비식별화기법에 대한 평가와 향후과제

서울대학교 통계학과
임요한

최혜정 (서울대 통계학과, 한국은행), 박민정 (통계청 통계개발원)

발표 개요

1. 들어가기 – 개인정보 보호의 전체 그림 보기

2. 통계적 비식별화조치
   (통계적노출제어, statistical disclosure control)

   - 노출위험(Risk) and 정보유용성(Utility)
   - 통계적노출제어
   - 고차원 자료

3. 재현자료(Synthetic Data)

4. 차등정보보호

5. 정리하기

1. 들어가기

금융빅데이터

금융 빅데이터의 사례

1. 카드회사의 고객 데이터 분석을 통한 마이크로마케팅

2. 카카오페이, 삼성월렛등의 전자결제 자료.

3. 카카오뱅크·K뱅크 등 인터넷 전문은행과 중금리 신용대출

[인베스트조선]

'빅데이터를 활용한 신용평가모델'에 기반하고 있는데, 이 모델이 개인정보 보호 규제를 받아야 하는데다 데이타유효성에 대한 검증도 부족

이를 위해 인터넷 은행 컨소시엄 참여한 금융 · 통신 · 엔터테인먼트 등의 기업들은 각각 보유한 고객정보를 종합해서 "얼마나 대출을 받을지", "얼마나 연체를 하지 않을지" 등을 예상하는 신용평가모델을 구축하기로 했다.

카카오뱅크는 사회관계서비스(SNS) 사용내역, 고객사 상품 결제 내역 등을 활용해서 '카카오스코어'란 모델을 만들 예정이다. 또 K뱅크는 KT의 통신요금 납부 내역, 비씨카드 거래 정보, KG이니시스와 다날 등 결제대행사(PG)의 결제 정보 등의 빅데이터를 활용, 'CSS'(Credit Scoring System)란 신용평가 모델을 구축할 계획이다.

국내에선 개인정보 보호법으로 인해 개인정보 "거래"가 불가능하다. 각 컨소시엄 내부 데이터 외엔 활용하기 어렵다. 활용할 수 있는 데이터 자체가 한정적이다.

금융업권 신용정보가 한 데 모아 관리하는 한국신용정보원이 올초 출범했지만 역할은 기존 금융권의 신용대출 한도, 금리, 연체율, 부도율 등의 데이터를 제공하는 데 그친다.

개인정보보호

개인정보란?

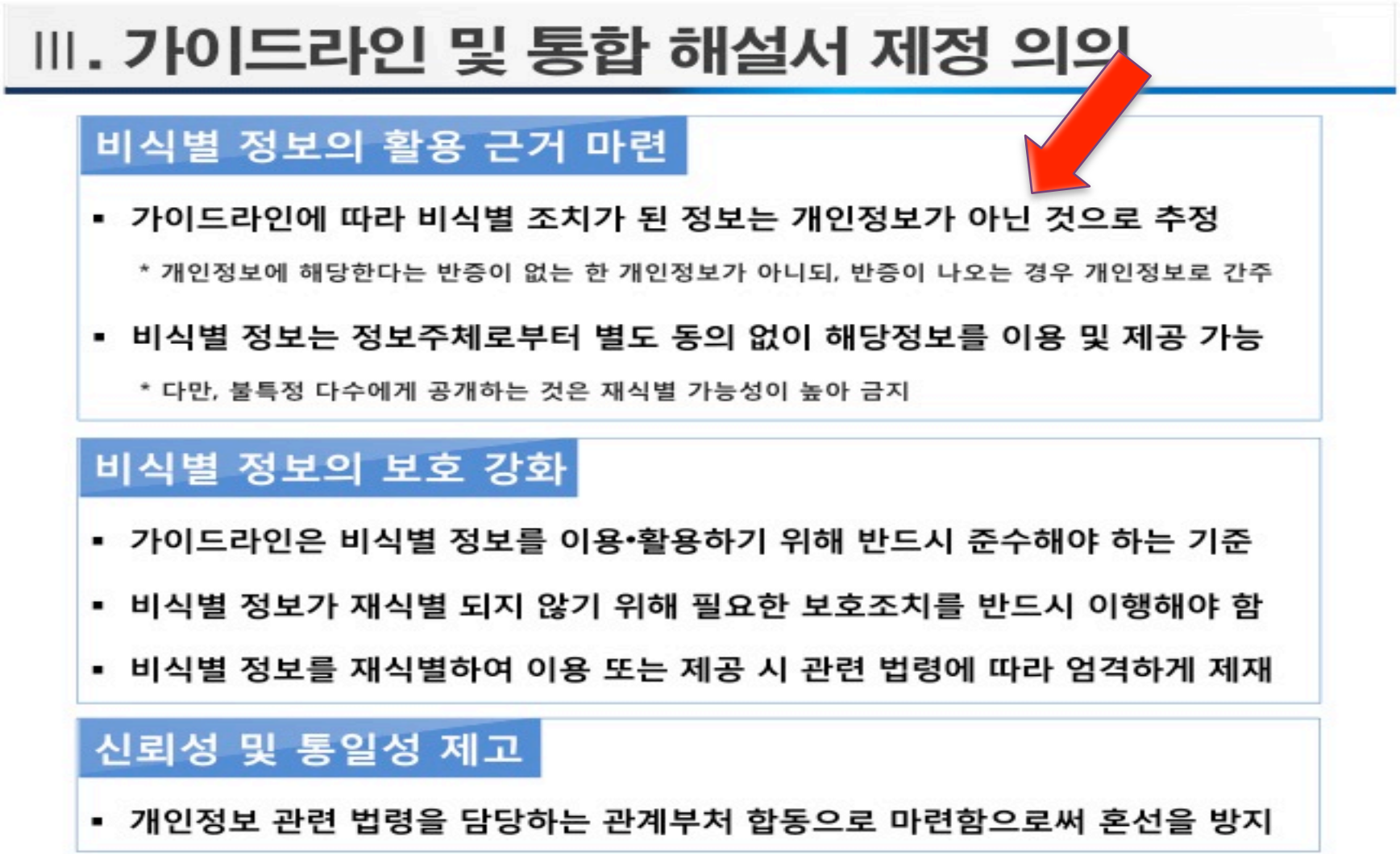
1. 「신용정보의 이용 및 보호에 관한 법률」개정(안): 2. “개인신용정보”란 신용정보 중 생존하는 개인에 관한 정보(사업을 경영하는 개인의 경우 그 사업에 관한 정보는 제외한다)로서 신용정보주체를 식별할 수 있는 정보(그 정보만으로 신용정보주체를 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 식별할 수 있는 정보를 포함한다)를 말한다. (법안 제2조제2호)

2. 「개인정보보호법」: 1. "개인정보"란 살아 있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보(해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것을 포함한다)를 말한다.(법 제2조제1호)

3. 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」: 6. "개인정보"란 생존하는 개인에 관한 정보로서 성명·주민등록번호 등에 의하여 특정한 개인을 알아볼 수 있는 부호·문자·음성·음향 및 영상 등의 정보(해당 정보만으로는 특정 개인을 알아볼 수 없어도 다른 정보와 쉽게 결합하여 알아볼 수 있는 경우에는 그 정보를 포함한다)를 말한다. (법 제2조제6호)

