낮은 속성들의 빈도차이를 제한하는 방법



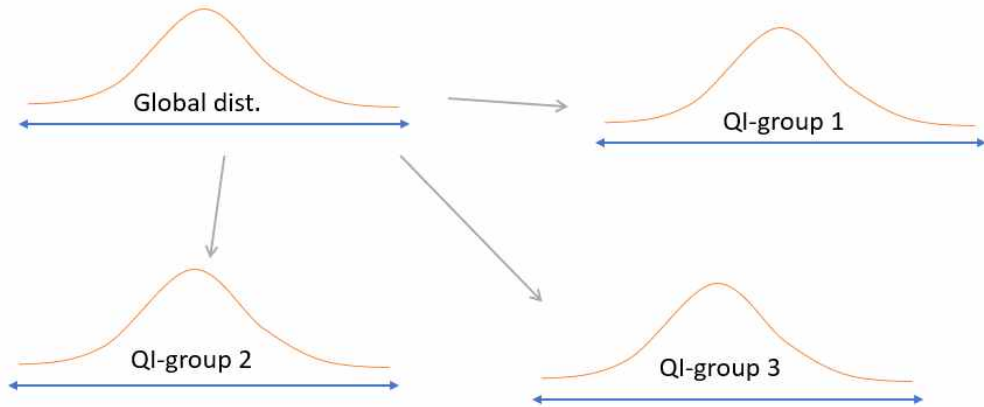
## $\delta$ -disclosure ( $\delta$ -노출)



(그림 3-10)  $\delta$ -노출의 쓸림현상

## t-closeness

- Build QI-groups so that
  - QI-group's SA-distribution is similar to global SA-distribution



## Differential Privacy

- 기본적으로 PPDM 모델에 초점
- Differential Privacy 정의
  - “한 개인의 데이터에의 포함여부와 상관없이 공격자에게 노출되는 정보의 수준은 그 비율적인 측면에서 매우 비슷하게 유지되어야한다.”
- 확률의 비율적 변화량  $< \exp(\epsilon)$
- $\epsilon$ -Differential Privacy

**Definition 2.** A randomized function  $\mathcal{K}$  gives  $\epsilon$ -differential privacy if for all data sets  $D_1$  and  $D_2$  differing on at most one element, and all  $S \subseteq \text{Range}(\mathcal{K})$ ,

$$\Pr[\mathcal{K}(D_1) \in S] \leq \exp(\epsilon) \times \Pr[\mathcal{K}(D_2) \in S] \quad (1)$$

(그림 3-13)  $\epsilon$ -차분 프라이버시의 정의