+ 금융위 가이드라인과 k-익명성

금융위원회(2016.06.09) [“개인정보 비식별 조치 가이드라인” 주요내용]에서 발췌

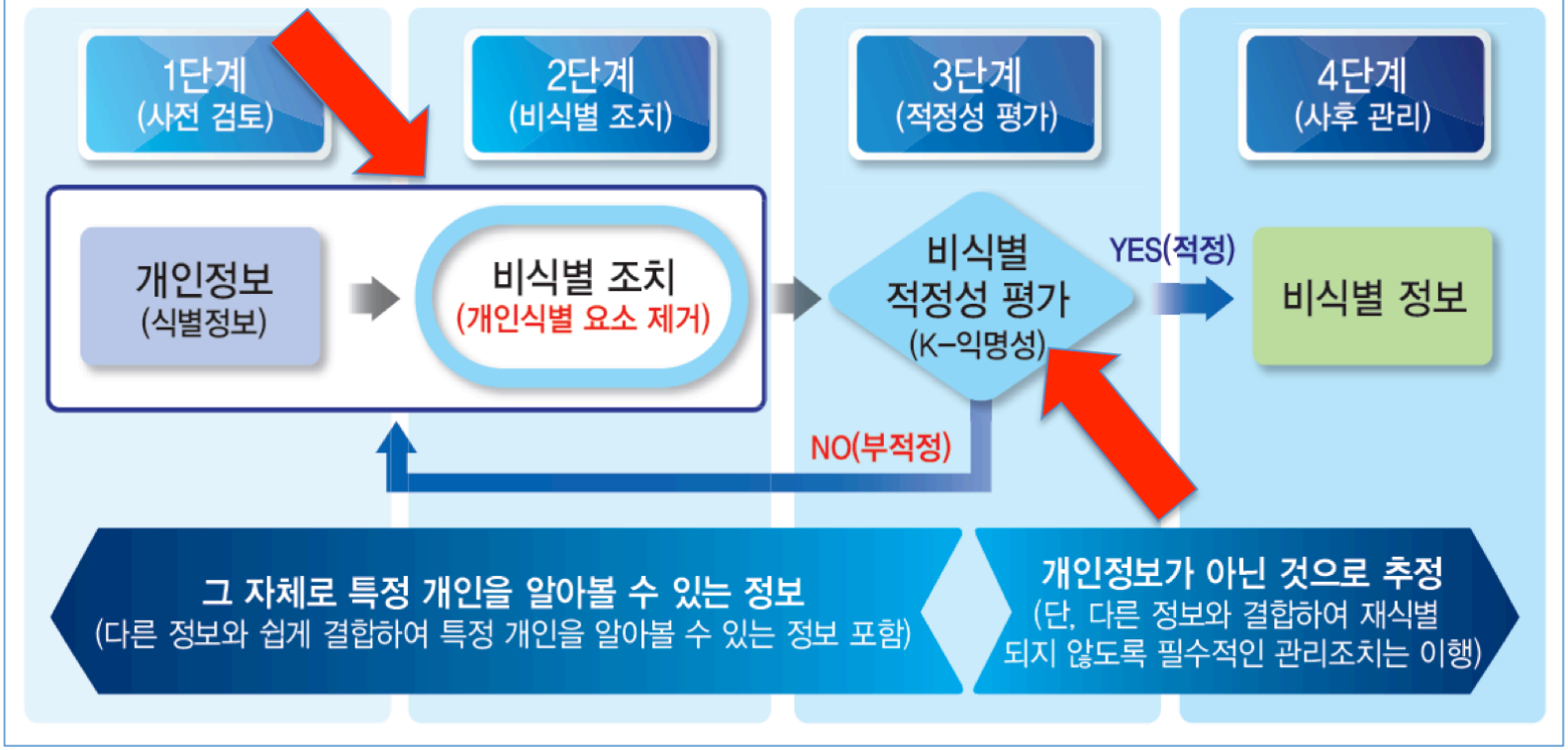


Ⅳ. 가이드라인 주요 내용

1. 비식별 조치 절차 및 기준

(1) 개 요

< 비식별 조치 및 사후관리 절차 >



이용자의 유형과 노출제어의 종류

종류	물리적 노출제어		통계적 노출제어	
이용자	심층이용자		불특정 다수	
전략/기법	접근제어	결과통제	비식별화	재현자료
배포형식	데이터센터		공공이용파일	

출처: 가계금융.복지조사 마이크로테이터 제공을 위한 매스킹 방안 평가 (2013) 박민정외 2인, 통계개발원.

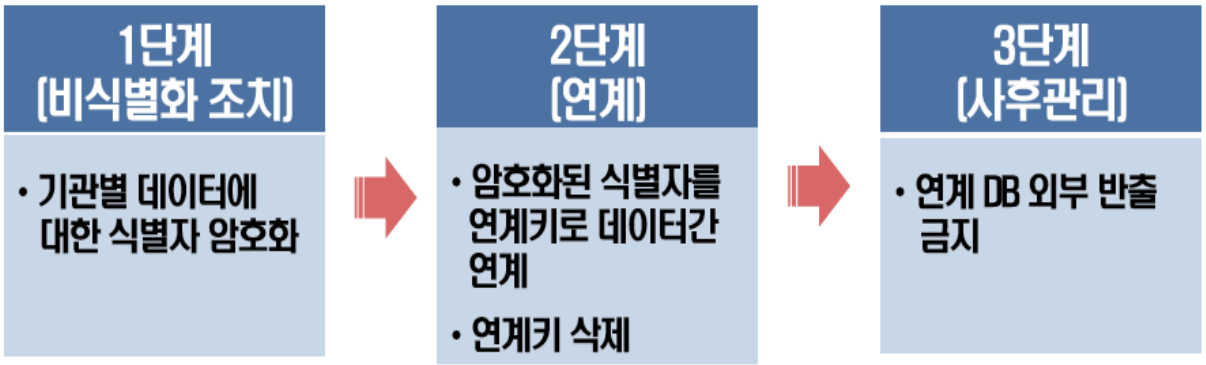
통계청 데이터센터

데이터센터의 제한된 영역에서 작업

연계데이터(분석용 DB)/결과물의 비식별화 조치

KCB-통계청 와의 시범사례 : 신혼부부 신용 DB구축

1. 비식별화 단계



구분	관련 여건	보완 조치 내용
연계를 위한 데이터 제공 양식	• 관련법상 제3자 제공 금지 규정 ⇒ KCB는 연계를 위한 1차 암호화 데이터의 제공에 한계 • 통계법상 통계자료 제공 가능 ⇒ 통계법상 식별할 수 없는 형태로 통계자료를 처리 후 제공(제31조)	• 통계청이 KCB에 데이터 제공 • KCB가 데이터간 연계 및 연계키 삭제 후 분석용 DB 작성 → 양기관 공유 * 분석용 DB는 개인정보가 아님(법률 자문 결과)
데이터 관리 및 분석 방식	• 상호간에 소관 데이터의 유출 우려	• 통계청의 독립된 제한 공간(데이터 센터)에서 자료 분석 및 통계 작성 → 민관기관의 서버를 가지고 입주 • 자료 접근권자 인가

[인용: 통계개발원 보고서]

발표의 범위/내용

이용자의 유형에 따른 관리가 필요

일반이용자 – 공공이용파일 – 개인정보 비식별화

심층이용자 – KCB가 데이터센터로부터 반출하는자료는 연구용 DB/분석결과물로 개인정보 비식별화된 자료

비식별화로 인하여 자료의 유용성 훼손이 심함.

2. K-익명성과 통계적노출제어

기본 개념1: 노출위험(disclosure Risk)

k-익명성, l-다양성, t-근접성

유일성(재식별 가능성)에 근거한 척도

의사결정이론에 근거한 척도

k-익명성, l-다양성, t-근접성: 개인정보비식별화 조치 가이드라인

프라이버시 보호 모델 : 재식별 가능성 검토 기법 * k, ℓ, t값은 전문가 등이 검토하여 마련

기법	의미	적용례
k-익명성	• 특정인임을 추론할 수 있는지 여부를 검토, 일정 확률수준 이상 비식별 되도록 함	• 동일한 값을 가진 레코드를 k개 이상으로 함. 이 경우 특정 개인을 식별할 확률은 1/k임
ℓ-다양성	• 특정인 추론이 안된다고 해도 민감한 정보의 다양성을 높여 추론 가능성을 낮추는 기법	• 각 레코드는 최소 l개 이상의 다양성을 가지도록 하여 동질성 또는 배경지식 등에 의한 추론 방지
t-근접성	• ℓ-다양성 뿐만 아니라, 민감한 정보의 분포를 낮추어 추론 가능성을 더욱 낮추는 기법	• 전체 데이터 집합의 정보 분포와 특정 정보의 분포 차이를 t이하로 하여 추론 방지

[금융위원회(2016.06.09) ["개인정보 비식별 조치 가이드라인" 주요내용]에서 발췌, 이후 동.]

예제: k-익명성과 l-다양성

개체번호	Key변수 1	Key 변수 2	민감변수	k-익명성, 빈도	l-다양성
1	1	1	50	3	2
2	1	1	50	3	2
3	1	1	42	3	2
4	1	2	42	1	1
5	2	2	62	2	1
6	2	2	62	2	1

유일성에 기반한 방법(general)

표본과 모집단: f_k, F_k

표본과 모집단의 크기: n, N

노출위험을 계측함에 있어 침입자(intruder)가 모집단 F_k 의 정보가 있다는 가정을 함

F_k 가 알려져 있는 경우 $r_i = 1/F_k$.

F_k 가 알려지지 않은 경우

$$r_i = \mathrm{E}\left(\frac{1}{F_k}\middle| f_k\right) = \sum_{h \geq f_k} \frac{1}{h} \mathrm{P}\left(\mathbf{F_k = h}\middle|\mathbf{f_k}\right)$$

k-익명성: 모집단 = 표본

+ 기본 개념 2: 비식별화기법

처리기법	예시	세부기술
가명처리 (Pseudonymization)	• 홍길동, 35세, 서울 거주, 한국대 재학 → 임꺽정, 30대, 서울 거주, 국제대 재학	① 휴리스틱 가명화 ② 암호화 ③ 교환 방법
총계처리 (Aggregation)	• 임꺽정 180cm, 홍길동 170cm, 이콩쥐 160cm, 김팥쥐 150cm → 물리학과 학생 키 합 : 660cm, 평균키 165cm	④ 총계처리 ⑤ 부분총계 ⑥ 라운딩 ⑦ 재배열
데이터 삭제 (Data Reduction)	• 주민등록번호 901206-1234567 → 90년대 생, 남자 • 개인과 관련된 날짜정보(합격일 등)는 연단위로 처리	⑧ 식별자 삭제 ⑨ 식별자 부분삭제 ⑩ 레코드 삭제 ⑪ 식별요소 전부삭제
데이터 범주화 (Data Suppression)	• 홍길동, 35세 → 홍씨, 30~40세	⑫ 감추기 ⑬ 랜덤 라운딩 ⑭ 범위 방법 ⑮ 제어 라운딩
데이터 마스킹 (Data Masking)	• 홍길동, 35세, 서울 거주, 한국대 재학 → 홍○○, 35세, 서울 거주, ○○대학 재학	⑯ 임의 잡음 추가 ⑰ 공백과 대체

1. 재코딩(recording): 특정 변수의 범주를 상위 범주로 묶는 것.

연령을 5세 단위로 묶어 제공. 특정 값을 이상/이하로 묶기(top-down coding).소득을 100만원 단위 분류 그리고 상위 소득자의 범주를 1000만원 이상으로 코딩(top coding)

2. 국소감추기(local suppression): 특정 셀을 부분적으로 통합하는 방법
변수를 재-범주화 하지 않고 특정 셀을 주변 셀들과 통합.

3. 국소통합(micro-aggregation): k 개 이상의 개체들을 한 그룹으로 묶고 각 그룹의 개체 값들을 그룹의 평균값이나 중앙값으로 대체. 주로 연속형 자료에 적용.

고(중)차원 자료에 대한 k-익명성

K-익명성을 통한 개인정보 비식별화 방법은

식별.민감 변수의 수가 많은 경우

정보손실량이 매우 크다.

"On k-Anonymity and the Curse of Dimensionality" VLDB 2005,
By Charu C. Aggarwal [G-citation number 572]

예제 1: 고차원 자료의 k 익명화, 범주형 변수, 국소 감추기

자료 : German Credit Data

변수 수: 20개, 관측치 개수 : 1000개

변수 순서 : Rondom Forest 이용한 중요도 순

X1 : account balance (4 levels), X2 : credit history (5 levels)
X3 : Purpose (10 levels), X4 : Savings account (5 levels)
…

R sdcMicro 패키지의 local suppression함수를 이용한 k -익명화

(참조) 설명변수 중 연속형 변수인 경우 recording을 통해 범주형 변수로 변환한 후 k-익명화 작업을 진행

k=2, key 변수=X1~X5:
5개의 변수에 대하여 k-익명화 절차 진행

obs.No	X1	X2	X3	X4
1	<0DM	critical account	NA	no savings account
2	0<=...<200DM	credits paid back till now	radio/television	<100DM
3	no account	critical account	NA	<100DM
4	<0DM	credits paid back till now	furniture/equipment	<100DM
5	<0DM	delay in paying off	NA	<100DM
6	no account	credits paid back till now	NA	no savings account
7	no account	credits paid back till now	NA	500<=...<1000 DM
8	0<=...<200DM	credits paid back till now	used car	<100DM
9	no account	credits paid back till now	NA	>=1000DM
10	0<=...<200DM	critical account	new car	<100DM
11	0<=...<200DM	credits paid back till now	new car	<100DM
12	<0DM	credits paid back till now	business	<100DM
13	0<=...<200DM	credits paid back till now	radio/television	<100DM
14	<0DM	critical account	new car	<100DM
15	<0DM	credits paid back till now	new car	<100DM
16	<0DM	credits paid back till now	NA	100<=...<500DM
17	no account	critical account	radio/television	no savings account
18	<0DM	no credit taken	business	NA
19	0<=...<200DM	credits paid back till now	used car	<100DM
20	no account	credits paid back till now	radio/television	500<=...<1000 DM

k=5, key 변수=X1~X20:
20개의 변수에 모두 대하여 k-익명화 절차 진행

obs.No	X1	X2	X3	X4
1	<0DM	critical account	radio/television	no savings account
2	0<=...<200DM	credits paid back till now	NA	<100DM
3	NA	critical account	NA	<100DM
4	<0DM	credits paid back till now	furniture/equipment	<100DM
5	<0DM	NA	NA	<100DM
6	no account	credits paid back till now	NA	no savings account
7	no account	credits paid back till now	NA	500<=...<1000 DM
8	NA	credits paid back till now	used car	<100DM
9	no account	credits paid back till now	radio/television	NA
10	NA	critical account	NA	<100DM
11	0<=...<200DM	credits paid back till now	new car	NA
12	<0DM	credits paid back till now	NA	<100DM
13	0<=...<200DM	credits paid back till now	radio/television	<100DM
14	<0DM	critical account	NA	NA
15	<0DM	credits paid back till now	new car	<100DM
16	<0DM	credits paid back till now	NA	NA
17	no account	critical account	radio/television	no savings account
18	NA	NA	NA	NA
19	0<=...<200DM	credits paid back till now	NA	<100DM
20	no account	credits paid back till now	radio/television	500<=...<1000 DM

Summary on missing rates:
범주형 자료의 k-익명화

Key 변수	K	X1 결측치수	X2 결측치수	결측치가 있는 관측치수	전체 결측치 비중 (%)
X1-X5	k=2	1	33	177	0.9
	k=3	3	67	273	1.5
	k=5	3	125	373	2.4
X1-X10	k=2	102	151	754	5.7
	k=3	154	239	905	8.1
	k=5	232	349	978	11.3
X1-X15	k=2	154	157	825	9.0
	k=3	222	224	959	12.4
	k=5	333	325	1000	16.8
X1-X20	k=2	183	132	935	15.6
	k=3	222	209	1000	21.6
	k=5	310	279	1000	27.2

예제 2: 연속형 변수의 k-익명화, 국소통합(microaggregation) 방법

R sdcMicro 패키지의 microaggregation함수를 이용

자료:

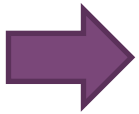
German credit data의 연속형 변수(3개)와 simulated data이용
- X1, X2, X3 : German credit data의 연속형 변수
  (X1 : duration, X2: credit amount, X3 : Age)
- X4, X5, …X20 는 독립인 [1,10]위의 uniform 난수

데이터 개수 : 1000개

k=10, key 변수 = X1~X20의 경우 비식별화 결과

원자료

obs. No	Du	CA	Age
1	6.0	1,169.0	67.0
2	48.0	5,951.0	22.0
3	12.0	2,096.0	49.0
4	42.0	7,882.0	45.0
5	24.0	4,870.0	53.0
6	36.0	9,055.0	35.0
7	24.0	2,835.0	53.0
8	36.0	6,948.0	35.0
9	12.0	3,059.0	61.0
10	30.0	5,234.0	28.0
11	12.0	1,295.0	25.0
12	48.0	4,308.0	24.0
13	12.0	1,567.0	22.0
14	24.0	1,199.0	60.0
15	15.0	1,403.0	28.0
16	24.0	1,282.0	32.0
17	24.0	2,424.0	53.0
18	30.0	8,072.0	25.0
19	24.0	12,579.0	44.0
20	24.0	3,430.0	31.0

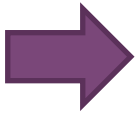


10-익명화 후 자료

obs. No	Du	CA	Age
1	16.4	1,703.4	50.0
2	47.3	7,027.0	30.5
3	15.7	2,314.4	39.1
4	37.2	6,317.5	32.7
5	39.0	4,185.2	47.6
6	15.7	2,314.4	39.1
7	17.2	1,754.2	42.3
8	27.0	4,561.5	26.1
9	13.0	2,293.3	59.3
10	24.3	5,059.7	39.1
11	14.1	1,821.5	31.4
12	48.6	5,711.0	29.5
13	14.7	2,174.4	28.6
14	24.0	2,931.5	47.0
15	16.8	2,169.0	32.2
16	19.4	3,346.7	30.2
17	23.4	4,554.7	31.7
18	43.2	10,447.9	30.1
19	19.2	9,100.5	50.8
20	16.6	2,512.8	32.5

원자료

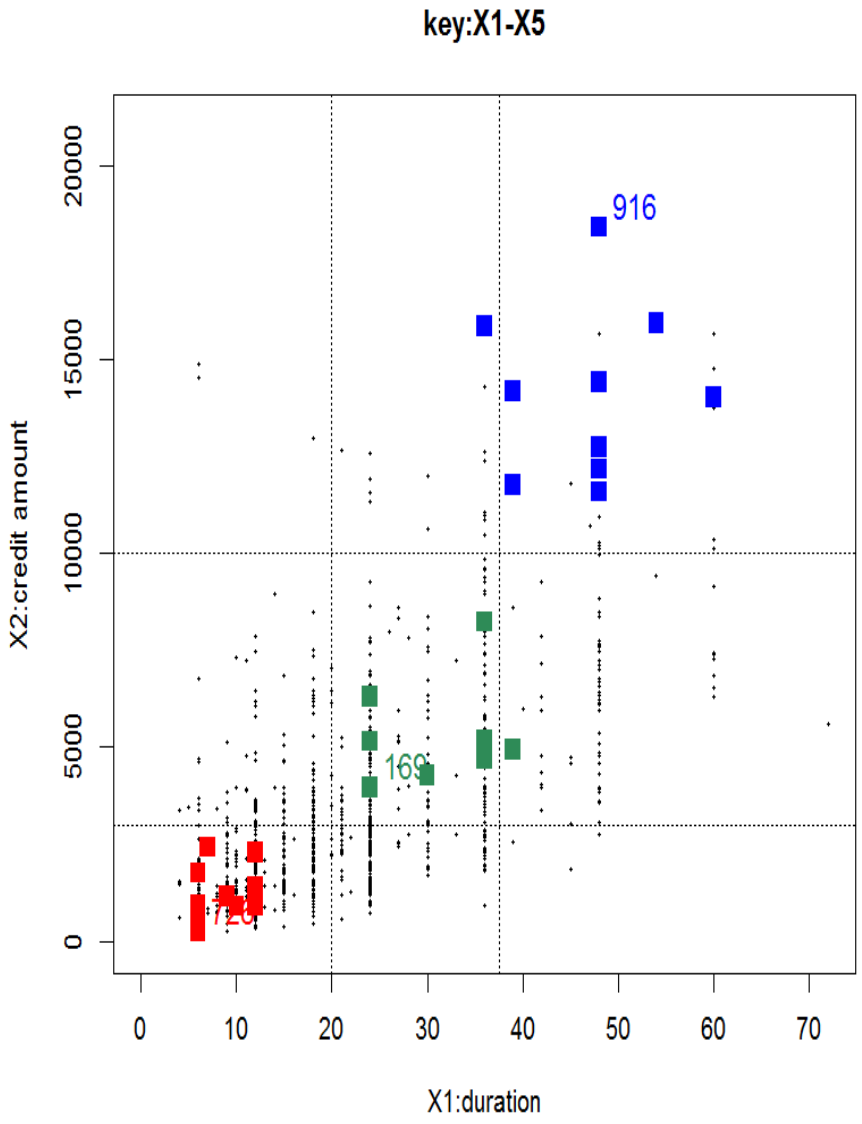
obs.No	Du	CA	Age
3	12.0	2,096.0	49.0
6	36.0	9,055.0	35.0
148	12.0	682.0	51.0
306	6.0	1,543.0	33.0
415	24.0	1,381.0	35.0
552	6.0	1,750.0	45.0
601	7.0	2,329.0	45.0
622	18.0	1,530.0	32.0
639	12.0	1,493.0	34.0
756	24.0	1,285.0	32.0
평균	15.7	2,314.4	39.1

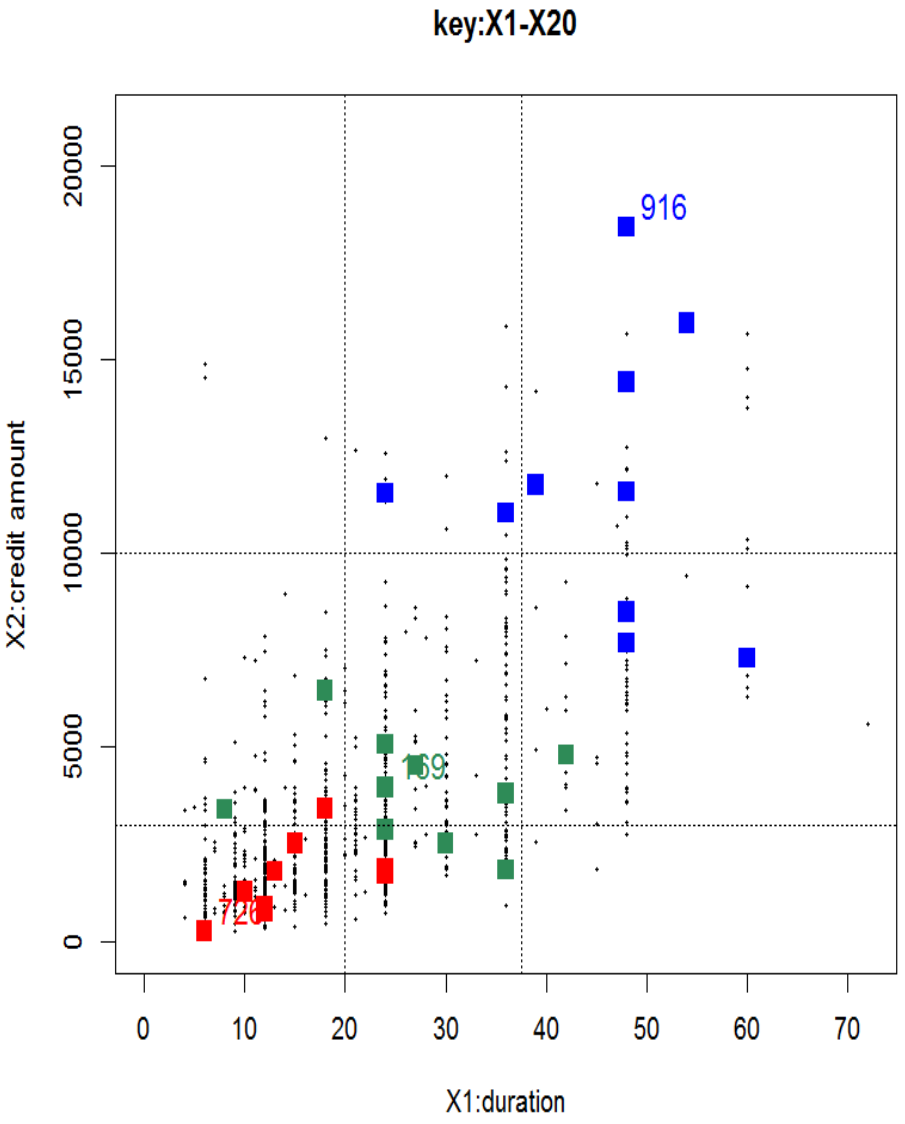


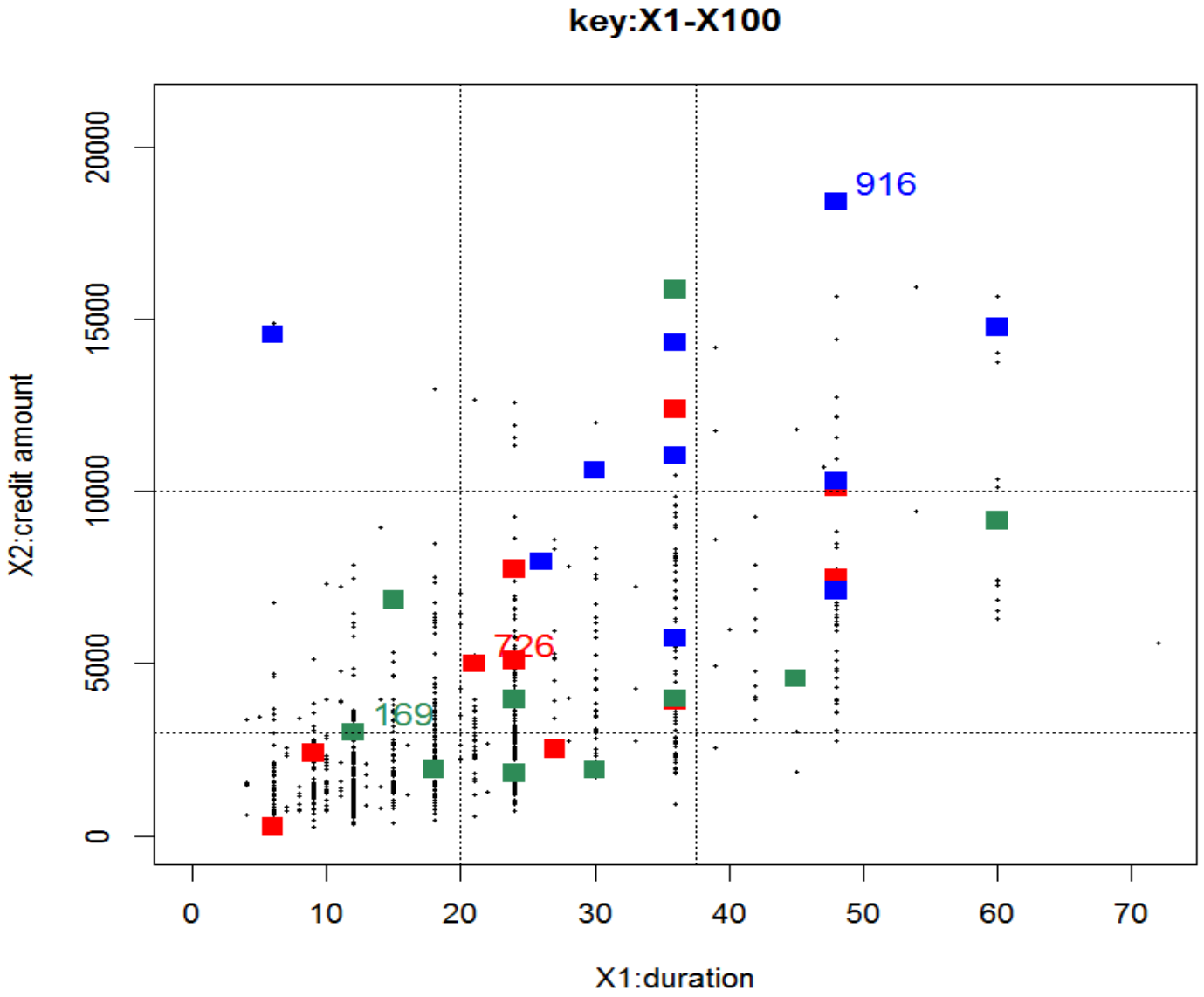
10-익명화 후 자료

obs.No	Du	CA	Age
3	15.7	2,314.4	39.1
6	15.7	2,314.4	39.1
148	15.7	2,314.4	39.1
306	15.7	2,314.4	39.1
415	15.7	2,314.4	39.1
552	15.7	2,314.4	39.1
601	15.7	2,314.4	39.1
622	15.7	2,314.4	39.1
639	15.7	2,314.4	39.1
756	15.7	2,314.4	39.1

(예시) k=10인 경우 관측치 169, 726, 916을 포함하는 군집 비교






key:X1-X100
X2:credit amount
X1:duration
0
5000
10000
15000
20000
0
10
20
30
40
50
60
70
916
726
169

3. 재현자료
(synthetic data)

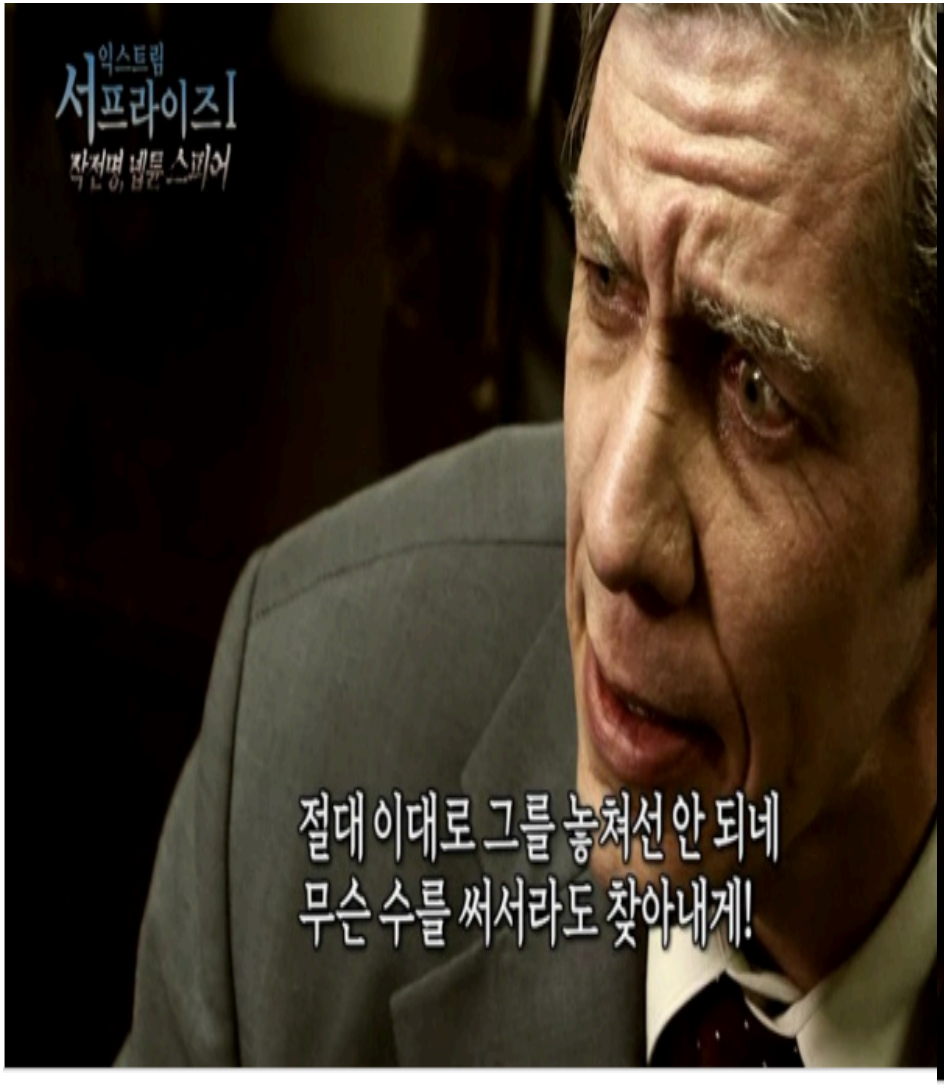
Synthetic Data 소개

학술 연구에서 제안된 방법론의 성능을 여러가지 시나리오를 가정한 가상자료(synthetic data, 재현자료)를 만들어 검증하는데서 착안

공개 또는 분석 대상이 되는 원자료와 통계적, 확률적 특성이 동일한 자료를 생성 이를 이용자에게 제공

재현의 정도에 따라 “부분-재현자료(partially synthetic data)”와 “전체-재현자료(fully synthetic data)”로 나누어짐

재현(synthetic): 시나리오=확률적분포, 등장인물=자료



정보의 손실을 줄이기 위하여 실제에서는 부분-재현자료, 즉 key-변수들과 민감변수들에 대하여만 synthetize 절차를 진행

부분-재현자료의 생성 절차

- $\mathbf{X}$는 원자료의 준식별.민감 변수들이고 $\mathbf{Y}$는 비식별변수
  $\mathbf{X}\big|\mathbf{Y}=\mathbf{y} \sim f\left(\mathbf{x}\big|\mathbf{Y}=\mathbf{y}\right)$.
- 원자료로 부터 $f\left(\mathbf{x}\big|\mathbf{Y}=\mathbf{y}\right)$를 추정: $\widehat{f}\left(\mathbf{x}\big|\mathbf{Y}=\mathbf{y}\right)$.
- $\widehat{f}\left(\mathbf{x}\big|\mathbf{Y}=\mathbf{y}\right)$를 이용하여 원자료와 같은 크기의 B개의 재현자료 $\left(\mathbf{X}^{(b)},\mathbf{Y}\right)$, $b=1,2,\ldots,B$를 생성함.

Synthetic Data 장점과 단점

장점:

1) 재현된 부분의 자료에는 원자료가 남아 있지 않아 개개인의 노출(re-identification)위험이 높지 않음. 비 민감정보의 존재로 여전히 노출위험은 존재함.

2) 정보의 유용성과 정보보호 능력 모두에서 전통적 방법에 대하여 우수 [여러 실험과 실증연구]

3) 다른 비식별화된 자료에 비하여 재현된 자료의 자료분석이 용이함.

단점:

자료를 재현할 때 사용되는 모형이 정확하여야 함.

특히 고차원(식별.민감변수의 수가 많은 경우) 참-모형을 특정함에 어려움이 있음.

즉, Synthetic data의 생성에서

- 원자료로 부터 $f(\mathbf{x}|\mathbf{Y}=\mathbf{y})$를 추정: $\hat{f}(\mathbf{x}|\mathbf{Y}=\mathbf{y})$.

이 어려움.

Synthetic data 사례

1. 해외 여러 공공기관에서 관심을 가지고 시도중이나 아직은 시범 단계임

2. SBB(SIPP Synthetic Beta):

   미국 Federal Privacy Council과 Census of Bureau

   Census of Bureau의 **Survey of Income and Program Participation(SIPP) 자료**와 과세 표준자료인 **Social Security Administration(SSA)/Internal Revenue Service(IRS)자료**와의 결합데이터 셋

재현자료 제공:

https://www.census.gov/programs-surveys/sipp/

3. 이 외에도 같은 Census of Bureau의 종단면경제활동자료, 종단면 고용가구 동적자료/독일의 고용 통계자료와 관련한 시범 사례들이 진행 중

Synthetic SIPP data 절차

데이터에 접근하기 위해서는 일정 형식으로 데이터를 요청하여 당국의 승인을 받아야 함

1. Application form 제출
   - 개인정보, 프로젝트 개관, 필요한 변수를 작성하여 제출
   - 필요한 변수는 SIPP 데이터 변수 설명(123개) 참조하여 작성

2. 승인여부 결정
   - 당국의 심사를 거쳐 5영업일 이내 통보

3. 승인이 되면 데이터 서버에 접속할 수 있는 개인 account 할당
   - 자료에 대한 접근만 가능하고 다운로드는 불가
   - 서버에서 SAS 및 Stata 활용하여 SSB data를 분석

4. 분석결과는 당국과의 공유 및 타당성 검증 필요
   - 당국은 이용자가 제공한 분석 코드(SAS 및 Stata code)를 활용하여 원자료를 분석한 결과를 이용자의 분석 결과와 비교하여 타당성 검증

Synthetic Data 예제: German Credit Data

- Data : German credit data(표본크기=1000)
  - 900개체 : training data
  - 100개체 : test data

- Synthetic variable
  - 반응변수 y (credit status : good, bad)
  - 연속형 설명변수 3개(duration, credit amount, age)

- Models for synthesis
  - 이산형 : f(y|나머지): logistic regression
  - 연속형 : f(C1|나머지), f(C2|나머지), f(C3|나머지): linear regression

Synthetic data

1. 900개의 training 데이터를 이용하여 synthesis 대상 변수에 대한 모델을 추정
2. 추정된 모델과 training 데이터를 이용하여 synthetic data set 생성
3. Synthetic data set을 이용하여 반응변수 y에 대한 모델(logistic regression)을 추정
4. 3 model과 original training data를 이용한 모델을 비교
   - Testing 데이터(100개)를 이용하여 오분류율 비교

Synthetic data : 예시

Original Training Data

obs.No	y	duration	credit.amount	age	account balance	Credit history	purpose	
1	good	6	1,169	67	<0 DM	critical account	radio/television	
2	bad	48	5,951	22	0<=...<200 DM	credits paid back till now	radio/television	
3	good	12	2,096	49	no account	critical account	education	
4	good	42	7,882	45	<0 DM	credits paid back till now	furniture/equipment	
6	good	36	9,055	35	no account	credits paid back till now	education	
7	good	24	2,835	53	no account	credits paid back till now	furniture/equipment	
8	good	36	6,948	35	0<=...<200 DM	credits paid back till now	used car	...
9	good	12	3,059	61	no account	credits paid back till now	radio/television	
10	bad	30	5,234	28	0<=...<200 DM	critical account	new car	
			...					

Synthetic Data

obs.No	y	duration	credit.amount	age	account balance	Credit history	purpose	
1	good	33	2,672	43	<0 DM	critical account	radio/television	
2	bad	32	4,751	39	0<=...<200 DM	credits paid back till now	radio/television	
3	good	10	839	46	no account	critical account	education	
4	good	41	4,753	23	<0 DM	credits paid back till now	furniture/ equipment	
6	good	24	8,606	31	no account	credits paid back till now	education	
7	good	9	2,957	46	no account	credits paid back till now	furniture/ equipment	
8	good	29	5,420	42	0<=...<200 DM	credits paid back till now	used car	...
9	good	32	3,685	41	no account	credits paid back till now	radio/television	
10	bad	19	3,407	33	0<=...<200 DM	critical account	new car	
	…							

Synthetic data

- Synthetic 데이터의 경우 생성한 100개 데이터 set의 평균, ( )내는 표준편차

혼동행렬(confusion matrix)

model	original data		synthetic data	
	true good	true bad	true good	true bad
pred. bad	56	15	55.87 (0.418)	14.34 (0.497)
pred. good	13	16	13.13 (0.418)	16.66 (0.497)

model	original data	synthetic data
오분류율	0.28	0.275

K-익명성의 대안으로서의 Synthetic Data

K-익명성과 비교하여

1. 고차원의 자료에 대한 어려움이 약하지만 여전히 존재

2. 데이터 센터의 심층이용자를 위한

   탐색적 자료로서의 물리적 비용 절감
   노출위험의 제어하에 결과값으로 제공가능

3. 일반이용자를 위한 공표자료로는 적합하여 보이지 않음

4. 차등정보보호 (Differential Privacy)

차등정보보호(Differential Privacy)

1. 정의

2. 차등정보보호를 보장하는 자료생성 기법:

   - 라플라스 기계(Laplance machine)
   - 평활히스토그램법

3. Local differential privacy

4. 활용사례

5. 개인정보비식별화를 위한 차등정보보호

차등정보보호 개념과 정의

개인정보 보호의 목적:

1. 어느 한 개인(나)의 자료가 공표되는 자료(Q)에 영향을 주지 않음

$$\mathbf{Q}\left(\mathbf{D}_{\text{all}}\right) = \mathbf{Q}\left(\mathbf{D}_{\text{all}-\text{me}}\right)$$

2. 공표자료 R의 공표여부가 외부침입자가 나의 기록을 찾을 확률을 변화시키지 않음

$$\mathrm{P}\left(\text{find me}\middle|\mathbf{R}\right) = \mathrm{P}\left(\text{find me}\right)$$

차등정보보호 정의

- $\kappa(\mathbf{X})$는 원자료 $\mathbf{X}$에 대하여 $\mathbf{Y}$를 결과 값으로 갖는 랜덤화 함수(randomized function)이다. 여기서 $\mathbf{Y}$는 익명화된 자료, 로지스틱 회귀계수와 같은 모집단의 모수 추정값등 공표되는 모든 값이 될 수 있다.

- $\kappa(\mathbf{X})$가 $\epsilon-$차등정보보호($\epsilon-$differential private)라 함은 다음으로 정의된다.

  오직 한 개의 개체(한 명의 개인)정보만 다른 두 데이터베이스 $\mathbf{X}_1$과 $\mathbf{X}_2$에 가 **임의의** $S \subset \mathrm{range}(\kappa)$에 대하여

$$\log\left(\frac{\mathrm{P}(\kappa(\mathbf{X_1}) \in S)}{\mathrm{P}(\kappa(\mathbf{X_2}) \in S)}\right) \leq \epsilon.$$

다음과 동치

$$\exp(-\epsilon) \leq \frac{\mathrm{P}(\kappa(\mathbf{X_1}) \in S)}{\mathrm{P}(\kappa(\mathbf{X_2}) \in S)} \leq \exp(\epsilon)$$

특정 개인 A가 포함된 원자료 $\mathbf{X}_1$ 과 포함되지 않은 원자료 $\mathbf{X}_2$로 부터 생성된 공표자료들, 즉 $\kappa(\mathbf{X}_1)$ 과 $\kappa(\mathbf{X}_2)$ 사이에 확률적으로 유의한 차이가 없음

즉, 한 명의 정보가 원자료(데이터베이스)에 포함되어 있는지의 여부가 공표 자료와 그 분석 결과에 유의한 영향을 미치지 못함.

차등정보보호 절차는:

위의 성질이 만족되도록 결과값에 NOISE를 추가하는 작업
NOISE의 종류와 크기는 결과값의 형태에 의존함.

예제 1: 라플라스 기계

- 원자료 $\mathbf{X}$에 대하여 $d-$차원 실수값을 갖는 함수 $f(\mathbf{X})$(공표할 대상)이 주어짐.
- 주어진 f에 대하여 "global sensitivity (GS)"

$$GS(f) = \sup_{\mathbf{x}_1 \sim \mathbf{x}_2} \left\| f(\mathbf{X}_1) - f(\mathbf{X}_2) \right\|_1$$

  를 정의함.
- 서로 독립인 모수가 $b = GS(f)/\epsilon$인 d개의 라플라스 난수,

$$p(x; b) = \frac{1}{2b} \exp\{ -|x|/b \},$$

  를 생성함. $\mathbf{W} = (W_1, W_2, \ldots, W_d)^{\mathrm{T}}$.
- $\epsilon-$차등정보보호 함수(절차) $\kappa(\mathbf{X})$는

$$\kappa(\mathbf{X}) = f(\mathbf{X}) + \mathbf{W}.$$

더해지는 라플라스 노이즈의 분산이 $2GS(f)^2/2$이고 이는 (1) 함수 민감도가 높은 f는 보다 큰 노이즈를 필요로 함, (2) 높은 수준의 보안(작은 ϵ)을 위하여도 분산이 큰 노이즈를 추가하여야 함.

자료의 평균:

예제 2: 평활 히스토그램:

- p개의 항목에 대하여 n명의 개인에 대하여 기록한 원자료 $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$.
- 각 $\mathbf{X}_i$가 IID $f(\mathbf{x})$를 따름. 편의상 기록치들은 $[0,1]^p$의 값을 가짐을 가정.
- 적절한 수준의 δ에 대하여

$$\widehat{f}_\delta(\mathbf{x}) = (1-\delta)\widehat{f}(\mathbf{x}) + \delta$$

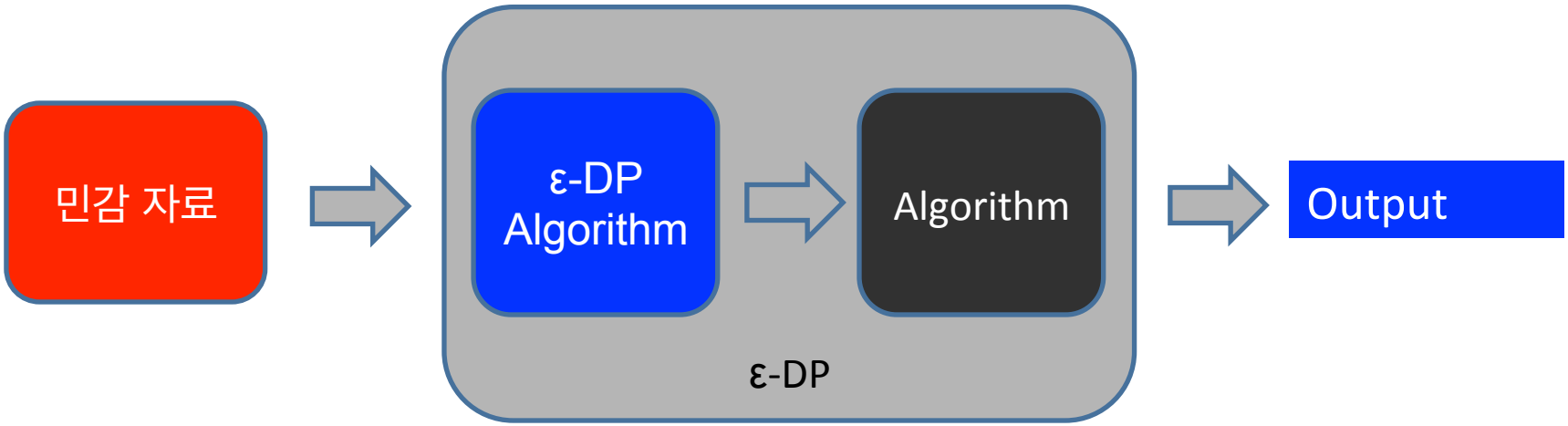
를 계산 이로 부터 n^*개의 난수를 생성 공표값으로 함.

- 위에서 함수 추정에서 δ는 $p-$차원 uniform난수의 분포이고 $\widehat{f}(x)$ 는 원자료를 이용한 $f(x)$의 추정량임.
- δ**의 결정**:

$$n^* \log\left(\frac{(1-\delta)p}{n\delta} + 1\right) \leq \epsilon.$$

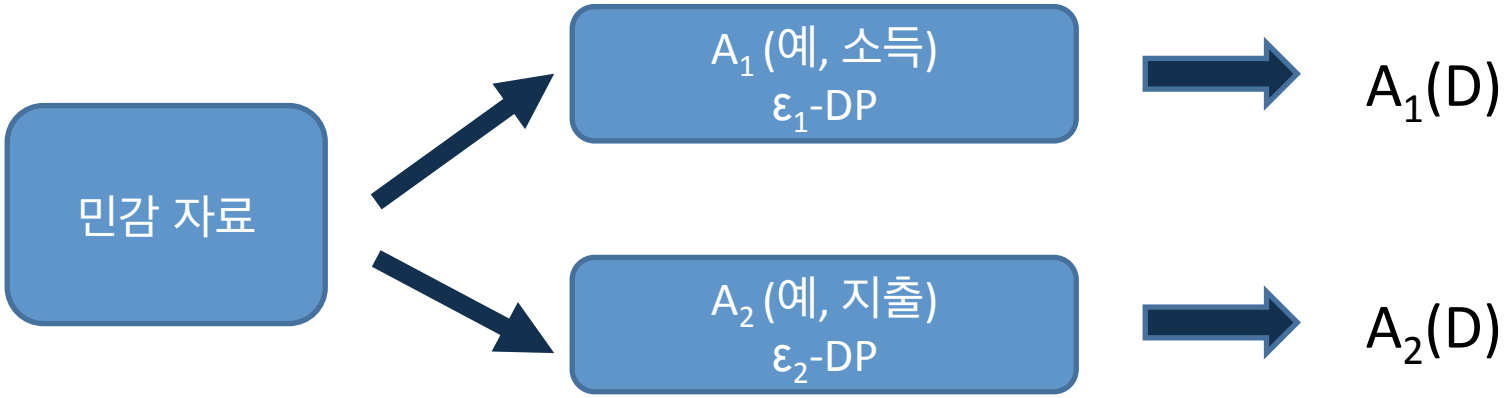
차등정보보호의 성질

Post-processing Invariance:



[그림은 K. Chaudhuri와 A.D. Sarwate의 슬라이드에서 따옴]

Composition:



If A_1 is ϵ_1-DP and A_2 is ϵ_2-DP, then the union $(A_1(D), A_2(D))$ is $(\epsilon_1 + \epsilon_2)$-DP

자료의 세로합 가로합 모두에 적용이 됨

[그림은 K. Chaudhuri와 A.D. Sarwate의 슬라이드에서 따옴]

Dependency to $\kappa(\mathbf{X})$:

차등정보절차가 공표되는 자료 또는 결과값의 형태 $\kappa(\mathbf{X})$

에 의존한다.

결과값의 형태 예로는 평균, 히스토그램, 분류자(classifier), 분산 등 매우 다양하게 고려되었다.

Local Differential Privacy

DP(차등정보보호)는 central의 존재를 가정하고
결과값에 잡음을 더하여 개인정보를 보호하는 절차임.

Epsilon ϵ 에 의하여 계측되는 노출위험을 제어하기
위한 noise의 종류와 양을 정하는 절차.

Local DP는 central로 부터 자료제공자의 개인정보보호를
위하여 적절한 noise가 더해진 자료를 central에 제공.

예제 3: randomized response와 비율의 추정

Randomized response:

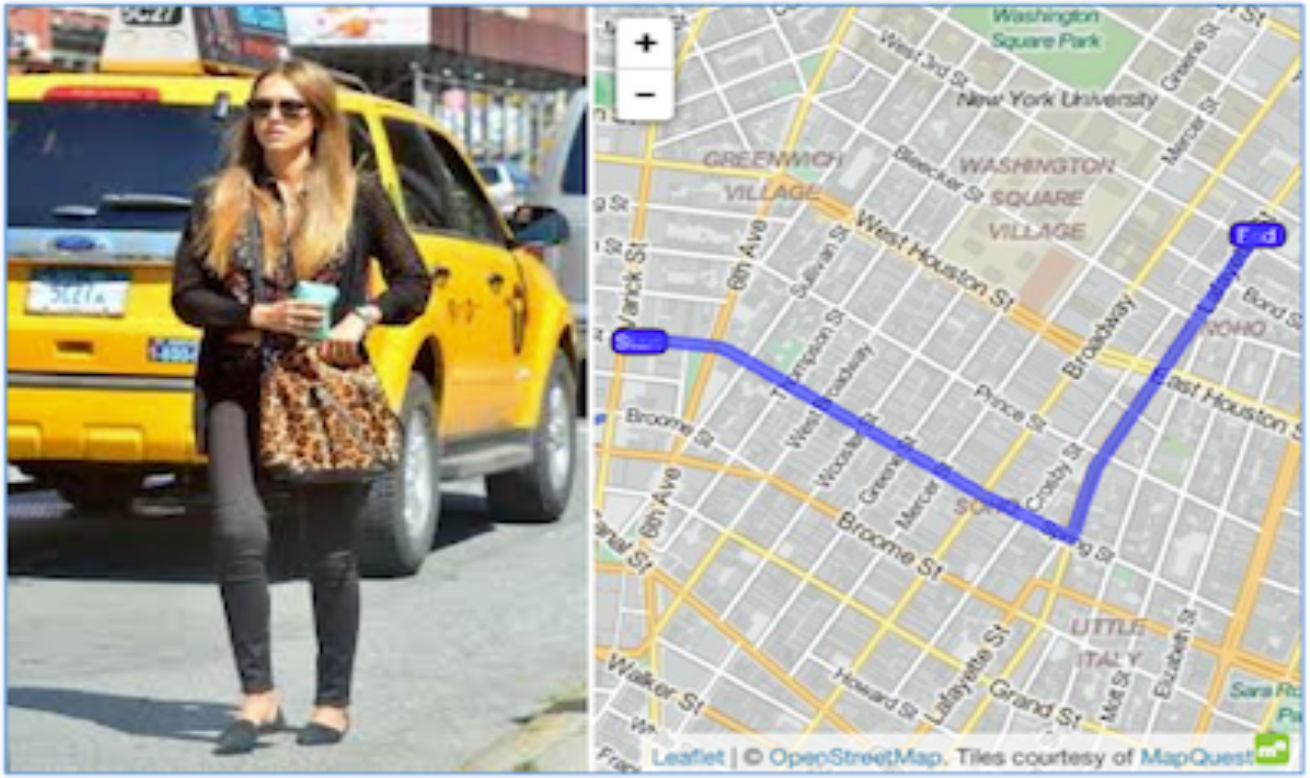
개인 응답자: 민감함 질문에 대하여 동전을 던져 앞면이 나오면 진실을/ 뒷면이 나오면 O/X중 임의로 선택된 답을 제공.

잡음이 추가된 자료로 부터 O에 대한 모비율을 정확히 추정가능.

Google사의 **Chrome** 브라우저: 크롬사용자들의 "예/아니오"형태의 자료에 차등정보보호 잡음이 추가된 상태로 구글 데이터베이스에 저장하고 구글 내부 이용자들은 응답에 대한 실제비율이 아닌 비율에 대한 추정량과 분산추정량을 이용.

예제 4: NYC taxicab data set

Riding with the stars, from research.neustar.biz



Jessica Alba (Click to Explore)

2013 NY taxicab dataset : pickup and drop off times, locations, fare and tip amounts를 익명화하여 제공

해커들 de-anonymize, 개별 기사의 연간 수입 계산 가능 (Driver privacy)

보조정보가 있으면 승객의 privacy 노출 가능 (Passenger privacy in the NYC taxicab dataset). 보조정보: the picture, some information from celebrity gossip blogs

```sql
SELECT D.dropoff_latitude, D.dropoff_longitude, F.total_amount, F.tip_amount
FROM tripData AS D, tripFare AS F
WHERE D.hack_license = F.hack_license AND D.pickup_datetime = F.pickup_datetime
  AND pickup_datetime > "2013-07-08 19:33:00" AND pickup_datetime < "2013-07-08 19:37:00"
  AND pickup_latitude > 40.719 AND pickup_latitude < 40.7204
  AND pickup_longitude > -74.0106 AND pickup_longitude < -74.01;
```

Jessica Alba의 노출정보

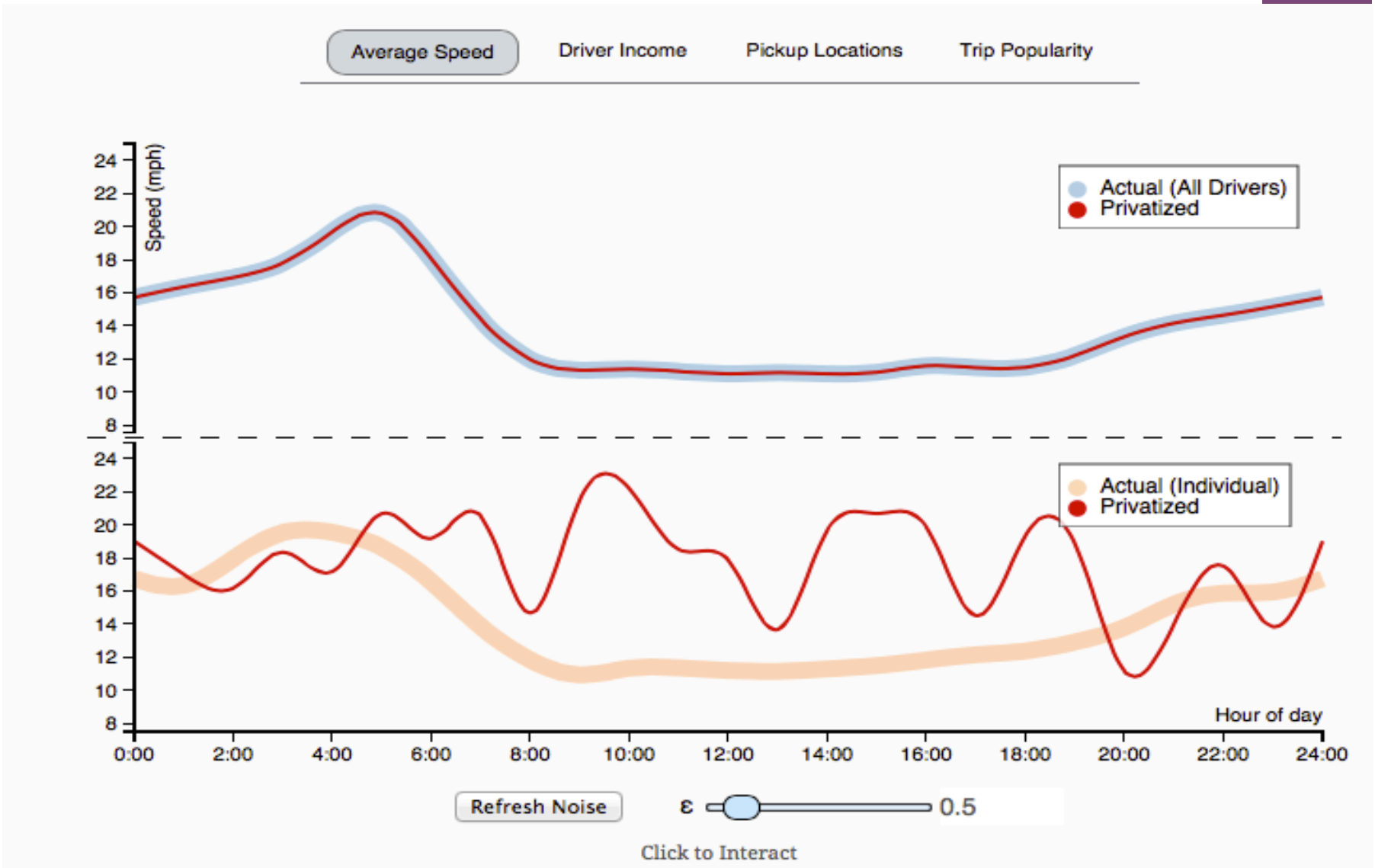
승하차 위치, 시간, 요금($9), 팁($0)

Differentially Privatized Trip (drop-off)

Total Fare:
$25 - $30

Tip Amount:
$6 - $8

NYC taxicab data: local DP



K-익명성의 대안으로서의 차등정보보호

Differentially Privatized Synthetic Data:

1. 심층이용자의 분석툴이 정하여져 있는경우 테이터센터의 연구용DB의 Local DP를 이용한 개인정보비식별화 절차

2. 민감정보에 대하여 개인들이 중앙에 LDP자료 제공.
중앙은 도표/평균/회귀식 추정등 특별한 목적으로만 제공된 자료 사용가능

3. 일반이용자를 위한 도표들을 위한 DP/local DP procedure 가능

5. 정리와 제언

정리 – 개인정보 비식별화

1. 개인정보보호에 있어 심층분석자와 일반분석자의 분리가 필요.

2. 심층분석자를 위한 데이터센터
   - 데이터센터 내에서 자료의 분석
   - 데이터센터 외부로 나오는 통합자료/결과물에 대한 개인정보 비식별화조치

3. 개인정보 비식별화조치에 따른 정보손실 불가피. 특히 고차원자료의 경우 정보손실이 큼. 특히 K-익명성의 규제 편의성과 정보손실량

4. 재현자료, 차등정보보호등 대안적 비식별화방법의 사용을 통한 시스템의 유연성 확보 필요

정리-데이터센터

5. 데이터센터의 체계

물리적 이동을 통한 제한된 영역에서의 분석

재현자료/차등정보보호등을 이용한 탐색적 자료분석
- 데이터 센터/심층이용자 모두의 비용절감

차등정보보호된 자료의 제공을 통한 분석의 유연성 확보

Thank you